Imię i nazwisko: | | | | | | | | | | | | | | | | | | | Nr indeksu: | | | | | | |

Egzamin z Wykładu Monograficznego p.t. "DATA MINING"

1. (6 pkt.) Firma X jest dostawcą usług połączeń bezprzewodowych (wireless) w USA, która ma 34.6 milionów klientów. Firma szacuje, że średnio 2% klientów rezygnuje ze usług tej firmy miesięcznie, i koszt pozyskania nowego klienta wynosi od 320 USD do 360 USD. Firma X decyduje się na strategię utrzymania swoich obecnych klientów i dowiedziała się, że metody data mining mogą być używane do (1) odkrywania różnych grup klientów z różnymi wzorcami zachowania i zapotrzebowania oraz do (2) wspomagania podejmowania decyzji marketingowych. Jako konsultant rozwiązań data mining'owych, proszę pomóc firmie X przy wyborze właściwych technik:

   (a) Do identyfikacji wzorca zachowań klientów, firma X powinna stosować technikę:
   (i) Klasyfikacji (ii) Grupowania (clustering) (iii) Reguły asocjacji
   (iv) Regresji liniowej (v) Żadna z tych metod (vi) Wszystkie

   (b) Do wykrywania docelowej grupy klientów dla wybranego produktu, firma X powinna stosować technikę:
   (i) Klasyfikacji (ii) Grupowania (clustering) (iii) Reguły asocjacji
   (iv) Regresji liniowej (v) Żadna z tych metod (vi) Wszystkie

   (c) Do znalezienia najlepszego planu taryfowego dla każdego klienta, firma X powinna stosować technikę:
   (i) Klasyfikacji (ii) Grupowania (clustering) (iii) Reguły asocjacji
   (iv) Żadna z tych metod (v) Wszystkie

2. (10 pkt.) Przedstawiona tablica decyzyjna zawiera przykłady klasyfikacji ludzi na 3 klasy: *normalna, niedowaga i nadwaga.*

Imię	Waga	Wzrost	Klasa
Kristina	160 lb	1.6 m	Normalna
Jim	210 lb	2.0 m	Normalna
Maggie	207 lb	1.9 m	Normalna
Martha	130 lb	1.8 m	Niedowaga
Stephanie	221 lb	1.7 m	Nadwaga
Bob	215 lb	1.8 m	Normalna
Kathy	178 lb	1.6 m	Normalna
Dave	138 lb	1.7 m	Niedowaga
Worth	160 lb	2.2 m	Niedowaga
Steven	190 lb	2.1 m	Normalna
Debbie	234 lb	1.8 m	Nadwaga
Todd	285 lb	1.9 m	Nadwaga
Kim	135 lb	1.9 m	Niedowaga
Amy	198 lb	1.8 m	Normalna
Lynette	289 lb	1.7 m	Nadwaga

Sklasyfikuj nowe przypadki za pomocą algorytmu 5 najbliższych sąsiadów (można wybrać optymalną dla siebie funkcję odległości i/lub stworzyć nowe atrybuty z istniejących):

(a) John [185 lb, 2.0 m]

(b) Kelly [165 lb, 1.5 m]

(c) Sam [180 lb, 2.4 m]

(d) Laura [195 lb, 1.8 m]

(e) Mike [220 lb, 1.7 m]

3. (6 pkt.)

   (a) Skonstruuj drzewo decyzyjne za pomocą miary *entropii* z następującej tablicy decyzyjnej. Narysuj wynikowe drzewo.

Instance	Attribute a_1	Attribute a_2	Class
1	T	F	-
2	T	F	-
3	F	F	+
4	F	T	-
5	F	T	+
6	T	T	-
7	F	T	+
8	F	T	+
9	F	F	+
10	T	F	-

   (b) Naszkicuj drzewo decyzyjne otrzymane w wyniku zastosowania algorytmu przycinania (pruning), przy dopuszczalnym błędzie w liściach = 75%, na drzewie otrzymanym w poprzednim kroku.

4. (6 pkt.) Dana jest tablica decyzyjna:

Instance	Attribute X	Attribute Y	Attribute Z	Class
1	F	F	F	+
2	F	F	T	-
3	F	T	T	-
4	F	T	T	-
5	F	F	T	+
6	T	F	F	+
7	T	F	T	-
8	T	F	T	-
9	T	T	F	+
10	T	F	T	+

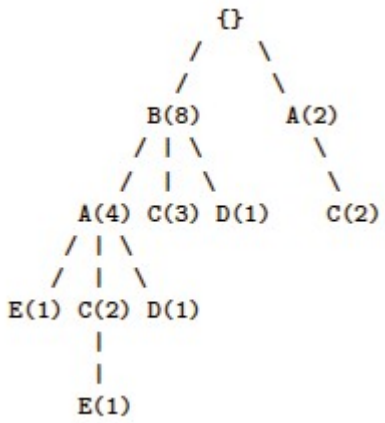
   (a) Oblicz rozkłady prawdopodobieństwa zmiennych losowych, potrzebnych do klasyfikacji nowych przypadków metodą *naiwnego Bayesa*.

