UNIWERSYTET GDAŃSKI

WYDZIAŁ ZARZĄDZANIA
Kierunek: Informatyka i Ekonometria
Specjalność: Big Data

Paweł Kińczyk
Nr albumu 276262

Zastosowanie uczenia maszynowego do predykcji celowego wywołania pożaru

Praca magisterska napisana
w Katedrze Statystyki
pod kierunkiem dr hab. Ewy Wycinki prof. uczelni

Sopot 2022

Załącznik nr 1 do zarządzenia Rektora UG nr 70/R/15

OŚWIADCZENIE

Ja, niżej podpisana(y), oświadczam, że przedłożona praca dyplomowa została przygotowana przeze mnie samodzielnie, nie narusza praw autorskich, interesów prawnych i materialnych innych osób.

17.09.2022r
data

podpis

Załącznik nr 3 do zarządzenia Rektora UG nr 70/R/15

OŚWIADCZENIE

Wyrażam zgodę / ~~nie wyrażam zgody*~~ na udostępnienie osobom zainteresowanym mojej pracy dyplomowej dla celów naukowo-badawczych.
Zgoda na udostępnienie pracy dyplomowej nie oznacza wyrażenia zgody na kopiowanie pracy dyplomowej w całości lub w części

* *niepotrzebne skreślić*

17.09.2022r
data

podpis

Spis treści

WSTĘP 5

1. FINANSOWY I SPOŁECZNY ASPEKT POŻARÓW 7

1.1. Specyfika pożarów 7

1.2. Koszty społeczne i ekonomiczne 9

1.3. Przestępczość ubezpieczeniowa 13

2. UCZENIE MASZYNOWE 21

2.1. Algorytmy uczenia maszynowego 21

2.2. Uczenie nadzorowane 23

2.3. Uczenie nienadzorowane 35

2.4. Uczenie przez wzmacnianie 40

2.5. Ewaluacja modeli 41

3. MODELE PREDYKCYJNE CELOWEGO PODPALENIA BUDYNKU 45

3.1. Charakterystyka danych 45

3.2. Budowanie modeli predykcyjnych 51

3.3. Ocena klasyfikatorów 56

PODSUMOWANIE 58

SPIS TABEL 60

SPIS RYSUNKÓW 61

BIBLIOGRAFIA 63

ZAŁĄCZNIKI 67

WSTĘP

Ryzyko pożarów towarzyszy ludziom od początku istnienia naszego gatunku. W miarę rozwoju techniki i usług pojawiły się rozwiązania mające na celu zapobieganie, zwalczanie i zabezpieczanie przed żywiołem. Jednym z nich są ubezpieczenia. Produkty tego typu pozwalają przenieść ryzyko strat poniesionych w wyniku pożaru na firmę ubezpieczeniową. Możliwość ta wykorzystywana jest przez oszustów do wyłudzania świadczeń finansowych. Dla przykładu, w Polsce, w roku 2013 wartość przestępczości ubezpieczeniowej z użyciem ognia wyniosła 10,6 miliona złotych[1]. Finansowo zasadnym jest, aby stale ulepszać metody rozpoznawania przestępstw ubezpieczeniowych. W niniejszej pracy, sprawdzono możliwość zastosowania uczenia maszynowego do predykcji zjawiska celowego wywołania pożaru. Wymagało to połączenia wiedzy z dziedziny ubezpieczeń, ochrony przeciwpożarowej oraz informatyki. Temat pracy został podjęty ze względu na koszty społeczne oraz finansowe zarówno samych pożarów jak i oszustw ubezpieczeniowych. Zwłaszcza w ubezpieczeniach zauważalny jest trend wzrostowy ilości wykrywanych tego typu przestępstw[2]. Dodatkowo wiele z tych przestępstw nie jest wykrywanych.

Postawiono hipotezę, że w tego typu problemie przydatne mogą być algorytmy uczenia maszynowego. Umożliwiają one, na podstawie dostępnych danych, wyszukiwać współzależności jakich człowiek nie jest w stanie zbadać samodzielnie w tak dużej liczbie przypadków. O znaczeniu problemu świadczy to, że firmy ubezpieczeniowe posiadają swoje własne systemy predykcyjne dzięki, którym wykrywają przestępstwa ubezpieczeniowe[3]. Tego typu narzędzia są pomocne nie tylko ubezpieczycielowi, ale również organom ścigania i straży pożarnej.

Biorąc pod uwagę ograniczony dostęp do wiedzy oraz duże możliwości algorytmów, w pracy postawiono dwa cele. Pierwszym było zweryfikowanie czy zasadnym jest stosowanie narzędzi big data w przewidywaniu celowych podpaleń. W przypadku wymiernych korzyści ich wykorzystania drugim celem było utworzenie ogólnodostępnego kodu wraz z bazą danych,

[1] Guzewski P., Wróblewski D., Małozięć D. ,*Czerwona księga pożarów*, Narodowe Centrum Badań i Rozwoju, Józefów 2014, s. 788
[2] Majewski P. ,*Analiza Danych Dotyczących Przestępstw Ubezpieczeniowych w 2020 roku*, Polska Izba Ubezpieczeń, Warszawa 2021, s. 18
[3] Helvestone M. ,*Consumer Protection in the Age of Big Data*, Washington University Law Review, Washington 2016, s. 881-887

który mógłby posłużyć do dalszego zwiększania efektywności wykrywania celowych podpaleń budynków. Kod został napisany tak, że pozwala na ciągłe uczenie algorytmów predykcji wraz z wgrywaniem kolejnych zaktualizowanych danych.

Pracę dla przejrzystego opisu stanu obecnej wiedzy na temat pożarów oraz przedstawienia procesu uczenia maszynowego podzielono na cztery rozdziały. W pierwszym rozdziale opisano zjawisko jakim jest pożar, jego przyczyny i koszty oraz przestępstwa ubezpieczeniowe. Kolejny rozdział traktuje o uczeniu maszynowym, rodzajach klasyfikatorów i sposobach ewaluacji modeli. Następnie, w trzecim rozdziale, opisano użyte dane, proces tworzenia algorytmów predykcyjnych oraz wyniki predykcji. W ostatnim rozdziale podsumowano wyniki pracy.

Do budowy algorytmu wykorzystano język programowania Python z pakietami Scikit-learn, LightGBM oraz Catboost. Algorytmy uczone i walidowane zostały na zagregowanych danych pochodzących z publicznych zasobów udostępnianych przez rząd Wielkiej Brytanii.

1. FINANSOWY I SPOŁECZNY ASPEKT POŻARÓW

Ludzie wykorzystują ogień już od 650-450 tys. lat p.n.e. Przyczynił się on do rozwoju naszego gatunku. Niestety, tak jak na początku jego używania, tak i dziś ogień, gdy przerodzi się w pożar jest nieobliczalny i trudny do opanowania.

Ogień jest sumą zjawisk towarzyszących fizyczno-chemicznemu procesowi spalania mianowicie - emisji promieniowania widzialnego, emisji promieniowania elektromagnetycznego (ciepło) oraz emisji akustycznej. Pożar natomiast jest niekontrolowanym, samoistnym procesem spalania w miejscu, które nie zostało do tego przeznaczone. W obu tych zjawiskach do ich zapoczątkowania potrzebny jest tak zwany trójkąt spalania: materiał palny, utleniacz oraz źródło energii cieplnej.

Dane zbierane w Polsce i w innych krajach wskazują, że nie można bagatelizować tego żywiołu. Pod koniec XX w. na całej kuli ziemskiej odnotowano około 70 mln pożarów, w których życie traciło 70 tys. ludzi.

1.1. Specyfika pożarów

Średnio co godzinę w wyniku pożarów ginie 8 osób, a znacznie więcej doznaje uszczerbku na zdrowiu, który uniemożliwia normalne funkcjonowanie[4]. Szacuje się, że straty spowodowane pożarami rocznie stanowią około 1% światowego PKB. To tak, jakby każdy z nas rocznie przez 3-4 dni pracował tylko i wyłącznie w celu zrekompensowania zniszczeń jakie przyniósł żywioł. W Polsce od lat 80 zauważa się ciągły wzrost liczby zaobserwowanych pożarów. Ze względu na swój charakter, obecną dostępność do źródeł ognia i zmiany klimatu może być to jedno z głównych zagrożeń współczesnej cywilizacji. Dlatego też powinniśmy stale analizować nowe przypadki pożarów i wyciągać z nich wnioski na przyszłość[5]. Aby zacząć opisywać specyfikę powstawania pożarów należy rozpocząć od przyczyn jego występowania. W różnych krajach sposób klasyfikacji powstawania pożarów różni się od siebie, natomiast przyczyny występowania mają swoje odpowiedniki w każdym systemie klasyfikacji. Ze względu na to, że niniejsza praca nie traktuje o prawodawstwie i systemach oceny przyczyn w różnych krajach, za przykład posłużę się Polską normą prawodawczą. Przypuszczalne przyczyny

[4] Brushlinsky N., Sokolov S., Wagner P. ,*Humanity and fires*, Fundacja Edukacja i Technika Ratownictwa, Warszawa 2010, s. 13.
[5] Guzewski P., Wróblewski D., Małozięć D. ,*Czerwona księga pożarów*, Narodowe Centrum Badań i Rozwoju, Józefów 2014, s. 5

pożarów zestawiono w tabeli 1. Przyczyny te przyporządkowane są poszczególnym grupom. Należy zauważyć jak bardzo złożonym zadaniem jest zapobieganie pożarom ze względu na ilość potencjalnych przyczyn.

Tabela 1. Przypuszczalne przyczyny pożarów stosowane podczas interwencji przez polską straż pożarną

Kod	Przyczyna	Grupa statystyczna
1	nieostrożność osób dorosłych (NOD) przy posługiwaniu się ogniem otwartym, w tym papierosami, zapałkami	nieostrożność
2	NOD przy wypalaniu pozostałości roślinnych na polach	
3	NOD przy posługiwaniu się substancjami łatwopalnymi i pirotechnicznymi	
4	NOD przy prowadzeniu prac pożarowo niebezpiecznych	
5	NOD w pozostałych przypadkach	
6	nieostrożność osób nieletnich (NON) przy posługiwaniu się ogniem otwartym, w tym papierosami, zapałkami	
7	NON przy wypalaniu pozostałości roślinnych na polach	
8	NON przy posługiwaniu się substancjami łatwopalnymi i pirotechnicznymi	
9	NON przy prowadzeniu prac pożarowo niebezpiecznych	
10	NON w pozostałych przypadkach	
11	wady urządzeń i instalacji elektrycznych, w szczególności przewodów, osprzętu oświetlenia, odbiorników bez urządzeń grzewczych	urządzenia elektryczne
12	nieprawidłowa eksploatacja urządzeń i instalacji elektrycznych	
13	wady elektrycznych urządzeń ogrzewczych, w szczególności: piecy, grzałek, kuchni	urządzenia ogrzewcze
14	nieprawidłowa eksploatacja elektrycznych urządzeń ogrzewczych	
15	wady urządzeń ogrzewczych na paliwo stałe	
16	nieprawidłowa eksploatacja urządzeń ogrzewczych na paliwo stałe	
17	wady urządzeń ogrzewczych na paliwo ciekłe	
18	nieprawidłowa eksploatacja urządzeń ogrzewczych na paliwo ciekłe	
19	wady urządzeń ogrzewczych na paliwo gazowe	
20	nieprawidłowa eksploatacja urządzeń ogrzewczych na paliwo gazowe	
21	wady urządzeń mechanicznych	inne
22	nieprawidłowa eksploatacja urządzeń mechanicznych	
23	wady procesów technologicznych	procesy technologiczne, magazynowanie
24	nieprzestrzeganie reżimów technologicznych	
25	nieprawidłowe magazynowanie substancji niebezpiecznych	
26	wady środków transportu	środki transportu
27	nieprawidłowa eksploatacja środków transportu	
28	samozapalenia biologiczne	inne
29	samozapalenia chemiczne	
30	wyładowania atmosferyczne	

Kod	Przyczyna	Grupa statystyczna
31	wady konstrukcji budowlanych	
32	nieprawidłowa eksploatacja konstrukcji budowlanych	
33	elektryczność statyczna	
34	podpalenia (umyślne), w tym akty terroru	podpalenia
35	pożary jako następstwo innych miejscowych zagrożeń	
36	inne przyczyny	inne
37	nieustalone	nieustalone

[Źródło: Opracowanie własne na podstawie Guzewski P., Wróblewski D., Małozięć D. ,Czerwona księga pożarów, Narodowe Centrum Badań i Rozwoju, Józefów 2014, s. 193]

Kolejnym aspektem dotyczącym pożarów, gdy już do nich dojdzie, jest grupa pożarowa. W Polskim prawodawstwie wyróżnia się następujące grupy pożarów[6]:

1. A - materiałów stałych, zwykle pochodzenia organicznego, których normalne spalanie zachodzi z tworzeniem żarzących się węgli;
2. B - cieczy i materiałów stałych topiących się;
3. C - gazów;
4. D - metali;
5. F - tłuszczów i olejów w urządzeniach kuchennych

Powyższy podział odnosi się do rodzaju palących się w pożarze materiałów. Wpływa on na sposób zabezpieczenia materiałów, zagrożeń pożarowych oraz metod zwalczania danego pożaru.

1.2. Koszty społeczne i ekonomiczne

Tabela 2. zawiera dane opracowane Wilmota z World Fire Statistics Center. Dla krajów z kompletem danych widać, że całkowity koszt pożaru jest trzykrotnie wyższy niż sama bezpośrednia strata w wyniku pożaru. Czyli sama wartość obiektu jaki został bezpowrotnie spalony jest tylko częścią wszystkich kosztów. Według tego zestawienia na Węgrzech odnotowuje się w ujętych latach największą liczbę zgonów w wyniku pożaru na 100 000 osób. Następne kraje to między innymi Finlandia, USA i Dania. Wielkości budynków, ich projekt oraz materiały konstrukcyjne mogą w znacznej mierze wpływać na duży odsetek ofiar.

[6] *Rozporządzeniem Ministra Spraw Wewnętrznych i Administracji z dnia 21 kwietnia 2006 r. w sprawie ochrony przeciwpożarowej budynków, innych obiektów budowlanych i terenów*, Warszawa 2006,

Tabela 2. Porównanie kosztów pożaru w wybranych krajach w latach 1991-2003

KRAJ	BEZPOŚREDNIE STRATY [%][1]	POŚREDNIE STRATY [%][1]	KOSZT SŁUŻB POŻAROWYCH [%][1]	KOSZTY ADMINISTRACJI BRANŻY UBEZPIECZEŃ [%][1]	KOSZT OCHRONY PRZECIWPOŻAROWEJ BUDYNKÓW [%][1]	ŁĄCZNY KOSZT POŻARU [%][1]	OFIARY POŻARU NA 100 000 OSÓB [%][1]
Austria	(79-80)	0.029 (79-80)	N.A.	0.14 (79-80)	NA.	N.A.	0,74
Belgia	0.40 (88-89)	N.A.	0.18 (87-89)	0.28	0.21 (87-88)	N.A.	1,47
Kanada	0.24	N.A.	0.16 (85)	0.21 (80-81)	0.34	N.A.	1,58
Dania	0.26	0.034	0.09 (87-88)	0.08 (87-88)	0.40 (86-88)	0.864	1,64
Finlandia	0.17 (88-89)	0.021	0.18 (85-86)	0.05	NA.	N.A.	2,18
Francja	0.23	0.037	N.A.	0.16 (79-80)	0.18	N.A.	1,26
Niemcy	0.20	0.037	N.A.	0.09	NA.	N.A.	1,17
Węgry	0.12 (86-88)	0.028	N.A.	0.01 (87—88)	0.42	N.A.	3,31
Japonia	0.08	0.016 (85-86)	0.27	0.	0.27	0.746	1,52
Holandia	0.19	0.03	0.16 (87—88)	0.04 (87—88)	0.32	0.740	0,63
Nowa Zelandia	0.20	N.A.	0.18	0.22	0.12	N.A.	0,92
Norwegia	0.24	0.005	0.12	0.	0.28	0.755	1,60
Hiszpania	0.12 (1984)	NA.	NA.	0.05 (86)	N.A.	N.A.	0,86
Szwecja	0.25	0.009	0.21	0.06	0.12	0.649	1,35
Szwajcaria	0.23 (1989)	0.095	N.A.	N.A.	0.29	N.A.	0,53
Wielka Brytania	0.19	0.019	0.27	0.1	0.14	0.729	1,49
USA	0.15	0.013	0.29	0.06	0.30	0.813	1,95
KOMENTARZ:							
[1] - Średni procent produktu krajowego brutto (1991-2003)							
N.A. = wyliczenia nieznane							
Rok danych jeśli jest spoza okresu 1991-2003 podawany jest obok w nawiasach							

[Źródło: Opracowanie własne na podstawie Ramachandran G. ,The Economics of Fire Protection, Routledge, New York 2003, s. 3.]

Innym miernikiem obrazującym koszty żywiołu jest udział kosztów PKB. Na rysunku 1. przedstawiono wieloletnie bezpośrednie straty spowodowane pożarami jako procentu PKB. Różnice w punktach procentowych dla poszczególnych krajów w dużej mierze odzwierciedlają stan ekonomiczny danego narodu, a nie same koszty pożarów. Prawdopodobieństwo wystąpienia pożaru powinno rosnąć wraz ze wzrostem działalności przemysłowej i handlowej, jednak trend ten nie jest zauważalny w Japonii. W tym kraju PKB rośnie szybciej niż straty wynikające z pożarów. Warto nadmienić, że PKB nie jest najlepszym narzędziem porównawczym ponieważ określa on tylko wartość dóbr i usług jakie wyprodukowano w ciągu jednego roku. Do szacowania prawdziwej procentowej wartości pożarów lepsze są zasoby trwałego kapitału brutto (GFCS). Ujmuje on wartość środków trwałych takich jak budynki, urządzenia i maszyny. Oszacowanie tego typu wskaźnika jest trudne, jednak można się przy jego szacowaniu oprzeć na rocznych informacjach dotyczących nakładów brutto na środki trwałe. Zestawienie tego typu wykonał Ramachandran gdzie porównał bezpośrednie straty w wyniku pożaru w latach 1963-2008 dla wybranych krajów. Graficzne zestawienie ukazane jest na rysunku 1. Można zauważyć globalną tendencję do zmniejszania się skutków, a tym samym kosztów pożarów. Znaczny postęp zgodnie z tym wskaźnikiem dokonała Norwegia.

W celu zmniejszenia strat spowodowanych pożarami, we wszystkich krajach, rządy mają obowiązek zarządzania oraz kontrolowania ryzyka stwarzanego przez pożary. Służy temu prawo oraz instytucje ku temu powołane. Na świecie wiele firm wypracowało plan zarządzania ryzykiem pożarowym. Plany te zawierają się w trzech fazach[7]:

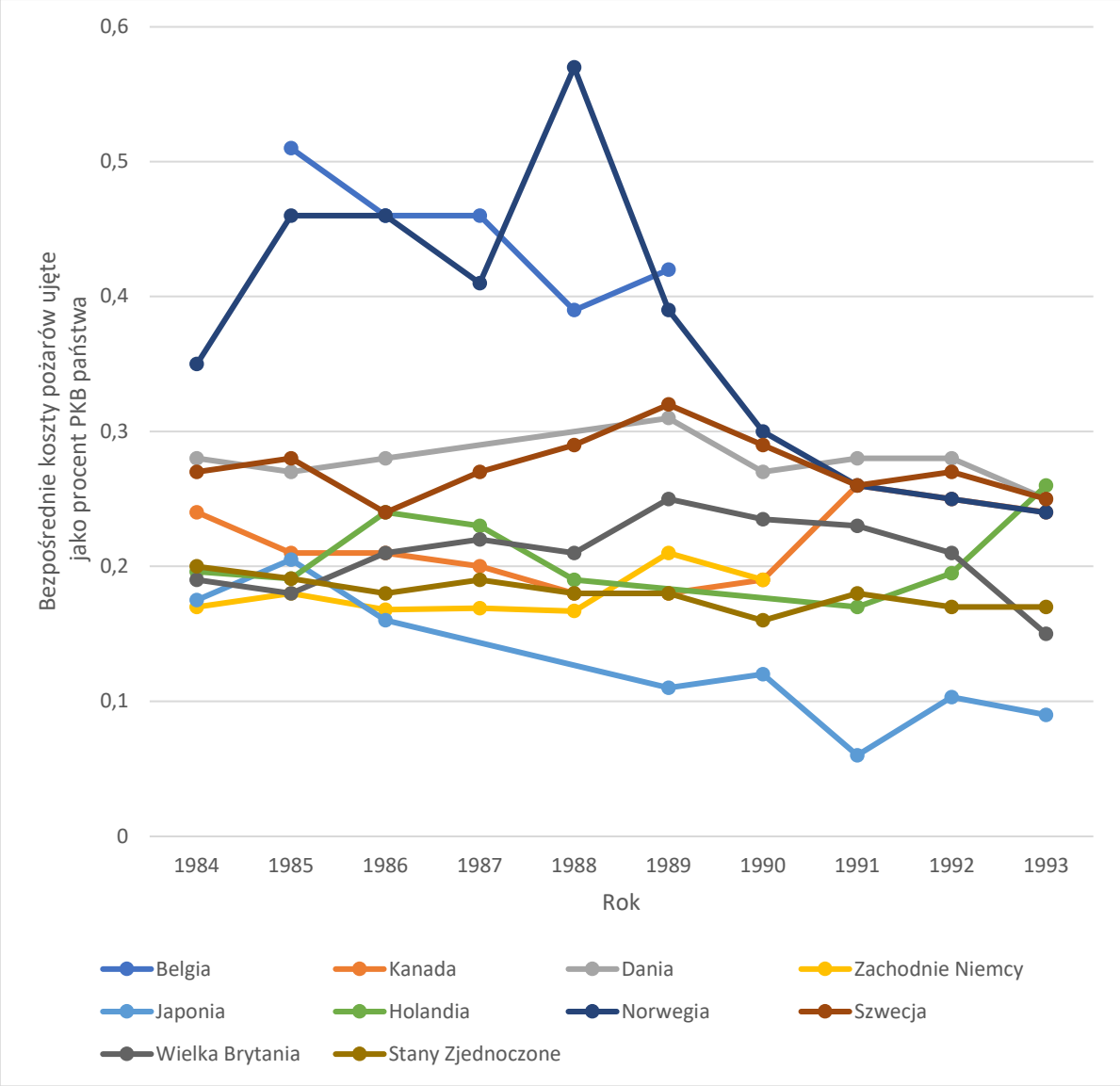
1. Przed pożarem – polega na identyfikacji źródeł zapłonu takie jak np.: nieostrożne usuwanie zapałek, wadliwe okablowanie.
2. W trakcie pożaru – zabezpieczenie obiektu w przypadku wybuchu pożaru np.: zamontowanie tryskaczy, hydrantów.
3. Po pożarze – dotyczy działań po pożarze czyli naprawy uszkodzeń części budynku w celu jego ponownego użytkowania

W pierwszej fazie oceniamy jakościowo i ilościowo elementy, które stwarzają możliwość spowodowania pożaru. Następnie szacujemy wielkość strat wynikających z tych

[7] Guzewski P., Wróblewski D., Małozięć D. ,*Czerwona księga pożarów*, Narodowe Centrum Badań i Rozwoju, Józefów 2014, s. 981

zagrożeń przy użyciu opracowanych modeli probabilistycznych. Jeśli przewidywane straty znacznie przewyższają dopuszczalny poziom sugeruje się usunięcie niektórych źródeł zapłonu. Jeśli natomiast mieścimy się w zakresie normy przechodzimy do drugiej fazy. W niej skupiamy się przede wszystkim na zabezpieczeniu ludzi oraz ograniczaniu strat w wyniku zdarzenia. Do tego działania służą środki ochrony przeciwpożarowej. Trzecia faza obejmuje naprawę zniszczeń, ocenę szkód i roszczenia o odszkodowania.

