Wybrane zadania przygotowujące do egzaminu z EUSI - cz. 2

dr Piotr Wąsiewicz

1. Ze zbioru treningowego podanego w tabeli poniżej wykreować metodą zstępującej konstrukcji drzewo decyzyjne (jak najmniej rozbudowane - minimalizacja entropii). Atrybut wiek zdyskretyzować korzystając z dwóch progów 30 i 65 lat. Atrybut ryzyko będzie kategorią.

x	wiek	samochód	ryzyko
1	18	maluch	duże
2	35	maluch	małe
3	50	sportowy	duże
4	66	minivan	duże
5	18	sportowy	duże
6	35	minivan	małe
7	60	maluch	małe
8	70	sportowy	duże
9	25	minivan	małe

Rozwiązanie:
Atrybut wiek otrzymuje po dyskretyzacji trzy wartości:
w_1: wiek < 30, w_2: wiek $\geqslant 30 \wedge$ wiek < 65, w_3: wiek $\geqslant 65$.
Najpierw obliczana jest informacja zawarta w zbiorze i entropie rozkładu wartości kategorii tzw. etykiet między wybrane przez wartości atrybutów podzbiory zbioru trenującego.

$I(P) = -\frac{|P^{mae}|}{|P|}\log_2(\frac{|P^{mae}|}{|P|}) - \frac{|P^{due}|}{|P|}\log_2(\frac{|P^{due}|}{|P|}) = -\frac{4}{9}\log_2(\frac{4}{9}) - \frac{5}{9}\log_2(\frac{5}{9}) = 0.991,$
$E_{\mathsf{wiek,w_1}}(P) = -\frac{|P^{mae}_{\mathsf{wiek,w_1}}|}{|P_{\mathsf{wiek,w_1}}|}\log_2(\frac{|P^{mae}_{\mathsf{wiek,w_1}}|}{|P_{\mathsf{wiek,w_1}}|}) - \frac{|P^{due}_{\mathsf{wiek,w_1}}|}{|P_{\mathsf{wiek,w_1}}|}\log_2(\frac{|P^{due}_{\mathsf{wiek,w_1}}|}{|P_{\mathsf{wiek,w_1}}|}) = -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) = 0.918,$
$E_{\mathsf{wiek,w_2}}(P) = -\frac{|P^{mae}_{\mathsf{wiek,w_2}}|}{|P_{\mathsf{wiek,w_2}}|}\log_2(\frac{|P^{mae}_{\mathsf{wiek,w_2}}|}{|P_{\mathsf{wiek,w_2}}|}) - \frac{|P^{due}_{\mathsf{wiek,w_2}}|}{|P_{\mathsf{wiek,w_2}}|}\log_2(\frac{|P^{due}_{\mathsf{wiek,w_2}}|}{|P_{\mathsf{wiek,w_2}}|}) = -\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4}) = 0.811,$
$E_{\mathsf{wiek,w_3}}(P) = -\frac{|P^{mae}_{\mathsf{wiek,w_3}}|}{|P_{\mathsf{wiek,w_3}}|}\log_2(\frac{|P^{mae}_{\mathsf{wiek,w_3}}|}{|P_{\mathsf{wiek,w_3}}|}) - \frac{|P^{due}_{\mathsf{wiek,w_3}}|}{|P_{\mathsf{wiek,w_3}}|}\log_2(\frac{|P^{due}_{\mathsf{wiek,w_3}}|}{|P_{\mathsf{wiek,w_3}}|}) = -\frac{0}{2}\log_2(\frac{0}{2}) - \frac{2}{2}\log_2(\frac{2}{2}) = 0,$
$E_{\mathsf{samochód,maluch}}(P) = -\frac{|P^{mae}_{\mathsf{samochód,maluch}}|}{|P_{\mathsf{samochód,maluch}}|}\log_2(\frac{|P^{mae}_{\mathsf{samochód,maluch}}|}{|P_{\mathsf{samochód,maluch}}|})$-
$\frac{|P^{due}_{\mathsf{samochód,maluch}}|}{|P_{\mathsf{samochód,maluch}}|}\log_2(\frac{|P^{due}_{\mathsf{samochód,maluch}}|}{|P_{\mathsf{samochód,maluch}}|}) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.918,$
$E_{\mathsf{samochód,minivan}}(P) = -\frac{|P^{mae}_{\mathsf{samochód,minivan}}|}{|P_{\mathsf{samochód,minivan}}|}\log_2(\frac{|P^{mae}_{\mathsf{samochód,minivan}}|}{|P_{\mathsf{samochód,minivan}}|}) -$
$\frac{|P^{due}_{\mathsf{samochód,minivan}}|}{|P_{\mathsf{samochód,minivan}}|}\log_2(\frac{|P^{due}_{\mathsf{samochód,minivan}}|}{|P_{\mathsf{samochód,minivan}}|}) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.918,$
$E_{\mathsf{samochód,sportowy}}(P) = -\frac{|P^{mae}_{\mathsf{samochód,sportowy}}|}{|P_{\mathsf{samochód,sportowy}}|}\log_2(\frac{|P^{mae}_{\mathsf{samochód,sportowy}}|}{|P_{\mathsf{samochód,sportowy}}|}) -$
$\frac{|P^{due}_{\mathsf{samochód,sportowy}}|}{|P_{\mathsf{samochód,sportowy}}|}\log_2(\frac{|P^{due}_{\mathsf{samochód,sportowy}}|}{|P_{\mathsf{samochód,sportowy}}|}) = -\frac{0}{3}\log_2(\frac{0}{3}) - \frac{3}{3}\log_2(\frac{3}{3}) = 0,$

Następnie obliczane są średnie ważone entropie:

$$E_{\mathsf{wiek}}(P) = \frac{|P_{\mathsf{wiek,w_1}}|}{|P|}E_{\mathsf{wiek,w_1}}(P) + \frac{|P_{\mathsf{wiek,w_2}}|}{|P|}E_{\mathsf{wiek,w_2}}(P) + \frac{|P_{\mathsf{wiek,w_3}}|}{|P|}E_{\mathsf{wiek,w_3}}(P) = \frac{3}{9}(0.918) + \frac{4}{9}(0.811) + \frac{2}{9}0 = 0,666,$$

$$E_{\mathsf{samochod}}(P) = \frac{|P_{\mathsf{samochod,maluch}}|}{|P|}E_{\mathsf{samochod,maluch}}(P) + \frac{|P_{\mathsf{samochod,minivan}}|}{|P|}E_{\mathsf{samochod,minivan}}(P) + \frac{|P_{\mathsf{samochod,sportowy}}|}{|P|}E_{\mathsf{samochod,sportowy}}(P) = \frac{3}{9}(0.918) + \frac{3}{9}(0.918) + \frac{3}{9}0 = 0,612,$$

I wartości infomacyjne dla poszczególnych atrybutów:

