

**UNIWERSYTET RZESZOWSKI**

**Kolegium Nauk Przyrodniczych**

Monika Bąk

Nr albumu: 101260

Kierunek: Informatyka

**PORÓWNANIE SKUTECZNOŚCI MIAR ENTROPII W UJĘCIU   
OPTYMISTYCZNYM I PESYMISTYCZNYM DO UZUPEŁNIANIA BRAKUJĄCYCH DANYCH**

Praca magisterska

Praca wykonana pod kierunkiem

Dr hab. Barbara Pękala, Prof. UR

Rzeszów, 2022

**Spis treści**

[1. Wstęp 2](#_Toc105773355)

[1.1. Cel i zakres pracy 3](#_Toc105773356)

[2. Analiza stanu wiedzy 5](#_Toc105773357)

[2.1. Pojęcie entropii 5](#_Toc105773358)

[2.1.1. Entropia Shannon’a 6](#_Toc105773359)

[2.1.2. Entropia rozmyta 7](#_Toc105773360)

[2.2. Miary entropii przedziałowej 7](#_Toc105773361)

[2.2.1. Entropia w ujęciu klasycznym 13](#_Toc105773362)

[2.2.2. Entropia w ujęciu optymistycznym 14](#_Toc105773363)

[2.2.3. Entropia w ujęciu pesymistycznym 16](#_Toc105773364)

[3. Część badawcza 16](#_Toc105773365)

[3.1. Wykorzystane bazy danych 17](#_Toc105773366)

[3.2. Sposoby uzupełniania brakujących danych z wykorzystaniem entropii 18](#_Toc105773367)

[3.3. Struktura projektu 20](#_Toc105773368)

[3.4. Zastosowanie algorytmu entropii w diagnostyce medycznej 22](#_Toc105773369)

[3.5. Analiza efektywności poszczególnych miar entropii 30](#_Toc105773370)

[4. Dyskusja, podsumowanie i wnioski końcowe 35](#_Toc105773371)

[5. Bibliografia 35](#_Toc105773372)

[6. Spis rysunków 36](#_Toc105773373)

[7. Spis tabel 37](#_Toc105773374)

# **Wstęp**

W dzisiejszym świecie, gdy jakość życia ludzi nieustannie się zmienia poprzez ciągły wzrost tępa egzystencji, jak również poziomu technologicznego, pojawia   
się coraz częściej temat niepewności w skutek uwidaczniających się nowych problemów życiowych. Często występujące dylematy nie są do końca możliwe   
do rozwiązania poprzez niewystarczająco dobrze sformułowane stwierdzenia, bądź jest to niewykonalne przy użyciu samej ludzkiej intuicji.   
W takich sytuacjach naprzeciw ludziom wychodzi pojęcie entropii, która pozwala określać precyzyjne sformułowania i wykonywać obliczenia w odniesieniu   
do jednego z najbardziej palących problemów życia ludzkiego.

Entropia innymi słowy zwana jest również, jako miara niepewności   
bądź miara uporządkowania, im mniejsza jest entropia tym badane zdarzenie   
jest bardziej uporządkowane, im większa spada jego uporządkowanie. Zerowa entropia mówi o doskonałym uporządkowaniu. Dzięki mierze entropii zmniejsza się poziom niewiedzy i niesie to pewną informację, przez co można uznać entropię również jako miarę informacji.

Samo pojęcie entropii w teorii informacji zapoczątkował Claude Shannon   
w swojej pracy „A Mathematical Theory of Communication”, przez co określana jest jako entropia Shannona bądź klasyczna entropia.

Z entropią można się spotkać w różnych dziedzinach nauki, na przykład   
w zbiorach rozmytych, gdzie przybiera formę entropii rozmytej, która   
w rzeczywistości różni się od pojęcia zapoczątkowanego przez Shannona, ponieważ zajmuje się niejasnościami i niejednoznacznymi niepewnościami,   
gdzie entropia klasyczna skupia się na niepewnościach probabilistycznych.

Miarę entropii można przedstawiać w ujęciach pesymistycznych   
i optymistycznych, co również wpływa na efektywność otrzymywanych wyników, niemniej jednak stosowana w różnych aspektach życia np. w diagnostyce medycznej, pozwala na porównanie i analizę otrzymanych wyników w dwóch skrajnych perspektywach.

# **Cel i zakres pracy**

Tematem pracy jest porównanie skuteczności miar entropii w ujęciu optymistycznym i pesymistycznym służących do uzupełniania brakujących danych.

Celem pracy jest stworzenie programu obliczającego wartości entropii   
w dwóch ujęciach: optymistycznym i pesymistycznym przy użyciu środowiska Python, który posłuży w przeprowadzonych badaniach jako metoda uzupełniania braków danych w dostępnej bazie, a następnie pozwoli   
na postawienie tezy i wysnucie wniosków z uzyskanych wyników.

Zakres pracy obejmuje: przygotowanie programu w środowisku Python, który wykorzystując ogólnodostępną bazę danych zawierającą informacje   
o diagnostyce raka piersi w stanie Wisconsin, przeprowadza na niej różnorakie operacje takie jak: obliczenie dolnego i górnego krańca zakresu poszczególnego atrybutu, normalizacji uzyskanych wyników w celu zbudowania przedziałów   
w zakresie [0;1], następnie dokonano losowego usunięcia po jednym atrybucie w wierszu w przedziale 5%, 10% i 20% zbioru, kolejnym etapem było uzupełnienie brakujących danych wartościami [0;1] oraz użycie algorytmu   
K-najbliższych sąsiadów a finalnie również entropii, przy wykorzystaniu różnych agregacji oraz wskaźników prawdopodobieństwa aby skorzystać z entropii   
w ujęciu optymistycznym jak i pesymistycznym. Uzyskane wyniki z dokonanych badań, to jest: accuracy – dokładność, sensitivity – wrażliwość,   
specificity – specyficzność oraz precision – precyzja, finalnie zestawiono   
i porównano w celu postawienia wniosków.[1]

# **Analiza stanu wiedzy**

Dokonanie realizacji postawionego zadania wymaga znajomości podstawowych pojęć z dziedziny teorii informacji, zbiorów rozmytych   
oraz samej entropii. Informacje dotyczące ostatniego pojęcia ściśle łączą się   
z pierwszym jak i z drugim, gdy jest brana pod uwagę entropia rozmyta. Zapoznanie ze wspomnianymi pojęciami podzielone zostało na dwa etapy. Pierwszy z nich dotyczy pojęcia entropii w teorii informacji, która łączy się   
z entropią Shannon’a, drugi etap natomiast dotyczy entropii rozmytej i metod uzupełniania brakujących danych z jej użyciem.

# **Pojęcie entropii**

Pojęcie entropii w teorii informacji można rozumieć jako średni poziom „informacji”, „niepewności” bądź „niespodzianki”, który jest nieodłącznie związany z możliwymi wynikami zmiennej. Wzór na to pojęcie zależny jest   
od dyskretnej zmiennej losowej X, która przyjmuje wartości w alfabecie Xi i jest dystrybuowana zgodnie z [2]

Gdzie p(x) oznacza prawdopodobieństwo zajścia zdarzenia *x*. W podanym wzorze oraz najczęściej w teorii informacji jako podstawę logarytmu najczęściej stosuję się 2, gdzie wówczas jednostką entropii jest bit. Jeśli podstawą logarytmu jest liczba *e* wtedy jednostka ta nazywa się nat, natomiast   
jeśli podstawa wynosi 10 w tej sytuacji mowa o jednostce dit lub hartley.   
W przypadku gdy prawdopodobieństwo zajścia zdarzenia losowego *x* wynosi , wartość , przyjmowana jest jako 0, co jest zgodne z granicą:

.

Pojęcie entropii zostało uogólnione w 1960 roku przez węgierskiego matematyka Alfreda Rényi na zestaw funkcji, które umożliwiły kwantyfikację różnorodności, niepewności lub losowości systemu.

Samo pojęcie entropii może być rozumiane jako niepewność, że pewne fundamentalne wydarzenie nastąpi w następnym czasie. Jeżeli entropia wynosi 0 to znaczy że, zdarzenie w zbiorze zdarzeń zachodzi z prawdopodobieństwem 1, bo z góry wiadomo co się stanie, więc nie ma niepewności.

Tabela 1. Własności entropii.[3]

|  |  |
| --- | --- |
| Własności entropii | |
| 1. | Nieujemna. |
| 2. | Maksymalna, jeśli prawdopodobieństwa zajść zdarzeń są takie same. |
| 3. | Równa 0, jeśli stany systemu przyjmują wartości tylko 0 lub tylko 1. |
| 4. | Własność superpozycji – w sytuacji gdy dwa systemy są niezależne, entropia sumy tych systemów równa się sumie entropii. |

## **Entropia Shannon’a**

Claude Shannon w swojej pracy z 1948 r. „A Mathemcatical Theory of Communication” wprowadził pojęcie entropii informacyjnej, przez co również określana jest jako entropia Shannon’a. Teoria Shannona definiuje system teleinformatyczny składający się z trzech elementów:

* Źródło danych;
* Kanał komunikacyjny;
* Odbiornik.

