v2017

Oppgave 1 - Various

a. Explain pageview in context of web usage mining.

   Pageview is a collection of resources representing a specific “user event”, example link click, viewing a product page, adding stuff to cart

b. One type of pre-processing in web usage mining is path completion. Why is this necessary,
   and how can it be done?

   This is by client side cachingwhen results in missing access refrences to those pages have been cached. ??

Oppgave 2 - Modeling

▼ Task

Miljøbomringen AS will soon be responsible for the tollbooths in all major cities in Norway, and want a data warehouse that can be used to analyze traffic, i.e. the toll passages. As part of this reorganization, all cars must have AutoPass (transponder) for automatic registration of passages. A customer may have several cars and must have one transponder for each car. The price for each passage changes dynamically/continuously for each station independently of others, based on time of day, pollution, traffic jams, etc.

An example of analyzes you should be able to do against the data warehouse:

□ Number of passages for each quarter for each station.

□ Number of passages for each quarter for each car.

□ Average number of passages per month.

□ Average price for cars for one particular station.

The description is somewhat imprecisely formulated and it is part of the task to select what should be included. We are primarily looking for you to show modeling principles for data warehousing. Explain any assumptions you find necessary to do.

Create a star schema for the described case.

Fact | Customer, Car, AutoPass, Time, Station | price

Customer < All < City < Address < Name

Car < All < Model < Type < Reg_nr

Time < All < Year < Quarter < Month < Day

Station < All < City < Address < Name

Oppgave 3 - OLAP

a. Explain roll-up and drill-down.

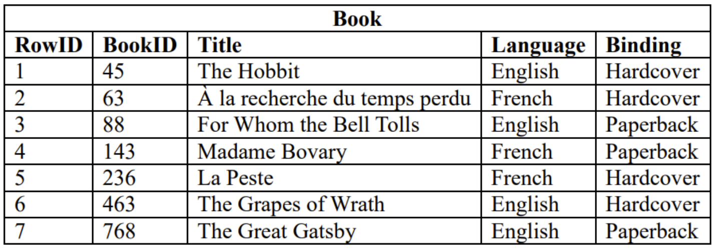
   Roll-up: Moving up in the concept-hierarkie, and performing dimention reduction. Move from detailed data to data with less details.

   Drill-down: Moving down in the concept-hierarkie, and performing a increase in dimention. Go from overview data to more spesiffic data

b. .

   ▼

   Given a dimension table Book in a data warehouse, we want to use bitmap indexes on the attributes Language and Binding in order to be able to perform queries more efficiently. Show structure and contents of the bitmap indexes based on the contents in the table below



Book

RowID	BookID	Title	Language	Binding
1	45	The Hobbit	English	Hardcover
2	63	À la recherche du temps perdu	French	Hardcover
3	88	For Whom the Bell Tolls	English	Paperback
4	143	Madame Bovary	French	Paperback
5	236	La Peste	French	Hardcover
6	463	The Grapes of Wrath	English	Hardcover
7	768	The Great Gatsby	English	Paperback

Language:
English: 1010011
French: 0101100

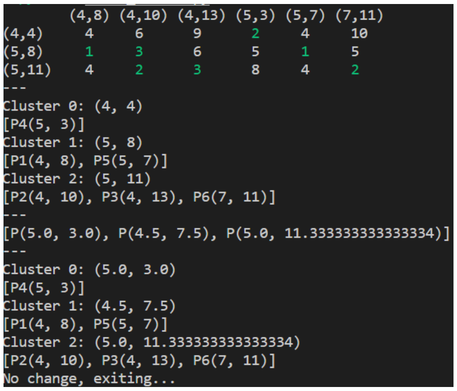
Binding:
Hardcover: 1100110
Paperback: 0011001

Oppgave 4 - Clustering

▼

Assume a two-dimensional dataset as shown in the table to the right. Perform clustering using K-means, with k=3 and intial centroids S1=(4,4), S2=(5,8) og S3=(5,11). Use Manhattan distance.