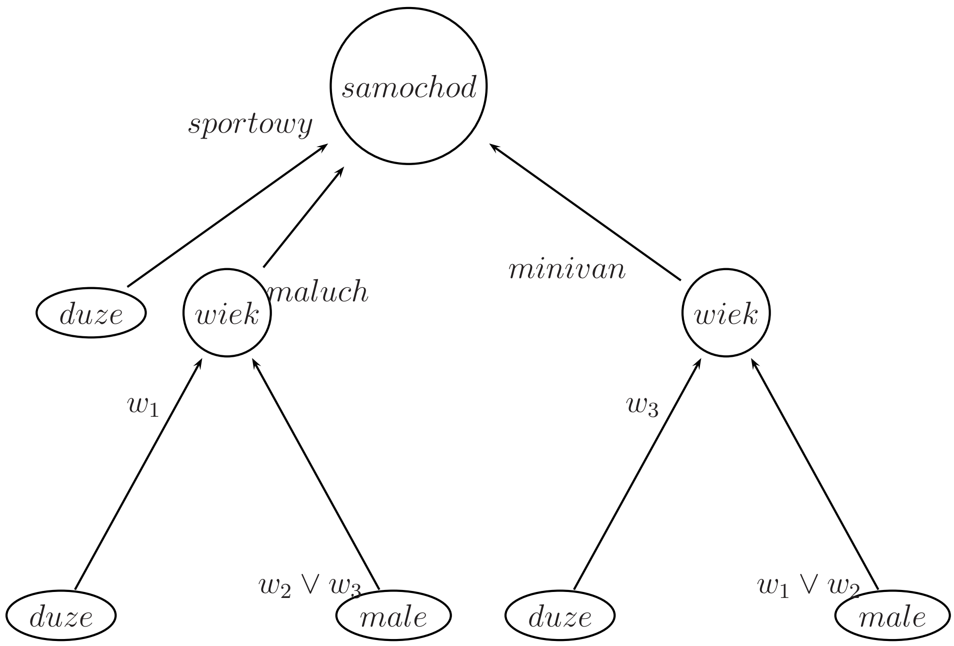
$$IV_{\mathsf{wiek}}(P) = -\frac{|P_{\mathsf{wiek,w_1}}|}{|P|}\log_2(\frac{|P_{\mathsf{wiek,w_1}}|}{|P|}) - \frac{|P_{\mathsf{wiek,w_2}}|}{|P|}\log_2(\frac{|P_{\mathsf{wiek,w_2}}|}{|P|}) - \frac{|P_{\mathsf{wiek,w_3}}|}{|P|}\log_2(\frac{|P_{\mathsf{wiek,w_3}}|}{|P|}) = -\frac{3}{9}\log_2(\frac{3}{9}) - \frac{4}{9}\log_2(\frac{4}{9}) - \frac{2}{9}\log_2(\frac{2}{9}) = 0,528 + 0,519 + 0,482 = 1,53,$$

$$IV_{\mathsf{samochód}}(P) = -\frac{|P_{\mathsf{samochód,maluch}}|}{|P|}\log_2(\frac{|P_{\mathsf{samochód,maluch}}|}{|P|}) - \frac{|P_{\mathsf{samochód,minivan}}|}{|P|}\log_2(\frac{|P_{\mathsf{samochód,minivan}}|}{|P|}) - \frac{|P_{\mathsf{samochód,sportowy}}|}{|P|}\log_2(\frac{|P_{\mathsf{samochód,sportowy}}|}{|P|}) = -\frac{3}{9}\log_2(\frac{3}{9}) - \frac{3}{9}\log_2(\frac{3}{9}) - \frac{3}{9}\log_2(\frac{3}{9}) = 0,528 + 0,528 + 0,528 = 1,584,$$

Na końcu współczynniki przyrostu informacji wynoszą odpowiednio:

$$\vartheta_{\mathsf{wiek}}(P) = \frac{I(P) - E_{\mathsf{wiek}}(P)}{IV_{\mathsf{wiek}}(P)} = \frac{0,991 - 0,666}{1,53} = 0,212$$

$$\vartheta_{\mathsf{samochód}}(P) = \frac{I(P) - E_{\mathsf{samochód}}(P)}{IV_{\mathsf{samochód}}(P)} = \frac{0,991 - 0,612}{1,584} = 0,239$$



Jak widać atrybut samochód ma większy współczynnik i wygrywa staje się pierwszym węzłem drzewa decyzyjnego, a jego trzy łuki biegnące do następników mają za nazwy jego wartości.

Dla wartości sportowy każdy przykład zawierający ją ma etykietę duże atrybutu ryzyko, stąd jej łuk kończy się liściem o wartości duże.

Dla wartości maluch jej łuk kończy się z braku jasnego wyboru etykiety tylko na podstawie wartości atrybutu samochód węzłem atrybutu wiek - ostatnim z dostępnych testów na drodze do określenia etykiety przykładu złożonego z testowanych dwóch atrybutów wiek i samochód. Poniżej zamieszczony został opis następników nowego węzła.

Przykłady z wartością w_1 atrybutu wiek i wartością maluch mają zawsze etykietę duże stąd łuk biegnący od węzła wiek o nazwie w_1 kończy się liściem duże, a dla innych wartości atrybutu wiek przy wartości maluch atrybutu samochód przykłady mają etykiety małe stąd odpowiednie liście.

Wracając do trzeciego łuku o nazwie minivan biegnącego od korzenia można zauważyć, że też z braku takich samych etykiet dla przykładów z wartością minivan i z dowolną wartością atrybutu wiek łuk ten kończy się węzłem o nazwie wiek i dalej zależności i liście są takie same jak dla węzła kończącego łuk maluch.

2. Za pomocą algorytmu sekwencyjnego pokrywania CN2 uzyskać <u>nieuporządkowany</u> zbiór zdaniowych reguł ze zbioru treningowego podanego w tabeli poniżej. Opisać dokładnie kolejne kroki algorytmu. Atrybut wiek zdyskretyzować korzystając z dwóch progów 30 i 65 lat. Atrybut ryzyko będzie kategorią. Dla ułatwienia założyć, że wszystkie kompleksy są istotne statystycznie oraz że kompleks warunkujący z reguły zdaniowej musi pokrywać przykłady tylko z jedną etykietą - jedną wartością kategorii.

x	wiek	samochód	ryzyko
1	18	maluch	duże
2	35	maluch	małe
3	50	sportowy	duże
4	66	minivan	duże
5	18	sportowy	duże
6	35	minivan	małe
7	60	maluch	małe
8	70	sportowy	duże
9	25	minivan	małe

Rozwiązanie:
Atrybut wiek otrzymuje po dyskretyzacji trzy wartości:

- w_1: wiek < 30,
- w_2: wiek $\geqslant 30 \wedge$ wiek < 65,
- w_3: wiek $\geqslant 65$.

Zbiór $\mathbb{S}$ kompleksów atomowych (czyli tylko z jednym selektorem nieuniwersalnym) ($\mathbb{S} = \{\mathbb{K}_1, \mathbb{K}_2, \mathbb{K}_3, \mathbb{K}_4, \mathbb{K}_5, \mathbb{K}_6, \mathbb{K}_7, \mathbb{K}_8, \mathbb{K}_9, \mathbb{K}_{10}, \mathbb{K}_{11}, \mathbb{K}_{12}\}$) jest następujący:

	$\mathbb{S} = \{$
$\mathbb{K}_1$	$< w_1, ? >$,
$\mathbb{K}_2$	$< w_2, ? >$,
$\mathbb{K}_3$	$< w_3, ? >$,
$\mathbb{K}_4$	$< w_1 \vee w_2, ? >$,
$\mathbb{K}_5$	$< w_2 \vee w_3, ? >$,
$\mathbb{K}_6$	$< w_1 \vee w_3, ? >$,
$\mathbb{K}_7$	$<?,$ maluch $>$,
$\mathbb{K}_8$	$<?,$ minivan $>$,
$\mathbb{K}_9$	$<?,$ sportowy $>$,
$\mathbb{K}_{10}$	$<?,$ maluch $\vee$ minivan $>$,
$\mathbb{K}_{11}$	$<?,$ minivan $\vee$ sportowy $>$,
$\mathbb{K}_{12}$	$<?,$ maluch $\vee$ sportowy $>\}$

Kolejne kroki algorytmu CN2

(a) Początkowo $R = \phi, P = T = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}, \mathbb{S}$

(b) Następuje wywołanie *znajdź-kompleks*(T, P).

