FINAL EXAMINATION TDT4300

SPRING 2022

INFORMATION

- Academic contact during examination: Dhruv Gupta
- E-mail: dhruv.gupta@ntnu.no
- Examination date: 24-05-2022
- Examination time (from-to): 15:00-18:00
- Permitted examination support material: Open book
- Language: English

- Checked By:
- Date:
- Signature:

1 DATAWAREHOUSES AND OLAP OPERATIONS

Exercise 1. United Nations (UN) in its initiative to address climate change collects meteorological data from all over the world. You work as a data scientist at the UN and are tasked with the objective of analyzing global weather reports to identify unusual weather patterns (e.g., heat waves). The meteorological departments of each country provide information for `temperature` in degree Celsius for each major city four times daily (morning, afternoon, evening, and night). To detect and understand unusual weather patterns they also provide information for `air quality` as an index value in $[0, 201)$, `wind speed` in km/h, `precipitation` in mm/hr, and `UV levels` as index values in $[1, 11)$. With this context answer following questions pertaining to data analysis:

1. You are tasked with the providing insights from this large weather data. You initially decide to use relational database techniques. What are some problems you would face when trying to cross-tabulate the data with traditional SQL?

2. Given the challenges with relational database techniques, you decide to store your data in a data warehouse. To this end, you must first design concept hierarchies for the different attributes in the dataset. Write down the hierarchies you propose. Remember to come up with these hierarchies you may need to map continuous attributes to categorical levels. State these mappings and any other assumption you make.

3. Based on the concept hierarchies created above, design a star schema for storing the data in a data warehouse.

4. Given the concept hierarchy you have designed, how many cuboids will be needed for the full materialization for the cube?

5. At the upcoming UN Climate Change Conference (COP26), you are tasked with presenting the insights you identified from the dataset. One of the important questions to answer is: what is rise in average temperature this year as compared to last five years across the world? In particular, identify the sequence of OLAP operations you need to perform on the base cuboid to find the rise in average temperature this year as compared to last decade across Europe when extreme conditions existed in terms of precipitation, air quality, uv levels, and wind speeds (i.e., all at one end of your concept hierarchies).

Hint: For the OLAP operations we are expecting operations of the type "Roll Up, Drill Down, Slice, Dice, and Pivot" to arrive at the sub-cube or cross-tab to visualize the answer. For example, for the query "what are the total computer sales by Florida for quarter Q1" the answer is:
Roll Up Location: City -> State; Roll Up Time : Weeks -> Quarter;
Dice: State = "Florida" AND Quarter = "Q1";

1

1. Analytical queries that require cross-tabulation, drill-down and roll-up using SQL become too cumbersome. For example, for a N dimensional roll-up requires writing a complex SQL-query consisting of N group bys and unions.

2. Concept Hierarchies:
   a) Time Concept Hierarchy: See Figure 1.
   b) Location Concept Hierarchy:
   city → state → country → continent → ALL.
   c) Precipitation Concept Hierarchy: See Figure 2.
   d) UV Index Hierarchy: See Figure 3.
   e) Air Quality Index Concept Hierarchy: See Figure 4.
   f) Wind Speed Concept Hierarchy: See Figure 5.

