

Data oddania: _____

Ocena: _____

Julia Szymańska 224441
Przemysław Zdrzałik 224466

Projekt 2. Podsumowania lingwistyczne relacyjnych baz danych

1. Cel

Celem projektu jest stworzenie aplikacji pozwalającej na opisanie językiem naturalnym zbioru danych zawierających liczbowe informacje o ponad 3 milionach wypadków samochodowych w 49 stanach Zjednoczonych Stanów Ameryki, mających miejsce od lutego 2016 do grudnia 2020[4]. Zbiór danych składa się z 47 kolumn. W celu wykonania opisu danych językiem naturalnym wykorzystamy metody logiki rozmytej. Logika rozmyta pozwala na opisanie wartości zapisanych językiem naturalnym za pomocą zrozumiałych określeń jak: mało, dużo, około połowy.

2. Charakterystyka podsumowywanej bazy danych














W programie został użyty zbiór danych[4] znajdujący się w pliku CSV, który został przekształcony w bazę danych.

Zbiór danych zawiera informacje o ponad 3 milionach wypadków samochodowych w 49 stanach Zjednoczonych Stanów Ameryki, mających miejsce od lutego 2016 do grudnia 2020. Spośród 47 kolumn znajdujących się w zbiorze danych, wybraliśmy następujące 11 kolumn:

- Czas rozpoczęcia - Start_Time - czas rozpoczęcia się wypadku w lokalnej strefie czasowej, przyjmuje wartości od 8 lutego 2016, do 31 grudnia 2020. Wartość kolumny zostanie zamieniona na wartość całkowitą oznaczającą liczbę sekund od początku 1970 roku.

- Czas zakończenia - End_Time - czas zakończenia się wypadku w lokalnej strefie czasowej, przyjmuje wartości od 8 lutego 2016, do 1 stycznia 2021. Wartość kolumny zostanie zamieniona na wartość całkowitą oznaczającą liczbę sekund od początku 1970 roku.
- Odległość - Distance - długość odcinka ulicy wyrażony w milach, na którego miał wpływ wypadek. Przyjmuje wartości zmiennoprzecinkowe od 0 do 334, gdzie zdecydowana większość danych mieści się w przedziale od 0.00 do 4.00.
- Temperatura - Temperature - temperatura powietrza wyrażona w Fahrenheit'ach, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od -16.00 do 104.00. Temperature można opisać jako bardzo zimną, zimną, umiarkowaną, ciepłą, bardzo ciepłą. Oczywiście jest to opis subiektywny.
- Temperatura odczuwalna - Wind_Chill - temperatura odczuwalna wyrażona w Fahrenheit'ach, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od -16.00 do 101.00. Temperaturę odczuwalną można opisać tak samo jak temperaturę.
- Wilgotność - Humidity - wilgotność powietrza wyrażona w procentach w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 4.00 do 100.00.
- Ciśnienie - Pressure - ciśnienie powietrza wyrażone w inches, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 27.00 do 32.00. Ciśnienie można opisać jako wysokie, umiarkowane lub niskie.
- Widoczność - Visibility - widoczność wyrażona w milach, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 0.00 do 12.00. Widoczność można opisać jako dobrą, ograniczoną, słabą.
- Prędkość wiatru - Wind_Speed - prędkość wiatru wyrażona w milach na godzinę, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 0.00 do 40.00. Wiatr można opisać jako słaby, umiarkowany, silny.
- Ilość opadów - Precipitation - ilość opadów wyrażona w inches, w momencie, gdy zdarzył się wypadek. Jeśli opady nie występowały to kolumna przyjmuje wartość nan. Przyjmuje wartości zmiennoprzecinkowe od 0.00 do 0.50.

Atrybutom nadawane są opisane zwyczajowe wartości lingwistyczne ze względu na zwiększenie przystępności i ułatwienie szybkiego zrozumienia atrybutu przez człowieka, kiedy ten atrybut nie musi być dokładnie opisany. Przykładowo temperatura, mimo że rozumiała dla człowieka w postaci liczbowej, jest łatwiejsza do szybszego zrozumienia w postaci tekstowej, a dla ludzi nie ma dużego znaczenia czy temperatura różni się o 1 czy 2 stopnie, wystarczy opisać ją słownie tak jak wcześniej podaliśmy jako bardzo zimną, zimną, umiarkowaną, ciepłą, bardzo ciepłą.

accidents		
  id		bigint
 severity		smallint
 start_time	timestamp with time zone	
 end_time	timestamp with time zone	
 distance		double precision
 temperature		double precision
 wind_chill		double precision
 humidity		double precision
 pressure		double precision
 visibility		double precision
 wind_speed		double precision
 precipitation		double precision

Powered by yFiles

Rysunek 1. Tabela reprezentująca omawiane dane wykonana w DBMS PostgreSQL

3. Atrybuty i liczności obiektów wyrażone zmiennymi lingwistycznymi

Poniżej zostaną przedstawione zmienne lingwistyczne dla jedenastu atrybutów z bazy danych wraz z przypisanymi etykietami w formie funkcji przynależności oraz wzorów analitycznych.

