FINAL EXAMINATION
TDT4300

AUTUMN 2021

INFORMATION

- Academic contact during examination: Dhruv Gupta
- E-mail: dhruv.gupta@ntnu.no
- Examination date: 09-August-2021
- Examination time (from-to): 09:00-13:00
- Permitted examination support material: Open book
- Language: English

- Checked By:
- Date:
- Signature:

1 DATAWAREHOUSES AND OLAP OPERATIONS

Exercise 1. **Book Franchise**. Shops belonging to well-known book franchise sell various kinds of `printed materials` (e.g., recipe books, children books, novels etc.). Items in each shop are provided by `publishers` located around the world. The franchise currently has bookshops only in `locations` across Europe. The franchise utilizes an out-dated method of maintaining its `sales` across stores. Concretely, each shop-owner sends out a summary of everyday's sales to the headquarters for analysis. The CEO of the franchise wants to adopt the data warehousing approach for data analytics. As a new data scientist at the franchise headquarters you are tasked with the implementation of their data warehouses. Answer the questions below and state any assumptions you have made to model the data warehouse.

1. Create the concept hierarchies for the different dimensions that are part of the above problem statement.
2. Create a star schema to implement the data warehouse.
3. Additionally, create a snowflake schema to implement the data warehouse. What is the one key feature that would make snowflake schema more useful for implementing the data warehouse as compared to star schema.

Solution 1

1. Concept Hierarchies.
   a) Printed Material: Name → Author → Type → Genre → ALL.
   b) Location: City → Country → Continent → ALL.
   c) Time: Day → Week → Month → Year → ALL.
   d) Shop: Location → Type → ALL.
2. Star Schema.

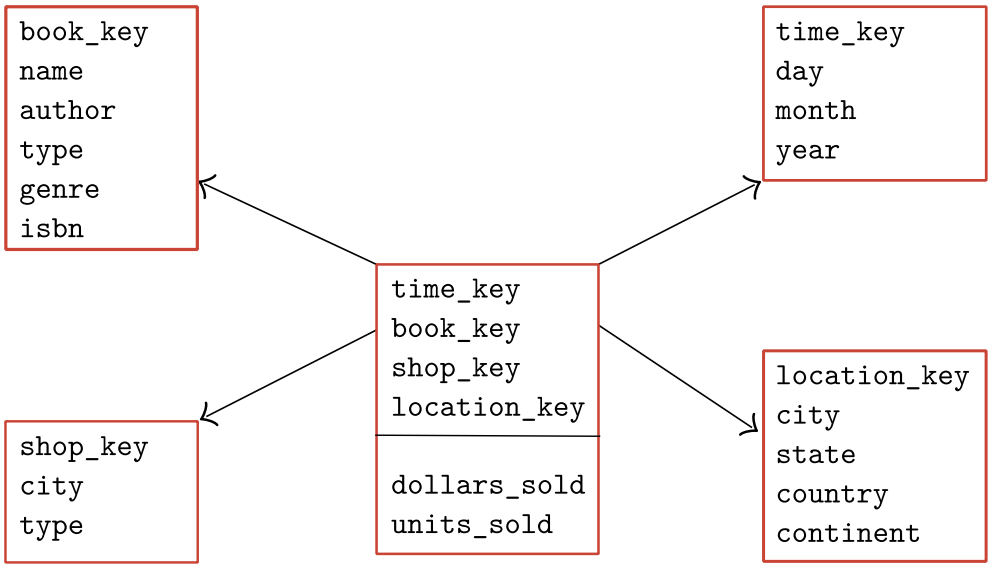


Figure 1: Star Schema.

