Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 9, 24–52 DOI: 10.5604/01.3001.0015.9862
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 20.12.2021, poprawiony/revised: 23.05.2022, zaakceptowany/accepted: 22.08.2022

Wpływ pandemii COVID-19 na stan zdrowia psychicznego społeczeństwa[1]

Aneta Ptak-Chmielewska[a], Karolina Baszniak[b], Jarosław Kurpanik[c]

Streszczenie. Pandemia COVID-19 odmieniła życie ludzi na całym świecie, m.in. wpłynęła na kondycję psychiczną i funkcjonowanie wielu rodzin. Głównym celem badania omawianego w artykule jest ocena wpływu pandemii COVID-19 na stan zdrowia psychicznego członków gospodarstw domowych. W badaniu posłużono się zbiorem danych pochodzących z ankiety *COVID Impact Survey*, przeprowadzonej w 2020 r. (w trakcie pierwszej fali pandemii) w Stanach Zjednoczonych wśród osób dorosłych przez organizację Data Foundation. Analizie poddano 6768 obserwacji. Oszacowano model regresji logistycznej oraz modele oparte na metodach data mining, takich jak: drzewa decyzyjne, wzmacnianie gradientowe, metoda *k*-najbliższych sąsiadów, sztuczne sieci neuronowe i metoda wektorów wspierających. Analiza skupień pozwoliła podzielić respondentów na grupy uwidaczniające cechy charakterystyczne i problemy członków gospodarstw domowych, a w utworzonym modelu uwzględniono kwestie zdrowia i zaburzeń psychicznych oraz ich związek z sytuacją finansową gospodarstw. Wyniki badania wskazują na to, że izolacja, zdalny tryb nauczania i pracy oraz mniejsza aktywność fizyczna przyczyniają się do pogarszania się stanu zdrowia psychicznego.

Słowa kluczowe: uczenie maszynowe, pandemia COVID-19, data mining, stan zdrowia psychicznego, gospodarstwo domowe, Stany Zjednoczone

JEL: C1, C45, I12

Influence of the COVID-19 pandemic on the mental health of society

Abstract. The COVID-19 pandemic changed the lives of people all over the world, by e.g. affecting the mental health and the functioning of many families. The main goal of the research presented in this paper is to assess the influence of the COVID-19 pandemic on the mental health of members of households. The research was performed on the basis of a data set from the *COVID Impact Survey* carried out by the Data Foundation think tank in 2020 (during the first wave of the COVID-19 pandemic) in the USA among adult respondents. The survey used 6,768

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na konferencji Multivariate Statistical Analysis MSA 2021, która odbyła się w dniach 8–10 listopada 2021 r. w Łodzi. / The article is based on a paper delivered at the Multivariate Statistical Analysis MSA 2021 conference, held on 8–10 November 2021 in Łódź.

[a] Szkoła Główna Handlowa w Warszawie, Instytut Statystyki i Demografii, Polska / SGH Warsaw School of Economics, Institute of Statistics and Demography, Poland.
ORCID: https://orcid.org/0000-0002-9896-4240. Autor korespondencyjny / Corresponding author, e-mail: aptak@sgh.waw.pl.

[b] ING Tech Poland. ORCID: https://orcid.org/0000-0003-3079-3526. E-mail: karolina.baszniak@ing.com.

[c] Uniwersytet Ekonomiczny w Katowicach, Wydział Informatyki i Komunikacji, Polska / University of Economics in Katowice, Faculty of Informatics and Communication, Poland.
ORCID: https://orcid.org/0000-0002-4288-4647. E-mail: jarosław.kurpanik@ue.katowice.pl.

observations. The authors estimated a model of logistic regression and models based on data mining methods, such as decision trees, XGBoost, the *k*-nearest neighbours method, artificial neural networks and a support vector machine. Cluster analysis allowed the division of the respondents into groups showing their characteristic features and problems, and the constructed model took into account their mental health issues and the relationship between those issues and the financial situation of households. The results demonstrate that isolation, remote education and work, and limited physical activity contribute to the worsening of the mental health of the population.

Keywords: machine learning, COVID-19 pandemic, data mining, mental health, household, USA

1. Wprowadzenie

Pandemia COVID-19, nazywana również pandemią koronawirusa, to globalna epidemia choroby zakaźnej wywoływanej przez koronawirusa SARS-CoV-2. Pierwsze przypadki zakażenia odnotowano w grudniu 2019 r. w chińskim mieście Wuhan, a 20 stycznia 2020 r. Światowa Organizacja Zdrowia (World Health Organization – WHO) ogłosiła wybuch epidemii – stanu zagrożenia zdrowia publicznego o zasięgu międzynarodowym (WHO, b.r.).

Pandemia COVID-19 odmieniła życie ludzi na całym świecie. Nie tylko wycisnęła piętno na zdrowiu i życiu wielu osób, lecz także wpłynęła na gospodarkę, kulturę, politykę, edukację, środowisko czy styl życia. Wybuch pandemii stał się zagrożeniem dla stabilności światowych gospodarek – prognozuje się, że pozostaną niestabilne aż do momentu wyraźnego poprawienia się wyników finansowych podmiotów gospodarczych (Scope Ratings, 2020).

Wśród sektorów gospodarki najbardziej dotkniętych wybuchem pandemii należy wymienić:

- branżę turystyczną – ucierpiała z powodu wprowadzania zakazów podróżowania oraz zamykania miejsc publicznych i atrakcji turystycznych (Turner, 2020);
- sektor detaliczny – na jego kondycję ekonomiczną wpłynęło skrócenie godzin otwarcia sklepów, a nawet ich całkowite zamknięcie. Według raportu Yelp (2020) ok. 60% amerykańskich firm, które zostały zamknięte od początku pandemii, nie wznowi już działalności. Wybuch epidemii podawany był także jako przyczyna braków w zaopatrzeniu. Wynikało to m.in. z paniki zakupowej, prowadzącej do opróżniania półek sklepowych z artykułów spożywczych oraz do zakłóceń w operacjach fabrycznych czy logistycznych (Tyko i in., 2020);
- branżę kulturalną i sektor edukacyjny – muzea, kina, teatry i inne instytucje kulturalne na całym świecie zostały tymczasowo zamknięte, a wystawy i koncerty – odwołane lub przełożone. Wiele instytucji i placówek edukacyjnych podjęło próbę świadczenia usług za pośrednictwem platform cyfrowych.

Pandemia miała negatywny wpływ na zdrowie ludzi w wielu aspektach, m.in. doprowadziła do ogólnego spadku liczby wizyt u lekarzy. Szacuje się, że w Stanach Zjednoczonych liczba konsultacji z powodu symptomów zawału serca zmniejszyła

się o 38%, a w Hiszpanii – o 40% (World Food Programme, 2021). Obostrzenia pandemiczne negatywnie oddziaływały również na zdrowie psychiczne ludzi na całym świecie. Dystans społeczny spotęgował poczucie samotności, przyczynił się do popadania w depresję, a w wyniku braku możliwości wyjścia z domu – nasilił przemoc domową (Surkova i in., 2020). Z badań ankietowych wynika, że w czerwcu 2020 r. 40% dorosłych Amerykanów doświadczało zaburzeń zdrowia psychicznego, a 11% poważnie rozważało podjęcie próby samobójstwa w ciągu ostatniego miesiąca (Czeisler i in., 2020).

Głównym celem badania omawianego w artykule jest ocena wpływu pandemii COVID-19 na stan zdrowia psychicznego członków gospodarstw domowych.

2. Badania społecznych i gospodarczych skutków pandemii COVID-19

Społeczne i gospodarcze następstwa pandemii są przedmiotem wielu prac badawczych. W Europie m.in. European Foundation for the Improvement of Living and Working Conditions (Eurofound) prowadzi badanie ankietowe *Living, working and COVID-19*, dotyczące doświadczeń z życia i pracy w czasie pandemii, a jego celem jest wsparcie decydentów w wyprowadzeniu krajów z kryzysu. Zagregowane dane interaktywne z tego badania są publikowane na stronie Eurofound[2]. Można je przeglądać w różnych zestawieniach i dla różnych krajów.

W lutym i marcu 2021 r. przeprowadzono trzecią rundę tego badania, a podsumowanie jej wyników opublikowano w raporcie *Living, working and COVID-19 (Update April 2021): Mental health and trust decline across EU as pandemic enters another year* (Eurofound, 2021). Przedstawiono w nim sytuację społeczną i ekonomiczną mieszkańców Europy podczas pierwszego roku życia zgodnie z restrykcjami wprowadzonymi po wybuchu pandemii COVID-19. Przeanalizowano podstawowe wnioski oraz prześledzono rozwój sytuacji w 27 krajach Unii Europejskiej od rozpoczęcia badania w kwietniu 2020 r. Skupiono się na problemach, z którymi zmagają się społeczeństwa w trakcie pandemii, jak np.: wzrost niepewności na rynku pracy wynikający z większego ryzyka jej utraty, spadek poziomu dobrobytu, pogorszenie w zakresie zapewnienia równości płci, spadek zaufania do instytucji publicznych, rosnące trudności z zachowaniem równowagi pomiędzy pracą a życiem prywatnym oraz wzrost nieufności wobec szczepień. Wyniki przedstawione w raporcie wskazują na potrzebę bardziej holistycznego podejścia do wsparcia tych grup ludności, które najbardziej ucierpiały w wyniku kryzysu.

Innym badaniem ankietowym, którego wyniki zostały wykorzystane w badaniu omawianym w artykule, jest *COVID Impact Survey*, przeprowadzone w Stanach

[2] Zob. https://www.eurofound.europa.eu/data/covid-19.

Zjednoczonych przez niezależną pozarządową organizację non profit Data Foundation. W związku z ograniczoną infrastrukturą monitorowania zdrowia Data Foundation zdecydowała się na przeprowadzenie statystycznie wiarygodnego badania ankietowego poruszającego tematykę zdrowia fizycznego i psychicznego, bezpieczeństwa ekonomicznego i żywnościowego oraz zatrudnienia. Zostało ono opracowane przy wsparciu czołowych krajowych ekspertów w dziedzinie zdrowia publicznego, ekonomii i nauk społecznych (Wozniak i in., 2020) w celu przekazania płynącej z niego wiedzy decydentom i amerykańskiej opinii publicznej, jak również zachęcenia rządu do wykonywania podobnych analiz na większą skalę.