Na podstawie znajdujących się w bazie danych pól Czas rozpoczęcia (Start_Time) oraz Czas zakończenia (End_Time) zostanie obliczony Czas utrudnień w ruchu drogowym (Duration) spowodowanych przez wypadek według wzoru:

$$Duration = End_Time - Start_Time \quad (1)$$

Przedstawienie Czasu utrudnień w ruchu drogowym (Duration) spowodowanych przez wypadek jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: krótki, średni, długi.

$$\mu_{czasTrwaniaPonizejGodziny}(x) = \begin{cases} \frac{1-x}{1} & \text{dla } 0 < x \leq 1 \end{cases} \quad (2)$$

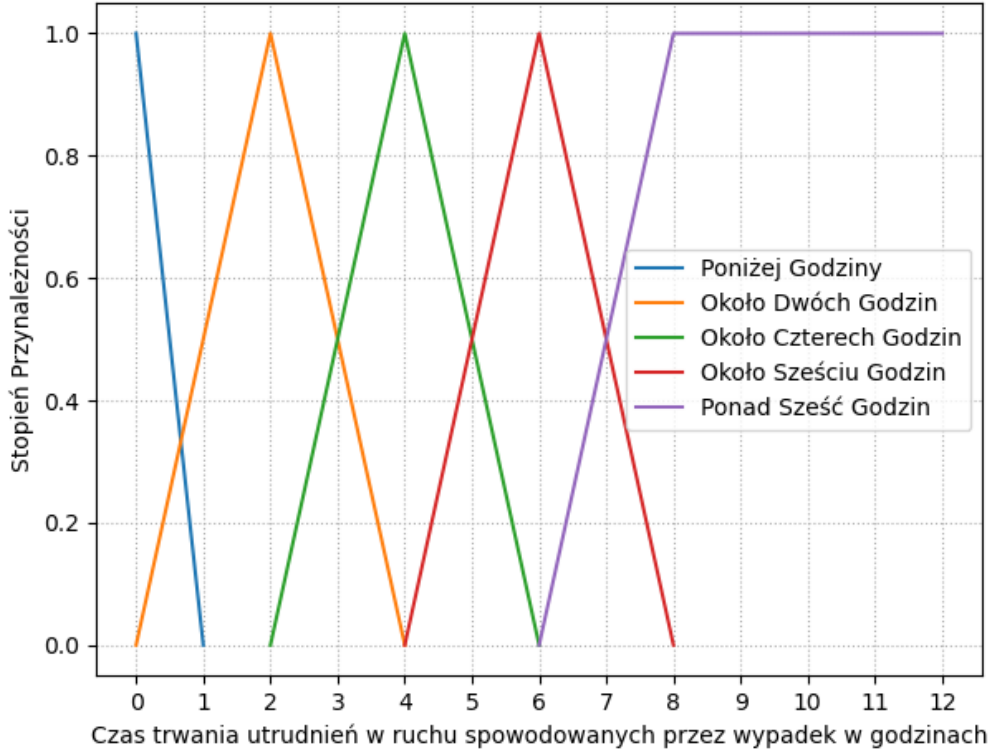
$$\mu_{czasTrwaniaOkoloDwochGodzin}(x) = \begin{cases} \frac{x}{2} & \text{dla } 0 < x \leq 2 \\ \frac{4-x}{2} & \text{dla } 2 < x \leq 4 \end{cases} \quad (3)$$

$$\mu_{czasTrwaniaOkoloCzterechGodzin}(x) = \begin{cases} \frac{x-2}{2} & \text{dla } 2 < x \leq 4 \\ \frac{6-x}{2} & \text{dla } 4 < x \leq 6 \end{cases} \quad (4)$$

$$\mu_{czasTrwaniaOkoloSzesciuGodzin}(x) = \begin{cases} \frac{x-4}{2} & \text{dla } 4 < x \leq 6 \\ \frac{8-x}{2} & \text{dla } 6 < x \leq 8 \end{cases} \quad (5)$$

$$\mu_{czasTrwaniaOPonadSzescGodzin}(x) = \begin{cases} \frac{x-6}{2} & \text{dla } 6 < x \leq 8 \\ 1 & \text{dla } 8 \leq x \end{cases} \quad (6)$$

gdzie: $\mu_{czasTrwaniaKrotki}$, $\mu_{czasTrwaniaSredni}$, $\mu_{czasTrwaniaDlugi}$ - funkcje przynależności, x - czas trwania wypadku.



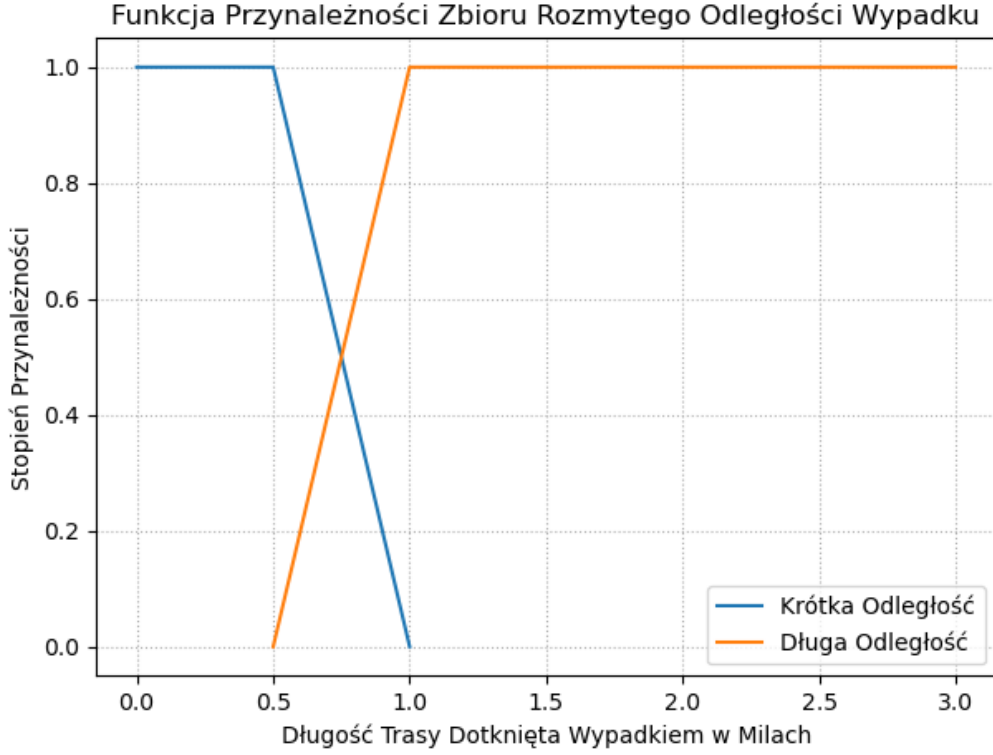
Rysunek 2. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej czas utrudnień w ruchu drogowym (Duration) spowodowanych przez wypadek.