- $S = \{<?>\} \neq \phi, k_* = <?>$
  $\vartheta_{k_*}(P) = -E_{k_*}(P) = \frac{|P^{mae}|}{|P|}\log_2(\frac{|P^{mae}|}{|P|}) + \frac{|P^{due}|}{|P|}\log_2(\frac{|P^{due}|}{|P|}) = \frac{5}{9}\log_2(\frac{5}{9}) + \frac{4}{9}\log_2(\frac{4}{9}) = -0.991,$
- $S' = \mathbb{S} = S \cap \mathbb{S},$
  Ze względu na to, że dąży się do uzyskania nieuporządkowanego zbioru reguł funkcje oceny kompleksów atomowych są liczone tylko raz w zbiorze T i potem cały czas wykorzystywane.
  $\vartheta_{\mathbb{K}_1}(T) = -E_{\mathbb{K}_1}(T) = \frac{|T^{mae}_{\mathbb{K}_1}|}{|T_{\mathbb{K}_1}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_1}|}{|T_{\mathbb{K}_1}|}) + \frac{|T^{due}_{\mathbb{K}_1}|}{|T_{\mathbb{K}_1}|}\log_2(\frac{|T^{due}_{\mathbb{K}_1}|}{|T_{\mathbb{K}_1}|}) = \frac{1}{3}\log_2(\frac{1}{3}) + \frac{2}{3}\log_2(\frac{2}{3}) = -0.918,$
  $\vartheta_{\mathbb{K}_2}(T) = -E_{\mathbb{K}_2}(T) = \frac{|T^{mae}_{\mathbb{K}_2}|}{|T_{\mathbb{K}_2}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_2}|}{|T_{\mathbb{K}_2}|}) + \frac{|T^{due}_{\mathbb{K}_2}|}{|T_{\mathbb{K}_2}|}\log_2(\frac{|T^{due}_{\mathbb{K}_2}|}{|T_{\mathbb{K}_2}|}) = \frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4}) = -0.811,$
  $\vartheta_{\mathbb{K}_3}(T) = -E_{\mathbb{K}_3}(T) = \frac{|T^{mae}_{\mathbb{K}_3}|}{|T_{\mathbb{K}_3}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_3}|}{|T_{\mathbb{K}_3}|}) + \frac{|T^{due}_{\mathbb{K}_3}|}{|T_{\mathbb{K}_3}|}\log_2(\frac{|T^{due}_{\mathbb{K}_3}|}{|T_{\mathbb{K}_3}|}) = \frac{0}{3}\log_2(\frac{0}{3}) + \frac{3}{3}\log_2(\frac{3}{3}) = 0,$
  $\vartheta_{\mathbb{K}_4}(T) = -E_{\mathbb{K}_4}(T) = \frac{|T^{mae}_{\mathbb{K}_4}|}{|T_{\mathbb{K}_4}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_4}|}{|T_{\mathbb{K}_4}|}) + \frac{|T^{due}_{\mathbb{K}_4}|}{|T_{\mathbb{K}_4}|}\log_2(\frac{|T^{due}_{\mathbb{K}_4}|}{|T_{\mathbb{K}_4}|}) = \frac{4}{7}\log_2(\frac{4}{7}) + \frac{3}{7}\log_2(\frac{3}{7}) = -0.985,$
  $\vartheta_{\mathbb{K}_5}(T) = -E_{\mathbb{K}_5}(T) = \frac{|T^{mae}_{\mathbb{K}_5}|}{|T_{\mathbb{K}_5}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_5}|}{|T_{\mathbb{K}_5}|}) + \frac{|T^{due}_{\mathbb{K}_5}|}{|T_{\mathbb{K}_5}|}\log_2(\frac{|T^{due}_{\mathbb{K}_5}|}{|T_{\mathbb{K}_5}|}) = \frac{3}{6}\log_2(\frac{3}{6}) + \frac{3}{6}\log_2(\frac{3}{6}) = -1,$
  $\vartheta_{\mathbb{K}_6}(T) = -E_{\mathbb{K}_6}(T) = \frac{|T^{mae}_{\mathbb{K}_6}|}{|T_{\mathbb{K}_6}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_6}|}{|T_{\mathbb{K}_6}|}) + \frac{|T^{due}_{\mathbb{K}_6}|}{|T_{\mathbb{K}_6}|}\log_2(\frac{|T^{due}_{\mathbb{K}_6}|}{|T_{\mathbb{K}_6}|}) = \frac{1}{5}\log_2(\frac{1}{5}) + \frac{4}{5}\log_2(\frac{4}{5}) = -0.721,$
  $\vartheta_{\mathbb{K}_7}(T) = -E_{\mathbb{K}_7}(T) = \frac{|T^{mae}_{\mathbb{K}_7}|}{|T_{\mathbb{K}_7}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_7}|}{|T_{\mathbb{K}_7}|}) + \frac{|T^{due}_{\mathbb{K}_7}|}{|T_{\mathbb{K}_7}|}\log_2(\frac{|T^{due}_{\mathbb{K}_7}|}{|T_{\mathbb{K}_7}|}) = \frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3}) = -0.918,$
  $\vartheta_{\mathbb{K}_8}(T) = -E_{\mathbb{K}_8}(T) = \frac{|T^{mae}_{\mathbb{K}_8}|}{|T_{\mathbb{K}_8}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_8}|}{|T_{\mathbb{K}_8}|}) + \frac{|T^{due}_{\mathbb{K}_8}|}{|T_{\mathbb{K}_8}|}\log_2(\frac{|T^{due}_{\mathbb{K}_8}|}{|T_{\mathbb{K}_8}|}) = \frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3}) = -0.918,$
  $\vartheta_{\mathbb{K}_9}(T) = -E_{\mathbb{K}_9}(T) = \frac{|T^{mae}_{\mathbb{K}_9}|}{|T_{\mathbb{K}_9}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_9}|}{|T_{\mathbb{K}_9}|}) + \frac{|T^{due}_{\mathbb{K}_9}|}{|T_{\mathbb{K}_9}|}\log_2(\frac{|T^{due}_{\mathbb{K}_9}|}{|T_{\mathbb{K}_9}|}) = \frac{0}{3}\log_2(\frac{0}{3}) + \frac{3}{3}\log_2(\frac{3}{3}) = 0,$
  $\vartheta_{\mathbb{K}_{10}}(T) = -E_{\mathbb{K}_{10}}(T) = \frac{|T^{mae}_{\mathbb{K}_{10}}|}{|T_{\mathbb{K}_{10}}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_{10}}|}{|T_{\mathbb{K}_{10}}|}) + \frac{|T^{due}_{\mathbb{K}_{10}}|}{|T_{\mathbb{K}_{10}}|}\log_2(\frac{|T^{due}_{\mathbb{K}_{10}}|}{|T_{\mathbb{K}_{10}}|}) = \frac{4}{6}\log_2(\frac{4}{6}) + \frac{2}{6}\log_2(\frac{2}{6}) = -0.918,$

