

**Mateusz Pazdan**

Realizacja algorytmu drzewa decyzyjnego, oraz badanie wpływu metaparametrów na zmiany poprawności kwalifikacji.

**Praca projektowa z modułu Sztuczna Inteligencja**

Rzeszów, 2024

Spis treści

[1. Wstęp 2](#_Toc165820510)

[1.1. Cel projektu 2](#_Toc165820511)

[1.2. Specyfikacja danych 2](#_Toc165820512)

[1.3. Przygotowanie danych 3](#_Toc165820513)

[2. Zagadnienia teoretyczne 4](#_Toc165820514)

[2.1. Drzewo decyzyjne 4](#_Toc165820515)

# Wstęp

## Cel projektu

Celem tego projektu jest zbadanie efektywności drzewa decyzyjnego w klasyfikacji danych medycznych związanych z przetrwaniem pacjentów po operacjach raka piersi w zależnośi od różnych metaparametrów, w celu lepszego zrozumienia ich wpływu na poprawnoś klasyfikacji. W ramach projektu zbadano wpływ niżej wymienonych metaparametrów.

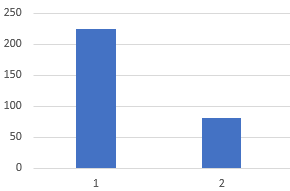
* Kryterium podziału (‘Criterion’)
* Maksymalna głębokoś drzewa (‘max\_depth’)
* Metaparamter Splitter (‘Splitter’)
* Minimalna liczba próbek wymagana do podziału węzła wewnętrznego (‘min\_samples\_split’)
* Minimalna liczba próbek wymagana do utworzenia liścia (‘min\_samples\_leaf’)
* Minimalna frakcja wag wymagana do podziału węzła (‘min\_weight\_fraction\_leaf’)
* Maksymalna liczba cech branych pod uwagę (‘max\_features’)
* Maksymalna liczba liści (max\_leaf\_nodes’)
* Minimalne zmniejszenie nieczynności wymagane do podziału węzła (‘min\_impurity\_decrease’)
* Minimalna złożoność kosztów cięcia (‘ccp\_alpha’)

## Specyfikacja danych

W Projekci wykorzystamy zbiór danych *Haberman's Survival* zawierający informacje dotyczące pacjentów, którzy przeszli operację raka piersi w szpitalu Uniwersytetu Świętej Franciszki w Chicago w latach 1958-1970. Podanych zbiór składa się z 306 rekordów opisancyh 4 cechami:

* *age* - Wiek pacjenta w momencie operacji.
* *Operation\_year*- Rok operacji w przedziale 1958 do 1970 jako 2 ostatnie cyfry.
* *Number of positive axillary nodes detected* - Liczba zarejestrowanych węzłów chłonnych.
* *Surtival status* - Status przeżycia po operacji, gdzie 1 (pacjent żył ponad 5 lat po operacji), a 2 (pacjent zmarł w ciągu 5 lat po operacji).

Układ cech kolumnowy, co oznacza, że każda cecha jest reprezentowana przez kolumnę, a każda obserwacja jest reprezentowana przez wiersz. Jako parametry przujmuje się kolumny 1-3. Kolumna 4 zawiera prawdziwą wartość klasyfikacji.



Rysunek 1 Ilość poszczególnych klasyfikatorów w zbiorze Haberman

## Przygotowanie danych

W zbiorze Haberman podczas przygtowywania danych nie napotkano na błędne rekordy lub brakuące dane. Ujednolicono dane przeprowadzając normalizację, według wzoru:

- Znormalizowana wartość cechy

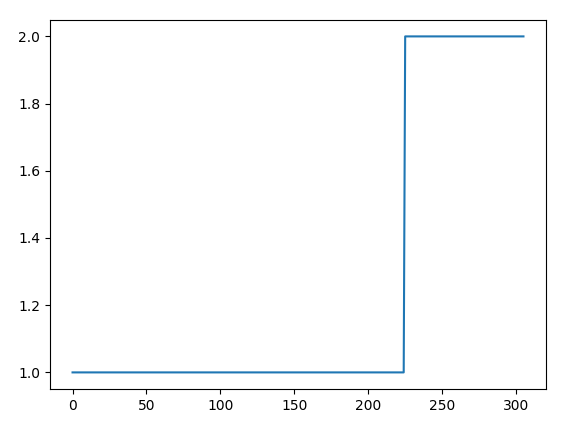
- maksymalna wartość znormalizowanej cechy (1)

- minimalna wartość znormalizowanej cechy (-1)

- maksymalna wartość cechy nieznormalizowana

- minimalna wartość cechy nieznormalizowana

Przed normalizacją wartości cechy 1 w przedziale [30, 83], cechy 2 w przedziale [58, 69], dla 3 cechy w przedziale [0, 52]. Dzięki normalizacji otrzymaliśmy znormalizowane wartości cech w przedziale od -1 do 1. Na koniec dane poddano sortowaniu, zostały one ułożone względem identyfikatorów statusu przeżycia, najpierw te z wartością „1”, a pózniej z wartością „2”, co przedstawia poniższy wykres.