3. Snowflake Schema.

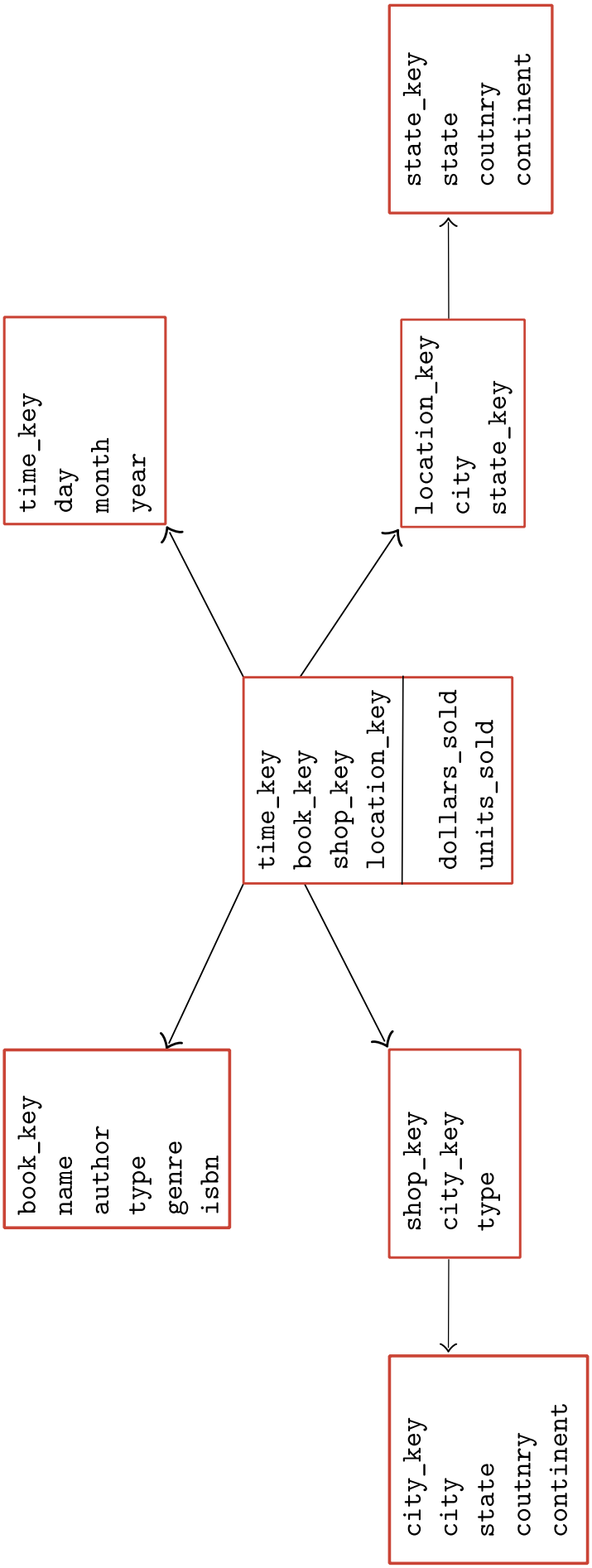


Figure 2: Snowflake Schema.

Exercise 2. Oracle has multiple software engineering teams across Norway. The following table details of some of the employees in the country. Due to the ongoing Covid-19 pandemic, Oracle would like to implement the hybrid-workplaces concept. Specifically, to gain insights into the working habits of its employees Oracle wants to utilize data warehouses which store its employee records. To speedup the query processing Oracle utilizes the concept of bitmap indexes in its data warehouse solution. How would the bitmap indexes look like for the sample data below? Having constructed the bitmap indexes answer the following questions to help Oracle.

Employee ID	Name	Gender	Dependents	Office	Title
1	Magnus	Male	Yes	Oslo	Developer
2	Kjetil	Male	Yes	Bergen	Tester
3	Anna	Female	No	Trondheim	Developer
4	Charlotte	Female	No	Oslo	Tester
5	John	Male	Yes	Oslo	Project manager
6	Birgit	Female	Yes	Bergen	Developer
7	Robert	Male	No	Trondheim	Developer

Show the bitmap indices on Gender, Office, and Title attributes and use them to:

1. Identify developers that have dependents.
2. Identify employees that are situated outside Trondheim.
3. Identify male employees that reside outside Trondheim.
4. Identify testers working in Oslo or Bergen.

Solution 2

• Bit Map Indexes for `Gender`.

	1	2	3	4	5	6	7
Male	1	1	0	0	1	0	1
Female	0	0	1	1	0	1	0

• Bit Map Indexes for `Dependents`.

	1	2	3	4	5	6	7
Male	1	1	0	0	1	1	0
Female	0	0	1	1	0	0	1

• Bit Map Indexes for `Office`.

	1	2	3	4	5	6	7
Oslo	1	0	0	1	1	0	0
Bergen	0	1	0	0	0	1	0
Trondheim	0	0	1	0	0	0	1

• Bit Map Indexes for `Title`.

	1	2	3	4	5	6	7
Developer	1	0	1	0	0	1	1
Tester	0	1	0	1	0	0	0
Project Manager	0	0	0	0	1	0	0

- Answer to part 1.

