Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską

Politechnika Świętokrzyska
Wydział Zarządzania i Modelowania Komputerowego

Kierunek studiów:

Inżynieria danych

Marzena Nowakowska

Materiały dydaktyczne do przedmiotu

ODKRYWANIE ZWIĄZKÓW W DANYCH WIELOWYMIAROWYCH

opracowane w ramach realizacji Projektu
„Dostosowanie kształcenia w Politechnice Świętokrzyskiej do potrzeb współczesnej gospodarki”
FERS.01.05-IP.08-0234/23

Kielce, 2025

Spis treści

Materiały dydaktyczne objęte licencją Creative Commons BY 4.0.

Licencja dostępna pod adresem: https://creativecommons.org/licenses/by/4.0/

1. Wprowadzenie

Współczesna analiza danych to nie tylko zbieranie informacji, ale przede wszystkim wydobywanie ukrytej wiedzy z danych wielowymiarowych, w których każdy obiekt opisany jest wieloma cechami lub zmiennymi. W takich danych mogą kryć się zależności trudne do zauważenia tradycyjnymi metodami, a ich odkrycie pozwala lepiej zrozumieć badane zjawiska, prognozować przyszłe zdarzenia i podejmować świadome decyzje.

W analizie danych wielowymiarowych można stosować wiele różnych narzędzi i technik uczenia maszynowego, takich jak:

- metody klasteryzacji do grupowania obiektów o podobnych cechach;
- redukcja wymiarowości w celu wizualizacji i uproszczenia złożonych danych;
- sieci neuronowe i głębokie uczenie do wykrywania nieliniowych zależności i prognozowania;
- regresje i uogólnione modele liniowe dla przewidywania wartości liczbowych lub szacowania prawdopodobieństw wystąpienia zjawiska;
- modele hybrydowe do analizy złożonych i hierarchicznych zależności między zmiennymi.

Dzięki poznaniu metod eksploracji danych i modelowania zjawisk opisanych przez te dane, studenci uczą się nie tylko identyfikować wzorce i zależności w danych, ale także dobierać odpowiednie metody do charakteru problemu, oceniać ich skuteczność i ograniczenia, co stanowi fundament zarówno dla badań naukowych, jak i praktycznych zastosowań analityki danych.

Materiały dydaktyczne, które są przedstawione w niniejszym opracowaniu obejmują wybrane studia przypadków i obejmują następujące zagadnienia:

- analiza asocjacji – wykrywanie powtarzających się wzorców i reguł zależności między zmiennymi;
- analiza sekwencji – badanie kolejności zdarzeń;
- drzewa decyzyjne i lasy losowe – metody klasyfikacji.

Analiza sekwencji została przygotowana z wykorzystaniem sztucznej inteligencji. Prezentacja pozostałych metod bazuje na wynikach własnych badań lub projektów dyplomowych realizowanych pod opieką autorki.

Autorka zachęca do aktywnego studiowania tych materiałów, eksperymentowania z różnymi metodami i narzędziami oraz samodzielnego odkrywania zależności w danych wielowymiarowych. Poprzez praktyczne ćwiczenia i analizę realnych lub syntetycznych przypadków można najlepiej zrozumieć moc i ograniczenia każdej

techniki, a także rozwijać umiejętności analityczne niezbędne w nowoczesnej nauce i pracy zawodowej w obszarze danych.

2. Analiza asocjacji

Zadaniem analizy asocjacji (*association analysis*) jest identyfikacja wspólnych wystąpień różnych wartości cech w jednym rekordzie zbioru danych. Taka technika badawcza nosi również nazwę analizy koszykowej (*market basket analysis*) i jest często stosowana w badaniach preferencji zakupowych klientów supermarketów. Odkrywanie zależności asocjacyjnych jest odkrywaniem zależności dotyczących istotnego współwystępowania pewnych określonych wartości atrybutów. Reguły asocjacyjne reprezentują w najogólniejszym pojęciu wiedzę o tym, że pewne wartości niektórych atrybutów występują łącznie z pewnymi wartościami innych atrybutów.

Analiza asocjacji jest stosowana nie tylko w kontekście organizacji towarów i przedstawienia oferty towarowej w marketach. Wykorzystuje się ją również do różnych badań związanych lub niezwiązanych z aktywnością człowieka. Poniżej zaprezentowano zagadnienie identyfikacji wzorców wypadków drogowych i przedstawienie tych wzorców w postaci reguł asocjacyjnych. Przedstawiono je na podstawie wyników zrealizowanych własnych badań naukowych autorki obejmujących analizy bezpieczeństwa ruchu drogowego (brd). W tym przypadku, wypadki drogowe zostały poddane analizom, takim jak transakcje w supermarkecie. Do przeprowadzenia badań wykorzystano moduł Enterprise Miner™ komercyjnego systemu analitycznego SAS®.

Dane pochodzą z powypadkowych raportów policyjnych. Zawierają one opis zdarzeń drogowych w postaci cech (zmiennych), takich jak: czas i miejsce zdarzenia, otoczenie drogi, warunki atmosferyczne, zachowanie kierującego, stan techniczny pojazdu, status wypadku. Większość cech wypadku drogowego ma charakter jakościowy (opisowy), dzięki czemu jest możliwe wykrycie współzależności między wartościami tych cech stosując analizę asocjacji. W tym kontekście wypadek to koszyk, a jego charakterystyki (wartości cech) definiują zawartość koszyka.

2.1. Dane o wypadkach drogowych do analizy koszykowej

Analizom poddano wybrany wycinek zbioru danych o zdarzeniach drogowych obejmujący wypadki z pieszymi z terenu województwa świętokrzyskiego, z pięciu lat początku XXI wieku, zarejestrowane na zamiejskich drogach jednojezdniowych, dwukierunkowych. Zbadano podzbiór danych o wypadkach z pieszymi spowodowane

przez kierujących pojazdami. W tabeli 1 zestawiono zmienne uwzględnione w analizie; zarówno cechy jak i ich wartości określono na podstawie raportów policyjnych. Wartości, których częstość występowania jest zbyt mała, aby wyodrębnić je w analizie zostały zagregowane. W szczególności:

- *Różne zachowania kierującego* jest *kategorią* powstałą w wyniku agregacji zachowań: *Gwałtowne hamowanie*, *Jazda bez wymaganego oświetlenia*, *Nieprzestrzeganie innych znaków i sygnałów drogowych*, *Nieprawidłowe skręcanie*, *Niezachowanie bezpiecznej odległości między pojazdami*, *Nieprawidłowe wymijanie* oraz *Nieudzielenie pierwszeństwa przejazdu*.

Zastosowano kodowanie nazw cech oraz ich wartości, przygotowując w ten sposób dane do struktury akceptowanej przez analizę koszykową.

Tabela 1. Wykaz cech wypadków drogowych z pieszymi w analizie koszykowej

Opis cechy	**Kod cechy**	**Wartości cechy**	**Kod wartości**
Cechy środowiska			
Stan nawierzchni	StNw	Sucha	Sch
		Nie sucha: mokra, ośnieżona lub oblodzona	NSch
		Inna	Inn
Pogoda	Pgd	Dobra, w tym oślepiające słońce	Dbr
		Zła: chmury, deszcz, opady śniegu, mgła/dym	Zl
		Inna	Inn
Oświetlenie	Osw	Światło dzienne	SwDz
		Brak oświetlenia (noc)	BrOs
		Słabe (niepełne) światło	SlOs
		Brak danych	BD
Pora roku	PrRk	Zima: Grudzień - Luty	Zm
		Wiosna: Marzec - Maj	Wsn
		Lato: Czerwiec - Sierpień	Lt
		Jesień: Wrzesień - Listopad	Jsn
Pora dnia	PrDn	Noc: [24:00, 6:00)	Nc
		Ranek: [6:00, 12:00)	Rnk
		Popołudnie: [12:00, 18:00)	Ppl
		Wieczór: [18:00, 24:00)	Wcz
Dzień tygodnia	DzTg	Dzień pracy	Prc
		Dzień wolny od pracy	Wln
Czynniki drogi			
Kategoria drogi	KtDr	Krajowa	Krj
		Wojewódzka	Wjw
		Powiatowa	Pwt
		Gminna	Gmn
		Brak danych	BD
Rodzaj obszaru	RdOb	Zabudowany	Zbd
		Niezabudowany	NZbd
Odcinek drogi	OdDr	Prosty odcinek drogi	Prs

Opis cechy	Kod cechy	Wartości cechy	Kod wartości
		Łuk poziomy	LkPz
		Pochylenie niwelety lub łuk pionowy	PchLkPn
		Obszar skrzyżowania	ObSk
Czynnik ludzki			
Grupa wiekowa kierującego	GrWkKr	[7, 15)*	02
		[15, 18)	03
		[18, 25)	04
		[25, 40)	05
		[40, 60)	06
		>= 60	07
Kierujący pod wpływem alkoholu lub innych środków odurzających	AlKr	Tak	T
		Nie	N
Płeć kierującego	PlKr	Kobieta	K
		Mężczyzna	M
		Brak danych	BD
Zachowanie kierującego	ZchKr	Jazda po niewłaściwej stronie drogi	JzNwSDr
		Nadmierna prędkość	NdPr
		Nieprawidłowe cofanie	NpCf
		Nieprawidłowe omijanie	NpOm
		Nieprawidłowe przejeżdżanie przejść dla pieszych	NpPPsz
		Nieprawidłowe wyprzedzanie	NpWp
		Zmęczenie, zaśnięcie lub ograniczenie sprawności psychomotorycznej	NspKr
		Różne zachowania (agregacja	RZchKr
		Inne zachowania (niezdefinowane)	InZchKr
		Współwina kierujących	Wsp
Cecha wypadku			
Status wypadku	StWp	Lekki	Lk
		Ciężki	Ck
		Śmiertelny	Sm

* *Może odnosić się do dziecka lub młodej osoby jadącej na rowerze lub motorynce.*

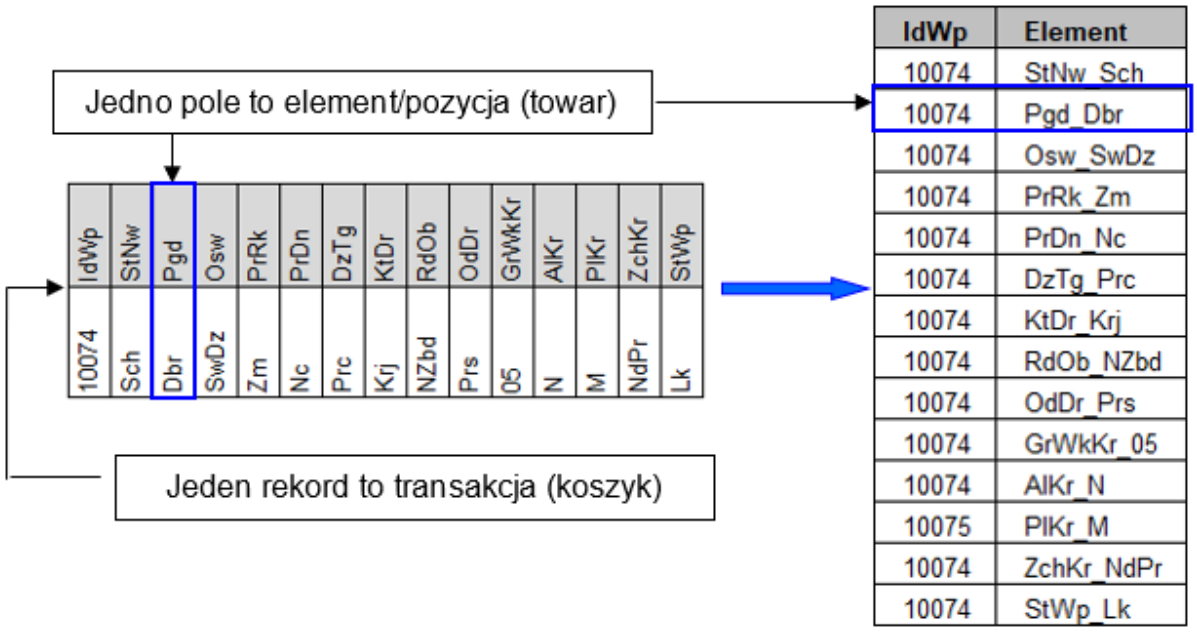
Skutki wypadku drogowego są zdefiniowane za pomocą zmiennej *Status wypadku* (*StWp*) przez następujące wartości informujące o ciężkości wypadku (wg Zarządzenia nr 635 Komendanta Głównego Policji z dnia 30 czerwca 2006 r. w sprawie metod i form prowadzenia przez Policję statystyki zdarzeń drogowych. Warszawa, 2006):

- *Lk*; wypadek lekki, gdy nie ma ofiar śmiertelnych ani ciężko rannych i przynajmniej jedna osoba jest lekko ranna, tzn. poniosła uszczerbek na zdrowiu naruszający czynności ciała lub rozstrój zdrowia na okres trwający nie dłużej niż 7 dni, zgodnie z orzeczeniem lekarza,

- *Ck*; wypadek ciężki, gdy nie ma ofiar śmiertelnych i przynajmniej jedna osoba jest ciężko ranna, tzn. doznała kalectwa, ciężkiej, nieuleczalnej lub długotrwałej choroby, w tym choroby psychicznej, całkowitej lub trwałej niezdolności do pracy, itp.,
- *Sm*; wypadek śmiertelny, gdy co najmniej jedna z ofiar została zabita na miejscu lub zmarła w ciągu 30 dni wskutek wypadku.

Ponieważ struktura danych do analizy asocjacji jest zdefiniowana przez dwa pola, zbiór danych o zdarzeniach drogowych jest przekształcony do takiej formy. Jedno pole to identyfikator wypadku drogowego (identyfikator koszyka), drugie jest zakodowaną wartością elementu charakteryzującego zdarzenie (odpowiednik towaru w koszyku) i ma postać: *Zmienna_WartośćZmiennej*. To drugie pole przechowuje więc informację dwojakiego rodzaju: o tym, która cecha opisuje zdarzenie (*Zmienna*) i o tym, jaka jest kategoria tej zmiennej (*WartośćZmiennej*). Ideę omawianej zmiany struktury danych przedstawiono na rys. 1.

[Tekst alternatywny. Dwa zestawienia danych przedstawione w układzie tabelarycznym. Po lewej stronie widoczny jest rekord dotyczący zdarzenia drogowego z piętnastoma polami opisującymi jego cechy. Po prawej stronie pokazano ten sam rekord w przekształconej formie – zawierającej dwa nagłówki: identyfikator zdarzenia oraz element opisujący zdarzenie poprzez połączenie cechy z jej wartością. Strzałka między tabelami wskazuje kierunek przekształcenia – od tabeli po lewej do tabeli po prawej.]



Rys. 1. Dostosowanie struktury danych o wypadkach drogowych do analizy asocjacji

Generalnie, analizy asocjacyjne nie uwzględniają następstwa czasowego badanych elementów, czyli określonego reżimu kolejności ich wystąpienia. W przypadku zdarzeń drogowych kolejność taka jest istotna – istnieje logiczna sekwencja tworzenia się informacji o wartościach cech opisujących wypadek. Przykładowo, znane są wartości takich zmiennych jak: kategoria drogi, obszar i odcinek drogi, warunki atmosferyczne, oświetlenie, pora roku, zanim nieprawidłowe zachowanie uczestnika ruchu drogowego spowoduje wypadek o określonym stopniu ciężkości. Stąd kolejność pojawienia się czynników generujących wypadek wpływa na konieczność takiego wyboru reguł asocjacyjnych, aby nie była zakłócona logika zależności.

