Zeszyty Naukowe SGSP 2022, Nr 81, s. 7–28
ISSN: 0239-5223; e-ISSN: 2720-0779
Creative Commons Attribution 4.0 International License (CC-BY)
DOI: 10.5604/01.3001.0015.8120

mgr inż. Jan Stefan Bihałowicz
Szkoła Główna Służby Pożarniczej
e-mail: jbihalowicz@sgsp.edu.pl
ORCID: 0000-0003-3465-5315

WYKORZYSTANIE ELEMENTÓW UCZENIA MASZYNOWEGO DO MODELOWANIA STĘŻENIA ZANIECZYSZCZEŃ ATMOSFERYCZNYCH: STUDIUM PRZYPADKU PYŁU $PM_{2.5}$ W SZCZECINIE

Abstrakt

W pracy przedstawiono możliwość modelowania stężeń zanieczyszczeń w lokalizacji o określonym, stałym profilu emisji przy wykorzystaniu modeli uczenia maszynowego. Jako zanieczyszczenie wybrano pył $PM_{2.5}$, a jako zmienne objaśniające przyjęto parametry metrologiczne mierzone na stacji synoptycznej. Przeprowadzono uczenie i walidację sześciu różnych modeli na podstawie obserwacji meteorologicznych zarejestrowanych w latach 2013–2018 na stacji IMGW-PIB w Szczecinie (Polska) oraz średniodobowych stężeń pyłu $PM_{2.5}$ z tego samego okresu zmierzonych na stacji GIOŚ w Szczecinie przy ul. Andrzejewskiego, podzielonych na trzy równoliczne klasy stężeń. Dwa modele, które dawały najdokładniejsze wyniki, zostały szczegółowo przedstawione. Czułość tych modeli, w zależności od klasy stężenia pyłu, zawierała się pomiędzy 0,484 a 0,711. Te dwa modele zostały zastosowane do identyfikacji wzrostu średniodobowych stężeń w trakcie zdarzenia nietypowego – pożaru składowisk odpadów. Stężenia przewidziane w dniach, w których trwał pożar, były zaniżone względem faktycznych stężeń, co pozwala na zastosowanie modeli w identyfikacji zjawisk atypowych, które mają wpływ na stężenia zanieczyszczeń w danym miejscu.

Słowa kluczowe: modelowanie matematyczne, uczenie maszynowe, zanieczyszczenia atmosferyczne, $PM_{2.5}$, pożar

APPLICATION OF MACHINE LEARNING IN AIR POLLUTANTS MODELING: A CASE STUDY OF $PM_{2.5}$ IN SZCZECIN (POLAND)

Abstract

The work presents the possibilities of using machine learning in modeling pollutant concentrations at locations with defined constant sources of emission. The $PM_{2.5}$ was chosen as the pollutant to be studied with meteorological variables as exogenous variables measured at a weather station. Six different models were implemented and cross-validated on meteorological data recorded in 2013-2018 at the Institute of Meteorology and Water Management station in Szczecin, Poland, and $PM_{2.5}$ concentrations from the same period divided into three classes, measured at the air quality station of the Chief Inspectorate of Environmental Protection (Poland) located at Andrzejewskiego Street in Szczecin. Two best-performing models were described in detail. The sensitivity of the models was found to vary from 0.484 to 0.711 depending on the class of $PM_{2.5}$ concentration. Those two models were then applied to identify increases in PM concentrations that were caused by an extraordinary incident – landfill fire. It was proven that the predicted values of concentration that occur during the fire were underestimated as compared to actual concentration levels and hence such models can be applied in the identification of abnormal phenomena that may affect the concentrations of pollutants in a given location.

Keywords: mathematical modeling, machine learning, air pollutants, $PM_{2.5}$, fire

1. Wprowadzenie

Jakość powietrza atmosferycznego jest aspektem zdrowia publicznego. Pomimo rosnącej świadomości społecznej i wysiłkom, dzięki którym stężenie pyłu zawieszonego (PM) jest monitorowane w coraz większej liczbie miejsc, wciąż są miasta i osiedla, w których jakość powietrza, a co najmniej stężenie PM, nie są monitorowane lub monitorowane okresowo za pomocą mobilnych stacji monitoringu jakości powietrza. Mobilne stacje monitoringu powietrza są lokalizowane na określony czas w miejscach, w których nie jest prowadzony stały monitoring jakości powietrza. W trakcie eksploatacji stacji w danym miejscu zbierane są dane dotyczące jakości powietrza i dane meteorologiczne; przeważnie zakłada się okres roku trwania takich pomiarów [1–3]. Uzyskanie danych z okresowego monitoringu jakości powietrza referencyjnymi metodami jest celowe, gdyż można w dalszej kolejności wykorzystać je do stworzenia modeli prognozujących jakość powietrza w danym miejscu również po zaprzestaniu monitoringu w danym receptorze, pod warunkiem, że zarówno profil emisyjny w otoczeniu receptora, jak i warunki meteorologiczne i topograficzne w otoczeniu nie uległy istotnym zmianom. Jednym z podejść, które może pozwolić na takie prognozowanie, jest uczenie maszynowe (ang. *machine learning*, ML). ML jest szeroko wykorzystywane do modelowania różnych procesów zachodzących w środowisku, takich jak zjawiska związane z burzami [4] i opadami [5], nasłonecznienie [6], poziom wód gruntowych [7] i stę-

żenia PM [8, 9]. Najczęściej stężenie PM jest modelowane na podstawie danych meteorologicznych [10–12], których pomiar jest łatwiejszy i powszechniejszy.

Rosnące możliwości obliczeniowe komputerów osobistych oraz powstające pakiety wolnego oprogramowania i oprogramowania open-source pozwalają na rozpowszechnienie możliwości modelowania ML; tym samym przestaje być ono zarezerwowane tylko dla tzw. superkomputerów. Niniejsza praca ma na celu zaprezentować możliwości wykorzystania różnych modeli klasyfikacyjnych ML do prognozowania stężeń zanieczyszczeń i proces wyboru modelu. Jako badane zanieczyszczenie wybrano pył $PM_{2.5}$ na stacji Głównego Inspektoratu Ochrony Środowiska (GIOŚ) w Szczecinie przy ul. Andrzejewskiego, a jako zmienne objaśniające przyjęto dane meteorologiczne mierzone przez stację synoptyczną Instytutu Meteorologii i Gospodarki Wodnej – Państwowego Instytutu Badawczego (IMGW-PIB). Do modelowania użyto pakietu narzędzi open-source Orange [13], który pozwala na eksplorację danych przy wykorzystaniu programowania wizualnego.

2. Materiały i metody

2.1. Dane o jakości powietrza – $PM_{2.5}$

Jakość powietrza atmosferycznego jest monitorowana w Polsce przez Główny Inspektorat Ochrony Środowiska [14], a dane dotyczące stężeń poszczególnych substancji są gromadzone i przekazywane do Europejskiej Agencji Środowiskowej (EEA) [15]. W pracy wykorzystano udostępnione przez EEA dane ze stacji w Szczecinie przy ul. Andrzejewskiego (53°22'51.5"N 14°39'48.1"E). Na tej stacji wykonywane są m.in. zautomatyzowane pomiary średnich stężeń pyłu zawieszonego $PM_{2.5}$ oraz PM_{10} z czasem uśredniania 1 h, manualne pomiary stężenia $PM_{2.5}$ oraz PM_{10} z czasem uśredniania 24 h oraz określana jest zawartość metali i wielopierścieniowych węglowodorów aromatycznych w pyle PM_{10}. W pracy wykorzystano codzienne dane o stężeniu pyłu $PM_{2.5}$ (pomiar manualny) z okresu 1.01.2013–31.12.2018 [15]. Czas pokrycia pomiarami w tym okresie wynosił 2144 dni (97,9%).

2.2. Dane meteorologiczne

Monitorowaniem, zbieraniem, przetwarzaniem i udostępnianiem danych dotyczących warunków atmosferycznych i meteorologicznych w Polsce zajmuje się Instytut Meteorologii i Gospodarki Wodnej – Państwowy Instytut Badawczy (IMGW-PIB). Dane ze stacji meteorologicznych udostępniane są przez portal Dane Publiczne IMGW-PIB [16]. W pracy wykorzystano dane ze stacji meteorolo-

Creative Commons Attribution 4.0 International License (CC-BY)

gicznej I rzędu (tzw. stacji synoptycznej) zlokalizowanej w Szczecinie, w dzielnicy Dąbie (53°23'43.1"N 14°37'21.8"E), których zestawienie przedstawiono w tab. 1. Ogółem zebrano 2191 rekordów danych meteorologicznych z okresu 1.01.2013–31.12.2018. Z tego zbioru pominięto 36 dni (1,6% ogółu), podczas których nie dokonano pomiaru co najmniej jednego z powyżej wymienionych parametrów. W dalszej analizie wykorzystano jedynie kompletne rekordy, tzn. kiedy jednocześnie były mierzone wszystkie parametry meteorologiczne i stężenie $PM_{2.5}$, tj. 2110 rekordów (96,3% kompletności danych).

