

BUSINESS INTELLIGENT DATA MINING

MINH HOẠ ỨNG DỤNG

TB: Hồ Thị Hoàng Vy

GIỚI THIỆU

walmart

- Diapers and Beer? "What is the relationship?"

Amazon



<http://www.sqlservercentral.com/articles/Microsoft+Association+algorithm/101807/>

SSAS-BI

Bốn bước để thực hiện khai phá dữ liệu với SSAS:

1. Define what we want to achieve
2. Prepare the data
3. Build the mining models
4. Deploy and maintain the models in production

BƯỚC 1: ĐỊNH NGHĨA YÊU CẦU

Nêu các câu hỏi mà chúng ta sẽ cố gắng trả lời:

Ví dụ:

Có sự tương quan giữa việc bán sản phẩm music, film, và audio book product types với sở thích khách hàng hay nghề nghiệp của khách hàng không?

BƯỚC 2: CHUẨN BỊ DỮ LIỆU

Trong bước này phải đảm bảo:

- Dữ liệu liên quan
- Kiểm tra chất lượng dữ liệu
 - → đã được làm sạch và sẵn sàng trong KDL
 - Nhưng để trả lời câu hỏi B1 thì cần chuẩn bị dữ liệu ra thành từng bảng chứa các dữ liệu về việc khách hàng mua music, film, audio book

BƯỚC 3: XÂY DỰNG MÔ HÌNH KHAI PHÁ

Bước này, sẽ tạo cấu trúc khai phá dữ liệu bao gồm các data mining models

1. Data mining models có thể dung các thuật toán khác nhau (mining algorithms)
2. Mining models có thể được xây dựng từ relational sources (dữ liệu quan hệ) hoặc từ OLAP cubes
3. Process mô hình và kiểm chứng việc thực thi

BƯỚC 4: TRIỂN KHAI VÀ BẢO TRÌ MÔ HÌNH

Sử dụng các mô hình để dự đoán (prediction) | sử dụng DMX (ngôn ngữ khai phá dữ liệu của SQLserver) hoặc sử dụng Prediction query builder

- 1 prediction là 1 sự tiên đoán (forecast | guess) về giá trị tương lai của 1 biến cụ thể
 - ✓ VD: dùng mô hình dự đoán doanh số của sản phẩm music trong quý tới
- Có thể cần phải xử lý mô hình đều đặn để cập nhật dữ liệu mới cho mô hình

MINH HOẠ ÁP DỤNG

(Amadeus Entertainment case study)

MS SQL analysis service

DANH SÁCH BÀI TẬP

Bài tập 0 – Gom cụm và cây quyết định

Bài tập 1 – Luật kết hợp

Bài tập 2 - Time Series Forecasting

Bài tập 3

Bài tập 4 – Cây quyết định

BƯỚC 1 – XÁC ĐỊNH YÊU CẦU

Cần: phân nhóm khách hàng tương đồng → phân tích và dự đoán hành vi của khách hàng của mỗi nhóm

➤ Đặc trưng để phân nhóm: dựa vào lịch sử mua hàng (music, film, and audio book) và thông tin nhân khẩu (interest, gender, and occupation)

❑ **Gom cụm khách hàng → clustering**

❑ **Dựa vào: gender, interest, occupation → dự đoán khách hàng có mua music không? → decision tree**

BƯỚC 2 – CHUẨN BỊ DỮ LIỆU

- ❑ Tạo table sau trong csdl DDS
- ❑ Nhập dữ liệu (thông tin cá nhân, kèm các xác định liệu khách hàng có mua (Y) | không mua (N) các sản phẩm music, film, audio book không
- ❑ Cách nhập liệu: truy vấn dữ liệu từ DDS

```
create table dm_purchase_pattern
( customer_key int
, gender        char(1)
, interest      varchar(30)
, occupation    varchar(50)
, music         char(1)
, film          char(1)
, audio_books   char(1)
, constraint pk_dm_purchase_pattern
primary key clustered (customer_key)
)
go
```