$\vartheta_{\mathbb{K}_{11}}(T) = -E_{\mathbb{K}_{11}}(T) = \frac{|T^{mae}_{\mathbb{K}_{11}}|}{|T_{\mathbb{K}_{11}}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_{11}}|}{|T_{\mathbb{K}_{11}}|}) + \frac{|T^{due}_{\mathbb{K}_{11}}|}{|T_{\mathbb{K}_{11}}|}\log_2(\frac{|T^{due}_{\mathbb{K}_{11}}|}{|T_{\mathbb{K}_{11}}|}) = \frac{2}{6}\log_2(\frac{2}{6}) + \frac{4}{6}\log_2(\frac{4}{6}) = -0.918$,

$\vartheta_{\mathbb{K}_{12}}(T) = -E_{\mathbb{K}_{12}}(T) = \frac{|T^{mae}_{\mathbb{K}_{12}}|}{|T_{\mathbb{K}_{12}}|}\log_2(\frac{|T^{mae}_{\mathbb{K}_{12}}|}{|T_{\mathbb{K}_{12}}|}) + \frac{|T^{due}_{\mathbb{K}_{12}}|}{|T_{\mathbb{K}_{12}}|}\log_2(\frac{|T^{due}_{\mathbb{K}_{12}}|}{|T_{\mathbb{K}_{12}}|}) = \frac{2}{6}\log_2(\frac{2}{6}) + \frac{4}{6}\log_2(\frac{4}{6}) = -0.918$

- $\mathbb{K}_9 =<?,$ sportowy $>$ ma największą wartość $\vartheta = 0$ w zbiorze $\mathbb{S}$ razem z $\mathbb{K}_3$, ale więcej przykładów pokrywa; $S = \{\mathbb{K}_9\}, k_* = \mathbb{K}_9$,

(c) $R = \{<?,$ sportowy $>\rightarrow$ duże$\}$, $P = \{1, 2, 4, 6, 7, 9\}$,

(d) $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),

- $S = \{<?>\} \neq \phi, k_* =<?>$ i $\vartheta_{k_*}(P) = -0.991$,
- $S' = \mathbb{S} = S \cap \mathbb{S}$,
  ze względu na użycie $\mathbb{K}_9$ wyklucza się wszystkie kompleksy atomowe z wartością atrybutu samochód = sportowy czyli $\mathbb{K}_9$,$\mathbb{K}_{11}$,$\mathbb{K}_{12}$, bo takich przykładów z wartością sportowy już w zbiorze P nie ma.
  W następnym kroku chcąc uzyskać najlepszy kompleks wykorzystuje się funckje oceny liczone jeden raz na początku.
- $\mathbb{K}_3 =< w_3, ? >$ ma największą wartość $\vartheta = 0$; $S = \{\mathbb{K}_3\}, k_* = \mathbb{K}_3$,

(e) $R = \{<?,$ sportowy $>\rightarrow$ duże, $< w3, ? >\rightarrow$ duże$\}$, $P = \{1, 2, 6, 7, 9\}$,

(f) $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),

- $S = \{<?>\} \neq \phi, k_* =<?>$ i $\vartheta_{k_*}(P) = -0.991$,
  ze względu na użycie $\mathbb{K}_3$ wyklucza się wszystkie kompleksy atomowe z wartością atrybutu wiek = w_3 czyli $\mathbb{K}_3$,$\mathbb{K}_5$,$\mathbb{K}_6$, bo takich przykładów z wartością w_3 już w zbiorze P nie ma.
- $\mathbb{K}_2 =< w_2, ? >$ ma wartość $\vartheta = -0.811$, ale przyjęto, że dla ułatwienia tworzy się reguły pokrywające przykłady tylko z jedną etykietą czyli dla kompleksów o wartości funkcji oceny 0, dlatego pętla wykonuje się dalej.
  $S = \{< w2, ? >\}$;
- Zgodnie z algorytmem CN2: $S' := S \cap \mathbb{S}$; $S' := S' - S - \{< \phi >\}$;
  Kompleks $\{< w2,$ maluch$\vee$minivan $>\}$ ma wartość funkcji oceny równą 0 i pokrywa najwięcej przykladów z P, gdyż mimo, że ocenia się według zbioru T (zbiór reguł nieuporządkowany), to trzeba tworzyć reguły pokrywające przykłady ze zbioru P i to jak najwięcej.

(g) $R = \{<?,$ sportowy $>\rightarrow$ duże, $< w3, ? >\rightarrow$ duże, $< w2,$ maluch$\vee$minivan $\rightarrow$ małe $>\}$, $P = \{1, 9\}$,

(h) $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),

- $S = \{<?>\} \neq \phi, k_* =<?>$ i $\vartheta_{k_*}(P) = -0.991$,
- Pozostały tylko dwa przykłady o różnych etykietach, aby kompleksy mogły uzyskać ocenę równą 0 muszą mieć identyczne wartości atrybutów, stąd powstają dwie nowe reguły.

(i) Ostatecznie
  $R = \{<?,$ sportowy $>\rightarrow$ duże,
  $< w3, ? >\rightarrow$ duże,
  $< w2,$ maluch $\vee$ minivan $\rightarrow$ małe $>$
  $< w1,$ minivan $\rightarrow$ małe $>$
  $< w1,$ maluch $\rightarrow$ duże $>$
  $\}$

W uzyskanym zbiorze reguł można reguły zamieniać miejscami, gdyż jest to zbiór nieuporządkowany.

3. Za pomocą algorytmu sekwencyjnego pokrywania CN2 uzyskać uporządkowany zbiór zdaniowych reguł ze zbioru treningowego podanego w tabeli poniżej. Opisać dokładnie kolejne kroki algorytmu. Atrybut wiek zdyskretyzować korzystając z dwóch progów 30 i 65 lat. Atrybut ryzyko będzie kategorią. Dla ułatwienia założyć, że wszystkie kompleksy są istotne statystycznie oraz że kompleks warunkujący z reguły zdaniowej musi pokrywać przykłady tylko z jedną etykietą - jedną wartością kategorii.

x	wiek	samochód	ryzyko
1	18	maluch	duże
2	35	maluch	małe
3	50	sportowy	duże
4	66	minivan	duże
5	18	sportowy	duże
6	35	minivan	małe
7	60	maluch	małe
8	70	sportowy	duże
9	25	minivan	małe

Rozwiązanie:

Atrybut wiek otrzymuje po dyskretyzacji trzy wartości:

- w_1: wiek < 30,
- w_2: wiek $\geqslant 30 \wedge$ wiek < 65,
- w_3: wiek $\geqslant 65$.

Zbiór $\mathbb{S}$ kompleksów atomowych (czyli tylko z jednym selektorem nieuniwersalnym) ($\mathbb{S} = \{\mathbb{K}_1, \mathbb{K}_2, \mathbb{K}_3, \mathbb{K}_4, \mathbb{K}_5, \mathbb{K}_6, \mathbb{K}_7, \mathbb{K}_8, \mathbb{K}_9, \mathbb{K}_{10}, \mathbb{K}_{11}, \mathbb{K}_{12}\}$) jest następujący:

	$\mathbb{S} = \{$
$\mathbb{K}_1$	$< w_1, ? >,$
$\mathbb{K}_2$	$< w_2, ? >,$
$\mathbb{K}_3$	$< w_3, ? >,$
$\mathbb{K}_4$	$< w_1 \vee w_2, ? >,$
$\mathbb{K}_5$	$< w_2 \vee w_3, ? >,$
$\mathbb{K}_6$	$< w_1 \vee w_3, ? >,$
$\mathbb{K}_7$	$<?, \text{maluch} >,$
$\mathbb{K}_8$	$<?, \text{minivan} >,$
$\mathbb{K}_9$	$<?, \text{sportowy} >,$
$\mathbb{K}_{10}$	$<?, \text{maluch} \vee \text{minivan} >,$
$\mathbb{K}_{11}$	$<?, \text{minivan} \vee \text{sportowy} >,$
$\mathbb{K}_{12}$	$<?, \text{maluch} \vee \text{sportowy} >\}$

Kolejne kroki algorytmu CN2

(a) Początkowo $R = \phi, P = T = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}, \mathbb{S}$

(b) Następuje wywołanie *znajdź-kompleks*(T, P).