Ankietę przeprowadzono na reprezentatywnej próbie losowej[3] z kilkunastu stanów i obszarów metropolitalnych. Wyniki dostarczają informacji na temat doświadczeń amerykańskiego społeczeństwa związanych z pandemią COVID-19.

Z kolei *Household Pulse Survey* to badanie przeprowadzone przez agencję rządową Stanów Zjednoczonych (United States Census Bureau) we współpracy z agencjami federalnymi, które dostarczyło danych na temat społecznego i ekonomicznego oddziaływania pandemii koronawirusa na amerykańskie gospodarstwa domowe. Próba biorąca udział w ankiecie została wybrana losowo, z zachowaniem reprezentatywności dla całego kraju. Kwestionariusz był krótki; respondenci otrzymywali go e-mailowo. Wyniki ankiety udostępniono zarówno w postaci danych jednostkowych, umożliwiających dalszą analizę przez badaczy, jak i w formie interaktywnego pulpitu nawigującego, pozwalającego na zapoznanie się z zagregowanymi wynikami zaprezentowanymi na wykresach i w tabelach.

W czerwcu 2020 r. Chetty i in. opracowali, na podstawie danych z sektora prywatnego, raport o ekonomicznych skutkach pandemii COVID-19 w Stanach Zjednoczonych pt. *The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data*. Badacze zamieścili w nim ogólnodostępną bazę danych, które w czasie rzeczywistym pokazują aktywność gospodarczą prywatnych firm. Publikowane statystyki dotyczą m.in. wydatków konsumentów, sytuacji biznesu i kluczowych wskaźników gospodarczych.

W grudniu 2020 r. ukazał się raport z badania wzdłużnego trzech holenderskich kohort kontrolnych na temat wpływu pandemii COVID-19 na zdrowie psychiczne ludzi z depresją, zaburzeniami lękowymi i obsesyjno-kompulsyjnymi oraz bez nich autorstwa badaczy związanych z międzynarodowym dziennikiem medycznym „The Lancet. Psychiatry". W kwietniu i maju 2020 r. badacze rozsyłali kwestionariusze internetowe do osób z zaburzeniami psychicznymi i do ludzi zdrowych, zawierające pytania na temat postrzegania wpływu pandemii na zdrowie psychiczne oraz radzenia sobie z lękiem przed infekcją, a także cztery zwalidowane skale oceniające wystę-

[3] Więcej informacji na stronie: https://www.covid-impact.org/about-the-survey-questionnaire.

powanie objawów depresji, zaburzeń lękowych, zmartwień i samotności używane we wcześniejszych badaniach. Liczbę i przewlekłość zaburzeń określono na podstawie diagnoz z poprzednich lat przy użyciu regresji logistycznej i modeli mieszanych.

Badanie wykazało, że w czasie pandemii COVID-19 u osób bez zaburzeń depresyjnych, lękowych lub obsesyjno-kompulsyjnych wystąpił większy wzrost częstości objawów depresji, zaburzeń lękowych, zmartwień i samotności niż u osób z zaburzeniami psychicznymi. W przypadku osób z największym obciążeniem psychicznym zauważono nawet niewielki spadek częstości występowania objawów (Pan i in., 2021).

Badania skutków pandemii COVID-19 są niezwykle istotne dla decydentów opracowujących strategię walki z jej negatywnymi skutkami oraz dla wszystkich, którym zrozumienie zmian, jakie przyniosły wydarzenia ostatnich dwóch lat, może ułatwić efektywne funkcjonowanie w życiu prywatnym i zawodowym.

3. Metody wykorzystane do analizy skutków pandemii COVID-19

W celu przestudiowania głównych problemów zdrowotnych, z jakimi zmagali się członkowie gospodarstw domowych w czasie pierwszej fali pandemii, wyniki ankiety *COVID Impact Survey* (COVID Impact Survey, b.r.) poddano analizie metodami statystycznymi (regresja logistyczna) oraz data mining.

Jednym z głównych skutków pandemii jest pogarszający się stan zdrowia psychicznego osób dotkniętych jej następstwami. Przy użyciu modelu predykcyjnego zidentyfikowano osoby, u których z dużym prawdopodobieństwem zostaną zdiagnozowane problemy psychiczne. Ma to bardzo duże znaczenie, ponieważ kondycja psychiczna jest nie mniej ważnym składnikiem ogólnego stanu zdrowia niż kondycja fizyczna. Choroby psychiczne, zwłaszcza depresja, zwiększają ryzyko wystąpienia wielu innych chorób, szczególnie przewlekłych, takich jak udar, cukrzyca typu II i choroby serca (National Institute of Mental Health, 2021). Szacunki sprzed pandemii COVID-19 pokazują, że z powodu samej depresji i lęków świat traci rocznie prawie bilion dolarów w obszarze produktywności ekonomicznej, a wyniki badań pozwalają stwierdzić, że każdy dolar wydany na rzetelną opiekę nad depresją i lękami zwróci się w przyszłości pięciokrotnie (WHO, 2020).

Badanie *COVID Impact Survey*, przeprowadzone wśród osób w wieku 18 lat i więcej, zostało oparte na danych z bazy National Survey Data wykorzystywanej do badań o zasięgu narodowym, reprezentacyjnych dla całej populacji gospodarstw domowych Stanów Zjednoczonych[4]. Podstawą konstrukcji próby reprezentatywnej stały się dane adresowe z bazy U.S. Postal Service (USPS) zawarte w Delivery

[4] Więcej informacji na stronie covid-impact.org.

Sequence File (DSF), uwzględniającej ok. 97% gospodarstw domowych w Stanach. Wyniki były ważone zmiennymi demograficznymi z *Current Population Survey 2020*.

Jednym z najistotniejszych kroków podczas budowania modelu i analizowania wpływu określonych czynników na zmienną objaśnianą jest wybór zmiennych objaśniających. Aby zapewnić jak najlepszą jakość modelu, na podstawie literatury przedmiotu dokonano statystycznego doboru zmiennych z uwzględnieniem zależności zmiennych objaśniających od zmiennej objaśnianej oraz wyeliminowano zmienne objaśniające skorelowane ze sobą.

Początkowo zbiór składał się z 7505 obserwacji. Po usunięciu brakujących wartości (odpowiedź „nie wiem"), przypadków pominięcia pytań lub odmowy odpowiedzi ostatecznie liczył 6655 obserwacji. Dodatkowo został podzielony na zbiory uczący (treningowy) i testowy w proporcji 75% do 25%[5] (losowanie warstwowe z warstwowaniem po zmiennej zależnej). Z tego względu w zbiorze treningowym znalazło się 4991 obserwacji, a w testowym – 1664.

Zmienną objaśnianą jest zmienna binarna PHYS3H, określająca, czy dana osoba ma zdiagnozowaną chorobę o podłożu psychicznym. Zmienne objaśniające zostały przedstawione w zestawieniu 1. Przyjęcie zmiennej objaśnianej PHYS3H jest ograniczeniem przeprowadzonego badania, ponieważ dokładna data diagnozy jest nieznana (mogła zostać postawiona jeszcze przed wybuchem pandemii), niemniej jednak wyniki analizy dostarczają cennych informacji, i to nie tylko z punktu widzenia polityki zdrowotnej.

Zestawienie 1. Zmienne objaśniające użyte do budowy modelu

Zmienne	Opis	Wartości	Typ
PHYS8	Czy swój stan zdrowia oceniłbyś jako: wyśmienity, bardzo dobry, dobry, umiarkowany czy zły? (American Psychiatric Association, 2019)	(1) Wyśmienity (2) Bardzo dobry (3) Dobry (4) Umiarkowany (5) Zły	porządkowa
PHYS1K	Czy w ciągu ostatnich siedmiu dni doświadczyłeś uczucia zmęczenia lub znużenia? (Kennedy, 2008)	(1) Tak (0) Nie	binarna
PHYS1Q	Czy w ciągu ostatnich siedmiu dni doświadczyłeś utraty apetytu? (American Psychiatric Association, 2019)	(1) Tak (0) Nie	binarna
SOC5A	Jak często w ciągu ostatnich siedmiu dni byłeś zdenerwowany, niespokojny? (Kanter i in., 2008)	(1) Wcale lub krócej niż dzień (2) 1–2 dni (3) 3–4 dni (4) 5–7 dni	porządkowa

[5] Typowa proporcja przyjmowana w data miningu.

Zestawienie 1. Zmienne objaśniające użyte do budowy modelu (dok.)

Zmienne	Opis	Wartości	Typ
SOC5C	Jak często w ciągu ostatnich siedmiu dni czułeś się samotny? (Weeks i in., 1980)	(1) Wcale lub krócej niż dzień (2) 1–2 dni (3) 3–4 dni (4) 5–7 dni	porządkowa
SOC5D	Jak często w ciągu ostatnich siedmiu dni czułeś brak nadziei, myśląc o przyszłości? (Healthline, b.r.)	(1) Wcale lub krócej niż dzień (2) 1–2 dni (3) 3–4 dni (4) 5–7 dni	porządkowa
ECON6B	Czy w ciągu ostatnich siedmiu dni otrzymałeś dodatkową pomoc żywieniową, zawnioskowałeś o nią lub starałeś się zakwalifikować do programu dodatkowej pomocy żywieniowej (*Supplemental Nutrition Program*)? (Brown i Moran, 1997)	(1) Otrzymałem (2) Zawnioskowałem (3) Próbowałem zawnioskować (4) Nie otrzymałem lub nie wnioskowałem	nominalna
ECON7_1	Czy obecnie, aby pokryć dodatkowy wydatek, musiałbyś skorzystać ze swojej karty kredytowej i mógłbyś spłacić dług dopiero w następnym okresie rozliczeniowym? (Brown i Moran, 1997)	(1) Tak (0) Nie	binarna
PHYS2_18	Czy uważasz, że gdy zostajesz w domu z powodu złego samopoczucia, wynika to z panowania pandemii koronawirusa?	(1) Tak (0) Nie	binarna
PHYS3M	Czy stwierdzono u ciebie nadwagę lub otyłość? (Pereira-Miranda i in., 2017)	(1) Tak (0) Nie	binarna
ECON6A	Czy w ciągu ostatnich siedmiu dni otrzymałeś zasiłek dla bezrobotnych, zawnioskowałeś o taki zasiłek lub starałeś się go otrzymać? (Brown i Moran, 1997)	(1) Otrzymałem (2) Zawnioskowałem (3) Próbowałem zawnioskować (4) Nie otrzymałem lub nie wnioskowałem	nominalna
AGE7	Wiek w latach	(1) 18–24 (2) 25–34 (3) 35–44 (4) 45–54 (5) 55–64 (6) 65–74 (7) 75+	porządkowa
HHINCOME	Dochód gospodarstwa domowego (roczny) w tys. USD	(1) poniżej 10 (2) 10–20 (3) 20–30 (4) 30–40 (5) 40–50 (6) 50–75 (7) 75–100 (8) 100–150 (9) 150 i więcej	porządkowa

Źródło: opracowanie własne na podstawie COVID Impact Survey (b.r.).

