h2021

Oppgave 1

concept	concept	concept	concept	fact
book	shop	locations	time	book
All	All	All	All	location
type	type	continent	year	shop
author	location	country	month	time
name		state	day	—-
		city		unit_sold
				dollars_sold

b)^

c)snowflake splitt everything into smaller pieces

Oppgave 2

Office
Oslo: 1001100
Bergen: 0100010
Trondheim: 0010001

Gender:
Male: 1100101
Female: 0011010

Title:
Developer: 1010011
Tester: 0101000
Project manager: 0000100

Dependent:
Yes: 1100110

No: 0011001

a)
1010011
1100110
1000010. first and second last

b)
$1111111 -$
$0010001 =$
1101110

c)1101110
1100101
1100100

d)

$1001100\ or$
0100010
1101110

1101110
0101000
0101000

Oppgave 3

Instance	Type	A	B	C	D	Class
1	High	3.5	4.0	7.0	-2.00	H
2	High	2.0	-4.0	4.0	2.00	H
3	Low	9.1	4.5	18.2	-2.25	L
4	High	2.0	-6.0	4.0	3.00	H
5	High	1.5	7.0	3.0	-3.50	H
6	High	7.0	-6.5	14.0	3.25	H
7	Low	2.1	2.5	4.2	-1.25	L
8	Low	8.0	-4.0	16.0	2.00	L

Table 1: Data for pre-processing.

Would remove Type as is same as Class.

Looks like C = 2*A, would remove A or C

D = -1/2*B, would remove B or D

Oppgave 4

Since feature is 10^6 we would need to run it $2^{10^6} times$.

Oppgave 5

Name: Nominal

Age: Ratio

Joining Date: Interval

Title: Ordinal

Oppgave 6

A	4
B	5
C	5
D	2
E	4
A, B	3
A, C	4
A, E	2
B, C	4
B, E	4
C, E	3
A, B, C	3
A, B, E	2
A, C, E	2
B, C, E	3
A, B, C, E	2

Oppgave 7

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

D is eliminted

Sorted list:

[[('C',), ('A',)],
[('B',), ('C',), ('E',)],
[('B',), ('C',), ('A',), ('E',)],
[('B',), ('E',)],
[('B',), ('C',), ('A',), ('E',)],
[('B',), ('C',), ('A',)]]

First:

root

|--C(1)

    |--A(1)

second:

root

|--C(1)

    |--A(1)

|--B(1)

    |--C(1)

        |--E(1)

Third:

root

|--C(1)

    |--A(1)

|--B(2)

    |--C(2)

        |--E(1)

        |--A(1)

            |--E(1)

Forth:

root

|--C(1)

    |--A(1)

|--B(3)

    |--C(2)

        |--E(1)

fifth:

root

|--C(1)

    |--A(1)

|--B(4)

    |--C(3)

        |--E(1)

Sixth:

root

|--C(1)

    |--A(1)

|--B(5)

    |--C(4)

        |--E(1)

|--A(1)

|--E(1)

|--E(1)

|--A(2)

|--E(2)

|--E(1)

|--A(3)

|--E(2)

|--E(1)

Projected FP-tree

Project E:

[('B',), ('C',), ('E',)],
[('B',), ('C',), ('A',), ('E',)],
[('B',), ('E',)],
[('B',), ('C',), ('A',), ('E',)]

BCE:1
BE:1
BCAE:2

root

|--B(4)

|--C(3)

|--E(1)

|--A(2)

|--E(2)

|--E(1)

Path B, C, A, E has highest count

Projected FP-tree root → B(4) → C(3) → A(2), A(2) is under support limit 3 and is removed

Frequente Path= {EB(4), EC(3), EBC(3)}

Project A:

[('C',), ('A',)],
[('B',), ('C',), ('A',)],
[('B',), ('C',), ('A',),],
[('B',), ('C',), ('A',)]

CA:1
BCA:3

root

|--C(1)

|--A(1)

|--B(3)

|--C(3)

|--A(3)

Test for projecttion AB:

B:3

Test for projection AC:
BC:3

B:1

root → B(3)

Frequnte Path{AC(4), ACB(3)}

Frequente Path {AB(3)}

Project C:

[('C',)],
[('B',), ('C',)],
[('B',), ('C',)]
[('B',), ('C',)],
[('B',), ('C',)]

root

|--C(1)

|--B(4)

|--C(4)

C:1
BC:4

Projected FP-tree: root → B(4)