2.2. Reguły analizy koszykowej w badanu wypadków drogowych

Klasyczna reguła asocjacyjna postaci A → B mówi, że *Jeżeli zbiór wartości A jest częścią jakiegoś zdarzenia, to zbiór wartości B też jest jego częścią*. *A* jest elementem warunkującym (poprzednikiem), a *B* elementem warunkowanym (następnikiem) reguły.

Wypadki drogowe są specyficznymi transakcjami różniącymi się od tych, występujących w klasycznej analizie asocjacji. Po pierwsze, każda transakcja (wypadek) ma taka samą liczbę elementów (cech wypadku drogowego), podczas gdy liczba rodzajów towarów w koszyku może być różna dla różnych klientów (koszyków). Po drugie, każdy element reguły składa się z dwóch części: jeden informuje o cesze wypadkowej a drugi o jej wartości, podczas gdy w analizie koszykowej element reguły jest po prostu nazwą zakupionego towaru. W prezentowanej analizie cecha wypadkowa może być dowolnym z czynników środowiska, drogi lub sprawcy wypadku. Razem ze swoją wartością może ona wystąpić tylko raz w poprzedniku reguły. Dwuelementowa reguła, rozważana w kontekście badań bezpieczeństwa ruchu drogowego (brd) ma postać: $A\_i \rightarrow B\_j$. Informuje ona o tym, że pewna okoliczność A_i, która charakteryzuje wypadek drogowy współwystępuje z zagrożeniem opisanym przez wartość *j* cechy *B*. Przykładowo, reguła *PrDn_Wcz* → *StWp_Sm* stwierdza, że jeżeli wypadek zdarza się w nocy to zazwyczaj jest śmiertelny. W analizach brd zagrożenie *B* może być wyrażone przez zachowanie użytkownika drogi, które powoduje wypadek lub status wypadku.

W miarach oceny reguł uwzględniona jest struktura danych o wypadkach drogowych opracowanych do analizy koszykowej.

- Wsparcie reguły *Support* wskazuje odsetek zdarzeń, dla których rozważane cechy *A* i *B* mają wartości odpowiednio *i* oraz *j*:

$$Support(A\_i \rightarrow B\_j) = \frac{N_{A\_i \wedge B\_j}}{N} \approx P(A\_i \wedge B\_j) \tag{1}$$

- Wiarygodność *Confidence* jest częstością względną, która informuje o tym, jaka część zdarzeń scharakteryzowanych przez okoliczność *A_i* jest opisana przez zagrożenie *B_j*:

$$\begin{aligned} Confidence(A\_i \rightarrow B\_j) &= \frac{Support(A\_i \rightarrow B\_j)}{Support(A\_i)} = \frac{N_{A\_i \wedge B\_j} / N}{N_{A\_i} / N} \\ &= \frac{N_{A\_i \wedge B\_j}}{N_{A\_i}} \approx P(B\_j / A\_i) \end{aligned} \tag{2}$$

- *Lift* informuje o sile związku między elementami *A_i* i *B_j*. Wartość większa od jedności wskazuje na asocjację pozytywną:

$$\begin{aligned} Lift(A\_i \rightarrow B\_j) &= \frac{Suport(A\_i \rightarrow B\_j)}{Support(A\_i) \cdot Support(B\_j)} \\ &= \frac{Confidence(A\_i \rightarrow B\_j)}{Support(B\_j)} \approx \frac{P(B\_j / A\_i)}{P(B\_j)} \end{aligned} \tag{3}$$

- *WWP* informuje o sile wpływu na zagrożenie *B_j* wartości *i* okoliczności *A* w porównaniu z inną wartością.

$$WWP(A\_i \rightarrow B\_j) = \frac{Confidence(A\_i \rightarrow B\_j)}{Confidence(A\_\neg i \rightarrow B\_j)} \approx \frac{P(B\_j / A\_i)}{P(B\_j / A\_\neg i)} \tag{4}$$

Im większe wsparcie i wiarygodność oraz im silniej większe od jedynki *Lift* i *WWP* tym reguła mocniejsza.

2.3. Identyfikacja związków między cechami wypadków drogowych

Eksperymenty badawcze zaplanowano, tak aby odpowiedzieć na pytanie: „Wśród analizowanych cech wypadku drogowego (włączając zachowanie sprawcy), które i

jak znacząco towarzyszą określonej ciężkości wypadków drogowych z pieszymi na drogach zamiejskich?". Zagrożenie *B* jest wyrażone w regułach asocjacyjnych za pomocą cechy *Status wypadku – StWp*.
Wygenerowano reguły dwuelementowe, tzn. reguły z jednym poprzednikiem i jednym następnikiem. Następnie rozbudowano poszukiwania w celu wykrycia asocjacji z dwoma elementami w poprzedniku. Ponieważ liczba generowanych reguł jest bardzo duża, zastosowano pewne filtry, aby wyodrębnić najważniejsze związki:

- rozważano wszystkie reguły dwuelementowe *element1* → *element 2* spełniające kryteria: *Support > 10%* i *WWP > 1*,
- reguły trzyelementowe *(element1, element 2)* → *element 3*, które spełniają kryterium *Lift > 1.05* zostały uporządkowane malejąco wg wartości miary *Support* i do dyskusji zestawiono 10 pierwszych z tak utworzonej listy.

Reguły asocjacyjne zawarte w tablicach prezentowanych dalej są pogrupowane według następnika reguły (tzn. według statusu wypadku) a następnie uporządkowane w obrębie każdej grupy rosnąco wg rozmiaru reguły (liczby elementów) i malejąco wg wsparcia. Zestawienie zawarto w tablicy 2.

Tablica 2. Reguły asocjacyjne dla wypadków drogowych z udziałem pieszych

Numer	Rozmiar	Element1	Element2	Liczebność	Support [%]	Conf [%]	Lift	WWP
Następnik reguły: StWp_Lk								
1	2	Pgd_Dbr		264	30,14	44,98	1,02	1,04
2	2	StNw_Sch		261	29,8	45,24	1,02	1,04
3	2	Osw_SwDz		191	21,8	52,04	1,16	1,34
4	2	KtDr_Pwt		177	20,2	47,96	1,08	1,14
5	2	DzTg_Wln		166	18,94	46,24	1,04	1,06
6	2	PrDn_Ppl		123	14,04	49,6	1,12	1,16
7	2	PrRk_Lt		109	12,44	45,22	1,02	1,02
8	2	GrWkKr_BD		103	11,76	48,58	1,10	1,12
9	2	Osw_SlOs		90	10,28	48,38	1,08	1,12
10	2	ZchKr_NpOm		88	10,04	54,32	1,22	1,28
11	3	OdDr_Prs	KtDr_Pwt	152	17,36	46,92	1,06	1,08
12	3	Osw_SwDz	OdDr_Prs	151	17,24	53,16	1,20	1,32
13	3	Osw_SwDz	RdOb_Zbd	150	17,12	50,5	1,14	1,22
14	3	RdOb_Zbd	KtDr_Pwt	148	16,9	47,74	1,08	1,12
15	3	Pgd_Dbr	Osw_SwDz	142	16,22	53,38	1,20	1,32
16	3	StNw_Sch	Osw_SwDz	136	15,52	53,34	1,20	1,30
17	3	Pgd_Dbr	KtDr_Pwt	133	15,18	49,08	1,10	1,16
18	3	StNw_Sch	KtDr_Pwt	129	14,72	48,86	1,10	1,14
19	3	RdOb_Zbd	DzTg_Wln	128	14,62	47,4	1,06	1,10
20	3	Pgd_Dbr	DzTg_Wln	116	13,24	48,34	1,08	1,12
Następnik reguły: StWp_Ck								
21	2	RdOb_Zbd		248	28,32	36,96	1,04	1,24
22	2	DzTg_Prc		185	21,12	35,78	1,02	1,04

Numer	Rozmiar	Element1	Element2	Liczebność	Support [%]	Conf [%]	Lift	WWP
23	2	Osw_SwDz		137	15,64	37,32	1,06	1,10
24	2	KtDr_Pwt		132	15,06	35,78	1,02	1,02
25	2	Osw_BrOs		113	12,9	35,88	1,02	1,02
26	2	StNw_NSch		105	11,98	35,96	1,02	1,02
27	2	Pgd_Zl		104	11,88	36,62	1,04	1,06
28	2	GrWkKr_05		90	10,28	37,34	1,06	1,08
29	2	PrRk_Lt		90	10,28	37,34	1,06	1,08
30	3	StNw_Sch	RdOb_Zbd	174	19,86	38,00	1,08	1,18
31	3	Osw_SwDz	RdOb_Zbd	118	13,48	39,74	1,12	1,20
32	3	Osw_SwDz	DzTg_Prc	92	10,5	39,82	1,12	1,18
33	3	Osw_BrOs	RdOb_Zbd	84	9,58	37,84	1,08	1,1
34	3	Pgd_Zl	StNw_NSch	79	9,02	37,44	1,06	1,08
35	3	RdOb_Zbd	GrWkKr_05	78	8,9	40,00	1,14	1,18
36	3	PrRk_Lt	OdDr_Prs	76	8,68	37,44	1,06	1,08
37	3	PrRk _Lt	RdOb_Zbd	76	8,68	42,94	1,22	1,28
38	3	OdDr_Prs	GrWkKr_05	73	8,34	38,82	1,10	1,14
39	3	RdOb_Zbd	KtDr_Wjw	64	7,3	39,76	1,12	1,16
Następnik reguły: StWp_Sm								
40	2	OdDr_Prs		152	17,36	21,28	1,06	1,38
41	2	Pgd_Dbr		119	13,58	20,28	1,00	1,02
42	2	DzTg_Prc		108	12,32	20,88	1,04	1,08
43	2	PrDn_Wcz		102	11,64	26,50	1,32	1,74
44	2	Osw_BrOs		96	10,96	30,48	1,50	2,12
45	3	OdDr_Prs	DzTg_Prc	94	10,74	22,32	1,10	1,22
46	3	OdDr_Prs	PrDn_Wcz	88	10,04	26,92	1,34	1,66
47	3	Osw_BrOs	OdDr_Prs	86	9,82	31,98	1,58	2,14
48	3	Pgd_Dbr	PrDn_Wcz	73	8,34	27,34	1,36	1,60
49	3	RdOb_Zbd	PrDn_Wcz	70	8	24,82	1,22	1,38
50	3	ZchKr_InZchKr	OdDr_Prs	68	7,76	29,70	1,46	1,76
51	3	StNw_Sch	PrDn_Wcz	68	7,76	25,46	1,26	1,42
52	3	StNw_Sch	Osw_BrOs	65	7,42	31,56	1,56	1,88
53	3	Pgd_Dbr	Osw_BrOs	64	7,3	31,84	1,58	1,90
54	3	Osw_BrOs	PrDn_Wcz	62	7,08	29,80	1,48	1,74

Wypadki lekkie

Dobre warunki środowiska zdefiniowane przez elementy: *Dobra pogoda*, *Sucha nawierzchnia jezdni*, *Światło dzienne*, *Dzień wolny od pracy*, *Popołudnie* oraz *Lato* są okolicznościami, w których nieprawidłowe zachowania kierujących powodują wypadki drogowe z lekkimi obrażeniami ciała ofiar. Reguły 1, 2, 3, 5, 6 i 7 definiujące ww. związki występują stosunkowo często: *Support* $\in$ (12%, 30%) i *Confidence* > 44%, ale mają zróżnicowane wartości siły asocjacji: *Lift* zmienia się od 1,02 do 1,16. Element *Słabe oświetlenie* również tworzy pozytywny związek asocjacyjny z wypadkiem lekkim, ale odpowiednia reguła 9 jest słabsza niż reguła 3 posiadająca *Światło dzienne* w poprzedniku (por. *Lift* i *WWP* w tabeli 2). Spośród czynników niezwiązanych ze środowiskiem, dwa definiują relacje z elementem *StWp_Lk*.

Jednym jest *Kategoria powiatowa drogi* w stosunkowo istotnej regule 4. Drugim jest zachowanie kierującego zdefiniowane jako *Nieprawidłowe omijanie*, które definiuje regułę 10 najsilniejszą w całej grupie wypadków lekkich: *Confidence* > 54%, *Lift* =1.22 i *WWP* = 1,28. Wspominane wcześniej dobre warunki środowiska występują w regułach trzyelementowych połączone między sobą i w kombinacji z elementami: *Prosty odcinek* drogi oraz *Obszar zabudowany*. Miary *Confidence* i *Lift* mają wartości podobne dla reguł dwu- i trzyelementowych, *WWP* jest większy dla reguł trzyelementowych.

Wypadki ciężkie

Wszystkie dwuelementowe reguły w grupie wypadków ciężkich spowodowanych przez kierującego są raczej słabe: *Lift* nie przekracza 1,06. Mają prawie takie same wartości miary *Confidence*, mieszczące się w przedziale od 35,8% do 37,3%. Reguła 21 przyciąga uwagę ze względu na największą wartość wskaźnika *WWP*. Informuje, że wypadki z pieszymi spowodowane przez kierujących mają status ciężkich o 25% częściej na obszarach zabudowanych niż niezabudowanych. Poprzedniki reguł dla wypadków ciężkich różnią się od poprzedników reguł dla wypadków lekkich w obrębie czynników środowiska następującymi wartościami: *Dzień tygodnia* (*Dzień pracy* vs. *Dzień wolny od pracy*), *Stan nawierzchni* (*Mokra* vs. *Sucha*), *Pogoda* (*Zła* vs. *Dobra*). Można zauważyć, że wśród wszystkich reguł trzyelementowych w analizowanej grupie:

- reguła 35 *(RdOb_Zbd, GrWkKr_05)* → *StWp_Ck* informuje, że 40% wypadków z pieszymi powodowanych na obszarach zabudowanych przez kierujących pojazdami w wieku 25-40 lat skutkuje ciężkimi obrażeniami ciała ofiar; wypadki ciężkie są warunkowane wspomnianymi okolicznościami o 18% częściej niż innymi kombinacjami wartości cech: *Grupa wiekowa kierującego* i *Rodzaj obszaru*,
- reguła 37 *(PrRk_Lt, RdOb_Zbd)* → *StWp_Ck* informuje, że 43% wypadków na obszarach zabudowanych w okresie letnim ma status ciężkich; wypadki takie są warunkowane wspomnianymi okolicznościami o 28% częściej niż innymi kombinacjami wartości cech: *Rodzaj obszaru* i *Pora roku*.

Reguły wymienione powyżej (35 i 37) oraz reguły: 30, 31, 33 i 39 wskazują na ważność cechy *Obszar zabudowany* w generowaniu ciężkich wypadków z udziałem pieszych. Ten element współwystępuje z innymi elementami w ponad połowie reguł trzyelementowych w tabeli 2. *Kategoria powiatowa drogi*, obecna w dwuelementowej regule 24, nie jest obecna w żadnej istotnej regule trzyelementowej. Jednakże, kategoria wyższa (*Kategoria wojewódzka* w połączeniu z czynnikiem *Obszar zabudowany*) tworzy pozytywne asocjacje ze statusem ciężkim wypadku: zgodnie z regułą 39 prawie 40% wypadków na drogach wojewódzkich przechodzących przez obszary zabudowane to wypadki ciężkie.