   (b) Sklasyfikuj przypadek [F,T,F] metodą *naiwnego Bayesa* na podstawie zadanej tablicy.

5. (6 pkt.) Skonstruuj binarne drzewo decyzyjne za pomocą miary *rozróżnialności*. Naszkicuj drzewo wynikowe:

Age	Income	Class
30	high	no
35	high	yes
40	medium	yes
40	low	yes
40	low	no
35	low	yes
30	medium	no
30	low	yes
30	medium	yes
35	medium	yes
35	high	yes
40	medium	no

6. (10 pkt.) Znaleźć wszystkie częste zbiory na podstawie przedstawionego drzewa FP-tree (przy minimalnym wsparciu = 2):



7. (6 pkt.) Dany jest zbiór liczb {1, 2, 3, 4, 5, 6, 10, 20, 30, 40, 50, 60}. Chcemy znaleźć 4 klastry algorytmem k-centroidów minimalizując *sumę kwadratów błędów*. Zainicjowano 4 pierwsze cetroidy {2}, {5}, {20}, {50}.

   (a) Podaj centroidy otrzymane w kolejnych krokach algorytmu.

   (b) Podaj *sumę kwadratów błędów* wynikowych klastrów.

Odpowiedzi:

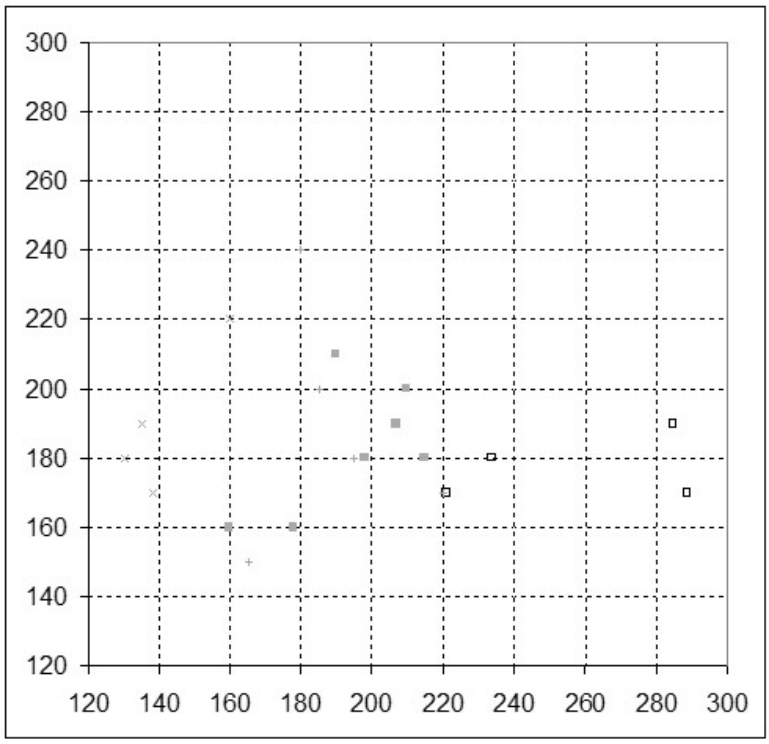
1. Max. 6 punktów

   (a) (ii) 2pkt; (iii) 1pkt.
   II - using information about customers' transactions a clustering algorithm can reveal customer groups with common usage patterns.

   (b) (i) 2pkt;
   I - The task above requires assigning a customer into one of two groups: likely or not likely to terminate contract. This is a classification task, hence in order to identify which customers should be targeted a classification algorithm can be applied to predict whether or not a customer is likely to terminate his/her contract.

(c) (i) 2pkt;
I - The task of assigning a customer into one of a set of plans is also a classification task.

2. (10 pkt.) po 2 punkty za każdą poprawną odpowiedź;



(a) John [185 lb, 2.0 m]: **Normalna**
Ratio = 185 / 2.0 = 92.5 Nearest neighbors are: Steven [90.5, Average], Kristina [100, Average], Dave [81.2, Underweight], Jim [105, Average] and Maggie [108.9, Average]. By simple majority voting, John should be labeled Average.

(b) Kelly [165 lb, 1.5 m]: **Normalna**
Ratio = 165 / 1.5 = 110.0 Nearest neighbors are: Amy [110, Average], Maggie [108.9, Average], Jim [105, Average] and Bob [119.4, Average]. By simple majority voting, Kelly should be labeled Average.

(c) Sam [180 lb, 2.4 m] **Niedowaga**
Ratio = 180 / 2.4 = 75.0 Nearest neighbors are: Worth [72.7, Underweight], Martha [72.2, Underweight], Kim [71.1, Underweight], Dave [81.2, Underweight] and Steven [90.5, Average]. By simple majority voting, Sam should be labeled Underweight.

(d) Laura [195 lb, 1.8 m]: **Normalna**
Ratio = 195 / 1.8 = 108.3 Nearest neighbors are: Maggie [108.9, Average], Amy [110, Average], Kathy [111.3, Average], Jim [105, Average] and Kristina [100.0, Average]. By simple majority voting, Laura should be labeled Average.