Shannon w swojej pracy poruszył „Podstawowy problem komunikacji”, który polega na tym, aby wygenerowane przez źródło dane, na podstawie sygnału otrzymanego przez kanał, były gotowe do zidentyfikowania przez odbiorcę. Matematyk rozważał różne sposoby kodowania, kompresowania i przesyłania wiadomości ze źródła danych i udowodnił w swoim słynnym twierdzeniu o kodowaniu źródłowym, że entropia reprezentuje absolutną matematyczną granicę tego, jak dobrze dane ze źródła mogą być bezstratnie skompresowane w całkowicie bezszumowym kanale. Wynik dla zaszumionych kanałów został dzięki temu znacznie wzmocniony przez Shannon’a, co przedstawił w „Twierdzeniu o kodowaniu zaszumionych kanałów”. [4]

## **Entropia rozmyta**

Pojęcie entropii rozmytej w rzeczywistości różni się od pojęcia zapoczątkowanego przez Shannona – entropii klasycznej, ponieważ zajmuje się niejasnościami i niejednoznacznymi niepewnościami, gdzie entropia Shannona skupia się na niepewnościach probabilistycznych.

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

# **Miary entropii przedziałowej**

Entropie są znaczącym aspektem w teorii zbiorów rozmytych, z powodu ich wysokiej użyteczności. Ponadto, w wielu przypadkach podnoszą efekty stosowania zbiorów rozmytych w różnych aplikacjach, gdyż uwzględniają niepewność związaną z budową dokładniej funkcji przynależności. W związku z tym istnieje szerokie zainteresowanie badaniami nad rozszerzeniem koncepcji entropii do radzenia sobie z interwałowymi zbiorami rozmytymi. [1]

W celu zbudowania miary entropii przedziałowej należy posłużyć się pojęciem pierwszeństwa oraz podobieństwa między interwałowymi zbiorami rozmytymi odzwierciedlającymi niepewność. W związku z tym aby trafnie zdefiniować i zrozumieć wspomniane wcześniej miary, na początku należy skupić się na samym pojęciu interwałowych zbiorów rozmytych – IVFS, oraz miar i wskaźników niezbędnych w manipulacji interwałami, to jest: różnorakie funkcje uporządkowania relacji i agregacji oraz miara pierwszeństwa i podobieństwa – Precedence.

Zbiór rozmyty o wartościach przedziałowych (IVFS) jest definiowany następująco: przez *LI* = {[*p*, ] : *p*, [0, 1], *p* } jest oznaczana rodzina przedziałów należących do przedziału jednostkowego. Jeśli , to definiowany jest zbiór rozmyty o wartościach przedziałowych *S* w *X* jako odwzorowanie   
 takie, że każde

*S*(*x*) = [*S*(*x*),

Oznacza stopień przynależności elementu do S.

Rodzina wszystkich IVFSs w *X* jest oznaczona przez IVFS(*X*). Należy założyć odzwierciedlenie aspektu zastosowań w zbiorze skończonym *X* = {*x1, …, xn*}. W przeciwieństwie do zbiorów rozmytych w przypadku IVFSs przynależność nie jest dokładnie wskazana – podana jest tylko górna i dolna granica możliwego członkostwa dlatego, interwałowe zbiory rozmyte są bardzo przydatne ze względu na niepewność informacji.

Każdy zbiór rozmyty *S* mógłby być oczywiście traktowany jako IVFS taki, że *S*(*x*) = . Zatem, FS(*X*) IVFS(*X*), gdzie FS(*X*) jest rodziną zbiorów rozmytych na *X.*

Podstawowe operacje w interwałowych zbiorach rozmytych dla zwane przecięciem, zjednoczeniem, dopełnieniem to:

Gdzie *N* jest rozmytą negacją, a w szczególności dla standardowej negacji jest podany wzór:

Który tworzy sieć rozdzielczą (IVFSs(*X*), ) i spełniają prawa Morgana. [5]

Po przedstawieniu głównej definicji interwałowych zbiorów rozmytych, należy skupić się na objaśnieniu uporządkowania relacji i agregacji w rodzinie przedziałów należących do przedziałów jednostkowych. W tej części omówione zostaną relacje między dowolnymi IVFSs: S(*x*) i T(*x*) dla dowolnego , stąd zostanie przyjęta następująca notacja:

Najbardziej znanym i często używanym porządkiem częściowym w *LI* jest:

Gdzie wtedy i tylko wtedy, gdy i .

Podstawowe operacje, tj. wspólne i spotykające się, określane są w *LI* poniżej:

Analogicznie do (IVFSs(*X*), ) struktura (*LI* jest siecią o rzędzie , gdzie a to wierzchołek i odpowiednio na dole (*LI,* ).

W wielu przypadkach porządek częściowy musi zostać rozszerzony do porządku liniowego, taka relacja nazywana jest porządkiem dopuszczalnym o czym mówi poniższe twierdzenie:

Niech będzie ciągłą funkcją agregacji, która dla wszystkich , , równości zachowują p = q. Jeżeli rząd na *LI* określany jest przez:

),

Wtedy, jest dopuszczalnym porządkiem w *LI.*

Tabela 2. Podstawowe przykłady rzędów dopuszczalnych w LI.[5]

|  |  |
| --- | --- |
| Podstawowe przykłady rzędów dopuszczalnych w *LI* | |
| Rząd Xu-Yager | lub |
| Rząd leksykograficzny |  |
| Rząd | Lub  Gdzie jest zdefiniowany jako dla każdego |

Następnie należy skupić się na pojęciu funkcji agregacji na *LI,* której opis będzie dotyczył agregacji związanych z i który oznacza ścisłą nierówność.

Pierwsza z definicji będzie dotyczyła agregacji związanej z i brzmi ona następująco: niech nazywana jest interwałową funkcją agregacji, jeśli rośnie względem rzędu (częściowo lub liniowo) tj.

i

nx

nx

Funkcja agregacji jest ukuta jako reprezentowalna w sytuacji, gdy istnieją funkcje agregacji przedstawiane w następujący sposób:

Powyższy wynik umożliwia scharakteryzowanie reprezentatywnych funkcji agregacji na *LI.*

Przykłady reprezentowalnych funkcji agregacji dotyczących są następujące:

* Reprezentowalna średnia arytmetyczna:
* Reprezentowalna średnia geometryczna:
* Reprezentowalna średnia potęgowa:
* Reprezentowalny produkt:
* Reprezentowalny środek produktu:
* Reprezentowalna średnia maksymalna:

Co więcej, operacja jest funkcją agregacji w odniesieniu do dla .[5]

Kolejnym narzędziem przydatnym do manipulacjami interwałami musi zostać wprowadzone zanim będzie można przedstawić perspektywę dotyczącą podobieństwa i entropii w środowisku IVFS i jest nim wskaźnik pierwszeństwa, którego definicja przedstawia się następująco:

Funkcja jest postrzegana jako wskaźnik pierwszeństwa, jeśli spełnia następujące warunki dla dowolnego

1. Jeśli
2. Jeśli
3. Jeśli

Poniżej przedstawiono wskaźnik pierwszeństwa w odniesieniu do

gdzie N(*x*) = . Ponadto funkcja poniżej jest wskaźnikiem pierwszeństwa w połączeniu z

Oraz użycie funkcji agregacji , można uzyskać następujący wskaźnik pierwszeństwa:

Gdzie

Jest rozmytą negacją w związku z [5]

Następnie należy skupić się na klasach miar podobieństwa między zbiorami interwałowych zbiorów rozmytych. W celu skonstruowania podobieństwa o wartościach przedziałowych potrzebne są wspomniane wcześniej funkcji agregacji i wskaźniki pierwszeństwa które uwzględniają szerokość interwałów.

Niech i *card*(*X*) = *n*. Dla potrzeba użyć następującego pojęcia częściowego rzędu dla i = 1, …,n gdzie to ten sam rodzaj rzędów (częściowy bądź liniowy) dla każdego   
*i* oraz *si* = *S*(*xi*), *ti* = *T*(*xi*). Należy zauważyć, że jeśli dla *i* = 1, …, *n* będzie brany ten dam porządek , to porządek pomiędzy IVFSs, *S, T* jest częściowy, ale nie musi być liniowy.

Niech będzie funkcją agregacji. Następna akcja która spełnia warunki:

1. .

Nazywana jest miarą podobieństwa dla *i* = 1, …, n.

Wszystkie wspomniane powyżej miary i wartości odnoszą się i łączą z przedziałowymi interwałowymi zbiorami rozmytymi (IVFS), które służą do konstrukcji rozmytej entropii przedziałowej, której klasyczne ujęcia zostaną poruszone w kolejnym podrozdziale. Inne ujęcia entropii rozmytej takie jak optymistyczne i pesymistyczne, które zostały zdefiniowane na rzecz prowadzonego badania, również zostaną omówione w dalszej części pracy.

## **Entropia w ujęciu klasycznym**

W tym podrozdziale zostaną przedstawione podstawowe przykłady przedziałowej entropii rozmytej z którymi można się spotkać w różnych literaturach. Konstruowane entropie o wartościach przedziałowych, potrzebują funkcji agregacji i miary pierwszeństwa w związku z szerokością interwałów, których znaczenie zostało już wcześniej omówione. Ogólna definicja entropii dla porządku częściowego bądź liniowego, brzmi następująco:

Niech *N* będzie silną (inwolucyjną) negacją z punktem równowagi Funkcja to entropia o wartościach interwałowych w IVFS(*X*) w odniesieniu do negacji *N* jeśli dla , jest:

Niech *Prec* będzie miarą podzbiorów, *N* będzie silną negacją z punktem równowagi *e* i *A* będzie idempotentną funkcją agregacji. Następnie funkcja   
jest entropią w odniesieniu do *N*.

Jeśli *S* jest ostry, to , dla każdego *i*, więc dla każdego *i* w konsekwencji otrzymujemy z własnością agregacji funkcję przez co została udowodniona entropia z punktu 1. Entropię z punktu 2 można uzyskać bezpośrednio przez punkt 3 oraz idempotentność funkcji agregacji *A*.