Tab. 1. Parametry mierzone na stacji meteorologicznej IMGW-PIB w Szczecinie w latach 2013–2018 wraz z ich oznaczeniami używanymi w pracy

Mierzony parametr	**Oznaczenie**	**Jednostka**
dobowa temperatura maksymalna	*TMAX*	°C
dobowa temperatura minimalna	*TMIN*	°C
średnia dobowa temperatura	*STD*	°C
minimalna temperatura przy gruncie	*TMNG*	°C
suma dobowa opadu	*SMDB*	mm
suma opadu w ciągu dnia	*WODZ*	mm
suma opadu w ciągu nocy	*WONO*	mm
rodzaj opadu	*OPAD*	mm
wysokość pokrywy śnieżnej	*PKNS*	cm
równoważnik wodny śniegu	*RWSN*	$\frac{\text{mm}}{\text{cm}}$
usłonecznienie	*USL*	h
czas trwania opadu deszczu	*DESZ*	h
czas trwania opadu śniegu	*SNEG*	h
czas trwania opadu deszczu ze śniegiem	*DISN*	h
czas trwania gradu	*GRAD*	h
czas trwania mgły	*GRAD*	h
czas trwania zamglenia	*MGLA*	h
czas trwania sadzi	*SADZ*	h
czas trwania gołoledzi	*GOLO*	h
czas trwania zamieci śnieżnej niskiej	*ZMNI*	h
czas trwania zamieci śnieżnej wysokiej	*ZMWS*	h
czas trwania zmętnienia	*ZMET*	h
czas trwania wiatru ≥ 10 m/s	*FF10*	h

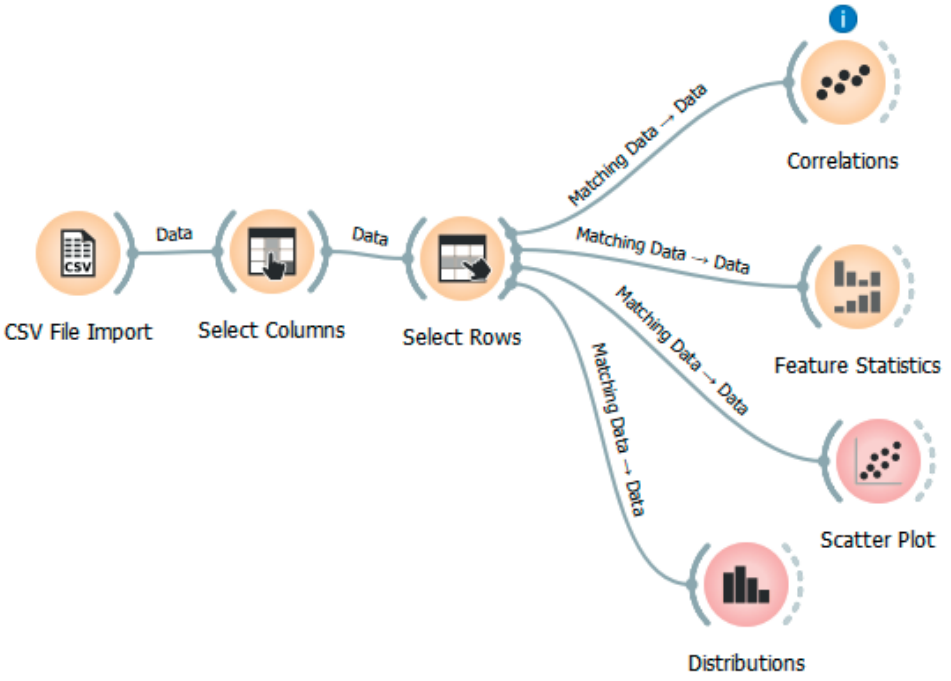
Mierzony parametr	Oznaczenie	Jednostka
czas trwania wiatru > 15 m/s	*FF15*	h
średnia dobowa prędkość wiatru	*FWS*	$\frac{\text{m}}{\text{s}}$
czas trwania burzy	*BRZA*	h
czas trwania rosy	*ROSA*	h
czas trwania szronu	*SZRO*	h
izoterma dolna	*IZD*	cm
izoterma górna	*IZG*	cm
średnie dobowe zachmurzenie ogólne	*NOS*	oktany
średnie dobowe ciśnienie pary wodnej	*CPW*	hPa
średnia dobowa wilgotność względna	*WLGS*	%
średnie dobowe ciśnienie na poziomie stacji	*PPPS*	hPa
wystąpienie pokrywy śnieżnej	*DZPS*	zmienna binarna 0/1
wystąpienie błyskawicy	*DZBL*	zmienna binarna 0/1
stan gruntu	*STAN*	zmienna binarna 0/1

Źródło: opracowanie własne

2.3. Orange

Orange jest narzędziem opartym na języku programowania Python służącym do ML, eksploracji i obrazowania danych [13]. Programowanie w tym narzędziu odbywa się wizualnie, tzn. funkcje są reprezentowane przez graficzne obiekty – widżety, a przekazywanie danych między nimi odbywa się wzdłuż graficznych linii, które są rysowane przez użytkownika. Przykładowy program służący do identyfikacji korelacji w zbiorze danych, pozyskania podstawowych statystyk, prezentowania punktów na wykresie punktowym oraz dopasowywania rozkładów empirycznych zaprezentowano na rys. 1. Zastosowania programu są bardzo szerokie, gdyż wśród widżetów można znaleźć funkcje dedykowane nadzorowanemu i nienadzorowanemu ML, bioinformatyce, analizie szeregów czasowych, informacji przestrzennej i eksploracji tekstów.

Creative Commons Attribution 4.0 International License (CC-BY)



Rys. 1. Przykładowy program w Orange pozwalający na wyznaczenie współczynników korelacji, parametrów rozkładów zmiennych, statystyk opisowych zmiennych oraz rysowanie wykresów punktowych

Źródło: opracowanie własne

2.4. Modele klasyfikacyjne

2.4.1. *k*-NN

Model *k* najbliższych sąsiadów (*k* nearest neighbors, *k*-NN) został wprowadzony w latach 50. XX wieku w Szkole Medycyny Lotniczej Sił Powietrznych Stanów Zjednoczonych [17]. Pozwala on na przewidywanie klasy przynależności punktu na podstawie odległości do *k* najbliższych sąsiadów. Danymi początkowymi jest zbiór punktów $X = \{x_i \in \mathbb{R}^n\}_{i \in I}$ należących do przestrzeni liniowej $\mathbb{R}^n$ (punkty *n* zmiennych objaśniających, *I* to zbiór indeksów) oraz przypisane im kategorie $Y = \{y_i\}_{i \in I}$ o wartościach $y_i \in C$. Razem stanowią one zbiór par $A = \{(x_i, y_i) \in \mathbb{R}^n \times C\}_{i \in I}$. Aby móc określać *k* najbliższych sąsiadów, zbiór *X* musi mieć metrykę $d(x_i, x_j)$. W szczególności przestrzeń $\mathbb{R}^n$ jako przestrzeń współrzędnych może mieć wprowadzoną jedną z wielu norm, a przez to wprowadzoną metrykę generowaną przez normę. W praktyce najczęściej używane są metryki generowane przez *p*-normy $d(x_i, x_j) = || x_j - x_{kp} || = (\sum_{l=1}^{n} | x_i^l - x_j^l |^p)^{\frac{1}{p}}$ oraz odległość Czebyszewa (maksimum)

$d(x_i, x_j) = \max_{l=1,\ldots,n} |x_i^l - x_j^l|$. Aby przewidzieć, do jakiej klasy $y_0 \in C$ należy punkt x_0, trzeba wyznaczyć odległości pomiędzy punktami zbioru X a x_0, tworząc zbiór $D = \{d(x_0, x_i)\}_{i \in I}$. Algorytm *k*-NN wymaga przekształcenia zbioru par uporządkowanych $D \times Y$ w zbiór skierowany $(D \times Y, \leq)$ z quasi-porządkiem zadanym przez relację niewiększości metryki. Z tak uporządkowanego zbioru można wybrać zbiór *k* najmniejszych elementów $\{(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), \ldots, (x_{(k)}, y_{(k)})\}$ – są to najbliżsi sąsiedzi x_0. Wartość y_0 określa się na podstawie najczęściej występującej wartości w zbiorze $\{y_i\}_{i \in (1),\ldots,(k)}$. Często, aby zminimalizować wpływ „odległych" najbliższych sąsiadów, tworzona jest klasyfikacja ważona, najczęstszą wagą jest odwrotność odległości od x_0.