Rysunek 1. Bezpośrednie straty w wyniku pożarów przedstawiony jako % PKB dla wybranych krajów w latach 1984-1993



[Źródło: Opracowanie własne na podstawie Ramachandran G. ,The Economics of Fire Protection, Routledge, New York 2003, s.4]

W Polsce wykorzystuje się dwa wzory pozwalające wyliczyć ryzyko pożarowe. W obu rodzajach przyjmuje się, że ryzyko pożarowe jest iloczynem prawdopodobnych strat w wyniku

nieszczęśliwego zajścia i częstotliwości powstawania pożarów dla konkretnego rodzaju obiektu[8].

Na całym świecie podejmuje się ograniczanie ryzyka i skutków pożarów. Niestety działania te wymagają znacznych pieniędzy na wszystkich poziomach gospodarki narodowej. W związku z tym pojawia pytanie: „Ile środków kontroli powinniśmy zastosować i jakim kosztem?”. W wyniku ciągłego rozwoju inżynierii sukcesywnie zmniejszamy koszty ochrony przeciwpożarowej, dzięki poprawie technik wnoszenia budynków, sprawniejszym detektorom oraz ciągłej poprawie efektywności instalacji przeciwpożarowych. Ta tendencja pozwala ograniczać występowanie pożarów co widać na danych z rysunku 1. ale należy zwrócić uwagę na poprawie efektywności pod względem ekonomicznym. Zasada, że inwestycja powinna przynosić zysk ma tu również zastosowanie[9].

Analizę ekonomiczną inwestycji opiera się na oszacowaniu kosztów i korzyści związanych z danym projektem. Problemem jaki napotyka się w szacunkach jest to, że wystąpienie pożaru jest zjawiskiem niepewnym, a koszty związane z ochroną przeciwpożarową ponoszone są bez względu na wystąpienie pożogi. W praktyce stosuje się metody, które redukują szacowane korzyści i koszty do poziomu rocznego. Na bazie tych wyliczeń projekt jest opłacalny jeśli wartość roczna korzyści przewyższa koszty. W takich przypadkach często opłacalne staje się ubezpieczanie majątku, dla którego występuje ryzyko pożaru. W wyniku takiej transakcji podmiot zyskuje ograniczenie wpływu straty w wyniku nieszczęśliwego zdarzenia kosztem składki ubezpieczeniowej[10]. Pozwala to zabezpieczyć majątek bez potrzeby ponoszenia ponadnormatywnych wydatków na ochronę przeciwpożarową.

1.3. Przestępczość ubezpieczeniowa

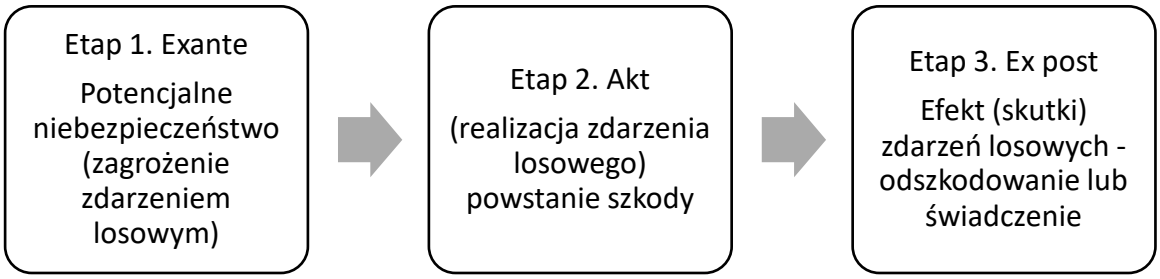
Posiadanie ubezpieczenia wiąże się z pokusą nadużyć, a w skrajnych przypadkach przestępczością ubezpieczeniową. Z danych Polskiej Izby Ubezpieczeń wynika, że tego typu

[8] Pecio M. ,*Ryzyko ubezpieczeniowe a ryzyko pożarowe*, Przestrzeń, Ekonomia, Społeczeństwo, 2015, s. 217
[9] Ramachandran G. ,*The Economics of Fire Protection*, Routledge, New York 2003, s. 7
[10] Soroka-Potrzebna H. ,*Ubezpieczenie jako metoda zarządzania ryzykiem w opinii eksportujących mikroprzedsiębiorstw*, Zeszyty naukowe Uniwersytetu Szczecińskiego nr 854, Szczecin 2015, s.437

zjawiska wykrywane są we wszystkich rodzajach ubezpieczeń[11]. Jednym z rodzajów wyłudzeń odszkodowań są celowe podpalenia.

Przestępczość ubezpieczeniowa należy do kategorii zwanej ryzykiem społecznym, zmiennym w czasie (ryzyko dynamiczne). W praktyce przedstawiane jest ono w trzech etapach. Na rysunku 2 przedstawiono poszczególne stadia. „Ex ante" uwzględnia endogeniczne i egzogeniczne zmienne wpływające za ryzyko wystąpienia zdarzenia. Łącznie z ujęciem ryzyka przestępczości przed zawarciem umowy. W etapie „Ex post" pod uwagę bierze się ryzyko przestępstwa, które może pojawić się podczas trwania umowy.[12]

Rysunek 2.. Ryzyko jako proces stadialny



[Źródło: Opracowanie własne na podstawie Płonka M., Oręziak B., Wielec M. ,Rynek Ubezpieczeniowy - Zapobieganie przyczynom przestępczości, Wydawnictwo Instytutu Wymiaru Sprawiedliwości, Warszawa 2021, s. 12

W każdym z tych etapów, zgodnie z ogólną teorią ryzyka, wyróżnia się dwa czynniki wpływające na prawdopodobieństwo zaistnienia zjawiska jakie obejmuje ubezpieczenie[13]:

1. Zdarzenie losowe, które jest niezależne od woli ubezpieczającego i ubezpieczonego, wpływające na wzrost kosztów po stronie ubezpieczającego.
2. Hazard, który zawiera szereg warunków i okoliczności wpływających na „rozmiar" oraz szansę zdarzenia losowego

Zatem hazard jest powiązany z potencjalną możliwością wystąpienia przestępczości ubezpieczeniowej. Wynika to z tego, że dążenie do zawarcia umowy ubezpieczenia oraz jej zawarcie wpływają na zachowanie ubezpieczonego. Świadomość posiadania ochrony

[11] Majewski P. ,*Analiza Danych Dotyczących Przestępstw Ubezpieczeniowych w 2020 roku*, Polska Izba Ubezpieczeń, Warszawa 2021, s. 12-17
[12] Płonka M., Oręziak B., Wielec M. ,*Rynek Ubezpieczeniowy - Zapobieganie przyczynom przestępczości*, Wydawnictwo Instytutu Wymiaru Sprawiedliwości, Warszawa 2021,.s. 22-25
[13] Michalski T., Pajewska R. ,*Ubezpieczenia gospodarcze*, Wydawnictwa Szkolne i Pedagogiczne S.A., Warszawa 2001,.s. 29-40

ubezpieczeniowej kreuje pokusę nadużyć, a w skrajnych przypadkach prowadzi do przestępczości ubezpieczeniowej. Wyróżnia się trzy rodzaje hazard, które zawarto w tabeli 3.

Tabela 3. Rodzaje hazardu w ubezpieczeniach.

Typ hazardu	Przykłady zachowania
Hazard fizyczny	Zewnętrzne uwarunkowania lub cechy fizyczne, które bezpośrednio wpływają na większe prawdopodobieństwo wystąpienia strat. Na przykład: liczba czujników dymu, stan urządzeń. Zachowania: korzystanie z uszkodzonego urządzenia, brak zachowania należytej ostrożności.
Hazard duchowy	Subiektywne zachowanie ubezpieczonego wynikające ze świadomości istnienia ochrony ubezpieczeniowej. Na przykład: niedbałość, zrezygnowanie z czynności zapobiegającym powstawanie ryzyka (stosowanie się do zaleceń itd.) Zachowania: brak stosowania zabezpieczenia dóbr/mienia ubezpieczonego, zaniechanie włączania systemów ochrony (alarm, czujniki dymu itd.).
Hazard moralny	Grupa predyspozycji danej osoby, objawiających się negatywnymi tendencjami. Na przykład: nieuczciwość, skłonność do wyłudzeń/defraudacji. Zachowania: celowe podpalenia w celu wyłudzenia ubezpieczenia, prowokowanie szkody.

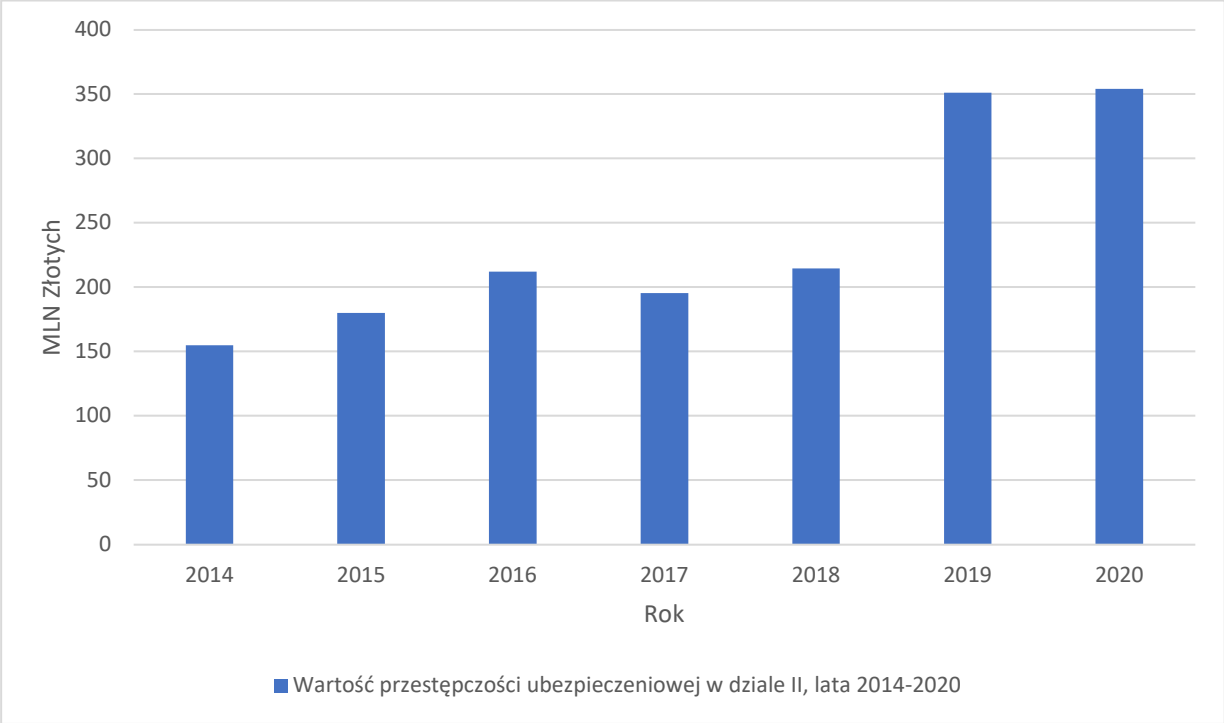
Źródło: Opracowanie własne na podstawie Płonka M., Oręziak B., Wielec M. ,Rynek Ubezpieczeniowy - Zapobieganie przyczynom przestępczości, Wydawnictwo Instytutu Wymiaru Sprawiedliwości, Warszawa 2021, s. 22-25.

Wszystkie rodzaje hazardu są negatywne dla ubezpieczyciela ponieważ prowadzą do zwiększenia kosztów z tytułu wypłaty odszkodowań. Nie wszystkie przypadki są jednak przestępstwami ubezpieczeniowymi. W głównej mierze do przestępstw zaliczane są zjawiska hazardu moralnego, gdzie podmiot ubezpieczony celowo podejmuje działania mające na celu nadużycie.

W raporcie Polskiej Izby Ubezpieczeniowej dotyczącego przestępstw ubezpieczeniowych na rok 2020 przedstawiono dane dotyczące skali wyłudzeń. W dziale I

obejmującym ubezpieczenia na życie wartość wykrytych wyłudzeń stanowią 0,269% łącznej sumy wypłaconych świadczeń. Dla działu II (pozostałe ubezpieczenia osobowe i majątkowe) wartość udaremnionych przestępstw wynosi 1,59% łącznej kwoty wypłat z tytułu ubezpieczenia. Z roku na rok wartość przestępczości ubezpieczeniowej w dziale II rośnie co przedstawione jest na rysunku 3.

Rysunek 3. Wartość przestępczości ubezpieczeniowej w dziale II w Polsce w latach 2014-2020



*[Źródło: Opracowanie własne na podstawie Majewski P. ,*Analiza Danych Dotyczących Przestępstw Ubezpieczeniowych w 2020 roku*, Polska Izba Ubezpieczeń, Warszawa 2021, s. 15]*

Podane wartości mogą być zaniżone ze względu na występowanie tak zwanej „ciemnej liczby" nieujawnionych spraw. Szacuje się, że w przypadku działu I, aż 87% spraw nie jest zgłaszana organom ścigania.[14] Czynniki wpływające na skalę tych zjawisk to między innymi[15]:

- Wpływ otoczenia na sprawcę – na przykład społeczne przyzwolenie na oszustwo, albo presja ze strony środowiska[16].

[14] Majewski P. ,*Analiza Danych Dotyczących Przestępstw Ubezpieczeniowych w 2020 roku*, Polska Izba Ubezpieczeń, Warszawa 2021, s. 12-17
[15] Buczkowski K. ,*Oszustwo asekuracyjne jako jeden z przejawów przestępczości ubezpieczeniowej*, Wolters Kluwer Polska SA, 2015, s. 132-134
[16] Połeć R., Lemańczyk M. ,*Przeciwdziałanie przestępczości ubezpieczeniowej w aspekcie jej społecznej akceptacji*, Studia oeconomica posnaniensia, 2015, s. 113-114

- Przepływ informacji – zwiększający się dostęp do wiedzy i informacji ułatwia wyszukiwanie metod wyłudzania ubezpieczeń, ale również umożliwia na łatwiejsze działanie grup zorganizowanych.
- Rozwój technologiczny – anonimowe kontakty marketingowe z ubezpieczalnią ułatwiają oszustwa, a dostęp do nowych technologii pozwala skutecznie unikać wykrycia przestępstwa.
- Prawo – luki prawne umożliwiają omijanie odpowiedzialności. Natomiast instytucje ubezpieczeniowe, które w swoim interesie mają wykrywanie przestępstw nie posiadają wystarczającego statusu prawnego takiego jak policja, który umożliwiałby skuteczne zwalczanie oszustw.
- Niska świadomość ubezpieczeniowa – ubezpieczony nie rozumie funkcji ochronnej ubezpieczenia, przez co brak wypłaconego świadczenia odczytuje jako swoją stratę. Takie rozumowanie prowadzi do pokusy nadużycia.

Działania mające ograniczyć zjawisko przestępstw ubezpieczeniowych odnoszą się do etapu „Ex ante" i mają na celu zapobieganie oraz do etapu „ Ex post" gdzie wykrywa się i aktywnie zwalcza nadużycia. Prewencyjne metody zarządzania ryzykiem przestępstw ubezpieczeniowych to[17]:

W charakterze systemowym

1. Regulacje prawne firm ubezpieczeniowych – licencjonowanie działalności i kontrola firm ubezpieczeniowych[18].
2. Etyka zawodowa – wypracowanie dobrych praktyk w firmach ubezpieczeniowych.
3. Współpraca organizacji w celu wykrywania i prewencji – zacieśnianie współpracy organizacji poprzez tworzenie wspólnych baz danych, wymianie informacji oraz wzajemnych ostrzeganiu[19].
4. Edukacja – edukacja ludzi w celu zwiększenia ich świadomości związanej z ubezpieczeniami oraz etyką.

[17] Płonka M., Oręziak B., Wielec M. ,*Rynek Ubezpieczeniowy - Zapobieganie przyczynom przestępczości*, Wydawnictwo Instytutu Wymiaru Sprawiedliwości, Warszawa 2021, s. 37-40
[18] Pecio M. ,*Nadzór nad rynkiem ubezpieczeniowym na przykładzie Polski i Stanów Zjednoczonych – model działania*, Wroclaw Economic Review, Wrocław 2009, s. 48-51
[19] Filipowska M. ,*Rynek ubezpieczeniowy wciąż zagrożony przestępczością*," Wroclaw Economic Review, Wroclaw Economic Review, Wrocław 2008, s. 152-153

5. Ochrona konsumencka – niweluje asymetrię informacji i środków między firmą ubezpieczeniową, a konsumentem.
6. Zagrożenie infamią – prowadzenie centralnego rejestru nieuczciwych kontrahentów oraz klientów.

   W charakterze partykularnym (w firmie ubezpieczeniowej)
7. Unikanie ryzyka – poprzez unikanie określonych grup klientów, którzy przejawiają większą szansę na nieuczciwe zachowania[20].
8. Zarządzanie ryzykiem i śledzenie ryzyka – wewnątrzzakładowe działania mające na celu minimalizację ryzyka przestępczości ubezpieczeniowej. Na przykład poprzez prowadzenie własnych analiz, które potem wykorzystywane są przy odmowach ubezpieczeń.
9. Współpraca z organami ścigania – współpraca z organami ścigania oraz prywatnymi firmami detektywistycznymi w celu identyfikacji wyłudzeń.
10. Opracowanie produktów ubezpieczeniowych – konstruowanie procesu wydawania ubezpieczenia oraz treści umowy, które skutecznie ograniczają możliwości oszustwa. Na przykład klauzule umowne i charakteryzowanie klienta na podstawie wywiadu i formularzy.

W odniesieniu do etapu „Ex post" kontroluje się skutki przestępstw. Działania tego typu opierają się na tworzeniu wewnętrznych procedur umożliwiających wykrywanie przestępstw. Do działań tego typu zalicza się[21]:

1. Zatrzymywanie ryzyka – poprzez utworzenie rezerwy składki.
2. Monitoring ryzyka – działania mające na celu ograniczenie strat finansowych. Wyróżnia się między innymi:
   a. Aktualizację ryzyka – obowiązek informowania ubezpieczającego o zmianach okolicznościowych[22].
   b. Obowiązek informowania o szkodzie – w celu ograniczania skali szkody i umożliwienia stałego monitoringu wypłacanych świadczeń.

[20] Pecio M. ,*Przeciwdziałanie przestępczości*, Teoria i Praktyka 2016 nr 4, 2016, s.126
[21] Strupczewski G. ,*Ryzyko i zarządzanie ryzykiem*, Kraków 2013, s. 36–37
[22] Cycoń M. ,*Umowa ubezpieczenia, Współczesne ubezpieczenia gospodarcze*, red. W. Sułkowska, Uniwersytet Ekonomiczny w Krakowie, Kraków, 2013, s. 137

c. Obniżanie wysokości odszkodowania – w przypadku ewidentnego naruszenia zapisów umowy przez ubezpieczonego.
d. Zgłaszanie czynów przestępczych

3. Penalizacja czynów przestępczych – określenie norm prawnych związanych z przestępstwami ubezpieczeniowymi oraz ich egzekwowania poprzez współpracę z organami ścigania.

Rynek ubezpieczeń jest jednym z rynków zaufania publicznego, dlatego przestępstwa ubezpieczeniowe są istotnym problem, który należy zwalczać. Ze względu na różne wymiary przestępstw i nadużyć należy przy wyborze narzędzi zapobiegawczych każdorazowo stosować porównania kosztów danego rozwiązania względem zysków jakie generuje. Barierą w niektórych przypadkach może być legalność stosowania pewnych rozwiązań na przykład przepisy RODO i prawa obywatelskie[23]. W takim przypadku zastosowanie metod uczenia maszynowego oraz analizy danych daje duży potencjał poprzez możliwość wyszukiwania współzależności, przy jednoczesnej anonimizacji oraz braku faworyzacji, ze względu na cechy danych osób.

Firmy ubezpieczeniowe korzystają z dużych zbiorów danych, ze względu na niższe koszty ich przechowywania oraz analizy. Dostęp do takiej liczby danych pozwala, przy wykorzystaniu analizy danych, na optymalizacji dochodów. Może być to osiągnięte dzięki:

- Wyszukiwaniu wzorców i korelacji – między klientami i ich potrzebami, albo w przypadku profilowania potencjalnych oszustów.
- Opracowywania nowych produktów – przy wykorzystaniu grupowaniu klientów w docelowe grupy i tworzenie spersonalizowanych usług.
- Przewidywanie – określanie potencjalnych zdarzeń, na przykład przestępstw ubezpieczeniowych w oparciu o historyczne dane.
- Monitorowanie zmian – dzięki stale aktualizowanej bazie danych możliwe jest stałe monitorowanie zachowań klientów i rynku

Na dzień dzisiejszy wiele firm ubezpieczeniowych wykorzystuje aktywnie analizę dużych danych w celu personalizacji reklam i modelowania ryzyka ubezpieczeniowego. Wykorzystanie

[23] Korff D., Georges M. ,*Podręcznik Inspektora Ochrony Danych*, Garante per la protezione dei dati personal, 2019, s. 44-46

tego typu narzędzi daje możliwości tworzenia lepszych produktów, jednak wiąże ze sobą pewne zagrożenia[24]. Agregacja dużych zbiorów danych stwarza ryzyko ich wycieku co skutkuje większą ilością poszkodowanych osób. Innym potencjalnym zjawiskiem jest rosnąca segmentacja klientów, która może doprowadzić do ograniczenia dostępu do usług lub wykluczenia niektórych klientów z rynku.