Jeśli to to jako konsekwencję można uzyskać:

Dla *i* = 1, …,*n*.   
W powyższym zbiorze można zauważyć zmienną *A*, która odnosi się do funkcji agregacji dla klasycznej przedziałowej entropii rozmytej, które zostały podane we wcześniejszym rozdziale, na przykład *Amean, Ameanpow* czy *Ameanmax.* Natomiast miara *Prec* dotyczy miar podobieństwa, której przykłady również jak w przypadku funkcji agregacji zostały podane we wcześniejszej części pracy.

# **Entropia w ujęciu optymistycznym**

Używając wzoru klasycznej rozmytej entropii przedziałowej można wyprowadzić go do obliczania jej w innych ujęciach. Pierwszym z nich użytym w przeprowadzonych badaniach jest entropia w ujęciu optymistycznym, nazwana „similarity possible”.

Podstawowym wzorem obliczającym entropię w ujęciu optymistycznym jest: ,  
gdzie:

* czyli są to przedziałowe zbiory rozmyte;
* ;
* – przedziałowe agregacje typu „possible”, gdzie , ;
* – precedence indicator typu „possible”, gdzie:
* Agregacje wykorzystane w obliczaniu precedence indicator, które prezentują się następująco: , gdzie:
  + , gdzie *A* to średnia geometryczna bądź potęgowa;
  + , gdzie *A* to średnia geometryczna bądź potęgowa.
* *N*- negacje wykorzystane do obliczania precedence indicator,   
  , gdzie:
  + ;

Po przedstawieniu wszystkich składowych do obliczenia entropii w ujęciu optymistycznym można zauważyć, że głównym aspektem który odróżnia ją od klasycznej są funkcje agregacji, negacji oraz precedence, czyli miara prawdopodobieństwa. Warto też zauważyć poprzez , że ujęcie optymistyczne polega na tym że przedziały agregacji użyte do obliczenia entropii w pewnym stopniu pokrywają się.

# **Entropia w ujęciu pesymistycznym**

Kolejnym ujęciem miary entropii jest ujęcie pesymistyczne, na rzecz badań nazwany „similarity necessery”, który również swoją nazwę zawdzięcza przedziałom agregacji, które w tym przypadku mogą nie mieć żadnych wspólnych części, co przedstawia się poniżej:

Funkcje agregacji i negacji w tym przypadku nie występują w wzorze na precedences stąd uzyskane wyniki różnią się w znacznym stopniu w odniesieniu do entropii w ujęciu optymistycznym, natomiast wspomniana funkcja precedence indicator wygląda następująco:

,

Gdzie: .

W tym przypadku również entropia w ujęciu pesymistycznym różni się od entropii klasycznej miarą precedence indicator, który jest specyficzny   
i zdefiniowany tylko dla „similarity necessary”.

# **Część badawcza**

Do realizacji badań zaimplementowano algorytm przedziałowej entropii rozmytej, do obliczenia której posłużono się miarami funkcji agregacji i miar podobieństwa odpowiednio dla entropii w ujęciu optymistycznym i pesymistycznym. Każda z nich została użyta do uzupełniania braków danych w wybranym na potrzeby badania zbiorze. Uzyskanie wyniki zostaną porównane pomiędzy sobą oraz z wynikami uzyskanymi w przypadku klasycznej entropii rozmytej.

## **Wykorzystane bazy danych**

Projekt został stworzony z wykorzystaniem ogólnodostępnej bazy danych zawierającej informacje o diagnostyce raka piersi w stanie Wisconsin. Użyty zbiór jest jednym z popularnych zestawów danych z UCI Machine Learning Repository, który zawiera informacje o 569 instancjach. Zebrane przez naukowców dane pochodzą z cyfrowego obrazu próbek, które zostały pobrane za pomocą cienkoigłowego aspiratu masy piersi i poprzez charakterystykę jąder komórkowych opisane w postaci dziesięciu cech o wartościach rzeczywistych jakich jak:

* Promień;
* Tekstura;
* Obwód;
* Pole;
* Gładkość (lokalna zmienność długości promieni);
* Zwartość, obliczana przez wzór:
* Wklęsłość;
* Punkty wklęsłe (liczba wklęsłych części konturu);
* Symetria;
* Wymiar fraktalny.

Każdy badany pacjent w bazie danych również posiada cechę odnoszącą się do diagnozy jaką otrzymał, która została przedstawiona w wartościach M – B, które oznaczają M –zmiany złośliwe, natomiast B – zmiany łagodne. Cały zestaw danych zawiera 212 rekordów z typem zmian złośliwych i 357 z łagodnymi zmianami nowotworowymi.

# **Sposoby uzupełniania brakujących danych z wykorzystaniem entropii**

Metod uzupełniania braków danych w dziedzinie systemów rozmytych   
jest wiele, każda z nich charakteryzuje się większą bądź mniejszą skutecznością,   
co można odczytać z uzyskanych lub ogólnodostępnych wyników badań.

Metody uzupełniania braków danych można rozróżnić, biorąc pod uwagę skuteczność ich działania na dwie grupy. Każda z poniżej wymienionych została przetestowana z użyciem tej samej bazy danych z której również skorzystano w tej pracy i która została opisana we wcześniejszym rozdziale. Poniżej przedstawione wybrane wyniki przykładowych metod, które będą użyte w celu późniejszego porównania ich z zdefiniowanymi nowymi miarami entropii w ujęciu optymistycznym i pesymistycznym.

Pierwsza grupa metod opiera się głównie na znajdowaniu najbardziej podobnych wartości, których przykładowe instrukcje są opisane poniżej:

* Instrukcja 1 wstawia dla brakującej wartości interwał [0;1];
* Instrukcja 2 wyszukuje najbardziej podobne *k* obiektów i wybiera ich wartość dla brakującego atrybutu, następnie oblicza i wstawia dla brakującej wartości minimum, maksimum lub średnią wybranych wartości brakującego atrybutu:.
* Instrukcja 3 znajduje najbardziej podobne obiekty podobnie do instrukcji 2, ale różni się tym że znajduje obiekty z tej samej klasy decyzyjnej co obiekt z brakującą wartością.

Wyniki osiągnięte poprzez użycie ostatniej instrukcji przedstawiają się następująco:

Tabela 3. Instrukcja 3 - uzupełnianie braków danych, Źródło: „Application of entropy measures with uncertainty in classification methods with missing data problem”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metoda** | **Dokładność** | **Wrażliwość** | **Specyfika** | **Precyzja** | **Braki Danych** |
| **Instrukcja 3** | 0,8555894 | 0,9531871 | 0,7877616 | 0,8871398 | 5% |
| 0,8450197 | 0,9529377 | 0,7587463 | 0,8750599 | 25% |
| 0,8399786 | 0,9511847 | 0,7513039 | 0,8683283 | 50% |

Do uzupełniania braków danych można również posłużyć się różnymi rodzajami entropii.

* Pierwszą propozycją do przeprowadzenia badań, może być wspomniana wcześniej klasyczna przedziałowa entropia rozmyta, której wzór przedstawia się następująco: .
* Kolejną entropią, którą można się posłużyć do badań jest entropia o wzorze:
* Przykład entropii trzeciej, która stosuje odlgegłości Hamminga, przedstawia się następująco:   
  , gdzie .

Wyniki uzyskanie z użycia wspomnianych entropii, przedstawiają się następująco:

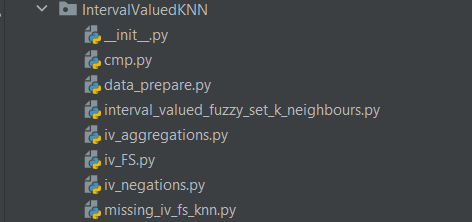
Tabela 4. Uzupełnianie braków danych z użyciem różnych entropii, Źródło: : „Application of entropy measures with uncertainty in classification methods with missing data problem”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Entropia** | **Dokładność** | **Wrażliwość** | **Specyfika** | **Precyzja** | **Braki Danych** |
| **Entropia I** | 0,8875097 | 0,9370335 | 0,8024434 | 0,8914741 | 5% |
| 0,8800044 | 0,9370166 | 0,7698350 | 0,879212 | 25% |
| 0,8778067 | 0,9370094 | 0,7770936 | 0,8775432 | 50% |
| **Entropia II** | 0,8870261 | 0,9369867 | 0,8011517 | 0,8907949 | 5% |
| 0,8742604 | 0,9352522 | 0,7523640 | 0,8717101 | 25% |
| 0,8683938 | 0,9342521 | 0,7482776 | 0,864492 | 50% |
| **Entropia III** | 0,8864288 | 0,9346241 | 0,8038857 | 0,8913553 | 5% |
| 0,8808355 | 0,9314817 | 0,777911 | 0,872423 | 25% |
| 0,8768056 | 0,9165073 | 0,8091392 | 0,8910009 | 50% |

Na potrzeby przeprowadzenia badań zdefiniowano kolejne miary entropii, które przedstawiają ją w ujęciu optymistycznym i pesymistycznym. Każda z nich swoją podstawę opiera na klasycznym wzorze przedziałowej entropii rozmytej, jednakże do obliczenia jej używa różnych funkcji agregacji, negacji i miary prawdopodobieństwa, o czym zostało wspomniane w rozdziale 2.2.2. oraz 2.2.3. gdzie miało miejsce zdefiniowanie wspomnianych pojęć. Wartości otrzymane z przeprowadzonych badań zostaną przedstawione oraz zestawione ze sobą w dalszej części pracy.

## **Struktura projektu**

Program służący do uzupełniania braków danych z użyciem entropii został stworzony przy użyciu środowiska Python. Stworzony projekt odpowiada za różnorakie operacje, które posłużyły do uzyskania oczekiwanych rezultatów stąd też składa się on z kilku modułów/klas, których wykaz przedstawiony jest na rysunku 1.