Algorytm *k*-NN do modelowania stężenia pyłu $PM_{2.5}$ zostanie wykorzystany przy założeniach, że:

- $k = 5$;
- zbiór *I* przebiega od $i = 1$ do $i = 2110$;
- punkty x_i są elementami przestrzeni $\mathbb{R}^{32} \times \{0,1\}^3$;
- jako normę przyjęto normę euklidesową (*p*-normę dla $p = 2$);
- dla przewidywania klasy przyjęto wagi $\frac{1}{d(x_0, x_i)}$.

2.4.2. Tree (drzewo)

Drzewa mają szerokie zastosowanie w różnych obszarach nauki. Mogą przedstawiać procesy decyzyjne, zależności i hierarchię oraz są stosowane jako struktura danych. Szczególne znaczenie mają drzewa binarne, które dla każdego wierzchołka mają dokładnie dwa liście (lewego i prawego syna). Drzewa binarne służą do przedstawienia modeli klasyfikacyjnych. Dane wejściowe dla algorytmu są takie jak dla algorytmu k-NN, tzn. pary $A = \{(x_i, y_i) \in \mathbb{R}^n \times C\}_{i \in I}$. Istotą modelu jest to, aby na podstawie warunków dotyczących zmiennych składających się na x_i stworzyć algorytm decyzyjny określający, do której klasy z *C* należy dany punkt. Liczba warunków (poziomów) jest ustalana arbitralnie, a warunki mogą dotyczyć różnych zmiennych lub różne warunki mogą kilkukrotnie dotyczyć tej samej zmiennej. Podział danych na liście może być dokonany przy pomocy różnych algorytmów – miar, które sprawdzają spójność danych wewnątrz liści. Dwiema najczęściej stosowanymi miarami są: niespójność Giniego i zysk informacyjny. Niespójność Giniego I_G opisuje, jak często popełniany byłby błąd w klasyfikowaniu danych, jeśli punkty byłyby zaklasyfikowane w sposób losowy, zgodny z rozkładem klas w liściu. Oznaczając jako p_c część elementów w liściu zaklasyfikowanych klasą *c*, to można wyrazić I_G jako $I_G = \sum_{c=1}^{|C|}(p_c \sum_{c \neq c'} p_{c'}) = 1 - \sum_{c=1}^{|C|} p_c^2$. Oznacza to, że jeśli wszystkie elementy zaklasyfikowane do jednego liścia miałyby tę samą klasę, to $I_G = 0$. Druga metoda oceny spójności liścia wywodzi się z teorii informacji. Każdej zmiennej losowej przyjmującej wartości dyskretne można przypisać miarę – entropię Sha-

Creative Commons Attribution 4.0 International License (CC-BY)

nona. Dla modelu klasyfikacyjnego, z klasami ze zbioru C można określić entropię zbioru B jako $H(B) = -\sum_{c=1}^{|C|} p_c \log p_c$. Zysk informacyjny IG to różnica między entropią drzewa (węzła) W a entropią liści (synów) L, $IG = H(T) - H(L_1) - H(L_2)$. Podział danych na liście następuje w wyniku maksymalizacji zysku informacyjnego. Na podstawie tak wyuczonego drzewa można dokonać przewidywania klasy ze zbioru C na podstawie dowolnych danych z przestrzeni punktów liniowej $\mathbb{R}^n$.

W ramach tworzenia drzewa klasyfikacyjnego dla pyłu $PM_{2.5}$ w programie Orange stosowane jest maksymalizowanie IG. Dodatkowo zostaną przyjęte ograniczenia:

- liście o mniej niż 5 elementach nie będą dzielone;
- w wyniku podziału nie może powstać liść mający mniej niż dwa elementy;
- maksymalna wysokość drzewa wynosi 100;
- liście nie będą dzielone, jeśli co najmniej 95% danych w liściu należy do jednej klasy.

2.4.3. Random Forest (losowy las decyzyjny)

Random Forest jest rozwinięciem koncepcji drzewa decyzyjnego. W tym algorytmie powstaje nie jedno, a wiele drzew klasyfikacyjnych. Tworzenie poszczególnego drzewa i określanie spójności liści jest analogiczne jak w powyżej omówionym przypadku. Każde drzewo uczone jest na podstawie własnej próbki danych, wylosowanych ze zbioru A, ze zwracaniem (metodą samowsporną, ang. *bootstrap*). W wyniku takiej konstrukcji „lasu", dla dowolnego punktu x_i każde drzewo może wskazywać na inną klasę należącą do zbioru C. Ostateczne zaklasyfikowanie może być uzyskane jako średnia z klas przydzielonych przez każde z drzew lub jako ta klasa, na którą większość drzew „zagłosowała". Tak wyuczony zbiór drzew pozwala lepiej przewidzieć klasę ze zbioru C dla nowego punktu niż pojedyncze drzewo [18].

Algorytm zaimplementowany w Orange przewiduje klasę na podstawie klasy wskazanej przez większość drzew. Przy klasyfikacji stężeń pyłu $PM_{2.5}$ przyjęto 10 drzew oraz że nie będą dzielone liście o liczbie elementów mniejszej niż 5.

2.4.4. Logistic Regression (regresja logistyczna)

Model regresji logistycznej ma zastosowanie do przewidywania zmiennych dyskretnych. W regresji logistycznej używa się pojęcia szansy l, która jest stosunkiem prawdopodobieństwa sukcesu p do prawdopodobieństwa porażki $1 - p$. W modelu logarytm naturalny szansy, który często nazywany jest funkcją logitową, jest opisany jako kombinacja liniowa n zmiennych $(x_i^1, x_i^2, \ldots, x_i^n)$, tj. $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{n} \beta_j x_i^j$. To równanie można również przekształcić, aby uzy-

skać wzór na $p = \frac{\exp(\beta_0 + \sum_{j=1}^{n} \beta_j x_i^j)}{1 + \exp(\beta_0 + \sum_{j=1}^{n} \beta_j x_i^j)}$. Uczenie modelu regresji logistycznej polega na wyznaczeniu współczynników β tej kombinacji. W przypadku gdy zbiór C posiada więcej niż dwie wartości, trzeba zastosować multimianową regresję logistyczną (ang. *multinomial logistic regression*). Dla każdej klasy c ze zbioru C można zapisać prawdopodobieństwo jej wystąpienia $p(C = c_k) = \frac{\exp(\beta_k x_i^j)}{1 + \exp(\beta_0 + \sum_{j=1}^{|C|-1} \beta_j x_i^j)}$ dla $k \leq |C| - 1$ oraz $p(C = c_{|C|}) = \frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{|C|-1} \beta_j x_i^j)}$ – przy takim sformułowaniu modelu multimianowego ważna jest niezależność zmiennych.

W modelu zastosowanym w niniejszej pracy wykorzystano multimianową regresję logistyczną, gdyż $|C| = 3$ oraz przyjęto regularyzację Tichonowa [19].

2.4.5. Naive Bayes (naiwny klasyfikator bayesowski)

Naiwny klasyfikator bayesowski jest szybkim algorytmem klasyfikującym dane przy założeniu spełnienia twierdzenia Bayesa i zakłada niezależność zmiennych opisujących. Twierdzenie Bayesa zastosowane do klasyfikacji na podstawie n elementowych punktów x_i pozwala zapisać prawdopodobieństwo wystąpienia klasy c_k przy punkcie danych x_i, które jest równe $p(c_k|x_i) = \frac{p(c_k)p(x_i|c_k)}{p(x_i)}$. Licznik tego ułamka może zostać wyrażony poprzez poszczególne zmienne $p(c_k)p(x_i|c_k) = p(c_k)p(x_i^1, x_i^2, \ldots, x_i^n|c_k) = p(c_k)p(x_i^1|c_k)p(x_i^2, \ldots, x_i^n|c_k, x_i^1) =$ $p(c_k)p(x_i^1|c_k)p(x_i^2|c_k, x_i^1)p(x_i^3, \ldots, x_i^n|c_k, x_i^1, x_i^2) = p(c_k)p(x_i^1|c_k)p(x_i^2|c_k, x_i^1) \cdot \ldots \cdot$ $p(x_i^n|c_k, x_i^1, x_i^2, \ldots, x_i^n)$. Kluczowym założeniem tego modelu, które często jest nazywane naiwnym, jest to, że zmienne są niezależne, tj. $\forall_{l \neq m} p(x_i^l|x_i^m) = p(x_i^l)$. Przy tym założeniu licznik ułamka na prawdopodobieństwo warunkowe upraszcza się do $p(c_k)p(x_i^1|c_k)p(x_i^2|c_k) \cdot \ldots \cdot p(x_i^n|c_k) = p(c_k)\prod_{j=1}^{n} p(x_i^j|c_k)$. Przyporządkowanie klasy w modelowaniu tym klasyfikatorem polega na znalezieniu takiego k, dla którego wyrażenie $p(c_k)\prod_{j=1}^{n} p(x_i^j|c_k)$ przyjmuje wartość maksymalną.