Frequnte path{CB(4)}

Project B:

[('B',)],
[('B',)],
[('B',)],
[('B',)],
[('B',)],

root

|--B(5)

B:5
Frequent path = {Ø}

Oppgave 8

```
---
[P(0, 0), P(3.25, 1.0)]
---
Cluster 0: (0, 0)
[P1(0, 2), P2(0, 0), P3(1.5, 0)]
Cluster 1: (3.25, 1.0)
[P4(5, 0), P5(5, 2)]
---
[P(0, 0), P(5.0, 1.0)]
---
Cluster 0: (0, 0)
[P1(0, 2), P2(0, 0), P3(1.5, 0)]
Cluster 1: (5.0, 1.0)
[P4(5, 0), P5(5, 2)]
No change, exiting...
```

Oppgave 9

	(5,8)	(6,7)	(6,5)	(2,4)	(3,4)	(5,4)	(7,4)	(9,4)	(3,3)	(8,2)	(7,5)
(5,8)	0	1	1	3	2	0	2	4	2	3	2
(6,7)	1	0	0	3	3	1	1	3	3	2	1
(6,5)	1	0	0	1	1	1	1	1	2	2	0
(2,4)	3	3	1	0	0	0	0	0	1	2	1
(3,4)	2	3	1	0	0	0	0	0	0	2	1
(5,4)	0	1	1	0	0	0	0	0	1	2	1
(7,4)	2	1	1	0	0	0	0	0	1	1	0
(9,4)	4	3	1	0	0	0	0	0	1	1	1
(3,3)	2	3	2	1	0	1	1	1	0	1	2
(8,2)	3	2	2	2	2	2	1	1	1	0	1
(7,5)	2	1	0	1	1	1	0	1	2	1	0
DENSITY:	4	6	9	8	8	10	10	9	7	5	9

Found points where density is above minptr and these are cors. Se that P1 and P10 both are neighbors to cores and is therefor BORDER

```
[P1(5, 8, PointType.BORDER),
 P2(6, 7, PointType.CORE),
 P3(6, 5, PointType.CORE),
 P4(2, 4, PointType.CORE),
 P5(3, 4, PointType.CORE),
 P6(5, 4, PointType.CORE),
 P7(7, 4, PointType.CORE),
 P8(9, 4, PointType.CORE),
 P9(3, 3, PointType.CORE),
 P10(8, 2, PointType.BORDER),
 P11(7, 5, PointType.CORE)]
```

Oppgave 10

Instance	Age	Car	Risk
1	25	Sports	L
2	20	Vintage	H
3	25	Sports	L
4	45	SUV	H
5	20	Sports	H
6	25	SUV	H

$GINI = 1 - \frac{2}{6}^2 - \frac{4}{6}^2 = 0.44$

	H:2, L:0	H:1, L:2	H:1, L0
Sorted	20	25	45
Split position	10	22.5	35
H ≤	0	2	3
L ≤	0	0	2
H >	4	2	1
L >	2	2	0
GINI ≤	$1 - 0 - 0 = 1$	$1 - \frac{2}{2}^2 - \frac{0}{2}^2 = 0.0$	$1 - \frac{3}{5}^2 - \frac{2}{5}^2 = 0.48$
GINI >	$1 - \frac{4}{6}^2 - \frac{2}{6}^2 = 0.44$	$1 - \frac{2}{4}^2 - \frac{2}{4}^2 = 0.5$	$1 - \frac{1}{1}^2 + \frac{0}{1}^2 = 0$

GINI	$\frac{0}{6} * 1 + \frac{6}{6} * 0.44 = 0.44$	$\frac{2}{6} * 0 + \frac{4}{6} * 0.5 = 0.33$	$\frac{1}{6} * 0 + \frac{5}{6} * 0.48 = 0.4$

	L	H
Sport	2	1
Vintage	0	1
SUV	0	2

$$GINI_{sport} = 1 - \frac{2}{3}^2 - \frac{1}{3}^2 = 0.44$$

$$GINI_{vintage} = 1 - \frac{0}{1}^2 - \frac{1}{1}^2 = 0.0$$

$$GINI_{SUV} = 1 - \frac{0}{2}^2 - \frac{2}{2}^2 = 0.0$$

$$GINI = \frac{3}{6} * 0.44 + \frac{1}{6} * 0 + \frac{2}{6} * 0 = 0.22$$

Best split on Car

Best splitt on Sport, all other cars are High

calculate AGE again