Do budowy modeli predykcyjnych wykorzystano metody i modele statystyczne oraz data mining: regresję logistyczną (ang. *logistic regression* – LR) drzewa decyzyjne (ang. *decision trees* – DT), wzmacnianie gradientowe (Extreme Gradient Boosting – XGBoost), metodę *k*-najbliższych sąsiadów (ang. k-*nearest neighbours* – KNN), sztuczne sieci neuronowe (ang. *neural networks* – NN) i metodę wektorów wspierających (Support Vector Machines – SVM). Szczegółowy opis metod data mining można znaleźć m.in. w następujących publikacjach polskojęzycznych: Frątczak (2012), Larose (2008), Lasek i Pęczkowski (2013) oraz anglojęzycznych: Alpaydin (2004), Kim i Sohn (2010), Rokach i Maimon (2014), Sarma (2007), Hastie i in. (2001).

3.1. Regresja logistyczna (LR)

Model regresji logistycznej jest używany wtedy, gdy zmienna objaśniana ma charakter binarny. Regresja logistyczna pozwala na obliczenie prawdopodobieństwa przewidywanego zdarzenia i opiera się na funkcji logistycznej następującej postaci:

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\alpha+\sum\beta_i x_i)}}, \tag{1}$$

gdzie:
α – stała regresji,
β_i – współczynnik regresji dla i-tej zmiennej objaśniającej,
$x_1, x_2, \ldots, x_k$ – zmienne niezależne.

S-kształtna funkcja logistyczna f, widoczna na wykresie, przyjmuje wartości z zakresu od 0 do 1.

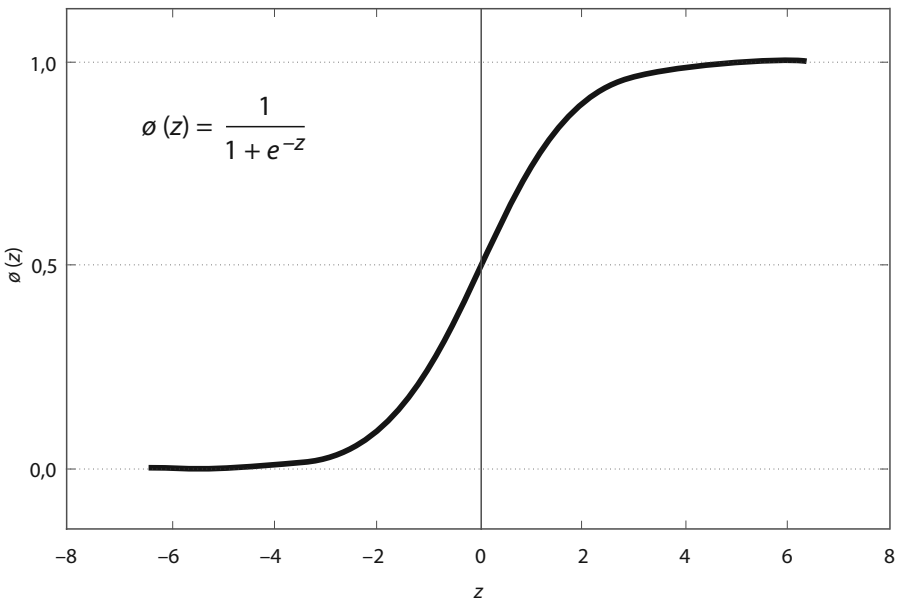
W regresji logistycznej wyniki są interpretowane w kategorii ilorazów szans. Szansę oblicza się, dzieląc prawdopodobieństwo sukcesu przez prawdopodobieństwo porażki, czyli:

$$odds = \frac{p}{1-p}. \tag{2}$$

Do selekcji zmiennych w regresji logistycznej stosuje się metody: wsteczną, postępującą lub krokową. W omawianym badaniu do selekcji zmiennych zastosowano metodę wsteczną (ang. *backward*), czyli wprowadzenie wszystkich zmiennych objaśniających, a następnie usuwanie po kolei najmniej istotnych statystycznie, przy czym istotność zazwyczaj mierzy się rezultatem odpowiedniego testu istotności (*t*-Studenta lub *F-score*) albo kryterium informacyjnym (np. kryterium informacyjnym Akaikego, Akaike information criterion – AIC). W tym przypadku wykorzysta-

no test *t*-Studenta. Ponadto częstą procedurą – uwzględnioną również tutaj – jest sprawdzenie występowania interakcji, czyli wpływu wybranych zmiennych niezależnych na inne zmienne.

Wykres funkcji logistycznej



Źródło: opracowanie własne.

3.2. Drzewa decyzyjne (DT)

Drzewa decyzyjne to narzędzie wykorzystywane do hierarchicznej segmentacji i podziału zbioru danych (schemat 1). Wyjściowym elementem drzewa jest korzeń, który zawiera cały zbiór danych. Kolejno, za pomocą odpowiednich reguł, dokonuje się podziału zbioru danych na węzły (segmenty) i podziału segmentów na subsegmenty. Segment z subsegmentami tworzy gałąź drzewa. Ostateczny subsegment, który nie podlega dalszemu podziałowi, to tzw. liść. Konkretna obserwacja jest przypisana tylko do jednego liścia. Model drzewa pokazuje przyporządkowanie danych do liści, służące do predykcji lub klasyfikacji.

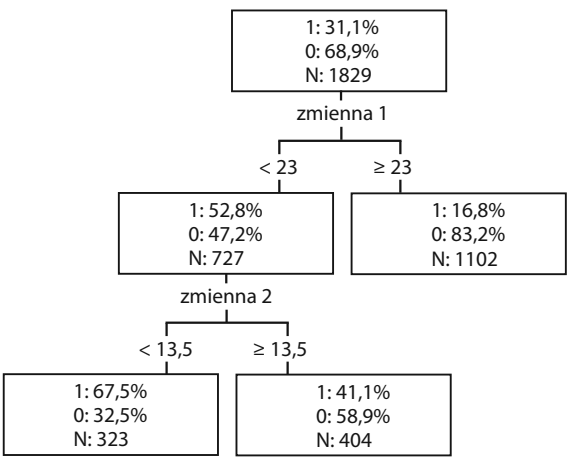
Do budowy drzew decyzyjnych niezbędne są duże zbiory danych z odpowiednio licznymi przypadkami zmiennej objaśnianej. Obserwacje odstające mogą zniekształcić wyniki modelu, ale największym zagrożeniem jest ryzyko przeuczenia modelu, czyli zbytniego dopasowania się modelu do danych uczących, które nie znajduje potwierdzenia w danych testowych. Drzewo decyzyjne nie zawiera oszacowania parametrów, ponieważ w przeciwieństwie do regresji opiera się na podziale zbioru

na oddzielne grupy. Reguły podziału mogą posłużyć do predykcji lub klasyfikacji nowych danych. Drzewo klasyfikacyjne służy do klasyfikacji danych w przypadku zmiennej objaśnianej binarnej lub porządkowej, natomiast drzewo regresyjne – do predykcji w przypadku zmiennej objaśnianej ilorazowej lub przedziałowej.

Podstawowe miary wykorzystywane w drzewie klasyfikacyjnym do pomiaru jakości podziału dla zmiennej zależnej binarnej lub porządkowej to:

- osiągnięty stopień separacji (test chi-kwadrat Pearsona);
- stopień redukcji zanieczyszczenia (entropia lub współczynnik Giniego).

Schemat 1. Przykładowe drzewo decyzyjne



Uwaga. N – liczebność.
Źródło: opracowanie własne.

Kryterium zatrzymania podziału może być osiągnięta minimalna liczba obserwacji w liściu, wielkość węzła lub liczba podziałów w ścieżce. Po zbudowaniu drzewo wymaga przycięcia do optymalnej wielkości.

Zaletami drzew decyzyjnych są przede wszystkim łatwa interpretowalność wyników oraz elastyczność modelu. Drzewa decyzyjne nie są wrażliwe na braki danych, nie wymagają założenia normalności rozkładów czy niezależności zmiennych objaśniających, którymi mogą być zmienne ciągłe, dyskretne, nominalne lub binarne, ani selekcji zmiennych, ponieważ automatycznie wybierane są tylko najistotniejsze zmienne, również z uwzględnieniem zależności nieliniowych.

Za największą wadę drzew decyzyjnych należy uznać ich niestabilność i skłonność do przetrenowania. Ponadto nie mają zastosowania w przypadku niedużych zbiorów danych. Ostateczne oszacowanie prawdopodobieństw jest zagregowane na poziomie subsegmentu końcowego.

3.3. Wzmacnianie gradientowe (XGBoost)

Algorytm XGBoost stanowi jedną z najpopularniejszych metod uczenia maszynowego. Bazuje na wielu klasyfikatorach, przy czym końcowa klasyfikacja poszczególnych obserwacji – na decyzjach podjętych przez pojedyncze drzewa decyzyjne. Podobnie jak w algorytmie AdaBoost wykorzystuje się strategię przyrostową, ponieważ jest mniej czasochłonna i skomplikowana niż trenowanie wszystkich drzew decyzyjnych równolegle (jak w przypadku lasów losowych).