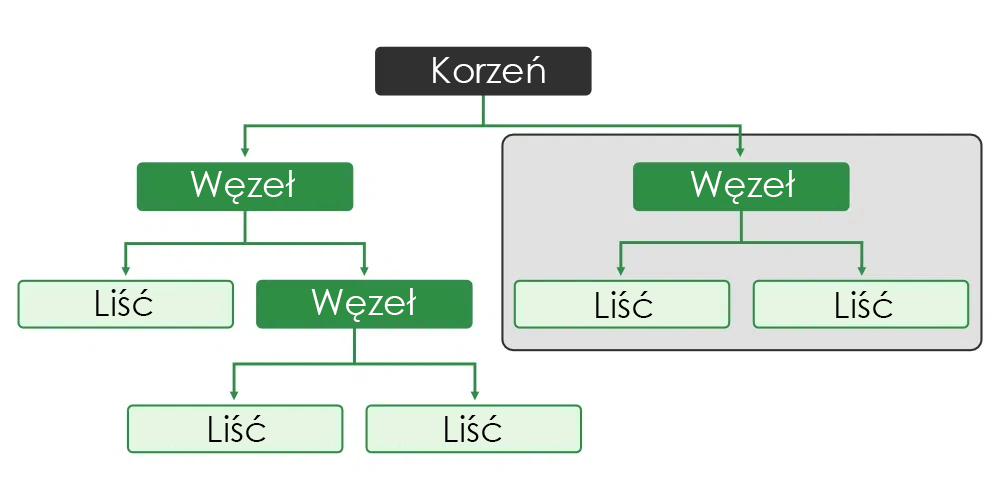


Rysunek 2 Wykres po rozkladu cechy wyjściowej po posortowaniu

# Zagadnienia teoretyczne

## Drzewo decyzyjne

Drzewo decyzyjne jest popularną strukturą używaną w uczeniu maszynowym. Opisuje ona w sposób graficzny klasyfikator, czyli funkcję, która zbiorowi cech obiektu przypisuje etykietę klasy. Reprezentuja ona możliwe rozwiązania w oparciu o określone warunki.



Rysunek 3 Reprezentacja drzewa decyzyjnego

Drzewo decyzyjne złożone jest z:

* **Węzłów** – reprezentują one testy przeprowadzane na wartościach atrybutów próbek zbioru danych. Tutaj dokonuje się podziału danych na mniejsze grupy w zależności od wyniku testu.
* **Gałęzi** - wychodzące z węzłów reprezentują różne wyniki przeprowadzanych testów. Każda gałąź prowadzi do kolejnego węzła lub liścia na podstawie rezultatu testu na nadrzędnym węźle.
* **Liści** - to węzły końcowe, które utożsamiane są z etykietami klasowymi. Liście są miejscem, gdzie podejmowana jest ostateczna decyzja dla danej ścieżki.

Kontrukcja drzewa decyzyjnego polega na kolejnym przechodzeniu przez poziomy drzewa podejmując decyzje o utworzeniu węzła lub liścia, rozpoczynając od korzenia, aby zapewnić możliwe wysoką sprawność klasyfikatora. W zależności od tego, czy kryterium stopu jest spełnione. Jeśli tak, tworzony jest liść i dobierana jest do niego etykieta, w przeciwnym przypadku tworzony jest węzeł i wybierany jest dla niego test. **Kryterium stopu** jest spełnione, jeśli:

* Aktualny zbiór przypadków zawiera wyłącznie przykłady jednej klasy,
* Wyraźna większość przypadków ma te samą klasę,
* Wyczerpał się zbiór testów.

Uzyskane drzewo powinno charakteryzować się możliwie niewielkim rozmiarem, uczenie polega zatem na wyborze takich testów, aby możliwie szybko i z jak najmniejszym błędem osiągnąć liście wskazujące etykiety klas.

Ważną rolę w tworzeniu drzewa decyzyjnego odgrywa wybranie odpowiedniego **testu**, ponieważ to on w największym stopniu decyduje o własnościach uzyskanego rozwiązania. Można wyróznić testy oparte o wiele bądź jeden atrybut. Częściej stosowane są te oparte o jeden atrybut, prowadzą one do dwóch możliwych rozwiązań. Ważnym problemem jest również wybór atrybutu. Najlepszy będzie taki, który zapewnia maksymalną rozróżnialność elementów różnych klas, a z drugiej jego użycie pozwala generować drzewo o prostej strukturze.