BƯỚC 2 – CHUẨN BỊ DỮ LIỆU

```
insert into dm_purchase_pattern
(customer_key, gender, interest, occupation, music, film, audio_books)
select c.customer_key, c.gender, c.interest1, c.occupation
, case sum(case p.product_category when 'Music' then 1 else 0 end)
  when 0 then 'N' else 'Y' end as music
, case sum(case p.product_category when 'Films' then 1 else 0 end)
  when 0 then 'N' else 'Y' end as films
, case sum(case p.product_category when 'Audio Books' then 1 else 0 end)
  when 0 then 'N' else 'Y' end as audio_books
from fact_product_sales f
join dim_product p on f.product_key = p.product_key
join dim_customer c on f.customer_key = c.customer_key
group by c.customer_key, c.gender, c.interest1, c.occupation
go
```

BƯỚC 2 (TT) – MÔ TẢ DL TRAIN

- Nếu đã tạo datasource và datasource view trong BIDS project → open project add thêm table vừa tạo vào DDS
- Right click mining structure → new mining structure → ...
- Xác định thông tin: input, predict, key



BƯỚC 2(TT) – CONTENT TYPE

Có 8 loại nội dung: discrete, continuous, discretized, key, key sequence, keytime, ordered, và cyclical

- discrete column: chứa 1 tập các giá trị cụ thể ví dụ: city, gender
- continuous column: chứa giá trị của 1 loại cụ thể (doanh thu: dữ liệu số)

BƯỚC 2 (TT)

Table 13-3. *Valid Data Types for Mining Structure Content Types*

Content Types	Date	Double	Long	Text	Boolean
Discrete	Yes	Yes	Yes	Yes	Yes
Continuous	Yes	Yes	Yes	No	No
Discretized	Yes	Yes	Yes	No	No
Key	Yes	Yes	Yes	Yes	No
Key Sequence	Yes	Yes	Yes	Yes	No
Key Time	Yes	Yes	Yes	No	No
Ordered	Yes	Yes	Yes	Yes	Yes
Cyclical	Yes	Yes	Yes	Yes	Yes

Data Mining Wizard

Specify Columns' Content and Data Type
Specify mining structure columns' content and data type.

Mining model structure:

Columns	Content Type	Data Type
Customer Key	Key	Long
Gender	Discrete	Text
Interest	Discrete	Text
Music	Discrete	Text
Occupation	Discrete	Text

Detect continuous or discrete for numeric columns:

< Back Next > Finish >> | Cancel

BƯỚC 3 – CHỌN MÔ HÌNH

Sử dụng thuật toán: decision tree → Chọn “Allow drill through” để khám phá dữ liệu