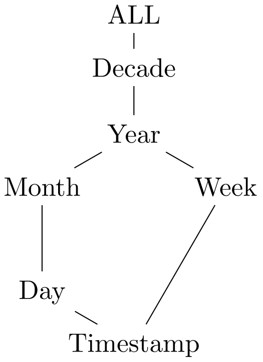


Figure 1: Time Concept Hierarchy

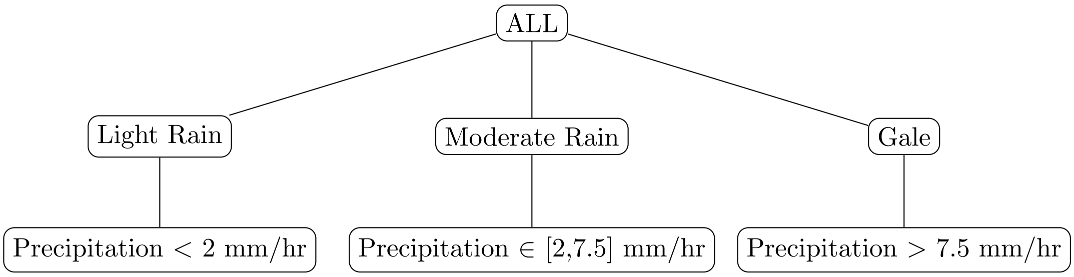


Figure 2: Concept Hierarchy for Precipitation Levels

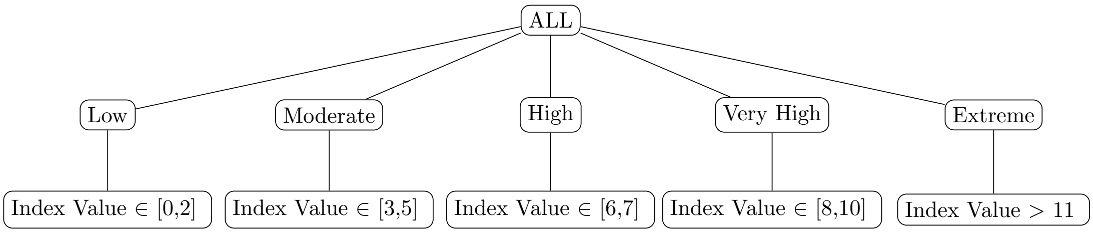


Figure 3: Concept Hierarchy for UV Index

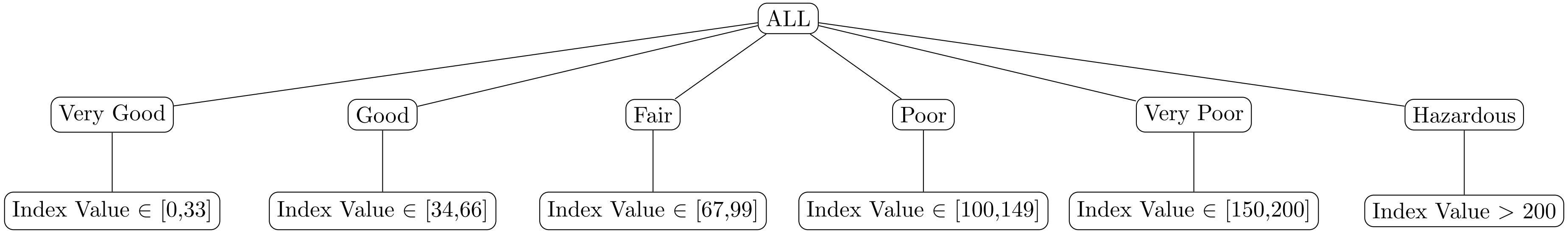


Figure 4: Concept Hierarchy for Air Quality Index

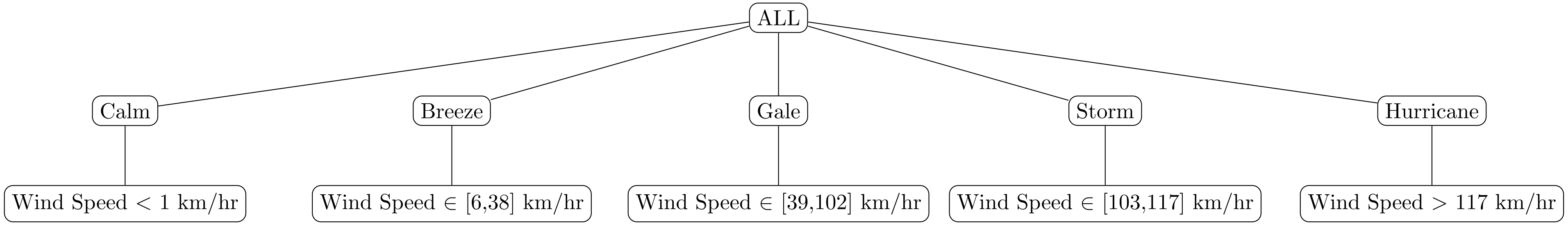

ALL
Calm
Breeze
Gale
Storm
Hurricane
Wind Speed < 1 km/hr
Wind Speed ∈ [6,38] km/hr
Wind Speed ∈ [39,102] km/hr
Wind Speed ∈ [103,117] km/hr
Wind Speed > 117 km/hr