	X	Y
P1	4	8
P2	4	10
P3	4	13
P4	5	3
P5	5	7
P6	7	11



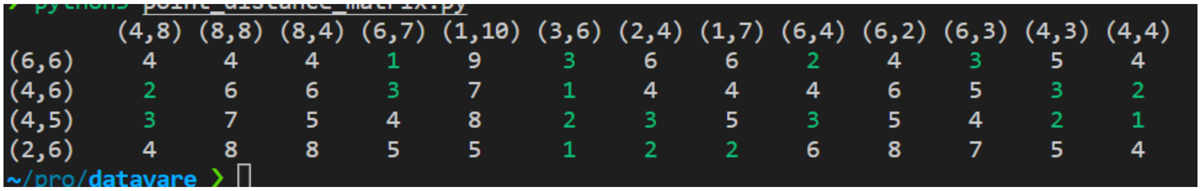
```
        (4,8) (4,10) (4,13) (5,3) (5,7) (7,11)
(4,4)     4     6      9      2     4     10
(5,8)     1     3      6      5     1     5
(5,11)    4     2      3      8     4     2
---
Cluster 0: (4, 4)
[P4(5, 3)]
Cluster 1: (5, 8)
[P1(4, 8), P5(5, 7)]
Cluster 2: (5, 11)
[P2(4, 10), P3(4, 13), P6(7, 11)]
---
[P(5.0, 3.0), P(4.5, 7.5), P(5.0, 11.333333333333334)]
---
Cluster 0: (5.0, 3.0)
[P4(5, 3)]
Cluster 1: (4.5, 7.5)
[P1(4, 8), P5(5, 7)]
Cluster 2: (5.0, 11.333333333333334)
[P2(4, 10), P3(4, 13), P6(7, 11)]
No change, exiting...
```

Oppgave 5 - Classification

a. .

▼

You are given a dataset of samples P1 = (4,8), P2 = (8,8), P3 = (8,4), P4 = (6,7), P5 = (1,10),
P6 = (3,6), P7 = (2,4), P8 = (1,7), P9 = (6,4), P10 = (6,2), P11 = (6,3), P12 = (4,3), and
P13=(4,4). The samples belong to three clusters C1 = {P1,P2,P3,P4}, C2 = {P5,P6,P7,P8}
and C3 ={P9,P10,P11,P12,P13}. Consider associated clusters as class labels. Classify the
samples A = (6,6), B = (4,6), C = (4,5), and D=(2,6) by employing the k-nearest neighbor (kNN) method. Use the Manhattan distance metric and k = 3. Describe how the results of the classification are achieved.