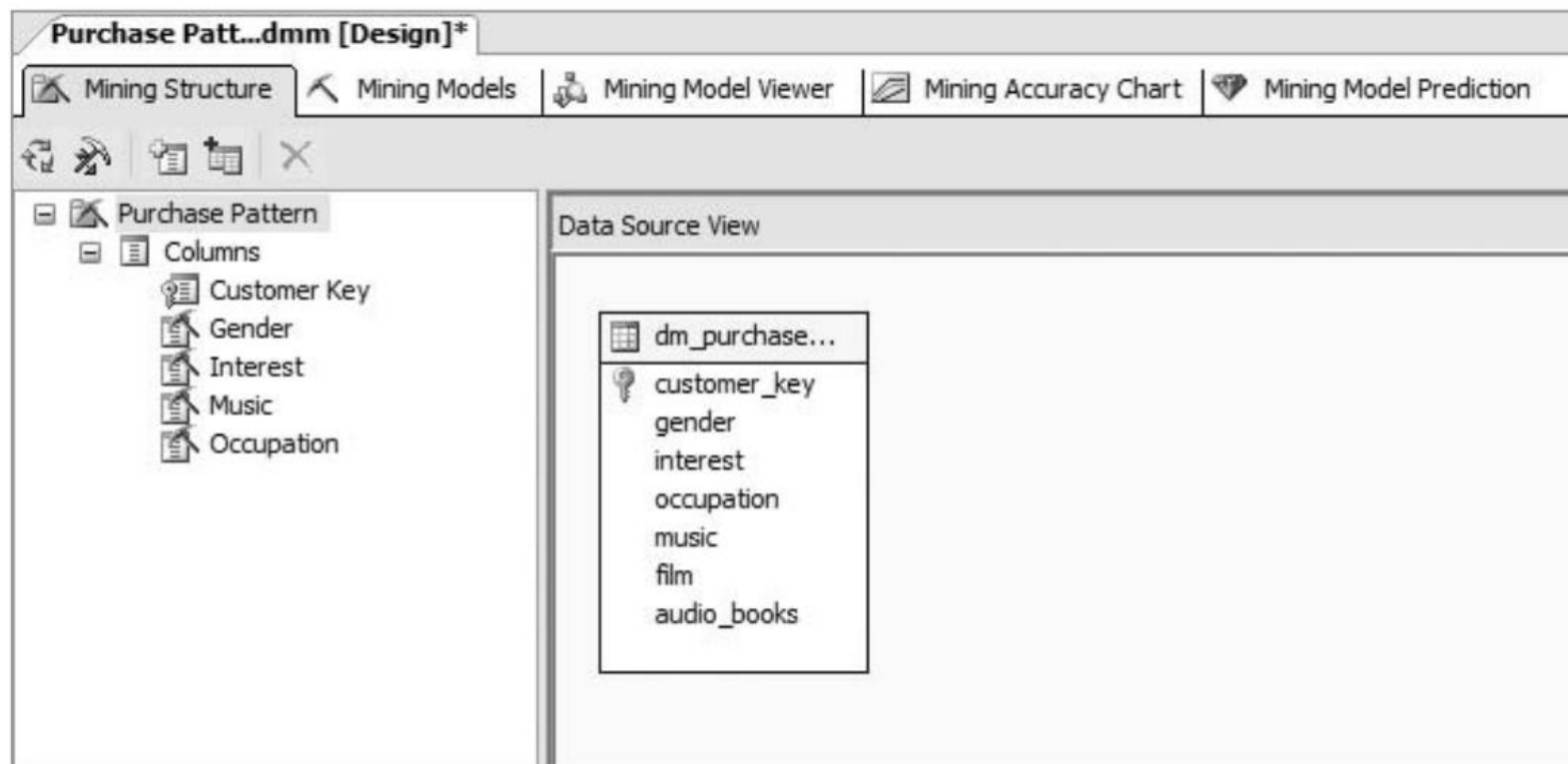


Figure 13-6. *Mining Structure designer*

BƯỚC 3 – CHỌN MÔ HÌNH

Tạo thêm một mô hình khác: clustering → gom cụm khách hàng

Purchase Patt...dmm [Design]*

Mining Structure | **Mining Models** | Mining Model Viewer | Mining Accuracy Chart | Mining Model Prediction

↺ ↻ ↷ ✕

Structure ▲	Decision Trees	Clustering
	🌳 Microsoft_Decision_Trees	🌀 Microsoft_Clustering
🔑 Customer Key	🔑 Key	🔑 Key
📄 Gender	📄 Input	📄 Input
📄 Interest	📄 Input	📄 Input
📄 Music	📄 PredictOnly	📄 PredictOnly
📄 Occupation	📄 Input	📄 Input

BƯỚC 4.1 – XÂY DỰNG MÔ HÌNH

Sau khi đã tạo các models, tiếp theo sẽ xây dựng và triển khai và xử lý các model

→ right click vào màn hình model vị trí bất kỳ → process, build and deploy models