Figure 5: Concept Hierarchy for Wind Speed

3. Star Schema: See Figure 6.

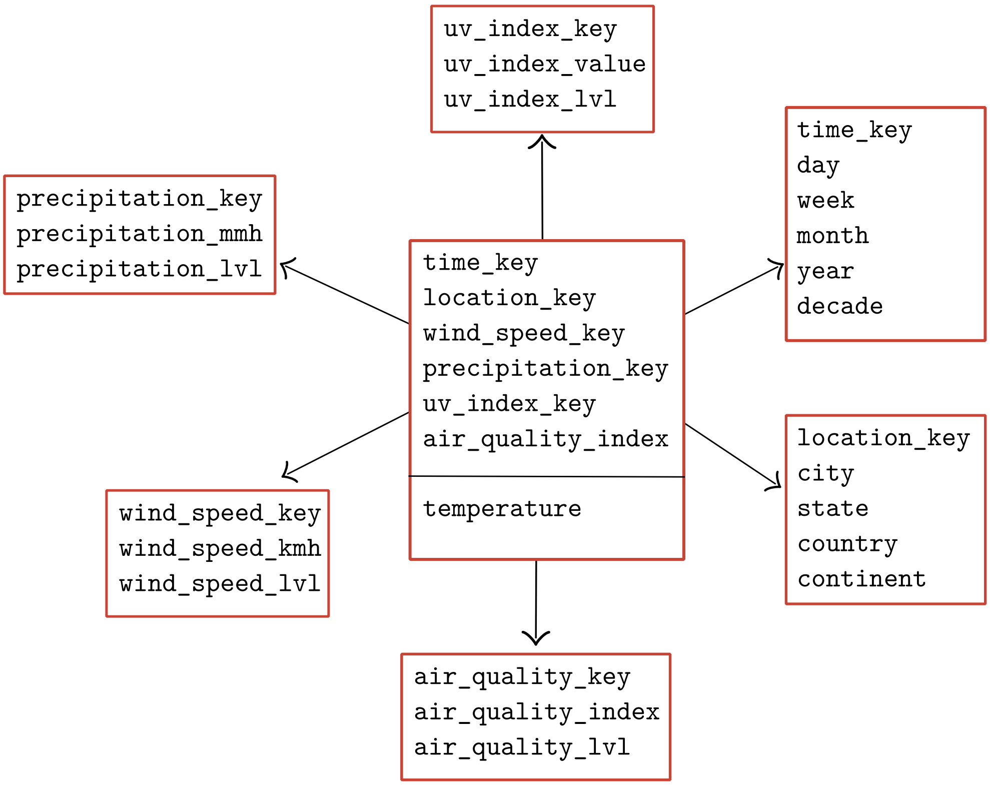


Figure 6: Star Schema.

4. Number of cuboids is given by the following formula:

$$\text{Total Number of Cubiods} = \prod_{i=1}^{n} \left(L_i + 1\right), \tag{1}$$

where, L_i is the number of levels in dimension i. We have following number of levels for each hierarchy:

   a) Time: 5.

   b) Location: 4.

   c) Precipitation: 2.

   d) UV Index: 2.

   e) Air Quality: 2.

   f) Wind Speed: 2.

Therefore, the total number of cuboids is:

$$\text{Total Number of Cubiods} = 5 \times 5 \times 3 \times 3 \times 3 \times 3 = 2430. \tag{2}$$

5. Following are the two sets of OLAP operations that would be needed to compute the two averages: once over the last decade and once over the current year.

ROLL-UP location	:city → continent.
ROLL-UP time	:timestamp → decade.
ROLL-UP precipitation	:precipitation_value → precipitation_level.
ROLL-UP uv_index	:uv_value → uv_level.
ROLL-UP air_quality	:air_quality_value → air_quality_level.
ROLL-UP uv_wind_speed	:wind_speed_value → wind_speed_level.
DICE	:continent = "Europe" AND time="2010s" AND precipitation_level = "Gale" AND uv_index_level = "Extreme" AND air_quality_level = "Hazardous" AND wind_speed_level = "hurricane".

ROLL-UP location	:city → continent.
ROLL-UP time	:timestamp → decade.
ROLL-UP precipitation	:precipitation_value → precipitation_level.
ROLL-UP uv_index	:uv_value → uv_level.
ROLL-UP air_quality	:air_quality_value → air_quality_level.
ROLL-UP uv_wind_speed	:wind_speed_value → wind_speed_level.
DICE	:continent = "Europe" AND time="2022" AND precipitation_level = "Gale" AND uv_index_level = "Extreme" AND air_quality_level = "Hazardous" AND wind_speed_level = "hurricane".

2 DATA

Exercise 2. You are working as a Data Engineer at Piazza, which hosts a platform where students belonging to a course can ask and discuss questions. You are tasked with building a system for analyzing student's posts on Piazza for the TDT4300 course. Following are the questions that arise in your attempt to build this system:

1. In order to analyze posts, first you need to model the posts attributes. What can be potential attributes used to model a post computationally?
2. What would be the associated type (e.g., discrete, continuous, interval etc.) of the attributes you have come up with?
3. One important functionality on Piazza is to search for similar posts based on text. Based on what you know about different types of datasets. How would model the text contained in the posts so that search functionality can be implemented? How would represent the following question snippet: "How do we solve the question on Apriori algorithm?".
4. In order to assist both the instructors and the students, Piazza provides a functionality to detect posts that ask the same question. Based on the text model chosen above what similarity measure will you choose to implement this duplicate-detection functionality?