Przedstawienie odległości jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: krótki, długi.

$$\mu_{OdlegloscKrotki}(x) = \begin{cases} 1 & \text{dla } x \leq 0.5 \\ \frac{1-x}{0.5} & \text{dla } 0.5 < x \leq 1 \end{cases} \quad (7)$$

$$\mu_{OdlegloscDlugi}(x) = \begin{cases} \frac{x-0.5}{0.5} & \text{dla } 0.5 < x \leq 1 \\ 1 & \text{dla } 1 \leq x \end{cases} \quad (8)$$

gdzie: $\mu_{OdlegloscKrotki}$, $\mu_{OdlegloscDlugi}$ - funkcje przynależności, x - odległość.



Rysunek 3. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej odległości.

Przedstawienie temperatury oraz temperatury odczuwalnej jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: bardzo zimno, zimno, umiarkowanie, ciepło, bardzo ciepło.

$$\mu_{temperaturaBardzoZimno}(x) = \begin{cases} 1 & \text{dla } x \leq 14 \\ \frac{23-x}{9} & \text{dla } 14 < x \leq 23 \end{cases} \quad (9)$$

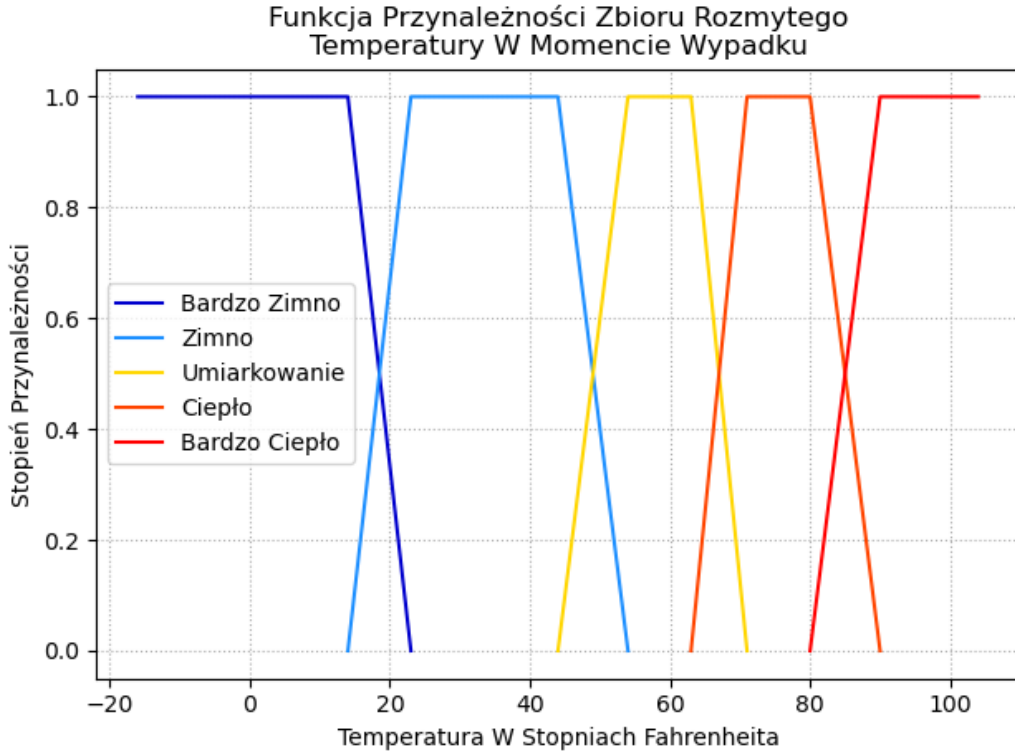
$$\mu_{temperaturaZimno}(x) = \begin{cases} \frac{x-14}{9} & \text{dla } 14 < x \leq 23 \\ 1 & \text{dla } 23 < x < 44 \\ \frac{54-x}{10} & \text{dla } 44 < x \leq 54 \end{cases} \quad (10)$$

$$\mu_{temperaturaUmiarkowanie}(x) = \begin{cases} \frac{x-44}{10} & \text{dla } 44 < x \leq 54 \\ 1 & \text{dla } 54 < x < 63 \\ \frac{71-x}{8} & \text{dla } 63 < x \leq 71 \end{cases} \quad (11)$$

$$\mu_{temperaturaCieplo}(x) = \begin{cases} \frac{x-63}{8} & \text{dla } 63 < x \leq 71 \\ 1 & \text{dla } 71 < x < 80 \\ \frac{90-x}{10} & \text{dla } 80 < x \leq 90 \end{cases} \quad (12)$$

$$\mu_{temperaturaBardzoCieplo}(x) = \begin{cases} \frac{x-80}{10} & \text{dla } 80 < x \leq 90 \\ 1 & \text{dla } 90 \leq x \end{cases} \quad (13)$$

gdzie: $\mu_{temperaturaBardzoZimna}$, $\mu_{temperaturaZimna}$, $\mu_{temperaturaUmiarkowana}$, $\mu_{temperaturaCiepła}$, $\mu_{temperaturaBardzoCiepła}$ - funkcje przynależności, x - temperatura, temperatura odczuwalna.