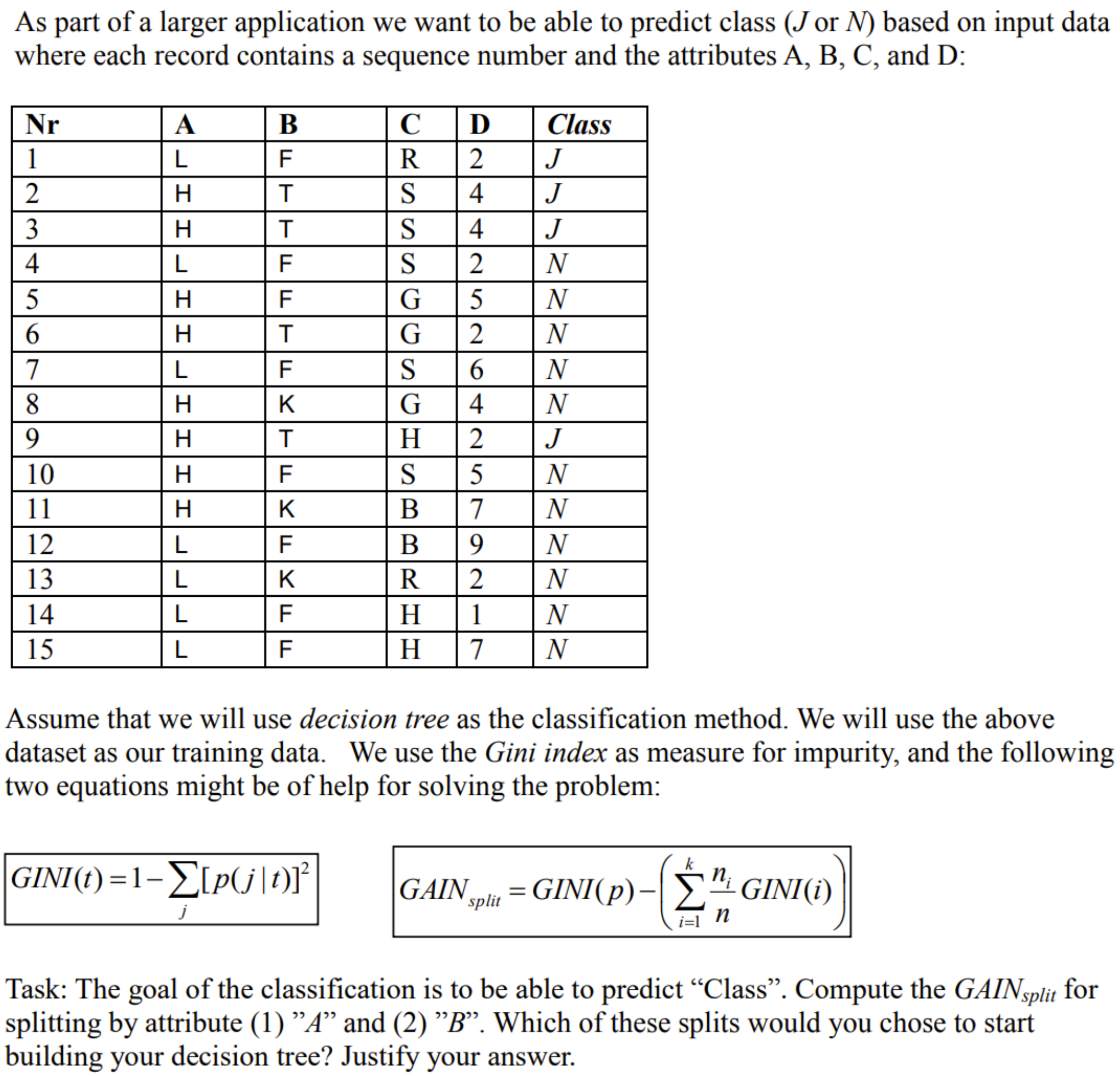


```
❯ python3 point_distance_matrix.py
        (4,8) (8,8) (8,4) (6,7) (1,10) (3,6) (2,4) (1,7) (6,4) (6,2) (6,3) (4,3) (4,4)
(6,6)     4     4     4     1     9      3     6     6     2     4     3     5     4
(4,6)     2     6     6     3     7      1     4     4     4     6     5     3     2
(4,5)     3     7     5     4     8      2     3     5     3     5     4     2     1
(2,6)     4     8     8     5     5      1     2     2     6     8     7     5     4
~/pro/datavare ❯ ▯
```

A is in C3, B is any cluster, C in C3 and D in C2

b. .

▼



As part of a larger application we want to be able to predict class (*J* or *N*) based on input data where each record contains a sequence number and the attributes A, B, C, and D:

Nr	**A**	**B**	**C**	**D**	***Class***
1	L	F	R	2	*J*
2	H	T	S	4	*J*
3	H	T	S	4	*J*
4	L	F	S	2	*N*
5	H	F	G	5	*N*
6	H	T	G	2	*N*
7	L	F	S	6	*N*
8	H	K	G	4	*N*
9	H	T	H	2	*J*
10	H	F	S	5	*N*
11	H	K	B	7	*N*
12	L	F	B	9	*N*
13	L	K	R	2	*N*
14	L	F	H	1	*N*
15	L	F	H	7	*N*

Assume that we will use *decision tree* as the classification method. We will use the above dataset as our training data. We use the *Gini index* as measure for impurity, and the following two equations might be of help for solving the problem:

$$\boxed{GINI(t) = 1 - \sum_j [p(j \mid t)]^2}$$

$$\boxed{GAIN_{split} = GINI(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \right)}$$

Task: The goal of the classification is to be able to predict "Class". Compute the $GAIN_{split}$ for splitting by attribute (1) "*A*" and (2) "*B*". Which of these splits would you chose to start building your decision tree? Justify your answer.