2

1. Following are potential attributes that can be used to model a post:
   a) Question ID
   b) Author Name / ID
   c) Number of Views
   d) Good Question Likes
   e) Question Text
   f) Question Tags
2. Following are the types associated with the attributes identified above:
   a) Question ID: Discrete and Nominal
   b) Author Name / ID: Discrete and Nominal
   c) Number of views: Discrete and Ordinal
   d) Good Question Likes: Discrete and Ordinal
   e) Question Text: Discrete and Nominal
   f) Question Tags: Discrete and Nominal
3. One would use a term-document matrix to model the text contained in the posts. For the text snippet following is the term-document matrix:

	how	do	we	solve	the	question	on	apriori	algorithm
d_1	1	1	1	1	1	1	1	1	1

Table 1: Term Document Matrix.

4. One would use the cosine similarity measure.

3 ASSOCIATION RULE ANALYSIS

Exercise 3. Compute the frequent itemsets for the transaction database given in Table 2 using the FPGrowth algorithm with minimum support equal to 3.

tid	**itemset**
t_1	ACDEF
t_2	ABCDE
t_3	BCF
t_4	ACDEF
t_5	DB

Table 2: Transaction database.

3 1. 1-itemset support values.

itemset	**Support**
A	3
B	3
C	4
D	4
E	3
F	3

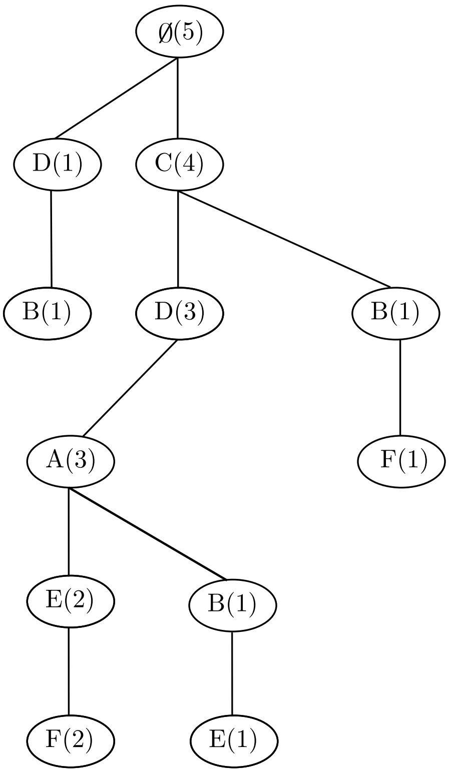
2. 1-itemset reordered based on support values.

itemset	**Support**
C	4
D	4
A	3
B	3
E	3
F	3

3. Reordered transaction database:

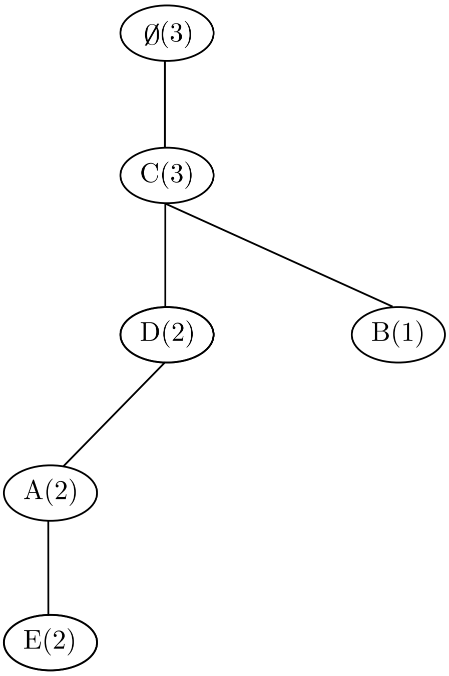
tid	**itemset**
t_1	CDAEF
t_2	CDABE
t_3	CBF
t_4	CDAEF
t_5	DB

4. FP-Tree for the entire transaction database.



5. Projected FP-Tree for: F

Path	Count
CDAEF	2
CBF	1



Frequent Itemsets: {F(3), FC (3)}.

6. Projected FP-Tree for: E

Path	Count
CDAEF	2
CDABE	1

$\phi(3) \rightarrow C(3) \rightarrow A(3) \rightarrow B(1)$.

Frequent Itemsets: {E(3), EA(3), ED(3), EC(3), EAD(3), EAC(3), ECD(3), EACD(3)}.

7. Projected FP-Tree for: A

Path	Count
CDA	3

$\phi(3) \rightarrow C(3) \rightarrow D(3)$.

Frequent Itemsets: {A(3), AC(3), AD(3), ACD(3)}.