Wzmacnianie gradientowe pozwala na agregację wielu stabilnych, ale niezbyt wydajnych klasyfikatorów. Główną koncepcją wzmacniania jest to, że w procesie iteracyjnym obserwacje otrzymują wagi wskazujące klasyfikatorowi, na których obserwacjach ma się koncentrować w kolejnej iteracji. Finalna decyzja klasyfikacyjna zależy od wyniku głosowania drzew klasyfikatorów.

Zaletę metody wzmacniania stanowi zdolność do redukcji błędu uczenia, który maleje w tempie wykładniczym. Od innych algorytmów XGBoost odróżnia zastosowanie kary, którą model otrzymuje za zbyt dużą liczbę liści w drzewie decyzyjnym. Takie podejście pozwala kontrolować złożoność modelu.

Algorytm składa się z dwóch części. W pierwszej minimalizuje się funkcję straty, czyli błąd (inaczej – funkcję kosztu), w drugiej kontrolowana jest złożoność modelu, co zapobiega przetrenowaniu.

3.4. Metoda *k*-najbliższych sąsiadów (KNN)

Model KNN jest jedną z najbardziej intuicyjnych i najprostszych metod uczenia z nadzorem. Nie uczy się do przodu, ale opóźnia uczenie i dokonuje klasyfikacji w momencie, kiedy otrzymuje żądanie zaklasyfikowania nowych danych. Z tego powodu jest określany jako uczenie oparte na przypadkach lub pamięci (ang. *instance-based learning, memory-based learning*). Polega na przypisaniu nowej niezaklasyfikowanej obserwacji do grupy, do której należy większość jej *k*-najbliższych sąsiadów. Metoda ta efektywnie redukuje błąd klasyfikacji, kiedy liczba obserwacji w próbie w zbiorze treningowym jest duża. Trafność klasyfikacji zależy od przyjętej wartości *k*, czyli optymalnej liczby sąsiadów, i w dużym stopniu od formuły odległości wykorzystanej do obliczenia dystansu między obiektami. W najprostszej wersji metody KNN dopasowanie zależy od liczebności kategorii najbliżej nowym danym.

3.5. Sztuczna sieć neuronowa (NN)

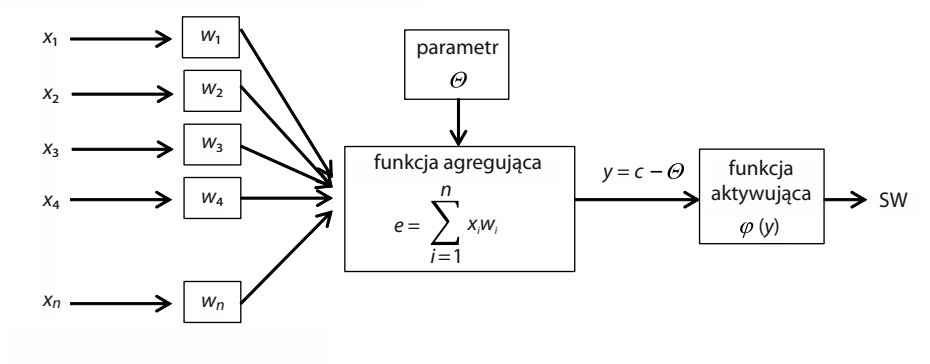
Konstrukcja modelu NN opiera się na podobieństwie do działania mózgu i systemu nerwowego organizmów żywych, czyli sieci neuronów i połączeń pomiędzy nimi. Wagi są modyfikowane podczas uczenia. Do neuronu dociera wiele sygnałów wej-

ściowych x_i, gdzie $i = 1, 2, ..., n$, ale każdy neuron ma jedno wyjście. Każdej zmiennej wejściowej przypisana jest waga w_i. Po przypisaniu wag początkowych neuron jest aktywowany poprzez funkcję aktywacji (e), będącej sumą ilorazu zmiennych wejściowych i przypisanych wag. Wówczas zmienna y jest wyznaczana jako różnica pomiędzy wartością e a wartością progu Θ. Sygnał wyjściowy zależy od pobudzenia neuronu i funkcji aktywacji $\varphi(y)$. Forma tej funkcji określa typ neuronu (zob. schemat 2).

Sztuczne sieci neuronowe wykazują się elastycznością i szybko adaptują się do zmian. Są odporne na zaszumioną informację wejściową i nie wymagają żadnych założeń, takich jak np. normalność rozkładów. Zmienne objaśniające mogą być jakościowe i ilościowe. Możliwe jest modelowanie każdego rodzaju zależności nieliniowych i nieciągłych.

Model NN ma jednak ograniczenia, a największym jest brak możliwości bezpośredniej interpretacji otrzymanych wyników. Przy bardzo skomplikowanej architekturze sieci proces jej uczenia może być czasochłonny lub nie osiągnąć optymalnej redukcji błędu. Sieć nie selekcjonuje automatycznie zmiennych istotnych dla modelu i jest wrażliwa na przetrenowanie.

Schemat 2. Przykładowa sieć neuronowa



Uwaga. SW – sygnał wyjściowy.
Źródło: opracowanie własne.

3.6. Metoda wektorów wspierających (SVM)

Metoda SVM, nazywana też maszyną wektorów wspierających, opiera się na koncepcji płaszczyzn decyzyjnych, które definiują granice decyzji. Płaszczyzna decyzyjna oddziela obiekty należące do różnych klas. Najprostszym przykładem jest klasyfikator liniowy, który dzieli obiekty na klasy za pomocą linii prostej. Większość zagadnień wymaga jednak zastosowania bardziej skomplikowanych struktur do uzyskania

optymalnego podziału, czyli klasyfikacji obiektów ze zbioru testowego na podstawie zbioru uczącego. Klasyfikator bazujący na oddzielnych liniach do wydzielenia różnych klas obiektów jest nazywany *klasyfikatorem hiperpłaszczyznowym*.

Metoda SVM posługuje się hiperpłaszczyznami w przestrzeni wielowymiarowej do podziału obserwacji na klasy. Wspiera zarówno zadanie regresji, jak i klasyfikacji z wieloma zmiennymi ciągłymi i dyskretnymi – w przypadku tych drugich automatycznie tworzone są zmienne binarne odpowiadające kategoriom. Do skonstruowania optymalnej hiperpłaszczyzny metodą wektorów wspierających wykorzystuje się iteracyjny algorytm uczenia do minimalizacji funkcji błędu.

3.7. Ocena modeli

Skuteczność algorytmu uczenia maszynowego może być oceniona za pomocą macierzy pomyłek (ang. *confusion matrix*), która jest przedstawiana w postaci tabeli. Sprawdzając skuteczność algorytmu na zbiorze testowym, otrzymuje się informacje o błędach, które model popełnił w trakcie nauki na zbiorze treningowym. Jeśli klasyfikator jest binarny, macierz pomyłek ma postać macierzy 2×2 (zestawienie 2).

Zestawienie 2. Macierz pomyłek

Wartości przewidywane	Wartości faktyczne	
	pozytywne	negatywne
Pozytywne	prawdziwe pozytywne (TP)	fałszywe pozytywne (FP)
Negatywne	fałszywe negatywne (FN)	prawdziwe negatywne (TN)

Źródło: opracowanie własne na podstawie: Ghoneim (2019).

Komórki (klasy) macierzy to:

- TP (ang. *true positive*) – liczba przypadków pozytywnych, które zostały poprawnie zaklasyfikowane jako pozytywne;
- FN (ang. *false negative*) – liczba przypadków pozytywnych, które zostały niepoprawnie zaklasyfikowane jako negatywne, czyli popełniono błąd drugiego rodzaju;
- FP (ang. *false positive*) – liczba przypadków negatywnych, które zostały niepoprawnie zaklasyfikowane jako pozytywne, czyli popełniono błąd pierwszego rodzaju;
- TN (ang. *true negative*) – liczba przypadków negatywnych, które zostały poprawnie zaklasyfikowane jako negatywne.

Na podstawie powyższych informacji można wprowadzić następujące miary (Ghoneim, 2019):

- trafność (ang. *accuracy*):

$$trafność = \frac{TP + TN}{TP + FN + FP + TN}; \tag{3}$$

- błąd klasyfikacji (ang. *error ratio*):

$$błąd\ klasyfikacji = \frac{FP + FN}{TP + FN + FP + TN}; \tag{4}$$

- precyzja (ang. *precision*) – udział poprawnych klasyfikacji przypadków pozytywnych w liczbie przewidywanych przypadków pozytywnych ogółem:

$$precyzja = \frac{TP}{TP + FP}; \tag{5}$$

- czułość (ang. *recall*) – udział poprawnych klasyfikacji przypadków pozytywnych w liczbie faktycznych przypadków pozytywnych ogółem:

$$czułość = \frac{TP}{TP + FN}; \tag{6}$$

- specyficzność (ang. *specificity*) – udział poprawnych klasyfikacji przypadków negatywnych w liczbie faktycznych przypadków negatywnych ogółem:

$$specyficzność = \frac{TN}{TN + FP}; \tag{7}$$

- *F-score* – średnia harmoniczna czułości oraz precyzji:

$$F\text{-}score = 2 \cdot \frac{precyzja \cdot czułość}{precyzja + czułość}. \tag{8}$$

Miara ta, ze względu na uśrednienie, jest właściwa w przypadku, gdy wartość precyzji i czułości jest podobna. Wartość *F-score* zawiera się w przedziale od 0 do 1, przy czym 1 oznacza model idealny.

4. Wyniki analizy eksploracyjnej na podstawie *COVID Impact Survey*

Do eksploracji danych wykorzystano analizę skupień, a do budowy modeli predykcyjnych – regresję logistyczną, drzewa decyzyjne, wzmacnianie gradientowe, metodę *k*-najbliższych sąsiadów, sztuczne sieci neuronowe i metodę wektorów wspierających. Wybór metod był podyktowany częstością ich stosowania w przypadku zagadnień klasyfikacji i predykcji.

