**WYŻSZA SZKOŁA INFORMATYKI**

**I ZARZĄDZANIA**

Z SIEDZIBĄ W RZESZOWIE

**Drzewa decyzji**

Prowadzący: dr inż. Mariusz Wrzesień Wykonawca: Oleh Korsunskyi w58957

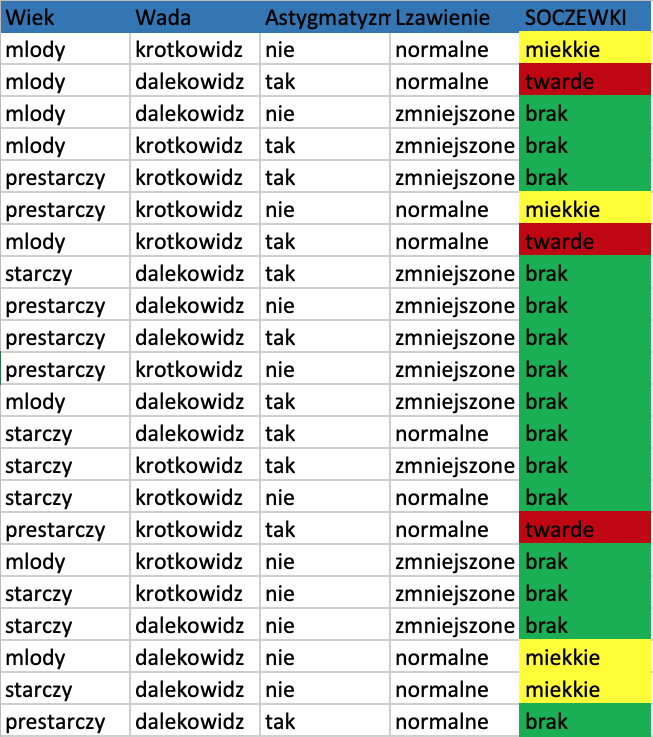
Rzeszów 2019

# Opis problemu

Celem laboratorium jest przygotowanie drzewa decyzji na podstawie danych dotyczących problemu doboru soczewek. Drzewo powinno być zbudowane na podstawie rekurencyjnego algorytmu, który buduje węzły decyzyjne, aż do momentu kiedy wszystkie przypadki ze zbioru uczącego zostaną pokryte. Atrybuty do węzłów decyzyjnych wybierane na podstawie kryterium przyrostu informacji (entropii). Drzewo decyzji jest jedną z alternatyw strukturalnej informacji atrybut-wartość. Ponadto należy obliczyć średnią liczbę pytań dla drzewa przedstawionego w prezentacji dla tego laboratorium.

# Badane zbiory danych

Dane, z których ma być wygenerowane drzewo decyzji, dotyczą problemu doboru soczewek. Są one przedstawione za pomocą tablicy decyzji, która składa się z pięciu kolumn i dwudziestu dwóch wierszy. Wszystkie wiersze to cechowy opis poszczególnych obiektów, natomiast kolumny to poszczególne cechy. Kolumna pierwsza reprezentuje atrybut porządkowy o nazwie „wiek”. Zawiera trzy wartości atrybutów, tj. młody, starczy i prestarczy. Kolumna druga reprezentuje atrybut binarny „wada wzroku”. Posiada dwie wartości, dalekowidz oraz krótkowidz. Kolejna kolumna reprezentuje atrybut binarny „astygmatyzm” i dysponuje także dwoma wartościami, tak i nie. Przedostatnia z kolumn reprezentuje atrybut binarny „łzawienie” wraz z dwoma wartościami, normalne i zmniejszone. Ostatnia kolumna jest kolumną decyzji. Składa się z trzech klas: brak, miękkie, twarde.



Średnią liczbę pytań trzeba obliczyć dla drzewa decyzji, które jest podane w prezentacji, na podstawie bazy Iris. Baza ta składa się ze 150 przypadków. Każdy przypadek jest opisany za pomocą 4 cech: petal length, petal width, sepal length, sepal width. Cały zbiór jest podzielony na 3 klasy: Iris-setosa, Iris-versicolor, Iris-virginica. Każda klasa jest przedstawiona 50 przypadkami.