`Title` = "`Developer`" AND `Dependents` = "`Yes`".

	1	2	3	4	5	6	7
`Title` = "`Developer`"	1	0	1	0	0	1	1
`Dependents` = "`Yes`"	1	1	0	0	1	1	0
AND	1	0	0	0	0	1	0

Answer = {Magnus, Birgit}.

- Answer to part 2.

¬(`Office` = "`Trondheim`").

	1	2	3	4	5	6	7
`Office` = "`Trondheim`"	0	0	1	0	0	0	1
¬	1	1	0	1	1	1	0

Answer = {Magnus, Kjetil, Charlotte, John, Birgit}.

- Answer to part 3.

`Gender` = "`Male`" AND $\big($¬(`Office` = "`Trondheim`")$\big)$.

	1	2	3	4	5	6	7
`Gender` = "`Male`"	1	1	0	0	1	0	1
¬`Office` = "`Trondheim`"	1	1	0	1	1	1	0
AND	1	1	0	0	1	0	0

Answer = {Magnus, Kjetil, John}.

- Answer to part 4.

`Title` = "`Tester`" AND $\big($(`Office` = "`Oslo`") OR (`Office` = "`Bergen`")$\big)$.

	1	2	3	4	5	6	7
`Office` = "`Oslo`"	1	0	0	1	1	0	0
`Office` = "`Bergen`"	0	1	0	0	0	1	0
OR	1	1	0	1	1	1	0

	1	2	3	4	5	6	7
`Title` = "`Tester`"	0	1	0	1	0	0	0
`Office` = "`Oslo`" OR "`Bergen`"	1	1	0	1	1	1	0
AND	0	1	0	1	0	0	0

Answer = {Kjetil, Charlotte}.

2 DATA

Exercise 3. Consider the data presented in the table below. What are some of the steps that you will take to clean the data before applying any classification methods (supervision is done based on the `Class` label) and why?

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.

Solution 3

1. Remove `Type` as it is correlated and same as `Class` attribute for training.
2. Similarly, attribute `A` and `C` are positively correlated. Remove `A` or `C`.
3. Similarly, attribute `B` and `D` are negatively correlated. Remove `B` or `D`.

Exercise 4. Consider a dataset that has 1×10^6 features and 1×10^9 instances. We need to apply a data mining algorithm to this massive dataset and measure its performance. Given its size, we would ideally select the most relevant subset of features to reduce computation cost. How many times would we need to run the algorithm for this ideal feature selection procedure? Do you think this is feasible? What would be some alternative methods of solving this problem (only name the methods)?

Solution 4

We need $2^n = 2^{10^6}$ iterations as there are 2^n subset of features. Other ways of feature selection or dimensionality reduction: PCA, LDA etc.

Exercise 5. Consider the following employee database. What is the attribute types for each of the attributes that are recorded for the employees.

Employee ID	Name	Age	Joining Date	Title
1	Magnus	23	01-09-2021	Developer
2	Kjetil	25	01-04-2020	Tester
3	Anna	30	01-01-2016	Developer
4	Charlotte	25	01-02-2019	Tester
5	John	35	01-06-2014	Project manager
6	Birgit	60	01-03-2000	Developer
7	Robert	55	01-07-2005	Developer

Solution 5

1. `Name` - nominal.
2. `Age` - ratio.
3. `Joining Date` - interval.
4. `Title` - Ordinal.

3 ASSOCATION RULE ANALYSIS

Exercise 6. Compute the frequent itemsets for the transaction database given in table below using the Apriori algorithm with minimum support equal to 3. While answering the question, also write down step-by-step procedure that would entail by applying the Apriori algorithm.

A	B	C	D	E
1	2	1	1	2
3	3	2	6	3
5	4	3		4
6	5	5		5
	6	6		

Solution 6

• Candidate 1-itemsets and frequent 1-itemsets.

C_1	Support
A	4
B	5
C	5
D	2
E	4

L_1	Support
A	4
B	5
C	5
E	4

• Candidate 2-itemsets and frequent 2-itemsets.

C_2	Support
AB	3
AC	4
AE	2
BC	4
BE	4
CE	3

L_2	Support
AB	3
AC	4
BC	4
BE	4
CE	3

• Candidate 3-itemsets and frequent 3-itemsets.

C_3	Support
ABC	3
BCE	3

L_3	Support
ABC	3
BCE	3

There is no need to scan the transaction database for frequent 4-itemsets as there are no candidates.