4.1. Analiza skupień

Analiza skupień to proces grupowania podobnych do siebie obserwacji w podzbiory wewnętrznie homogeniczne i jednocześnie odróżniające się od pozostałych (heterogeniczne), który zachodzi na podstawie określonych miar podobieństwa i przyjętego algorytmu. Podejście to ma szerokie zastosowanie w analityce biznesowej, a także w badaniach społecznych i demograficznych. Grupowanie ma na celu sortowanie różnych obiektów w taki sposób, aby stopień ich podobieństwa był maksymalny, jeśli należą do tej samej grupy (Panek i Zwierzchowski, 2013).

W przypadku danych pochodzących z ankiety badającej wpływ pandemii COVID-19 na zdrowie psychiczne i społeczne, a także sytuację ekonomiczną i finansową można dzięki segmentacji podzielić respondentów na mniejsze, jednorodne grupy i określić ich cechy charakterystyczne. To ułatwia uzyskanie odpowiedzi na pytania, czy reakcje ludzi na skutki pandemii były podobne i jakie czynniki okazały się w tych przypadkach istotne.

W omawianym badaniu do analizy skupień wykorzystano wybrane zmienne, głównie quasi-ciągłe:

- SOC3A (Jak często w ciągu ostatniego miesiąca komunikowałeś się z bliskimi telefonicznie lub za pośrednictwem internetu (w tym mailowo)?);
- PHYS8 (Czy swój stan zdrowia oceniłbyś jako: wyśmienity, bardzo dobry, dobry, umiarkowany czy zły?);
- SOC5A (Jak często w ciągu ostatnich siedmiu dni byłeś zdenerwowany, niespokojny?);
- SOC5B (Jak często w ciągu ostatnich siedmiu dni miałeś stany depresyjne?);
- SOC5C (Jak często w ciągu ostatnich siedmiu dni czułeś się samotny?);
- ECON4B (Pomyśl o tym, co wydarzy się za trzy miesiące – jak bardzo jest prawdopodobne, że będziesz wówczas zatrudniony?);
- ECON5AA (Wskaż, na ile prawdziwe dla ciebie jest stwierdzenie: „Martwimy się, że skończą się nam zapasy jedzenia, zanim zdobędziemy środki, aby kupić kolejne produkty”);

- ECON5AB (Wskaż, na ile prawdziwe dla ciebie jest stwierdzenie: „Jedzenie, które zakupiliśmy poprzednim razem, nie wystarczyło nam, ale nie mieliśmy środków, aby kupić kolejne produkty");
- PHYS3H (Czy lekarz lub pracownik medyczny powiedział ci kiedykolwiek, że cierpisz z powodu złego stanu psychicznego?);
- AGE7 (wiek; zmienna ciągła);
- HHINCOME (dochód gospodarstwa domowego).

Po usunięciu obserwacji brakujących pozostałe 6768 poddano analizie. W pierwszym kroku oceniono korelację między zmiennymi. Quasi-ciągły charakter zmiennych pozwolił na zastosowanie współczynnika korelacji liniowej Pearsona. W tabl. 1 zamieszczono wyniki obliczeń dla poszczególnych par wybranych zmiennych.

Tabl. 1. Współczynniki korelacji Pearsona zmiennych quasi-ciągłych wykorzystanych w grupowaniu

Zmienne	SOC3A	PHYS8	SOC5A	SOC5B	SOC5C	ECON4B	ECON5AA	ECON5AB	HHINCOME	HHSIZE1
SOC3A	1,000	0,083	–0,006	–0,023	–0,007	0,095	–0,031	–0,036	–0,102	–0,016
PHYS8	0,083	1,000	0,179	0,167	0,165	0,205	–0,180	–0,179	–0,280	–0,018
SOC5A	–0,006	0,179	1,000	0,595	0,575	–0,011	–0,174	–0,164	–0,097	0,033
SOC5B	–0,023	0,167	0,595	1,000	0,590	–0,027	–0,183	–0,162	–0,102	0,019
SOC5C	–0,007	0,165	0,575	0,590	1,000	–0,033	–0,202	–0,172	–0,103	0,027
ECON4B	0,095	0,205	–0,011	–0,027	–0,033	1,000	–0,045	–0,050	–0,214	–0,186
ECON5AA ...	–0,031	–0,180	–0,174	–0,183	–0,202	–0,045	1,000	**0,806**	0,411	–0,144
ECON5AB ...	–0,036	–0,179	–0,164	–0,162	–0,172	–0,050	**0,806**	1,000	0,386	–0,124
HHINCOME	–0,102	–0,280	–0,097	–0,102	–0,103	–0,214	0,411	0,386	1,000	0,104
HHSIZE1	–0,016	–0,018	0,033	0,019	0,027	–0,186	–0,144	–0,124	0,104	1,000

Uwaga. Pogrubieniem zaznaczono zmienne silnie skorelowane.

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Analizując te wyniki, można dostrzec pewne związki między zmiennymi. Dla przykładu zmienna HHINCOME wykazuje ujemną korelację ze zmienną PHYS8 – wyższe dochody gospodarstwa domowego częściej wiążą się z ocenami stanu zdrowia jako wyśmienity i bardzo dobry. Zbadanie korelacji zmiennych w zbiorze stanowiło ułatwienie przy wyborze zmiennych do segmentacji (pominięcie zmiennych silnie skorelowanych).

Skala wartości wybranych zmiennych nie była bardzo zróżnicowana, w związku z czym nie przeprowadzono ich standaryzacji, ponieważ nie poprawiłaby ona wyników analizy. Do grupowania wykorzystano natomiast metodę hierarchiczną Warda, która opiera się na łączeniu skupień o minimalnej sumie kwadratów odchyleń między środkami ciężkości tych skupień. Wykorzystano kryterium CCC (Cubic Clustering Criterion) oraz miary pseudo-t^2 i pseudo-R. Kierując się tymi przesłankami,

zdecydowano się na utworzenie pięciu skupień. Ponadto – w celu pogłębienia analizy – zastosowano iteracyjną metodę *k*-średnich.

Aby lepiej określić profil osób znajdujących się w poszczególnych skupieniach, obliczono średnie wartości zmiennych w grupach (tabl. 2).

Tabl. 2. Średnie wartości zmiennych w poszczególnych skupieniach

Zmienne	Skupienia				
	1	2	3	4	5
SOC3A	1,424	1,387	1,430	1,788	1,324
PHYS8	2,317	1,965	2,555	1,788	2,446
SOC5A	1,510	1,419	1,831	1,788	2,026
SOC5B	1,473	1,416	1,891	1,292	2,044
SOC5C	1,445	1,418	1,902	1,260	2,022
ECON4B	4,668	1,396	2,352	4,718	1,989
ECON5AA	2,921	2,963	2,372	2,669	2,695
ECON5AB	2,955	2,985	2,487	2,725	2,777
PHYS3H	0,114	0,117	0,255	0,155	0,246
HHINCOME	6,850	7,542	2,791	2,891	6,377
HHSIZE1	1,995	2,341	2,620	1,465	3,756

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Po przeanalizowaniu wyników dotyczących poszczególnych skupień można zauważyć, że respondenci w dwóch licznych grupach (skupienie 1 i 2) rzadko deklarowali problemy ze zdrowiem psychicznym czy fizycznym (zob. tabl. 3). Widoczna jest także zależność pomiędzy dochodami gospodarstwa domowego a oceną stanu zdrowia ankietowanych.

Tabl. 3. Liczebność grup i opis cech charakterystycznych wyróżnionych skupień

Skupienie	Liczebność	Opis
1	1336	osoby o stosunkowo wysokich dochodach i dobrym zdrowiu psychicznym, jednak często obawiające się o swoje zatrudnienie w najbliższym czasie
2	2187	osoby o najwyższych dochodach, najlepiej oceniające swój stan zdrowia, najmniej martwiące się o zatrudnienie w najbliższym czasie
3	1578	osoby o najniższych dochodach, często deklarujące problemy finansowe, najgorzej oceniające stan swojego zdrowia psychicznego
4	709	osoby w wielu przypadkach samotne, martwiące się o zatrudnienie, nie najgorzej oceniające stan swojego zdrowia
5	958	osoby otoczone bliskimi, często deklarujące uczucie zdenerwowania czy niepokoju, stosunkowo często sygnalizujące gorszy stan zdrowia psychicznego

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Interesująca jest zmienna ECON4B – oceniane przez respondentów prawdopodobieństwo bycia zatrudnionym w ciągu najbliższych trzech miesięcy. Dotyczące jej wyniki mogą mieć związek z aktualną sytuacją pandemiczną. Osoby najbardziej

dotknięte problemami z zatrudnieniem wypadały najgorzej w odpowiedziach na pytania dotyczące stanów depresyjnych, samotności i niepokoju. Oznacza to, że negatywne skutki pandemii miały wpływ na zdrowie psychiczne wielu osób.

4.2. Analiza i kategoryzacja zmiennych

Dla części zmiennych przeprowadzono kategoryzację. Dzięki łączeniu nieznacznie różniących się kategorii, także pod względem rozkładu zmiennej objaśnianej, usunięto grupy o niskiej liczebności (poniżej 5%). Zmienne AGE7 i HHINCOME skategoryzowano, aby uzyskać monotoniczność emiprycznego logitu. Zmienne ostatecznie uwzględnione w modelu przedstawiono w zestawieniu 3.

Zestawienie 3. Zmienne objaśniające po kategoryzacji

Zmienne	Opis	Wartości	Typ
PHYS8	Czy swój stan zdrowia oceniłbyś jako: wyśmienity, bardzo dobry, dobry, umiarkowany czy zły?	(1) Wyśmienity (2) Bardzo dobry (3) Dobry (4) Umiarkowany (5) Zły	nominalna
PHYS1K	Czy w ciągu ostatnich siedmiu dni doświadczyłeś uczucia zmęczenia lub znużenia?	(1) Tak (0) Nie	binarna
PHYS1Q	Czy w ciągu ostatnich siedmiu dni doświadczyłeś utraty apetytu?	(1) Tak (0) Nie	binarna
SOC5A	Jak często w ciągu ostatnich siedmiu dni byłeś zdenerwowany, niespokojny?	(0) Prawie wcale (1) Co najmniej jeden dzień	binarna
SOC5C	Jak często w ciągu ostatnich siedmiu dni czułeś się samotny?	(0) Prawie wcale (1) Co najmniej jeden dzień	binarna
SOC5D	Jak często w ciągu ostatnich siedmiu dni czułeś brak nadziei, myśląc o przyszłości?	(0) Prawie wcale (1) Co najmniej jeden dzień	binarna
ECON6B	Czy w ciągu ostatnich siedmiu dni otrzymałeś dodatkową pomoc żywieniową, zawnioskowałeś o nią lub starałeś się zakwalifikować do programu dodatkowej pomocy żywieniowej (*Supplemental Nutrition Program*)?	(1) Zawnioskowałem lub planuję to zrobić (0) Nie wnioskowałem	binarna
ECON7_1	Czy obecnie, aby pokryć dodatkowy wydatek, musiałbyś skorzystać ze swojej karty kredytowej i mógłbyś spłacić dług dopiero w następnym okresie rozliczeniowym?	(1) Tak (0) Nie	binarna
PHYS2_18	Czy uważasz, że gdy zostajesz w domu z powodu złego samopoczucia, wynika to z panowania pandemii koronawirusa?	(1) Tak (0) Nie	binarna
PHYS3M	Czy stwierdzono u ciebie nadwagę lub otyłość?	(1) Tak (0) Nie	binarna

Zestawienie 3. Zmienne objaśniające po kategoryzacji (dok.)