Rysunek 1. Struktura projektu – klasy

Cały pakiet o nazwie IntervalValuedKNN zawiera 8 modułów, które służą do:

* \_init\_.py – służy do inicjalizacji klasy interval\_valued\_fuzzy\_set\_k\_neighbours.py;
* cmp.py – klasa obsługująca błędy przybliżeń podczas operacji matematycznych;
* data\_prepare.py – klasa zawierająca funkcje związane z przygotowaniem zbiorów danych;
* interval\_valued\_fuzzy\_set\_k\_neighbours.py – klasa, która zawiera cały algorytm KNN, którym należy się posłużyć przy usuwanie oraz uzupełnianiu braków danych;
* iv\_aggregations.py – definiuje funkcje agregacji, które są używane do obliczania entropii;
* iv\_FS.py – klasa definiująca przedziałowy zbiór rozmyty wraz z porządkami;
* iv\_negations.py – definiuje funkcje negacji, używane we wzorze obliczającym entropię;
* missing\_iv\_fs\_knn.py – główna klasa, służąca do uruchomienia KNN, Entropii z odpowiednimi parametrami.

Cały stworzony program działa w oparciu o wiele bibliotek, które oferuje środowisko Python, aby projekt był czytelniejszy i sprawniej działał. Poniżej zostały wymienione przykładowe biblioteki użyte na rzecz pracy:

* „numpy” – służy do obsługi dużych, wielowymiarowych tabel i macierzy;[6]
* „pandas” – oferuje struktury danych i operacje służące do manipulowania różnymi tabelami liczbowymi i szeregami czasowymi;[7]
* „tabulate” – pozwala na tworzenie i wyświetlanie tabel;[8]
* „datetime” – oferuje wiele metod zwracania informacji o dacie (rok, miesiąc dzień, godzina, minuta, sekunda i mikrosekunda);[9]
* „scikit” – pozwala na proste i wydajne analizowanie danych;[10]
* „sklearn” – pozwala na oblczanie różnych wartości z dziedziny klasyfikacji wielokryterialnej, jak np. dokładność podzbioru, czy może posłużyć przy normalizacji danych;[11]
* „math” – oferuje funkcje podobne do kalkulatora, pozwala na skorzystanie z różnorakich obliczeń, na przykład: logarytm oraz zawiera niektóre stałe matematyczne takie jak: liczba pi;[12]
* „openpyxl” – pozwala na pracę z arkuszami kalkulacyjnymi Excel, umożliwia wczytywanie z nich wartości oraz wprowadzanie nowych i tworzenia tabel.[13]

Projekt oprócz wykorzystanych modułów korzysta również z podstawowych funkcji jakie oferuje Python, jak połączenie z bazą danych czy najprostsze wyświetlanie na ekran otrzymanych wyników.

# **Zastosowanie algorytmu entropii w diagnostyce medycznej**

Celem pracy było wykorzystanie algorytmu entropii w uzupełnianiu braków danych w zbiorach danych, które w przypadku prowadzonych badań dotyczył diagnostyki medycznej, dokładnie diagnostyki raka piersi. Aby dokonać pomiarów należało przeprowadzić szereg różnych działań, począwszy od przygotowania danych, przetestowania algorytmu z różnymi parametrami oraz finalnie z uzyskanych entropii obliczenie wartości dokładności, czułości, swoistości i precyzji, które ukazują informację w jakim stopniu wykorzystany algorytm poradził sobie z uzupełnianiem braków danych.

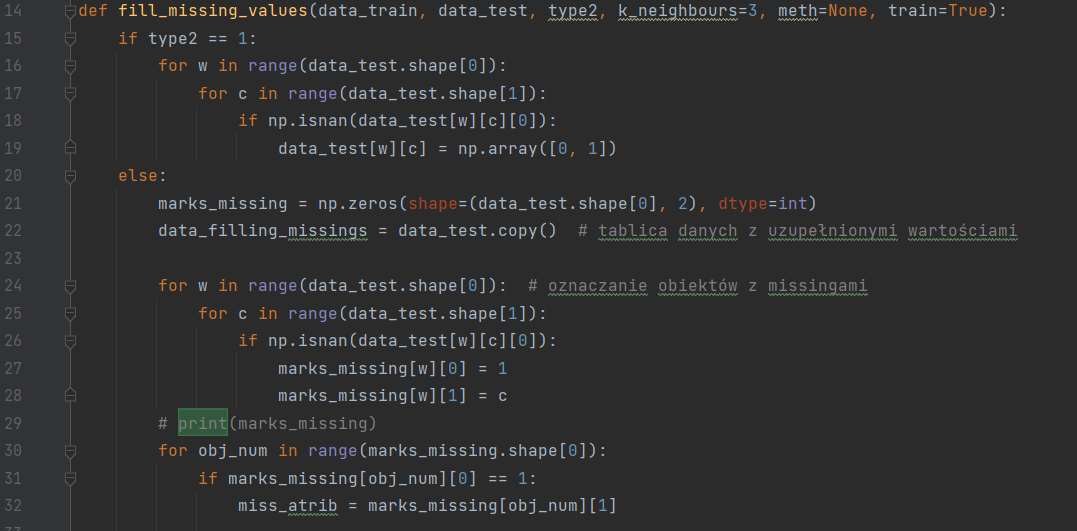
Pierwszym etapem w celu implementacji algorytmu było połączenie z bazą danych, która została zaimportowana z biblioteki „sklearn.datasets”.



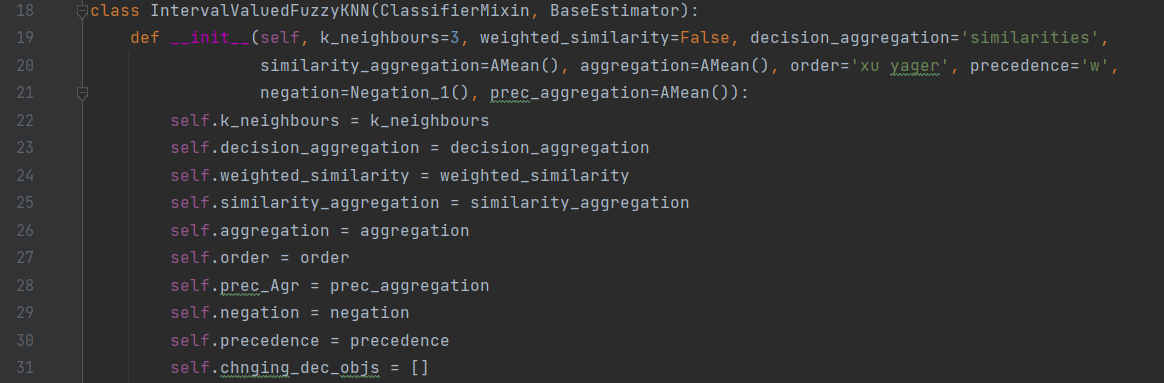
Rysunek 2. Import bazy danych

Pobrany zbiór został znormalizowany w celu zbudowania przedziałów w zakresie [0;1] oraz podzielony na część szkoleniową – 70% całości i testową – 30% całości, aby następnie dokonać testowania algorytmu. Na potrzeby przeprowadzenia badań z przygotowanego zbioru dokonano kolejno w przedziale 5%, 25% i 50% całego zbioru usunięcia po jednym atrybucie, aby następnie użyć zaimplementowanego algorytmu do uzupełniania braków danych.

Kolejnym etapem było sprawdzenie całego zbioru i wstawienia w miejsca brakujących danych 1, w przeciwnym przypadku 0, w celu późniejszego użycia algorytmu K-najbliższych sąsiadów do znalezienia obiektów o podobnych atrybutach.



Rysunek 3. Oznaczanie brakujących wartości

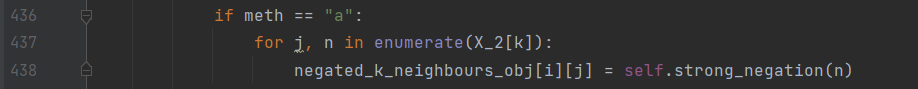


Rysunek 4. Deklaracja klasy obliczającej KNN

Po znalezieniu obiektów o podobnych atrybutach wykorzystano następnie algorytm do obliczania entropii, który w swoim działaniu wykorzystywał przedziałową funkcję agregacji, precedence indicator oraz agregacje służące do jego obliczenia oraz porządki prawdopodobieństwa. Algorytm entropii przedstawia się następująco:



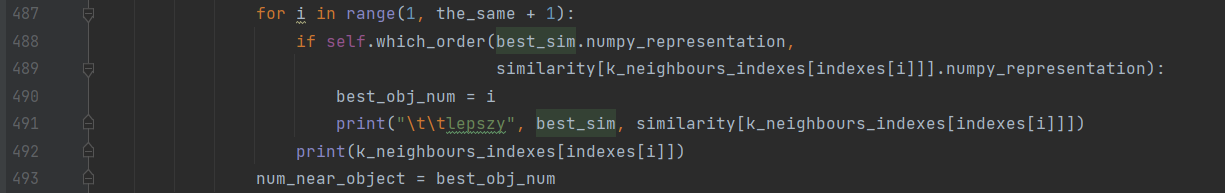
Rysunek 5. Algorytm entropii część 1



Rysunek 6. Algorytm entropii część 2



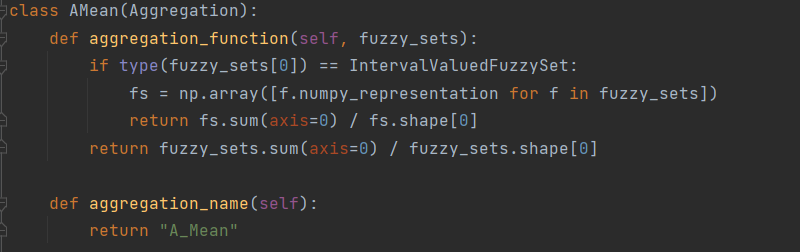
Rysunek 7. Algorytm entropii część 3



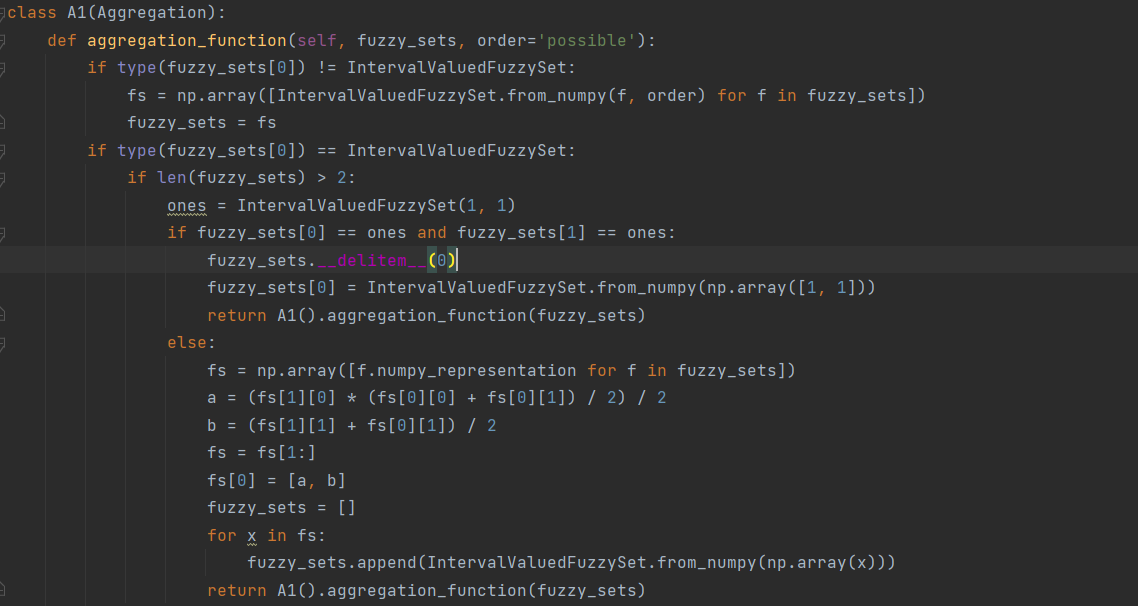
Rysunek 8. Algorytm entropii część 4

W celu obliczenia entropii posługiwano się agregacjami wymienionymi w rozdziale poświęconym definiowaniu entropii w ujęciu optymistycznym oraz pesymistycznym. Dokonano również obliczeń entropii w ujęciu klasycznym w celu późniejszego porównania jej z wartościami uzyskanymi z miar entropii zdefiniowanych na rzecz prowadzonych badań.

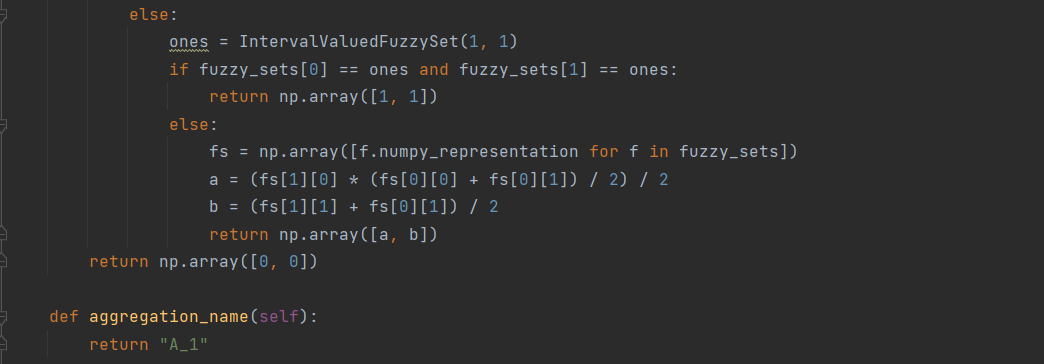
Entropia klasyczna została obliczona z użyciem funkcji agregacji: AMean, AMeanPow oraz AMeanMax i odpowiadającym im miar precedence indicator, użyto również porządków lex2, partial oraz xu yager.