8. Projected FP-Tree for: D

Path	Count
D	1
CD	3

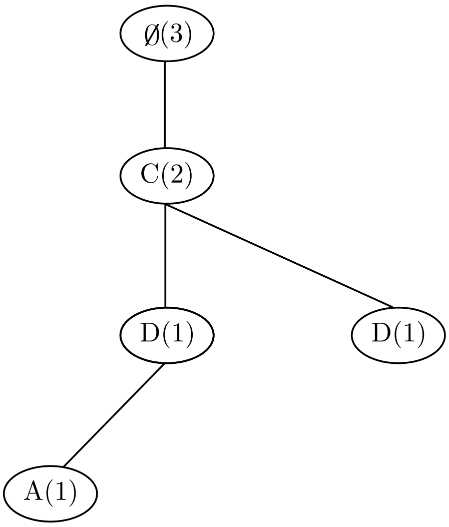
$\phi(3) \rightarrow C(3)$.

Frequent Itemsets: {D(4), DC(3)}.

9. Projected FP-Tree for: B

Path	Count
CB	1
DB	1
CDAB	1

Frequent Itemset: $\{B(3)\}$.



a) Projected FP-Tree for: BA

Path	Count
CDA	1

$\phi(1) \rightarrow C(1) \rightarrow D(1)$.

Frequent Itemsets: None.

b) Projected FP-Tree for: BD

Path	Count
CD	1
D	1

$\phi(2) \rightarrow C(1)$.

Frequent Itemsets: None.

c) Projected FP-Tree for: BC

Path	Count
C	2

$\phi(2)$.

Frequent Itemsets: None.

4 CLUSTERING

Exercise 4. Dendograms succinctly capture the result of a hierarchical agglomerative clustering. Fundamentally, a dendogram is a binary tree. For a given set of n points, how many possible dendograms can be enumerated? Hint: A tree with m nodes contains $m-1$ edges. Furthermore, a binary tree with n leaf nodes has $n-1$ internal nodes.

4 A tree with m nodes contains $m-1$ edges. Also, a binary tree with n leaf nodes has $n-1$ internal nodes.

Consider that a dendogram at a given stage of clustering has t leaf nodes (corresponding to the individual data points). Then it has a total of $t+t-1 = 2\cdot t-1$ nodes. This implies it has a total of $2\cdot t-1-1 = 2\cdot t-2$ edges.

To continue with the hierarchical clustering algorithm we need to grow the dendogram by another leaf node: $t+1$. This gives us $2\cdot t-2+1 = 2\cdot -1$ edges. Consequently, the following number of dendograms that we can enumerate:

$$\prod_{t=1}^{n-1}(2\cdot t-1) = 1\cdot 3\cdot 5\cdot 7\ldots(2\cdot n-3). \tag{3}$$

Exercise 5. Consider the points in Figure 7 and take the following distance measure:

$$L_{min} = \min_{i=1}^{2}\{|x_i - y_i|\}$$

Using $\epsilon = 2$, `minpts` $= 2$, and L_{min}, answer following questions if we are to apply DB-Scan clustering algorithm.

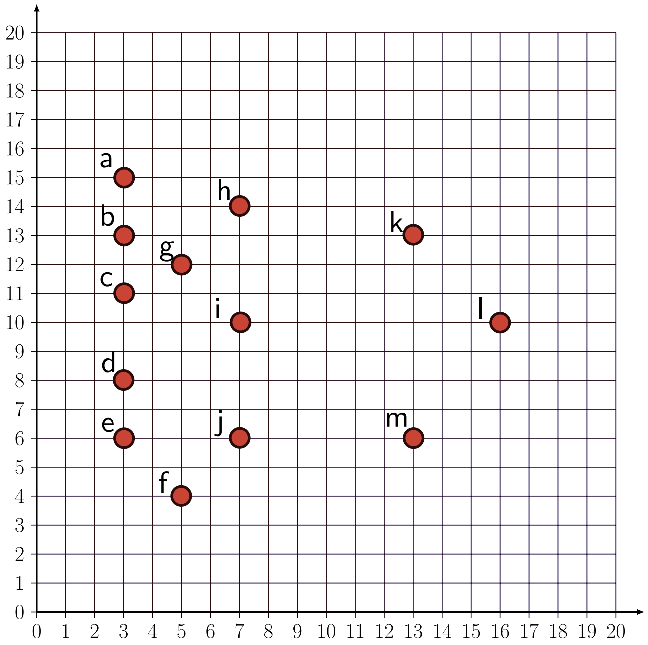


Figure 7: Figure for density based clustering.