Drzewo dla którego trzeba policzyć średnią liczbę pytań



# Obliczenia

Tu zostanie opisany algorytm wyboru atrybutu dla pierwszego węzła decyzyjnego, czyli dla korzenia. Inne atrybuty są wybierane w taki sam sposób.

Najpierw liczymy entropię przed podziałem dla całego zbioru.

Info[15,4,3]= -4/22 \* LOG(4/22;2) - 3/22\*LOG(3/22;2)-15/22\*LOG(15/22;2)

Dalej szukamy wszystkie możliwe podziały.

|  |  |  |  |
| --- | --- | --- | --- |
| Wiek | Wada | Astygmatyzm | Lzawienie |
| mlody | krotkowidz | nie | normalne |
| prestarczy | dalekowidz | tak | zmniejszone |
| starczy |  |  |  |

Dla każdego podziału należy policzyć entropię informacji, i wybrać atrybut z największym przyrostem informacji. Wzór entropii:

Kolejno dla atrybutów wykonano obliczenia:

Dla Wiek = młody

Info([2,2,4])= -2/8\*LOG(2/8;2)-2/8\*LOG(2/8;2)-4/8\*LOG(4/8;2)= 1,5

Dla Wiek = prestarczy

Info([1,1,5]) =-1/7\*LOG(1/7;2)-1/7\*LOG(1/7;2)-5/7\*LOG(5/7;2)= 1,14

Dla Wiek = starczy

Info([1,0,6])= -1/7\*LOG(1/7;2)-6/7\*LOG(6/7;2) = 0,59 bitów

Wartość dla atrybutu Wiek

Info([2,2,4],[1,1,5],[1,0,6])=(8/22)\*1,5 + (7/22)\*1,14 + (7/22)\*0,59= 1,09 bitu

Wartość dla całej tablicy decyzji

Info([4,3,15])=entropy(4/22,3/22,15/22)=-4/22 log2(4/22) -3/22 log2(3/22) -15/22 log2(15/22) =1,22bitu

Przyrost informacji= Info([15,4,3]) - Info([2,2,4],[1,1,5],[1,0,6])=1,5 – 1,09 = 0,11

Analogiczne obliczenia wykonujemy dla pozostałych atrybutów.

Na podstawie wszystkich obliczeń można zauważyć, że największy Przyrost Informacji ma atrybut Łzawienie, czyli jest to najistotniejszy z atrybutów opisujących. W związku z powyższym będzie on korzeniem naszego drzewa.

Wybranie atrybutu Łzawienie jako korzenia spowodowało utworzenie dwóch podzbiorów z czego jeden jest liściem, który opisuje 12 z 15 obiektów klasy Brak.

Tak samo stosujemy ten algorytm dla wszystkich podziałów i dla wszystkich podzbiorów, aż do momentu kiedy cały zbiór uczący będzie opisany drzewem

Do sprawdzenia struktury drzewa stosuje się parametr zwany średnią liczbą pytań, którego wzór wygląda następująco:

gdzie:

Sk(xi) – rząd i-tej alternatywy

(pi) - prawdopodobieństwo i-tej alternatywy

Poniższe obliczenia dotyczą drzewa przedstawionego w prezentacji do tego laboratorium.