298 customers

with interest = 'Pop Music', 264 of them (88 percent) purchased music products

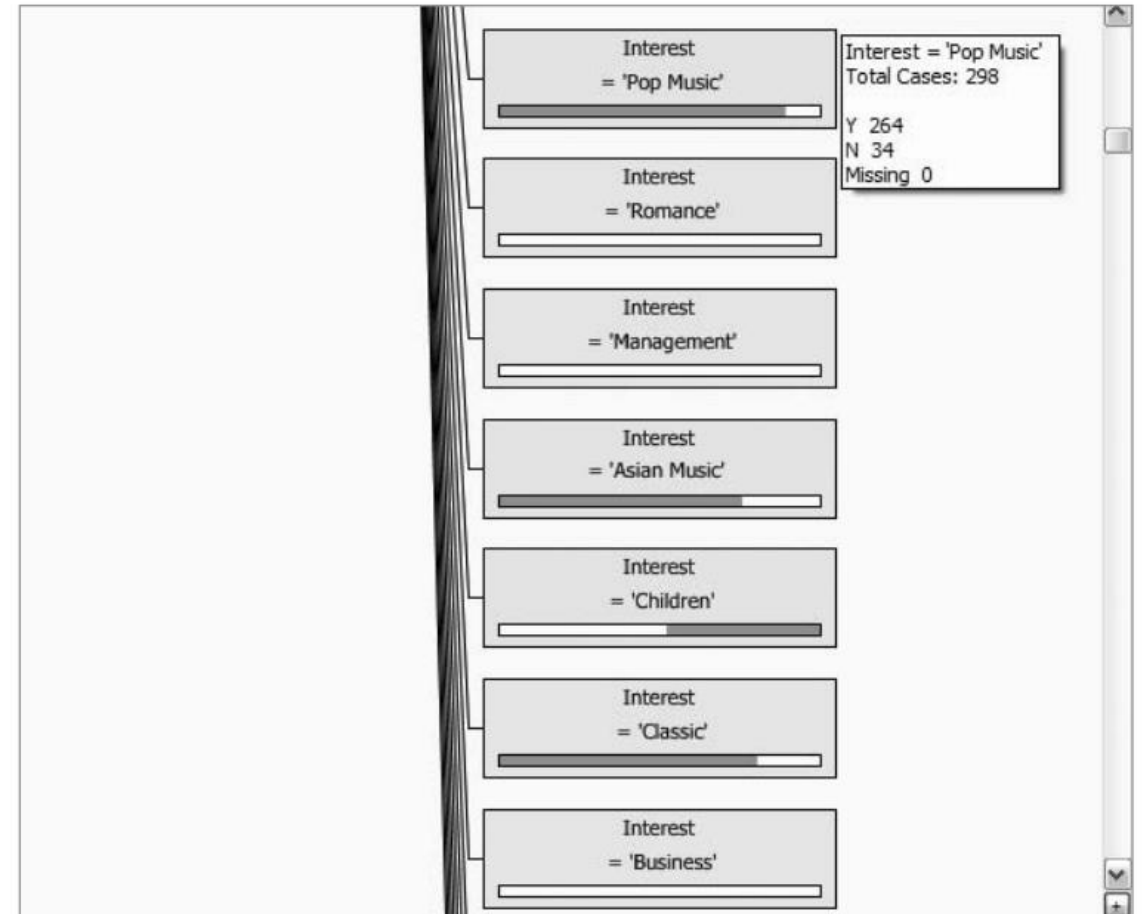