1. Find all the core points for the set of points shown in Figure 7.
2. Two points are said to be density reachable if there is a sequence of core points leading up to the destination. For example, v is density reachable from u if there is a sequence of core points from u to v. Is the point j density reachable from point b in Figure 7? Explain your reasoning.
3. Is density reachability a symmetric relationship? Explain your reasoning.

	a	b	c	d	e	f	g	h	i	j	k	l	m
a													
b	0												
c	0	0											
d	0	0	0										
e	0	0	0	0									
f	2	2	2	2	2								
g	2	1	1	2	2	0							
h	1	1	3	4	4	2	2						
i	4	3	1	2	4	2	2	0					
j	4	4	4	2	0	2	2	0	0				
k	2	0	2	5	7	8	1	1	3	6			
l	5	3	1	2	4	6	2	4	0	4	3		
m	9	7	5	2	0	2	6	6	4	0	0	3	

4 Distance Matrix
Density

	a	b	c	d	e	f	g	h	i	j	k	l	m
Density	9	9	10	11	9	11	12	8	8	8	7	5	6

1. All points are core points.
2. Yes.
3. No, because of the asymmetric nature of the relationship there may not be a seqeuence of core points from a border point to a core point.

5 CLASSIFICATION

Exercise 6. Construct a decision tree using the Hunt's Algorithm for the dataset given in Table 3 where the attribute **Class** is the classification label for each record. Use the Gini index for determining the best split points.

Instance	a_1	a_2	a_3	**Class**
1	M	X	A	YES
2	F	Y	B	YES
3	M	Y	C	YES
4	F	Y	C	YES
5	M	X	C	YES
6	F	Y	D	NO
7	M	Y	A	NO
8	F	X	A	NO
9	M	Y	A	NO
10	F	X	C	NO

Table 3: Table for decision tree based exercise.

5 1. Determination of root node split.

- Gini at root node:

Class=YES	5
Class=NO	5
Gini Index	0.50

– Split a_1:

$a_1 = M$	
YES	3
NO	2

$$\text{Gini} = 1 - \frac{3^2}{5^2} - \frac{2^2}{5^2} = \frac{12}{25} = 0.48.$$

$a_1 = F$	
YES	2
NO	3

$$\text{Gini} = 1 - \frac{2^2}{5^2} - \frac{3^2}{5^2} = \frac{12}{25} = 0.48.$$

$$\begin{aligned}\text{Gini} &= \frac{5}{10} \cdot \frac{12}{25} + \frac{5}{10} \cdot \frac{12}{25} = \frac{12}{25} \\ &= 0.48.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= 0.50 - 0.48 \\ &= 0.02.\end{aligned}$$

– Split a_2:

$a_2 = X$	
YES	2
NO	2

$$\text{Gini} = 1 - \frac{2^2}{4^2} - \frac{2^2}{4^2} = \frac{1}{2} = 0.5.$$

$a_2 = Y$	
YES	3
NO	3

$$\text{Gini} = 1 - \frac{2^2}{4^2} - \frac{2^2}{4^2} = \frac{1}{2} = 0.5.$$

$$\begin{aligned}\text{Gini} &= \frac{4}{10} \cdot \frac{1}{2} + \frac{6}{10} \cdot \frac{1}{5} = \frac{1}{2} \\ &= 0.50.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= 0.50 - 0.5 \\ &= 0.00.\end{aligned}$$

– Split a_3:

$a_3 = A$	
YES	1
NO	3

$$\text{Gini} = 1 - \frac{1^2}{4^2} - \frac{3^2}{4^2} = \frac{3}{8} = 0.375.$$

$a_3 = B$	
YES	1
NO	0

$$\text{Gini} = 1 - \frac{1^2}{1^2} = 0.00.$$

$a_3 = C$	
YES	3
NO	1

$$\text{Gini} = 1 - \frac{3^2}{4^2} - \frac{1^2}{4^2} = \frac{3}{8} = 0.375.$$

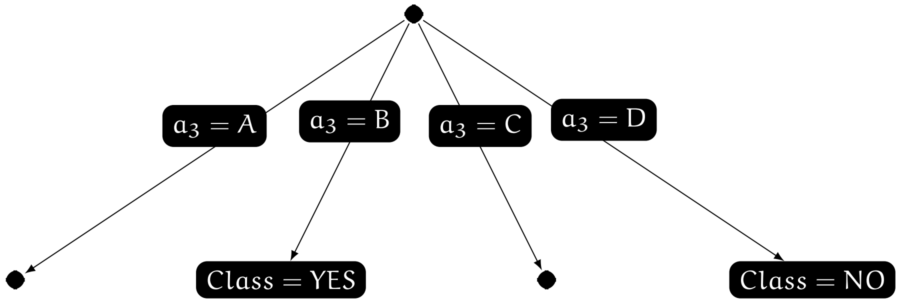
$a_3 = D$	
YES	0
NO	1

$$\text{Gini} = 1 - \frac{1^2}{1^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{4}{10} \cdot \frac{3}{8} + \frac{1}{10} \cdot 0 + \frac{4}{10} \cdot \frac{3}{8} + \frac{1}{10} \cdot 0 = \frac{3}{10} \\ &= 0.30.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= 0.50 - 0.30\\ &= 0.20.\end{aligned}$$

Split on $a_3 \in \{A, B, C, D\}$.