Wypadki śmiertelne

Jest oczywiste, że reguły dla wypadków śmiertelnych mają niższe liczebności niż dla wypadków ze statusem lżejszym (wypadków śmiertelnych jest mniej niż ciężkich i lekkich). Jednakże miary *Lift* i *WWP* dla tych reguł mają generalnie większe wartości, zwłaszcza dla reguł trzyelementowych. Prawie wszystkie elementy warunkujące status śmiertelny należą do grupy czynników środowiska. Są to: *Dobra pogoda*, *Dzień pracy*, *Wieczór*, *Brak oświetlenia w nocy*. Słaba widoczność, wyrażona przez *PrDn_Wcz* (wieczór) oraz *Osw_BrOs* (brak oświetlenia drogi), występująca samodzielnie w regułach odpowiednio 43 i 44 oraz w połączeniu z elementami: *Prosty odcinek drogi* (reguły: 46 i 47), *Dobra pogoda* (reguły: 48 i 53), *Obszar zabudowany* (reguła 49) oraz *Sucha nawierzchnia jezdni* (reguły: 51 i 52) odgrywa kluczową rolę w definiowaniu okoliczności towarzyszących śmierci na drodze z powodu nieprawidłowych zachowań kierujących. Wypadki śmiertelne są:

- dwa razy częstsze przy braku oświetlenia w nocy niż przy innym oświetleniu (por. *WWP* w regułach: 44, 47 i 53),
- od 40% do 70% częstsze pod koniec dnia (wieczorem) niż o innej porze (por. *WWP* dla reguł: 43, 46, 48, 49 i 51).

Zmienna *Zachowanie kierującego* jest obecna tylko w jednej, dość silnej trzyelementowej regule asocjacyjnej 50 (*Lift* = 1,5, *WWP* = 1,8). Współwystępuje z elementem *Prosty odcinek drogi*. Niestety, wartość *Inne zachowania* tej zmiennej nie pozwala określić rodzaju nieprawidłowego zachowania sprawcy.

Zidentyfikowano następujący porządek czynników tworzących pozytywne asocjacje z ciężkością wypadku, poczynając od najbardziej znaczących: cechy środowiska, cechy drogi i niektóre cechy sprawcy wypadku. Ogólnie można stwierdzić, że wartości miar prawdopodobieństwa reguł asocjacyjnych, wsparcia i wiarygodności, zmniejszają się wraz ze wzrostem ciężkości wypadku. Natomiast siła asocjacji i wskaźnik wpływu poprzednika zmieniają się w sposób odwrotny. W grupie reguł dla wypadków spowodowanych przez kierujących, asocjacje dla statusu lekkiego i ciężkiego mają mniejszą ważność niż dla statusu śmiertelnego.

2.4. Uwagi końcowe

Analiza asocjacji to metoda eksploracji danych służąca do odkrywania zależności i współwystępowania elementów w zbiorach danych. Choć często kojarzona jest z analizą koszykową w handlu detalicznym, jej zastosowania wykraczają daleko poza obszar zakupów. Oprócz ww. zademonstrowanej analizie bezpieczeństwa (na przykładzie ruchu drogowego) można ja stosować np. w powiazań między genami, analizie wzorców ruchu i powiązanych zdarzeń w logistyce, identyfikacji schematów oszust finansowych.

3. Analiza sekwencji

Analiza sekwencji (*sequence analysis*) stanowi rozwinięcie analizy asocjacji, chociaż jest również traktowana jak rozwinięcie innych metod eksploracji danych. Podczas gdy reguły asocjacyjne pozwalają na identyfikację współwystępowania określonych elementów w zbiorach danych, analiza sekwencyjna wprowadza dodatkowy wymiar – porządek czasowy zdarzeń. Dzięki temu możliwe staje się nie tylko wskazanie, że pewne elementy występują wspólnie, ale także określenie, w jakiej kolejności pojawiają się w procesie lub przebiegu zdarzeń.

Analiza sekwencji ma zastosowanie w wielu dziedzinach, obejmując nauki przyrodnicze, obszary techniczne i społeczne. Niektóre wymieniono poniżej.

- Medycyna. Badanie kolejności występowania objawów, diagnoz i procedur, co wspiera prognozowanie przebiegu chorób oraz dobór odpowiednich terapii.
- Bioinformatyka. Analiza sekwencji DNA, RNA i białek w celu wykrywania powtarzalnych motywów oraz zrozumienia zależności biologicznych.
- Edukacja. Monitorowanie aktywności studentów w systemach e-learningowych i identyfikacja ścieżek prowadzących do lepszych efektów uczenia się.
- Procesy biznesowe. Identyfikacja powtarzalnych schematów działań w logistyce, produkcji, obsłudze klienta czy zarządzaniu projektami.
- Predykcja awarii w systemach technicznych. Analiza sygnałów i zdarzeń na liniach produkcyjnych, w sieciach energetycznych czy systemach dystrybucji ciepła, umożliwiająca przewidywanie usterek oraz wdrażanie strategii konserwacji predykcyjnej.
- Kryminalistyka. Odtwarzanie sekwencji zdarzeń w sprawach kryminalnych, identyfikacja powtarzalnych schematów działań sprawców oraz wspieranie procesu profilowania.
- Cyberbezpieczeństwo. Analiza logów systemowych i sieciowych w celu wykrywania charakterystycznych sekwencji działań atakujących, co pozwala na szybszą detekcję i przeciwdziałanie zagrożeniom.
- Analiza strumienia kliknięć w sieci. Badanie kolejności odwiedzanych stron internetowych pomaga w ulepszaniu nawigacji, poprawie doświadczenia użytkowników oraz zwiększaniu współczynnika konwersji.
- Przetwarzanie języka naturalnego (NLP). Wydobywanie sekwencji służy do odkrywania wzorców w danych tekstowych, co może usprawniać tłumaczenia maszynowe, streszczanie tekstów oraz analizę sentymentu.

Ważnym celem analizy sekwencyjnej jest odkrywanie wzorców sekwencyjnych (odkrywanie sekwencji) – powtarzalnych schematów zdarzeń – oraz formułowanie reguł sekwencyjnych, które umożliwiają prognozowanie przyszłych zdarzeń na

podstawie dotychczasowych obserwacji. Oba pojęcia bywają używane zamiennie, ale mają różne znaczenia. Analiza sekwencji obejmuje wszystkie metody badania danych sekwencyjnych, takie jak:

- grupowanie (klasteryzacja) sekwencji,
- modelowanie procesów,
- odkrywanie wzorców i reguł sekwencyjnych.

Analiza sekwencji stanowi szerokie podejście do pracy z danymi sekwencyjnymi, natomiast odkrywanie sekwencji jest jedną z metod w jej ramach, skoncentrowaną na znajdowaniu wzorców i reguł.

W kolejnych podrozdziałach tego rozdziału jest zaprezentowane praktyczne wykorzystanie analizy sekwencji do identyfikacji wzorców i reguł sekwencyjnych w procesie leczenia cukrzycy typu 2. Do tego celu zostały sztucznie wygenerowane dane. W całej pracy wykorzystano język Python i jego biblioteki do analizy i wizualizacji danych (pandas, matplotlib, seaborn, networkx).

3.1. Dane o terapiach cukrzycy typu 2

Dane dotyczące przebiegu terapii pacjentów z cukrzycą typu 2 zostały sztucznie wygenerowane, tak aby odwzorowywały typowe schematy leczenia, a jednocześnie nie odnosiły się do rzeczywistych pacjentów ani dokumentacji medycznej. Zawierają one informacje o kolejnych etapach terapii, takich jak zastosowanie leków, badania kontrolne czy hospitalizacje. Wykorzystano je do identyfikacji określonych wzorców i reguł sekwencyjnych w procesie leczenia. Wycinek danych jest zaprezentowany w tabeli 3. Ich struktura jest adekwatna do wykonania analizy sekwencji. Ta struktura została uwzględniona w programie w języku Python opracowanym do wykonania analizy sekwencji.

Tabela 3. Wycinek danych o terapiach cukrzycy typu 2 do analizy sekwencji

IdPac	Miesc	KodLeku	HbA1C	Waga	BMI	FazaLeczen
P008	0	GLP1+MET+SGLT2	6.9	98	31.2	Start
P008	5	MET	6.5	97	27	Korekta
P008	12	DPP4+INS+MET	6.1	97	31.7	Eskalacja
P008	32	INS+SU	6.5	88	28.4	Korekta
P009	0	DPP4+GLP1+SU	7.6	105	33.4	Start
P009	6	INS+SU	7.2	103	29.3	Korekta
P009	10	GLP1	7.6	90	32.2	Korekta
P009	15	DPP4+INS+SGLT2	6.8	102	36.2	Eskalacja
P009	25	MET+SGLT2	6.6	101	36.6	Spadek
P009	32	GLP1+INS	6	99	37.4	Stop

Charakterystyka zbioru

- 78 pacjentów (identyfikator pacjenta: *IdPac*), 392 obserwacje; pacjenci mają sekwencje terapii o różnej liczbie elementów.
- Obserwacja: 3 lata (36 miesięcy)
- Każdy pacjent ma przypisaną terapię (*KodLeku*): wprowadzenie leku na początku (*Miesc* = 0), zmianę leków lub modyfikacje dawek. Kody leków:
  - DPP4 inhibitor DPP-4,
  - GLP1 analog GLP-1,
  - INS insulina,
  - MET metformina,
  - SGLT2 inhibitor SGLT2,
  - SU sulfonylomocznik.
- Zmienna *HbA1c* odnosi się do poziomu hemoglobiny glikowanej – formy hemoglobiny, która powstaje, gdy glukoza we krwi wiąże się z hemoglobiną w czerwonych krwinkach. Ponieważ czerwone krwinki żyją około 120 dni, HbA1c odzwierciedla średni poziom glukozy we krwi z ostatnich 2–3 miesięcy, dlatego wskaźnik jest stosowany w monitorowaniu i diagnostyce cukrzycy typu 2. Badanie HbA1c ma istotne znaczenia w monitorowaniu długoterminowej kontroli cukrzycy, oceny skuteczności leczenia oraz diagnozowania i oceny ryzyka powikłań, takich jak uszkodzenia nerek czy oczu. Typowe wartości HbA1c są następujące:
  - < 5.7% prawidłowy poziom,
  - 5.7–6.4% stan przedcukrzycowy,
  - 6.5-7% cukrzyca (diagnostyczny próg),
  - > 7.0% niewyrównana cukrzyca (wymagająca intensyfikacji leczenia).
- Zmienne *Waga* oraz *BMI* informują odpowiednio o wadze pacjenta w kilogramach oraz indeksie masy ciał. Oba te parametry sa istotne z punktu widzenia leczenia cukrzycy.
- odnosi się do poziomu hemoglobiny glikowanej – formy hemoglobiny, która powstaje, gdy glukoza we krwi wiąże się z hemoglobiną w czerwonych krwinkach. Ponieważ czerwone krwinki żyją około 120 dni, HbA1c odzwierciedla średni poziom glukozy we krwi z ostatnich 2–3 miesięcy, dlatego wskaźnik
- Znaczenie wartości zmiennej *FazaLeczen*:
  - *Start* – rozpoczęcie terapii lekowej lub podjęcie nowego leczenia,
  - *Korekta*– dostosowanie dawki lub zmiana parametrów terapii (np. mała korekta),
  - *Eskalacja* – zwiększenie intensywności leczenia, np. dodanie nowego leku lub zwiększenie dawki,

- *Spadek* – zmniejszenie intensywności terapii, np. zmniejszenie dawki lub ograniczenie leków,
- *Stop* – zakończenie procesu intensyfikacji lub modyfikacji terapii i ustabilizowanie pacjenta na określonej kombinacji leków; nie oznacza rezygnacji z leczenia, ale osiągnięcie docelowej fazy terapeutycznej. dalsze obserwacje mogą pokazać kontynuację tej terapii, korekty dawek lub – jeśli wystąpią nowe okoliczności kliniczne – przejście do kolejnej fazy.

W celu podkreślenia istoty struktury, po wygenerowaniu, dane w tabeli 3 uporządkowano najpierw wg identyfikatora pacjenta a następnie wg kolejnego numeru miesiąca leczenia. Każdy pacjent ma przypisaną sekwencję zakodowanych zdarzeń wskazujących sekwencję leczenia; w tym przypadku: uporządkowany ciąg różnych leków stosowanych w określonej kolejności w leczeniu choroby. W tabeli 3, sekwencja *GLP1+MET+SGLT2 → MET → DPP4+INS+MET → INS+SU* leczenia pacjenta *P008* ma długość 4 – składa się z czterech elementów i może być opisana następująco.

1. Terapia lekowa została rozpoczęta przy *HbA1c* równym 6,9%. Pacjent miał zaordynowaną kombinację leków: analog GLP-1, metforminą i inhibitor SGLT2 (*GLP1+SGLT2+SU*) – bardzo nowoczesne, agresywne podejście, sugerujące pacjenta z wysokim ryzykiem sercowo-naczyniowym i/lub otyłością.
2. W piątym miesiącu terapii zastąpione dotychczasowe leki metforminą (*MET*). Wcześniejsza terapia została odstawiona (np. z powodu nietolerancji GLP-1 lub SGLT2, kosztów albo działań niepożądanych), pozostawiono tylko metforminę. To uproszczenie schematu mogło być związane ze stanem klinicznym pacjenta.
3. Po roku (12. miesiąc) od zdiagnozowania choroby, do metforminy dołączono inhibitor DPP-4 oraz insulinę (*DPP4+INS+MET*). To może oznaczać, że kontrola cukrzycy była niezadowalająca. DPP-4 działa łagodniej niż GLP-1, ale może być lepiej tolerowany. Wprowadzona kombinacja leków wskazuje na potrzebę wzmocnienia leczenia z powodu progresji choroby metabolicznej.
4. Po następnych dwudziestu miesiącach (32. miesiąc leczenia), w zaordynowanym zestawie leków wprowadzono korektę, pozostawiając insulinę a inne leki zastępując sulfonylomocznikiem. Pacjent przeszedł na schemat bardziej klasyczny i tańszy. Sulfonylomoczniki stymulują wydzielanie insuliny, a podawana insulina uzupełnia niedobory wynikające z wyczerpania trzustki. Możliwe, że odstawienie metforminy było implikowane np. nietolerancją żołądkowo-jelitową albo niewydolnością nerek.

3.2. Wzorce w sekwencjach terapii dla pacjentów

Analiza sekwencji jest narzędziem, które pozwala nie tylko śledzić pojedyncze zdarzenia, ale przede wszystkim odkrywać wzorce w kolejności ich występowania.

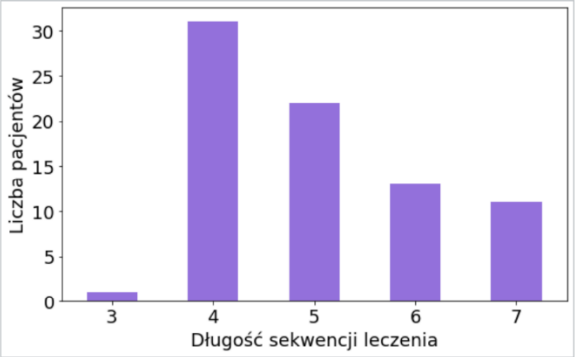
W kontekście leczenia chorób przewlekłych, takich jak cukrzyca typu 2, szczególnego znaczenia nabiera możliwość uchwycenia, jak pacjent przechodzi przez kolejne etapy terapii – od leczenia początkowego, przez modyfikacje dawkowania i zmiany kombinacji leków, aż po sytuacje wymagające intensyfikacji lub ustabilizowania leczenia.