Exercise 7. Compute the frequent itemsets for the transaction database given in the table below using the FPGrowth algorithm with minimum support equal to 3. Show the FPGrowth procedure step-by-step including the building of the FP-Tree and the projected FP Trees.

A	B	C	D	E
1	2	1	1	2
3	3	2	6	3
5	4	3		4
6	5	5		5
	6	6		

Solution 7 • 1-itemset support values.

1-Itemset	Support
A	4
B	5
C	5
D	2
E	4

• 1-itemset reordered based on support values.

1-Itemset	Support
B	5
C	5
A	4
E	4
D	2

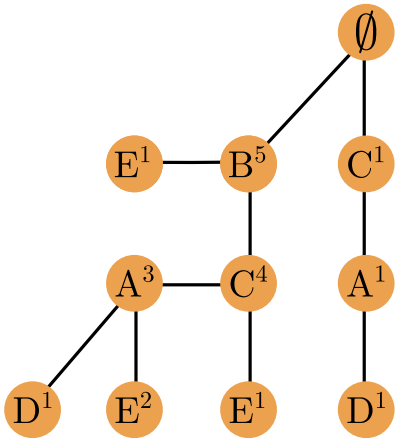
• Transaction database.

tid	Transaction
1	ACD
2	BCE
3	ABCE
4	BE
5	ABCE
6	ABCD

• Reordered transaction database.

tid	Transaction
1	CAD
2	BCE
3	BCAE
4	BE
5	BCAE
6	BCAD

• FP-Tree for the entire transaction database.



1. **Project on** D**.**
We can remove D from the FP-Tree as its support (2) is less than min. support (3).

Frequent Path = {∅}

2. **Project on** E**.**

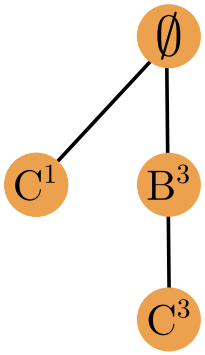
Path	Count
BCAE	2
BCE	1
BE	1

– Projected FP-Tree for E: $\emptyset \rightarrow B^4 \rightarrow C^3 \rightarrow A^2$.
Since, support for A is less than minimum support we can remove it for frequent itemset generation.
Frequent Path = {EB(4), EC(3), EBC(3)}

4. **Project on** A**.**

Path	Count
CA	1
BCA	3

– Projected FP-Tree for A:



2.1 **Project on** AC**.**

Path	Count
BC	3
B	1

– Projected FP-Tree for AC: $\emptyset \rightarrow B^3$.
Frequent Pattern = $\{AC(4), ACB(3)\}$.

2.2 **Project on** AB**.**

Path	Count
B	3

Frequent Pattern = $\{AB(3)\}$.

4. **Project on** C**.**

Path	Count
C	1
BC	4

– Projected FP-Tree for C: $\emptyset \rightarrow B^4$.
Frequent Path = $\{CB(4)\}$

5. **Project on** B**.**

Path	Count
B	5

Frequent Path = $\{\emptyset\}$

4 CLUSTERING

Exercise 8. **K-Means Clustering.** Consider the dataset given below. Apply the k-Means Clustering algorithm assuming $k = 2$ and initial cluster assignments are: $C_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4\}$ and $C_2 = \{\mathbf{x}_3, \mathbf{x}_5\}$. Show the iterations of the k-Means Clustering algorithm until it converges. Utilize the L_1 norm (also known as the Manhattan Distance) for distance computations.

Instance	X_1	X_2
$\mathbf{x}_1$	0	2
$\mathbf{x}_2$	0	0
$\mathbf{x}_3$	1.5	0
$\mathbf{x}_4$	5	0
$\mathbf{x}_5$	5	2

Table 2: Table for K-means based exercise.