Rysunek 4. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej temperatury.

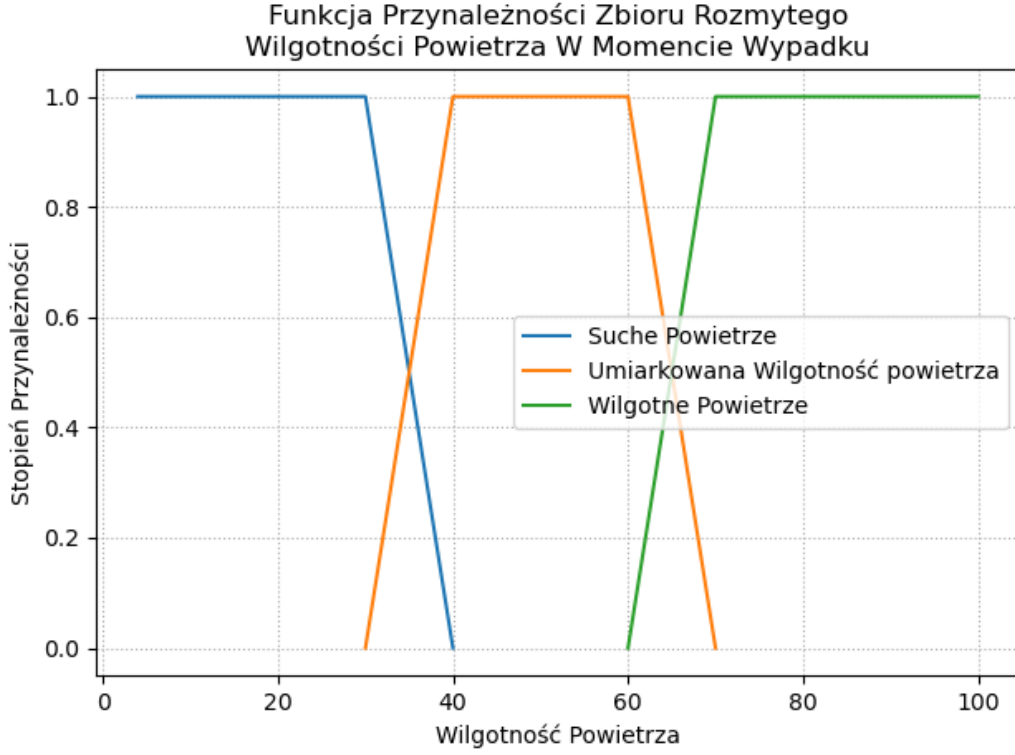
Przedstawienie wilgotności jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: suche, umiarkowane, wilgotne.

$$\mu_{wilgotnoscSuche}(x) = \begin{cases} 1 & \text{dla } x \leq 30 \\ \frac{30-x}{9} & \text{dla } 30 < x \leq 40 \end{cases} \quad (14)$$

$$\mu_{wilgotnoscUmiarkowane}(x) = \begin{cases} \frac{x-30}{10} & \text{dla } 30 < x \leq 40 \\ 1 & \text{dla } 40 < x < 60 \\ \frac{70-x}{10} & \text{dla } 60 < x \leq 70 \end{cases} \quad (15)$$

$$\mu_{wilgotnoscWilgotne}(x) = \begin{cases} \frac{x-60}{10} & \text{dla } 60 < x \leq 70 \\ 1 & \text{dla } 70 \leq x \end{cases} \quad (16)$$

gdzie: $\mu_{wilgotnoscSuche}$, $\mu_{wilgotnoscUmiarkowane}$, $\mu_{wilgotnoscWilgotne}$ - funkcje przynależności, x - wilgotność.



Rysunek 5. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej wilgotności.

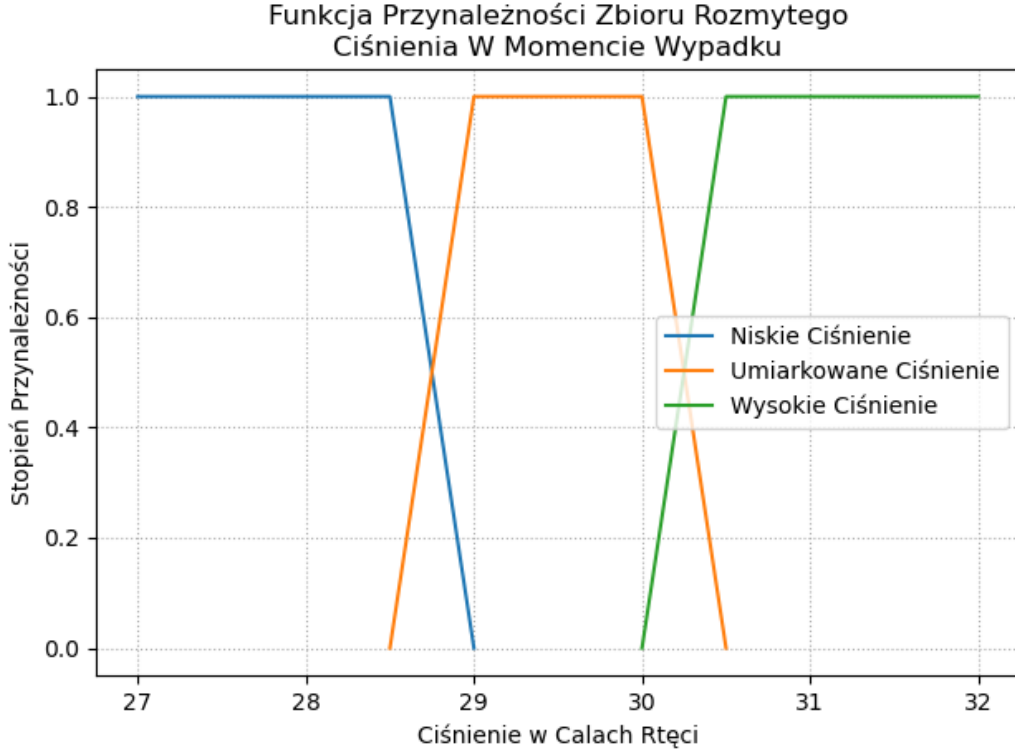
Przedstawienie ciśnienia jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: niskie, umiarkowane, wysokie.