W modelowaniu stężenia $PM_{2.5}$ w programie Orange nie przyjęto żadnych szczegółowych ograniczeń i założeń poza przytoczonymi powyżej.

2.4.6. Neural Network (sieć neuronowa)

Sztuczny neuron, używany w uczeniu maszynowym, to funkcja matematyczna mająca modelować działanie biologicznego neuronu. Podobnie jak w przypadku

Creative Commons Attribution 4.0 International License (CC-BY)

neuronu, sztuczny neuron otrzymuje jeden lub więcej argumentów (odpowiedniki potencjałów na dendrytach) i na podstawie kombinacji liniowej tych argumentów neuron jest aktywowany lub nie i otrzymywana jest odpowiedź. Aktywacja neuronu zależy od funkcji aktywacji neuronu. Najczęściej funkcjami aktywacyjnymi są funkcje grzbietowe takie jak funkcja logistyczna, funkcja Heaviside'a, funkcja prostownika (ang. *rectified linear unit*, ReLU) albo funkcje radialne. Przykładem sieci neuronowej, która jest zaimplementowana w Orange, jest perceptron wielowarstwowy. W tej sieci występują co najmniej trzy warstwy neuronów: warstwa wejściowa (neurony tej warstwy nie mają funkcji aktywacji), co najmniej jedna warstwa ukryta i warstwa wyjściowa. Każdy neuron z danej warstwy dostarcza argumenty do wszystkich neuronów następnej warstwy, przy czym każdy neuron następnej warstwy ma własny współczynnik wagi (współczynnik w kombinacji liniowej) dla argumentu. Uczenie modelu polega na wyznaczeniu tych wag tak, aby zminimalizować różnice w dopasowaniach w całej sieci za pomocą algorytmu propagacji wstecznej [20].

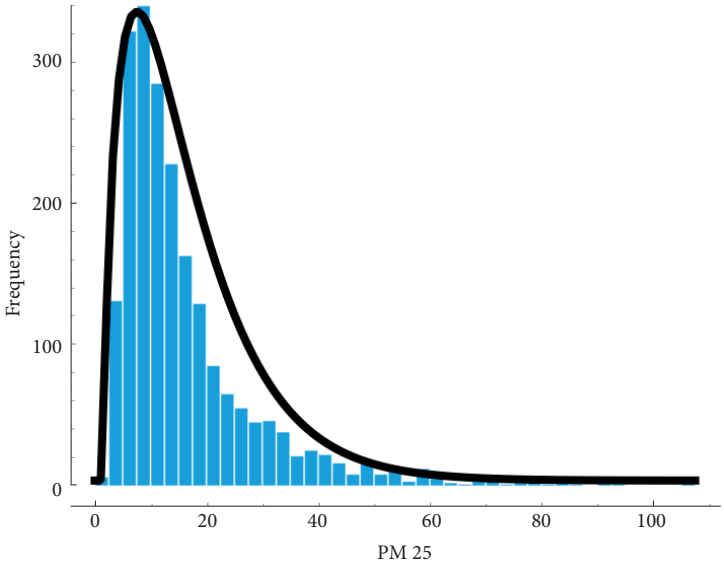
Do modelowania stężenia pyłu $PM_{2.5}$ użyto sieci neuronowej z jedną ukrytą warstwą zawierającą 100 neuronów. Wszystkie neurony były aktywowane funkcją ReLU, a wagi wyznaczane przy pomocy solvera Adam [21].

3. Wyniki i dyskusja

3.1. Statystyki opisowe

Przeanalizowano stężenia pyłu $PM_{2.5}$ na stacji w Szczecinie w latach 2013–2018 i ich rozkład został przedstawiony na rys. 2. Rozkład ten jest prawostronnie asymetryczny, współczynnik skośności zdefiniowany jako trzeci moment centralny m_3 przez sześcian odchylenia standardowego s^3 wynosi $A_3 = \frac{m_3}{s^3} = 2{,}30$. Średnie stężenie wynosiło $\langle PM_{2.5} \rangle = 16{,}2 \frac{\mu g}{m^3}$, podczas gdy mediana wynosiła $Me_{PM_{2.5}} = 12{,}1 \frac{\mu g}{m^3}$, a wartość najczęściej występująca to $Mo_{PM_{2.5}} = 9{,}97 \frac{\mu g}{m^3}$. Kurtoza rozkładu wynosi Kurt = 7,06, co oznacza, że zgodnie z rozkładami Pearsona stężenie $PM_{2.5}$ może zostać opisane za pomocą rozkładu Pearsona I [22] (uogólnienia rozkładu beta [23]). Za pomocą programu Orange sparametryzowano rozkład, uzyskując parametry kształtu $\alpha = 1{,}66, \beta = 5{,}76 \cdot 10^8$ (rys. 2). Minimalne zmierzone stężenie wynosiło $Min_{PM_{2.5}} = 1{,}63 \frac{\mu g}{m^3}$, maksymalne $Max_{PM_{2.5}} = 105{,}55 \frac{\mu g}{m^3}$, a pierwszy i trzeci kwartyl wynosiły odpowiednio $Q1_{PM_{2.5}} = 7{,}94 \frac{\mu g}{m^3}$ i $Q3_{PM_{2.5}} = 19{,}4 \frac{\mu g}{m^3}$. Do dalszej analizy dane podzielono na trzy równoliczne klasy stężeń dobowych $PM_{2.5}$, oznaczane *I*, *II* oraz *III*, tj. $I = \left[0; 9{,}265\right) \frac{\mu g}{m^3}$, $II = \left[9{,}265; 16{,}425\right) \frac{\mu g}{m^3}$ oraz $III = \left[16{,}425; +\infty\right) \frac{\mu g}{m^3}$.

W związku z tym, iż modele ML, które są uczone na zrównoważonych zbiorach albo prawie zrównoważonych (mających zbliżone ilości wystąpień) są dokładniejsze, w pracy użyto podziału zrównoważonego zamiast dodatkowych algorytmów równoważących próbkowanie zbioru danych [24].



Rys. 2. Rozkład dobowych stężeń pyłu $PM_{2.5}$ zmierzonych w latach 2013–2018 na stacji w Szczecinie przy ul. Andrzejewskiego wraz z dopasowaną krzywą rozkładu beta

Źródło: opracowanie własne

3.2. Korelacje

Wszystkie parametry meteorologiczne mierzone na stacji w Szczecinie zostały przetestowane na istnienie korelacji pomiędzy nimi. Z racji, iż większość parametrów nie posiada rozkładu normalnego, przyjęto jako miarę korelacji współczynnik korelacji rang Spearmana ρ. Istotność współczynnika testowano na podstawie zmiennej z, która podlega rozkładowi normalnemu $N(0;1)$, tj. $z = \sqrt{\frac{n-3}{1{,}06}}\operatorname{atanh}(\rho) \sim N(0;1)$, gdzie n to liczba pomiarów. Wartości współczynnika ρ wraz z poziomem istotności przedstawiono w tab. 1. Z racji wysokiej liczby pomiarów (2110 punktów), 28 zmiennych wykazywało korelację na poziomie istotności poniżej $\alpha = 0{,}01$, jedna na poziomie poniżej $\alpha = 0{,}05$ i jedna na poziomie poniżej $\alpha = 0{,}1$. Stężenie $PM_{2.5}$ było najbardziej skorelowane z czasem trwania zamglenia *ZMGL*. Zamglenie mierzone na stacjach IMGW może być związane ze spadkiem widzialności, co do której wykazano, że jej spadek związany jest z pogorszeniem jakości powietrza atmosferycznego [25–27].