Solution 8 1 **Iteration 1.**

$$
\begin{aligned}
k &= 2 \\
C_1 &= \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4\} \\
C_2 &= \{\mathbf{x}_3, \mathbf{x}_5\}
\end{aligned}
$$

Initial Centroids:

$$
\begin{aligned}
C_1 &= \big(\text{median}(0,0,5), \text{median}(0,0,2)\big) = (0,0). \\
C_2 &= \big(\text{median}(1.5,5), \text{median}(0,2)\big) = (3.25,1).
\end{aligned}
$$

Distances of the data points to the two centroids:

id	X_1	X_2	d_1	d_2
$\mathbf{x}_1$	0	2	2	4.25
$\mathbf{x}_2$	0	0	0	4.25
$\mathbf{x}_3$	1.5	0	1.5	2.75
$\mathbf{x}_4$	5	0	5	2.75
$\mathbf{x}_5$	5	2	7	2.75

$$
\begin{aligned}
C_1 &= \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} \\
C_2 &= \{\mathbf{x}_4, \mathbf{x}_5\}
\end{aligned}
$$

Updated Centroids:

$$
\begin{aligned}
C_1 &= \big(\text{median}(0,0,1.5), \text{median}(0,0,2)\big) = (0,0). \\
C_2 &= \big(\text{median}(5,5), \text{median}(0,2)\big) = (5,1).
\end{aligned}
$$

2 **Iteration 2.**
Distances of the data points to the two centroids:

$$
\begin{aligned}
C_1 &= \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} \\
C_2 &= \{\mathbf{x}_4, \mathbf{x}_5\}
\end{aligned}
$$

id	X_1	X_2	d_1	d_2
$\mathbf{x}_1$	0	2	2	6
$\mathbf{x}_2$	0	0	0	6
$\mathbf{x}_3$	1.5	0	1.5	4.5
$\mathbf{x}_4$	5	0	5	1
$\mathbf{x}_5$	5	2	7	2

Updated Centroids:

$$C_1 = \big(\text{median}(0,0,1.5), \text{median}(0,0,2)\big) = (0,0).$$
$$C_2 = \big(\text{median}(5,5), \text{median}(0,2)\big) = (5,1).$$

Centroids and cluster assignments repeat hereafter.

Exercise 9. **DBScan Algorithm.** DBScan algorithm allows the discovery of clusters of arbitrary shapes. We would like to process the given set of points in the figure below with the DBScan algorithm. To do so, use the following parameters: $\texttt{eps} = 1$ and $\texttt{minpts} = 6$. Also, consider the following distance function L_{min},

$$L_{min}(x,y) = \min_{i=1}^{d} \{|x_i - y_i|\}.$$

As an example computation of distances, consider two points $x = \langle 1,2\rangle$ and $y = \langle 2,4\rangle$. Then, L_{min} is computed as:

$$\begin{aligned} L_{min}(x,y) &= \min_{i=1}^{2} \{|x_i - y_i|\} \\ &= \min\{|1-2|, |2-4|\} \\ &= \min\{1,2\} \\ &= 1. \end{aligned}$$

Specifically, identify the noise points, border points, and core points of the clusters while answering the question.

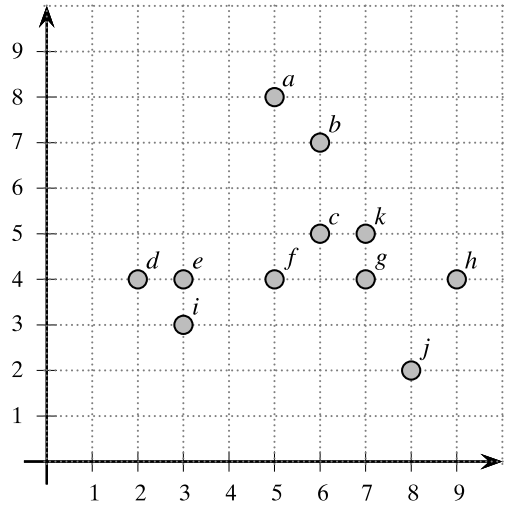


Figure 3: Figure for the hierarchical agglomerative clustering and density based clustering.

Solution 9 Distance Matrix

	a	b	c	d	e	f	g	h	i	j	k
a											
b	1										
c	1	0									
d	3	3	1								
e	2	3	1	0							
f	0	1	1	0	0						
g	2	1	1	0	0	0					
h	4	3	1	0	0	0	0				
i	2	3	2	1	0	1	1	1			
j	3	2	2	2	2	2	1	1	1		
k	2	1	1	1	1	0	1	2	1	0	

Density

	a	b	c	d	e	f	g	h	i	j	k
Density	4	6	6	8	8	10	10	9	7	5	9

- Core Points = {b,c,d,ef,g,h,i,k}, as their density (number of points within their `eps` radius) is $\geqslant 6$.
- Border Points = {a, j}, as they are in the vicinity (within `eps` radius) of a core point but their density is < 6.
- Noise Points = {$\emptyset$}.