Analiza wzorców w sekwencjach leczenia pacjentów umożliwia wyprowadzenie z danych surowych użyteczne informacje o charakterze diagnostycznym i terapeutycznym. Nie ogranicza się do pojedynczych terapii, ale pozwala uchwycić powtarzalne schematy, które mogą mieć znaczenie kliniczne – na przykład często obserwowane przejścia między terapiami lub typowe ciągi zdarzeń charakterystyczne dla określonych grup pacjentów.

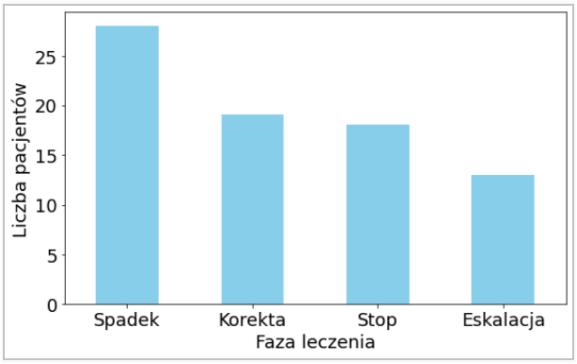
Rys. 2 ilustruje różnice w długości sekwencji terapeutycznych wśród pacjentów. Rozkład długości sekwencji pokazuje, jak wielu pacjentów przechodzi przez prostsze (krótsze) lub bardziej złożone (dłuższe) schematy leczenia. Średnia długość terapii pacjentów leczonych przez trzy lata jest równa 5, co oznacza, że jeden pacjent miał w analizowanym okresie średnio około 5 razy modyfikowaną swoją terapie. Największą grupę (31 osób) obejmują terapie o długości czterech etapów. Liczba pacjentów z coraz dłuższą sekwencją leczenia stopniowo się zmniejsza – kolejno 22, 13 i 11 osób dla sekwencji terapii 5-, 6- i 7-elementowych. Z punktu widzenia praktyki klinicznej zmiany terapii w pierwszych latach od rozpoczęcia leczenia mogą być uzasadnione, ponieważ lekarze dobierają różne kombinacje leków, aby osiągnąć docelowe wartości glikemii i stabilizację stanu pacjenta. Często udaje się opanować glikemię na wcześniejszych etapach. Dlatego większość pacjentów nie musi dochodzić aż do 6. czy 7. linii leczenia. Coraz mniejsza liczba pacjentów dla coraz dłuższych terapii jest zazwyczaj implikowana faktem, że część osób osiąga kontrolę glikemii wcześniej i tylko mniejszość wymaga wielokrotnej modyfikacji schematu. Może być również tak, że niektórzy przestają być obserwowana (np. zmiana lekarza) a inni rezygnują z terapii (np. z powodu działań niepożądanych).

Rys. 3 przedstawia jaki był etap terapii po trzech latach leczenia. W grupie 78 osób zarejestrowano 28 przypadków poprawy stanu zdrowia (faza leczenia = *Spadek*;) oraz 18 przypadków ustabilizowania leczenia (faza leczenia = *Stop*), co łącznie stanowi blisko 60% ogółu pacjentów. U ponad 16% leczonych (faza leczenia = *Eskalacja;* 13 osób) nie uzyskano pozytywnych rezultatów, co świadczy o nieefektywności dotychczasowej terapii i trudnej do kontrolowania cukrzycy oraz oznacza konieczność zintensyfikowania lub modyfikacji dotychczasowego procesu leczenia.

[Tekst alternatywny. Dwa wykresy słupkowe obok siebie. Na obu wykresach na osi pionowej jest odłożona liczba pacjentów a na osi poziomej: na lewym wykresie – liczby całkowite określające długość sekwencji leczenia, na prawym wykresie – opisy końcowych faz leczenia w sekwencjach terapii pacjentów.]



Rys 2. Rozkład długości sekwencji terapii pacjentów



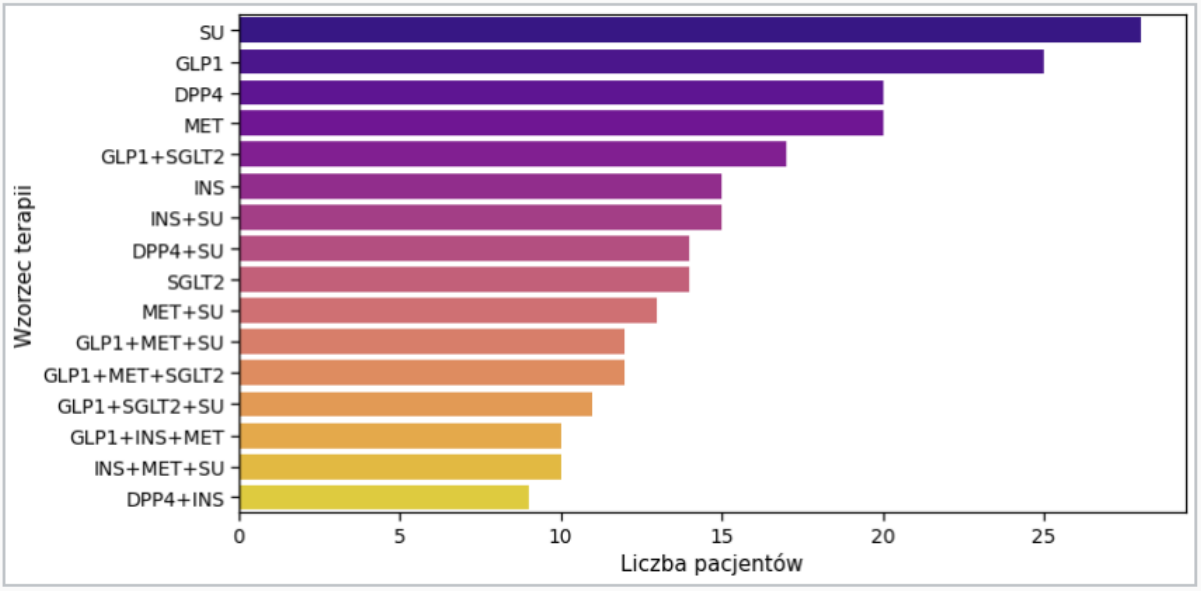
Rys. 3. Rozkład faz leczenia pacjentów po trzech latach terapii

Każdy etap każdej sekwencj terapii jest jednoelementowy; składa się albo z pojedycznego leku albo z kombinacji leków. Może być traktowany jako najprostszy wzorzec sekwencyjny (długości 1). Rys. 4 przedstawia takie jednoelementowe wzorce sekwencji, występujące w minimum 1/5 ogólnej liczby 78 sekwencji terapii. Najczęściej stosowane terapie wskazują, które leki i ich kombinacje dominują w praktyce leczenia.

Wystąpienie pojedynczych leków informuje o stosowaniu monoterapii w sekwencji leczenia; sulfonylomocznik (*SU*; 30 razy) jest popularny w monoterapii, gdy brak dostępności innych leków lub gdy inne leki są przeciwwskazane albo nieskuteczne u danego pacjenta, metforminę (MET, 26 razy) najczęściej aplikuje się w początkowej fazie leczenia, insulina stosuje się jako insuliną bazalną lub bolusową (INS, 17 razy) – w zasadzie nie występuje w początkowej fazie leczenia, pozostałe leki: GLP1 (25 razy), DPP4 (21 razy), SGLT2 (17 razy) są wykorzystywane, gdy pacjent ma choroby towarzyszące. Wszystkie sześć leków stosowanych w monoterapii znajduje się w pierwszej ósemce najczęstszych wzorców leczenia cukrzycy typu 2. Popularność monoterapii wynika z jej licznych zalet: jest bezpieczna i łatwa do monitorowania, pozwala ocenić reakcję pacjenta, skutecznie kontroluje glikemię w początkowej fazie choroby oraz umożliwia stopniową i kontrolowaną intensyfikację terapii. Kombinacje 2-3 leków od *GLP1+MET+SU* (18 razy) do *INS+SGLT2* (8 razy) świadczą o intensyfikacji terapii, w zależności od tolerancji pacjenta, ryzyka sercowo-naczyniowego i osiąganej kontroli glikemii. W całej bazie sekwencji nie ma terapii obejmującej więcej niż trzy leki jednocześnie. Najczęściej pojawiają się *SU* i *MET*, leki będące standardem w klasycznych schematach leczenia. Jednocześnie wiele

wzorców zawiera nowoczesną kombinację *GLP1* i *SGLT2*, odzwierciedlając trend w kierunku terapii spersonalizowanej, zwłaszcza u pacjentów z otyłością lub ryzykiem sercowo-naczyniowym.

[Tekst alternatywny. Kolorowy wykres słupkowy horyzontalny. Przedstawia najczęstsze jednoelementowe wzorce terapii w sekwencjach terapii. Na osi pionowej są odłożone wzorce terapii a na osi poziomej liczba pacjentów.]



Rys. 4. Top 20% najczęstszych jednoelementowych wzorców terapii

3.3. Reguły sekwencyjne w terapiach pacjentów

W zbiorze 78 sekwencji leczenia pacjentów 61 to sekwencje unikatowe. Różnią się one zarówno zastosowanymi terapiami, jak i ich kolejnością. Na rys. 5 pokazano 16 sekwencji pojawiających się najczęściej. Najliczniejsza ścieżka terapeutyczna (DPP4+INS+SGLT2 →, GLP1+SGLT2 → GLP1+MET+SU → SU) wystąpiła zaledwie 6 razy (7,7%). Tylko 8% nieunikatowych sekwencji (pięć pierwszych na wykresie) jest aplikowana 28% pacjentów (22 osoby). Większość sekwencji występuje tylko raz – pojawia się tzw. „długi ogon” rzadkich wzorców, co może utrudniać wyciąganie wiarygodnych wniosków statystycznych dla konkretnych rozwiązań terapeutycznych.

Obraz wskazuje na bardzo dużą różnorodność strategii leczenia cukrzycy typu 2, z częstymi terapiami wieloskładnikowymi, co może być spowodowane różnymi profilami pacjentów (wiek, BMI, choroby współistniejące), różnymi strategiami lekarzy, różnymi momentami włączenia nowych leków. W sekwencjach dominują kombinacje leków: DPP4 + INS + SGLT2, GLP1 + SGLT2, INS + SU, zwłaszcza INS

(insulina) oraz GLP1, świadcząc o częstszym stosowaniu wielolekowego podejścia niż monoterapię w zaawansowanych etapach i odzwierciedla potrzebę intensyfikacji terapii, gdy pojedyncze leki zawodzą. „Długi ogon" rzadkich sekwencji to efekt indywidualizacji leczenia (np. pacjenci z chorobami nerek, serca, różnym BMI) świadczy, że lekarze dostosowują leczenie do profilu pacjenta, a nie stosują sztywnego schematu. Terapie skojarzone z insuliną i inkretynami (naśladują lub wzmacniają działanie naturalnych hormonów jelitowych, inkretyn, zwiększając wydzielanie insuliny i hamując glukagon w sposób zależny od poziomu glukozy, tu: GLP1, DPP4 oraz kombinacje leków z tymi preparatami) odzwierciedlają eskalację leczenia. Duża heterogeniczność sekwencji terapii może oznaczać, że w praktyce trudno wyłonić typową ścieżkę pacjenta i warto analizować dane na poziomie klas leków lub strategii eskalacji.

[Tekst alternatywny. Kolorowy wykres słupkowy horyzontalny. Przedstawia najczęstsze sekwencje terapii. Na osi pionowej są odłożone sekwencje terapii a na osi poziomej liczba pacjentów.]

Rys. 5. Top 16 najczęstszych sekwencji terapii

Analiza reguł sekwencyjnych (przejść) pozwala ujawnić typowe ścieżki przebiegu terapii, najczęściej jej eskalacji. Tabela 4 pokazuje przejścia między bezpośrednio sąsiadującymi wzorcami leków w terapiach mającymi liczność co najmniej 3, co stanowi 3,85% wsparcia w bazie danych sekwencji terapii. Najliczniejszymi były przejścia, w których poprzednik ma jeden element i następnik jeden element. Ilustrację graficzną reguł sekwencyjnych przedstawia rys. 6.

Uporządkowane malejąco według wsparcie reguły wskazują cztery pierwsze najczęstsze przejścia terapeutyczne, stanowiące razem około 32% wszystkich zmian terapii. Wg tych reguł złożone terapie są stosunkowo często upraszczane (kombinacja leków do monoterapii) i dość powszechnie występują zmiany między terapiami inkretynowymi (GLP1, DPP4) a terapiami z SGLT2.

Tabela 4. Najczęstsze sąsiadujące przejścia terapii - reguły sekwencyjne

Nr	Reguła sekwencyjna			Wsparcie	Wsparcie [%]	Wiarygodność [%]
1	GLP1+MET+SU	→	SU	7	8,97	64
2	DPP4+INS+SGLT2	→	GLP1+SGLT2	6	7,69	75
3	GLP1+SGLT2	→	GLP1+MET+SU	6	7,69	40
4	GLP1+SGLT2+SU	→	DPP4	6	7,69	55
5	INS+SU	→	DPP4+INS+SU	5	6,41	45
6	DPP4+INS+SU	→	GLP1	5	6,41	71
7	GLP1	→	MET+SU	5	6,41	21
8	DPP4	→	DPP4+GLP1+MET	5	6,41	31
9	DPP4+GLP1+MET	→	GLP1+INS+MET	5	6,41	62
10	GLP1+INS+MET	→	INS+MET+SGLT2	5	6,41	50
11	GLP1	→	INS+MET+SU	4	5,13	17
12	SU	→	LP1+INS+SGLT2	4	5,13	22
13	INS+SU	→	GLP1+SU	3	3,85	27
14	GLP1+SU	→	LP1+SGLT2+SU	3	3,85	43
15	GLP1+SGLT2+SU	→	DPP4+SU	3	3,85	27
16	GLP1+INS	→	INS	3	3,85	75
17	GLP1+INS+SGLT2	→	INS+SGLT2	3	3,85	33
18	INS+SGLT2	→	MET	3	3,85	43

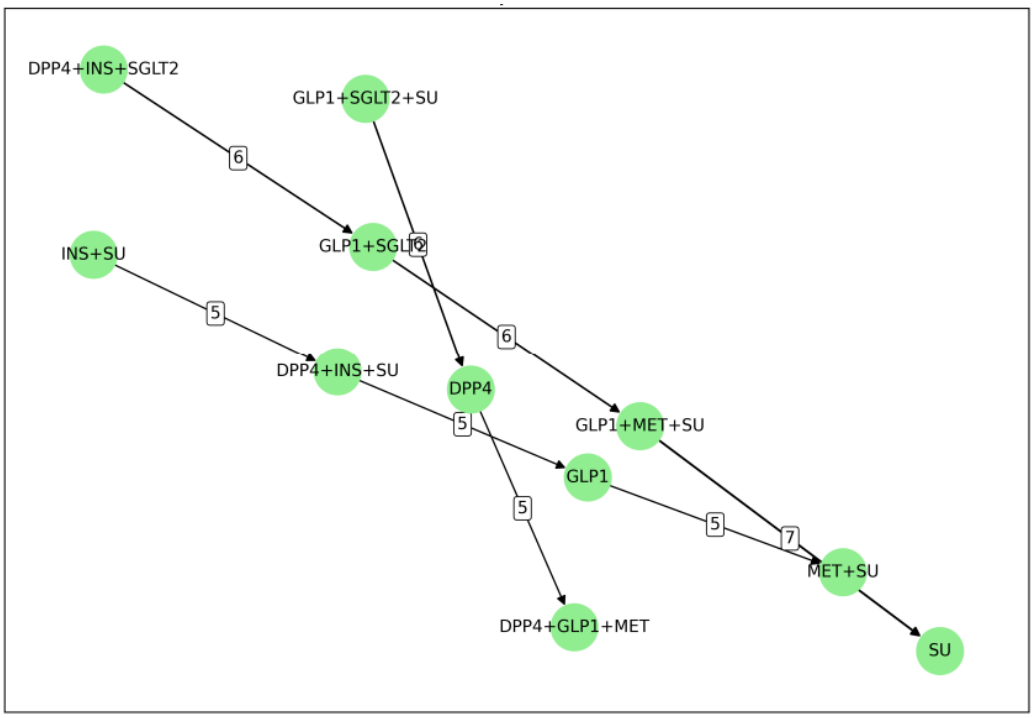
Wysoka wiarygodność sugeruje silny wzorzec decyzji terapeutycznej w opisanych poniżej przypadkach:

- Wsparcie 75%
  - DPP4+INS+SGLT2 → GLP1+SGLT2; w terapiach z SGLT2 skojarzonym z nhibitorem DPP4 i insuliną bardzo często te dwa ostatnie są zastępowane agonistą GLP1 świadcząc o intensyfikacj leczenia inkretynowego.
  - GLP1+INS → INS; w ¾ przypadków terapii skojarzonej z GLP1 następuje rezygnacja z GLP1 i pozostaje tylko insulina, co może oznaczać progresję choroby, jednak układ takich przejść występuje bardzo rzadko (tylko w trzech terapiach).
- Wsparcie 71%
  - DPP4+INS+SU → GLP1; złożoną terapię z inhibitorem DPP-4, insuliną i pochodną sulfonylomocznika w 70% przypadków zastępuje się leczeniem

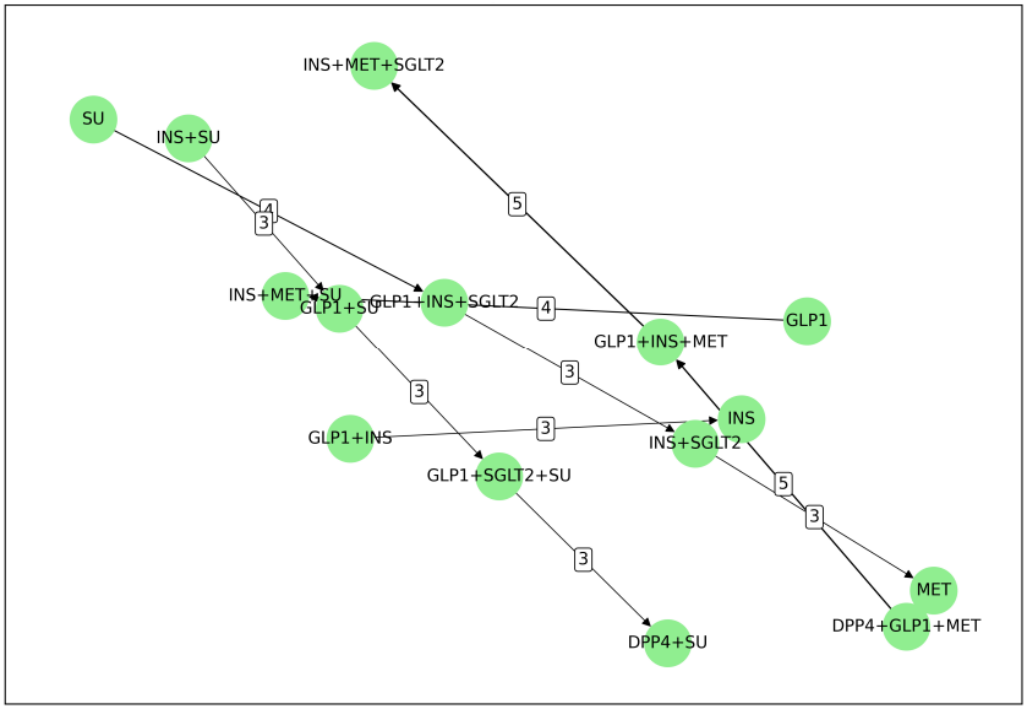
opartym na agonistach GLP-1, najprawdopodobniej w celu poprawy skuteczności i bezpieczeństwa terapii.

- Wsparcie 64%
  - GLP1+MET+SU → SU; dość duże (64%) prawdopodobieństwo uproszczenia terapii do monoterapii SU. Może oznaczać nietolerancję lub decyzję kliniczną o zmianie strategii leczenia.
- Wsparcie 62%
  - DPP4+GLP1+MET → GLP1+INS+MET; dodanie insuliny do istniejącego leczenia inkretynowego skazuje na eskalację terapii.

[Tekst alternatywny. Dwa wykresy w formie grafów skierowanych przedstawiające przejścia między elementami terapii zestawionymi w tabeli 4. Wykres 6a ilustruje przejścia 1-8, rys. 6b przejścia 8-18.]



Rys. 6a. Graf najczęstszych przejść w sekwencjach terapii: przejścia 1-8



Rys. 6b. Graf najczęstszych przejść w sekwencjach terapii: przejścia 8-18

Reguły z wiarygodnością > 60% wskazują powtarzalne decyzje terapeutyczne – można je traktować jako standardowe ścieżki leczenia w populacji, pod warunkiem ich potwierdzenia w znacząco większej wartości wskaźnika wsparcia.
Reguły o niskiej wiarygodności mogą świadczyć o indywidualizacji leczenia lub braku ustalonych standardów dalszej eskalacji.

Analizowano dane syntetyczne stworzone sztucznie przez model, służące wyłącznie do celów ilustracyjnych w analizie sekwencji. Zostały skonstruowane tak, aby wyglądały sensownie i realistycznie w kontekście leczenia cukrzycy typu drugiego. Ich sensowność wynika z logicznego układu kolejnych terapii, np. najpierw terapia łagodniejsza, później intensyfikacja leczenia. Obejmowały schematy monoterapii, terapii skojarzonych, eskalacji do insuliny czy zmian między inhibitorami DPP-4 a agonistami GLP-1. Zbiór hipotetycznych, nieprawdziwych sekwencji terapii cukrzycy typu 2 zawiera informacje przedstawiające złożony charakter choroby: z jednej strony częstą eskalację terapii w miarę progresji choroby, a z drugiej – upraszczania leczenia, gdy intensywna terapia okazuje się zbędna lub źle tolerowana.

3.4. Uwagi końcowe

Wszystkie przedstawione wnioski są oparte wyłącznie na danych sztucznie wygenerowanych przy użyciu technologii sztucznej inteligencji i w żadnym przypadku

nie odzwierciedlają rzeczywistych danych klinicznych, przebiegu choroby ani istniejących obserwacji klinicznych cukrzycy typu drugiego. Nie powinny być traktowane jako wskazówki, zalecenia ani odkrycia dotyczące leczenia cukrzycy typu drugiego. Jakiekolwiek próby przeniesienia tych wyników do praktyki klinicznej mogą prowadzić do błędnych wniosków i niebezpiecznych decyzji. Prezentowane przykłady mają jedynie na celu zilustrowanie metod analizy sekwencji i pokazać, w jaki sposób narzędzia sztucznej inteligencji mogą być wykorzystane w edukacji i nauce analityki danych.

4. Drzewa decyzyjne

Drzewa decyzyjne (*decision trees*) są definiowane poprzez strukturę drzewiastą i należą do technik (narzędzi) uczenia nadzorowanego. To oznacza, że jedna ze zmiennych definiujących zbiór danych będących przedmiotem badań jest zmienną celu (objaśnianą, w nomenklaturze drzew decyzyjnych – decyzją). Budowę drzewa zaczyna się w jego korzeniu, a kończy po osiągnięciu węzła terminalnego, czyli liścia. Sposób konstrukcji drzew zależy od metody podziału zbioru danych w węzłach i charakteru decyzji. Gdy decyzja jest zmienną jakościową (opisową), powstaje drzewo klasyfikacyjne; do popularnych algorytmów jego budowy zalicza się: CHAID, CART, C4.5, QUEST. Gdy decyzja jest zmienną ilościową (liczbową), powstaje drzewo regresyjne; do popularnych algorytmów jego budowy zalicza się: CART, C4.5.

Techniki drzew decyzyjnych mogą być wykorzystane do identyfikacji związków przyczynowo-skutkowych wypadków drogowych i przedstawienia ich w postaci reguł języka potocznego lub zdań logicznych. Każde zdarzenie drogowe jest opisane przez wiele cech, w większości o charakterze jakościowym. Model drzewiasty można wykorzystać do wskazania tych cech charakteryzujących zdarzenia drogowe, które wywierają znaczący wpływ na skutki tych zdarzeń. Model taki zbudowano w oparciu o dane o wypadkach i kolizjach zarejestrowanych na skrzyżowaniach ulic układu podstawowego w Kielcach w latach 1999-2001. Przedmiotem analiz były te zdarzenia drogowe, w których brały udział tylko pojazdy. Wszystkie eksperymenty przeprowadzono w środowisku GUI Enterprise Miner™ systemu SAS®.

4.1. Dane o zdarzeniach drogowych na skrzyżowaniach ulicznych

Przed przeprowadzeniem analiz dane o zdarzeniach drogowych zostały poddane przekształceniom w celu uporządkowania i wyczyszczenia zbioru. Poniżej opisano

analizowane cechy (zgodnie z policyjną Kartą Zdarzenia Drogowego – KZD) oraz podano rozkłady ich wartości.

- *WLOTY* – zawiera informacje, z jakich wlotów nadjechały pojazdy biorące udział w zdarzeniu. Przyjmuje wartości:
  - *S* (37,8%) – jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżały z tego samego wlotu,
  - *R* (25,0%) – jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżały z przeciwległych (równoległych) wlotów,
  - *P* (37,2%) – jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżały z wlotów poprzecznych.
- *RELACJE* – określa kombinacje relacji pojazdów biorących udział w zdarzeniu. Przyjmuje wartości: *L-L* (4,3%), *L-P* (2,0%), *L-W* (39,0%), *P-P* (6,1%), *P-W* (5,6%), *W-W* (43,0%), gdzie litery oznaczają relacje pojazdu: *L* – skręcającego w lewo, *P* – skręcającego w prawo, *W* – jadącego na wprost.
- *SYGNALIZACJA* (KZD) – informuje o obecności na skrzyżowaniu sygnalizacji świetlnej. Przyjmuje wartości:
  - Jest (51,8%),
  - *Nie ma* (48,2%) – wynik agregacji dwóch wartości cechy z KZD: *brak* i *nie działa*.
- *TYP*_SKRZYŻOWANIA – *zawiera* informacje o typie skrzyżowania, na którym miało miejsce zdarzenie. Możliwe wartości tej cechy są następujące: *Skanalizowane* (75,0%), *Zwykłe* (8,5%), *Zwykłe o poszerzonych wlotach* (16,5%).
- *ZACHOWANIE_KIEROWCY* (KZD) – cecha opisująca zachowanie kierowcy. Przyjmuje wartości:
  - *Niezachowanie bezpiecznej odległości między pojazdami* (14,2%),
  - *Nieprawidłowe skręcanie* (4,4%),
  - *Nieprawidłowe manewry* (4,4%) – wynik agregacji kilku wartości cechy z KZD: *Nieprawidłowe wyprzedzanie*; *Nieprawidłowe przejeżdżanie przejść dla pieszych*; *Nieprawidłowe omijanie*; *Nieprawidłowe cofanie*; *Nieprawidłowe wymijanie*,
  - *Nieudzielnie pierwszeństwa* (57,7%) – wynik agregacji dwóch wartości cechy z KZD: *Wjazd przy czerwonym świetle*; *Nieudzielenie pierwszeństwa przejazdu*,
  - *Niedostosowanie prędkości* (13,8%) – wynik agregacji dwóch wartości cechy z KZD: *Niedostosowanie prędkości do warunków ruchu*; *Gwałtowne hamowanie*,
  - *Inne zachowania* (5,5%) – wynik agregacji kilku wartości cechy z KZD: *Zmęczenie, zaśnięcie*; *Ograniczenie sprawności psychomotorycznej*; *Nieprzestrzeganie innych przepisów drogowych*; *Jazda po niewłaściwej stronie*; *Inne*.

- *RODZAJ_ZDARZENIA* (KZD) – cecha opisująca rodzaj zdarzenia drogowego. Przyjmuje wartości:
  - *Zderzenie czołowe* (10,3%),
  - *Zderzenie boczne* (55,4%),
  - *Zderzenie tylne* (31,4%),
  - *Pozostałe zdarzenia* (2,9%) – *wynik agregacji kilku wartości cechy z KZD: Najechanie na unieruchomiony pojazd; Najechanie na drzewo, słup, inny obiekt drogowy*; *Najechanie na zaporę kolejową*; *Najechanie na dziurę, wybój, garb*; *Najechanie na zwierzę*; *Wywrócenie się pojazdu*; *Wypadek z pasażerem*; *Inne rodzaje*.
- *STATUS_ZDARZENIA* – informuje o klasyfikacji zdarzenia. Przyjmuje wartości: *Wypadek* (11,7%) i *Kolizja* (88,3%).
- *STAN_NAWIERZCHNI* (KZD) – *cecha* opisująca stan nawierzchni w chwili zajścia zdarzenia. Przyjmuje wartości:
  - *Sucha* (60,2%),
  - *Mokra* (33,4%),
  - *Oblodzona, zaśnieżona* (5,7%),
  - *Inna* (0,7%) – wynik agregacji kilku wartości cechy z KZD: *Kałuże*, *rozlewiska*; *Zanieczyszczona*; *Inny stan*.
- *WARUNKI_ATMOSFERYCZNE* (KZD) – cecha opisana przez wartości:
  - *Pochmurno* (14,4%),
  - *Dobre warunki atmosferyczne* (64,0%),
  - *Opady* (15,7%) – wynik agregacji dwóch wartości cechy z KZD: *Opady deszczu* oraz *Opady śniegu, gradu*,
  - *Pozostałe* (5,9%) – wynik agregacji kilku wartości cechy z KZD: *Oślepiające słońce*; *Silny wiatr*; *Mgła, dym*; *Brak danych*.
- *OŚWIETLENIE* (KZD) – opisuje warunki widoczności na miejscu zdarzenia. Przyjmuje wartości:
  - *Światło dzienne* (80,4%),
  - *Noc-oświetlenie* (19,5%) – wynik agregacji kilku wartości cechy z KZD: *Zmrok, świt*; *Noc – droga oświetlona*; *Noc – droga niedostatecznie oświetlona*; *Noc – droga nieoświetlona*. Agregowanie do wartości *Noc – oświetlenie* jest w tym przypadku uzasadnione, ponieważ skrzyżowania ulic układu podstawowego są zawsze oświetlone. Jeśli zdarzenie nastąpiło w nocy na skrzyżowaniu ulic, to było to miejsce oświetlone,
  - *Brak danych* (0,1%).
- *GODZINA* – cecha opisująca godzinę zdarzenia zagregowana, tak aby była określona pora doby. Przyjmuje wartości:
  - *1* (3,0%) – *godziny* w przedziale $[0^{00}$; $6^{00})$ zostały sklasyfikowane jako okres nocy,

- *2* (24,7%) – *godziny* w przedziale [6^{00} ; 12^{00}) zostały sklasyfikowane jako okres przedpołudniowy,
- *3* (44,5%) – *godziny* w przedziale [12^{00} ; 18^{00}) zostały sklasyfikowane jako okres popołudniowy,
- *4* (27,8%) – *godziny* w przedziale [18^{00} ; 24^{00}) zostały sklasyfikowane jako okres wieczorny.