Zmienne	Opis	Wartości	Typ
ECON6A	Czy w ciągu ostatnich siedmiu dni otrzymałeś zasiłek dla bezrobotnych, zawnioskowałeś o taki zasiłek lub starałeś się go otrzymać?	(1) Zawnioskowałem lub zamierzam zawnioskować (0) Nie zawnioskowałem	binarna
AGE7	Wiek w latach	(1) 18–24 (2) 25–34 (3) 35–44 (4) 45–54 (5) 55–64 (6) 65–74 (7) 75+	porządkowa
HHINCOME	Dochód gospodarstwa domowego (roczny) w tys. USD	(1) poniżej 10 (2) 10–20 (3) 20–30 (4) 30–40 (5) 40–50 (6) 50–75 (7) 75–100 (8) 100–150 (9) 150 i więcej	porządkowa

Źródło: opracowanie własne.

4.3. Model regresji logistycznej

Pierwszy etap polegał na zbudowaniu modelu z efektami głównymi, co pozwoliło na badanie istotności zmiennych objaśniających. Identyfikowano potencjalne zmienne zakłócające i zmienne do usunięcia z modelu. Przypomnijmy, że zmienną objaśnianą jest zmienna binarna PHYS3H. Przyjmuje ona wartość 1, gdy dana osoba ma zdiagnozowaną chorobę o podłożu psychicznym.

Za zmienną zakłócającą uznawano taką, której usunięcie wpływało na zmianę oszacowania parametru przy zmiennej o ponad 10%. W przypadku gdy usunięcie wybranej zmiennej nie oddziaływało wyraźnie na oszacowanie ocen parametrów przy pozostałych zmiennych, była ona kwalifikowana jako możliwa do usunięcia z modelu.

W kolejnych krokach z modelu usuwano zmienne nieistotne statystycznie, w pierwszej kolejności zmienne o wyższym p-value w teście istotności. W związku z tym, że jedną z takich zmiennych jest HHINCOME, istotna z punktu widzenia badania, postanowiono nie eliminować jej z modelu, a jedynie zbadać jej wpływ na zmienną ECON6A.

Następnie zbadano efekt usunięcia zmiennej PHYS1K. Jej usunięcie nie wpłynęło znacząco na oszacowania parametrów pozostałych zmiennych, co oznacza, że PHYS1K jest zmienną możliwą do usunięcia z modelu. Zdecydowano się również zbadać wpływ, jaki wywarłoby usunięcie zmiennych istotnych statystycznie na osza-

cowania parametrów pozostałych zmiennych. Redukując model kolejno o zmienne ECON6A, ECON7_1, PHYS2_18, PHYS8, ECON6B, PHYS3M i AGE7, zidentyfikowano zmiany oszacowania parametrów przy wielu zmiennych. Z tego powodu zbadano wprowadzenie do modelu 29 interakcji.

Dalszym etapem budowy modelu było wprowadzenie interakcji drugiego rzędu, wykrytych na podstawie różnic oszacowania parametrów wynikających z usunięcia wybranej zmiennej. Zastosowano metodę selekcji krokowej, która polega na wprowadzaniu lub usuwaniu zmiennych z modelu. Wymuszono dodanie zmiennej HHINCOME, aby zyskać szansę na zidentyfikowanie efektów zmiennych modyfikujących. Dla tego testu został przyjęty poziom istotności 0,05.

Wyniki analizy, przedstawione w tabl. 4, wskazują na istotność jednej interakcji – ECON6A i ECON6B – oraz dziesięciu innych zmiennych objaśniających. Jednak żadna ze zmiennych możliwych do usunięcia z modelu nie stworzyła interakcji istotnych statystycznie, przez co w modelu nie wystąpiły żadne zmienne modyfikujące efekty główne. Dołączenie interakcji do modelu sprawiło, że nieistotna statystycznie okazała się zmienna PHYS2_18, mimo że nie stworzyła z innymi zmiennymi żadnych interakcji.

Tabl. 4. Analiza istotności zmiennych dla modelu z interakcjami

Zmienne	Stopnie swobody	Chi-kwadrat Walda	*p*-value
Zasiłek dla bezrobotnych (ECON6A)	1	8,4	0,0038
Dochód gospodarstwa (HHINCOME)	1	0,7	0,3975
Stan zdrowia (PHYS8)	4	42,3	<0,0001
Uczucie zmęczenia (PHYS1K)	1	3,4	0,0637
Utrata apetytu (PHYS1Q)	1	14,5	0,0001
Zdenerwowanie lub niepokój (SOC5A)	1	19,9	<0,0001
Samotność (SOC5C)	1	37,7	<0,0001
Brak nadziei na przyszłość (SOC5D)	1	32,8	<0,0001
Program pomocy żywieniowej (ECON6B)	1	14,7	0,0001
Dług na karcie kredytowej (ECON7_1)	1	4,3	0,0383
Wpływ pandemii koronawirusa na złe samopoczucie (PHYS2_18)	1	3,8	0,0512
Nadwaga lub otyłość (PHYS3M)	1	31,1	<0,0001
Wiek (AGE7)	1	69,9	<0,0001
Zasiłek dla bezrobotnych wśród osób korzystających z programu pomocy żywieniowej (ECON6B · ECON6A)	1	5,5	0,0193

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Do oceny jakości modelu wykorzystano miary oparte na macierzy pomyłek na zbiorze testowym. Warto zwrócić uwagę na to, że przygotowanie modelu nie służyło uzyskaniu maksymalnej wartości predykcji. Jego celem była analiza wpływu pandemii COVID-19 na stan zdrowia psychicznego respondentów.

Model regresji logistycznej cechuje bardzo wysoka (na poziomie ponad 86%) trafność, czyli udział poprawnych klasyfikacji. Przy wysokiej precyzji (ponad 73%) uzyskano czułość na poziomie 31%, co oznacza, że 31% faktycznie chorych respondentów model prawidłowo zaklasyfikował jako chorych. Natomiast specyficzność modelu na poziomie 98% wskazuje, że dobór charakterystyk daje lepsze możliwości klasyfikacji respondentów zdrowych (tabl. 5).

Tabl. 5. Ocena modelu regresji logistycznej

Wyszczególnienie	Klasa faktyczna		Wartość miary
	chorzy	zdrowi	
Klasa przewidywana: chorzy	89	33	.
zdrowi	197	1345	.
Trafność	.	.	0,86
Precyzja	.	.	0,73
Czułość	.	.	0,31
Specyficzność	.	.	0,98
F-score	.	.	0,43

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Ze względu na wyraźną dysproporcję pomiędzy grupą respondentów chorych a grupą respondentów zdrowych wyniki oceny modelu należy jednak interpretować ostrożnie. Wartość *F-score* jest w pełni interpretowalna tylko przy podobnej wartości precyzji i czułości.

Wyniki finalnego modelu zostały poddane analizie i interpretacji. W pierwszej kolejności zbadano oceny parametrów i ilorazów szans (tabl. 6). Jedyną zmienną, która okazała się nieistotna statystycznie na poziomie istotności 0,1, jest HHINCOME.

Tabl. 6. Oceny parametrów i ilorazy szans w modelu regresji

Zmienne	Ocena parametru	Iloraz szans
Wyraz wolny	–1,77	.
Zasiłek dla bezrobotnych (ECON6A)	–1,00	0,37
Dochód gospodarstwa (HHINCOME)	–0,02	0,98
Stan zdrowia (PHYS8)[a]: (1) Wyśmienity	–0,52	0,59
(3) Dobry	0,25	1,29
(4) Umiarkowany	0,54	1,72
(5) Zły	0,83	2,30
Utrata apetytu (PHYS1Q)	0,48	1,62
Zdenerwowanie lub niepokój (SOC5A)	0,23	1,26
Samotność (SOC5C)	0,33	1,39
Brak nadziei na przyszłość (SOC5D)	0,30	1,35
Program pomocy żywieniowej (ECON6B)	–0,19	0,83
Dług na karcie kredytowej (ECON7_1)	–0,20	0,82
Wpływ pandemii koronawirusa na złe samopoczucie (PHYS2_18)	0,25	1,28

Tabl. 6. Oceny parametrów i ilorazy szans w modelu regresji (dok.)

Zmienne	Ocena parametru	Iloraz szans
Nadwaga lub otyłość (PHYS3M)	0,51	1,66
Wiek (AGE7)	–0,22	0,80
Zasiłek dla bezrobotnych wśród osób korzystających z programu pomocy żywieniowej (ECON6B · ECON6A)	0,23	.

a Kategoria referencyjna: Stan zdrowia (PHYS8) – (2) Bardzo dobry.

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Powyższe wyniki pozwalają wyciągnąć wiele wniosków. U osób oceniających swój stan zdrowia jako wyśmienity szanse zdiagnozowania chorób o podłożu psychicznym są o ponad 40% mniejsze niż u osób określających swój stan zdrowia jako bardzo dobry (przy innych czynnikach niezmienionych). U osób oceniających swój stan zdrowia jako dobry szanse te są o prawie 30% wyższe, natomiast u osób oceniających swoje zdrowie najsłabiej – blisko dwuipółkrotnie wyższe (przy założeniu *ceteris paribus*) niż u osób określających swój stan zdrowia jako bardzo dobry.