5 CLASSIFICATION

Exercise 10. To build decision trees, the simplest algorithm to use is the Hunt's algorithm. For the dataset given in Table 3, create a decision tree using the Hunt's algorithm. For the evaluation criteria utilize the Gini Index. Having constructed the decision tree, what class does the instance $\langle \texttt{Age} = 27, \texttt{Car} = \texttt{Vintage} \rangle$ belong to?

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

Table 3: Table for decision tree based exercise.

Solution 10 1. Determination of root node split.

- Split on `Age`:

	2×H		2×L/H		H			
Data Point	20		25		45			
Split Point	10		22.5		35		50	
	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$
Class='L'	0	2	0	2	2	0	2	0
Class='H"	0	4	2	2	3	1	4	0
Gini Index	$0.\bar{4}$		$0.\bar{3}$		0.4		$0.\bar{4}$	

– Split on comparing `Age` with 10:

Age $\leqslant$ 10	
L	0
H	0

$$\text{Gini} = 1 - 0 - 0 = 1.$$

Age $>$ 10	
L	2
H	4

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{2}{6}\right)^2 \\ &= 1 - \frac{1}{9} - \frac{4}{9} \\ &= \frac{9-5}{9} = \frac{4}{9} \\ &= 0.\bar{4}.\end{aligned}$$

$$\begin{aligned}\text{Gini} &= \frac{0}{6} \cdot 1 + \frac{6}{6} \cdot \frac{4}{9} \\ &= 0.\bar{4}.\end{aligned}$$

– Split on comparing `Age` with 22.5:

Age $\leqslant$ 10	
L	0
H	2

$$\text{Gini} = 1 - 0 - 1 = 0.$$

Age $>$ 10	
L	2
H	2

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \\ &= 1 - \frac{1}{4} - \frac{1}{4} \\ &= \frac{1}{2} \\ &= 0.5.\end{aligned}$$

$$\begin{aligned}\text{Gini} &= \frac{2}{6} \cdot 0 + \frac{4}{6} \cdot \frac{1}{2} \\ &= \frac{2}{6} = \frac{1}{3} = 0.\bar{3}.\end{aligned}$$

– Split on comparing `Age` with 35:

Age $\leqslant$ 35	
L	2
H	3

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 \\ &= 1 - \frac{4}{25} - \frac{9}{25} \\ &= \frac{25-13}{25} = \frac{12}{25} \\ &= 0.48.\end{aligned}$$

Age $>$ 35	
L	0
H	1

$$\text{Gini} = 1 - 0 - 1 = 0$$

$$\begin{aligned}\text{Gini} &= \frac{5}{6} \cdot \frac{12}{25} + \frac{1}{6} \cdot 0 \\ &= \frac{2}{5} = 0.4.\end{aligned}$$

– Split on comparing `Age` with 50:

`Age` $\leqslant 50$
L	2
H	4

$$\begin{aligned}\text{Gini} &= 1-\left(\frac{2}{6}\right)^2-\left(\frac{2}{6}\right)^2\\ &= 1-\frac{1}{9}-\frac{4}{9}\\ &= \frac{9-5}{9}=\frac{4}{9}\\ &= 0.\bar{4}.\end{aligned}$$

`Age` > 50
L	0
H	0

$$\text{Gini} = 1-0-0 = 1.$$

$$\begin{aligned}\text{Gini} &= \frac{6}{6}\cdot\frac{4}{9}+\frac{0}{6}\cdot 1\\ &= 0.\bar{4}.\end{aligned}$$

• Split on `Car`:
– Multi-way split on `Car`:

`Car` = `Sports`
L	2
H	1

$$\begin{aligned}\text{Gini} &= 1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2\\ &= 1-\frac{1}{9}-\frac{4}{9}\\ &= \frac{9-5}{9}=\frac{4}{9}\\ &= 0.\bar{4}.\end{aligned}$$

`Car` = `Vintage`
L	0
H	1

$$\text{Gini} = 1-0-1 = 0$$

`Car` = `SUV`
L	0
H	2

$$\text{Gini} = 1-0-1 = 0$$

$$\text{Gini} = \frac{3}{6}\cdot\frac{4}{9}+0+0 = \frac{2}{9} = 0.\bar{2}$$

– Best binary split on `Car`:

`Car` = `Sports`
L	2
H	1

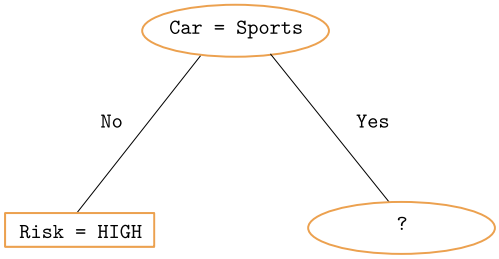
$$\begin{aligned}\text{Gini} &= 1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2\\ &= 1-\frac{1}{9}-\frac{4}{9}\\ &= \frac{9-5}{9}=\frac{4}{9}\\ &= 0.\bar{4}.\end{aligned}$$

`Car` $\in$ {`SUV`, `Vintage`}
L	0
H	3

$$\text{Gini} = 1-0-1 = 0$$

$$\text{Gini} = \frac{3}{6}\cdot\frac{4}{9}+0 = \frac{2}{9} = 0.\bar{2}$$

• Decision tree based on splitting on `Car`.



2. Determination of next node split.
• Split on `Age`:

	H	2×L	
Data Point	20	25	
Split Point	10	22.5	30
	⩽ | >	⩽ | >	⩽ | >
Class='L'	0 | 2	0 | 2	2 | 0
Class='H"	0 | 1	1 | 0	1 | 0
Gini Index	$0.\bar{4}$	$0.\bar{3}$	0.4

– Split on comparing `Age` with 10:

Age ⩽ 10	
L	0
H	0

$$\text{Gini} = 1 - 0 - 0 = 1.$$

Age > 10	
L	2
H	1

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \\ &= 1 - \frac{4}{9} - \frac{1}{9} \\ &= \frac{9-5}{9} = \frac{4}{9} \\ &= 0.\bar{4}.\end{aligned}$$

$$\begin{aligned}\text{Gini} &= \frac{0}{3} \cdot 1 + \frac{3}{3} \cdot \frac{4}{9} \\ &= 0.\bar{4}.\end{aligned}$$

– Split on comparing `Age` with 22.5:

Age ⩽ 22.5	
L	0
H	1

$$\text{Gini} = 1 - 0 - 1 = 0.$$

Age > 10	
L	2
H	0

$$\text{Gini} = 1 - 1 - 0 = 0.$$

$$\text{Gini} = 0.$$

– Split on comparing Age with 30:

Age $\leqslant 35$	
L	2
H	1

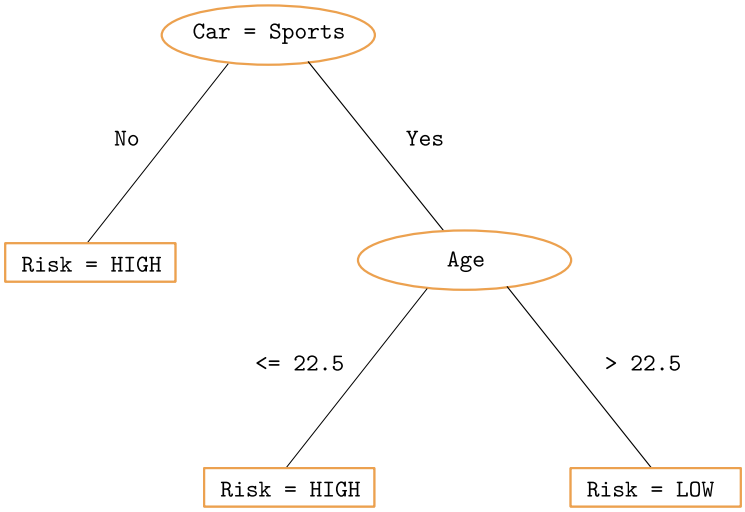
$$\begin{aligned}\text{Gini} &= 1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\\ &= 1-\frac{4}{9}-\frac{1}{9}\\ &= \frac{9-5}{9}=\frac{4}{9}\\ &= 0.\bar{4}.\end{aligned}$$

Age > 35	
L	0
H	0

$$\text{Gini} = 1-0-0=1$$

$$\begin{aligned}\text{Gini} &= \frac{3}{3}\cdot\frac{4}{9}+\frac{0}{3}\cdot 1\\ &= \frac{4}{9}=0.\bar{4}.\end{aligned}$$

• Final decision tree based on splitting on Age.



• The given instance $\langle \texttt{Age} = 27, \texttt{Car} = \textit{Vintage}\rangle$ would be classified with the risk label of High.