- $S = \{<? >\} \neq \phi, k_* =<? >$
  $\vartheta_{k_*}(P) = -E_{k_*}(P) = \frac{|P^{mae}|}{|P|}\log_2(\frac{|P^{mae}|}{|P|}) + \frac{|P^{due}|}{|P|}\log_2(\frac{|P^{due}|}{|P|}) = \frac{5}{9}\log_2(\frac{5}{9}) + \frac{4}{9}\log_2(\frac{4}{9}) = -0.991,$
- $S' = \mathbb{S} = S \cap \mathbb{S},$
  $\vartheta_{\mathbb{K}_1}(P) = -E_{\mathbb{K}_1}(P) = \frac{|P^{mae}_{\mathbb{K}_1}|}{|P_{\mathbb{K}_1}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_1}|}{|P_{\mathbb{K}_1}|}) + \frac{|P^{due}_{\mathbb{K}_1}|}{|P_{\mathbb{K}_1}|}\log_2(\frac{|P^{due}_{\mathbb{K}_1}|}{|P_{\mathbb{K}_1}|}) = \frac{1}{3}\log_2(\frac{1}{3}) +$

$\frac{2}{3}\log_2(\frac{2}{3}) = -0.918$,

$\vartheta_{\mathbb{K}_2}(P) = -E_{\mathbb{K}_2}(P) = \frac{|P^{mae}_{\mathbb{K}_2}|}{|P_{\mathbb{K}_2}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_2}|}{|P_{\mathbb{K}_2}|}) + \frac{|P^{due}_{\mathbb{K}_2}|}{|P_{\mathbb{K}_2}|}\log_2(\frac{|P^{due}_{\mathbb{K}_2}|}{|P_{\mathbb{K}_2}|}) = \frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4}) = -0.811$,

$\vartheta_{\mathbb{K}_3}(P) = -E_{\mathbb{K}_3}(P) = \frac{|P^{mae}_{\mathbb{K}_3}|}{|P_{\mathbb{K}_3}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_3}|}{|P_{\mathbb{K}_3}|}) + \frac{|P^{due}_{\mathbb{K}_3}|}{|P_{\mathbb{K}_3}|}\log_2(\frac{|P^{due}_{\mathbb{K}_3}|}{|P_{\mathbb{K}_3}|}) = \frac{0}{3}\log_2(\frac{0}{3}) + \frac{3}{3}\log_2(\frac{3}{3}) = 0$,

$\vartheta_{\mathbb{K}_4}(P) = -E_{\mathbb{K}_4}(P) = \frac{|P^{mae}_{\mathbb{K}_4}|}{|P_{\mathbb{K}_4}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_4}|}{|P_{\mathbb{K}_4}|}) + \frac{|P^{due}_{\mathbb{K}_4}|}{|P_{\mathbb{K}_4}|}\log_2(\frac{|P^{due}_{\mathbb{K}_4}|}{|P_{\mathbb{K}_4}|}) = \frac{4}{7}\log_2(\frac{4}{7}) + \frac{3}{7}\log_2(\frac{3}{7}) = -0.985$,

$\vartheta_{\mathbb{K}_5}(P) = -E_{\mathbb{K}_5}(P) = \frac{|P^{mae}_{\mathbb{K}_5}|}{|P_{\mathbb{K}_5}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_5}|}{|P_{\mathbb{K}_5}|}) + \frac{|P^{due}_{\mathbb{K}_5}|}{|P_{\mathbb{K}_5}|}\log_2(\frac{|P^{due}_{\mathbb{K}_5}|}{|P_{\mathbb{K}_5}|}) = \frac{3}{6}\log_2(\frac{3}{6}) + \frac{3}{6}\log_2(\frac{3}{6}) = -1$,

$\vartheta_{\mathbb{K}_6}(P) = -E_{\mathbb{K}_6}(P) = \frac{|P^{mae}_{\mathbb{K}_6}|}{|P_{\mathbb{K}_6}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_6}|}{|P_{\mathbb{K}_6}|}) + \frac{|P^{due}_{\mathbb{K}_6}|}{|P_{\mathbb{K}_6}|}\log_2(\frac{|P^{due}_{\mathbb{K}_6}|}{|P_{\mathbb{K}_6}|}) = \frac{1}{5}\log_2(\frac{1}{5}) + \frac{4}{5}\log_2(\frac{4}{5}) = -0.721$,

$\vartheta_{\mathbb{K}_7}(P) = -E_{\mathbb{K}_7}(P) = \frac{|P^{mae}_{\mathbb{K}_7}|}{|P_{\mathbb{K}_7}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_7}|}{|P_{\mathbb{K}_7}|}) + \frac{|P^{due}_{\mathbb{K}_7}|}{|P_{\mathbb{K}_7}|}\log_2(\frac{|P^{due}_{\mathbb{K}_7}|}{|P_{\mathbb{K}_7}|}) = \frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3}) = -0.918$,

$\vartheta_{\mathbb{K}_8}(P) = -E_{\mathbb{K}_8}(P) = \frac{|P^{mae}_{\mathbb{K}_8}|}{|P_{\mathbb{K}_8}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_8}|}{|P_{\mathbb{K}_8}|}) + \frac{|P^{due}_{\mathbb{K}_8}|}{|P_{\mathbb{K}_8}|}\log_2(\frac{|P^{due}_{\mathbb{K}_8}|}{|P_{\mathbb{K}_8}|}) = \frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3}) = -0.918$,

$\vartheta_{\mathbb{K}_9}(P) = -E_{\mathbb{K}_9}(P) = \frac{|P^{mae}_{\mathbb{K}_9}|}{|P_{\mathbb{K}_9}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_9}|}{|P_{\mathbb{K}_9}|}) + \frac{|P^{due}_{\mathbb{K}_9}|}{|P_{\mathbb{K}_9}|}\log_2(\frac{|P^{due}_{\mathbb{K}_9}|}{|P_{\mathbb{K}_9}|}) = \frac{0}{3}\log_2(\frac{0}{3}) + \frac{3}{3}\log_2(\frac{3}{3}) = 0$,

$\vartheta_{\mathbb{K}_{10}}(P) = -E_{\mathbb{K}_{10}}(P) = \frac{|P^{mae}_{\mathbb{K}_{10}}|}{|P_{\mathbb{K}_{10}}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_{10}}|}{|P_{\mathbb{K}_{10}}|}) + \frac{|P^{due}_{\mathbb{K}_{10}}|}{|P_{\mathbb{K}_{10}}|}\log_2(\frac{|P^{due}_{\mathbb{K}_{10}}|}{|P_{\mathbb{K}_{10}}|}) = \frac{4}{6}\log_2(\frac{4}{6}) + \frac{2}{6}\log_2(\frac{2}{6}) = -0.918$,