Reasumując, pożary są zagrożeniem dla życia i mienia człowieka. W wielu krajach straty wynikające z ich występowania są kosztem sięgającym w niektórych latach ponad 0,6% PKB. Z drugiej strony ciągłe zwiększanie ochrony przeciwpożarowej jest opłacalne jedynie do pewnego momentu, w którym to koszty ochrony przewyższają możliwe zyski z tytułu zabezpieczenia przed ryzykiem wystąpienia pożaru. W takich przypadkach firmy oraz klienci indywidualni wybierają możliwość zabezpieczenia się przed ryzykiem pożaru poprzez ubezpieczenia. Dla firm ubezpieczeniowych głównym dochodem są składki klientów, ale aby przynosić dochody, niezbędne jest odpowiednie oszacowywanie ryzyka wystąpienia nieszczęśliwego wypadku oraz wykrywanie wyłudzeń. Ze względu na globalizację, zmiany technologiczne oraz rosnącą liczbę klientów klasyczne metody szacowania ryzyka mogą nie wystarczyć do utrzymania konkurencyjności firmy. W tego typu przypadkach rozwiązaniem może być wykorzystanie algorytmów sztucznej inteligencji oraz uczenia maszynowego w celu analizowania dużych zbiorów danych jakie zbierane są przez przedsiębiorstwa świadczące takie usługi.

[24] Łańcucki J. ,*Klient na cyfrowym rynku ubezpieczeniowym*, Prawo asekuracyjne, Poznań,2018, s. 6

2. UCZENIE MASZYNOWE

Uczenie maszynowe jest współcześnie jednym praktycznych zastosowań sztucznej inteligencji. Polega na uczeniu maszyny rozpoznawania i przewidywania określonych zjawisk. W swojej „nauce" komputer wykorzystuje zbiór danych przygotowany przez człowieka.[25] Następnie w oparciu o wzory matematyczne „trenuje" rozpoznawanie poszukiwanych elementów. Wykorzystuje do tego prawidłowości występujące w otrzymanych danych. W wyniku czego osoba wykonująca uczenie maszynowe otrzymuje model, który potrafi na nowym zbiorze informacji oszacować poszukiwane wyniki.[26]

2.1. Algorytmy uczenia maszynowego

W uczeniu maszynowym głównym problemem jest wybór algorytmu. Ich podział można wykonać w oparciu o typ uczenia jaki wykorzystuje:

Uczenie nadzorowane – bazuje na sklasyfikowanych danych to znaczy takich, które opisane są za pomocą etykiet (data urodzenia, płeć, wiek). Na ich podstawie algorytm wyszukuje zależności w zbiorze w celu oszacowania szukanej wartości.[27] Zgodnie z metodyką uczymy się oszacowania wejścia x do wyjścia y poprzez zbiór danych uczących.

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

Gdzie:

N - liczba przykładów uczących

x – dane wejścia

y – dane wyjścia

Wejście x_i jest wektorem wartości odzwierciedlających próbkę. Liczby te nazywamy atrybutami lub cechami. W praktyce mogą to być pojedyncze wartości, ale również obrazy, tekst itp. Wartość y w tym przypadku może być wartością dyskretną, przyjmującą wartości pewnego skończonego zbioru, na przykład płeć, lub przyjmować wartości liczbowe np. dochód.

[25] *Czym jest uczenie maszynowe?*, https://www.sap.com/poland/insights/what-is-machine-learning.html. [Data uzyskania dostępu: 5 luty 2022]
[26] Cichosz P., *Systemy uczące się*, Warszawa, WNT, 2000,
[27] Grzyb M. ,*Uczenie maszynowe - Sekrety doboru odpowiedniego algorytmu*, *s. 4*

Uczenie nienadzorowane – ma zastosowanie gdy dane wejściowe nie są sklasyfikowane za pomocą etykiet, a wynik uczenia nie jest jednoznacznie określony. W takim przypadku głównym celem może być organizacja danych lub opisanie struktur danych wejściowych.

$$D = \{x_i\}_{i=1}^{N}$$

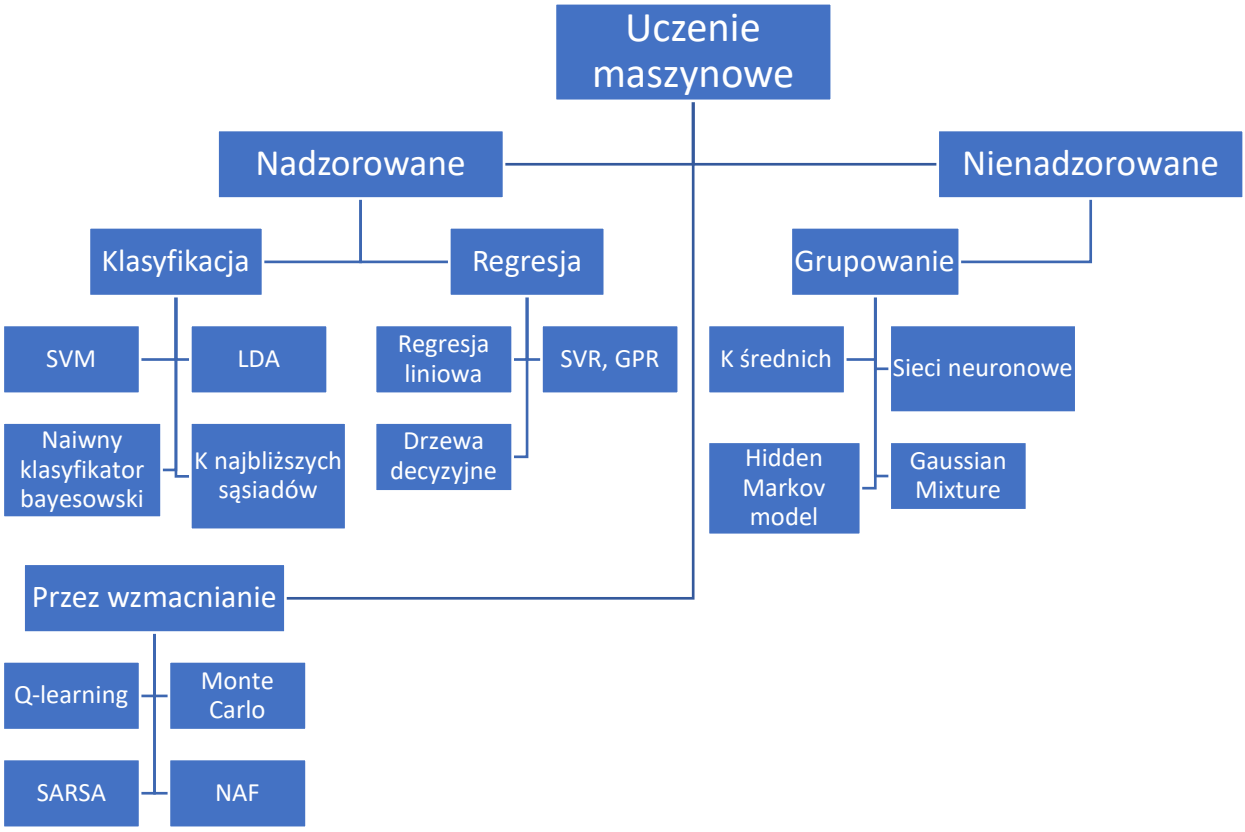
Gdzie:

N - liczba przykładów uczących

x – dane wejścia

Uczenie przez wzmacnianie – używane jest kiedy maszyna ma działać w środowisku nieznanym (nieopisanym). Jedyną informacją jaką otrzymuje ucząca się jednostka obliczeniowa jest impuls oznaczający „nagrodę" lub „karę".

Rysunek 4. Podział sposobów uczenia maszynowego



[Źródło: Opracowanie własne na podstawie Podział sposobów uczenia maszynowego, https://www.pngitem.com/middle/hRJJRJJ_types-of-machine-learning-machine-learning-types-of/ [Data uzyskania dostępu: 15 czerwiec 2022]]

Ze względu na typ problemu:

Algorytmy grupujące – zadaniem algorytmu jest przyporządkowanie danych zmiennych wejściowych w klastry[28].

Algorytmy klasyfikacyjne – klasyfikują wartość szukaną. W przypadku gdy algorytm przewiduje wartość z zakresu np. prawda/fałsz mówi się wówczas o klasyfikacji dwuklasowej. Jeśli natomiast zmienna szukana przyjmuje więcej niż dwie wartości definiujemy ją jako klasyfikację wieloklasową.

Algorytmy regresyjne – algorytmy, które mają oszacować wartość liczbową przewidywanej cechy

Inne – algorytmy, które nie zaliczają się do powyższych grup. Są to między innymi algorytmy rekomendacyjne lub algorytmy do rozpoznawania zdjęć

2.2. Uczenie nadzorowane

Algorytmy używane w tego typu uczeniu maszynowym tworzą funkcję, która korzysta z danych wejściowych w celu oszacowania wyników w przyszłych danych[29]. Zgodnie z rysunkiem 4 należy zauważyć, że uczenie nadzorowane dzieli się na dwie podgrupy. Klasyfikacja odpowiada za przewidzenie wartości dyskretnej. Regresja natomiast odpowiada za prognozowanie wartości liczbowych. Poniżej przedstawiono klasyfikatory dla omawianej grupy.

Naiwny klasyfikator bayesowski[30]

Naiwny klasyfikator Bayesa opiera swoje działanie na twierdzeniu Bayesa. Jest to prosty algorytm klasyfikacyjny, który nadaje się do zbiorów o wielu wymiarach na wejściu. Zakłada zupełną niezależność każdej ze zmiennych. Twierdzenie Bayesa[31]:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

[28] Rządca K. ,*Algorytmy grupowania danych*, 2004, https://www.mimuw.edu.pl/~krzadca/magister.pdf. [Data uzyskania dostępu: 1 czerwiec 2022]
[29] IBM ,*Supervised Learning*, https://www.ibm.com/cloud/learn/supervised-learning. [Data uzyskania dostępu: 2 czerwiec 2022]
[30] ang. Naive Bayes
[31] Berrar D. ,*Theorem and Naive Bayes Classifier*, Tokyo Institute of Technology, Tokio 2018, s. 7-8

Gdzie:

$P(B|A)$ – prawdopodobieństwo a posteriori klasy (B) danego predyktora (A)

$P(A|B)$ – prawdopodobieństwo predyktora danej klasy

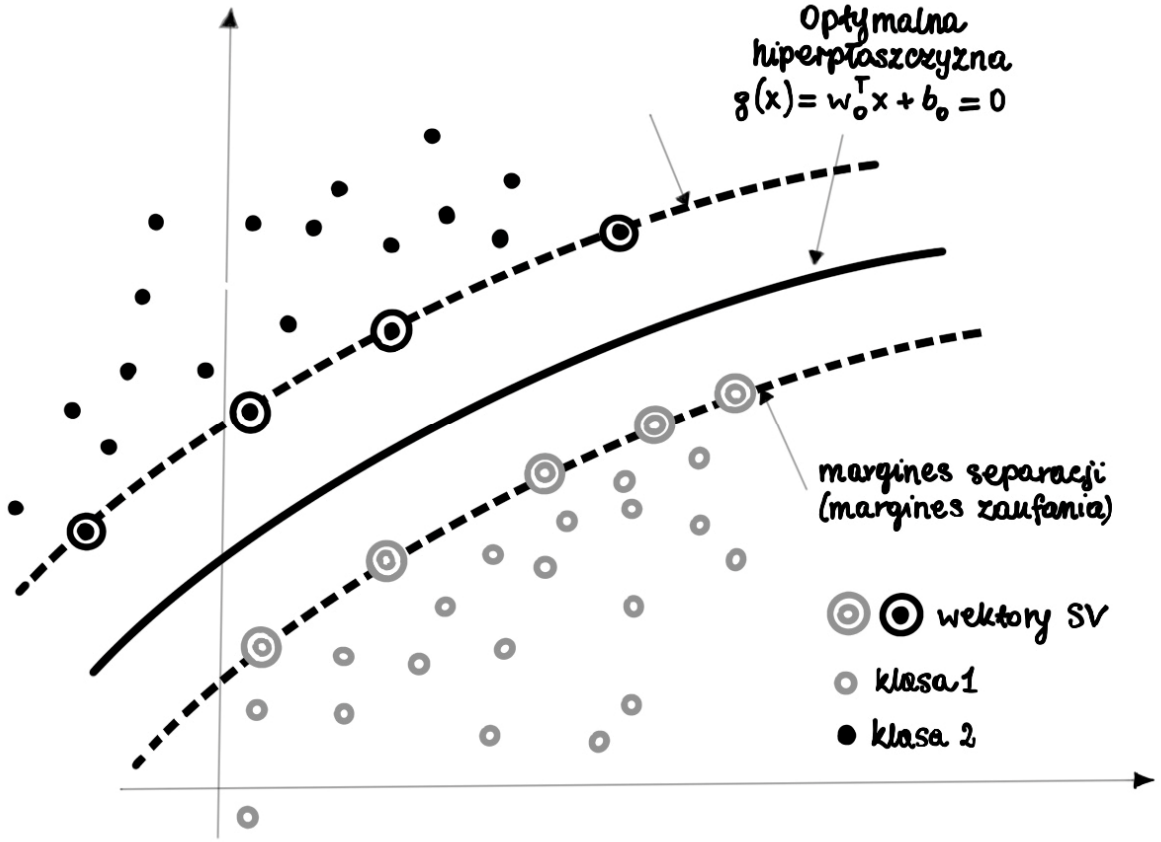
$P(B)$ – prawdopodobieństwo a priori klasy

$P(A)$ – prawdopodobieństwo a priori predyktora

Maszyna wektorów nośnych[32]

Popularny algorytm opracowany w latach dziewięćdziesiątych. Nadal używany jest w przypadku potrzeby wysokowydajnego algorytmu z niewielkim dostrojeniem. Metoda ta pokazywała nowe podejście ze względu na budowę klasyfikatora. Znajduje zastosowanie w problemach klasyfikacyjnych i regresji. Celem tej metody jest poszukiwanie hiperpłaszczyzny, która rozdzieli dwie klasy.

Rysunek 5. Optymalna hiperpłaszczyzna o maksymalnym marginesie separacji

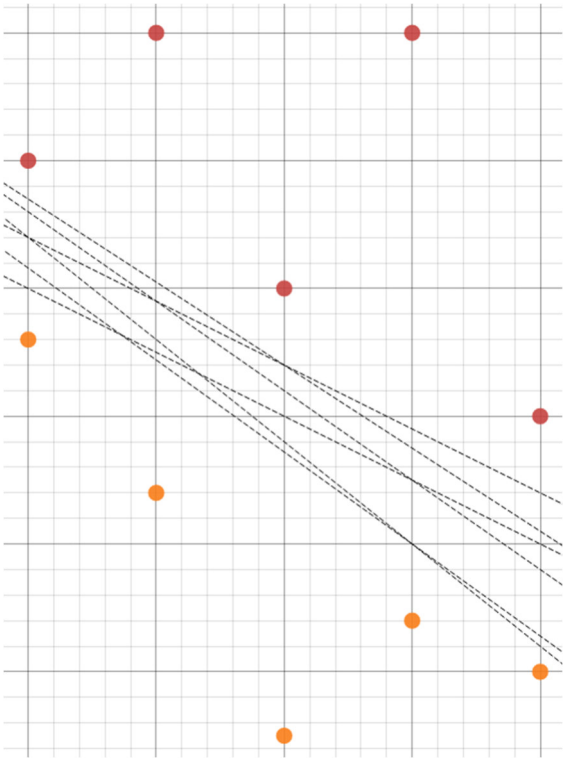


[Źródło: Opracowanie własne na podstawie Mrówczyńska M. ,Klasyfikatory neuronowe typu SVM w zastosowaniu do klasyfikacji przemieszczeń pionowych na obszarze LGOM, Zeszyty Naukowe Instytutu Gospodarki Surowcami Mineralnymi i Energi Polskiej Akademii Nauk, 2014, s. 71.]

[32] ang. Support Vector Machine (SVM)

Hiperpłaszczyzna musi rozdzielać klasy w taki sposób, żeby maksymalizować odległości między punktami z przeciwnych klas[33]. Hiperpłaszczyznę przedstawiono na rysunku 3.2. Znajdowanie takiej granicy to problem optymalizacyjny. Należy dobrać z prób najodpowiedniejszą linię. Przykład takiego procesu znajduje się na rysunku 5.

Rysunek 6. Dostępne hiperpłaszczyzny



[Źródło: Opracowanie własne na podstawie Stefanowski J. ,SVM - Support Vector Machines Metoda wektorów nośnych, Institute of Computing Sciences, Poznań University of Technology, http://www.cs.put.poznan.pl/jstefanowski/ml/SVM.pdf [Data uzyskania dostępu: 10 luty 2022], s. 9]

Z dostępnych linii wybierana jest taka, która pozwala uzyskać najlepsze efekty. Kryteria doboru takiej linii zależą od miar na jakie optymalizowany jest algorytm.

K najbliższych sąsiadów[34]

Metoda należąca do najprostszych ze wszystkich algorytmów uczenia maszynowego. Przechowuje wszystkie wzorce uczenia w odniesieniu, do których określa odległość wzorca testowego. Metodologia ta w pierwszej kolejności polega na przypisaniu zestawu n wartości

[33] Bottou L., Lin Ch. ,*Support Vector Machine*, , 2007, https://leon.bottou.org/publications/pdf/lin-2006.pdf [Data uzyskania dostępu: 5 czerwiec 2022], s. 3-4
[34] ang. k nearest neighbours

opisujących każdą z danych. Następnie dla szukanego nowego punktu danych, przeszukuje się cały zbiór zestawu uczącego w celu znalezienia najbardziej podobnych instancji[35]. Sprawdzenie jakie elementy znajdują się najbliżej szukanej wartości wykonuje się na przykład przy użyciu odległości euklidesowej[36]:

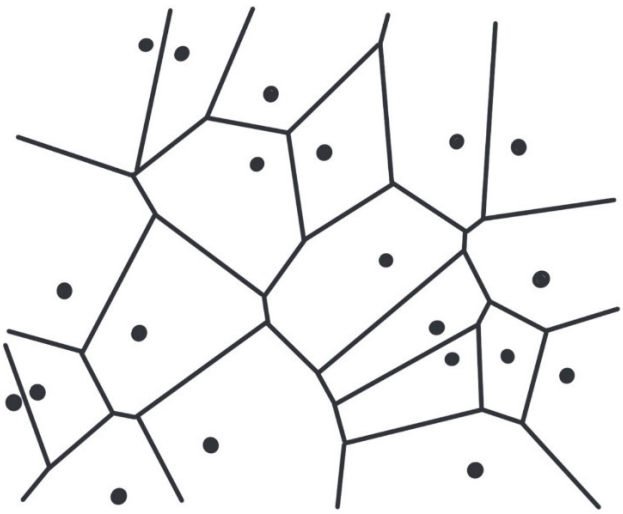
$$Odległość\ Euklidesowa\ (a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Gdzie:

a, b – punkty względem, których sprawdzana jest odległość

i – wszystkie atrybuty wejściowe

Rysunek 7. Schemat granic decyzyjnych



[Źródło: Opracowanie włąsne na podstawie Shwarts S., Ben D. ,Understanding Machine Learning, Cambridge University Press, New York 2014, s. 220.]

Na rysunku 7 przedstawiony jest schemat granic decyzyjnych wynikających z ich odległości euklidesowej. Przewidywana wartość zostanie przyporządkowana do punktu, w których granicy decyzyjnej się znajdzie.