$$\mu_{cisnienieNiskie}(x) = \begin{cases} 1 & \text{dla } x \leq 28.5 \\ \frac{28.5-x}{0.5} & \text{dla } 28.5 < x \leq 29 \end{cases} \quad (17)$$

$$\mu_{cisnienieUmiarkowane}(x) = \begin{cases} \frac{x-28.5}{0.5} & \text{dla } 28.5 < x \leq 29 \\ 1 & \text{dla } 29 < x < 30 \\ \frac{30.5-x}{0.5} & \text{dla } 30 < x \leq 30.5 \end{cases} \quad (18)$$

$$\mu_{cisnienieWysokie}(x) = \begin{cases} \frac{x-30}{1.5} & \text{dla } 30.5 < x \leq 32 \\ 1 & \text{dla } 32 \leq x \end{cases} \quad (19)$$

gdzie: $\mu_{cisnienieNiskie}$, $\mu_{cisnienieUmiarkowane}$, $\mu_{cisnienieWysokie}$ - funkcje przynależności, x - ciśnienie.



Rysunek 6. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej ciśnienia.

Przedstawienie widoczności jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: brak, słaba, ograniczona, dobra.

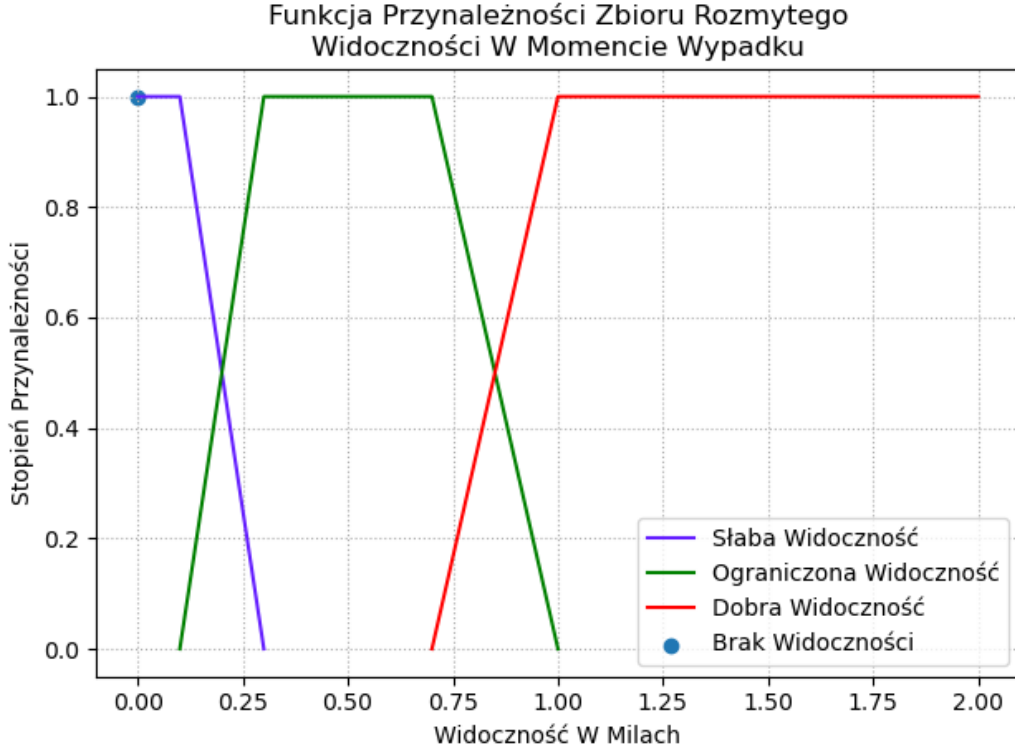
$$\mu_{widocznoscBrak}(x) = 1 \quad dla \quad x = 0 \quad (20)$$

$$\mu_{widocznoscSlaba}(x) = \begin{cases} 1 & dla \quad x \leq 0.1 \\ \frac{0.1-x}{0.2} & dla \quad 0.1 < x \leq 0.3 \end{cases} \quad (21)$$

$$\mu_{widocznoscOgraniczona}(x) = \begin{cases} \frac{x-0.1}{0.2} & dla \quad 0.1 < x \leq 0.3 \\ 1 & dla \quad 0.3 < x < 0.7 \\ \frac{1-x}{0.3} & dla \quad 0.7 < x \leq 1 \end{cases} \quad (22)$$

$$\mu_{widocznoscDobra}(x) = \begin{cases} \frac{x-0.7}{0.3} & dla \quad 0.7 < x \leq 1 \\ 1 & dla \quad 1 \leq x \end{cases} \quad (23)$$

gdzie: $\mu_{widocznoscBrak}$, $\mu_{widocznoscSlaba}$, $\mu_{widocznoscOgraniczona}$, $\mu_{widocznoscDobra}$ - funkcje przynależności, x - widoczność.