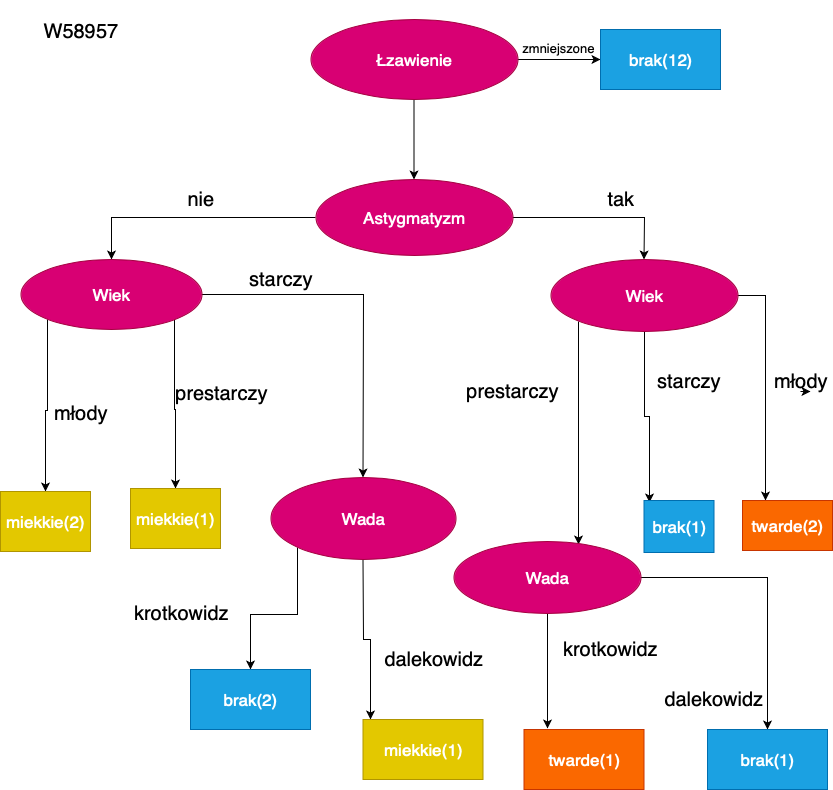
Liczba wszystkich przypadków: 150

E(Sk)=3\*(43/150) +4\*(2/150) +4\*(1/150) +3\*(2/150) +5\*(2/150) +5\*(2/150) +5\*(1/150) +5\*(47/150) +1\*(50/150) = 3,05

Średnia liczba pytań dla drzewa przedstawionego w prezentacji wynosi w przybliżeniu 3,05. Oznacza to, że aby klasyfikować obiekt musimy średnio przejść prze 3 węzły decyzyjne.

# Wyniki

W wyniku opisanego powyżej algorytmu powstało następujące drzewo decyzji:



# Wnioski

Drzewo decyzji to graficzny sposób przedstawienia wiedzy. Dzięki takiemu rozwiązaniu, zbudowanemu na podstawie zbioru uczącego, można sklasyfikować nowe obiekty, które nie brały udziału w procesie tworzenia drzewa.

Dla budowania drzewa został wykorzystany algorytm rekurencyjny, który dzielił zbiór pierwotny na mniejsze zbiory za pomocą wartości atrybutów aż do momentu kiedy w jednym zbiorze zostawały obiekty tylko jednej klasy, wtedy algorytm budował liść z etykietą tej klasy.

Dla danych z powyższego zbioru uczącego otrzymano drzewo składające się z 6 węzłów decyzyjnych oraz 9 węzłów terminalnych, inaczej mówiąc liści.

Drzewa decyzji są świetnym rozwiązaniem dla problemów klasyfikacji obiektów. Istnieje dużo różnych rodzajów i algorytmów drzew decyzji takich jak: ID3, C4.5. Również jest bardzo dużo kryteriów informatywności. Postać graficzna drzewa decyzji jest czytelna i zrozumiała dla człowieka, a poza tym te drzewa można mapować w reguły.

Największą wadą algorytmów budowania drzew decyzji jest to, że te algorytmy są algorytmami zachłannymi. To powoduje sytuacje w których algorytm od razu próbuje oddzielić jedną klasę, lokalnie to najlepsze rozwiązanie, ale globalnie to może wszystko zepsuć i my już nie dostaniemy optymalną strukturę drzewa.

A ocenić strukturę drzewa pomaga nam metoda walidacji zwana średnia liczba pytań. Metoda ta pokazuje przez jaką średnią liczbę węzłów decyzyjnych musi przejść obiekt w danym drzewie decyzji, zanim dostanie się do liścia. Za pomocą tej metody możemy porównywać różne drzewa decyzji i szukać najlepszej struktury. Im mniejszą średnią liczbę pytań ma drzewo tym lepiej.