Creative Commons Attribution 4.0 International License (CC-BY)

Tab. 2. Korelacje rang Spearmana ρ pomiędzy stężeniem pyłu $PM_{2.5}$ i zmiennymi meteorologicznymi mierzonymi na stacji w Szczecinie w latach 2013–2018

zmienna	ρ	zmienna	ρ	zmienna	ρ	zmienna	ρ	zmienna	ρ	zmienna	ρ
ZMGL	0,542	*CPW*	-0,263	*WODZ*	-0,208	*ROSA*	-0,144	*SADZ*	0,083	*DISN*	-0,022
IZD	0,416	*DESZ*	-0,249	*PPPM*	0,187	*WONO*	-0,138	*FF10*	-0,082	*ZMWS*	0,019
TMNG	-0,349	*ZMET*	0,238	*PPPS*	0,186	*SNEG*	0,125	*GOLO*	0,073	*NOS*	0,012
TMIN	-0,339	*TMAX*	-0,237	*PKSN*	0,183	*SZRO*	0,117	*USL*	-0,063	*DZBL*	0,010
FWS	-0,325	*MGLA*	0,223	*DZPS*	0,183	*RWSN*	0,115	*GRAD*	-0,043	*BRZA*	0,009
STD	-0,283	*SMDB*	-0,211	*WLGS*	0,172	*IZG*	0,111	*ZMNI*	0,036	*FF15*	0,002

Kolorem zielonym oznaczono korelacje na poziomie istotności $\alpha = 0{,}01$, kolorem żółtym $\alpha = 0{,}05$ a pomarańczowym $\alpha = 0.1$

Źródło: opracowanie własne

3.3. Modelowanie stężeń

3.3.1. Porównanie modeli

Za pomocą modeli klasyfikacyjnych modelowano stężenie pyłu $PM_{2.5}$. Modele były poddane pięciokrotnemu sprawdzaniu krzyżowemu. Przyjęto dwie miary dokładności modeli klasyfikacyjnych: *AUC* i *CA*. *AUC* to pole powierzchni pod krzywą charakterystyki operacyjnej odbiornika (ROC), uśrednione po wszystkich klasach, a *CA* to precyzja klasyfikacji modelu. Porównanie modeli (tab. 2) wskazuje, że najlepiej działającymi modelami pod względem *AUC* (*AUC* > 0,9) są model regresji logistycznej i model sieci neuronowej. Podobnie w CA oba te modele wykazują najwyższe wartości wśród badanych modeli, *CA* > 0,87. Dodatkowo, na podstawie miary *AUC* wyznaczono prawdopodobieństwo, że model X jest lepszy niż model Y i przedstawiono je w tab. 3. Prawdopodobieństwo, że model regresji logistycznej jest lepszy niż każdy z pozostałych modeli, przekracza $p > 0{,}5$, jednakże jedynie

Tab. 3. Ocena działania modeli klasyfikacyjnych,
AUC – pole powierzchni pod krzywą ROC, CA – precyzja klasyfikacji

	AUC	CA
Regresja logistyczna	0,919	0,879
Sieć neuronowa	0,916	0,882
Losowy las decyzyjny	0,894	0,867
kNN	0,864	0,856
Naiwny klasyfikator bayesowski	0,819	0,791
Drzewo	0,713	0,715

Źródło: opracowanie własne

w przypadku czterech modeli $p \geq 0{,}9$. Przyjmując jako wartość graniczną $p_{gr} = 0{,}9$, nie można odrzucić modelu sieci neuronowej jako gorszego od modelu regresji logistycznej, zwłaszcza że wykazuje on minimalnie wyższą precyzję klasyfikacji niż model regresji logistycznej (tab. 2). W dalszej części pracy zostanie omówiona ocena jedynie tych dwóch modeli: regresji logistycznej i sieci neuronowej.

Tab. 4. Prawdopodobieństwo, że model X (w wierszu) jest lepszy niż model Y (w kolumnie) do opisu danych

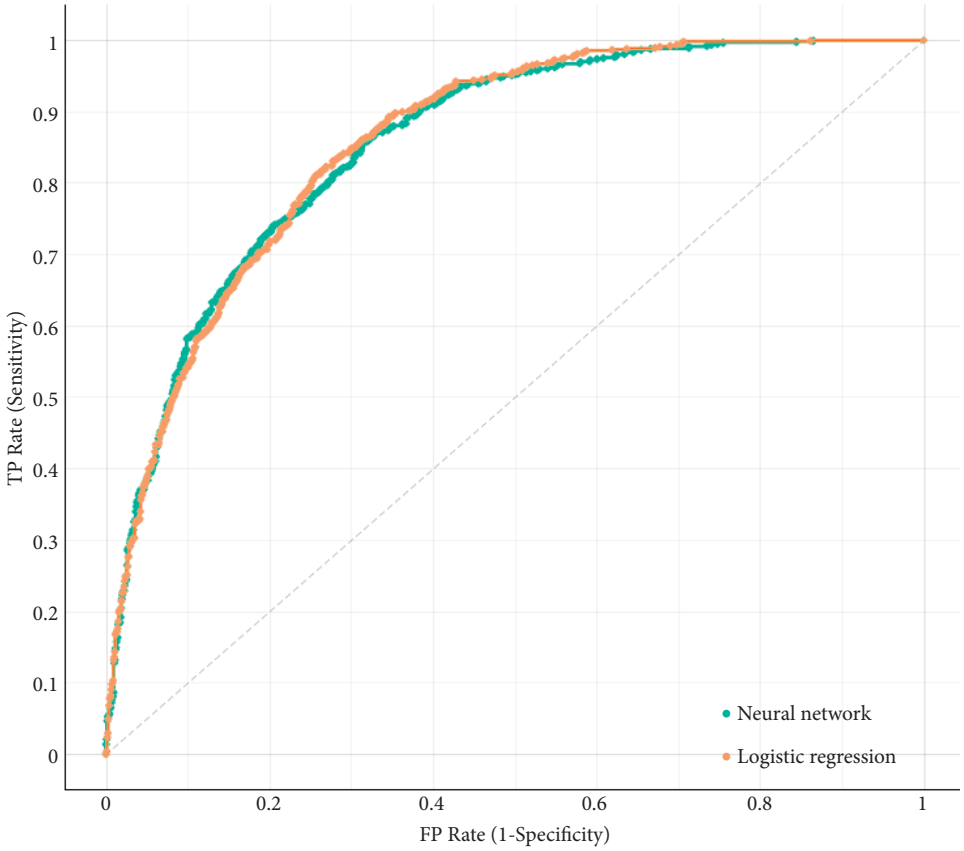
X \ Y	Regresja logistyczna	Sieć neuronowa	Losowy las decyzyjny	kNN	Naiwny klasyfikaror bayesowski	Drzewo
Regresja logistyczna		0,700	0,900	0,985	0,999	0,999
Sieć neuronowa	0,300		0,744	0,917	0,996	0,997
Losowy las decyzyjny	0,100	0,256		0,925	1,000	0,998
kNN	0,015	0,083	0,075		0,985	0,996
Naiwny klasyfikator bayesowski	0,001	0,004	0,000	0,015		0,986
Drzewo	0,001	0,003	0,002	0,004	0,014	