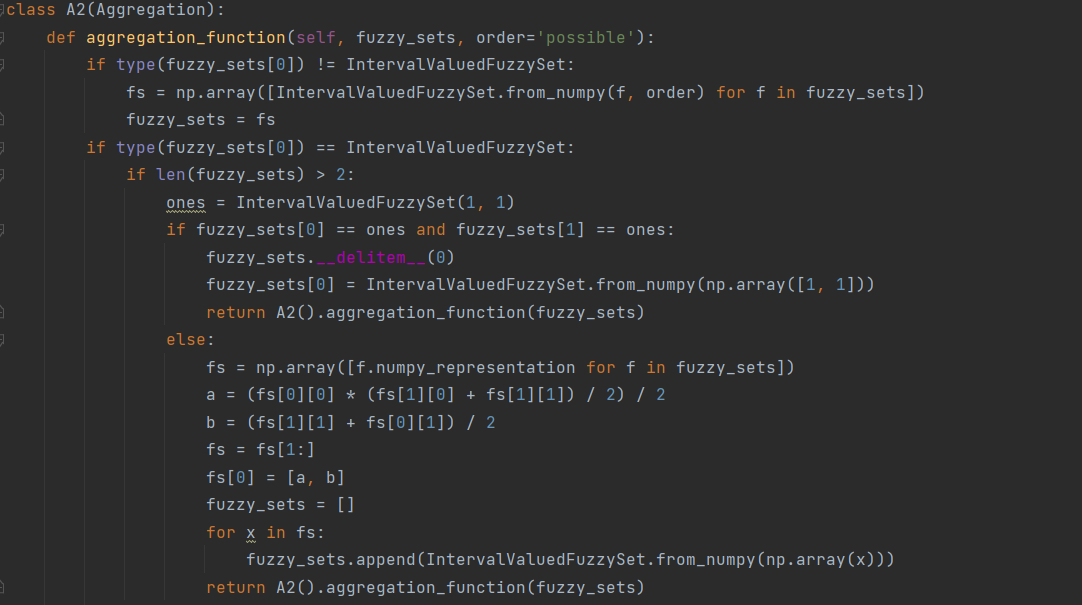
Rysunek 9. Agregacja AMean



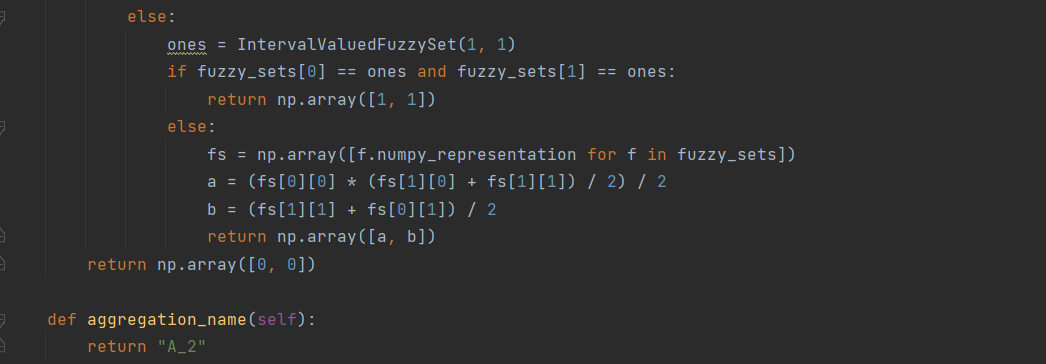
Rysunek 10. Agregacja AI - część 1



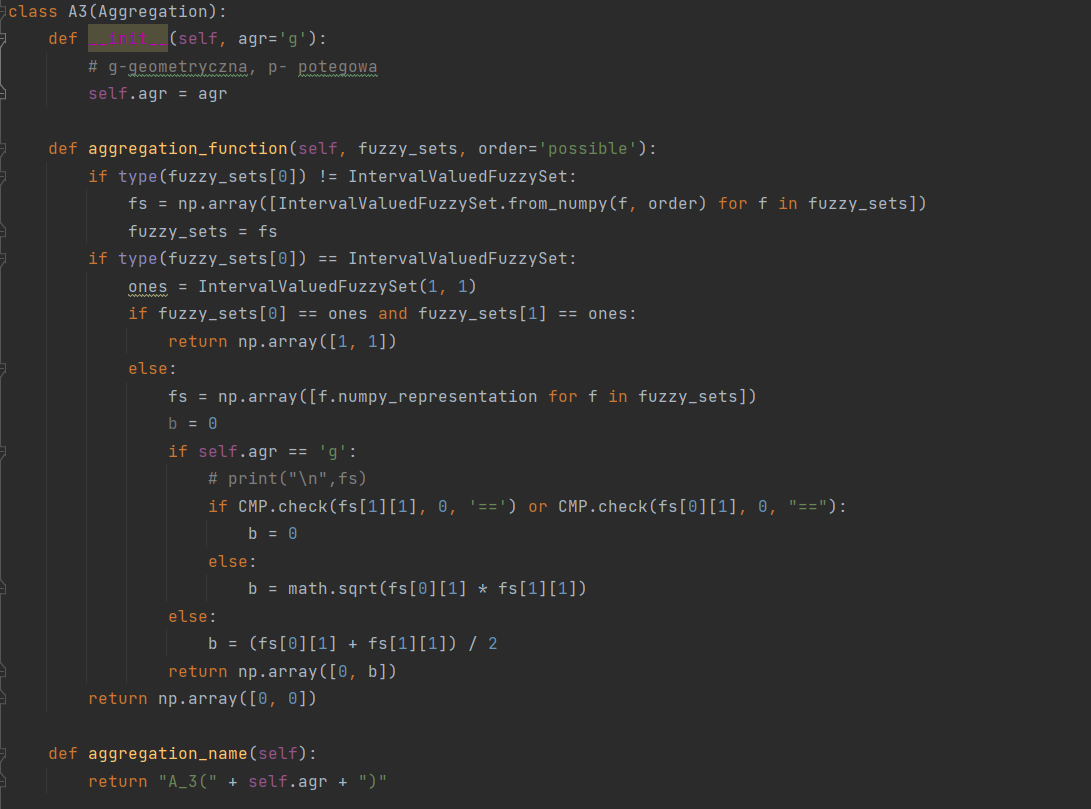
Rysunek 11. Agregracja A1 - część 2



Rysunek 12. Agregacja A2 - część 1



Rysunek 13. Agregacja A2 - część 2

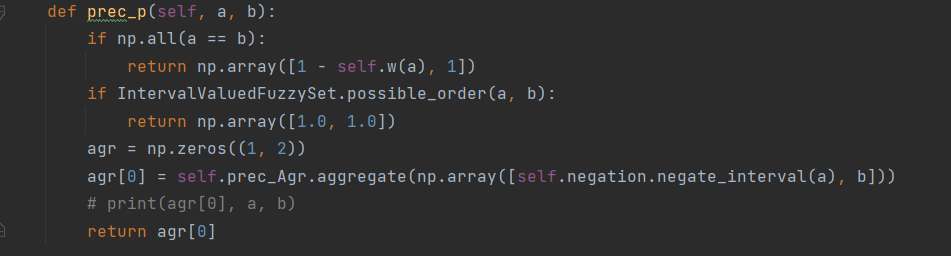


Rysunek 14. Agregacja A3

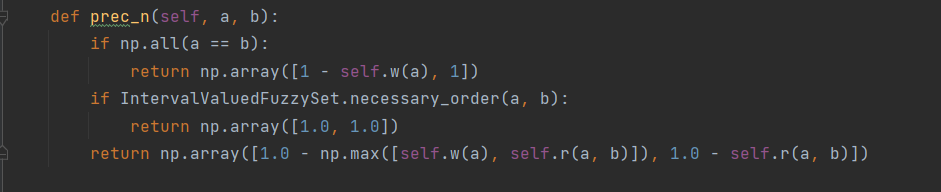


Rysunek 15. Agregacja A4

Funkcje obliczające miarę precedence przedstawiają się następująco:

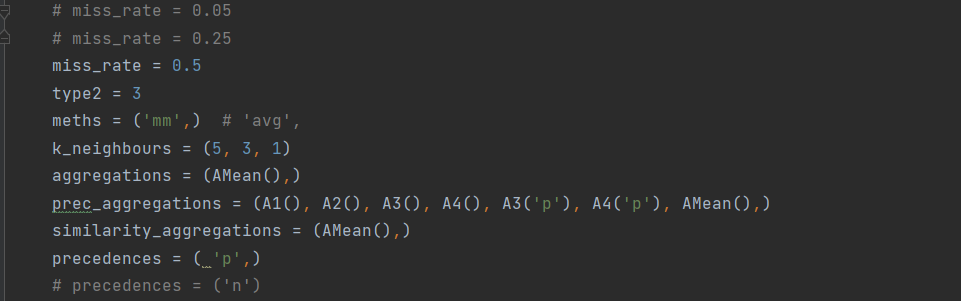


Rysunek 16. Funkcja precedence wykorzystywana w obliczaniu entropii w ujęciu optymistycznym

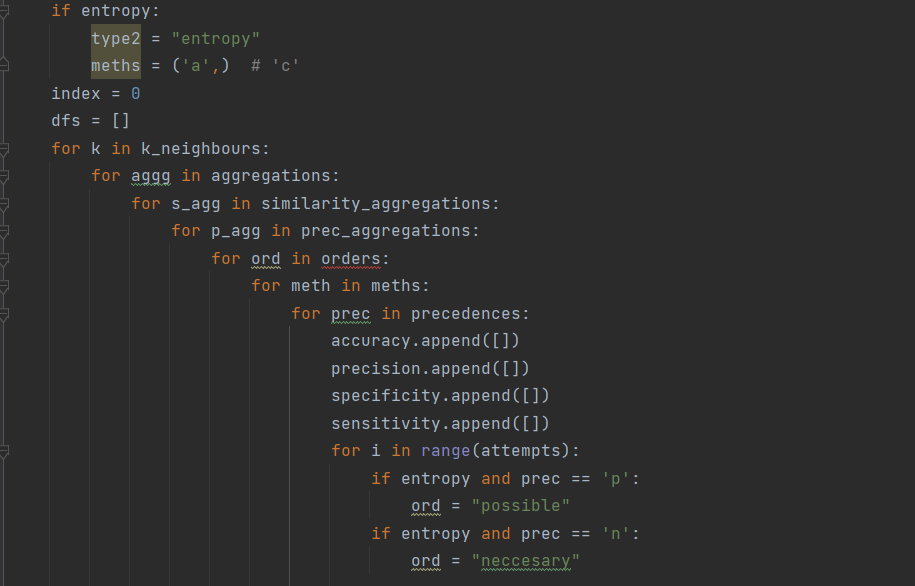


Rysunek 17. Funkcja precedence wykorzystywana w obliczaniu entropii w ujęciu pesymistycznym.

Finalnie w celu uruchomienia algorytmu należało podać odpowiednie nazwy funkcji, które odnosiły się odpowiednio do entropii w ujęciu optymistycznym i pesymistycznym. Przypisane wartości następnie były wykorzystywane w pętlach, które wywoływały konkretne funkcje, które później pozwalały obliczyć wartość entropii. Przeprowadzone działania zostały również powtórzone przy obliczaniu klasycznej entropii.



Rysunek 18. Wybranie odpowiednich funkcji do obliczania entropii

 Rysunek 19. Wywołanie entropii z odpowiednimi parametrami

Cały algorytm był testowany przy K-najbliższych sąsiadów równych odpowiednio 5, 3 oraz 1 i z użyciem 10-krotnej walidacji krzyżowej, która polega na podziale zbioru na 10 podzbiorów, a następnie przeprowadzenia na zbiorze uczącym analiz, natomiast zbiór testowy służy do potwierdzenia wiarygodności uzyskanych wyników.[14] Wykorzystanie kroswalidacji pozwoliło na uzyskanie z jej użyciem dziesięciu wartości z których obliczano ich średnią arytmetyczną i przedstawiano jako wartość uzyskaną przez algorytm w postaci poszczególnych parametrów: accuracy – dokładność, sensitivity – czułość, specificity – swoistość oraz precision – precyzja. Taki sprawdzian krzyżowy był wykonywany kolejno dla każdej obliczanej entropii w zależności od parametrów wykorzystywanych we wzorze.

Otrzymane wyniki w postaci 4 wyżej wymienionych parametrów pozwalają na określenie skuteczności użytych metod do uzupełniania braków danych, ponieważ są to miary jakości, które umożliwiają postawienie poprawnej hipotezy. Do obliczeń tych parametrów należy posłużyć się czterema zmiennymi, które przedstawiają się następująco:

* TP – prawdziwie dodatni, który w przypadku tej pracy odnosi się do osób chorych na raka piersi z diagnozą złośliwą którzy zostali poprawnie zidentyfikowani.
* TN – prawdziwie ujemny, który odnosi się do osób chorych na raka piersi z diagnozą łagodną którzy zostali poprawnie zdiagnozowani.
* FP – fałszywie dodani, odnosi się do osób chorych na raka łagodnego, którzy zostali błędnie zdiagnozowani (tzn. przypisani do diagnozy złośliwej).
* FN – fałszywie ujemny, odnosi się do osób chorych na raka złośliwego, którzy zostali błędnie zdiagnozowani (tzn. przypisani do raka łagodnego).

Wzory na miary jakości oparte na powyższych wartościach:

* Dokładność – accuracy:
* Precyzja – precision: , która określa jaka część uzyskanych wyników wskazanych przez klasyfikator jako dodatnie jest faktycznie dodatnia.
* Czułość – sensitivity: , która określ jaką część dodatnich wyników wykrył klasyfikator.
* Swoistość – specificity: , która określa jaką część ujemnych wyników wykrył klasyfikator. [15]

# **Analiza efektywności poszczególnych miar entropii**

Wynik przedstawiające miary jakości otrzymane w skutek implementacji algorytmu entropii służącego do uzupełniania braków danych w dwóch ujęciach pozwoliły dokonać analizy skuteczności użytych metod.

Początkowo w celu dokonania badań należy skupić się na wynikach jakie otrzymano kolejno dla entropii optymistycznej, a następnie entropii pesymistycznej na zbiorze danych z 5%, 25% oraz 50% brakujących wartości. Przedstawione tabele pozwolą wybrać najlepszy wynik osiągnięty z obliczeń dokonanych z użyciem entropii optymistycznej oraz agregacji użytych w precedence do jej uzyskania. Wybrane wartości zostaną wówczas porównane z wynikami otrzymanymi w przypadku entropii pesymistycznej oraz klasycznej z odpowiednimi brakami danych, co następnie pozwoli na wysnucie wniosków i postawienie tezy.