Z analizy wykluczono rekordy, dla których brakowało informacji o wlocie lub relacji pojazdu biorącego udział w zdarzeniu. Przygotowany zbiór danych liczył 2027 rekordów. Cały zbiór dzielono na podzbiory: treningowy, walidacyjny i testowy. Liczba obserwacji w zbiorach powstałych w wyniku podziału jest zależna od rodzaju eksperymentu.
Dane zostały przygotowane, tak aby uwzględnić ewentualną nierównomierność rozkładów decyzji, dla których prowadzono eksperymenty. W przypadku dużej dysproporcji rozkładu decyzji zastosowano próbkowanie warstwowe dla poszczególnych kategorii decyzji.

4.2. Związki przyczynowo-skutkowe zdarzeń drogowych na skrzyżowaniach ulicznych

Pewne cechy charakteryzujące zdarzenia drogowego, które odgrywają istotną rolę w diagnozie stanu bezpieczeństwa ruchu drogowego, zostały zdefiniowane jako zmienne objaśniane (czyli decyzje) w eksperymentach prowadzonych przy wykorzystaniu drzew decyzyjnych. Należą do nich:

- status zdarzenia (np. wypadek, kolizja),
- rodzaj zdarzenia (np. zderzenie czołowe, wywrócenie pojazdu),
- zachowanie kierowcy (np. nieudzielenie pierwszeństwa przejazdu, nieprawidłowe manewry).

Jako zmienne objaśniające wybrane zostały inne cechy opisujące zdarzenie drogowe oraz cechy geometryczno-ruchowe miejsca zdarzenia. Zadaniem każdego eksperymentu jest wygenerowanie wiarygodnych reguł definiujących związki przyczynowo-skutkowe między badanymi cechami.
Modele zbudowane dla wszystkich eksperymentów posiadają wspólne charakterystyki, realizując następujący wspólny schemat działania:

- podział warstwowy zbioru źródłowego wg decyzji (cechy objaśnianej): 70% – zbiór treningowy, 30% – zbiór walidacyjny,
- parametry drzewa:
  - minimalna liczba obserwacji w liściu: 20,
  - liczba obserwacji wymagana dla podziału węzła: 40,
  - maksymalna liczba gałęzi wychodzących z węzła: 6,
  - maksymalna głębokość drzewa: 10.

- zastosowanie dwóch różnych algorytmów podziału drzewa decyzyjnego wykorzystując metody: minimum entropii i chi-kwadrat,
- każdy z ww. algorytmów był wielokrotnie stosowany do wygenerowanych losowo ze zbioru źródłowego zbiorów treningowych i walidacyjnych.

Wybierane reguły są nie tylko zdefiniowane przez najczystsze liście, ale również spełniają warunki powtarzalności wyników eksperymentu.

Eksperyment 1

Zadaniem eksperymentu jest sprawdzenie czy istnieją reguły, które na podstawie określonych wartości cech objaśniających pozwolą sklasyfikować status zdarzenia drogowego: wypadek czy kolizja.
Parametry eksperymentu

- Zmienne objaśniające: *WLOTY, RELACJE*, *SYGNALIZACJA*, *TYP_SKRZYŻOWANIA*, *ZACHOWANIE_KIEROWCY*, *STAN_NAWIERZCHNI*, *WARUNKI_ATMOSFERYCZNE*, *OŚWIETLENIE*, *GODZINA*
- Decyzja: *STATUS_ZDARZENIA*

Zmienna decyzyjna ma bardzo nierównomierny rozkład: blisko 12% zdarzeń drogowych stanowią wypadki, pozostałe 88% to kolizje. W eksperymentach niwelujących tak dużą różnicę w rozkładzie wartości zmiennej celu *STATUS_ZDARZENIA* zastosowano próbkowanie z następującymi proporcjami wypadków do kolizji: 25% na 75%, 30% na 70%, 40% na 60%. Wykorzystano technikę, w której wylosowano określoną liczbę przykładów z kategorii większościowej (*STATUS_ZDARZENIA* = *Kolizja*) i wszystkie przykłady z kategorii mniejszościowej (*STATUS_ZDARZENIA* = *Wypadek*).

Drzewa będące wynikiem eksperymentów składały się z korzenia lub zawierały bardzo zanieczyszczone liście. Wyniki takie otrzymano zarówno dla różnych testowanych rozkładów zmiennej celu i różnych wariantów par zbiorów: treningowego i walidacyjnego. Dla analizowanego zbioru danych o zdarzeniach drogowych nie zidentyfikowano więc wyrazistych związków przyczynowo-skutkowych dotyczących statusu zdarzenia drogowego.

Eksperyment 2

Zadaniem eksperymentu jest sprawdzenie czy istnieją reguły, które pozwolą zdiagnozować czy rodzaj zdarzenia drogowego na skrzyżowaniu jest zdeterminowany określonymi wartościami cech objaśniających.
Parametry eksperymentu:

- Zmienne objaśniające: *WLOTY, RELACJE*, *SYGNALIZACJA*, *TYP_SKRZYŻOWANIA*, *ZACHOWANIE_KIEROWCY*,
- Decyzja: *RODZAJ_ZDARZENIA*.

W wielokrotnym powtarzaniu eksperymentów dla różnych par zbiorów: treningowego i walidacyjnego uzyskano podobne wzorce zdarzeń drogowych. Zarówno w przypadku metody minimum entropii jak i chi-kwadrat otrzymano bardzo proste drzewa o stosunkowo czystych liściach. Miary oceny tych drzew na i obu zbiorach wahały się nieznacznie:

- wartości wskaźnika *Proportion Correctly Classified* (*PCC*): 72% ÷ 76%,
- wartości na przekątnej macierzy pomyłek: *Zderzenie czołowe*: 15% ÷ 30%, *Zderzenie tylne*: 85% ÷ 90%, *Zderzenie boczne*: 85% ÷ 90%.

W przypadku obu metod podział zbioru na pierwszym poziomie drzewa został przeprowadzony na podstawie wartości cechy *WLOTY*; wskaźnik ważności cechy *Importance* = 1, na drugim poziomie podział dokonał się tylko dla jednego węzła, dla którego *WLOTY* = *S*. Cechą dzielącą zbiór jest *ZACHOWNIE_KIEROWCY*; ważność tej cechy dla metody minimum entropii: 0,47 ÷ 0,49, a dla metody chi-kwadrat: 0,42 ÷ 0,45.

W przypadku metody entropii wygenerowane zostały następujące reguły:

Reguła 1: Jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżają z wlotów poprzecznych, to rodzajem zdarzenia jest zderzenie boczne (86%-89% przypadków w zbiorach walidacyjnych).

Reguła 2: Jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżają z wlotów przeciwnych/równoległych, to rodzajem zdarzenia może być zderzenie boczne lub czołowe (odpowiednio 68%-71% i 28%-31% przypadków w zbiorach walidacyjnych).

Reguła 3: Jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżają z tych samych wlotów i kierowca nie zachowuje bezpiecznej odległości, to rodzajem zdarzenia jest zderzenie tylne (więcej niż 90% przypadków w zbiorach walidacyjnych).

Reguła 4: Jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżają z tych samych wlotów i kierowca nie dostosowuje prędkości do warunków ruchu, to rodzajem zdarzenia jest zderzenie tylne (87%-94% przypadków w zbiorach walidacyjnych).

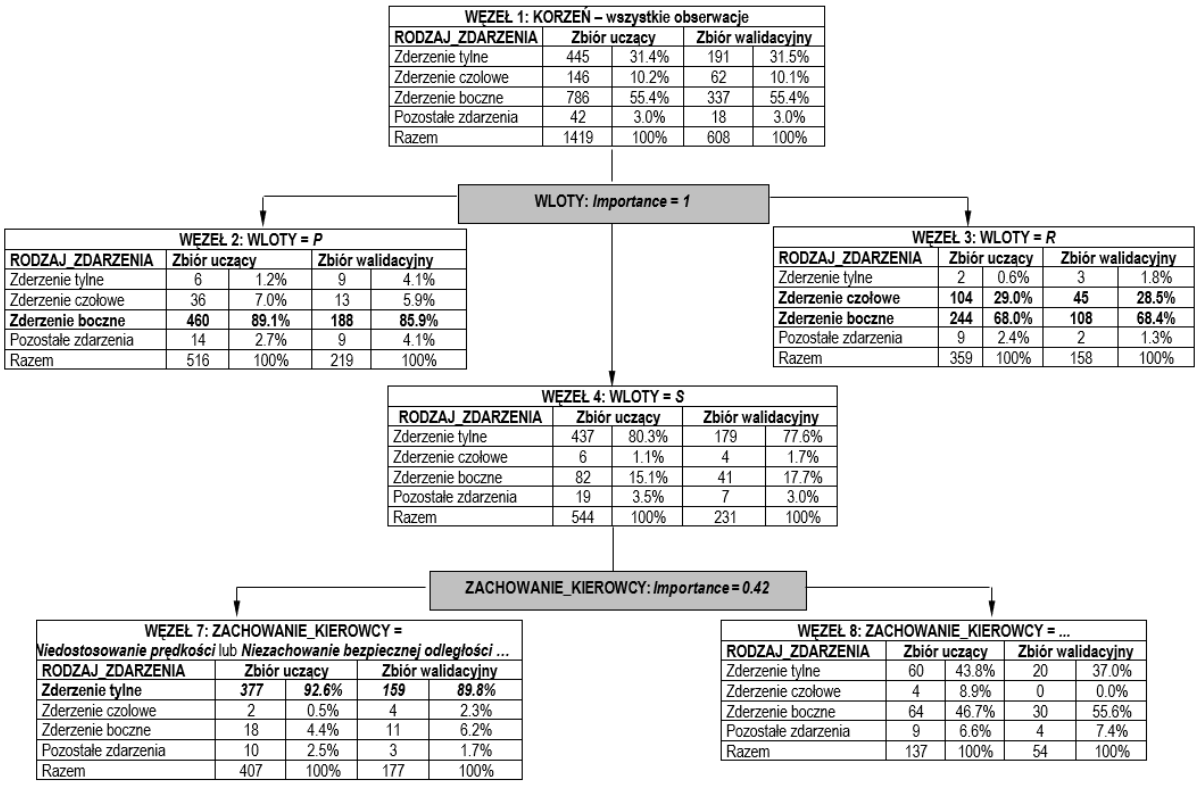
W przypadku zastosowania metody chi-kwadrat otrzymano trzy reguły. Dwie z nich są takie same jak reguły 1 i 2 w metodzie minimum entropii. Trzecia reguła została wygenerowana dla wartości cechy *WLOTY* = *S*:

Reguła 5: Jeżeli pojazdy biorące udział w zdarzeniu nadjeżdżają z tych samych wlotów i kierowca nie zachowuje bezpiecznej odległości lub kierowca nie dostosowuje prędkości do warunków ruchu, to rodzajem zdarzenia jest zderzenie tylne (ponad 90% przypadków w zbiorach walidacyjnych).

Z reguł 3, 4 i 5 wynika, że w przypadku zdarzeń na skrzyżowaniach ulicznych można agregować dwie wartości cechy ZACHOWANIE_KIEROWCY: *Niezachowanie bezpiecznej odległości między pojazdami* oraz *Niedostosowanie prędkości*.

Na rys. 7 przedstawiono przykładowe drzewo decyzyjne składające się z czterech liści, powstałe w wyniki zastosowania kryterium podziału chi-kwadrat. W prezentowanym modelu, zmienna *WLOTY* tworzy najsilniejszy związek z decyzją *RODZAJ_ZDARZENIA*. Test tożsamościowy dla tej zmiennej dzieli wszystkie obserwacje treningowe na trzy podzbiory, z których każdy reprezentuje indywidualnie testowaną wartość cechy: *P*, *R* i *S*. Na drugim poziomie drzewa znajduje się jeden prawie czysty liść zdefiniowany przez WĘZEŁ 2. Opisuje on okoliczności typu zdarzenia zderzenia bocznego, który jest typowy dla wlotów poprzecznych na skrzyżowanie (*WLOTY* = *P*). Na drugim poziomie drzewa WĘZEŁ 3 jest również liściem. Wskazuje on związek między zbliżaniem się do skrzyżowania z przeciwległych wlotów (*WLOTY* = *R*) jako okoliczność, a zderzeniami bocznymi jako skutkami zdarzeń na skrzyżowaniu, dużo rzadziej występują zderzenie czołowe. W WĘZLE 4 następuje dalszy podział drzewa prowadzący do trzeciego poziomu, a *ZACHOWANIE_KIEROWCY* jest drugą zmienną silnie powiązaną z decyzją. Na trzecim poziomie występują dwa liście. WĘZEŁ 7 jest prawie czystym liściem okoliczności zderzenia tylnego opisanych przez zmienne *WLOTY* i *ZACHOWANIE_KIEROWCY* równe odpowiednio: S i *Niedostosowanie prędkości* lub *Niezachowanie bezpiecznej odległości między pojazdami*. Liść w WĘZLE 8 nie wskazuje kategorii modalnej zmiennej *RODZAJ_ZDARZENIA* powiązanej z konkretną wartością *ZACHOWANIE_KIEROWCY*.

[Tekst alternatywny. Drzewo decyzyjne o dwóch poziomach. Na pierwszym poziomie podział zbioru danych według zmiennej *WLOTY*, na drugim poziomie – według zmiennej *ZACHOWANIE_KIEROWCY*. Drzewo ma cztery liście większościowe klasyfikujące wszystkie rodzaje zderzeń na skrzyżowaniu.]



Rys. 7. Drzewo decyzyjne dla eksperymentu nr 2 – kryterium podziału Chi-2

Eksperyment 3

Zadaniem eksperymentu jest sprawdzenie, czy istnieją reguły, które pozwolą zdiagnozować które czynniki związane z geometrią i organizacją ruchu na skrzyżowaniu, czyli czynniki które kształtuje projektant, mogą wpływać zachowanie kierującego, który brał udział w zdarzeniu.
Parametry eksperymentu:

- zmienne objaśniające: *WLOTY, RELACJE*, *SYGNALIZACJA*, *TYP_SKRZYŻOWANIA*,
- decyzja: *ZACHOWANIE_KIEROWCY*.

Wyniki otrzymane z zastosowaniem metod minimum entropii i chi-kwadrat są bardzo podobne dla różnych zestawów zbiorów uczącego i testowego. Miary oceny uzyskanych drzew na dla obu zbiorów wahały się nieznacznie:

- wartości wskaźnika *Proportion Correctly Classified* (*PCC*): 72% ÷ 76%,

- wartości na przekątnej macierzy pomyłek: *Zderzenie czołowe*: 10% ÷ 30%, *Zderzenie tylne*: 85% ÷ 90%, *Zderzenie boczne*: 85% ÷ 90%.

Najważniejszą cechą dzielącą zbiór uczący jest *WLOTY*; jej wskaźnik *Importance* ma wartość 1. Druga pod względem ważności jest cecha *RELACJE*. Jej wskaźnik ważności waha się dla obu metod i wielokrotnego powtarzania eksperymentu od 0,61 do 0,72.