Źródło: opracowanie własne

3.3.2. Charakterystyka operacyjna odbiornika (ROC)

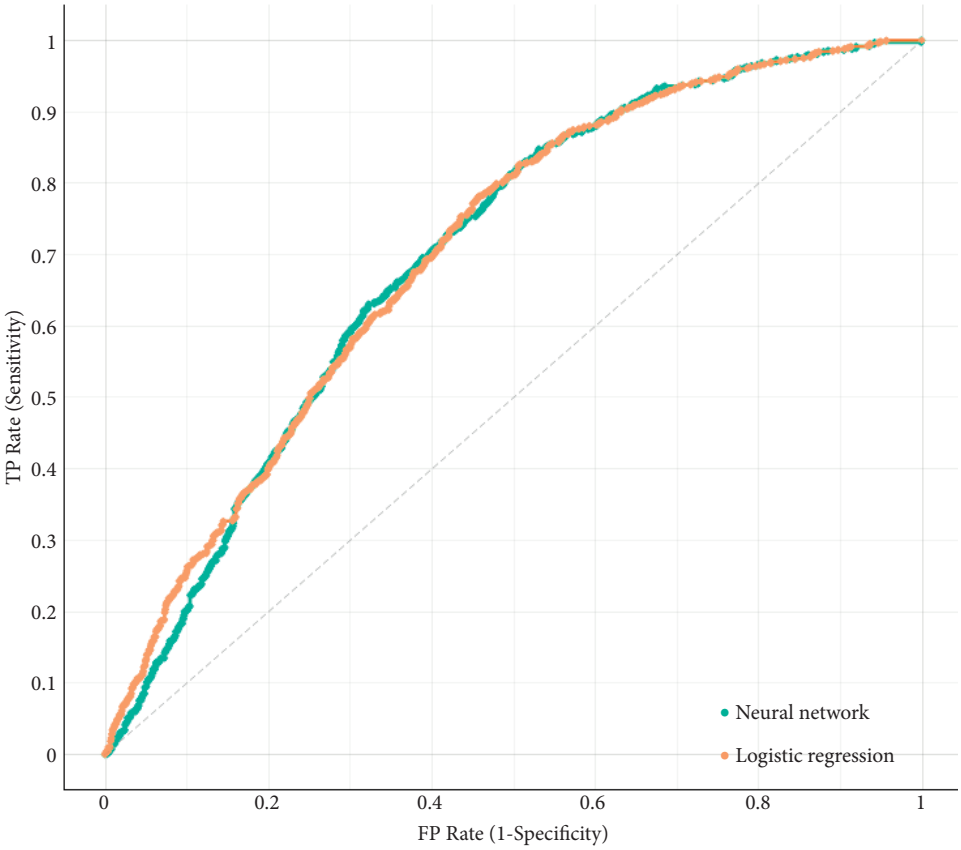
Jedną z metod oceny modelu klasyfikacyjnego jest wykreślenie charakterystyki operacyjnej odbiornika. Przedstawia ona graficznie ilość poprawnie zaklasyfikowanych punktów pomiarowych (ang. *true positive rate*, TPR) do danej klasy w funkcji błędnie zaklasyfikowanych (ang. *false positive rate*, FPR). W przypadku w pełni losowego klasyfikatora o równych prawdopodobieństwach klasyfikacji $\frac{1}{2}$ uzyskalibyśmy prostą o nachyleniu 45° na wykresie ROC, dla takiej prostej parametr $AUC = 0{,}5$. Im model lepiej klasyfikuje punkty, tym krzywa ROC jest bardziej stroma dla małych FPR, a pole powierzchni pod wykresem $AUC > 0{,}5$. Dla idealnego klasyfikatora $AUC \to 1$. Na rys. 3–5 przedstawiono ROC dla poszczególnych klas stężeń. Krzywe dla obu modeli, regresji logistycznej i sieci neuronowej, są bardzo zbliżone do siebie jedynie na rys. 4. Widać niewielką, acz systematyczną przewagę modelu regresji logistycznej nad modelem sieci neuronowej. Największe *AUC* obserwujemy dla klasy *III*, a najniższe dla klasy *II*. Nie najlepsze osiągi w „środkowej" klasie mogą być spowodowane tym, że jest ona najwęższa wśród tych klas (rozpiętość 7,16 $\frac{\mu g}{m^3}$), podczas gdy inne klasy były szersze.

Creative Commons Attribution 4.0 International License (CC-BY)



Rys. 3. Charakterystyka operacyjna odbiornika (ROC) dla klasyfikacji stężenia pyłu $PM_{2.5}$ w klasie $I = \left[0; 9{,}265\right) \frac{\mu g}{m^3}$ przez model sieci neuronowej (*neural network*) i regresji logistycznej (*logistic regression*)

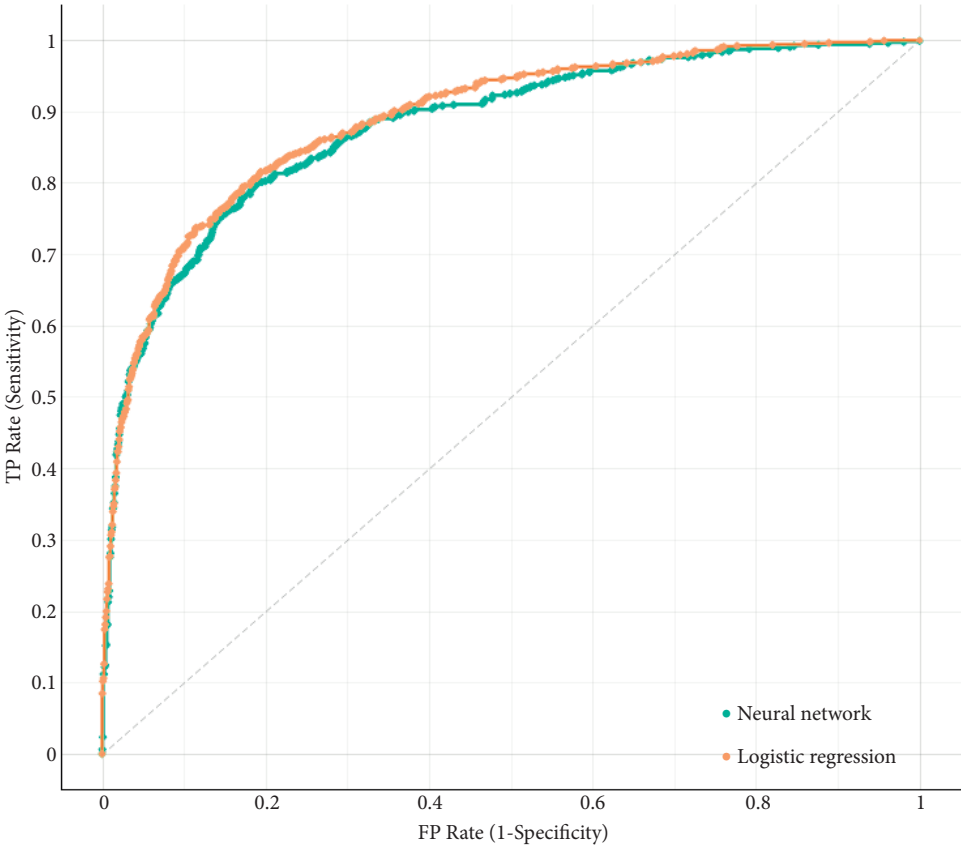
Źródło: opracowanie własne



Rys. 4. Charakterystyka operacyjna odbiornika (ROC) dla klasyfikacji stężenia pyłu $PM_{2.5}$ w klasie $II = \left[9{,}265; 16{,}425\right) \frac{\mu g}{m^3}$ przez model sieci neuronowej (*neural network*) i regresji logistycznej (*logistic regression*)

Źródło: opracowanie własne

Creative Commons Attribution 4.0 International License (CC-BY)



Rys. 5. Charakterystyka operacyjna odbiornika (ROC) dla klasyfikacji stężenia pyłu $PM_{2.5}$ w klasie $III = \left[16,425; +\infty\right) \frac{\mu g}{m^3}$ przez model sieci neuronowej (*neural network*) i regresji logistycznej (*logistic regression*)

Źródło: opracowanie własne

3.3.3. Macierz pomyłek

Macierz pomyłek służy do przedstawienia skuteczności modelu ML poprzez przedstawienie, ile danych zostało poprawnie zaklasyfikowanych (liczby na przekątnej macierzy), a ile niepoprawnie (pozostałe liczby). W tab. 4 przedstawiono macierz pomyłek dla modeli regresji logistycznej i sieci neuronowej. Wartości na diagonali dla regresji logistycznej i dla sieci neuronowej są bardzo zbliżone, co oznacza, że obydwa modele charakteryzują się bardzo zbliżoną czułością, dla klasy *I* wynosiła ona 0,706 i 0,688 odpowiednio, dla klasy *II* wynosiła 0,484 i 0,490 odpowiednio, a dla klasy *III* 0,711 i 0,708 odpowiednio. Obydwa modele miały tendencję do klasyfikowania punktów z klasy *II* do klasy niższej niż do klasy wyższej (półtora raza częściej). W pracy [12] opisano modele drzewa RUSBosted Tree – modele dla dwóch miast w Ekwadorze, Cotocollao i Belisario. Algorytm RUSBoost [28] został opracowany na potrzeby klasyfikacji silnie skośnych i niezbalansowanych danych. Modele tych miast również dzieliły dane na trzy klasy, $< 10 \frac{\mu g}{m^3}$. $[10; 25] \frac{\mu g}{m^3}$, $> 25 \frac{\mu g}{m^3}$, które zostały stworzone na podstawie aktów prawnych dotyczących jakości powietrza w Ekwadorze. Czułość modeli została przedstawiona w macierzach pomyłek. Czułość modelu dla miasta Cotocoallo wynosiła 0,763, 0,288 i 0,734 dla odpowiednich klas, podczas gdy dla miasta Belisario wynosiła odpowiednio 0,848, 0,535 i 0,484. Modele przedstawione w niniejszej pracy dla Szczecina posiadają porównywalne czułości w skrajnych klasach jak model dla miasta Cotocoallo, podczas gdy w środkowej klasie cechują się one zdecydowanie lepszą czułością. W porównaniu z modelem dla miasta Belisario czułość modelu dla Szczecina jest lepsza dla klasy obejmującej najwyższe stężenia, porównywalna dla środkowej klasy i gorsza dla klasy z najniższymi stężeniami.