Entropia w ujęciu optymistycznym była obliczana z użyciem zbioru danych z brakiem danych w zakresie 5%, 25%, 50%, oraz z użyciem odpowiednim jej parametrom.

Wyniki uzyskane z badania z zbiorem danych który zawierał po jednym losowo usuniętym atrybucie w 5% zbioru prezentują się następująco:

Tabela 5. Entropia optymistyczna z brakiem danych w 5%

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **accuracy** | **sensitivity** | **specificity** | **precision** | **aggr\_prec** | **k** |
| **1** | 0,82346623 | 0,754781219 | 0,942055829 | 0,956779522 | A\_1 | 5 |
| **2** | 0,819351901 | 0,750150358 | 0,93901461 | 0,954426577 | A\_2 | 5 |
| **3** | **0,840683537** | **0,770207236** | 0,962388079 | 0,972133903 | A\_3(g) | 5 |
| **4** | 0,802923977 | 0,725570454 | 0,937810189 | 0,953787467 | A\_4(g) | 5 |
| **5** | 0,820842793 | 0,742320094 | 0,956035134 | 0,965990315 | A\_3(p) | 5 |
| **6** | 0,812280702 | 0,742463904 | 0,934853632 | 0,953081619 | A\_4(p) | 5 |
| **7** | **0,840350877** | 0,767029943 | **0,967494877** | **0,975496442** | A\_Mean | 5 |
| **8** | 0,867757138 | 0,816945188 | 0,956181556 | 0,969451956 | A\_1 | 3 |
| **9** | 0,870082559 | 0,821654552 | 0,954710968 | 0,96873662 | A\_2 | 3 |
| **10** | **0,878307534** | **0,831383704** | **0,95935532** | **0,971852893** | A\_3(g) | 3 |
| **11** | 0,839181287 | 0,797590269 | 0,912532412 | 0,941677949 | A\_4(g) | 3 |
| **12** | 0,866590987 | 0,81314505 | 0,959104102 | 0,971525523 | A\_3(p) | 3 |
| **13** | 0,843274854 | 0,804125432 | 0,912532412 | 0,941964846 | A\_4(p) | 3 |
| **14** | 0,869590643 | 0,818864207 | 0,957970396 | 0,970643486 | A\_Mean | 3 |
| **15** | 0,898080495 | 0,898711809 | 0,898670865 | 0,938607167 | A\_1 | 1 |
| **16** | 0,902160303 | **0,910464068** | 0,888984072 | 0,933633248 | A\_2 | 1 |
| **17** | 0,893567251 | 0,898190076 | 0,885990116 | 0,931158664 | A\_3(g) | 1 |
| **18** | 0,835087719 | 0,83978213 | 0,827906241 | 0,896905291 | A\_4(g) | 1 |
| **19** | 0,894561404 | 0,895721605 | 0,894084063 | 0,935429261 | A\_3(p) | 1 |
| **20** | 0,835087719 | 0,83978213 | 0,827906241 | 0,896905291 | A\_4(p) | 1 |
| **21** | **0,902923977** | 0,900879071 | **0,907189006** | **0,943276987** | A\_Mean | 1 |

Przedstawiona tabela pokazuje, że w przypadku *k* równego 5 można zauważyć że entropia z użyciem miary precedence indicator z agregacją A3, która w swoim wzorze zawiera średnią geometryczną(g) charakteryzuje się najlepszym wynikiem w przypadku miary dokładności i czułości, co pozwala stwierdzić że najlepiej poradziła sobie z uzupełnieniem danych dotyczących osób chorych na raka z diagnozą złośliwą w sposób poprawny. Z kolei entropia z użyciem miary precedence indicator z agregacją AMean, osiągnęła najlepszą wartość miary swoistości i precyzji, co określa w jakim stopniu pozytywnie określiła osoby chore na raka złośliwego oraz również na raka z diagnozą łagodną.

W przypadku k równego 3, wszystkie cztery miary jakości okazały się najlepsze dla entropii z użyciem miary precedence indicator z agregacją A3(g), czyli poradziła sobie ona w tym przypadku najlepiej z uzupełnianiem braków danych.

Dla k równego 1 można zauważyć rozbieżność wyników pomiędzy wynikami. Najlepiej w przypadku miary dokładności, specyficzności oraz precyzji poradziła sobie entropia z użyciem agregacji AMean, natomiast w przypadku miary czułości była to miara entropii z użyciem agregacji A1.

Zestawiając wszystkie wartości przedstawiające skuteczność uzupełniania danych w zbiorze danych z 5% brakiem obiektów, można zauważyć że algorytm entropii działa najlepiej gdy *k* dla K-najbliższych sąsiadów jest równe 1 jeśli będzie brana pod uwagę miara dokładności, z kolei jeśli uwaga zostanie skupiona na pozostałych miarach jakości, algorytm najlepiej zadziałał gdy k było równe 5.

Wyniki uzyskane z badania ze zbiorem danych który zawierał po jednym losowo usuniętym atrybucie w 25% zbioru z użyciem entropii optymistycznej prezentują się następująco:

Tabela 6. Entropia optymistyczna z brakiem danych 25%

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **accuracy** | **sensitivity** | **specificity** | **precision** | **aggr\_prec** | **k** |
| **1** | 0,81687 | 0,744293 | 0,946218 | 0,959231 | A\_1 | 5 |
| **2** | 0,817544 | 0,745446 | 0,945979 | 0,959619 | A\_2 | 5 |
| **3** | 0,829601 | 0,763689 | 0,947054 | 0,961938 | A\_3(g) | 5 |
| **4** | 0,78538 | 0,696831 | 0,944166 | 0,956164 | A\_4(g) | 5 |
| **5** | 0,810757 | 0,732376 | 0,951294 | 0,962523 | A\_3(p) | 5 |
| **6** | 0,792982 | 0,710761 | 0,940791 | 0,954847 | A\_4(p) | 5 |
| **7** | **0,845614** | **0,785887** | **0,951571** | **0,966386** | A\_Mean | 5 |
| **8** | 0,85666 | 0,803534 | **0,952365** | 0,966753 | A\_1 | 3 |
| **9** | 0,858999 | 0,807283 | **0,952365** | **0,966898** | A\_2 | 3 |
| **10** | 0,861404 | 0,818982 | 0,937556 | 0,958016 | A\_3(g) | 3 |
| **11** | 0,793567 | 0,72942 | 0,912199 | 0,935515 | A\_4(g) | 3 |
| **12** | 0,852362 | 0,800412 | 0,945484 | 0,962706 | A\_3(p) | 3 |
| **13** | 0,799415 | 0,739344 | 0,91056 | 0,935075 | A\_4(p) | 3 |
| **14** | **0,879532** | **0,841254** | 0,949424 | 0,965949 | A\_Mean | 3 |
| **15** | 0,892322 | 0,897778 | 0,885893 | 0,93118 | A\_1 | 1 |
| **16** | 0,892903 | 0,892731 | **0,895398** | 0,936101 | A\_2 | 1 |
| **17** | 0,893567 | 0,909538 | 0,86686 | 0,922946 | A\_3(g) | 1 |
| **18** | 0,798246 | 0,79028 | 0,817209 | 0,885881 | A\_4(g) | 1 |
| **19** | 0,892807 | 0,895705 | 0,890305 | 0,933899 | A\_3(p) | 1 |
| **20** | 0,798246 | 0,79028 | 0,817209 | 0,885881 | A\_4(p) | 1 |
| **21** | **0,911111** | **0,922781** | 0,891218 | **0,936772** | A\_Mean | 1 |

Patrząc na powyższą tabelę można zauważyć, że dla k równego 5 najlepsze miary jakości osiąga entropia z użyciem precedence indicator z agregacją AMean.

Jeśli zostanie użyty algorytm z k równym 3, to w tym przypadku najlepszą miarą dokładności i czułości charakteryzuje się tak samo jak w przypadku gdy k jest równe 5 entropia z agregacją AMean. Natomiast w przypadku miar swoistości i precyzji najlepszą wartość uzyskuje entropia z agregacją A2, jednakże porównując ją do wartości otrzymanych na rzecz entropii z agregacją AMean można zauważyć że są to niewielkie różnice w wartościach, stąd jest możliwe stwierdzenie że najlepszą skuteczność w tym przypadku ma entropia z agregacją AMean.

W przypadku gdy k jest równe 1 najlepsze wyniki osiąga również entropia z użyciem precedence indicator z agregacją AMean.

Zestawiając wszystkie pomiary uzyskane z zastosowania entropii w ujęciu optymistycznej do uzupełniania braków danych w 25% całego zbioru, wolno stwierdzić że najlepsze wyniki zostały osiągnięte w przypadku gdy *k* dla K-najbliższych sąsiadów jest równe 3.