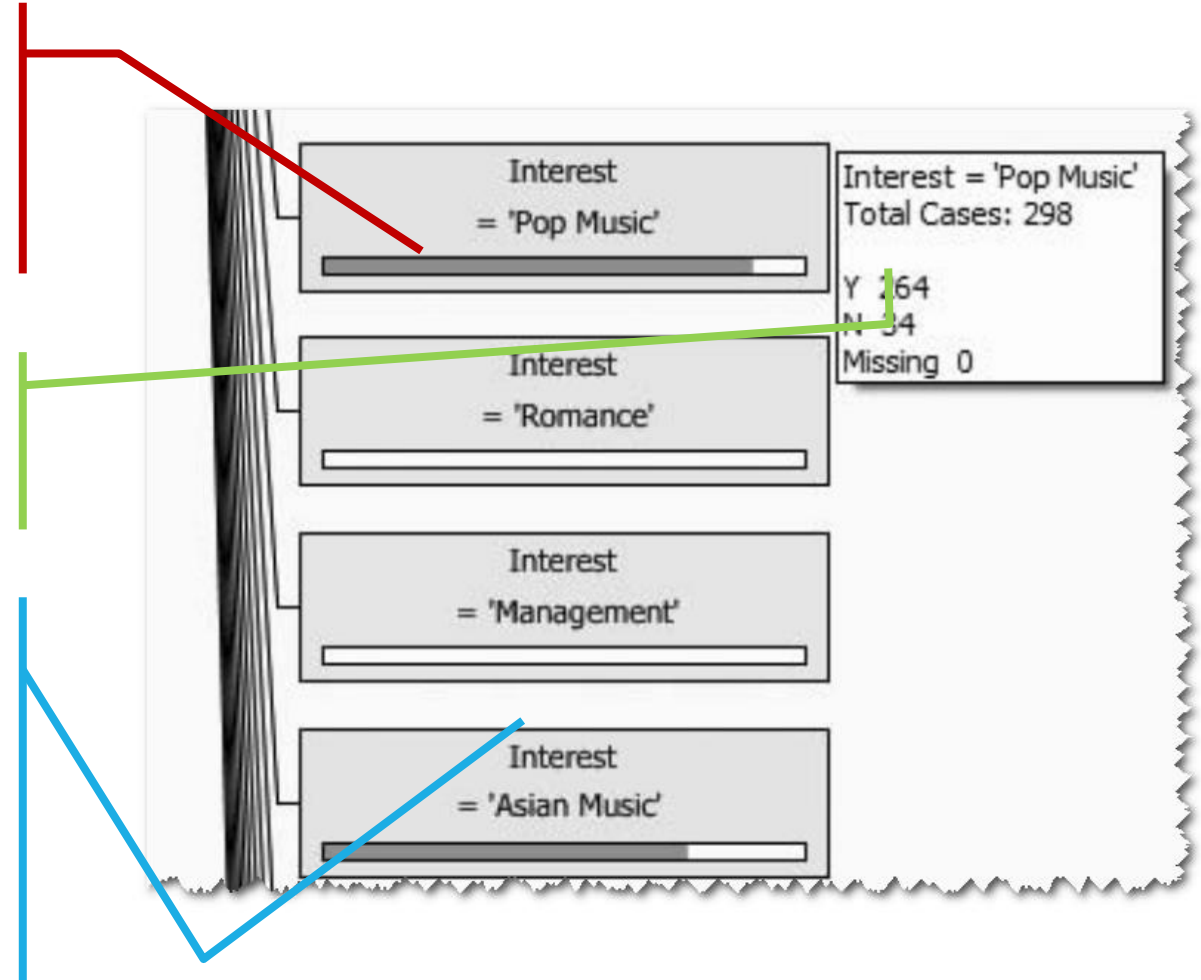
Figure 13-10. Examining the correlation between interest and music purchase