Rysunek 7. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej widoczności.

Przedstawienie predkości wiatru jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: brak, słaby, umiarkowany, silny, wicher, huragan.

$$\mu_{wiatrBrak}(x) = 1 \quad \text{dla } x = 0 \quad (24)$$

$$\mu_{wiatrSłaby}(x) = \begin{cases} 1 & \text{dla } x \leq 3 \\ \frac{3-x}{0.5} & \text{dla } 3 < x \leq 3.5 \end{cases} \quad (25)$$

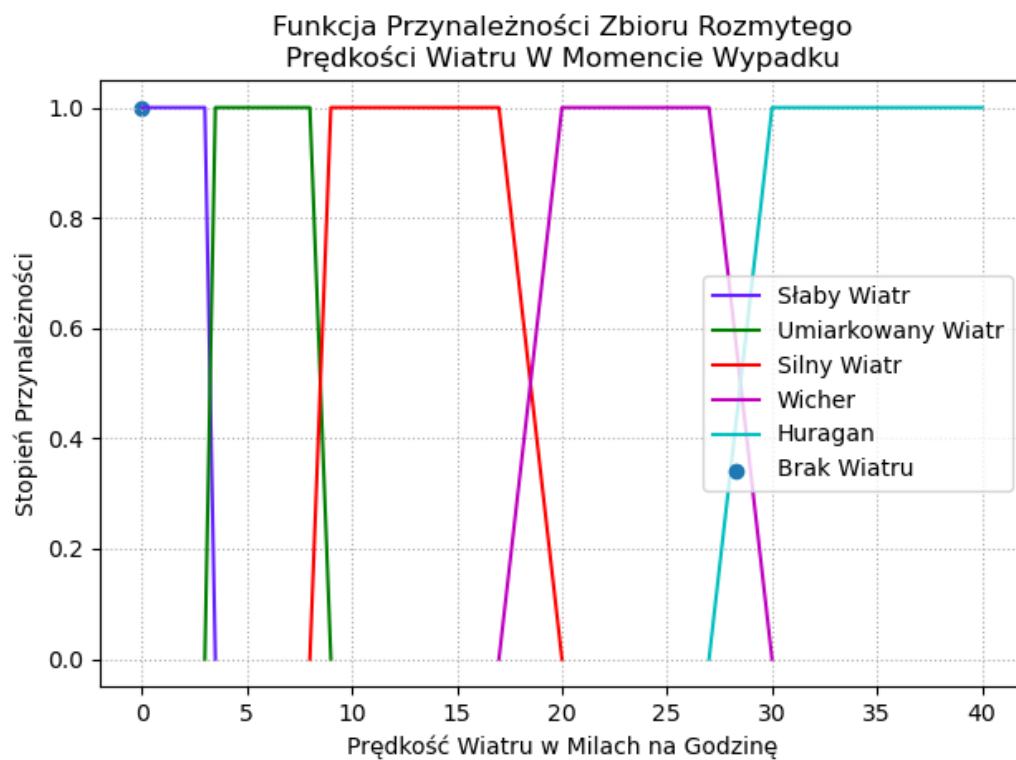
$$\mu_{wiatrUmiarkowany}(x) = \begin{cases} \frac{x-3}{0.5} & \text{dla } 3 < x \leq 3.5 \\ 1 & \text{dla } 3.5 < x < 8 \\ \frac{9-x}{1} & \text{dla } 8 < x \leq 9 \end{cases} \quad (26)$$

$$\mu_{wiatrSilny}(x) = \begin{cases} \frac{x-8}{1} & \text{dla } 8 < x \leq 9 \\ 1 & \text{dla } 9 < x < 17 \\ \frac{20-x}{3} & \text{dla } 17 < x \leq 20 \end{cases} \quad (27)$$

$$\mu_{wiatrWicher}(x) = \begin{cases} \frac{x-17}{3} & \text{dla } 17 < x \leq 20 \\ 1 & \text{dla } 20 < x < 27 \\ \frac{30-x}{3} & \text{dla } 27 < x \leq 30 \end{cases} \quad (28)$$

$$\mu_{wiatrHuragan}(x) = \begin{cases} \frac{x-40}{10} & \text{dla } 30 < x \leq 40 \\ 1 & \text{dla } 40 \leq x \end{cases} \quad (29)$$

gdzie: $\mu_{wiatrBrak}$, $\mu_{wiatrSlaby}$, $\mu_{wiatrUmiarkowany}$, $\mu_{wiatrSilny}$, $\mu_{wiatrWicher}$, $\mu_{wiatrHuragan}$
- funkcje przynależności, x - prędkość wiatru.



Rysunek 8. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej prędkości wiatru.

Przedstawienie opadów jako zmiennej lingwistycznej. Do zmiennej lingwistycznej zostały dopasowane etykiety: brak, niewielkie, umiarkowane, duże.