Użycie entropii w ujęciu optymistycznym do uzupełniania braków danych w zbiorze danych z brakiem losowych atrybutów w 50% zbioru, prezentuje następujące wyniki:

Tabela 7. Entropia optymistyczna z brakiem danych 50%

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **accuracy** | **sensitivity** | **specificity** | **precision** | **aggr\_prec** | **k** |
| **1** | 0,83614 | 0,775169 | 0,938968 | 0,954596 | A\_1 | 5 |
| **2** | 0,833213 | 0,769402 | 0,940532 | 0,955473 | A\_2 | 5 |
| **3** | 0,833729 | 0,767759 | 0,944961 | 0,959189 | A\_3(g) | 5 |
| **4** | 0,817544 | 0,763215 | 0,90956 | 0,93527 | A\_4(g) | 5 |
| **5** | 0,825504 | 0,757429 | 0,939917 | 0,955389 | A\_3(p) | 5 |
| **6** | 0,82807 | 0,779262 | 0,911102 | 0,93755 | A\_4(p) | 5 |
| **7** | **0,857895** | **0,803022** | **0,949684** | **0,964319** | A\_Mean | 5 |
| **8** | 0,861899 | 0,81786 | 0,937078 | 0,956879 | A\_1 | 3 |
| **9** | 0,863068 | 0,81683 | 0,941676 | 0,959919 | A\_2 | 3 |
| **10** | 0,861844 | 0,817351 | 0,936244 | 0,95694 | A\_3(g) | 3 |
| **11** | 0,846199 | 0,800126 | 0,922958 | 0,948385 | A\_4(g) | 3 |
| **12** | 0,862363 | 0,813435 | **0,94414** | 0,961848 | A\_3(p) | 3 |
| **13** | 0,852047 | 0,812281 | 0,918457 | 0,945247 | A\_4(p) | 3 |
| **14** | **0,877778** | **0,839065** | 0,94249 | **0,962472** | A\_Mean | 3 |
| **15** | 0,892233 | 0,913447 | 0,855644 | 0,91542 | A\_1 | 1 |
| **16** | 0,884018 | 0,902277 | 0,852416 | 0,912997 | A\_2 | 1 |
| **17** | **0,901754** | **0,916255** | **0,87709** | **0,926818** | A\_3(g) | 1 |
| **18** | 0,819298 | 0,843203 | 0,773584 | 0,870949 | A\_4(g) | 1 |
| **19** | 0,890388 | 0,906514 | 0,862903 | 0,918254 | A\_3(p) | 1 |
| **20** | 0,819298 | 0,843203 | 0,773584 | 0,870949 | A\_4(p) | 1 |
| **21** | 0,896491 | 0,911477 | 0,870513 | 0,92382 | A\_Mean | 1 |

Dokonując analizy powyższej tabeli można zauważyć, że najlepsze miary jakości osiąga entropia z użyciem precedence indicator o agregacji AMean dla k równego 5, co pozwala stwierdzić że uzupełnianie braków danych z jej użyciem przebiegło najlepiej w rozważanym przypadku.

Jeśli k dla K-najbliższych sąsiadów jest równe 3 wówczas najlepsze efekty odnosi również entropia z agregacja AMean, jednak można zauważyć małą rozbieżność w uzyskanych wynikach porównując ją z entropią która posługuje się precedence indicator o agregacjach A1, A2 czy A3, ponieważ są to różnice w tysięcznych częściach ułamków.

Gdy zostanie wzięte pod uwagę k równe 1, można stwierdzić że najlepiej poradziła sobie entropia z użyciem precedence indicator o agregacji A3, która w swoim wzorze posługuje się średnią geometryczną.

Biorąc pod uwagę całość uzyskanych wyników jest możliwe stwierdzenie że najlepsze wyniki w uzupełnianiu danych są osiągane jeśli k dla K-najbliższych sąsiadów jest równe 1, jednakże tylko wtedy gdy porównywane będą miary jakości uwzględnione całościowo, ponieważ najlepszym wynikiem dla dokładności charakterysuje się entropia dla k równego 1, dla czułości – k równe 1, swoistość – k równe 5, natomiast precyzja jest najlepsza dla k równego 5.

W celu zestawienia najlepszych osiągniętych wyników uzyskanych przy użyciu entropii w ujęciu optymistycznym otrzymano poniższą tabelę:

Tabela 8. Najlepsze wyniki uzyskane przy użyciu entropii w ujęciu optymistycznym.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **accuracy** | **sensitivity** | **specificity** | **precision** | **k** | **braki danych** |
| **1** | 0,840351 | 0,76703 | **0,967495** | **0,975496** | 5 | 5% |
| **2** | 0,845614 | 0,785887 | 0,951571 | 0,966386 | 5 | 25% |
| **3** | **0,857895** | **0,803022** | 0,949684 | 0,964319 | 5 | 50% |
| **4** | 0,878308 | 0,831384 | **0,959355** | **0,971853** | 3 | 5% |
| **5** | **0,879532** | **0,841254** | 0,949424 | 0,965949 | 3 | 25% |
| **6** | 0,877778 | 0,839065 | 0,94249 | 0,962472 | 3 | 50% |
| **7** | 0,902924 | 0,900879 | **0,907189** | **0,943277** | 1 | 5% |
| **8** | **0,911111** | **0,922781** | 0,891218 | 0,936772 | 1 | 25% |
| **9** | 0,901754 | 0,916255 | 0,87709 | 0,926818 | 1 | 50% |

Dokonując analizy uzyskanych najlepszych wyników przy użyciu entropii w ujęciu optymistycznym można stwierdzić, że w przypadku k dla K-najbliższych sąsiadów równych 5 następuje poprawa w uzupełnianiu braków danych jeśli do analizy wykorzystano zbiór w którym występuje 50% braków danych gdy jest brana pod uwagę miara dokładności i czułości. W przypadku miary swoistości i precyzji, najlepsze wyniki są osiągane w sytuacji gdy baza danych ma tylko 5% braków.

Jeśli *k* jest równe 3 lub 1 najlepsze wyniki biorąc pod uwagę miarę dokładności i czułości można osiągnąć przy korzystaniu ze zbioru danych w którym jest 25% braków. Natomiast miara swoistości i precyzji w obydwu przypadkach najlepsza jest dla zbioru który posiada tylko 5% braków.

# **Dyskusja, podsumowanie i wnioski końcowe**

# **Bibliografia**

[1] Pękala Barbara, Rak Ewa, Kosior Dawid, Mrukowicz Marcin, Bazan Jan, *„Application of similarity measures with uncertainty in classification methods”*

[2], [4] <https://en.m.wikipedia.org/wiki/Entropy_%28information_theory%29>

[3] <https://pl.wikipedia.org/wiki/Entropia_(teoria_informacji)>

[5] Pękala Barbara, Kosior Dawid, Dyczkowski Krzysztof, Szkoła Jarosław, „Application of entropy measures with uncertainty in classification methods with missing data problem”

[6] https://pl.wikipedia.org/wiki/NumPy

[7] https://en.wikipedia.org/wiki/Pandas\_(software)

[8] https://pypi.org/project/tabulate/

[9] https://www.w3schools.com/python/python\_datetime.asp

[10] https://scikit-learn.org/stable/

[11] https://scikit-learn.org/0.15/modules/model\_evaluation.html

[12] http://books.icse.us.edu.pl/runestone/static/thinkcspy/PythonModules/  
Themathmodule.html

[13] <https://openpyxl.readthedocs.io/en/stable/>

[14] <https://pl.wikipedia.org/wiki/Sprawdzian_krzy%C5%BCowy>

[15] Jaworski W., „Miary jakości”, Instytut Informatyki, Uniwersytet Warszawski

# **Spis rysunków**

[Rysunek 1. Struktura projektu – klasy 21](#_Toc105773274)

[Rysunek 2. Import bazy danych 23](#_Toc105773275)

[Rysunek 3. Oznaczanie brakujących wartości 23](#_Toc105773276)

[Rysunek 4. Deklaracja klasy obliczającej KNN 23](#_Toc105773277)

[Rysunek 5. Algorytm entropii część 1 24](#_Toc105773278)

[Rysunek 6. Algorytm entropii część 2 24](#_Toc105773279)

[Rysunek 7. Algorytm entropii część 3 24](#_Toc105773280)

[Rysunek 8. Algorytm entropii część 4 24](#_Toc105773281)

[Rysunek 9. Agregacja AMean 25](#_Toc105773282)

[Rysunek 10. Agregacja AI - część 1 25](#_Toc105773283)

[Rysunek 11. Agregracja A1 - część 2 26](#_Toc105773284)

[Rysunek 12. Agregacja A2 - część 1 26](#_Toc105773285)

[Rysunek 13. Agregacja A2 - część 2 26](#_Toc105773286)

[Rysunek 14. Agregacja A3 27](#_Toc105773287)

[Rysunek 15. Agregacja A4 27](#_Toc105773288)

[Rysunek 16. Funkcja precedence wykorzystywana w obliczaniu entropii w ujęciu optymistycznym 28](#_Toc105773289)

[Rysunek 17. Funkcja precedence wykorzystywana w obliczaniu entropii w ujęciu pesymistycznym. 28](#_Toc105773290)

[Rysunek 18. Wybranie odpowiednich funkcji do obliczania entropii 28](#_Toc105773291)

[Rysunek 19. Wywołanie entropii z odpowiednimi parametrami 29](#_Toc105773292)

# **Spis tabel**

[Tabela 1. Własności entropii.[3] 6](#_Toc105773242)

[Tabela 2. Podstawowe przykłady rzędów dopuszczalnych w LI.[5] 9](#_Toc105773243)

[Tabela 3. Instrukcja 3 - uzupełnianie braków danych, Źródło: „Application of entropy measures with uncertainty in classification methods with missing data problem” 19](#_Toc105773244)

[Tabela 4. Uzupełnianie braków danych z użyciem różnych entropii, Źródło: : „Application of entropy measures with uncertainty in classification methods with missing data problem” 19](#_Toc105773245)

[Tabela 5. Entropia optymistyczna z brakiem danych w 5% 31](#_Toc105773246)

[Tabela 6. Entropia optymistyczna z brakiem danych 25% 32](#_Toc105773247)

[Tabela 7. Entropia optymistyczna z brakiem danych 50% 33](#_Toc105773248)

[Tabela 8. Najlepsze wyniki uzyskane przy użyciu entropii w ujęciu optymistycznym. 35](#_Toc105773249)