(e) Mike [220 lb, 1.7 m]: **Nadwaga**
Ratio = 220 / 1.7 = 129.4 Nearest neighbors are: Debbie [130, Overweight], Stephanie [130, Overweight], Bob [119.4, Average], Kathy [111.3, Average] and Amy [110, Average]. By simple majority voting, Mike should be labeled Average. However, if we follow weighted voting, Mike should be labeled Overweight.

3. (6 pkt.) a1 a2

   N1 (T) N2 (F) N1 (T) N2 (F)

   + 0 5 + 3 2

   - 4 1 - 2 3

   Entropy(N1) = 0 Entropy(N1) = 0.97

   Entropy(N2) = 0.65 Entropy(N1) = 0.97

4. (6 pkt.)

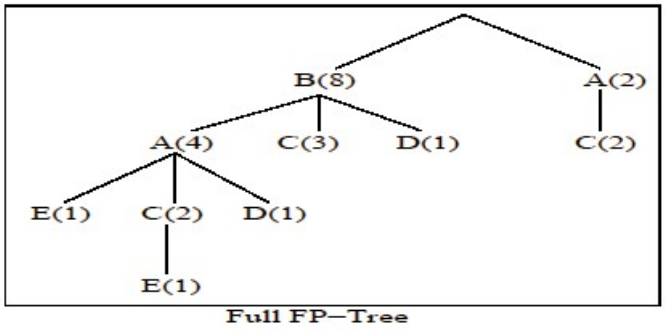
```
a)
P(X' | +) = 2 + (5)(1/5) / 5 + 5 = 3 / 10
P(X' | -) = 3 + (5)(1/5) / 5 + 5 = 4 / 10
P(Y | +) = 1 + (5)(1/5) / 5 + 5 = 2 / 10
P(Y | -) = 2 + (5)(1/5) / 5 + 5 = 3 / 10
P(Z' | +) = 3 + (5)(1/5) / 5 + 5 = 4 / 10
P(Z' | -) = 0 + (5)(1/5) / 5 + 5 = 1 / 10
b)
P(+ | X',Y,Z')
= P(X' | +) * P (Y | +) * P(Z' | +) * P(+) / P(X') * P(Y) * P(Z')
P(- | X',Y,Z')
= P(X' | -) * P (Y | -) * P(Z' | -) * P(-) / P(X') * P(Y) * P(Z')
Since the denominators are same, we can compare these values by just comparing the
numerators.
P(+ | X',Y,Z')
~= (3/10) * (2/10) * (4/10) * (5/10)
= 120 / 10000
P(- | X', Y, Z')
~= (4/10) * (3/10) * (1/10) * (5/10)
= 60 / 10000
Since, P(+ | X', Y, Z') > P(- | X', Y, Z'),
hence the instance should be classified as +.
```

5. (6 pkt.) Skonstruuj binarne drzewo decyzyjne za pomocą miary *rozróżnialności*. Naszkicuj drzewo wynikowe:

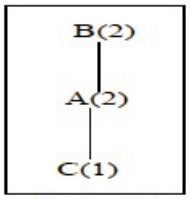
Age	Income	Class
30	high	no
35	high	yes
40	medium	yes
40	low	yes
40	low	no
35	low	yes
30	medium	no
30	low	yes
30	medium	yes
35	medium	yes
35	high	yes
40	medium	no

6. (10 pkt.) B(8) C(7) A(6) D(2) E(2) AE(2) BE(2) BAE(2) BD(2) CA(4) BCA(2) BC(5) BA(4)

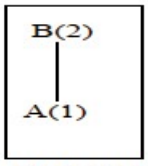
a) Consider the stage shown in Figure 1. We first project the full tree on the item E. We output E(2), and since only one path remains in the FPTree for E, we output all frequent combinations: AE(2), BE(2), BAE(2).
Next project on D(2). Only remaining frequent items is B, so we get BD(2).
Next Project on C(7). In the new tree, first project on A(4), to get CA(4), and BCA(2).
Next project on B(5) and output BC(5).
Project on A(6), to get BA(4). Finally output B(8).



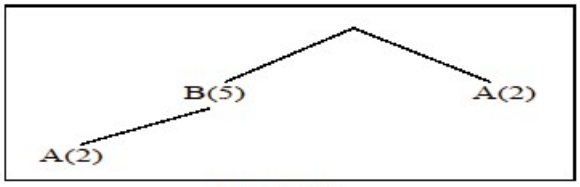
Full FP–Tree



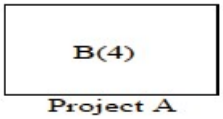
Project E



Project D



Project C



Project A

7. (6 pkt.) Initial centroids: 2, 5, 20, 50

1, 2, 3, 4, 5, 6, 10, 20, 30 and 40, 50, 60.

2 6.25 25 50

(1,2, 3, 4) (5,6,10) (20, 30) (40, 50, 60)

2.5 7 25 50