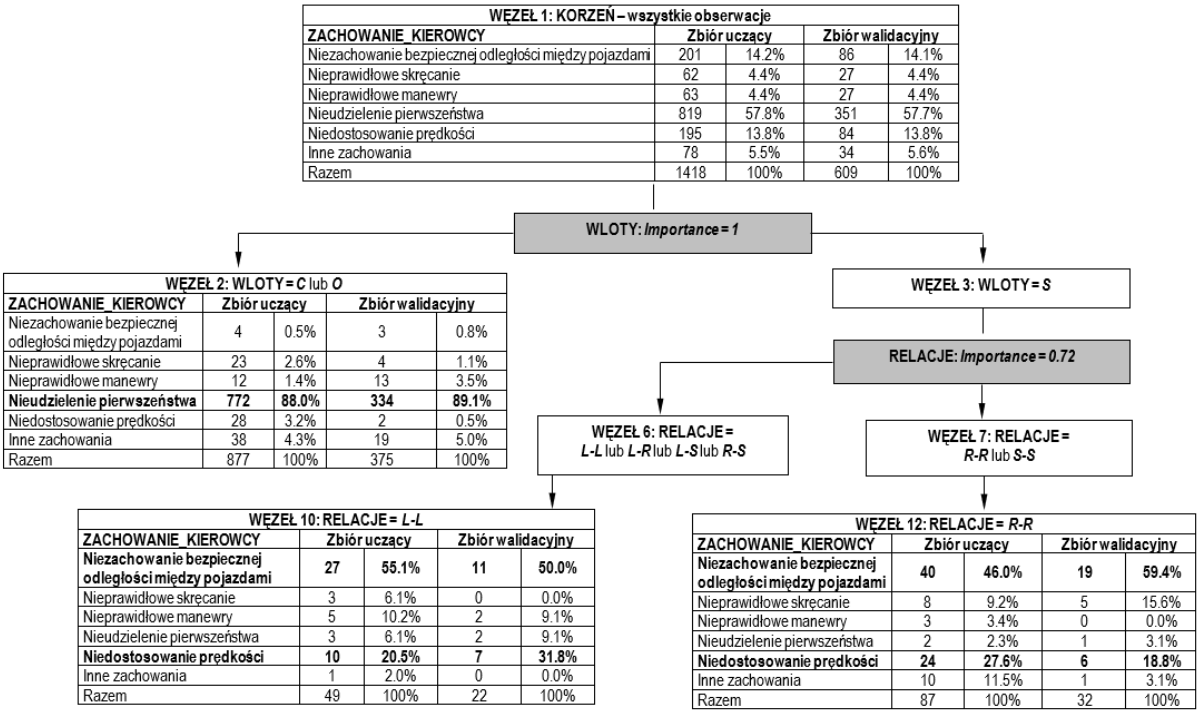
Niezależnie od zastosowanej metody poddziału drzewa uzyskano powtarzalne wyniki o takich samych regułach. Ich ilustracją jest rys. 8. Na pierwszym poziomie drzewa dla wartości cechy *WLOTY* równej *P* lub *R* powstają stosunkowo czyste liście, co pozwala na ustalenie reguły:

Reguła 1: Jeżeli kierowcy nadjeżdżają z wlotów poprzecznych lub równoległych, to przyczyną zdarzenia jest nieudzielenie pierwszeństwa przejazdu (86%-95% przypadków w zbiorach walidacyjnych).

Dla wartości cechy *WLOTY = S* następuje dalszy podział drzewa wg cechy *RELACJE*. Liście wynikowe na drugim poziomie są zanieczyszczone. Jednak dla kombinacji relacji *L-L* i *P-P* można zdefiniować regułę, w której, w przypadku skrzyżowań, jest uzasadnione połączenie dwóch wartości decyzji: *Niezachowanie bezpiecznej odległości między pojazdami*, *Niedostosowanie prędkości do warunków ruchu*):

Reguła 2: Jeżeli kierowcy nadjeżdżają z tego samego wlotu i oba pojazdy skręcają w tym samym kierunku, to przyczyną zdarzenia w większości przypadków jest niezachowanie bezpiecznej odległości między pojazdami lub niedostosowanie prędkości do warunków ruchu (nie mniej niż 70% przypadków w zbiorach walidacyjnych).

[Tekst alternatywny. Drzewo decyzyjne o dwóch poziomach. Na pierwszym poziomie podział zbioru danych według zmiennej *WLOTY*, na drugim poziomie – według zmiennej *RELACJE*. Drzewo ma trzy liście większościowe klasyfikujące rodzaje zachowań kierującego: *Nieudzielnie pierwszeństwa*, Nie*zachowanie bezpiecznej odległości między pojazdami* i *Niedostosowanie prędkości*.]



Rys. 8. Drzewo decyzyjne dla eksperymentu nr 3 – kryterium podziału Chi-2

Eksperyment 4

Zadaniem eksperymentu jest sprawdzenie czy czynniki, których projektant nie może kształtować mogą wpływać na zachowanie kierującego pojazdem biorącego udział w zdarzeniu.

Parametry eksperymentu:

- zmienne objaśniające: *STAN_NAWIERZCHNI*, *OŚWIETLENIE*, *GODZINA*, *WARUNKI_ATMOSFERYCZNE*,
- decyzja: *ZACHOWANIE_KIEROWCY*.

Drzewa otrzymane przy pomocy obu metod i wielokrotnym powtarzaniu eksperymentów mają tylko korzeń. Wskaźnik *PCC* jest niski i wynosi 58%. W przypadku drzew, które nie mają gałęzi, *PCC* jest zawsze równy odsetkowi

najliczniejszej kategorii zmiennej celu, a sam korzeń zawiera tylko informacje o rozkładzie decyzji.

4.3. Wnioski

Wykonane eksperymenty związane z analizą bezpieczeństwa ruchu drogowego pozwalają na sformułowanie wniosków, które mogą weryfikować obiegowe opinie na temat związków przyczynowo-skutkowych zdarzeń drogowych na skrzyżowaniach ulicznych.

1. Dla wypadków i kolizji na skrzyżowaniach ulicznych nie zidentyfikowano różniących się okoliczności przyczynowych, stąd zdarzenia obu statusów powinny być analizowane łącznie, aby móc zdiagnozować wzorce zdarzeń na takich skrzyżowaniach, określając ich typologię.
2. Zarówno typ skrzyżowania jak i obecność na nim sygnalizacji świetlnej nie okazały się cechami mającymi istotny wpływ na rodzaj zdarzenia drogowego czy zachowanie kierowcy biorącego udział w zdarzeniu.
3. Warunki zewnętrzne, takie jak pogoda, pora doby, warunki oświetlenia nie determinują błędnych zachowań kierowców skutkujących zdarzeniem drogowym na skrzyżowaniach.
4. W analizie bezpieczeństwa ruchu drogowego na skrzyżowaniach ulicznych następujące zachowania kierowców mogą być traktowane łącznie (agregowane): *niezachowanie bezpiecznej odległości między pojazdami* oraz *niedostosowanie prędkości do warunków ruchu*.

4.4. Uwagi końcowe

Główną zaletą drzew decyzyjnych jest prostota ich interpretacji oraz czytelność nawet dla osób, które nie posiadają eksperckiej wiedzy. Umożliwiają one sprawną implementację procesu klasyfikacji, czyli zastosowanie hipotezy uzyskanej w wyniku uczenia. Algorytmy takiego klasyfikatora działają bardzo dobrze dla zmiennych objaśniających różnych typów.

Największą wadą drzew decyzyjnych jest tendencja do nadmiernego dopasowywania, tzn. wykreowania takiego modelu, który daje bardzo dobre dopasowanie na zbiorze treningowym (uczącym), czyli takim, na podstawie którego jest budowany, ale złe dopasowanie na zbiorze zewnętrznym (testowym), tzn. nieuczestniczącym w procesie budowy drzewa, co może skutkować słabą wydajnością. Ponadto drobne zmiany w danych treningowych mogą znacząco wpłynąć na budowę drzewa decyzyjnego. Czyni to ten algorytm niestabilnym.

5. Lasy losowe

Lasy losowe (*random forests*) są uogólnieniem idei drzew decyzyjnych. Generalnie, są stosowane w zadaniach klasyfikacji, tzn. takich, w których zmienna celu ma charakter jakościowy. Lasy losowe zalicza się do procedur agregujących; ich działanie polega na klasyfikacji za pomocą grupy (rodziny) drzew decyzyjnych. Każde drzewo dostarcza unikatowy "głos" dla najbardziej popularnej kategorii określonego wektora danych (przykładu) zbioru uczącego. Końcowa decyzja jest podejmowana w wyniku głosowania większościowego nad klasami wskazanymi przez poszczególne drzewa decyzyjne.

Każde z drzew lasu losowego jest konstruowane na podstawie próby bootstrapowej, powstałej przez wylosowanie ze zwracaniem N rekordów ze zbioru uczącego o długości N. W każdym węźle podział jest dokonywany jedynie na podstawie *k* losowo wybranych cech. Ich liczba jest znacznie mniejsza od szerokości zbioru p (czyli liczby wszystkich zmiennych branych pod uwagę w badaniu): k << p; zazwyczaj k = $\sqrt{p}$. Dzięki tej własności lasy losowe bardzo dobrze się sprawdzają, gdy liczba zmiennych jest bardzo duża. Do wyboru zmiennej dzielącej zbiór danych w określonym węźle każdego drzewa lasu losowego stosuje się określone kryterium oceny (test), podobnie jak w drzewie decyzyjnym. Test polega na obliczeniu miary znaczenia w zbiorze danych węzła każdej kandydującej zmiennej wejściowej (objaśniającej) i wybraniu tej, która w danym momencie jest najważniejsza z punktu widzenia podziału zbioru. Jedną z najpopularniejszych miar jest wskaźnik Giniego.

W tym rozdziale przedstawiono wykorzystanie lasów losowych do klasyfikacji stanu zdrowia psychicznego na podstawie preferencji muzycznych. Zadanie jest omówione zgodnie z koncepcją dyplomowej pracy inżynierskiej studenta kierunku inżynieria danych, Michała Wilczyńskiego, realizowanej pod opieką autorki niniejszych materiałów dydaktycznych. Do przeprowadzenia badań wykorzystano pakiet *sckit-learn* (wersja 1.3.2) języka programowania Python, zawierający szeroki zakres algorytmów uczenia maszynowego oraz dodatkowo biblioteki *pandas* i *matplotlib*.

5.1. Dane do klasyfikacji stanu zdrowia psychicznego na podstawie preferencji muzycznych

Dane zawierają wyniki badań ankietowych przeprowadzonych w 2022 roku. Są dostępne na platformie Kaggle: (https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results/). Zbiór zawiera 736 rekordów i jest opisany przez 31 zmiennych (cech) ilościowych i jakościowych. W procesie wstępnej eksploracji i przygotowania danych do analiz, w tym czyszczenia danych, wykonano następujące operacje:

- wartości niektórych zmiennych zagregowano zgodnie z ich znaczeniem merytorycznym (zniwelowanie bardzo małych liczności niektórych kategorii),
- utworzono dwie zmienne kompozytowe, każdą z dwóch skorelowanych zmiennych jakościowych,
- poddano kategoryzacji zmienne ilościowe,
- usunięto zmienne o niskiej jakości (dużo braków, błędne wartości) lub takich, których rozkład wartości w znaczącej większości (nie mniej niż 80%) skupia się na jednej kategorii,
- usunięto sprzeczne rekordy lub zawierające wartości budzące wątpliwości (weryfikacja krzyżowa).

W wyniku uzyskano zbiór danych o długości 706 rekordów, bardziej zwarty niż pierwotny (mniejsza liczba zmiennych i bardziej zrównoważone rozkłady ich wartości), zawierający 20 zmiennych przygotowanych do budowy klasyfikatora o nazwie *Anxiety*. Ich wykaz oraz przyjmowane przez nie wartości są zestawione w tabeli 5. Ponieważ funkcja do budowania lasów losowych w Pyhtonie (*RandomForestClassifier*) akceptuje tylko zmienne mające wartości liczbowe, w przypadku zmiennych jakościowych, w kolumnie *Zakres wartości* podano numeryczne kody dla wartości opisowych tych zmiennych.

Tabela 5. Charakterystyka zmiennych modelu lasu losowego *Anxiety*

Lp.	Kod zmiennej	Znaczenie zmiennej	Zakres wartości
Zmienne wejściowe (objaśniające)			
1	Age	Wiek osoby badanej,	Liczba całkowita z przedziału [10, 89].
2	Prm_Spt	Użytkowanie Spotify jako głównej platformy streamingowej,	Yes – 1, No – 0
3	Prm_Oth	Użytkowanie innej platformy streamingowej niż Spotify,	Yes – 1, No – 0
4	Hr_P_D	Godziny spędzane w ciągu dnia na słuchaniu muzyki,	Liczba z przedzialu (0, 24]
5	Msc_Crt	Aktywna twórczość muzyczna osoby badanej (gra na instrumencie lub komponowanie muzyki)	Yes – 1, No – 0
6	Msc_Exp	Eksploracja nowych doznań muzycznych (słuchanie muzyki w języku obcym, nieopanowanym przez ankietowanego, rozwijanie horyzontów muzycznych)	Yes – 1, No – 0
7	Fv_Cls	Ulubiona muzyka klasyczna,	Yes – 1, No – 0
8	Fv_Trd	Ulubiona muzyka z gatunków tradycyjnych,	Yes – 1, No – 0
9	Fv_MdEi	Ulubiona muzyka nowoczesna i elektroniczna,	Yes – 1, No – 0
10	Fv_RpHp	Ulubiona muzyka z grupy rap i hip hop,	Yes – 1, No – 0

Lp.	Kod zmiennej	Znaczenie zmiennej	Zakres wartości
11	Fv_Wrd	Ulubiona muzyka światowa,	Yes – 1, No – 0
12	Fv_Mtl	Ulubiona muzyka metalowa,	Yes – 1, No – 0
13	Fb_Rck	Ulubiona muzyka rockowa,	Yes – 1, No – 0
14	Frq_Cls	Częstość słuchania muzyki klasycznej,	Never – 1, Rarely – 2, Sometimes – 3, Very frequently – 4
15	Frq_Trd	Częstość słuchania muzyki z gatunków tradycyjnych (Country, Folk, Gospel, Jazz, R&B)	Never – 1, Rarely – 2, Sometimes – 3, Very frequently – 4
16	Frq_MdEi	Częstość słuchania muzyki nowoczesnej i elektronicznej (EDM, Loft, Video game music)	Never – 1, Rarely – 2, Sometimes – 3, Very frequently – 4
17	Frq_RpHp	Częstość słuchania muzyki z grupy rap i hip hop,	Never – 1, Rarely – 2, Sometimes – 3, Very frequently – 4
18	Frq_Wrd	Częstość słuchania muzyki światowej (Latin, Pop, K pop)	Never – 1, Rarely – 2, Sometimes – 3, Very frequently – 4
19	Frq_Mtl	Częstość słuchania muzyki metalowej,	Never – 1, Rarely – 2, Sometimes – 3, Very frequently – 4
20	Frq_Rck	Częstość słuchania muzyki rockowej.	Never – 1, Rarely – 2, Sometimes – 3, Very frequently – 4
Zmienna wyjściowa (objaśniana, decyzja)			
21	Anxiety		No threat – brak znaczącego zagrożenia chorobą Threat – znaczące zagrożenie chorobą

Cecha *Anxiety o*kreślaja stan zdrowia psychicznego osoby ankietowanej. Respondenci sami oceniali swoje zdrowie psychiczne w skali od 0 do 10, gdzie 0 wskazuje na brak objawów danej przypadłości, podczas gdy 10 oznacza zaawansowane objawy. Wartości przez nich podawane nie zostały zweryfikowane przez profesjonalistów w dziedzinie psychologii. Ze względu na informację, jaką prezentują te zmienne, zostały one kandydatami do klasyfikacji przez lasy losowe. Wartości decyzji zostały przekształcone – kategorie reprezentowane przez numery 0, 1, 2, 3, 4 oraz 5 zastąpiono przez 0 (*No Threat*; brak znaczącego zagrożenia daną chrobą psychiczną), natomiast kategorie 6, 7, 8, 9, 10 zastąpiono przez 1 (*Threat*; znaczące zagrożenie wystąpienia choroby). Zaproponowany sposób agregacji uwzględnia fakt, że dla danych o niewielkiej liczbie rekordów i wielu możliwych

klasach wyjściowych (w tym przypadku aż 11) trudno jest zbudować skuteczny model (klasyfikacyjny

Pierwotne dane są dzielone w taki sposób, że 30% (212 rekordów) stanowi zbiór testowy, a 70% (494 rekordy) – zbiór uczący. Zbiór uczący jest używany do trenowania modelu, a zbiór testowy – do oceny jakości modelu i weryfikacji, jak radzi on sobie z nieznanymi wcześniej danymi.