Zmienna HHINCOME okazała się statystycznie nieistotna. Zmienna ECON6A została włączona do modelu w interakcji ze zmienną ECON6B, co utrudnia interpretację tej drugiej.

W przypadku zmiennej PHYS1Q okazało się, że szanse zdiagnozowania choroby psychicznej są ponadpółtorakrotnie wyższe u osoby, która w ciągu ostatnich siedmiu dni doświadczyła utraty apetytu, niż u osoby, która nie miała takiego problemu (przy założeniu *ceteris paribus*).

Dla zmiennych SOC5A, SOC5C i SOC5D otrzymano następujące wyniki: u osób, które w ciągu ostatnich siedmiu dni były zdenerwowane lub zaniepokojone, szanse zdiagnozowania chorób o podłożu psychicznym są o ok. 26% wyższe, u osób, które w ostatnim czasie czuły się samotne – o blisko 40% wyższe, a u osób martwiących się o swoją przyszłość – o ponad 35% wyższe niż u respondentów, którzy takich problemów nie zgłosili (przy innych czynnikach niezmienionych).

W przypadku osób, u których stwierdzono nadwagę lub otyłość, szanse zdiagnozowania choroby psychicznej są o ponad 66% wyższe niż w przypadku osób z prawidłową wagą lub niedowagą. Natomiast wśród tych, którzy nieprzewidziane wydatki musieli opłacać z karty kredytowej, szanse te są o ok. 18% niższe.

Jeśli chodzi o zmienną AGE7, to prawdopodobieństwo zdiagnozowania choroby psychicznej spada z wiekiem o 20% rocznie (*ceteris paribus*).

Bardzo istotna z punktu widzenia analizy jest zmienna PHYS2_18. Szanse zdiagnozowania choroby o podłożu psychicznym u osób, które zauważają u siebie gorsze samopoczucie spowodowane pandemią, są o blisko 30% wyższe niż u osób, które tego nie dostrzegają (przy innych czynnikach niezmienionych).

4.4. Model drzewa decyzyjnego

W budowie drzewa decyzyjnego posłużono się metodą przycinania *prune costcomplexity*, która wykorzystuje analizę kosztów na podstawie walidacji krzyżowej. Miarą błędu jest współczynnik błędnych klasyfikacji. Liczba liści w modelu wyniosła 10. Statystyki dopasowania drzewa przedstawiono w tabl. 7.

Tabl. 7. Ocena modelu drzewa decyzyjnego

Wyszczególnienie	Klasa faktyczna		Wartość miary
	chorzy	zdrowi	
Klasa przewidywana: chorzy	40	17	.
zdrowi	246	1361	.
Trafność	.	.	0,84
Precyzja	.	.	0,70
Czułość	.	.	0,14
Specyficzność	.	.	0,99
F-score	.	.	0,23

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Metoda przycinania drzewa niemal całkowicie niweluje problem nadmiernego dopasowania do zbioru treningowego, ponieważ statystyki dopasowania osiągają bardzo zbliżone wyniki zarówno w zbiorze treningowym, jak i testowym.

Model drzewa decyzyjnego cechuje wysoka trafność (ponad 84%), ale bardzo niska czułość, czyli zdolność do poprawnej klasyfikacji respondentów chorych. Specyficzność modelu bliska 1 wskazuje na wysoką precyzję poprawnej klasyfikacji respondentów zdrowych.

4.5. Model XGBoost

Użycie modelu XGBoost wiąże się w pierwszej kolejności z doborem odpowiedniego wzmacniacza, na podstawie którego dobierane są pozostałe parametry. W omawianym przypadku najlepsze wyniki uzyskano dla wzmacniacza gbtree wraz z maksymalną głębokością drzewa oraz minimalną sumą wag instancji ustawioną na 2.

Wprowadzenie losowości do drzew decyzyjnych i oparcie modelu na algorytmie XGBoost przyniosło wzrost trafności modelu i ponaddwukrotny wzrost czułości w porównaniu z drzewem decyzyjnym (tabl. 8).

Tabl. 8. Ocena modelu XGBoost

Wyszczególnienie	Klasa faktyczna		Wartość miary
	chorzy	zdrowi	
Klasa przewidywana: chorzy	82	32	.
zdrowi	204	1346	.
Trafność	.	.	0,86
Precyzja	.	.	0,72
Czułość	.	.	0,29
Specyficzność	.	.	0,98
F-score	.	.	0,41

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

4.6. Model *k*-najbliższych sąsiadów

Zastosowanie algorytmu *k*-najbliższych sąsiadów jest związane głównie z odpowiednim doborem parametru *k* wyznaczającego liczbę sąsiadów, do których odległość od danego elementu zbioru jest najmniejsza. Dobór wartości parametru *k* należy przeprowadzić ostrożnie, ponieważ zbyt duża jego wartość może się przyczynić do rozmycia heterogeniczności obszarów podziału, a także skutkować występowaniem większej liczby błędów klasyfikacji rzadszych wzorców. W trakcie eksperymentów przetestowano wartości *k* z przedziału od 1 do 30, a najlepsze wyniki uzyskano dla $k = 24$.

Tabl. 9. Ocena modelu *k*-najbliższych sąsiadów

Wyszczególnienie	Klasa faktyczna		Wartość miary
	chorzy	zdrowi	
Klasa przewidywana: chorzy	53	15	.
zdrowi	233	1363	.
Trafność	.	.	0,85
Precyzja	.	.	0,78
Czułość	.	.	0,19
Specyficzność	.	.	0,99
F-score	.	.	0,30

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Model *k*-najbliższych sąsiadów, podobnie jak model drzewa decyzyjnego, charakteryzuje się niską czułością i bardzo wysoką specyficznością (tabl. 9).

4.7. Model sztucznej sieci neuronowej

W zastosowanej w badaniu sztucznej sieci neuronowej każdy neuron z poprzedniej warstwy jest połączony z każdym neuronem z warstwy kolejnej. Występują tu trzy warstwy; pierwsza składa się z 26 neuronów, druga – z 13, a ostatnia (wyjściowa)

– z 1. Jako funkcję agregującą przyjęto funkcję liniową, a jako funkcję aktywacji – funkcję logitową.

Tabl. 10. Ocena modelu sztucznej sieci neuronowej

Wyszczególnienie	Klasa faktyczna		Wartość miary
	chorzy	zdrowi	
Klasa przewidywana: chorzy	95	37	.
zdrowi	191	1341	.
Trafność	.	.	0,86
Precyzja	.	.	0,72
Czułość	.	.	0,33
Specyficzność	.	.	0,97
F-score	.	.	0,45

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Okazało się, że model sztucznej sieci neuronowej cechuje się wysoką trafnością, ale jednocześnie najwyższą czułością i wysoką specyficznością. Model poprawnie klasyfikuje chorych w 33% przypadków (tabl. 10).

4.8. Model oparty na metodzie wektorów wspierających

Przeprowadzono eksperymenty, w których przetestowano kilka rodzajów funkcji podziału przestrzeni decyzyjnej w celu wyznaczenia granicy przynależności klasowej poszczególnych punktów, m.in. liniową, sigmoidalną, wielowymiarową oraz radialną funkcję bazową, dla której uzyskano najlepsze wyniki klasyfikacji.

Tabl. 11. Ocena modelu opartego na metodzie wektorów wspierających

Wyszczególnienie	Klasa faktyczna		Wartość miary
	chorzy	zdrowi	
Klasa przewidywana: chorzy	35	8	.
zdrowi	251	1370	.
Trafność	.	.	0,84
Precyzja	.	.	0,81
Czułość	.	.	0,12
Specyficzność	.	.	0,99
F-score	.	.	0,21

Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Czułość modelu opartego na metodzie wektorów wspierających okazała się najniższa. Model najlepiej rozpoznaje i klasyfikuje respondentów zdrowych, natomiast w przypadku respondentów chorych jego trafność klasyfikacji jest najniższa (tabl. 11).

4.9. Porównanie modeli

Do porównania modeli zastosowano miarę precyzji i czułości oraz uśrednioną miarę trafności predykcji, jaką jest *F1-score* na zbiorze testowym.

Tabl. 12. Porównanie modeli

Miary	RL	DT	XGBoost	KNN	NN	SVM
Trafność	**0,86**	0,84	**0,86**	0,85	**0,86**	0,84
Precyzja	0,73	0,70	0,72	0,78	0,72	**0,81**
Czułość	0,31	0,14	0,29	0,19	**0,33**	0,12
F1-score	0,43	0,23	0,41	0,30	**0,45**	0,21

Uwaga. Najwyższe wartości zaznaczono pogrubieniem.
Źródło: opracowanie własne z wykorzystaniem SAS Enterprise Guide.

Wyniki zastosowania metod uczenia maszynowego charakteryzują się tylko nieznacznie większą precyzją klasyfikacji niż wyniki regresji logistycznej. Gdy weźmie się pod uwagę wszystkie zastosowane modele, to okazuje się, że najlepsze (ale porównywalne z regresją) wyniki daje sztuczna sieć neuronowa (tabl. 12).

5. Podsumowanie

Głównym celem badania omawianego w artykule było ustalenie wpływu pandemii COVID-19 na stan zdrowia psychicznego członków gospodarstw domowych. Przeprowadzone analizy umożliwiły m.in. ocenę oddziaływania skutków rozprzestrzeniania się koronawirusa na gospodarstwa domowe. To ważne, ponieważ badacze mogą przekazać zdobytą wiedzę decydentom i opinii publicznej. Analizowano głównie wyniki badań i dane pochodzące ze Stanów Zjednoczonych, gdzie w ramach prac zleconych przez rząd i organizacje pozarządowe przeprowadzano ankiety dotyczące statusu zatrudnienia, bezpieczeństwa żywnościowego, warunków mieszkaniowych, zdrowia fizycznego, dostępu do opieki zdrowotnej czy problemów związanych z edukacją.