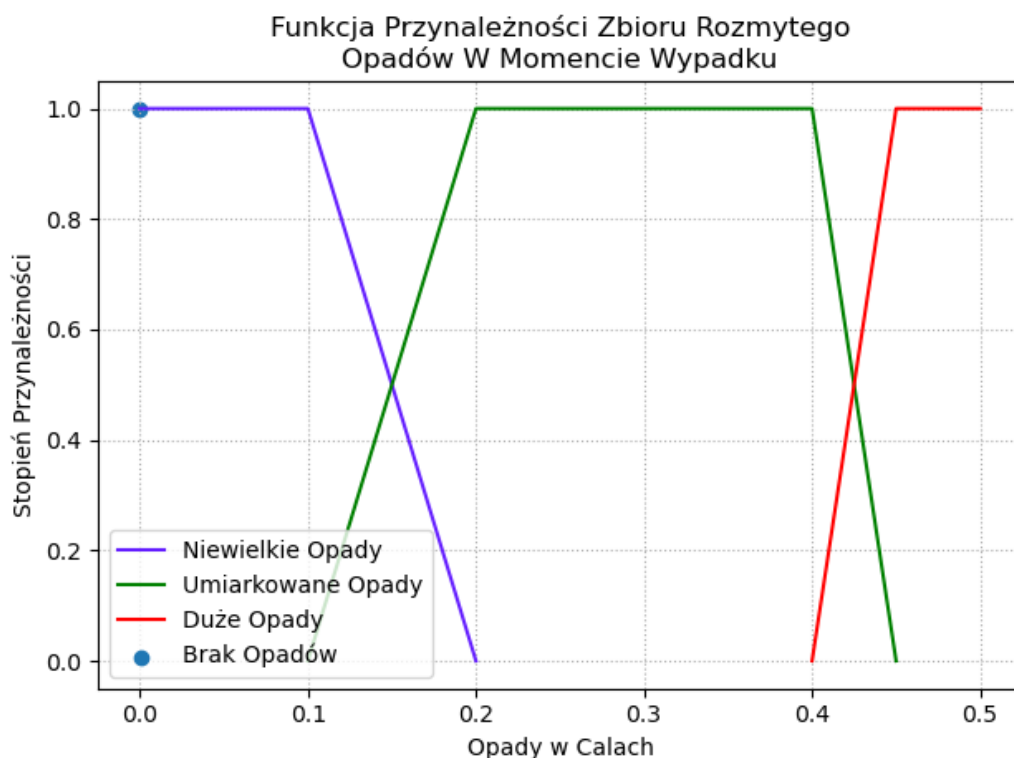
$$\mu_{opadyBrak}(x) = 1 \quad dla \quad x = 0 \quad (30)$$

$$\mu_{opadyNiewielkie}(x) = \begin{cases} 1 & dla \quad x \leq 0.1 \\ \frac{0.1-x}{0.1} & dla \quad 0.1 < x \leq 0.2 \end{cases} \quad (31)$$

$$\mu_{opadyUmiarkowane}(x) = \begin{cases} \frac{x-0.1}{0.1} & dla \quad 0.1 < x \leq 0.2 \\ 1 & dla \quad 0.2 < x < 0.4 \\ \frac{0.45-x}{0.05} & dla \quad 0.4 < x \leq 0.45 \end{cases} \quad (32)$$

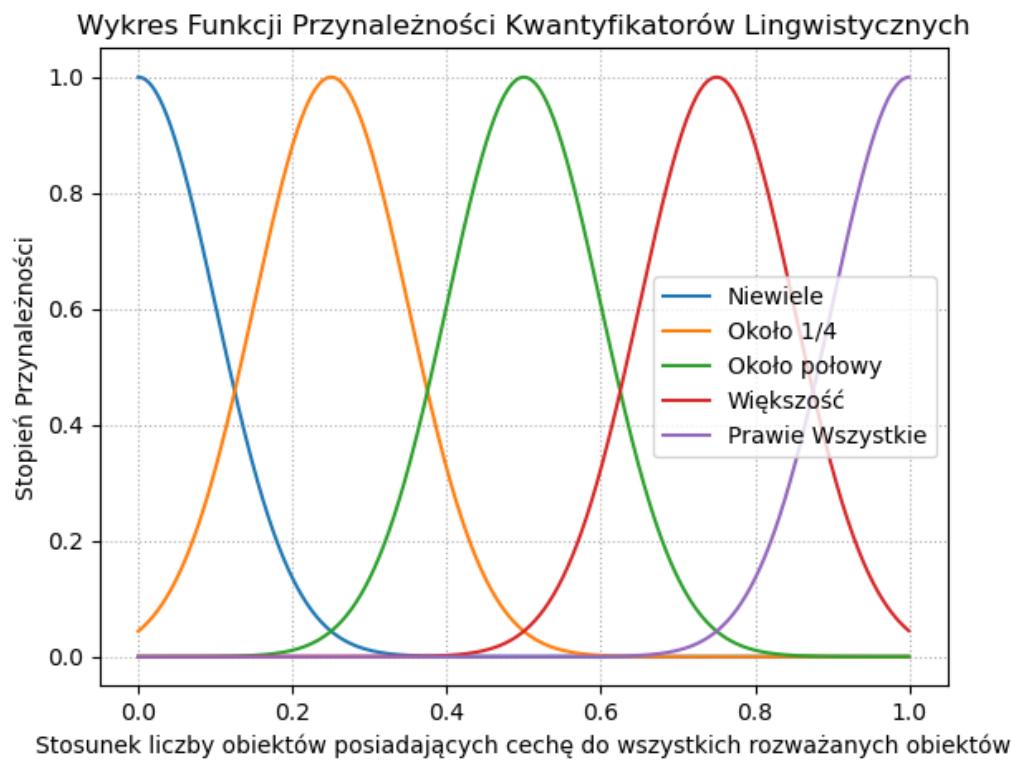
$$\mu_{opadyDuze}(x) = \begin{cases} \frac{x-0.45}{0.05} & dla \quad 0.4 < x \leq 0.45 \\ 1 & dla \quad 0.45 \leq x \end{cases} \quad (33)$$

gdzie: $\mu_{opadyBrak}$, $\mu_{opadyNiewielkie}$, $\mu_{opadyUmiarkowane}$, $\mu_{opadyDuze}$ - funkcje przynależności, x - opady.