Regresja liniowa

Regresja liniowa jest często wybieranym modelem, ponieważ jego reprezentacją jest niezłożone równanie. Wzór łączy w sobie określony zbiór wartości wejściowych (x), którego

[35] Brownlee J. ,*Master Machine Learning Algorithms*, Jason Brownlee, 2016, s. 102
[36] Shwarts S., Ben D. ,*Understanding Machine Learning*, Cambridge University Press, New York 2014,

rozwiązaniem jest przewidywany wynik dla tego zestawu wartości wejściowych (y). Zarówno wartości wejściowe (x), jak i wartości wyjściowe są numeryczne. Równanie liniowe przypisuje jeden współczynnik skali do każdej wartości wejściowej lub kolumny, zwany współczynnikiem reprezentowanym przez grecką literę Beta (β). Dodatkowy współczynnik ($B0$) to wartość, nadająca linii dodatkowy stopień swobody (np. poruszanie się w górę i w dół na wykresie dwuwymiarowy) i jest często nazywany przecięciem lub współczynnikiem odchylenia. Przykład uzyskanej linii regresji na podstawie danych przedstawiono na rysunku 8. Metoda opiera się na wzorze[37]:

$$y = B0 \quad + B1 \quad \times x$$

Gdzie:

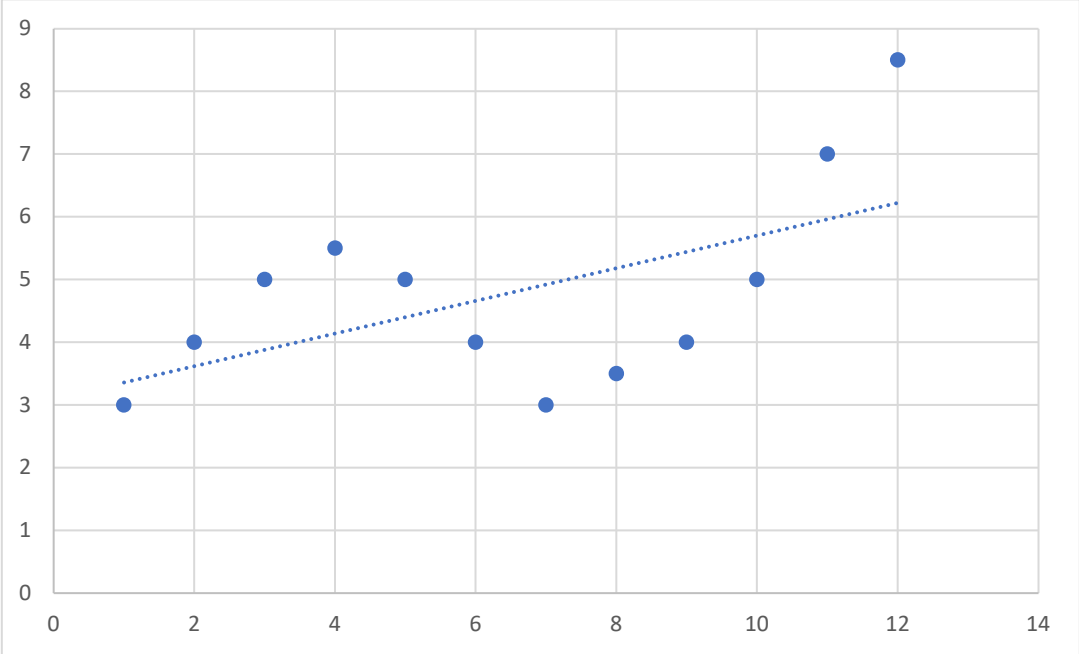
y – przewidywany wynik

$B0$ – dodatkowy współczynnik

$B1$ - współczynnik

x – wartość wejściowa

Rysunek 8. Linia regresji



[Źródło: Opracowanie własne przy wykorzystaniu programu excel]

[37] Brownlee J. ,*Master Machine Learning Algorithms*, Jason Brownlee, 2016, s. 34-36

Kiedy problem dotyczy wyższych wymiarów i mamy więcej niż jedno wejście x linia nazywana jest płaszczyzną lub hiperpłaszczyzną[38]. Przykład przedstawia rysunek 9. Postać wzoru dla wielokrotnej regresji:

$$y = B0 \quad + B1 \quad \times x^1 + \cdots + BN \quad \times x^n$$

Gdzie:

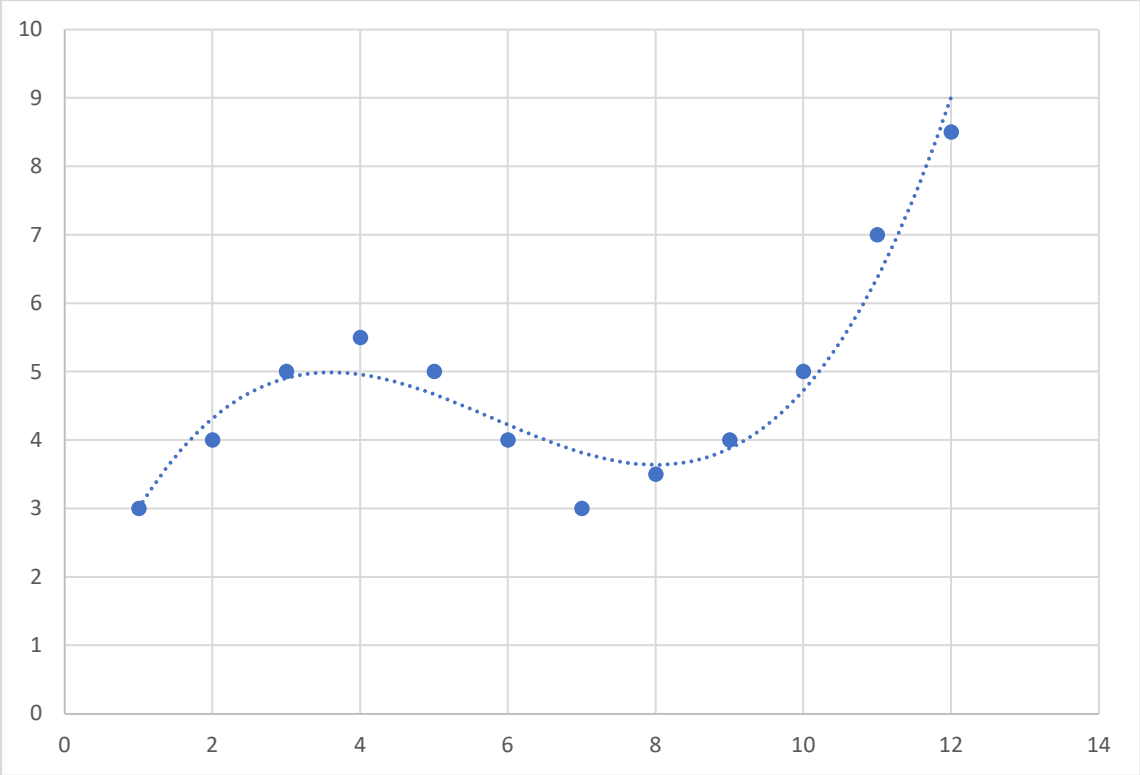
y – przewidywany wynik

$B0$ – dodatkowy współczynnik

BN – współczynnik dla kolejnych wartości wejściowych

x^n – wartości wejściowe

Rysunek 9. Linia regresji wielokrotnej



[Źródło: Opracowanie własne przy wykorzystaniu programu excel]

Linia regresji dopasowała się do otrzymanych danych. Dzięki jej wyznaczeniu możliwe jest oszacowanie wartości dla innych danych, które nie występują w badanym zbiorze.

[38] Gülden K., Neşe G. ,*A study on multiple linear regression analysis*, Elsevier Ltd, Sakarya 2013, s. 235

Regresja wielomianowa

Regresja wielomianowa jest rozwinięciem regresji liniowej. W modelu tym zależność między zmienną wejściową x, a wyjściową y modelowana jest jako wielomian n-tego stopnia. Na rysunku 10 przedstawiono przykład regresji wielomianowej. Ogólny wzór wykorzystywany w tym przypadku:

$$y = B0 \quad + B1 \quad \times x^1 + B2 \quad \times x^2 + \ B3 \quad \times x^3 + \cdots + BN \quad \times x^n$$

Gdzie:

y – poszukiwana funkcja

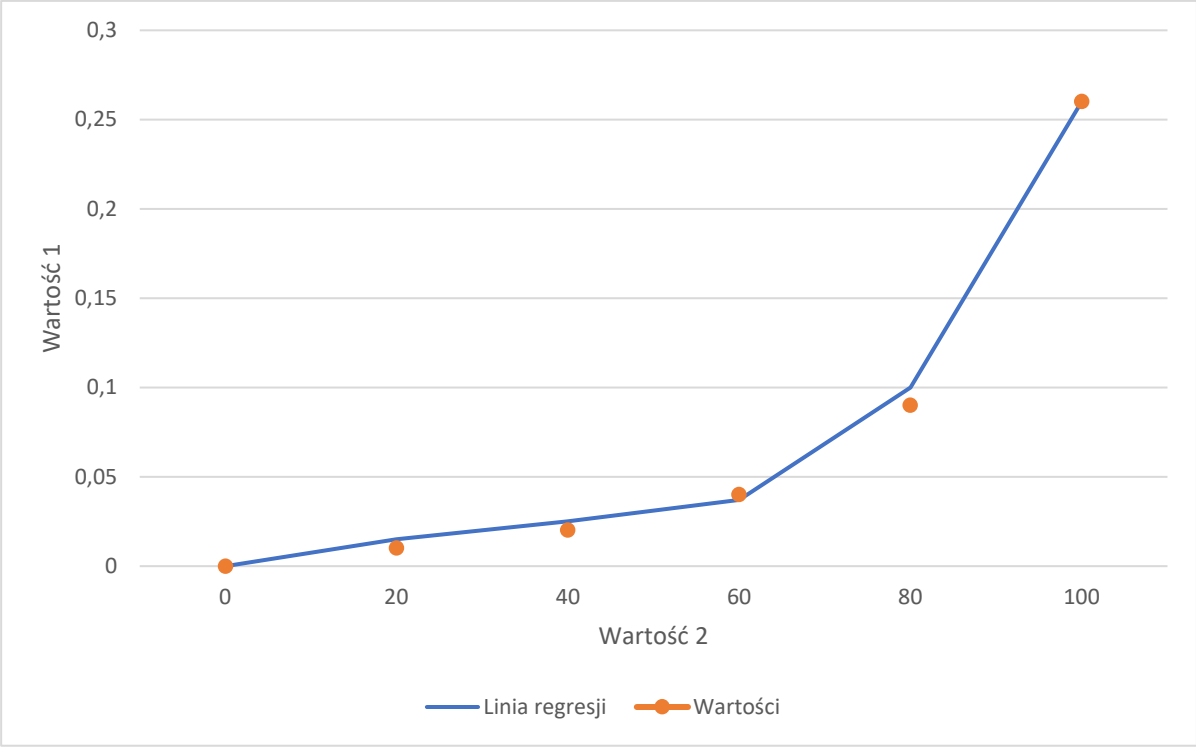
$B0$ – dodatkowy współczynnik

x^n – n wartość wejściowa

BN - współczynnik dla n wartości wejściowej

Regresja dopasowuje się do danych. Wskutek czego otrzymujemy linie regresji (regresja krzywoliniowa) łączące dane w zbiorze.

Rysunek 10. Wykres regresji krzywoliniowej



[Źródło: Opracowanie własne na podstawie https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/ [Data uzyskania dostępu: 13 maj 2022]]

Regresja logistyczna

Regresja logistyczna to technika wywodząca się ze statystyki. Metoda służy do rozwiązywania problemów z klasyfikacją binarną[39]. Nazwa algorytmu pochodzi od funkcji logistycznej, na jakiej opierają się obliczenia. Funkcja ta nazywana również sigmoidalną, powstała w celu opisywania populacji w środowisku[40]. Krzywa przyjmuje kształt litery „S" o wartościach rzeczywistych, które potem przypisuje w przedziale [0,1], lecz nigdy nie osiągnie granic. Wzór na krzywą:

$$\Phi_{sig}(x) = \frac{1}{1 + exp(-x)}$$

Gdzie:

$\Phi_{sig}(x)$ – funkcja sigmoidalna

$exp(-x)$ – funkcja ekspotencjalna

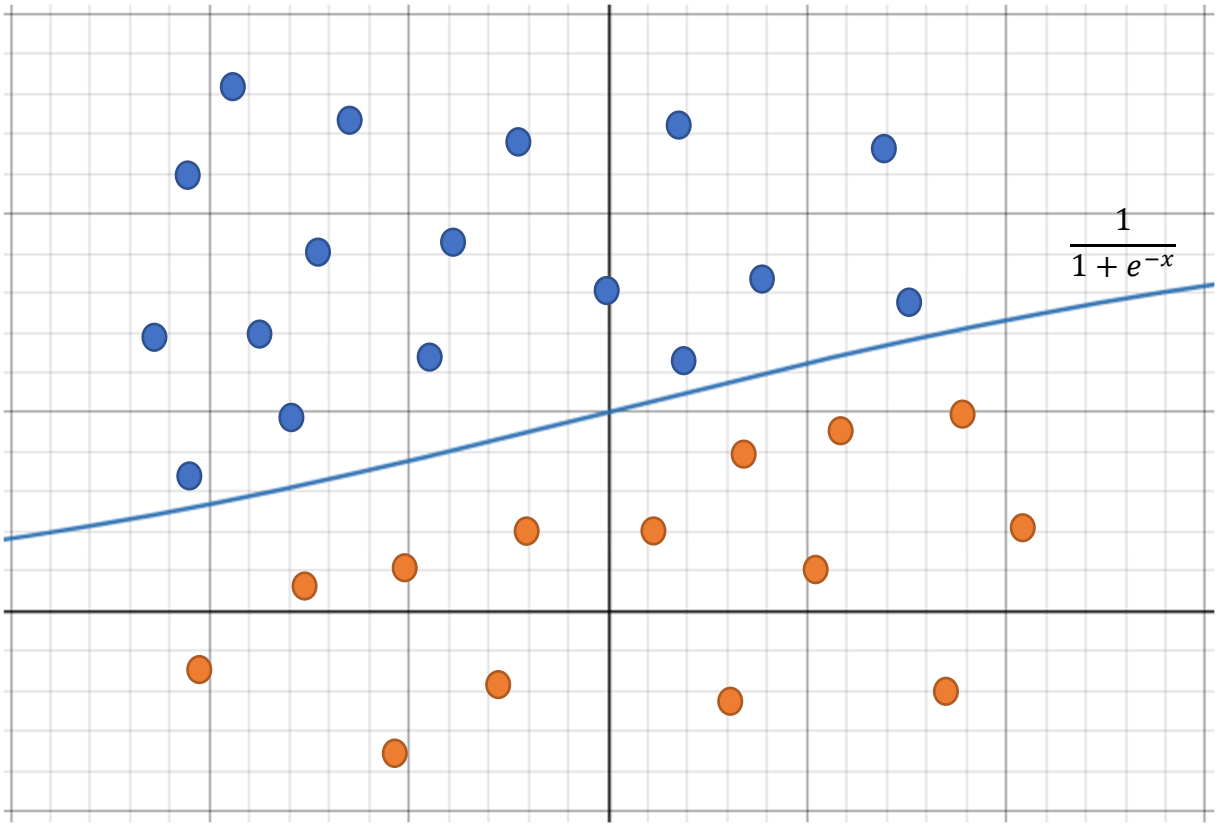
Tak samo jak w regresji liniowej wynik można przedstawić za pomocą równania. Każda z wartości wejściowych x łączone są liniowo za pomocą wag[41]. Rozpatrując wejściową postać $B0 \quad + B1 \quad \times x$ równanie regresji logistycznej przyjmie dla tego przykładu formę:

$$p(class = 0) = \frac{1}{1 + exp^{-(B0 + B1 \times x)}}$$

Wynikiem takiego algorytmu będzie prawdopodobieństwo wyjścia klasy domyślnej (w tym przypadku klasy 0). Rysunek 11 przedstawia przykład podziału zbioru przy pomocy regresji logistycznej.

[39] Klasyfikacja polegająca na przewidywaniu, do której z dwóch klas należy wystąpienie danych
[40] S Shwarts S., Ben D. ,*Understanding Machine Learning*, Cambridge University Press, New York 2014, s. 97-99
[41] Ezukwoke K., Zareian S. ,*Logistic regression and kernel logistic regression A comparative study of logistic regression and kernel logistic regression for binary classification*, University Jean Monnet, France 2019,. s. 2-3

Rysunek 11. Regresja logistyczna



[Źródło: Opracowanie własne na podstawie https://aigeekprogrammer.com/pl/klasyfikacja-binarna-regresja-logistyczna-keras/ [Data uzyskania dostępu: 13 maj 2022]]

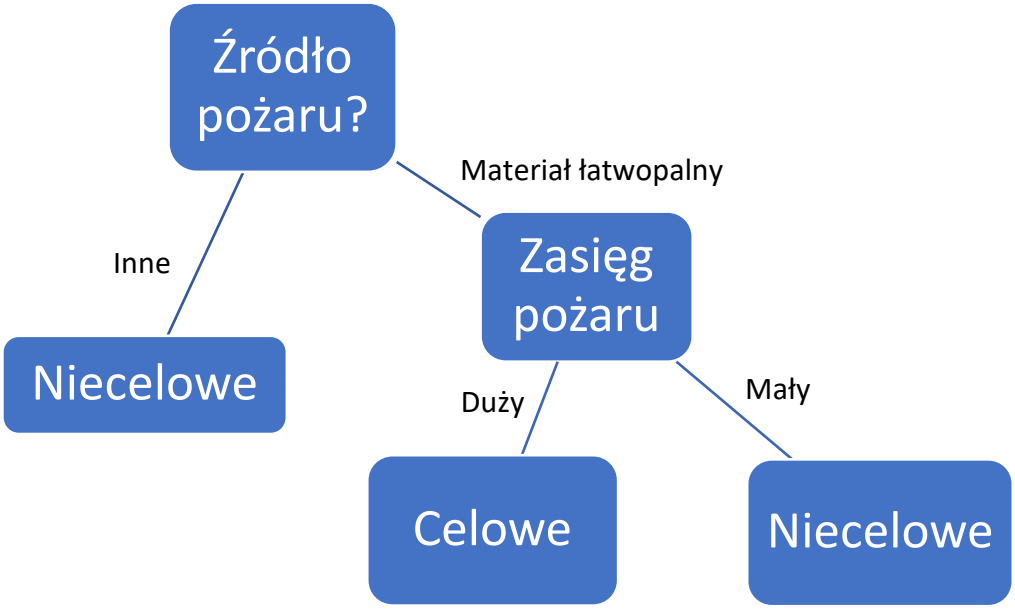
Obliczona linia regresji logistycznej oddziela dwa różniące się od siebie zbiory w klasyfikacji binarnej.

Drzewa decyzyjne[42]

Drzewa decyzyjne znane są również pod nazwą CART[43]. Pojęcie to obejmuje algorytmy drzew klasyfikacji i regresji. W celu przybliżenia modelu posłużono się drzewem binarnym (y ={0,1}), ale metoda pozwala używać go również w innych problemach. W każdym węźle od „korzenia" do „liścia" ścieżka algorytmu dzielona jest zwykle w oparciu o jedną z klas (x), lub z góry określonym zbiorze zasad dzielenia[44]. Przykładem obrazującym działanie algorytmu może być uproszczone drzewo decyzyjne na rysunku 12.

[42] ang. Decision Trees
[43] ang. Classification And Regression Trees
[44] Shwarts S., Ben D. ,*Understanding Machine Learning*, Cambridge University Press, New York 2014, s. 212-214

Rysunek 12. Przykład drzewa decyzyjnego



[Źródło: Opracowanie własne]

Algorytm ma posłużyć do oszacowania czy dany pożar został wywołany celowo. W pierwszej kolejności weryfikowane jest źródło pożaru. Jeśli źródłem pożaru jest inny materiał niż łatwopalny to zakłada się, że dany pożar został wywołany niecelowo bez dodatkowych testów. Jeśli natomiast źródłem jest materiał łatwopalny, to w kolejnym kroku sprawdzany jest zasięg pożaru. Na tym etapie, dla wariantu, że zasięg pożaru jest duży, predykcja dotycząca celowego podpalenia oznacza go jako celowe. Powyższy przykład pokazuje główną zaletę drzew decyzyjnych. Czyli łatwość w zrozumieniu i interpretacji klasyfikatora.

Uczenie binarne drzewa decyzyjnego opiera się na procesie dzielenia przestrzeni wyjściowej w oparciu o algorytm zachłanny[45], który trenowany jest na testowym zbiorze. Wartości i klasy zestawiane są ze sobą, a algorytm sprawdza punkty podziału za pomocą funkcji kosztu[46]. W wyniku czego wybierany jest punkt podziału o najniższym koszcie. Ta zasada jest powtarzana, aż drzewo uzyska swoją ostateczną formę. Dla regresji funkcja kosztu wygląda następująco:

[45] Algorytm zachłanny – algorytm w każdym kroku dokonuje najlepszego w danym momencie rozwiązania częściowego

[46] Chee S., Peck Y., Massoud M. ,*Predictive Analytics in Business Analytics: Decision Tree, Advances in Decision Sciences*, Tajwan 2022, s. 14-15

$$\sum_{i=1}^{n}(y_i - prediction)^2$$

Gdzie:

y_i – wartość wyjściowa dla próby testowej

$prediction$ – prognoza wykonana przez model dla próby testowej

W modelu CART do obliczania punktów podziału wykorzystywana jest funkcja Gini:

$$Gini(split) = \sum_{k=1}^{n} (p_k \times (1 - p_k))$$

Gdzie:

p_k – liczba instancji treningowych należących do klasy k

k – klasa

Punkt podziału, który posiada wszystkie klasy tego samego typu (idealna czystość klas), będzie posiadał Gini=0. Natomiast najgorszy podział będzie gdy Gini=0,5. Zadaniem algorytmu jest wybieranie kolejnych podziałów stosując się do jak najmniejszych kosztów (wartości Gini). W procedurze tej określa się kryterium stopu poprzez wykorzystanie minimalnej liczby możliwych instancji treningowych przypisanych do węzłów drzewa. W przypadku gdy liczba instancji nie osiągnie pewnego minimum algorytm nie dzieli węzła dalej tylko określa go jako końcowy[47].

Las losowy oraz Bagging (ang. Random Forest)

Las losowy oraz Bagging to metody opierające się na podobnych zasadach działania. Fundamentem dla obu algorytmów jest tak zwana metoda Bootstrap, która pozwala szacować wartości z wielu próbek danych. Przykładowo dla próbki z 100 wartościami oszacowujemy średnie z próby. Możemy to wykonać poprzez policzenie średniej z całego zbioru otrzymując wynik obarczony błędem[48]:

[47] Brownlee J. ,*Master Machine Learning Algorithms*, Jason Brownlee, 2016, s.72-75
[48] Brownlee J. ,*Master Machine Learning Algorithms*, Jason Brownlee, 2016, s. 126-127

$$\text{mean}\ (x) = \frac{1}{100} \times \sum_{i=1}^{100} x_i$$

Możliwe jest zniwelowanie błędu wykorzystując metodę bootstrap:

1. Utworzenie np. 1000 losowych podprób z naszego zbioru danych (możliwe jest wybranie wielokrotnie tej samej wartości)
2. Obliczenie średniej z każdej podpróby
3. Obliczenie średniej wszystkich otrzymanych średnich podprób i użycie jej jako szacunkowej średniej dla danych.

Na przykład z 3 prób uzyskano wartości średnie 2.3, 4.5, 3.3. Na ich podstawie można przyjąć szacunkową wartość średniej dla próby równą 3.367. Tej metody można użyć do obliczania innych wartości niż średnia. Wykorzystuje się ją do wyliczeń odchylenia standardowego albo wyuczonych współczynników w algorytmach uczenia maszynowego.

Bootstrap Aggregation (Bagging) korzysta z metody bootstrap w taki sposób, że łączy przewidywanie pochodzące z wielu wykonanych uczeń maszynowych na zbiorze w celu uzyskania lepszej prognozy. Bagging znajduje zastosowanie w przypadku danych o wysokiej wariancji[49]. Polega na wykonaniu kilku modeli, na przykład drzew decyzyjnych, a ilość wykonywanych predykcji określa się w oparciu o próbę walidacyjną, względem której określamy jak dobrze model końcowy potrafi przewidywać wynik.