$\vartheta_{\mathbb{K}_{11}}(P) = -E_{\mathbb{K}_{11}}(P) = \frac{|P^{mae}_{\mathbb{K}_{11}}|}{|P_{\mathbb{K}_{11}}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_{11}}|}{|P_{\mathbb{K}_{11}}|}) + \frac{|P^{due}_{\mathbb{K}_{11}}|}{|P_{\mathbb{K}_{11}}|}\log_2(\frac{|P^{due}_{\mathbb{K}_{11}}|}{|P_{\mathbb{K}_{11}}|}) = \frac{2}{6}\log_2(\frac{2}{6}) + \frac{4}{6}\log_2(\frac{4}{6}) = -0.918$,

$\vartheta_{\mathbb{K}_{12}}(P) = -E_{\mathbb{K}_{12}}(P) = \frac{|P^{mae}_{\mathbb{K}_{12}}|}{|P_{\mathbb{K}_{12}}|}\log_2(\frac{|P^{mae}_{\mathbb{K}_{12}}|}{|P_{\mathbb{K}_{12}}|}) + \frac{|P^{due}_{\mathbb{K}_{12}}|}{|P_{\mathbb{K}_{12}}|}\log_2(\frac{|P^{due}_{\mathbb{K}_{12}}|}{|P_{\mathbb{K}_{12}}|}) = \frac{2}{6}\log_2(\frac{2}{6}) + \frac{4}{6}\log_2(\frac{4}{6}) = -0.918$

- $\mathbb{K}_9 =<?, \text{sportowy}>$ ma największą wartość $\vartheta = 0$ w zbiorze $\mathbb{S}$ razem z $\mathbb{K}_3$, ale więcej przykładów pokrywa; $S = \{\mathbb{K}_9\}, k_* = \mathbb{K}_9$,

(c) $R = \{<?, \text{sportowy}> \rightarrow \text{duże}\}$, $P = \{1, 2, 4, 6, 7, 9\}$,

(d) $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),

- $S = \{<?>\} \neq \phi, k_* =<?>$ i $\vartheta_{k_*}(P) = -0.918$,
- $S' = \mathbb{S} = S \cap \mathbb{S}$,

  ze względu na użycie $\mathbb{K}_9$ wyklucza się wszystkie kompleksy atomowe z wartością atrybutu samochód = sportowy czyli $\mathbb{K}_9$,$\mathbb{K}_{11}$,$\mathbb{K}_{12}$, bo takich przykładów z wartością

sportowy już w zbiorze P nie ma.
Dla zbioru uporządkowanego trzeba wartość funkcji oceny kompleksów atomowych obliczać przed każdym wyborem najlepszego kompleksu.
$\vartheta_{\mathbb{K}_1}(P) = -1$, $\vartheta_{\mathbb{K}_2}(P) = 0$, $\vartheta_{\mathbb{K}_3}(P) = 0$, $\vartheta_{\mathbb{K}_4}(P) = -0,721$, $\vartheta_{\mathbb{K}_5}(P) = -0,811$, $\vartheta_{\mathbb{K}_6}(P) = -0.918$, $\vartheta_{\mathbb{K}_7}(P) = -0.918$, $\vartheta_{\mathbb{K}_8}(P) = -0.918$, $\vartheta_{\mathbb{K}_{10}}(P) = -0.918$,

- $\mathbb{K}_2 =< w_2, ? >$ ma największą wartość $\vartheta = 0$ razem z $\mathbb{K}_3$, ale więcej przykładów pokrywa; $S = \{\mathbb{K}_2\}, k_* = \mathbb{K}_2$,

(e) $R = \{<?, \text{sportowy} > \rightarrow \text{duże}, < w2, ? > \rightarrow \text{małe}\}$, $P = \{1, 4, 9\}$,

(f) $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),

- $S = \{<? >\} \neq \phi, k_* =<? >$ i $\vartheta_{k_*}(P) = -0.918$,
  ze względu na użycie $\mathbb{K}_2$ wyklucza się wszystkie kompleksy atomowe z wartością atrybutu wiek $= w_2$ czyli $\mathbb{K}_2$,$\mathbb{K}_4$,$\mathbb{K}_5$, bo takich przykładów z wartością w_2 już w zbiorze P nie ma.
  $\vartheta_{\mathbb{K}_1}(P) = -1$, $\vartheta_{\mathbb{K}_3}(P) = 0$, $\vartheta_{\mathbb{K}_6}(P) = -0.918$, $\vartheta_{\mathbb{K}_7}(P) = 0$, $\vartheta_{\mathbb{K}_8}(P) = -1$, $\vartheta_{\mathbb{K}_{10}}(P) = -0.918$,
- $\mathbb{K}_3 =< w_3, ? >$ ma największą wartość $\vartheta = 0$ razem z $\mathbb{K}_7$ i tyle samo przykładów pokrywa, ale trzeba wybrać i można zauważyć, że w zbiorze T pokrywa tylko przykłady o jednej etykiecie; $S = \{\mathbb{K}_3\}, k_* = \mathbb{K}_3$,

(g) $R = \{<?, \text{sportowy} > \rightarrow \text{duże}, < w2, ? > \rightarrow \text{małe}, < w3, ? > \rightarrow \text{duże}\}$, $P = \{1, 9\}$,

(h) $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),

- $S = \{<? >\} \neq \phi, k_* =<? >$ i $\vartheta_{k_*}(P) = -1$,
- $\mathbb{K}_8 =<?, \text{minivan} >$ ma największą wartość $\vartheta = 0$ razem z $\mathbb{K}_7$ i tyle samo przykładów pokrywa, ale trzeba wybrać go wybrać, aby ostatni przykład miał etykietę duże; $S = \{\mathbb{K}_8\}, k_* = \mathbb{K}_8$,

(i) $R = \{<?, \text{sportowy} > \rightarrow \text{duże}, < w2, ? > \rightarrow \text{małe}, < w3, ? > \rightarrow \text{duże}, <?, \text{minivan} > \rightarrow \text{małe}\}$, $P = \{1\}$,

(j) $P \neq \phi \Rightarrow$*znajdź-kompleks*(T, P),

- $S = \{<? >\} \neq \phi, k_* =<? >$ i $\vartheta_{k_*}(P) = 0$,
  Kompleks k_* tym razem ma największą wartość funkcji oceny i zostaje częścią reguły.

(k) Ostatecznie
$R = \{<?, \text{sportowy} > \rightarrow \text{duże},$
$< w2, ? > \rightarrow \text{małe},$
$< w3, ? > \rightarrow \text{duże},$
$<?, \text{minivan} > \rightarrow \text{małe},$
$<? > \rightarrow \text{duże}\}$

W uzyskanym zbiorze reguł NIE można reguł zamieniać miejscami, gdyż jest to zbiór uporządkowany. Najpierw nowe przykłady klasyfikuje reguła pierwsza, jak ona zawiedzie to druga itd.

4. Za pomocą algorytmu sekwencyjnego pokrywania AQ uzyskać nieuporządkowany zbiór zdaniowych reguł ze zbioru treningowego podanego w tabeli poniżej. Opisać dokładnie kolejne kroki algorytmu. Atrybut wiek zdyskretyzować korzystając z dwóch progów 30 i 65 lat. Atrybut ryzyko będzie kategorią. Ziarna pozytywne należy wybierać po kolei ze zbioru P przykładów nie pokrytych przez znalezione reguły. Ziarna negatywne po kolei ze zbioru T z pozycji pod ziarnem pozytywnym, a jak się skończy tabela to wybierać proszę ziarna negatywne jak najbardziej podobne do ziaren pozytywnych (jak najwięcej takich samych wartości atrybutów).

x	wiek	samochód	ryzyko
1	18	maluch	duże
2	35	maluch	małe
3	50	sportowy	duże
4	66	minivan	duże
5	18	sportowy	duże
6	35	minivan	małe
7	60	maluch	małe
8	70	sportowy	duże
9	25	minivan	małe

Rozwiązanie:
Atrybut wiek otrzymuje po dyskretyzacji trzy wartości:

- w_1: wiek < 30,
- w_2: wiek $\geqslant 30 \land$ wiek < 65,
- w_3: wiek $\geqslant 65$.