5.2. Budowa lasu losowego dla klasyfikacji stanu zdrowia psychicznego

Przyjęto, że kryterium podziału zbioru w węzłach każdego drzewa jest wskaźnik Giniego. Jest on miarą heterogeniczności (zanieczyszczenia) węzła ze względu na zmienną objaśnianą *Y* (decyzję). Na danym poziomkie w drzewie klasyfikacyjnym do podziału zbioru jest wybierana ta cecha X_i, dla której wskaźniik $Gini(Y|X_i)$ jest najmniejszy.

$$Gini(Y|X_i) = 1 - \sum_{j \in I} \frac{n_j}{n} Gini(X_{ij}|Y)$$
$$Gini(X_{ij}|Y) = \sum_{d \in C} p^2\ (d|X_{ij}) \quad (5)$$

gdzie:
- *I* jest zbiorem kategorii (wartości) *i*-tej zmiennej wejściowej,
- *n* jest liczbą obserwacji w zbiorze podlegającym podzieleniu,
- n_j jest liczbą obserwacji, dla których zmienna X_i ma wartość *j*,
- $p(d|X_{ij})$ jest zaobserwowanym prawdopodobieństwem kategorii *d* zmiennej objaśnianej *Y* pod warunkiem, że zmienna X_i ma wartość *j*.

Budowa lasu losowego wymaga określenia następujących wartości jego hiperparametrów:
- *max_depth* – maksymalna liczba poziomów (głębokość) drzewa decyzyjnego,
- *min_samples_split* – minimalna liczba obserwacji wymaganych do podzielenia zbioru w węźle drzewa,
- *class_weight* – para wag dla poszczególnych kategorii *Treat* oraz *No Threat* zmiennej objaśnianej, wprowadzona ze względu na znaczące różnice w liczebności tych kategorii (niezbalansowanie zbioru danych).

W poszukiwaniu najlepszego wyniku klasyfikacji, przeprowadzone eksperymenty dla różnych wartości hiperparametrów. Przyjęto, że sprawdzane wartości *max_depth* to liczby całkowite z przedziału [4, 10]. Przedział wartości dla *min_samples_split*

obejmował liczby całkowite z zakresu [10, 40]. W procesie poszukiwań, wartości wag dla *Threat* i *No Threat* są wyrażone zależnością:

$$weight_i = \frac{n}{n_c * n_i} \tag{6}$$

gdzie:
- n jest liczbą obserwacji w zbiorze uczącym,
- n_c jest liczbą kategorii zmiennej objaśnianej,
- n_i jest liczbą obserwacji z kategorią i w zbiorze uczącym,

i zmieniają się wraz ze zmianą wartości n_{Threat} od 1 do n-1 z jednoczesną zmianą wartości $n_{No\ Threat}$ od n-1 do 1. To oznacza, że dla zbioru uczącego o liczności 494, pary wag w procesie iteracyjnym są wyznaczone dla następujących par liczności kategorii *Threat* i *No Threat*: (1, 493), (2, 492), ..., (493, 1). Wyszukiwanie hiperparametrów wykonano poprzez sprawdzanie ich wszystkich możliwych kombinacji.
Zestaw hiperparametrów uznawano za najlepszy, jeżeli spełniony został szereg warunków:

- sprawdzenie, z pomocą metody *score*, czy las losowy nie dopasował się zbyt mocno do danych uczących,
- miary TPR (*True Positive Rate*) oraz TNR (*True Negative Rate*) są większe niż 0,5, zgodnie z założeniem, że docelowy model powinien mieć minimum 50% skuteczności w klasyfikacji kategorii pozytywnych (w tym przypadku *Threat*) oraz negatywnych (*No Threat*).
- TPR powinno być większe niż TNR, dzięki czemu model uzyskuje większą skuteczność w klasyfikacji sytuacji, będących poważnym zagrożeniem dla zdrowia
- model w bieżącym kroku iteracyjnym ma osiągać lepsze wyniki niż model z poprzednich kroków uznawany za najskuteczniejszy.

Najlepsze wyniki modelu *Anxiety* uzyskano dla następujących wartości hiperparametrów: *max_depth* = 7, *min_samples_split* = 14, *class_weight* (*Threat*, *No Threat*) = (1,43605; 0,76706).

Tabela 6 zawiera informacje o jakości klasyfikacji wybranego modelu *Anxiety*. Wskaźnik *score* dla danych treningowych na poziomie 0,842 oznacza, że osiągnął on aż 84,2% skuteczności w klasyfikacji dla danych uczących. Skuteczność dla danych testowych wyniosła 63,7%. Warto zauważyć, że skuteczność poprawnego wskazywania ryzyka wystąpienia stanów lękowych w zbiorze testowym jest na poziomie 70,6% i jest ona o wiele większa niż skuteczność wskazywania braku

zagrożenia zdrowia 53,5%. Czyni to ten model skutecznym klasyfikatorem zagrożeń zdrowia.

Tabela 6. Wyniki klasyfikacji lasu losowego dla modelu *Anxiety*

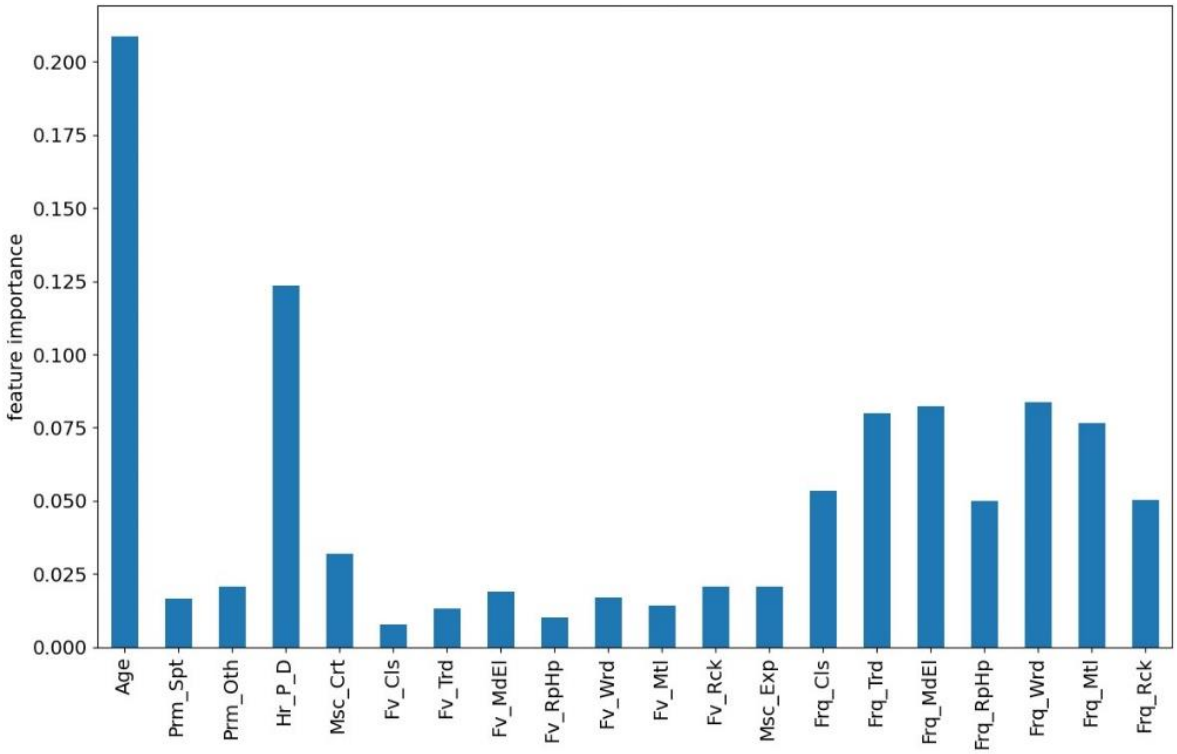
Miara	Rodzaj zbioru	
	Uczący	Testowy
Liczebność zbioru	494	212
Score [%]	0,842	0,637
Liczba poprawnie sklasyfikowanych wartości *Threat*	253	89
Odsetek poprawnie sklasyfikowanych wartości *Threat*	83,22%	70,63%
Liczba poprawnie sklasyfikowanych wartość *No Threat*	163	46
Odsetek niepoprawnie sklasyfikowanych wartości *No Threat* [%]	85,79%	53,49%
Liczba niepoprawnie sklasyfikowanych wartości *Threat*	51	37
Liczba niepoprawnie sklasyfikowanych wartość *No Threat*	27	40

W tabeli 7 zawarto wartości miary *importance* wyznaczone dla zmiennych użytych w procesie modelowania. Ich ilistracja graficzne jest przedstawiona na rys. 9. Najbardzoej znaczącymi cechami okazały się kolejno: wiek osoby badanej (*Age*) oraz czas, który dziennie przeznacza ona na słuchaniu muzyki (*Hr_P_D*). Cechy dotyczące częstości słuchania konkretnych grup muzycznych, które mają największy udział w klasyfikacji, odnoszą się do: muzyki tradycyjnej, nowoczesnej muzyki elektronicznej, muzyki światowej oraz metalowej (*Froq_Trd*, *Frq_MdEl*, *Frq_Wrd* oraz *Frq_Mtl*). Ich ważność jest na podobnym poziomie, przy czym najistotniejszą z nich okazała się częstość słuchania przez ankietowanego muzyki z kategorii „muzyka światowa” (*Frq_Wrd*).

Tabela 7. Ważność cech w budowaniu modelu *Anxiety*

Cecha	Importance	Cecha	Importance	Cecha	Importance
Age	0,208685	Fv_Cls	0,007597	Frq_Cls	0,053335
Prm_Spt	0,016608	Fv_Trd	0,013328	Frq _Trd	0,080060
Prm_Oth	0,020627	Fv_MdEl	0,018797	Frq _MdEl	0,082267
Hr_P_D	0,123608	Fv_RpHp	0,009935	Frq _RpHp	0,050108
Msc_Crt	0,031908	Fv_Wrd	0,016978	Frq _Wrd	0,083688
Msc_Exp	0,020761	Fv_Mtl	0,014132	Frq _Mtl	0,076378
		Fv_Rck	0,020827	Frq _Rck	0,050373

[Tekst alternatywny. Wykres słupkowy ilustrujący ważność cech modelu *Anxiety* klasyfikującego zagrożenie chorobą. Na osi poziomej jest odłożonych dwadzieścia cech biorących udział w budowie modelu lasu losowego. Na osi pionowej – wartość wskaźnika *importance*. Dyskusja rysunku jest w tekście opracowania.]



Rys. 9. Ważność cech dla modelu *Anxiety*

5.3. Wnioski

Najważniejszymi cechami przy klasyfikacji stanu psychicznego osoby okazały się wiek osoby badanej oraz czas, jaki spędza ona na słuchaniu muzyki. Wartości miary *importance* wskazują, że częstotliwość słuchania gatunków muzycznych ma również stosunkowo duży wpływ na klasyfikację. Eksperyment wykazał także, że ulubiony gatunek muzyczny, aktywność muzyczna osoby badanej (*Music creator*), podejście do odkrywania muzyki (*Explores music*) oraz preferowane źródło muzyki (*Streaming service*) nie mają dużego wpływu na klasyfikację.

Model *Anxiety* uzyskał wysoką skuteczność w klasyfikacji dla zbioru treningowego, co może być wynikiem nadmiernego dopasowania do danych. Skuteczność dla zbioru testowego na poziomie 70,6% w klasyfikacji osób, u których występuje ryzyko wystąpienia zaburzeń lękowych sugeruje, że model ten ma potencjał w praktycznym zastosowaniu. Warto podjąć dodatkowe kroki w celu jego dalszej optymalizacji, aby

podnieść skuteczność w klasyfikacji osób, u których nie występuje ryzyko wystąpienia zaburzenia.

5.4. Uwagi końcowe

Lasy losowe mają identyczne zalety jak drzewa decyzyjne. Algorytm takich klasyfikatorów ma jednak kilka dodatkowych atutów. Jednym z nich jest o wiele większa precyzja niż pojedyncze drzewo decyzyjne. Dzięki temu jest to bardzo dobry model predykcyjny. Jako że składa się on z wielu drzew decyzyjnych, których wyniki są agregowane, jest on bardziej odporny na nadmierne przeuczanie.
Główną wadą lasów losowych jest to, że uczą się one wolniej i zajmują więcej pamięci komputera niż drzewa decyzyjne. Kolejną wadą tego modelu jest fakt, że bardzo trudno (czasem niemożliwe) jest wygenerowanie reguł decyzyjnych, tak jak w pojedynczym drzewie. Modele lasów losowych są przez to trudniejsze w interpretacji niż pojedyncze klasyfikatory drzewiaste i traktuje się je więc przed wszystkim jako narzędzie klasyfikacji (prognozowanie wartości cech jakościowych),

6. Literatura

1. Gonzales A., Guruswamy G., Smith S.R. (2023). Synthetic data in health care: A narrative review. PLOS Digit Health 2(1): e0000082. https://doi.org/10.1371/journal.pdig.0000082.
2. Major H., Nowakowska M. (2003). Wykorzystanie techniki drzew decyzyjnych w analizach brd na skrzyżowaniach ulicznych. X Jubileuszowe SASForum w Polsce, Warszawa, 2-4.04.2003.
3. Major H., Nowakowska M. (2006). Cause-effect relationships of road incidents using decision tree method. ARCHIVES OF CIVIL ENGINEERING, LII, 3, Warszawa, 2006, s. 641-656.
4. Nowakowska M. (2009). Wykrywanie wzorców zagrożeń w interakcjach transportu drogowego z ruchem pieszym z wykorzystaniem analizy koszykowej. II Międzynarodowa Konferencja „Problemy Eksploatacji I Zarządzania Zrównoważonym Transportem" organizowana przez: Politechnika Świętokrzyska, Wydział Mechatroniki I Budowy Maszyn oraz Polskie Naukowo-Techniczne Towarzystwo Eksploatacyjne, Kielce – Wólka Milanowska, 15 -17 września 2009.
5. Savaré L., Ieva F., Corrao G., & Lora A. (2022). Mining and evaluation of patients' diagnostic-therapeutic paths through state sequences analysis. arXiv. https://arxiv.org/abs/2209.04384
6. Wilczyński M. (2024). Wykorzystanie lasów losowych w klasyfikacji stanu zdrowia psychicznego na podstawie preferencji muzycznych. Dyplomowa praca inżynierska na kierunku studiów I stopnia inżynieria danych, opieka: Nowakowska M. (WZiMK), Politechnika Świętokrzyska.