Lasy losowe są metodą ulepszoną względem metody bagging, która obarczona jest znacznym podobieństwem strukturalnym drzew decyzyjnych, a co za tym idzie, wysoką korelacją w ich przewidywaniach. W algorytmie lasu losowego uczenie się poddrzew wykonywane jest tak, aby ograniczyć korelację pomiędzy poddrzewami. W CART algorytm dokonuje podziału na podstawie wszystkich zmiennych. Natomiast lasy losowe zmieniają tą procedurę, w taki sposób, że algorytm ograniczony jest do losowej próbki funkcji do przeszukania. Ilość funkcji m, które można przeszukać przy każdym podziale określa się jako parametr algorytmu[50]. Dobra wartość domyślna dla modelu to:

Problem klasyfikacji: $m = \sqrt{p}$

[49] Louppe G. ,*Understanding Random Forest*, University of Liège, Liège 2014, s. 69-71
[50] Brownlee J. ,*Master Machine Learning Algorithms*, Jason Brownlee, 2016, s. 128

Problem regresji: $m = \frac{p}{3}$

Gdzie:

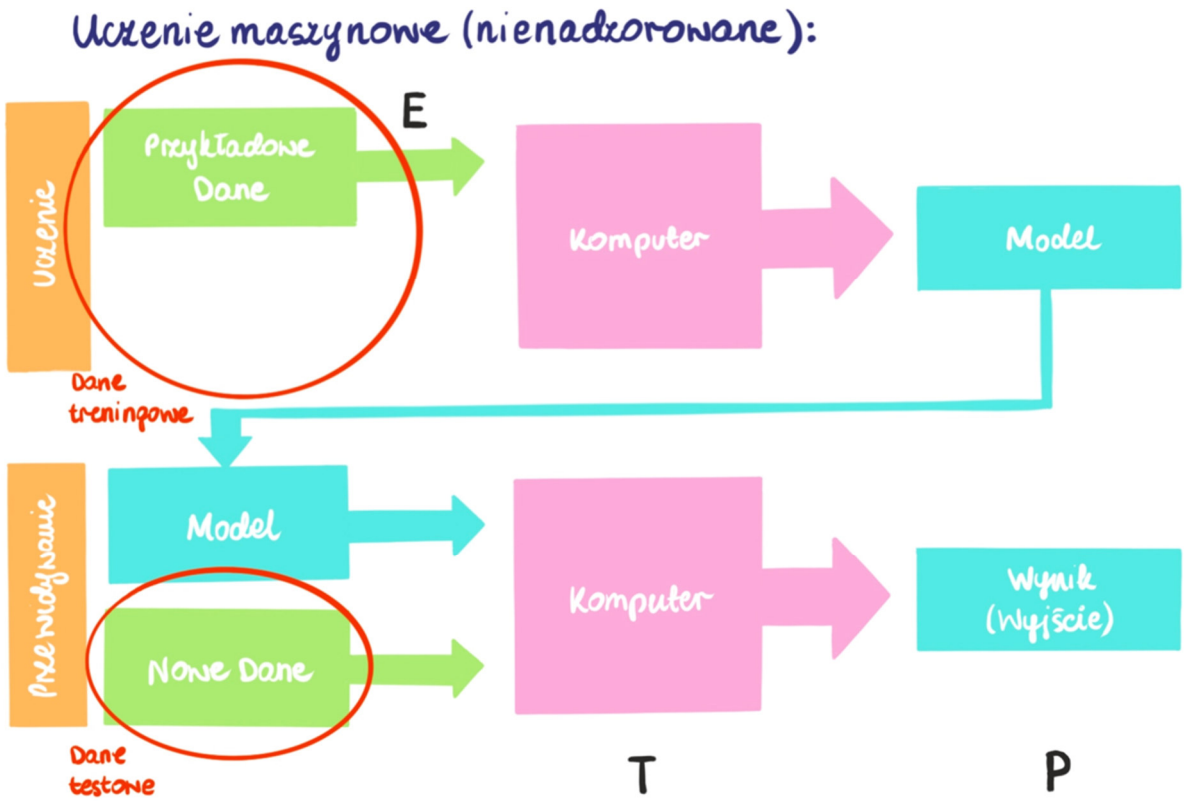
m – ilość cech losowo wybranych

p – liczba zmiennych wejściowych

2.3. Uczenie nienadzorowane

W uczeniu nienadzorowanym algorytm uczy się na podstawie zbioru danych, który składa się jedynie z cech oraz atrybutów wejściowych, ale nie zawiera etykiet(nie jest oznaczone przez człowieka). Model tworzony jest przy minimalnej ingerencji człowieka[51]. W związku z tym różnicą względem uczenia nadzorowanego jest to, że algorytm w tej metodzie wyszukuje wzorce z dostarczonych danych, a nie z wcześniej podanych przykładów (zbioru testowego). Na rysunku 13. przedstawiony jest schemat działania uczenia nienadzorowanego.

Rysunek 13. Model działania nienadzorowanego systemu uczącego



[Źródło: Opracowanie własne na podstawie Uczenie maszynowe nienadzorowane, https://neptune.ai/blog/life-cycle-of-a-machine-learning-project [Data uzyskania dostępu: 15 czerwiec 2022]]

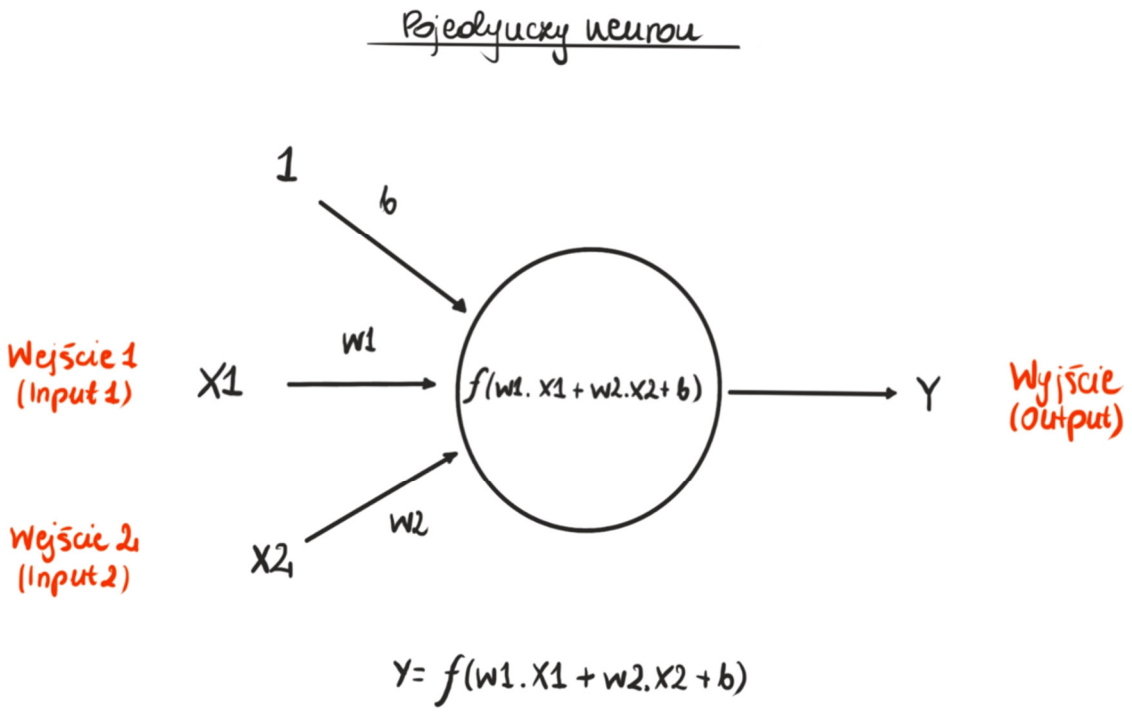
[51] Sawka K. ,*Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow*, Helion, Gliwice 2020,

Sztuczne sieci neuronowe

Metoda ta powstała poprzez wzorowanie się na procesach biologicznych zachodzących w ludzkich mózgach. W uproszczonym modelu ludzkiego mózgu składa się on z wielu urządzeń obliczeniowych „neuronów" połączonych ze sobą złożoną siecią komunikacyjną. Dzięki temu możliwe są skomplikowane obliczenia. Sztuczne sieci neuronowe, zaproponowane w połowie XX wieku, to graf skierowany, którego węzły odpowiadają neuronom, a krawędzie łączą wyjścia neuronów z wejściem innego[52].

W sieci podstawową jednostką obliczeniową jest tak zwany perceptron, który symbolizuje neuron[53]. Węzeł neuronu przyjmuje wektor $\vec{x}$, który jest zbiorem cech wejściowych poszczególnych $x_1, x_2, x_3, \cdots$. W perceptronie, jako dane wejściowe, uzyskuje ważoną sumę z wartości przychodzących do neuronu. Następnie neuron zwraca wyjściową wartość binarną[54].

Rysunek 14. Matematyczna reprezentacja jednego neuronu



[Źródło: Opracowanie własne na podstawie Matematyczna reprezentacja jednego neuronu, https://towardsdatascience.com/power-of-a-single-neuron-perceptron-c418ba445095 [Data uzyskania dostępu: 24 czerwiec 2022]]

[52] Louppe G. ,*Understanding Random Forest*, University of Liège, Liège 2014, s. 22-23
[53] Gurney K. ,*An introduction to neural networks*, Taylor & Francis e-Library, Londyn 2004, s. 30
[54] Shwarts S., Ben D. ,*Understanding Machine Learning*, Cambridge University Press, New York 2014, s. 229-230

Jak przedstawione jest to na rysunku 14., każde wejście do węzła neuronu x_i posiada przypisaną wagę. Wartość ta przypisywana jest w zależności od znaczenia względem innych danych wejściowych. Wprowadzone jest również dodatkowe wejście zwane „bias”. Wartość wejścia wynosi 1, a jego waga to b. Wykorzystanie dodatkowej wartości wejścia pozwala zwiększyć możliwość uczenia się oraz umożliwia przesunięcie progu aktywacji w zależności od wagi b.[55] Wewnątrz perceptronu wykonywana jest funkcja nieliniowa nazywana funkcją aktywacji. Wzór funkcji wygląda następująco:

$$Y = f(w_1x_1 + w_2x_2 + \cdots + b) = f\left(\sum_{i=1}^{n} , w_ix_i + b\right)$$

Gdzie:

w_1 – waga wejścia

x_1 – wejście do neuronu

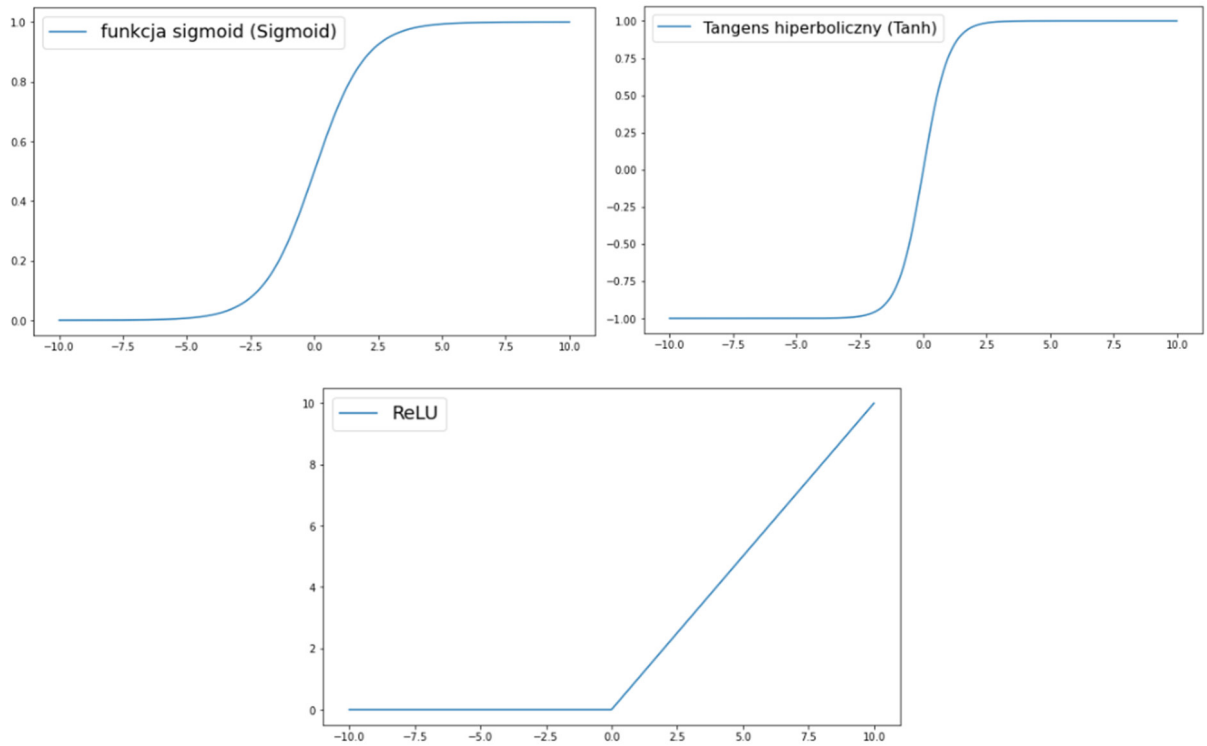
f - funkcja

b – bias

[55] Haykin S. ,*Neural Networks and Learning Machines, Pearson Prentice Hall*, New Jersey 2009, s. 49

Najpopularniejsze funkcje aktywacji jakie używane są do tego typu obliczeń przedstawiono na rysunku 15.

Rysunek 15. Najpopularniejsze funkcje aktywacji



*[Źródło: Opracowanie własne na podstawie Najpopularniejsze funkcje aktywacji, https://cdn-images-1.medium.com/max/1000/1*nFmfF-SAMZc9_jtFNullRg.png [Data uzyskania dostępu: 24 lipiec 2022]*

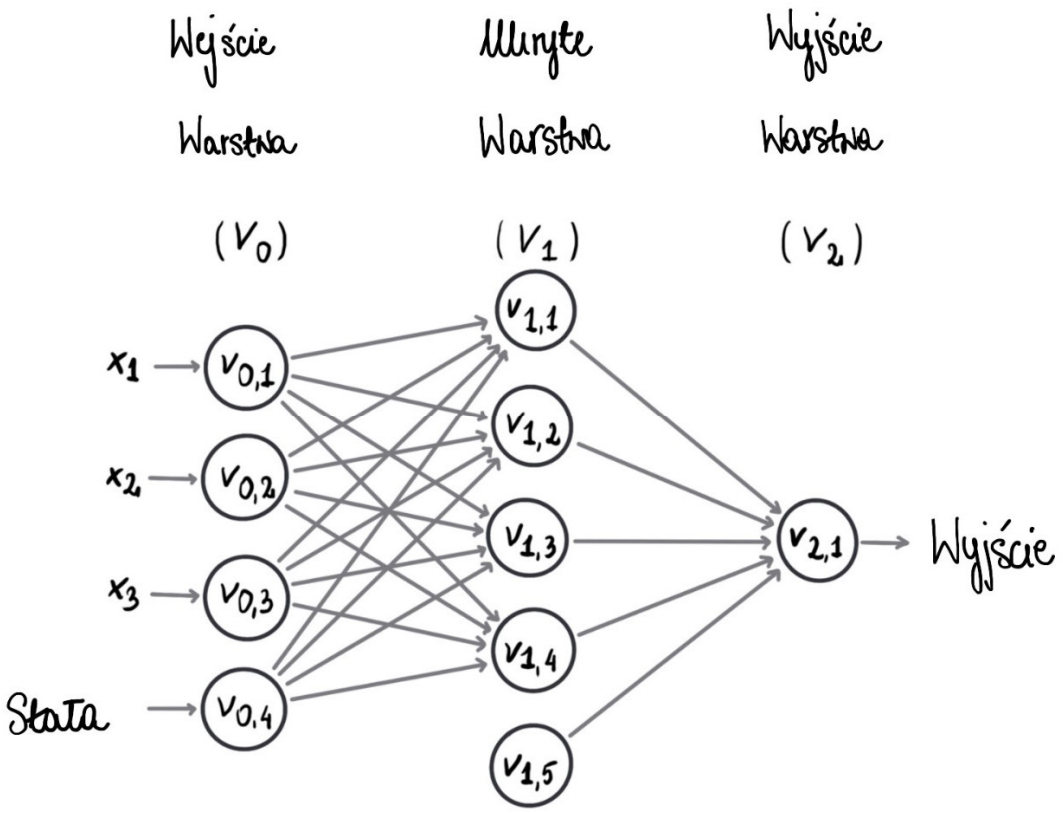
Przedstawione powyżej funkcje to od lewej:

- ReLU: $f(x) = max(0, x)$, zwraca wartość poniżej zera równe 0
- Funkcja sigmoidalna: $\sigma(x) = \frac{1}{1+e^{-x}}$, zwraca wartość od 0 do 1
- Tangens hiperboliczny: $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$, zwraca wartość od -1 do 1

Podstawową strukturą sieci neuronów jest typ „feedforward”, który nie zawiera cykli. Przepływ danych, na takim schemacie, odbywa się od skrajnych lewych neuronów zwanych wejściowymi. Na rysunku 3.13. można również zobaczyć neuron stały, który zawsze wyprowadza wartość 1 (stała). Ich zadaniem jest dostarczenie zewnętrznych danych do neuronów ukrytych. Neurony te nazywają się tak z powodu tego, że nie mają połączenia ze światem zewnętrznym. Należy zwrócić uwagę, że w warstwie ukrytej również umieszczony jest jeden neuron, którego zadanie jest zwracanie stałej wartości. Ostatnią warstwą jest warstwa wyjściowa, która w przypadku rysunku 16. zawiera jeden neuron wyjściowy (może

również zawierać ich wiele). Neuron lub neurony w tej warstwie odpowiedzialne są za przesyłanie informacji obliczonych przez sieć na zewnątrz[56].

Rysunek 16. Przykład sztucznej sieci neuronowej



[Źródło: Opracowanie własne na podstawie Shwarts S., Ben D. ,Understanding Machine Learning, Cambridge University Press, New York 2014, s.230.]

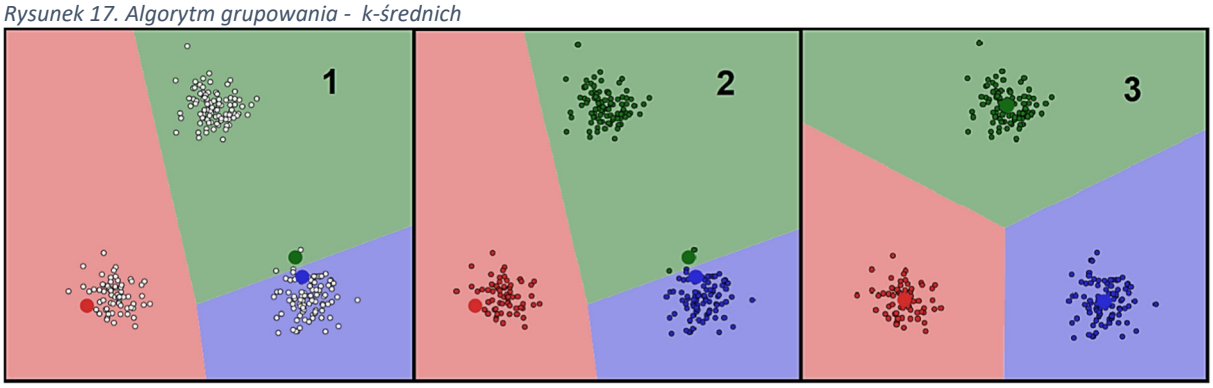
Algorytm k-średnich (centroidów)

Jest to algorytm należący do grupy modeli analizy skupień. Oznacza to, że sposób jego działania polega na szukaniu i dzieleniu grup obiektów podobnych do siebie[57]. Ostatecznym wynikiem algorytmu jest uzyskanie podziału punktów w taki sposób, żeby punkty z danego klastra były do siebie jak najbardziej podobne i tym samym jak najbardziej różniły się od innych grup. Podział odbywa się dzięki wyznaczeniu punktów ciężkości dla klastrów, tak zwanych centroidów, które „przyciągają" do siebie najbliżej leżące punkty do swojej grupy. Następnie proces ten jest powtarzany, aż do uzyskania idealnego podziału.

[56] Shwarts S., Ben D. ,*Understanding Machine Learning*, Cambridge University Press, New York 2014, s. 230
[57] Kajstura A. ,*Metoda k-średnich*, https://www.statystyka.az.pl/analiza-skupien/metoda-k-srednich.php. [Data uzyskania dostępu: 3 marzec 2022]

Proces działania algorytmu można przedstawić za pomocą rysunku 17. W przestrzeni zbioru umieszczane są w losowych miejscach centroidy(Rys 17 krok 1). Ich liczba zależy od przyjętych założeń. W omawianym przypadku ich ilość wynosi trzy. Po wprowadzeniu punktów wyliczane są najczęściej za pomocą odległości euklidesowej odległości analizowanych punktów danych od centroidów. W oparciu, o otrzymane wartości średniej arytmetycznej dystansu między wszystkimi punktami danych, centroidy są przesuwane (Rys 17 krok 2). Krok drugi powtarzany jest do momentu uzyskania założonego warunku końcowego. Na przykład uzyskania podziału grup, w takim stopniu, że między kolejnymi obliczeniami nie następuje zmiana przynależności punktów danych[58]. Wówczas uzyskuje się stan przedstawiony na Rys 17 w kroku 3.