Kolejne kroki algorytmu AQ

(a) Początkowo $R = 0, P = T = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

(b) Następuje wywołanie *znajdź-kompleks*(T, P).
  - $x_s = 1, c(x_s) =$ duże, $x_n = 2, c(x_n) =$ małe, $S = \{<?>\}$
  - powstaje częściowa gwiazda S': $S = S \cap S' = \{< w_1 \lor w_3, ? >\}$;
  - gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii małe, wybór następnego ziarna negatywnego $x_n = 6$
  - $S' = \{< w_1 \lor w_3, ? >, <?, \text{maluch} \lor \text{sportowy} >\}$
  - $S = S \cap S' = \{< w_1 \lor w_3, ? >, < w_1 \lor w_3, \text{maluch} \lor \text{sportowy} >\}$
  - $S = \{k_1, k_2\}$, $v_{k_1} = |T^{\text{duże}}_{k_1}| + (|T^{\text{małe}}| - |T^{\text{małe}}_{k_1}|) = 4 + (4 - 1) = 7$, $v_{k_2} = 3 + 4 = 7$
    Wartości funkcji oceny dla dwóch uzyskanych kompleksów ze zbioru S są takie same, ale k_2 pokrywa wyłącznie przykłady o jednej etykiecie duże, stąd on wchodzi w skład nowej reguły:

(c) $R = \{< w_1 \lor w_3, \text{maluch} \lor \text{sportowy} > \rightarrow \text{duże}\}$

(d) $P = \{2, 3, 4, 6, 7, 9\}$, dla $P \neq 0$ *znajdź-kompleks*(T, P)
  - $x_s = 2, c(x_s) =$ małe, $x_n = 3, c(x_n) =$ duże, $S = \{<?>\}$
  - powstaje częściowa gwiazda S': $S = S \cap S' = \{<?, \text{maluch} \lor \text{minivan} >\}$;
  - gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii duże, wybór następnego ziarna negatywnego $x_n = 4$
  - $S' = \{< w_1 \lor w_2, ? >, <?, \text{maluch} \lor \text{sportowy} >\}$
  - $S = S \cap S' = \{< w_1 \lor w_2, \text{maluch} \lor \text{minivan} >, <?, \text{maluch} >\}$
  - $S = \{k_1, k_2\}$, $v_{k_1} = |T^{\text{małe}}_{k_1}| + (|T^{\text{duże}}| - |T^{\text{duże}}_{k_1}|) = 4 + 5 = 9$, $v_{k_2} = 2 + (5 - 1) = 6$
    Kompleks k_1 ma lepszą wartość funkcji oceny, stąd pozostaje w składzie gwiazdy (jej parametr $m = 1$).
    $S = \{< w_1 \lor w_2, \text{maluch} \lor \text{minivan} >\}$.
  - gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii duże (ze zbioru T), wybór następnego ziarna negatywnego $x_n = 5$
  - $S' = \{< w_2 \lor w_3, ? >, <?, \text{maluch} \lor \text{minivan} >\}$
  - $S = S \cap S' = \{< w_2, \text{maluch} \lor \text{minivan} >, < w_1 \lor w_2, \text{maluch} \lor \text{minivan} >\}$

- $S = \{k_1, k_2\}$, $v_{k_1} = |T^{\mathsf{ma\l{}e}}_{k_1}| + (|T^{\mathsf{du\dot{z}e}}| - |T^{\mathsf{du\dot{z}e}}_{k_1}|) = 3 + 5 = 8$, $v_{k_2} = 4 + (5 - 2) = 7$ Kompleks k_1 nie dosyć, że ma lepszą wartość funkcji oceny, to jeszcze pokrywa wyłącznie przykłady o jednej etykiecie małe (ze zbioru T), stąd on wchodzi w skład nowej reguły:

(e) $R = \{< w_1 \vee w_3, \mathsf{maluch} \vee \mathsf{sportowy} > \rightarrow \mathsf{du\dot{z}e}, < w_2, \mathsf{maluch} \vee \mathsf{minivan} > \rightarrow \mathsf{ma\l{}e}\}$

(f) $P = \{3, 4, 9\}$, dla $P \neq 0$ *znajdź-kompleks*(T, P)

- $x_s = 3, c(x_s) = \mathsf{du\dot{z}e}$, $S = \{<? >\}$, $x_n = 6$
- $S = S \cap S' = \{<?, \mathsf{maluch} \vee \mathsf{sportowy} >\}$
- gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii małe ze zbioru T, wybór następnego ziarna negatywnego $x_n = 7$
- $S' = \{<?, \mathsf{sportowy} \vee \mathsf{minivan} >\}$
- $S = S \cap S' = \{<?, \mathsf{sportowy} >\}$ Kompleks z S pokrywa wyłącznie przykłady o jednej etykiecie duże (ze zbioru T), stąd on wchodzi w skład nowej reguły:

(g) $R = \{< w_1 \vee w_3, \mathsf{maluch} \vee \mathsf{sportowy} > \rightarrow \mathsf{du\dot{z}e}, < w_2, \mathsf{maluch} \vee \mathsf{minivan} > \rightarrow \mathsf{ma\l{}e}, <?, \mathsf{sportowy} > \rightarrow \mathsf{du\dot{z}e}\}$

(h) $P = \{4, 9\}$, dla $P \neq 0$ *znajdź-kompleks*(T, P)

- $x_s = 4, c(x_s) = \mathsf{du\dot{z}e}$, $S = \{<? >\}$, $x_n = 9$
- $S = S \cap S' = \{< w2 \vee w3, ? >\}$
- gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii małe ze zbioru T, wybór następnego ziarna negatywnego $x_n = 6$
- $S' = \{< w1 \vee w3, ? >\}$
- $S = S \cap S' = \{< w3, ? >\}$
  Kompleks z S pokrywa wyłącznie przykłady o jednej etykiecie duże (ze zbioru T), stąd on wchodzi w skład nowej reguły:

(i) $R = \{< w_1 \vee w_3, \mathsf{maluch} \vee \mathsf{sportowy} > \rightarrow \mathsf{du\dot{z}e}, < w_2, \mathsf{maluch} \vee \mathsf{minivan} > \rightarrow \mathsf{ma\l{}e}, <?, \mathsf{sportowy} > \rightarrow \mathsf{du\dot{z}e}, < w3, ? > \rightarrow \mathsf{du\dot{z}e}\}$

(j) $P = \{9\}$, dla $P \neq 0$ *znajdź-kompleks*(T, P)