Total:

J: 4, N:11$GINI = 1 - \frac{4}{15}^2 - \frac{11}{15}^2 = 0.39$

A:

	J	N
L	1	6
H	3	5

$$GINI_L = 1 - \frac{1}{7}^2 - \frac{6}{7}^2 = 0.25$$

$$GINI_H = 1 - \frac{3}{8}^2 - \frac{5}{8}^2 = 0.47$$

$$GINI_A = \frac{7}{15} * 0.25 + \frac{8}{15} * 0.47 = 0.37$$

$$GAIN = 0.39 - 0.37 = 0.02$$

B

	J	N
T	3	1
F	1	7
K	0	3

$$GINI_T = 1 - \frac{3}{4}^2 - \frac{1}{4}^2 = 0.38$$

$$GINI_F = 1 - \frac{1}{8}^2 - \frac{7}{8}^2 = 0.22$$

$$GINI_K = 1 - \frac{0}{3}^2 - \frac{3}{3}^2 = 0.0$$

$$GINI_B = \frac{4}{15} * 0.38 + \frac{8}{15} * 0.22 + \frac{3}{15} * 0 = 0.21$$

$$GAIN = 0.39 - 0.21 = 0.18$$

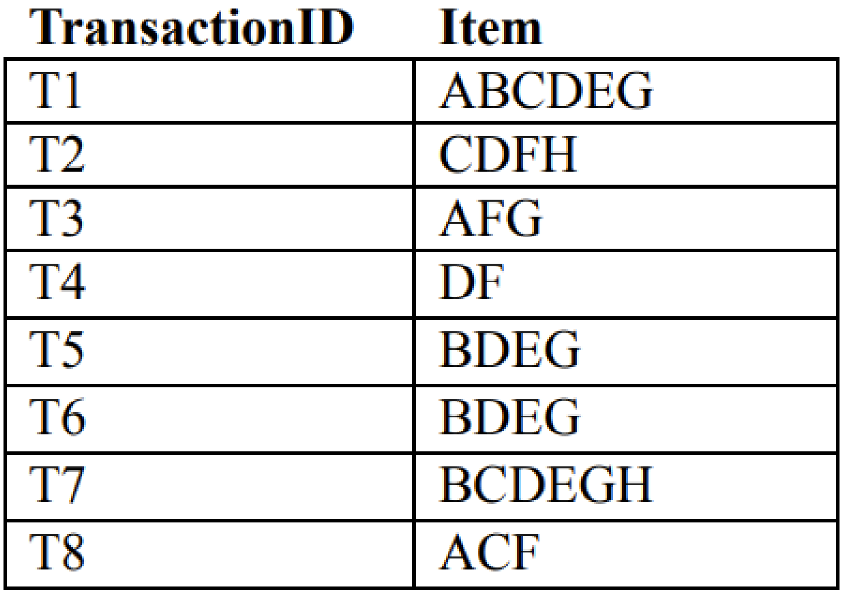
We select to splitt on B because it has a higher GAIN.

Oppgave 6

a. .

▼ task

Assume the market basket data below. Use the apriori-algorithm to find all frequent itemsets with minimum support of 50 % (i.e., minimum support count is 4). Use the Fk-1×Fk-1 method for candidate generation.



TransactionID	Item
T1	ABCDEG
T2	CDFH
T3	AFG
T4	DF
T5	BDEG
T6	BDEG
T7	BCDEGH
T8	ACF

A	3
B	4
C	4
D	6
E	4
F	4
G	5
H	2
BC	2
BD	4
BE	4
BF	0
BG	4
CD	3
CE	2
CF	2
CG	2
DE	4
DF	2
DG	4
EF	0
EG	4
FG	1
BDE	4
BDG	4
BEG	4
DEG	4
BDEG	4

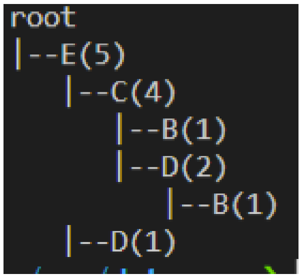
b. .

▼

Assume the market basket data below. You are now going to use the FP-growth-algorithm in order to find all frequent itemsets with minimum support of 40 % (i.e., minimum support count is 2).

1. Construct a FP tree based on the dataset.
2. Find frequent itemsets using the FP-growth-algorithm. Use table notation with the following columns in order to show the result:
    - Item
    - ”Conditional pattern base”
    - ”Conditional FP-tree”
    - Frequent itemsets

[(('E',), 5), (('C',), 4), (('D',), 3), (('B',), 2)]



```
root
|--E(5)
   |--C(4)
      |--B(1)
      |--D(2)
         |--B(1)
   |--D(1)
```

	Conditional patter base	Conditial FP-tree	
B	{(EC):1, (ECD):1}	{(EC):2}|B	B, BE, BC, BCE
D	{(EC):2, (E):1}	{(E):3}|D	D, DE
C	{(E):4}	{(E):4}|C	E, EC
E	Ø	Ø	E