Rysunek 17. Algorytm grupowania - k-średnich



[Źródło: Opracowanie własne na podstawie Algorytm k-średnich , http://itcraftsman.pl/algorytm-k-srednich-uczenie-nienadzorowane/ [Data uzyskania dostępu: 24 maj 2022]]

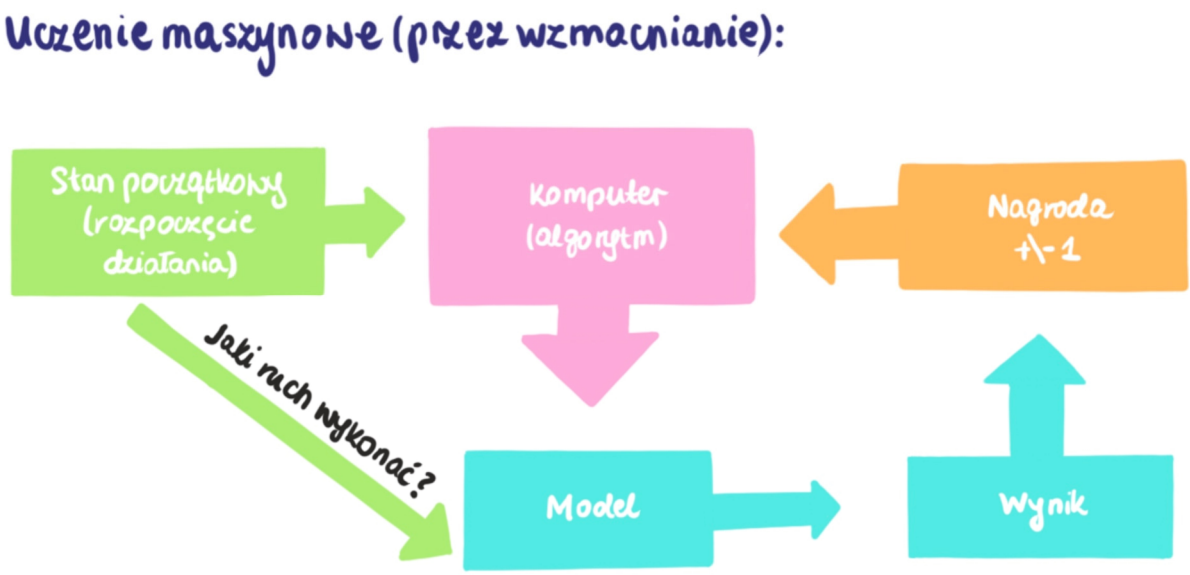
2.4. Uczenie przez wzmacnianie

Uczenie przez wzmacnianie jest trzecim rodzajem uczenia maszynowego. Od uczenia nadzorowanego różni się tym, że dane uzyskiwane nie są opisane i cały proces uczenia odbywa się bez nadzoru człowieka. Za to główną różnicą względem metody uczenia nienadzorowanego jest to, że celem algorytmu nie jest znalezienie ukrytych zależności w zbiorze danych, a maksymalizacja sygnału nagrody[59]. Modele tego typu nie potrzebują danych wejściowych. Jedyną informacją jaką otrzymuje jest informacja o tak zwanej „nagrodzie", albo „karze". Algorytm uzyskuje taki wynik po wykonaniu cyklu uczenia. W przypadku uzyskania pozytywnego wyniku program wie, że jego działania idą w dobrą stronę, natomiast negatywny

[58] Shwarts S., Ben D. ,*Understanding Machine Learning*, Cambridge University Press, New York 2014, s. 268-275
[59] Sutton R., Bart A. ,*Reinforcement Learning: An Introduction*, The MIT Press, Londyn 2015, s. 3-5

wynik oznacza potrzebę zmiany. W uproszczeniu, podejście do problemu od tej strony, można nazwać metodą prób i błędów. Schemat działania przedstawiono na rysunku 18.

Rysunek 18. Uczenie maszynowe przez wzmacnianie



[Źródło: Opracowanie własne na podstawie Uczenie maszynowe przez wzmacnianie, https://techvidvan.com/tutorials/reinforcement-learning/ [Data uzyskania dostępu: 13 maj 2022]]

Model uczenia przez wzmacnianie jest tym lepszy im więcej cykli nauki program odbył. Pozwala mu to na nauczenie się większej ilości wariantów w celu podejmowania odpowiednich działań tak, aby zmaksymalizować pozytywny bodziec (nagrodę) lub zminimalizować negatywny (karę).

2.5. Ewaluacja modeli

Ewaluacja modeli[60] ma na celu ocenę jakości i przydatności zbudowanych modeli predykcyjnych. Ze względu na różne typy problemów oraz danych w praktyce wykorzystywane jest wiele miar służących do oceny wyników. Ich wybór zależy od celu jaki ma spełniać model.

Metody ewaluacji

Przetrzymanie[61] polega na podziale dostępnego zbioru danych na dwa podzbiory. Najczęściej w proporcji 1:2, 1:3 na kolejno zbiór testowy i zbiór uczący. Zbiór uczący służy do trenowania modelu opartego o wybrany klasyfikator, natomiast po przeprowadzeniu

[60] Ang. Model evaluation
[61] Ang. Hold-out

dopasowania jakość modelu sprawdzana jest w oparciu o dane z podzbioru testowego. Wówczas określana jest jego zdolność do poprawnej predykcji.

Odmianą metody przetrzymania jest ewaluacja na k-częściach[62], która polega na podziale dostępnego zbioru danych na podzbiór testowy oraz k zbiorów uczących. Analogicznie jak w metodzie przetrzymania modele trenowane są na k zbiorach uczących, a następnie wybierany jest, ten który daje najlepsze wyniki na zbiorze testowym.

Metryki ewaluacji

Metryki ewaluacji służą do przedstawienia jakości zbudowanych modeli w oparciu o dostępne znane wyniki w wykorzystywanym zbiorze danych oraz predykcji zbudowanych modeli.

Tabela 3. Tablica pomyłek

		Klasa predykowana	
	Populacja	Klasyfikacja pozytywna	Klasyfikacja negatywna
Klasa rzeczywista	Stan pozytywny	Prawdziwie dodatni (TP)	Fałszywie ujemny (FN)
	Stan negatywny	Fałszywie dodatni (FP)	Prawdziwie ujemny (TN)

[Źródło: Opracowanie własne]

Metryki bazują na tablicy pomyłek, którą przedstawiono w tabeli 3. Dla problemów niebinarnych wykorzystuje się rozbudowane macierze pomyłek, których w tej pracy, ze względu na badaną binarną cechę, nie przedstawiono. Dla problemu binarnego wyróżnić można:

- Czułość[63] - opisuje odsetek prawdziwie dodatnich wyników w stosunku do łącznej ilości zaklasyfikowanych stanów pozytywnych. Opiera się na wzorze:

$$\text{Czułość } = \frac{TP}{TP + FN}$$

[62] Ang. K-fold evaluation
[63] Ang. Recall

- Swoistość[64] - opisywana jest jako odsetek prawdziwie negatywnych wyników w stosunku do łącznej wartości stanu negatywnego. Wzór na swoistość:

$$\text{Swoistość } = \frac{TN}{TN + FP}$$

- Precyzja[65] - stosunek prawdziwie dodatnich predykcji do sumy klasyfikacji pozytywnej. Zgodnie ze wzorem:

$$\text{Precyzja } = \frac{TP}{TP + FP}$$

- Dokładność[66] - suma prawdziwie pozytywnych i prawdziwie negatywnych klasyfikacji w stosunku do całej populacji w danych. Obliczana jest ona ze wzoru:

$$\text{Dokładność } = \frac{TP + TN}{Suma\ populacji}$$

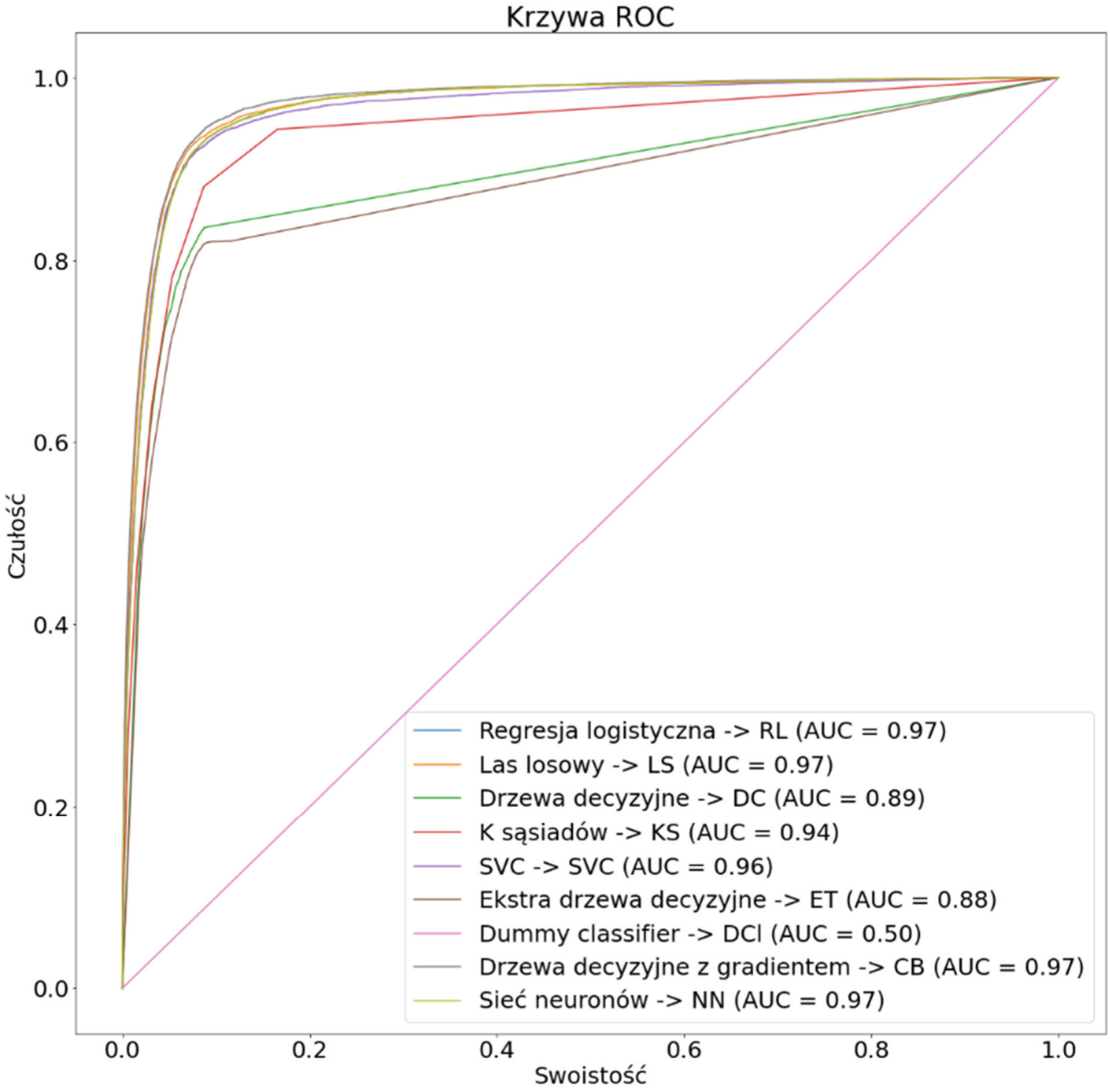
- F1 – miara ta opiera się na porównaniu wskaźnika precyzji i czułości przy użyciu średniej harmonicznej. Wzór pozwalający obliczyć tą miarę to:

$$F = 2 \cdot \frac{\text{Precyzja } \cdot \text{Czułość}}{\text{Precyzja } + \text{ Czułość}}$$

[64] Ang. Specifinicy
[65] Ang. Precision
[66] Ang. Accuracy

Innym powszechnie używanym wskaźnikiem jest krzywa ROC. Wykres ten zestawia ze sobą dwie miary. Czułość jako oś Y oraz swoistość jako oś X. Na przykładowej krzywej ROC przedstawionej na rysunku 19 pokazany jest przebieg krzywej dla kilku modeli.

Rysunek 19. Przykładowa krzywa ROC dla kilku modeli, którą uzyskano podczas prac według rozdziału trzeciego



[Źródło: Opracowanie własne na podstawie wyników niniejszej pracy]

W idealnym modelu krzywa ta powinna przechodzić przez punkt (0, 1), co oznaczałoby, że klasyfikator nie myli się w predykcji. Graficzne przedstawienie tych dwóch parametrów pozwala na porównanie jakości różnych klasyfikatorów.

3. MODELE PREDYKCYJNE CELOWEGO PODPALENIA BUDYNKU

Na podstawie dostępnej wiedzy i danych, w niniejszym rozdziale, wykonano algorytm w języku python, który umożliwił predykcję czy dany pożar został wywołany celowo. Posłużyły do tego ogólnodostępne dane publikowane przez rząd Wielkiej Brytanii. Przy ich pomocy wykonano optymalizację parametrów poszczególnych klasyfikatorów i porównano ich wyniki na zbiorze testowym.

3.1. Charakterystyka danych

Dane użyte w niniejszej pracy pochodzą z oficjalnych stron rządu Wielkiej Brytanii. Wykorzystano trzy bazy danych:

1. "Dane dotyczące pożarów I ratownictwa w Anglii, na rok kończący w marcu 2021r"[67] pobranych dnia 15.01.2022r ze strony: https://www.gov.uk/government/statistics/fire-statistics-incident-level-datasets
2. „Regionalny produkt krajowy brutto: wszystkie regiony"[68] pobranych dnia 15.01.2022r ze strony: https://www.ons.gov.uk/economy/grossdomesticproductgdp/datasets/regionalgrossdomesticproductallnutslevelregions
3. „Estymowana populacja Wielkiej Brytanii"[69] pobranych dnia 15.01.2022r ze strony: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland

Ze względu na to, że podstawowe dane na jakich się opierano czyli "Dane dotyczące pożarów i ratownictwa w Anglii, na rok kończący w marcu 2021r" nie zawierały informacji ekonomicznych i demograficznych postanowiono dodać je na bazie innych danych - w oparciu o FRS_TERYTORIUM jednostek straży pożarnej opracowano dwie dodatkowe zmienne odpowiadające za produkt krajowy brutto na głowę (GDP_NA_GŁOWĘ) oraz zagęszczenie ludności (ZAGĘSZCZENIE_LUDNOŚCI). Wszystkie etykiety przedstawiono w tabeli 4.

[67] Ang. Fire and rescue incident level dataset England, year ending March 2021
[68] Ang. Regional gross domestic product: all ITL regions
[69] Ang. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland

Tabela 4. Etykiety danych wykorzystanych w pracy

Zmienna w języku angielskim	**Zmienna w języku polskim**	**Zmienna wykorzystana do budowania modeli predykcyjnych**
FRS_TERRITORY	FRS_TERYTORIUM	Wykluczona
E_CODE_TERRITORY	E_KOD_TERYTORIUM	Wykluczona
FINANCIAL_YEAR	ROK_FINANSOWY	Wykluczona
GDP_PER_CAPITA	GDP_NA_GŁOWĘ	Ujęta
POPULATION_DENSITY	ZAGĘSZCZENIE_LUDNOŚCI	Ujęta
WEEKDAY_WEEKEND	DNI POWSZEDNIE_WEEKEND	Ujęta
DAY_NIGHT	PORA_DNIA	Ujęta
DWELLING_TYPE	RODZAJ_OBIEKTU	Wykluczona, do analizy wykorzystano tylko domy jednorodzinne
BUILDING_SPECIAL_CONSTRUCTION_DESCRIPTION	SPECJALNA_BUDOWA_OBIEKTU	Ujęta
OCCUPANCY_TYPE	ILOŚĆ_MIESZKAŃCÓW_OBIEKTU	Ujęta
OCCUPIED_NORMAL	CZY_BUDYNEK_JEST_NORMALNIE_ZAJĘTY	Ujęta
ALARM_SYSTEM	CZY_CZUJNIK_DYMU_JEST_W_OBIEKCIE	Ujęta
NO_ALARM	LICZBA_CZUJNIKÓW_DYMU	Ujęta
ALARM_SYSTEM_TYPE	RODZAJ_CZUJNIKA_DYMU_W_MIESZKANIU	Ujęta
ALARM_REASON_FOR_POOR_OUTCOME	POWÓD_ZŁEGO_ZADZIAŁANIA_ALRMU	Ujęta
IGNITION_TO_DISCOVERY	CZAS_DO_ZAUWAŻENIA_POŻARU	Ujęta
DISCOVERY_TO_CALL	CZAS_MIĘDZY_ODKRYCIEM_A_ZGŁOSZENIEM	Ujęta
LATE_CALL	PÓŹNE_WYWOŁANIE	Ujęta
ACCIDENTAL_OR_DELIBERATE	PRZYPADKOWE_CZY_CELOWE	Ujęta
IGNITION_POWER	CO_NAPĘDZA_ZAPŁON	Ujęta

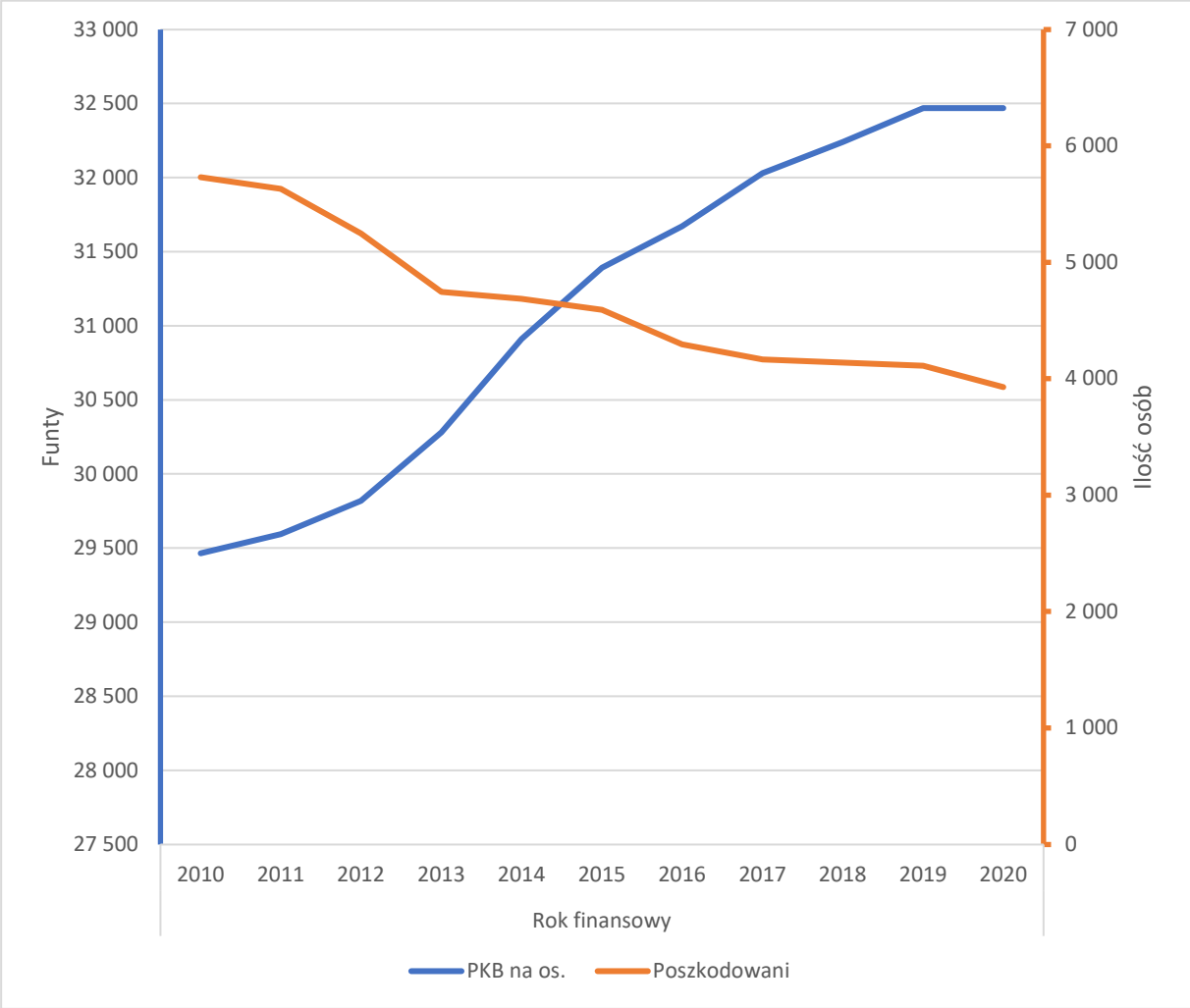
Zmienna w języku angielskim	Zmienna w języku polskim	Zmienna wykorzystana do budowania modeli predykcyjnych
SOURCE_OF_IGNITION	ŹRÓDŁO_POŻARU	Ujęta
FIRE_START_LOCATION	LOKACJA_ROZPOCZĘCIA_POŻARU	Ujęta
FIRE_SIZE_ON_ARRIVAL	ZASIĘG_OGNIA_PO_PRZYBYCIU	Ujęta
OTHER_PROPERTY_AFFECTED_ON_ARRIVAL	CZY_SĄSIEDNI_OBIEKT_ZOSTAŁ_DOTKNIĘTY_POŻAREM	Ujęta
ITEM_IGNITED	PRZEDMIOT_KTÓRY_ZAPŁONĄŁ	Ujęta
ITEM_CAUSING_SPREAD	PRZEDMIOT_ROZPRZESTRZENIAJĄCY	Ujęta
RAPID_FIRE_GROWTH	SZYBKI_WZROST_POŻARU	Ujęta
VEHICLES	POJAZDY_STRAŻY	Wykluczona
VEHICLES_CODE	KOD_POJAZDÓW	Ujęta
PERSONNEL	PERSONEL	Wykluczona
PERSONNEL_CODE	KOD_PERSONELU	Ujęta
RESPONSE_TIME	CZAS_ODPOWIEDZI	Wykluczona
RESPONSE_TIME_CODE	KOD_CZAS_ODPOWIEDZI	Ujęta
TIME_AT_SCENE	CZAS_INCYDENTU	Wykluczona
TIME_AT_SCENE_CODE	KOD_CZAS_INCYDENTU	Ujęta
FATALITY_CASUALTY	POSZKODOWANI_MARTWI	Ujęta
RESCUES	URATOWANI	Ujęta
EVACUATIONS	EWAKUOWANI	Wykluczona
EVACUATIONS_CODE	KOD_EWAKUOWANI	Ujęta
FIRE_DAMAGE_EXTENT	OBSZAR USZKODZONY PRZEZ POŻAR	Wykluczona
FIRE_DAMAGE_EXTENT_CODE	KOD_OBSZAR_USZKODZONY_PRZEZ_POŻAR	Ujęta
TOTAL_DAMAGE_EXTENT	OBSZAR USZKODZONY PRZEZ POŻAR_ORAZ_WODĘ	Wykluczona
TOTAL_DAMAGE_EXTENT_CODE	KOD_OBSZAR USZKODZONY PRZEZ POŻAR_ORAZ_WODĘ	Ujęta
SPREAD_OF_FIRE	ZASIĘG_POŻARU_PO_ZGASZENIU	Wykluczona
OTHER_PROPERTY_AFFECTED_CLOSE	SĄSIEDNI_OBIEKT_USZKODZONY	Ujęta

[Źródło: Opracowanie własne na podstawie opracowywanych danych]

Przed przystąpieniem do wykonywania modeli przeprowadzono eksplorację danych. Miało to na celu przyjrzeć się strukturze danych, liczby szukanych przypadków oraz określenie danych jakie będą wykorzystywane w modelowaniu. W tym celu sporządzono wykresy i zestawienia przedstawiające poszczególne dane.