- $x_s = 9, c(x_s) = \mathsf{du\dot{z}e}$, $S = \{<? >\}$, $x_n = 4$
- $S = S \cap S' = \{< w1 \vee w2, ? >\}$
- gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii duże ze zbioru T, wybór następnego ziarna negatywnego $x_n = 1$
- $S' = \{<?, \mathsf{minivan} \vee \mathsf{sportowy} >\}$
- $S = S \cap S' = \{< w1 \vee w2, \mathsf{minivan} \vee \mathsf{sportowy} >\}$
- gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii duże ze zbioru T, wybór następnego ziarna negatywnego $x_n = 5$
- $S' = \{<?, \mathsf{minivan} \vee \mathsf{maluch} >\}$
- $S = S \cap S' = \{< w1 \vee w2, \mathsf{minivan} >\}$
  Kompleks z S pokrywa wyłącznie przykłady o jednej etykiecie małe (ze zbioru T), stąd on wchodzi w skład nowej reguły:

(k) Ostatecznie

$$\begin{aligned}
R = \{& < w_1 \vee w_3, \mathsf{maluch} \vee \mathsf{sportowy} > \rightarrow \mathsf{du\dot{z}e},\\
& < w_2, \mathsf{maluch} \vee \mathsf{minivan} > \rightarrow \mathsf{ma\l{}e},\\
& <?, \mathsf{sportowy} > \rightarrow \mathsf{du\dot{z}e},\\
& < w3, ? > \rightarrow \mathsf{du\dot{z}e},\\
& < w1 \vee w2, \mathsf{minivan} > \rightarrow \mathsf{ma\l{}e}\}
\end{aligned}$$

W uzyskanym zbiorze reguł można reguły zamieniać miejscami, gdyż jest to zbiór nieuporządkowany.

5. Za pomocą algorytmu sekwencyjnego pokrywania AQ uzyskać uporządkowany zbiór zdaniowych reguł ze zbioru treningowego podanego w tabeli poniżej. Opisać dokładnie kolejne kroki algorytmu. Atrybut wiek zdyskretyzować korzystając z dwóch progów 30 i 65 lat. Atrybut ryzyko będzie kategorią. Ziarna pozytywne należy wybierać po kolei ze zbioru P przykładów nie pokrytych przez znalezione reguły. Ziarna negatywne po kolei ze zbioru P z pozycji pod ziarnem pozytywnym, a jak się skończy zbiór P to wybierać proszę ziarna negatywne ze zbioru T jak najbardziej podobne do ziaren pozytywnych (jak najwięcej takich samych wartości atrybutów).

x	wiek	samochód	ryzyko
1	18	maluch	duże
2	35	maluch	małe
3	50	sportowy	duże
4	66	minivan	duże
5	18	sportowy	duże
6	35	minivan	małe
7	60	maluch	małe
8	70	sportowy	duże
9	25	minivan	małe

Rozwiązanie:
Atrybut wiek otrzymuje po dyskretyzacji trzy wartości:

- w_1: wiek < 30,
- w_2: wiek $\geqslant 30 \wedge$ wiek < 65,
- w_3: wiek $\geqslant 65$.

Kolejne kroki algorytmu AQ

(a) Początkowo $R = 0, P = T = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

(b) Następuje wywołanie *znajdź-kompleks*(T, P).
- $x_s = 1, c(x_s) = \text{duże}$, $x_n = 2, c(x_n) = \text{małe}$, $S = \{<? >\}$
- powstaje częściowa gwiazda S': $S = S \cap S' = \{< w_1 \vee w_3, ? >\}$;
- gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii małe, wybór następnego ziarna negatywnego $x_n = 6$
- $S' = \{< w_1 \vee w_3, ? >, <?, \text{maluch} \vee \text{sportowy} >\}$
- $S = S \cap S' = \{< w_1 \vee w_3, ? >, < w_1 \vee w_3, \text{maluch} \vee \text{sportowy} >\}$
- $S = \{k_1, k_2\}$, $v_{k_1} = |T^{\text{duże}}_{k_1}| + (|T^{\text{małe}}| - |T^{\text{małe}}_{k_1}|) = 4 + (4 - 1) = 7$, $v_{k_2} = 3 + 4 = 7$
Wartości funkcji oceny dla dwóch uzyskanych kompleksów ze zbioru S są takie same, ale k_2 pokrywa wyłącznie przykłady o jednej etykiecie duże, stąd on wchodzi w skład nowej reguły:

(c) $R = \{< w_1 \vee w_3, \text{maluch} \vee \text{sportowy} > \rightarrow \text{duże}\}$

(d) $P = \{2, 3, 4, 6, 7, 9\}$, dla $P \neq 0$ *znajdź-kompleks*(P, P)
- $x_s = 2, c(x_s) = \text{małe}$, $x_n = 3, c(x_n) = \text{duże}$, $S = \{<? >\}$
- powstaje częściowa gwiazda S': $S = S \cap S' = \{<?, \text{maluch} \vee \text{minivan} >\}$;
- gwiazda w dalszym ciągu pokrywa przykłady z T o kategorii duże, wybór następnego ziarna negatywnego $x_n = 4$
- $S' = \{< w_1 \vee w_2, ? >, <?, \text{maluch} \vee \text{sportowy} >\}$
- $S = S \cap S' = \{< w_1 \vee w_2, \text{maluch} \vee \text{minivan} >, <?, \text{maluch} >\}$

- $S = \{k_1, k_2\}$, $v_{k_1} = |P_{k_1}^{\mathsf{małe}}| + (|P^{\mathsf{duże}}| - |P_{k_1}^{\mathsf{duże}}|) = 4 + 2 = 6$, $v_{k_2} = 2 + 2 = 4$
  Kompleks k_1 nie dosyć, że ma lepszą wartość funkcji oceny, to jeszcze pokrywa wyłącznie przykłady o jednej etykiecie małe (ze zbioru P), stąd on wchodzi w skład nowej reguły:

(e) $R = \{< w_1 \vee w_3, \mathsf{maluch} \vee \mathsf{sportowy} > \rightarrow \mathsf{duże}, < w_1 \vee w_2, \mathsf{maluch} \vee \mathsf{minivan} > \rightarrow \mathsf{małe}\}$

(f) $P = \{3, 4\}$, dla $P \neq 0$ *znajdź-kompleks*(P, P)

- $x_s = 3, c(x_s) = \mathsf{duże}$, $S = \{<?>\}$ Gwiazda S pokrywa przykłady o jednej etykiecie duży i kompleks $<?>$ wchodzi w skład nowej reguły:
- $R = \{< w_1 \vee w_3, \mathsf{maluch} \vee \mathsf{sportowy} > \rightarrow \mathsf{duże}, < w_1 \vee w_2, \mathsf{maluch} \vee \mathsf{minivan} > \rightarrow \mathsf{małe}, <?> \rightarrow \mathsf{duże}\}$
- ewentualnie, gdy $x_n = 9$, to
- $S = S \cap S' = \{< w_2 \vee w_3, ? >, <?, \mathsf{maluch} \vee \mathsf{sportowy} >\}$
  Kompleks k_1 pokrywa wszystkie przykłady ze zbioru P i wchodzi w skład nowej reguły:

(g) Ostatecznie

$$\begin{aligned} R = \{ & < w_1 \vee w_3, \mathsf{maluch} \vee \mathsf{sportowy} > \rightarrow \mathsf{duże}, \\ & < w_1 \vee w_2, \mathsf{maluch} \vee \mathsf{minivan} > \rightarrow \mathsf{małe}, \\ & < w_2 \vee w_3, ? > \rightarrow \mathsf{duże}\} \end{aligned}$$

W uzyskanym zbiorze reguł NIE można reguł zamieniać miejscami, gdyż jest to zbiór uporządkowany. Najpierw nowe przykłady klasyfikuje reguła pierwsza, jak ona zawiedzie to druga itd.