Tab. 5. Macierz pomyłek dla modeli regresji logistycznej i sieci neuronowej

		Przewidziane						Σ
		Regresja logistyczna			Sieć neuronowa			
		< 9,265	**9,265 - 16,425**	**≥ 16,425**	**< 9,265**	**9,265 - 16,425**	**≥ 16,425**	
Faktyczne	**< 9,265**	496	186	21	484	194	25	**703**
	9,265 - 16,425	223	341	140	213	345	146	**704**
	≥ 16,425	50	153	500	49	156	498	**703**
Σ		**769**	**680**	**661**	**746**	**695**	**669**	**2110**

Źródło: opracowanie własne

Creative Commons Attribution 4.0 International License (CC-BY)

3.4. Zastosowanie prognozowania stężeń – pożar składowiska odpadów

Macierze pomyłek, które opisują czułość każdego z omówionych modeli, powstały w wyniku porównania faktycznej klasy średniego dobowego stężenia $PM_{2.5}$ z klasą prognozowaną przez model. Dla każdego dnia można podać prawdopodobieństwo, że wynik stężenia pyłu należy do najniższej klasy p_1, do środkowej klasy p_2 i do najwyższej klasy p_3. Suma tych prawdopodobieństw, dla każdego z modeli z osobna, wynosi $p_1 + p_2 + p_3 = 1$. Pomimo iż w wyniku modelowania wskazano tylko na jedną klasę, tę, dla której prawdopodobieństwo p było największe, można uzyskać informację również nt. prawdopodobieństwa przynależności do innych klas. W dniu 20 września 2018 r. w Szczecinie miał miejsce pożar składowiska złomu [29, 30], który objął 4620 m^2 powierzchni i trwał, łącznie z dogaszaniem, ponad 18 godzin [31]. W trakcie pożaru Wojewódki Inspektorat Ochrony Środowiska informował o znacznym pogorszeniu jakości powietrza [29]. Pożar ten odbywał się w odległości 1,23 km od stacji pomiaru jakości powietrza w Szczecinie przy ul. Andrzejewskiego. W wyniku modelowania dyspersji atmosferycznej pyłu PM_{10} wyemitowanego w tym pożarze [32] określono, że w punkcie lokalizacji stacji GIOŚ jednogodzinowe średnie stężenia wzrosły o więcej niż $100\,\frac{\mu g}{m^3}$. Brakuje szacunków wzrostu stężenia pyłu $PM_{2.5}$ z powodu braku informacji o rozkładzie rozmiarów cząstek pyłu emitowanych w wyniku tego pożaru, jednakże należy podejrzewać, że podobnie jak w przypadku pożarów innych odpadów, pył $PM_{2.5}$ stanowi istotną część pyłu PM_{10} [33], w związku z czym stężenie pyłu $PM_{2.5}$ również wzrosło. W dniu 20 września 2018 r. średniodobowe stężenie pyłu $PM_{2.5}$ należało do najwyższej klasy stężeń *III*. Model regresji logistycznej zaklasyfikował na podstawie zmiennych meteorologicznych średnie dobowe stężenie tego dnia do klasy *II*, ponieważ prawdopodobieństwo przynależności do tej klasy wynosiło 0,491, a prawdopodobieństwo przynależności do faktycznej klasy *III* wynosiło 0,124. Podobnie następnego dnia, 21 września 2018 r., kiedy wciąż trwała akcja ratowniczo-gaśnicza, model regresji logistycznej zaklasyfikował stężenie do klasy *II* z prawdopodobieństwem 0,511, podczas gdy prawdopodobieństwo przynależności do klasy *III*, do której należało faktyczne stężenie, wynosiło 0,188. Model sieci neuronowej prognozował średniodobowe stężenie dnia 20 września jako *I* z prawdopodobieństwem 0,469, a jako *III* z prawdopodobieństwem 0,110. Podobnie następnego dnia model ten niepoprawnie zaklasyfikował średniodobowe stężenie jako *II* z prawdopodobieństwem 0,894, podczas gdy prawdopodobieństwo zaklasyfikowania do poprawnej klasy wynosiło 0,053. Obydwa modele zaniżyły średniodobowe stężenie pyłu $PM_{2.5}$ w trakcie trwania pożaru składowiska odpadów, podczas gdy kolejnego dnia, 22 września, obydwa poprawnie przewidziały klasę stężenia, z prawdopodobieństwem 0,682 dla modelu regresji logistycznej i 0,935 dla modelu sieci neuronowej. W trakcie nieprzewidywalnego zjawiska emisji dużej ilości pyłu do atmosfery – pożaru – obydwa modele się nie sprawdziły.

Skupienie się na tym fragmencie macierzy pomyłek, gdzie stężenia są wysokie, a modele nie identyfikują ich jako wysokie, czyli na prawej dolnej trójkątnej części macierzy, poniżej diagonali, pozwala identyfikować sytuacje ekstremalne, epizody smogowe, pożary itp.

4. Podsumowanie

Uczenie maszynowe może być zastosowane do modelowania stężeń zanieczyszczeń z dobrym skutkiem. Proces uczenia nie obejmuje jedynie samego uczenia, ale także proces doboru odpowiedniego modelu. Modele o najlepszej charakterystyce uzyskane w wyniku procesu doboru modelu miały czułość pomiędzy 0,484 a 0,711 w zależności od modelu i klasy stężenia pyłu. Wartość 0,484 w przypadku podziału na trzy klasy może być odbierana jako lepsza niż czułość w pełni losowego modelu (0,333), jednakże mogłaby ona być lepsza. Zbalansowanie klas stężeń powoduje, że szerokość środkowej klasy jest najmniejsza ze wszystkich. Może mieć to wpływ na niską czułość w środkowej klasie. W wyniku zestawienia przewidywań modeli z informacjami o pożarze składowiska odpadów pozwalało ocenić skuteczność przewidywania modelu w trakcie zdarzeń losowych, nieprzewidywalnych, którym towarzyszy emisja pyłu. Jak można było przewidzieć, stężenie pyłu $PM_{2.5}$ przewidziane przez modele było zaniżone względem faktycznego. Pozwala to spojrzeć na modelowanie z innej strony – podczas gdy modele dążą do uzyskania największej dokładności, można zacząć przyglądać się danym, dla których model nie zaprognozował stężeń poprawnie. Macierze pomyłek tych modeli wskazują, że częściej, około 1,15 raza, klasa stężenia była zaniżana niż zawyżana. Analiza sytuacji, w których stężenie pyłu jest zaniżone przez modele, może mieć zastosowanie do identyfikowania anomalii w profilu emisji takich jak wystąpienie pożaru, epizodu smogowego. Wskazana jest kontynuacja prac nad modelami przewidywania stężeń zanieczyszczeń, w większej liczbie klas, przy wykorzystaniu innych modeli ML. Wykorzystanie w uczeniu modelu zbioru pozbawionego zdarzeń ekstremalnych, takich jak np. duże i bardzo duże pożary [34] lub inne epizodyczne emitery zanieczyszczeń znajdujące się w bardzo bliskim otoczeniu stacji pomiaru jakości powietrza, pozwoli skonstruować model, który dobrze wykrywa anomalie. Model wyuczony na zbiorze bez anomalii będzie pozwalał na analizę stężeń na bieżąco, wskazywanie, czy zarejestrowane stężenie zgadza się z przewidywanym. Dokładna analiza każdej z przyszłych anomalii pozwoli na identyfikację jej źródła i ewentualne podjęcie działań zapobiegających ponownemu wystąpieniu niekontrolowanego wzrostu stężeń spowodowanego przez zidentyfikowane źródło.

Creative Commons Attribution 4.0 International License (CC-BY)

Podziękowania

Autor chciałby podziękować dr hab. inż. Wioletcie Roguli-Kozłowskiej, prof. ucz. ze Szkoły Głównej Służby Pożarniczej oraz st. bryg. dr. hab. inż. Adamowi Krasuskiemu, prof. ucz. ze Szkoły Głównej Służby Pożarniczej za pomoc i cenne uwagi podczas przygotowania niniejszej pracy.