Rysunek 20. pokazuje średnie PKB na osobę w stosunku do liczby ofiar lub poszkodowanych w pożarach. Zauważono, że obie wartości są ze sobą związane. Wzrost PKB na osobę łączy się ze spadkiem liczby osób jakie ucierpiały w wyniku pożaru.

Rysunek 20. Wykres przedstawiający średnie PKB na osobę oraz liczby poszkodowanych lub ofiar pożarów między 2010, a 2020 rokiem w Wielkiej Brytanii

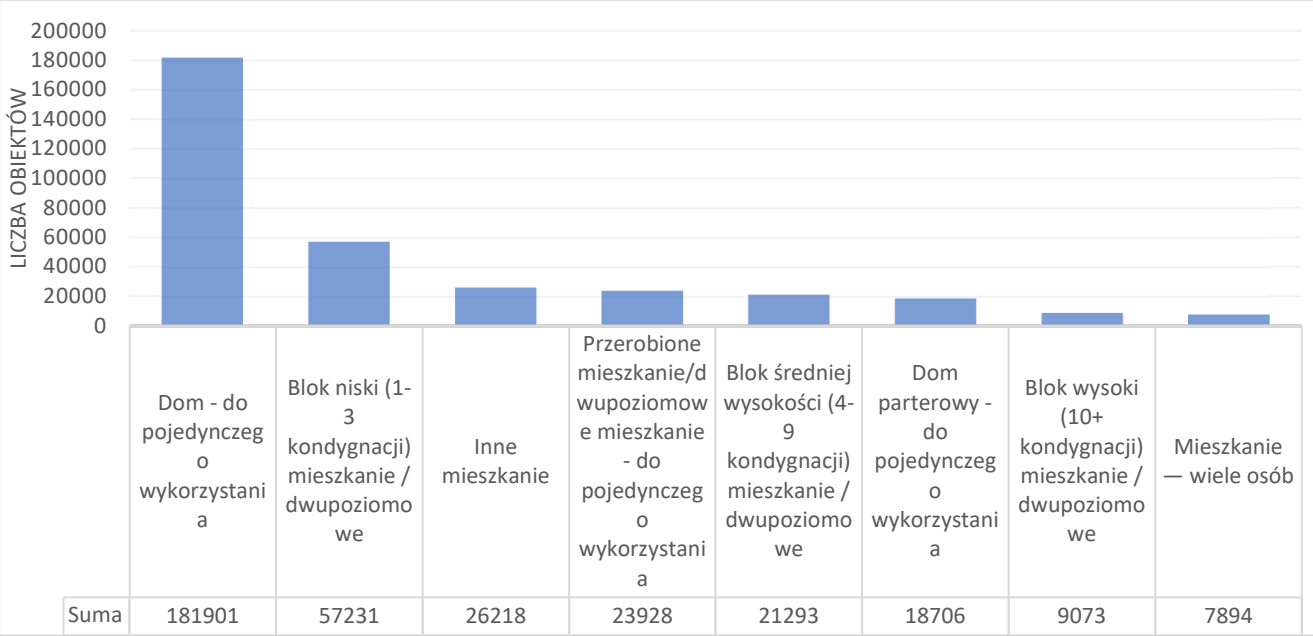


[Źródło: Opracowanie własne na podstawie danych wykorzystanych w pracy]

Wykres słupkowy na rysunku 21 przedstawia ilość obiektów, w których wybuchły pożary. Z zestawienia wynika, że w okresie od 2010 do 2020 roku, pożary najczęściej wybuchały w domach jednorodzinnych. Ze względu na dysproporcję między domami jednorodzinnymi, a resztą wyróżnionych obiektów postanowiono budowę modeli oprzeć tylko

i wyłącznie o domy jednorodzinne. Miało to za zadanie zwiększyć precyzję modeli, dla tego konkretnego przypadku.

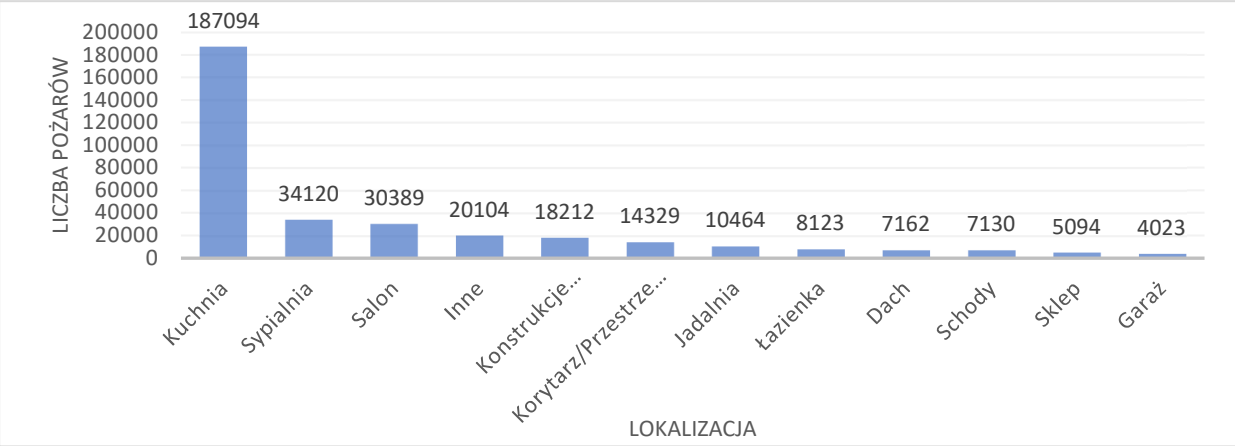
Rysunek 21. Suma obiektów w jakich odnotowano pożar w latach 2010-2020 w Wielkiej Brytanii



[Źródło: Opracowanie własne na podstawie danych wykorzystanych w pracy]

Najczęstszą lokalizacją, w której dochodziło do zaprószenia ognia jest kuchnia. Łączna ilość zdarzeń to ponad 180 tysięcy. Kolejnym pomieszczeniem, gdzie najczęściej dochodzi do rozpoczęcia pożaru jest sypialnia i jest to około 32 tysięcy przypadków. Trzecim miejscem co do ilości rozpoczęcia zdarzenia jest salon. Ilość wyników dla tego pomieszczenia to około 28 tysięcy.

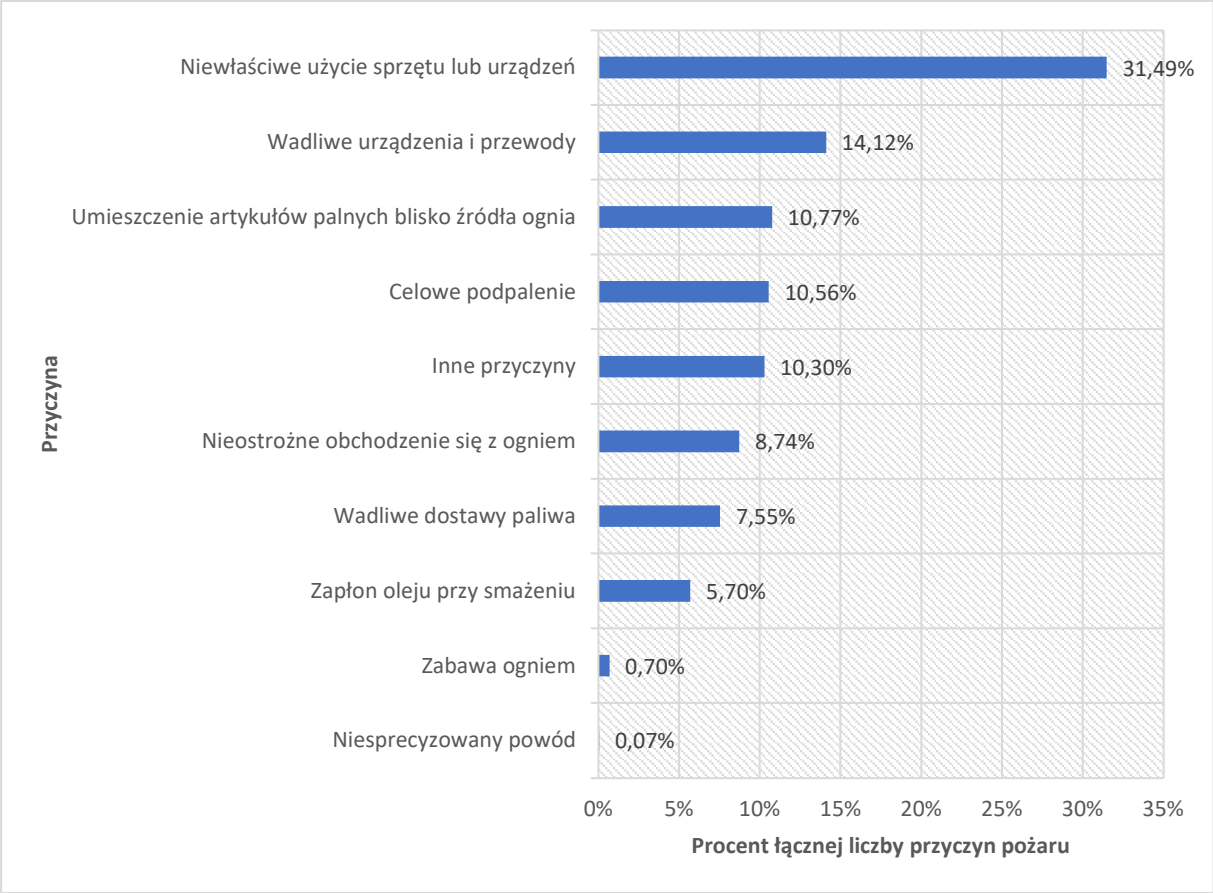
Rysunek 22. Liczba pożarów ze względu na lokalizację rozpoczęcia w latach 2020-2020 w Wielkiej Brytanii



[Źródło: Opracowanie własne na podstawie danych wykorzystanych w pracy]

Z rysunku 23, na którym przedstawiono przyczyny pożarów wynika, że w 31,49% wystąpienie pożaru wynika z niewłaściwego używania sprzętu lub urządzeń. Za 10,56% zdarzeń odpowiadają celowe podpalenia. W podobnej ilości przypadków - 10,77%, powodem pożaru jest umieszczanie artykułów palnych obok źródeł ognia.

Rysunek 23. Przyczyny pożarów w ujęciu procentowym w latach 2010-2020 w Wielkiej Brytanii



[Źródło: Opracowanie własne na podstawie danych wykorzystanych w pracy]

Ze względu na to, że część danych nie wiąże się z tym czy pożar został wywołany celowo oraz część danych została przedstawiona na dwa sposoby(w formie kodu i w formie tekstowej) wyłączono je z dalszej analizy, zgodnie z tabelą 4.

Wykorzystywane dane wymagały również usunięcia braków oraz przetłumaczenia z języka angielskiego. W kolejnym kroku dokonano podziału danych na numeryczne i kategoryczne w celu ich przetworzenia. Dane jakościowe zostały zamienione na zmienne binarne, aby umożliwić modelowanie badanego zjawiska. Przykładowe wartości po przekształceniu, w formie tabelarycznej, przedstawia załącznik nr 1.

3.2. Budowanie modeli predykcyjnych

Do modelowania, czy pożar został wywołany celowo, wykorzystano środowisko Python wraz z dostępnymi bibliotekami takimi jak Scikit-learn, LightGBM oraz Catboost. Utworzono w ten sposób modele wykorzystujące różne klasyfikatory w celu wstępnego wyboru najlepszego. Dane podzielono na grupę uczącą i testową. Na ich bazie wykonano dopasowanie i sprawdzenie wyników dla poszczególnych modeli. Ze względu na ograniczone zasoby obliczeniowe wybór najlepszego z klasyfikatorów podzielono na dwa etapy. W pierwszym etapie za pomocą sprawdzania krzyżowego, dla wybranych parametrów, zweryfikowano wynik dokładności modeli predykcyjnych. Na podstawie otrzymanych wyników wybrano parametry klasyfikatorów dające najwyższe wyniki i porównano ich jakość predykcji.

Regresja logistyczna

Klasyfikator regresji logistycznej osiągnął najwyższy wynik dokładności dla parametrów zawartych w tabeli 4.2.

Tabela 5. Zestawienie trzech najlepszych wyników i parametrów klasyfikatora regresji logistycznej

Numer modelu	**1**	**2**	**3**
Kara	l2	l2	l2
Odwrotność siły regularyzacji	1,1	1,1	2
Maksymalna ilość iteracji	1000	1000	1000
Algorytm optymalizacyjny	newton-cg	lbfgs	lbfgs
Wynik dopasowania	0,94454	0,94454	0,94454

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

W przypadku tego klasyfikatora najwyższe wyniki dokładności predykcji nie różnią się od siebie.

Las losowy

Klasyfikator wykorzystujący las losowy uzyskał najwyższe wyniki dla parametrów przedstawionych w tabeli 6.

Tabela 6. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora lasu losowego

Numer modelu	**1**	**2**	**3**
Liczba drzew w iteracji	300	200	300
Kryterium podziału	Gini	Log_loss	Log_loss
Bootstrap	Tak	Tak	Tak
Wynik dopasowania	0,948799	0,948791	0,948768

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Najlepszą dokładnością cechuje się klasyfikator numer jeden.

Drzewa decyzyjne

W przypadku tego klasyfikatora najwyższy poziom dokładności otrzymał on dla parametrów zawartych w tabeli 7.

Tabela 7. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora drzewa decyzyjne

Numer modelu	**1**	**2**	**3**
Kryterium podziału	Gini	Entropia	Entropia
Liczba funkcji uwzględnianych w podziale	Pierwiastek z liczby cech podczas dopasowania	Pierwiastek z liczby cech podczas dopasowania	Pierwiastek z liczby cech podczas dopasowania
Minimalna liczba próbek węźle „liścia”	14	12	16
Sposób podziału w każdym węźle	Losowy	Losowy	Losowy
Wynik dopasowania	0,9307	0,9298	0,9308

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Najwyższą dokładność wykazuje model wykorzystujący klasyfikator numer 3.

K najbliższych sąsiadów

Ten klasyfikator swoje najwyższe wyniki dokładności otrzymał dla parametrów zestawionych w tabeli numer 8.

Tabela 8. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora K najbliższych sąsiadów

Numer modelu	**1**	**2**	**3**
Liczba sąsiadów	6	7	8
Wynik dopasowania	0,929918	0,933084	0,931909

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Najwyższy wynik dokładności uzyskał klasyfikator numer dwa.

SVC

Klasyfikator SVC najwyższe wyniki uzyskał dla parametrów przedstawionych w tabeli 9.

Tabela 9. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora SVC

Numer modelu	**1**	**2**	**3**
Typ jądra	Radialna funkcja bazowa[70]	Wielomian	Liniowy
Wynik dopasowania	0,9480	0,9477	0,9441

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Najwyższą dokładność uzyskał klasyfikator numer 1.

Ekstra drzewa decyzyjne[71]

Najwyższe wyniki dokładności przy użyciu tego klasyfikatora otrzymano dla parametrów przedstawionych w tabeli 10.

Tabela 10.. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora ekstra drzewa decyzyjne

Numer modelu	**1**	**2**	**3**
Kryterium podziału	Gini	Gini	Gini
Liczba funkcji uwzględnianych w podziale	Pierwiastek z liczby cech podczas dopasowania	Pierwiastek z liczby cech podczas dopasowania	Liczby cech widocznych podczas dopasowania

[70] Ang. Radial Basis Function
[71] Ang. Extra tree classifier

Numer modelu	1	2	3
Minimalna liczba próbek węźle „liścia”	2	6	2
Sposób podziału w każdym węźle	Losowy	Losowy	Losowy
Wynik dopasowania	0,9323	0,9315	0,9320

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Model wykorzystujący klasyfikator numer 1 uzyskał najwyższą dokładność.

Drzewa decyzyjne z gradientem

Klasyfikator drzew decyzyjnych z gradientem osiągnął najwyższe wyniki dla parametrów zawartych w tabeli 11.

Tabela 11. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora drzewa decyzyjne z gradientem

Numer modelu	1	2	3
Iteracje	1000	1000	1000
Głębokość	4	6	8
Wynik dopasowania	0,9503	0,9503	0,9504

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Najlepsze dopasowanie uzyskał klasyfikator o parametrach numer trzy.

Sieć neuronów

Klasyfikator wykorzystujące sieć neuronów osiągnął najlepsze wyniki dla parametrów przedstawionych na rysunku 12.

Tabela 12. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora sieć neuronów

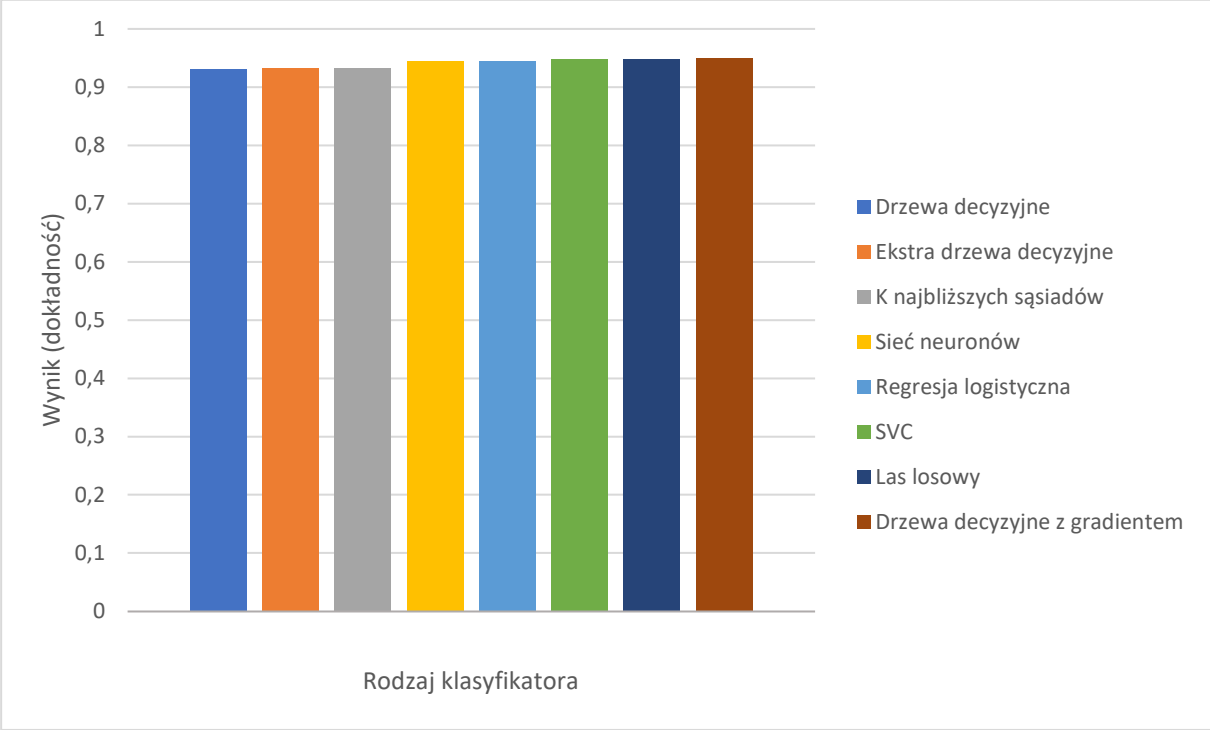
Numer modelu	1	2	3
Maksymalna liczba iteracji	500	750	1000
Funkcja aktywacji	Tożsamość	Tożsamość	Tożsamość
Wynik dopasowania	0,9444	0,9444	0,9443

[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Najwyższy wynik osiągnął klasyfikator 1. Zestawienie wyników dokładności dla najlepszych parametrów w przypadku poszczególnych klasyfikatorów przedstawiono na rysunku 24.

Wyniki dokładności poszczególnych parametrów dla zbioru uczącego mieszczą się w przedziale 0,93 do około 0,95.

Rysunek 24. Wyniki dokładności dla klasyfikatorów z najlepszymi parametrami



[Źródło: Opracowanie własne na podstawie wyników sprawdzania krzyżowego parametrów]

Najlepszym klasyfikatorem dla badanego zbioru były drzewa decyzyjne z gradientem.

3.3. Ocena klasyfikatorów

Z wykorzystaniem otrzymanych w pierwszej części najlepszych parametrów dla klasyfikatorów przeprowadzono predykcje na zbiorze testowym. Miary jakie zostały uwzględnione to: precyzja, czułość, dokładność, miara F1 oraz czas obliczeń. Wyniki przedstawiono liczbowo oraz kolorystycznie, za pomocą skali zielono-żółto-pomarańczowo-czerwonej, gdzie zielony oznacza najkorzystniejsze wyniki, a czerwony najmniej korzystne z punktu widzenia poszczególnych miar.