• Split on $a_3 = A$:
Gini at root node:

$a_2 = X$	
YES	1
NO	3

$$\text{Gini} = 1 - \frac{1^2}{4^2} - \frac{3^2}{4^2} = \frac{3}{8} = 0.375.$$

– Split a_1:

$a_1 = M$	
YES	1
NO	2

$$\text{Gini} = 1 - \frac{1^2}{3^2} - \frac{2^2}{3^2} = \frac{4}{9} = 0.\bar{4}.$$

$a_1 = F$	
YES	0
NO	1

$$\text{Gini} = 1 - \frac{1^2}{1^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{3}{4} \cdot \frac{4}{9} + \frac{1}{4} \cdot 0 = \frac{1}{3}\\ &= 0.50.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= \frac{3}{8} - \frac{1}{3} = \frac{1}{24} = 0.041\bar{6}\\ &= 0.00.\end{aligned}$$

– Split a_2:

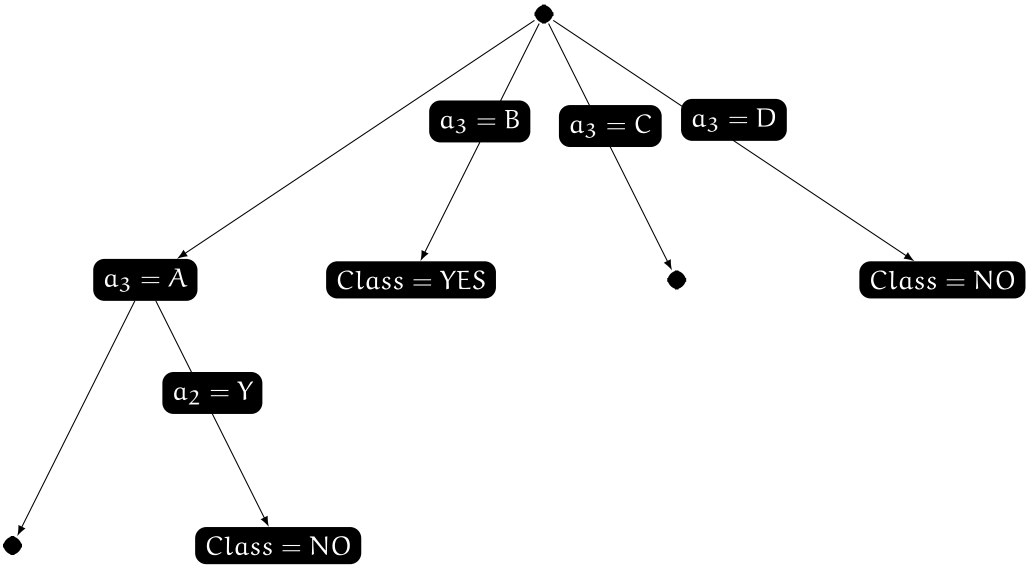
$a_2 = X$	
YES	1
NO	1

$$\text{Gini} = 1 - \frac{1^2}{2^2} - \frac{1^2}{2^2} = \frac{1}{2} = 0.50.$$

$a_1 = Y$	
YES	0
NO	2

$$\text{Gini} = 1 - \frac{2^2}{2^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{2}{4} \cdot \frac{1}{2} + \frac{2}{4} \cdot 0 = \frac{1}{4}\\ &= 0.50.\end{aligned}$$



$$\begin{aligned}\text{Gain} &= \frac{3}{8} - \frac{1}{4} = \frac{1}{8} = 0.125\\ &= 0.00.\end{aligned}$$

Split on a_2.
• Split on $a_3 = C$:
Gini at root node:

$a_2 = X$	
YES	3
NO	1

$$\text{Gini} = 1 - \frac{3^2}{4^2} - \frac{1^2}{4^2} = \frac{3}{8} = 0.375.$$

– Split a_1:

$a_1 = M$	
YES	2
NO	0

$$\text{Gini} = 1 - \frac{2^2}{2^2} = 0.00.$$

$a_1 = F$	
YES	1
NO	1

$$\text{Gini} = 1 - \frac{1^2}{2^2} - \frac{1^2}{2^2} = \frac{1}{2} = 0.50.$$

$$\begin{aligned}\text{Gini} &= \frac{2}{4} \cdot 0 + \frac{2}{4} \cdot \frac{1}{2} = \frac{1}{4}\\ &= 0.25.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= \frac{3}{8} - \frac{1}{4} = \frac{1}{8} = 0.125\\ &= 0.00.\end{aligned}$$