Jak twierdzą badacze publikujący w dzienniku medycznym „The Lancet", u osób, które przed wybuchem pandemii nie miały zaburzeń depresyjnych, lękowych lub obsesyjno-kompulsyjnych, w czasie jej trwania stwierdzono dość intensywne występowanie ich objawów. Co więcej, w czerwcu 2020 r. 40% dorosłych Amerykanów doświadczało zaburzeń psychicznych, a 11% ujawniło, że w ostatnim miesiącu poważnie rozważało podjęcie próby popełnienia samobójstwa.

Warto podkreślić, że choroby psychiczne są jednym z bagatelizowanych skutków pandemii COVID-19. Izolacja, zdalna edukacja i praca, ograniczenie kontaktów

międzyludzkich czy zmniejszona aktywność fizyczna, a w czasie pierwszej fali pandemii także panika zakupowa i strach przed zachorowaniem, wycisnęły piętno na kondycji psychicznej ludzi na całym świecie. Skutkiem tych zjawisk będą ukryte koszty pandemii, które będziemy ponosić w przyszłości. Wyniki badań dotyczących omawianej tematyki mogą okazać się pomocne podczas opracowywania strategii łagodzących negatywne skutki pandemii.

Analiza skupień pozwoliła na stworzenie pięciu grup, z których dwie najbardziej liczne obejmowały osoby oceniające swój stan zdrowia jako wyśmienity lub bardzo dobry. W tych skupieniach średnia dochodów gospodarstw domowych była najwyższa, co skłoniło badaczy do przeprowadzenia dalszej analizy sytuacji finansowej respondentów i zbadania jej związku z możliwą chorobą o podłożu psychicznym. Z tego względu do modeli została wprowadzona zmienna HHINCOME, określająca dochody ankietowanych.

Kolejny interesujący wynik uzyskany na podstawie grupowania jest taki, że często stany niepokoju lub zdenerwowania identyfikowano u respondentów otoczonych bliskimi – mieszkających z kilkoma innymi osobami lub utrzymujących częsty kontakt z rodziną. Może to oznaczać, że nie tylko samotność, lecz także obawa o najbliższych wpływają na stan zdrowia psychicznego. Natomiast w skupieniu, w którym znalazło się wiele osób obawiających się o swoje zatrudnienie, samoocena stanu zdrowia dokonana przez respondentów była dobra.

W badaniu zastosowano także modele data mining oraz model regresji logistycznej. Uzyskane wyniki unaoczniły, że wśród osób, które w ciągu ostatnich siedmiu dni doświadczyły uczucia zdenerwowania lub niepokoju, szanse zdiagnozowania chorób o podłożu psychicznym są o ok. 26% wyższe wśród tych, którzy w ostatnim czasie czuli się samotni (o blisko 40%) i martwili się o swoją przyszłość (o ponad 35%) niż wśród osób, które oceniły swój stan zdrowia jako bardzo dobry (*ceteris paribus*). Dodatkowo szanse te rosną znacząco w przypadku osób otyłych lub z nadwagą. Izolacja, zdalny tryb nauczania i pracy oraz zmniejszona aktywność fizyczna niewątpliwie przyczyniają się zatem do pogarszania się stanu zdrowia psychicznego.

Zastosowane w badaniu modele data mining dostarczyły wyników tylko nieznacznie lepszych, o wyższej precyzji klasyfikacji, niż model regresji logistycznej. Z porównania wszystkich zastosowanych modeli wynika, że najlepsze – ale porównywalne z rezultatami regresji – wyniki uzyskano za pomocą sztucznej sieci neuronowej.

Bibliografia

Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press.

American Psychiatric Association. (2019, April). *Warning Signs of Mental Illness*. https://psychiatry.org/patients-families/warning-signs-of-mental-illness.

Brown, G. W., Moran, P. M. (1997). Single mothers, poverty and depression. *Psychological medicine*, *27*(1), 21–33. https://doi.org/10.1017/s0033291796004060.

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M., The Opportunity Insights. (2020). *The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data* (NBER Working Paper 27431). http://www.nber.org/papers/w27431.

COVID Impact Survey. (b.r.). *Reliable information about the impacts of the COVID-19 pandemic*. Pobrane 10 października 2021 r. z https://www.covid-impact.org/.

Czeisler, M. É., Lane, R. I., Petrosky, E., Wiley, J. F., Christensen, A., Njai, R., Weaver, M. D., Robbins, R., Facer-Childs, E. R., Barger, L. K., Czeisler, C. A., Howard, M. E., Rajaratnam, S. M. W. (2020). *Mental Health, Substance Use, and Suicidal Ideation During the COVID-19 Pandemic*. Centers for Disease Control and Prevention.

European Foundation for the Improvement of Living and Working Conditions. (2021). *Living, working and COVID-19 (Update April 2021): Mental health and trust decline across EU as pandemic enters another year*. Publications Office of the European Union. https://doi.org/10.2806/76802.

Frątczak, E. (red.). (2012). *Zaawansowane metody analiz statystycznych*. Oficyna Wydawnicza Szkoły Głównej Handlowej w Warszawie.

Ghoneim, S. (2019, April 2). *Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on?*. https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124.

Hastie, T., Tibshirani, R., Friedman, J. H. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Predition*. Springer-Verlag.

Healthline. (b.r.). *Sings of Depression*. https://www.healthline.com/health/depression/recognizing-symptoms#lostinterest.

Kanter, J. W., Busch, A. M., Weeks, C. E., Landes, S. J. (2008). The Nature of Clinical Depression: Symptoms, Syndromes, and Behavior Analysis. *The Behavior Analyst*, *31*(1), 1–21. https://doi.org/10.1007/BF03392158.

Kennedy, S. H. (2008). Core symptoms of major depressive disorder: relevance to diagnosis and treatment. *Dialogues in Clinical Neuroscience*, *10*(3), 271–277. https://doi.org/10.31887/DCNS.2008.10.3/shkennedy.

Kim, H. S., Sohn, S. Y. (2010). Support Vector Machines for Default Prediction of SMEs Based on Technology Credit. *European Journal of Operational Research*, *201*(3), 838–846. https://doi.org/10.1016/j.ejor.2009.03.036.

Larose, D. T. (2008). *Metody i modele eksploracji danych*. Wydawnictwo Naukowe PWN.

Lasek, M., Pęczkowski, M. (2013). *Enterprise Miner. Wykorzystanie narzędzi* Data Mining *w systemie* SAS. Wydawnictwa Uniwersytetu Warszawskiego. https://doi.org/10.31338/uw.9788323527701.

National Institute of Mental Health. (2021). *Chronic Illness and Mental Health Recognizing and Treating Depression*. https://www.nimh.nih.gov/sites/default/files/documents/health/publications/chronic-illness-mental-health/21-mh-8015-chronicillness-mentalhealth_1.pdf.

Pan, K.-Y., Kok, A. A. L., Eikelenboom, M., Horsfall, M., Jörg, F., Luteijn, R. A., Rhebergen, D., van Oppen, P., Giltay, E. J., Penninx, B. W. J. H. (2021). The mental health impact of the COVID-19 pandemic on people with and without depressive, anxiety or obsessive-compulsive disorders: a longitudinal study of three Dutch case-control cohorts. *The Lancet. Psychiatry*, *8*(2), 121–129. https://doi.org/10.1016/S2215-0366(20)30491-0.

Panek, T., Zwierzchowski, J. (2013). *Statystyczne metody wielowymiarowej analizy porównawczej. Teoria i zastosowania*. Oficyna Wydawnicza Szkoły Głównej Handlowej w Warszawie.

Pereira-Miranda, E., Costa, P. R. F., Queiroz, V. A. O., Pereira-Santos, M., Santana, M. L. P. (2017). Overweight and Obesity Associated with Higher Depression Prevalence in Adults: A Systematic Review and Meta-Analysis. *Journal of the American Collage of Nutrition*, *36*(3), 223–233. https://doi.org/10.1080/07315724.2016.1261053.

Rokach, L., Maimon, O. (2014). *Data mining with decision trees. Theory and Applications* (2nd edition). World Scientific Publishing.

Sarma, K. S. (2007). *Predictive Modeling with SAS Enterprise Miner. Practical Solutions for Business Applications*. SAS Institute.

Scope Ratings. (2020, February 28). *Scope affirms China's sovereign rating at A+ and maintains the Outlook at Negative*. Pobrane 11 marca 2020 r. z https://scoperatings.com/#!search/research/detail/162598EN.

Surkova, E., Nikolayevsky, V., Drobniewski, F. (2020). False-positive COVID-19 results: hidden problems and costs. *The Lancet. Respiratory Medicine*, *8*(12), 1167–1168. https://doi.org/10.1016/S2213-2600(20)30453-7.

Turner, B. (2020, April 3). *'Most significant crisis in the history of travel': where to now for tourism?*. https://www.smh.com.au/business/the-economy/most-significant-crisis-in-the-history-of-travel-where-to-now-for-tourism-20200227-p5450j.html.

Tyko, K., Guynn, J., Snider, M. (2020, February 28). *Coronavirus fears empty store shelves of toilet paper, bottled water, masks as shoppers stock up*. https://eu.usatoday.com/story/money/2020/02/28/coronavirus-2020-preparation-more-supply-shortages-expected/4903322002/.

Weeks, D. G., Michela, J. L., Bragg, M. E. (1980). *Relation Between Loneliness and Depression: A Structural Equation Analysis*. American Psychological Association.

World Food Programme. (2021, May 5). *Global Report on Food Crises – 2021*. https://www.wfp.org/publications/global-report-food-crises-2021.

World Health Organization. (b.r.). *Coronavirus disease (COVID-19) pandemic*. Pobrane 10 października 2021 r. z https://www.who.int/.

World Health Organization. (2020, October 5). *COVID-19 disrupting mental health services in most countries, WHO survey*. https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey.

Wozniak, A., Willey, J., Benz, J., Hart, N. (2020). *COVID Impact Survey*. National Opinion Research Center.

Yelp Economic Average. (2020, September). *Local Economic Impact Report*. https://www.yelpeconomicaverage.com/business-closures-update-sep-2020.

Yuen, K. F., Wang, X., Ma, F., Li, K. X. (2020). The Psychological Causes of Panic Buying Following a Health Crisis. *International Journal of Environmental Research and Public Health*, *17*(10), 1–14. https://doi.org/10.3390/ijerph17103513.