Praca została zrealizowana w ramach projektu *Wpływ pożarów składowisk odpadów na jakość powietrza atmosferycznego – metodyka oraz oszacowanie wartości emisji* 2020/37/N/ST10/02997 przyznanego autorowi w ramach konkursu PRELUDIUM 19 finansowanego przez Narodowe Centrum Nauki (Kraków, Polska).

Realizacja pracy została wsparta przez środki Ministerstwa Spraw Wewnętrznych i Administracji oraz Ministerstwa Edukacji i Nauki na prowadzenie działalności badawczej przekazane Szkole Głównej Służby Pożarniczej.

References/Bibliografia

1. *GIOŚ Przeniesienie mobilnej stacji monitoringu powietrza z Ostrowca Świętokrzyskiego do Sandomierza*, https://powietrze.gios.gov.pl/pjp/rwms/content/show/2352 (dostęp: 14.11.2021).
2. *GIOŚ Przeniesienie mobilnej stacji monitoringu powietrza z Jędrzejowa do Opatowa*, https://powietrze.gios.gov.pl/pjp/rwms/content/show/2353 (dostęp: 14.11.2021).
3. *GIOŚ Przeniesienie mobilnej stacji monitoringu powietrza z Połczyna-Zdroju do Kołobrzegu*, https://powietrze.gios.gov.pl/pjp/rwms/content/show/2418 (dostęp: 14.11.2021).
4. Tang R., Zeng F., Chen Z., Wang J.-S., Huang C.-M., Wu Z., *The Comparison of Predicting Storm-Time Ionospheric TEC by Three Methods: ARIMA, LSTM, and Seq2Seq*, „Atmosphere (Basel)" 2020, 11, 316, doi:10.3390/atmos11040316.
5. Poornima S., Pushpalatha M., *Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units*, „Atmosphere (Basel)" 2019, 10, 668, doi:10.3390/atmos10110668.
6. Voyant C., Notton G., Kalogirou S., Nivet M.-L., Paoli C., Motte F., Fouilloy A., *Machine learning methods for solar radiation forecasting: A review*, „Renew. Energy" 2017, 105, 569–582, doi:10.1016/j.renene.2016.12.095.
7. Nikolos I.K., Stergiadi M., Papadopoulou M.P., Karatzas G.P., *Artificial neural networks as an alternative approach to groundwater numerical modelling and environmental design*, „Hydrol. Process" 2008, 22, 3337–3348, doi:10.1002/hyp.6916.
8. Shahriar S.A., Kayes I., Hasan K., Hasan M., Islam R., Awang N.R., Hamzah Z., Rak A.E., Salam M.A., *Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for Atmospheric $PM_{2.5}$ Forecasting in Bangladesh*, „Atmosphere (Basel)" 2021, 12, 100, doi:10.3390/atmos12010100.
9. Wang P., Zhang H., Qin Z., Zhang G., *A novel hybrid-Garch model based on ARIMA and SVM for PM 2.5 concentrations forecasting*, „Atmos. Pollut. Res." 2017, 8, 850–860, doi:10.1016/j.apr.2017.01.003.

10. Chen G., Li S., Knibbs L.D., Hamm N.A.S., Cao W., Li T., Guo J., Ren H., Abramson M.J., Guo Y., *A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information*, „Sci. Total Environ." 2018, 636, 52–60, doi:10.1016/j.scitotenv.2018.04.251.
11. Rahimpour A., Amanollahi J., Tzanis C.G., *Air quality data series estimation based on machine learning approaches for urban environments*, „Air Qual. Atmos. Heal." 2021, 14, 191–201, doi:10.1007/s11869-020-00925-4.
12. Kleine Deters J., Zalakeviciute R., Gonzalez M., Rybarczyk Y., *Modeling PM 2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters*, „J. Electr. Comput. Eng." 2017, 2017, 1–14, doi:10.1155/2017/5106045.
13. Demšar J., Curk T., Erjavec A., Gorup Č., Hočevar T., Milutinovič M., Možina M., Polajnar M., Toplak M., Starič A. et al., *Orange: Data mining toolbox in python*, „J. Mach. Learn. Res." 2013, 14(35):2349–2353.
14. GIOŚ Portal Jakość Powietrza GIOŚ, http://powietrze.gios.gov.pl/pjp/home (dostęp: 2.06.2020).
15. *EEA Explore air pollution*, https://www.eea.europa.eu/themes/air/explore-air-pollution-data (dostęp: 12.08.2021).
16. IMGW-PIB Dane publiczne IMGW-PIB, https://dane.imgw.pl/ (dostęp: 12.08.2021).
17. Fix E., Hodges J.L., *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*, Randolph Field, Texas, 1951.
18. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York: New York, NY, 2009; ISBN 978-0-387-84857-0.
19. Tikhonov A.N., *Solution of incorrectly formulated problems and the regularization method*, „Dokl. Akad. Nauk SSSR" 1963, 151.
20. Rumelhart D.E., Hinton G.E., Williams R.J., *Learning representations by back-propagating errors*, „Nature" 1986, 323, 533–536, doi:10.1038/323533a0.
21. Kingma D.P., Ba J.L., *Adam: A Method for Stochastic Optimization*, Int. Conf. Learn. Represent. 2015.
22. Pearson K., *III. Contributions to the mathematical theory of evolution*, „Proc. R. Soc. London" 1894, 54, 329–333, doi:10.1098/rspl.1893.0079.
23. Johnson N.L., Kotz S., Balakrishnan N., *Continuous univariate distributions*, 1994.
24. Haibo He, Garcia E.A., *Learning from Imbalanced Data*, „IEEE Trans. Knowl. Data Eng." 2009, 21, 1263–1284, doi:10.1109/TKDE.2008.239.
25. Majewski G., Czechowski P., Badyda A., Brandyk A., *Effect of air pollution on visibility in urban conditions. Warsaw case study*, „Environ. Prot. Eng." 2014, 40, 47–64, doi:10.5277/epe140204.
26. Majewski G., Rogula-Kozłowska W., Czechowski P., Badyda A., Brandyk A., *The Impact of Selected Parameters on Visibility: First Results from a Long-Term Campaign in Warsaw, Poland*, „Atmosphere (Basel)" 2015, 6, 1154–1174, doi:10.3390/atmos6081154.
27. Majewski G., Szeląg B., Mach T., Rogula-Kozłowska W., Anioł E., Bihałowicz J., Dmochowska A., Bihałowicz J.S., *Predicting the Number of Days With Visibility in a Specific Range in Warsaw (Poland) Based on Meteorological and Air Quality Data*, „Front. Environ. Sci." 2021, 9, doi:10.3389/fenvs.2021.623094.

Creative Commons Attribution 4.0 International License (CC-BY)

28. Seiffert C., Khoshgoftaar T.M., Hulse J. Van, Napolitano A., *RUSBoost: A Hybrid Approach to Alleviating Class Imbalance*, „IEEE Trans. Syst. Man, Cybern. – Part A Syst. Humans" 2010, 40, 185–197, doi:10.1109/TSMCA.2009.2029559.
29. Jaszczyński M., *POŻAR w Szczecinie przy ul. Pomorskiej. Policja: Zamknijcie okna. Doszło do eksplozji [ZDJĘCIA]*, „Głos Szczeciński" 2018.
30. Masternak O., *Pożar w Szczecinie przy ul. Pomorskiej. Nowe informacje, stan powietrza i oświadczenie firmy [ZDJĘCIA]*, „Głos Szczeciński" 2018.
31. Bihałowicz J.S., Rogula-Kozłowska W., Krasuski A., *Contribution of landfill fires to air pollution – An assessment methodology*, „Waste Manag." 2021, 125, 182–191, doi: 10.1016/j.wasman.2021.02.046.
32. Bihałowicz J.S., Rogula-Kozłowska W., Krasuski A., Salamonowicz Z., *The critical factors of landfill fire impact on air quality*, „Environ. Res. Lett." 2021, 16, 104026, doi:10.1088/1748-9326/ac27cd.
33. Bihałowicz J., Rogula-Kozłowska W., Krasuski A., Majder-Łopatka M., Walczak A., Mach T., *Aerosol from waste wood fires: number and volume size distribution. In Proceedings of the 3rd Symposium* [in:] Drzeniecka-Osiadacz A., Korzystka-Muskała M., Sawiński T., Bilińska D., Kubicka J. (ed.), *Air Quality and Health Book of Abstracts*, Wrocław 2021, p. 72–73.
34. KG PSP Interwencje PSP: lata 2010–2019 zestawienia, https://www.kgpsp.gov.pl/panstwowa_straz_pozarna/interwencje_psp (dostęp: 18.01.2020).