– Split a_2:

$a_2 = X$	
YES	1
NO	1

$$\text{Gini} = 1 - \frac{1^2}{2^2} - \frac{1^2}{2^2} = \frac{1}{2} = 0.50.$$

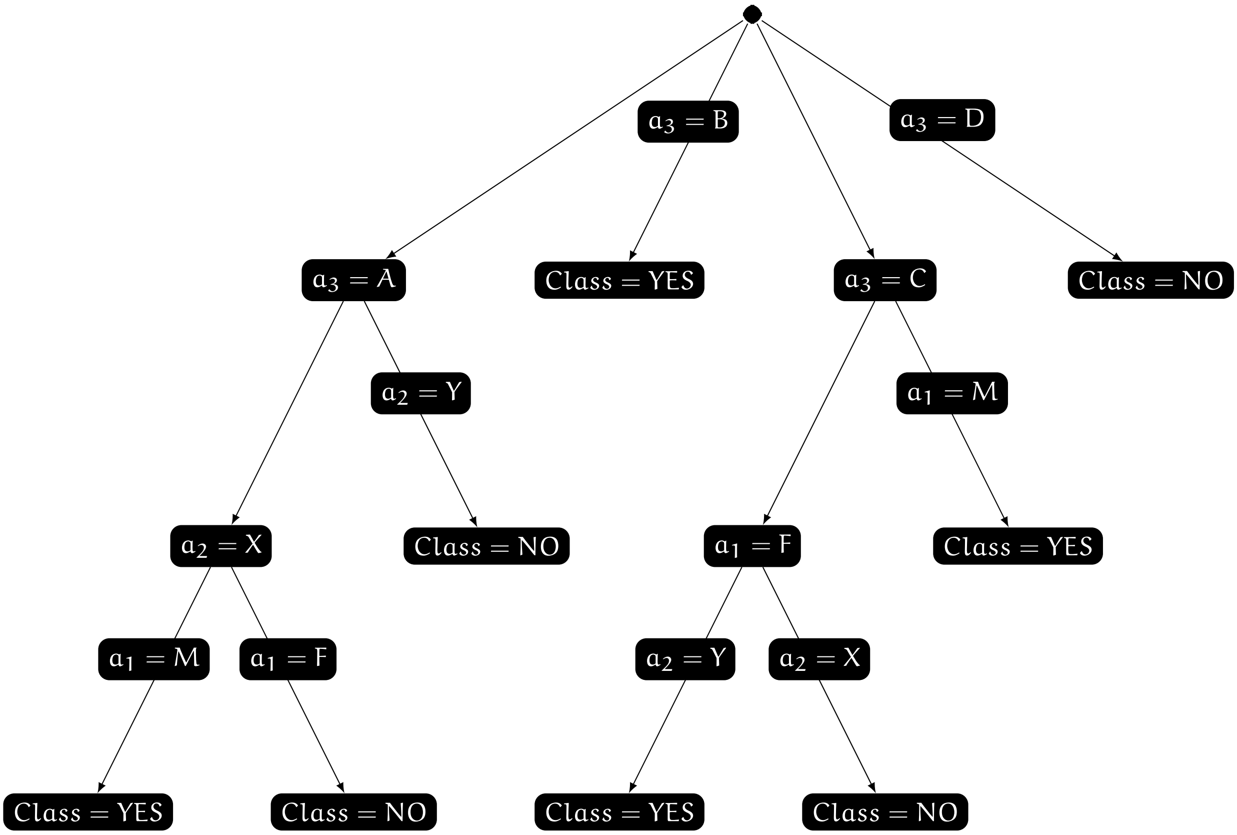
$a_1 = Y$	
YES	2
NO	0

$$\text{Gini} = 1 - \frac{2^2}{2^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{2}{4} \cdot \frac{1}{2} + \frac{2}{4} \cdot 0 = \frac{1}{4}\\ &= 0.50.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= \frac{3}{8} - \frac{1}{4} = \frac{1}{8} = 0.125\\ &= 0.00.\end{aligned}$$

Split on a_1 or a_2.
Final decision tree:



$a_3 = B$
$a_3 = D$
$a_3 = A$
Class = YES
$a_3 = C$
Class = NO
$a_2 = Y$
$a_1 = M$
$a_2 = X$
Class = NO
$a_1 = F$
Class = YES
$a_1 = M$
$a_1 = F$
$a_2 = Y$
$a_2 = X$
Class = YES
Class = NO
Class = YES
Class = NO