Tabela 13. Zestawienie wyników dla wykorzystanych klasyfikatorów

Klasyfikator	Precyzja	Czułość	Dokładność	Miara F1	Czas obliczeń [sekund]*
Regresja Logistyczna	0,753	0,732	0,945	0,742	15
Las losowy	0,777	0,741	0,949	0,759	383
Drzewa decyzyjne	0,696	0,659	0,932	0,677	32
K najbliższych sąsiadów	0,713	0,640	0,933	0,674	2547
Ekstra drzewa decyzyjne	0,692	0,580	0,926	0,631	7
SVC	0,750	0,758	0,946	0,754	2882
Drzewa decyzyjne z gradientem	0,780	0,751	0,950	0,766	35
Sieć neuronów	0,746	0,743	0,945	0,744	35
* - czas obliczeń na komputerze z procesorem Intel Core i7-8750H 2.2GHz i 8GB DDR4 RAM					

[Źródło: Opracowanie własne na podstawie wyników walidacji modeli na próbie testowej]

W tabeli 13. zestawiono miary jakie otrzymano dla poszczególnych klasyfikatorów na zbiorze testowym. Najlepsze wyniki otrzymał klasyfikator drzew decyzyjnych z gradientem, który dla precyzji, dokładności i miary F1 osiągnął najwyższe wyniki. Najbardziej czuły okazał się klasyfikator SVC. Najkrótszy czas obliczeń uzyskał klasyfikator ekstra drzew decyzyjnych jednak jego predykcje uzyskały najniższe miary. Omawiając miary jakie uzyskały poszczególne klasyfikatory należy zwrócić uwagę, że w przypadku predykcji wszystkich wartości w zbiorze testowym, jako pożary nie wywołane celowo, otrzymamy dokładność na poziomie 0,89.

Macierz pomyłek dla takiego wariantu pokazuje tabela 14.

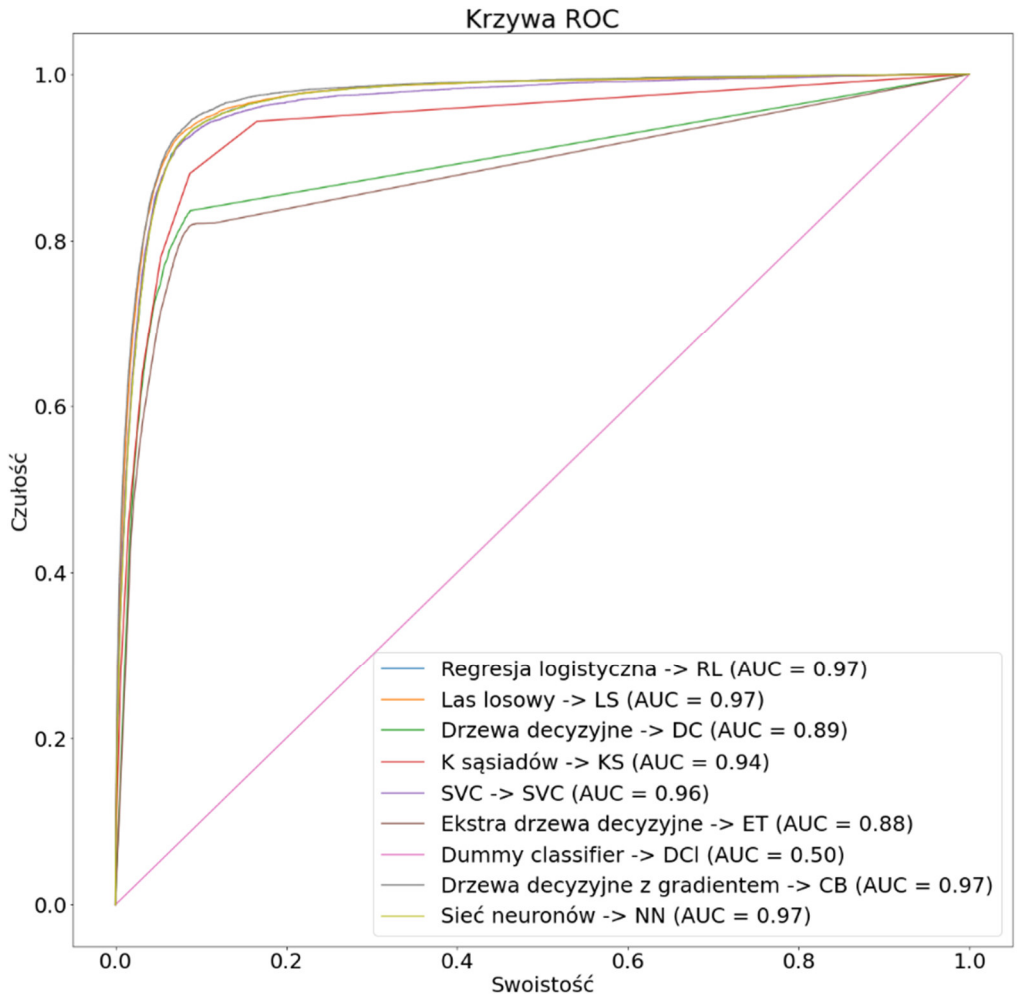
Tabela 14. Macierz pomyłek dla przypadku, w którym zakłada się, że wszystkie pożary wywołane są przez przypadek.

		Klasa predykowana	
	Populacja	Klasyfikacja pozytywna	Klasyfikacja negatywna
Klasa rzeczywista	Stan pozytywny	0,89	0,00
	Stan negatywny	0,11	0,00

[Źródło :Opracowanie własne na podstawie wyników modelu]

Krzywa ROC dla każdego klasyfikatora przedstawiona została na rysunku 25. Najwyższa wartość AUC to 0,97 i osiągają ją cztery klasyfikatory. Dla wszystkich klasyfikatorów czułość wzrasta w znacznym stopniu do 0,1 swoistości.

Rysunek 25. Krzywa ROC dla klasyfikatorów



[Źródło: Opracowanie własne na podstawie wyników walidacji modeli na próbie testowej]

Obserwując przebiegi krzywych można uznać, że klasyfikatory otrzymują lepsze wyniki niż przy losowej predykcji. Wszystkie wartości AUC są większe od 0,5, a więc modele są lepsze od losowej predykcji. Kształt krzywych można podzielić na trzy grupy. Pierwszą, o najgorszym przebiegu, jest grupa zawierająca modele oparte na klasyfikatorze drzew decyzyjnych oraz ekstra drzew decyzyjnych. Drugą grupę reprezentuje model oparty na klasyfikatorze k sąsiadów. Ostatni zbiór zawiera modele wykorzystujące resztę klasyfikatorów. Najwyższy wynik osiągnęły klasyfikatory regresji logistycznej, lasu losowego, drzew decyzyjnych z gradientem oraz sieci neuronów, który wyniósł 0,97 pola pod krzywą ROC (AUC). Biorąc pod uwagę wszystkie miary uwzględniane w tabeli 13 oraz przebieg krzywych ROC stwierdzono, że najlepszym klasyfikatorem jest klasyfikator drzew decyzyjnych z gradientem.

PODSUMOWANIE

Problem pożarów mimo rozwoju technik zapobiegania i zwalczania ognia wciąż stanowi znaczny koszt finansowy, a przede wszystkim społeczny. Dla przykładu w latach 1984-1993 dla Stanów Zjednoczonych koszt pożarów wahał się w okolicach 0,2% PKB[72]. Aby zapobiec ryzyku pożaru firmy oraz klienci indywidualni zabezpieczają dobra materialne i życie ludzkie poprzez stosowanie ochrony przeciwpożarowej. Z punktu widzenia ekonomicznego tego typu środki opłacalne są jedynie do momentu, w którym koszt ich stosowania nie przewyższy kosztów jakie można ponieść w przypadku pożaru. Dodatkowym zabezpieczeniem w takim przypadku okazują się ubezpieczenia od pożarów. W taki sposób, poprzez opłacanie składki ubezpieczeniowej, klient przenosi ryzyko na firmę ubezpieczeniową.

Przedsiębiorstwa świadczące produkty ubezpieczeniowe zarabiają dzięki odpowiedniemu szacowaniu ryzyka, tak aby składki klientów były wyższe niż potencjalne straty spowodowane wypłacaniem świadczeń. Globalizacja, zmiany technologiczne i rosnące ilości informacji wywierają dodatkową presję na firmach z tego sektora, powodując potrzebę stosowania nowych rozwiązań, aby móc konkurować o pozycję na rynku. Duże ilości danych, jakie otrzymują przedsiębiorstwa ubezpieczeniowe, nie są możliwe do analizy bez pomocy narzędzi jakie oferuje big data. Wykorzystanie tej technologii pozwala na optymalizację szacowania ryzyka przy tworzeniu ofert ubezpieczeniowych, ale również daje możliwość wyszukiwania potencjalnych oszustw i wyłudzeń.

[72] Ramachandran G. ,*The Economics of Fire Protection*, Routledge, New York 2003, s.4

W niniejszej pracy wykorzystano algorytmy uczenia maszynowego w celu budowy modelu, umożliwiającego predykcję czy dany pożar został wywołany celowo czy nie. Cel pracy jakim było utworzenie takiego modelu został osiągnięty. Dla najlepiej dopasowanych klasyfikatorów miara dokładności wynosiła 0,95. Ze względu na mały odsetek celowo wywołanych pożarów w całym zbiorze danych, ich predykcja jest trudna. Z tego powodu miara F1 nie osiąga wysokiego poziomu i wynosi dla najlepszego klasyfikatora jedynie 0,766. Klasyfikatorem tym okazały się drzewa decyzyjne z gradientem.

Na podstawie otrzymanych wyników stwierdzono, że wykorzystanie modelowania do sprawdzenia możliwości zajścia przestępstwa ma potencjał. W celu podniesienia jakości predykcji można rozszerzyć dostępne dane o informacje dotyczące ubezpieczonego, albo podzielić dostępne dane na mniejsze zbiory, uwzględniając na przykład, jedynie konkretne źródła ognia.

SPIS TABEL

Nazwa tabeli	***Strona***
Tabela 1. Przypuszczalne przyczyny pożarów stosowane podczas interwencji przez polską straż pożarną	*8*
Tabela 2. Porównanie kosztów pożaru w wybranych krajach(okres lat)	*10*
Tabela 4. Rodzaje hazardu w ubezpieczeniach	*15*
Tabela 3. Tablica pomyłek	*42*
Tabela 4. Etykiety danych wykorzystanych w pracy	*46*
Tabela 5. Zestawienie wyników i parametrów dla trzech najlepszych klasyfikatorów regresji logistycznej	*51*
Tabela 6. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora lasu losowego	*52*
Tabela 7. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora drzewa decyzyjne	*52*
Tabela 8. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora K najbliższych sąsiadów	*53*
Tabela 9. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora SVC	*53*
Tabela 10.. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora ekstra drzewa decyzyjne	*53*
Tabela 11. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora drzewa decyzyjne z gradientem	*54*
Tabela 12. Zestawienie trzech najlepszych wyników i parametrów dla klasyfikatora sieć neuronów	*54*
Tabela 13. Zestawienie wyników dla wykorzystanych klasyfikatorów	*56*

SPIS RYSUNKÓW

Nazwa rysunku	***Strona***
Rysunek 1. Bezpośrednie straty w wyniku pożarów przedstawiony jako % PKB	*12*
Rysunek 2. Ryzyko jako proces stadialny	*14*
Rysunek 3. Wartość przestępczości ubezpieczeniowej w dziale II w Polsce w latach 2014-2020	*16*
Rysunek 4. Podział sposobów uczenia maszynowego	*22*
Rysunek 5. Optymalna hiperpłaszczyzna o maksymalnym marginesie separacji	*24*
Rysunek 6. Dostępne hiperpłaszczyzny	*25*
Rysunek 7. Schemat granic decyzyjnych	*26*
Rysunek 8. Linia regresji	*27*
Rysunek 9. Linia regresji wielokrotnej	*28*
Rysunek 10. Wykres regresji krzywoliniowej	*29*
Rysunek 11. Regresja logistyczna	*31*
Rysunek 12. Przykład drzewa decyzyjnego	*32*
Rysunek 13. Model działania nienadzorowanego systemu uczącego	*35*
Rysunek 14. Matematyczna reprezentacja jednego neuronu	*36*
Rysunek 15. Najpopularniejsze funkcje aktywacji	*38*
Rysunek 16. Przykład sztucznej sieci neuronowej	*39*
Rysunek 17. Algorytm grupowania k-średnich	*40*
Rysunek 18. Uczenie maszynowe przez wzmacnianie	*41*
Rysunek 19. Przykładowa krzywa ROC dla kilku modeli	*44*
Rysunek 20. Wykres przedstawiający średnie PKB na osobę oraz liczby poszkodowanych lub ofiar pożarów między 2010, a 2020 rokiem w Wielkiej Brytanii	*48*

Rysunek 21. Suma obiektów w jakich odnotowano pożar w latach 2010-2020 w Wielkiej Brytanii	*49*
Rysunek 22. Liczba pożarów ze względu na lokalizację rozpoczęcia w latach 2020-2020 w Wielkiej Brytanii	*49*
Rysunek 23. Przyczyny pożarów w ujęciu procentowym w latach 2010-2020 w Wielkiej Brytanii	*50*
Rysunek 24. Wyniki dokładności dla klasyfikatorów z najlepszymi parametrami	*55*
Rysunek 25. Krzywa ROC dla klasyfikatorów	*57*

BIBLIOGRAFIA

1. Berrar D. ,*Theorem and Naive Bayes Classifier*, Tokyo Institute of Technology, Tokio 2018,
2. Bottou L., Lin Ch. ,*Support Vector Machine*, 2007, https://leon.bottou.org/publications/pdf/lin-2006.pdf [Data uzyskania dostępu: 5 czerwiec 2022]
3. Brownlee J. ,*Master Machine Learning Algorithms*, Jason Brownlee, 2016,
4. Brushlinsky N., Sokolov S., Wagner P. ,*Humanity and fires*, Fundacja Edukacja i Technika Ratownictwa, Warszawa 2010,
5. Buczkowski K. ,*Oszustwo asekuracyjne jako jeden z przejawów przestępczości ubezpieczeniowej*, Wolters Kluwer Polska SA, 2015,
   Cichosz P., *Systemy uczące się*, Warszawa, WNT, 2000,
6. Chee S., Peck Y., Massoud M. ,*Predictive Analytics in Business Analytics: Decision Tree, Advances in Decision Sciences*, Tajwan 2022,
7. Cycoń M. ,*Umowa ubezpieczenia, Współczesne ubezpieczenia gospodarcze*, red. W. Sułkowska, Uniwersytet Ekonomiczny w Krakowie, Kraków 2013,
8. Ezukwoke K., Zareian S. ,*Logistic regression and kernel logistic regression A comparative study of logistic regression and kernel logistic regression for binary classification*, University Jean Monnet, France 2019,
9. Filipowska M. ,*Rynek ubezpieczeniowy wciąż zagrożony przestępczością*," Wroclaw Economic Review, Wroclaw Economic Review, Wrocław 2008,
10. Grzyb M. ,*Uczenie maszynowe - Sekrety doboru odpowiedniego algorytmu*, https://mateuszgrzyb.pl/wp-content/uploads/2020/05/uczenie_maszynowe__sekrety_doboru_odpowiedniego_algorytmu.pdf. [Data uzyskania dostępu: 10 luty 2022]
11. Gülden K., Neşe G. ,*A study on multiple linear regression analysis*, Elsevier Ltd, Sakarya 2013,
12. Gurney K. ,*An introduction to neural networks*, Taylor & Francis e-Library, Londyn 2004,

13. Guzewski P., Wróblewski D., Małozięć D. ,*Czerwona księga pożarów*, Narodowe Centrum Badań i Rozwoju, Józefów 2014,
14. Haykin S. ,*Neural Networks and Learning Machines, Pearson Prentice Hall*, New Jersey 2009,
15. Helvestone M. ,*Consumer Protection in the Age of Big Data*, Washington University Law Review, Washington, 2016,
16. Korff D., Georges M. ,*Podręcznik Inspektora Ochrony Danych*, Garante per la protezione dei dati personal, 2019,
17. Louppe G. ,*Understanding Random Forest*, University of Liège, Liège 2014,
18. Łańcucki J. ,*Klient na cyfrowym rynku ubezpieczeniowym*, Prawo asekuracyjne, Poznań 2018,
19. Majewski P. ,*Analiza Danych Dotyczących Przestępstw Ubezpieczeniowych w 2020 roku*, Polska Izba Ubezpieczeń, Warszawa 2021,
20. Michalski T., Pajewska R. ,*Ubezpieczenia gospodarcze*, Wydawnictwa Szkolne i Pedagogiczne S.A., Warszawa 2001,
21. Mrówczyńska M. ,*Klasyfikatory neuronowe typu SVM w zastosowaniu do klasyfikacji przemieszczeń pionowych na obszarze LGOM*, Zeszyty Naukowe Instytutu Gospodarki Surowcami Mineralnymi i Energi Polskiej Akademii Nauk, 2014,
22. Pecio M. ,*Nadzór nad rynkiem ubezpieczeniowym na przykładzie Polski i Stanów Zjednoczonych – model działania*, Wroclaw Economic Review, Wrocław 2009,
23. Pecio M. ,*Ryzyko ubezpieczeniowe a ryzyko pożarowe*, Przestrzeń, Ekonomia, Społeczeństwo, 2015,
24. Pecio M. ,*Przeciwdziałanie przestępczości*, Teoria i Praktyka 2016 nr 4, 2016,
25. Płonka M., Oręziak B., Wielec M. ,*Rynek Ubezpieczeniowy - Zapobieganie przyczynom przestępczości*, Wydawnictwo Instytutu Wymiaru Sprawiedliwości, Warszawa, 2021,
26. Połeć R., Lemańczyk M. ,*Przeciwdziałanie przestępczości ubezpieczeniowej w aspekcie jej społecznej akceptacji*, Studia oeconomica posnaniensia 2015,
27. Ramachandran G. ,*Fire Losses in Different Countries*, , Borehamwood 1970,
28. Ramachandran G. ,*The Economics of Fire Protection*, Routledge, New York 2003,

29. Rządca K. ,*Algorytmy grupowania danych*, 2004, https://www.mimuw.edu.pl/~krzadca/magister.pdf. [Data uzyskania dostępu: 1 czerwiec 2022]
30. Sawka K. ,*Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow*, Helion, Gliwice 2020,
31. Shwarts S., Ben D. ,*Understanding Machine Learning*, Cambridge University Press, New York 2014,
32. Soroka-Potrzebna H. ,*Ubezpieczenie jako metoda zarządzania ryzykiem w opinii eksportujących mikroprzedsiębiorstw*, Zeszyty naukowe Uniwersytetu Szczecińskiego nr 854, Szczecin, 2015,
33. Strupczewski G. ,*Ryzyko i zarządzanie ryzykiem*, Kraków 2013,
34. Sutton R., Bart A. ,*Reinforcement Learning: An Introduction*, The MIT Press, Londyn 2015,
35. Wu J. ,*Advances in K-means Clustering*, Springer-Verlag, Berlin 2012,

Źródła internetowe

1. IBM ,*Supervised Learning*, https://www.ibm.com/cloud/learn/supervised-learning. [Data uzyskania dostępu: 2 czerwiec 2022]
2. Kajstura A. ,*Metoda k-średnich*, https://www.statystyka.az.pl/analiza-skupien/metoda-k-srednich.php. [Data uzyskania dostępu: 3 marzec 2022]
3. Stefanowski J. ,*SVM - Support Vector Machines Metoda wektorów nośnych*, Institute of Computing Sciences, Poznań University of Technology, http://www.cs.put.poznan.pl/jstefanowski/ml/SVM.pdf [Data uzyskania dostępu: 10 luty 2022]
4. *Czym jest uczenie maszynowe?*, https://www.sap.com/poland/insights/what-is-machine-learning.html. [Data uzyskania dostępu: 5 luty 2022]
5. *Sztuczne sieci neuronowe*, https://www.sztucznainteligencja.org.pl/kurs/sztuczna-inteligencja-dla-poczatkujacych/sztuczne-sieci-neuronowe/. [Data uzyskania dostępu: 02 marzec 2022]

6. *Podział sposobów uczenia maszynowego*, https://www.pngitem.com/middle/hRJJRJJ_types-of-machine-learning-machine-learning-types-of/ [Data uzyskania dostępu: 15 czerwiec 2022]
7. *Uczenie maszynowe nienadzorowane*, https://neptune.ai/blog/life-cycle-of-a-machine-learning-project [Data uzyskania dostępu: 15 czerwiec 2022]
8. *Matematyczna reprezentacja jednego neuronu*, https://towardsdatascience.com/power-of-a-single-neuron-perceptron-c418ba445095 [Data uzyskania dostępu: 24 czerwiec 2022]
9. *Najpopularniejsze funkcje aktywacji*, https://cdn-images-1.medium.com/max/1000/1*nFmfF-SAMZc9_jtFNullRg.png [Data uzyskania dostępu: 24 lipiec 2022]
10. *Algorytm k-średnich* , http://itcraftsman.pl/algorytm-k-srednich-uczenie-nienadzorowane/ [Data uzyskania dostępu: 24 maj 2022]
11. *Uczenie maszynowe przez wzmacnianie*, https://techvidvan.com/tutorials/reinforcement-learning/ [Data uzyskania dostępu: 13 maj 2022]
12. https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/ [Data uzyskania dostępu: 13 maj 2022]
13. https://aigeekprogrammer.com/pl/klasyfikacja-binarna-regresja-logistyczna-keras/ [Data uzyskania dostępu: 13 maj 2022]

ZAŁĄCZNIKI

Załącznik nr 1: Przykładowa baza danych

	ZAG_LUDNOŚCI	GDP_NA_OS	DNI POWSZEDNIE_WEEKEND	PORA_DNIA	SPECJALNA_BUDOWA_OBIEKTU	ILOŚĆ_MIESZKAŃCÓW_OBIEKTU	CZY_BUDYNEK_JEST_NORMALNIE_ZAJĘTY
185609	709	32358	1	Popołudnie	0	3 lub więcej osób dorosłych poniżej wieku emer...	1
56081	681	29354	1	Popołudnie	0	Samotna osoba w wieku emerytalnym	1
142215	148	24593	1	Wieczór	0	3 lub więcej osób dorosłych poniżej wieku emer...	1
302354	5700	55092	1	Popołudnie	0	Samotna osoba w wieku emerytalnym	1
143385	316	19367	1	Noc	0	Inne	1
...	...	...	...	...	...	...	...
304216	2222	29057	1	Wieczór	0	Samotna osoba poniżej wieku emerytalnego	1
51738	5218	47812	1	Poranek	0	Samotna osoba w wieku emerytalnym	1
334415	391	25766	1	Popołudnie	1	Para z dzieckiem/dziećmi	1
323142	163	24431	0	Popołudnie	0	3 lub więcej osób dorosłych z dzieckiem/dziećmi	1
198448	421	30380	1	Popołudnie	0	Samotna osoba poniżej wieku emerytalnego	1