Rysunek 9. Wykres funkcji przynależności zbiorów rozmytych ilustrujących wartości zmiennej lingwistycznej opadów.

Do kwantyfikatora lingwistycznego zostały dopasowane etykiety: niewiele, około 1/4, około połowy, większość, prawie wszystkie.



Rysunek 10. Wykres funkcji przynależności zbioru rozmytego opadów.

4. Narzędzia obliczeniowe: projekt (wybór, implementacja) i diagram UML pakietu obliczeń rozmytych. Diagram UML generatora podsumowań

4.1. Diagram pakietu obliczeń rozmytych

Diagram UML i zwięzły opis pakietu obliczeń rozmytych: źródło pakietu (zewnętrzny/własny/hybrydowy), przypis do literatury. Krótka charakterystyka najważniejszych klas i podstawowych dla zadania ich metod.

Sekcja uzupełniona jako efekt zadania Tydzień 10 wg Harmonogramu Zajęć na WIKAMP KSR.

4.2. Diagram UML generatora podsumowań. Krótka instrukcja użytkownika

Diagram UML generatora podsumowań (warstwy obliczeniowej oraz interfejsu użytkownika). Krótki ilustrowany opis jak użytkownik może korzystać z aplikacji, w szczególności wprowadzać parametry podsumowań, odczytywać wyniki oraz definiować własne etykiety i kwantyfikatory. Wersja JRE i inne wymogi niezbędne do uruchomienia aplikacji przez użytkownika na własnym komputerze.

Sekcja uzupełniona jako efekt zadania Tydzień 11 wg Harmonogramu Zajęć na WIKAMP KSR.

5. Jednopodmiotowe podsumowania lingwistyczne. Miary jakości, podsumowanie optymalne

Wyniki kolejnych eksperymentów wg punktów 2.-4. opisu projektu 2. Listy podsumowań jednopodmiotowych i tabele/rankingi podsumowań dla danych atrybutów obowiązkowe i dokładnie opisane w „captions” (tytułach), konieczny opis kolumn i wierszy tabel. Dla każdego podsumowania podane miary jakości oraz miara jakości podsumowania optymalnego.

Sekcja uzupełniona jako efekt zadania Tydzień 11 wg Harmonogramu Zajęć na WIKAMP KSR.

6. Wielopodmiotowe podsumowania lingwistyczne i ich miary jakości

Wyniki kolejnych eksperymentów wg punktów 2.-4. opisu projektu 2. Uzasadnienie i metoda podziału zbioru danych na rozłączne podmioty. Listy podsumowań wielopodmiotowych i tabele/rankingi podsumowań dla danych atrybutów obowiązkowe i dokładnie opisane w „captions” (tytułach), konieczny opis kolumn i wierszy tabel. Konieczne uwzględnienie wszystkich 4-ch form podsumowań wielopodmiotowych.

****** Możliwe sformułowanie zagadnienia wielopodmiotowego podsumowania optymalnego ******.

******Ewentualne wyniki realizacji punktu „na ocenę 5.0” wg opisu Projektu 2. i ich porównanie do wyników z części obowiązkowej******.

Sekcja uzupełniona jako efekt zadania Tydzień 12 wg Harmonogramu Zajęć na WIKAMP KSR.

7. Dyskusja, wnioski

Dokładne interpretacje uzyskanych wyników w zależności od parametrów klasyfikacji opisanych w punktach 3.-4 opisu Projektu 2. Szczególnie istotne są wnioski o charakterze uniwersalnym, istotne dla podobnych zadań. Omówić i wyjaśnić napotkane problemy (jeśli były). Każdy wniosek/problem powinien mieć poparcie w przeprowadzonych eksperymentach (odwołania do konkretnych wyników: tabel i miar jakości). Ocena które wybrane kwantyfikatory, sumaryzatory, kwalifikatory i/lub ich miary jakości mają małe albo duże znaczenie dla wiarygodności i jakości otrzymanych agregacji/podsumowań. Dla końcowej oceny jest to najważniejsza sekcja sprawozdania, gdyż prezentuje poziom zrozumienia rozwiązywanego problemu.

****** Możliwości kontynuacji prac w obszarze logiki rozmytej i wnioskowania rozmytego, zwłaszcza w kontekście pracy inżynierskiej, magisterskiej, naukowej, itp. ******

Sekcja uzupełniona jako efekt zadań Tydzień 11 i Tydzień 12 wg Harmonogramu Zajęć na WIKAMP KSR.

8. Braki w realizacji projektu 2.

Wymienić wg opisu Projektu 2. wszystkie niezrealizowane obowiązkowe elementy projektu, ewentualnie podać merytoryczne (ale nie czasowe) przyczyny tych braków.

Literatura

- [1] A. Niewiadomski, Zbiory rozmyte typu 2. Zastosowania w reprezentowaniu informacji. Seria „Problemy współczesnej informatyki” pod redakcją L. Rutkowskiego. Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2019.
- [2] S. Zadrozny, Zapytania nieprecyzyjne i lingwistyczne podsumowania baz danych, EXIT, 2006, Warszawa
- [3] A. Niewiadomski, Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2008.
- [4] 2021 Kaggle Inc [internetowa społeczność związana z analizą danych], US Accidents (3 million records – updated) A Countrywide Traffic Accident

Dataset (2016 - 2020) [przełączany 24 kwietnia 2021], Dostępny w: <https://www.kaggle.com/sobhanmoosavi/us-accidents>

Literatura zawiera wyłącznie źródła recenzowane i/lub o potwierdzonej wiarygodności, możliwe do weryfikacji i cytowane w sprawozdaniu.