BƯỚC 4.1 – XÂY DỰNG MÔ HÌNH

how many customers in that segment
purchased music products

how many customers in that segment
didn't purchase music products

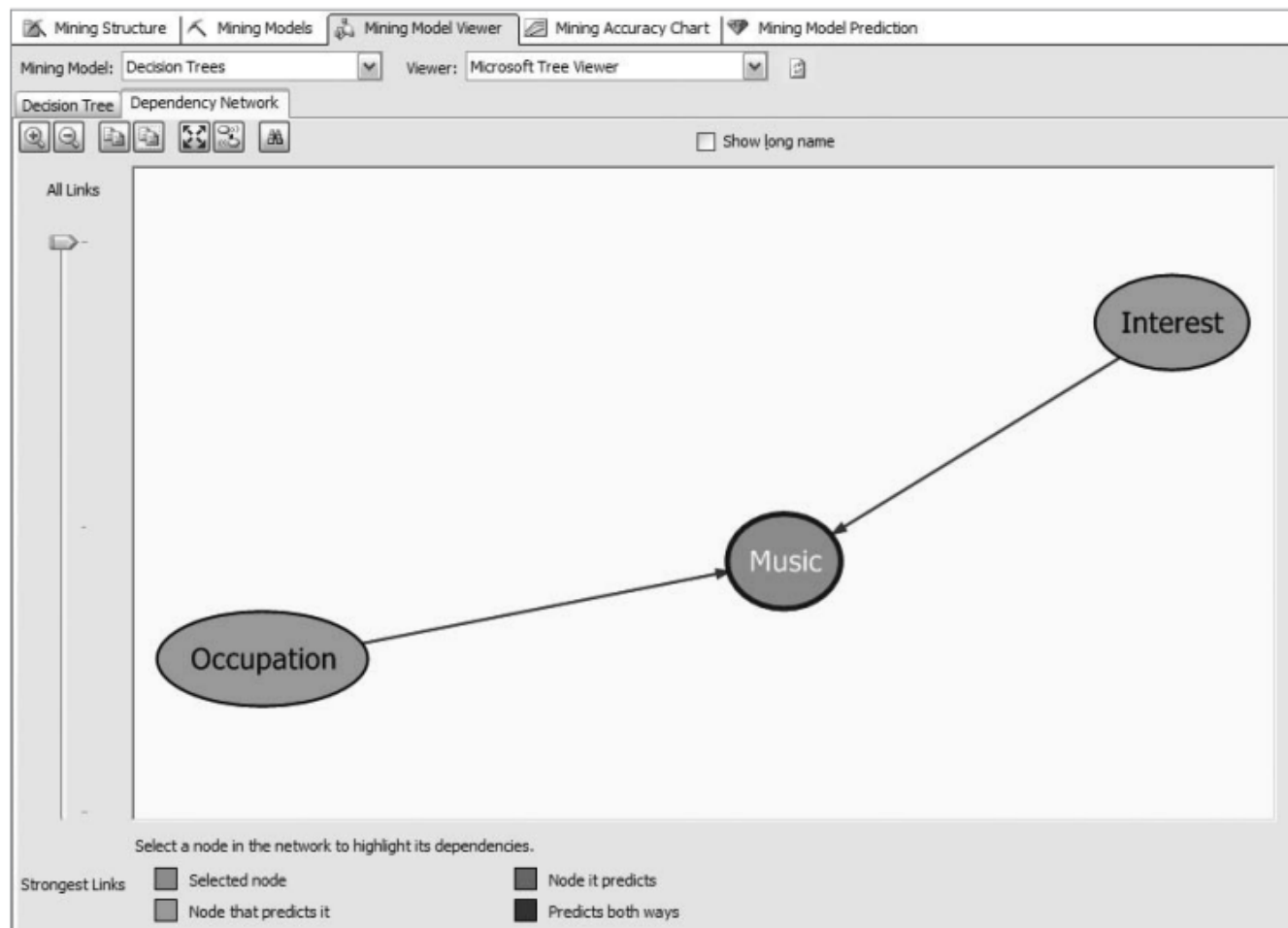
The group of customers who Are
interested in **management**, **none of
them purchased** the music products



BƯỚC 4.1 – XÂY DỰNG MÔ HÌNH

Tab dependency network:

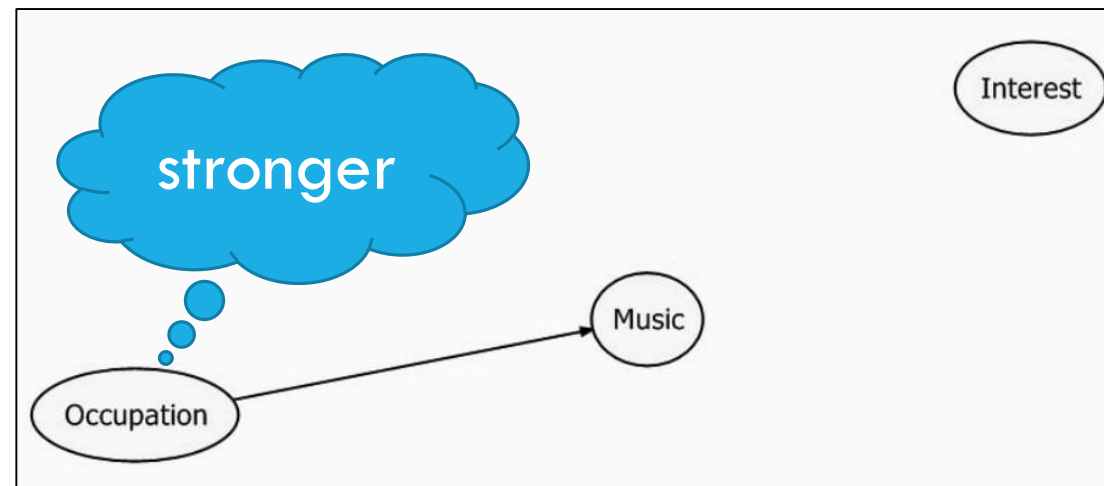
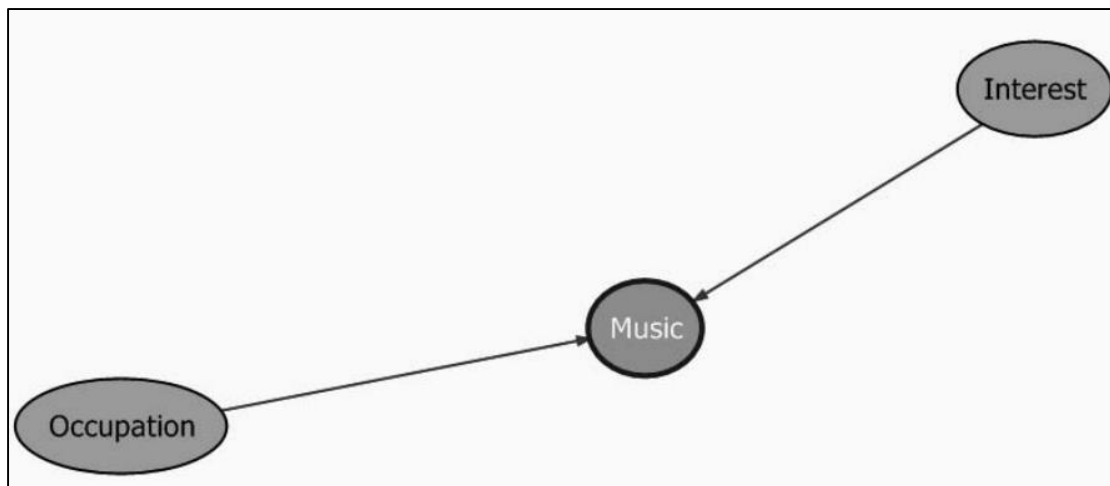
- Thể hiện mối liên hệ giữa các yếu tố nhân khẩu (gender, interest, and occupation) và hành vi “music purchases”
- → độ liên kết mạnh | yếu



BƯỚC 4.1 – XÂY DỰNG MÔ HÌNH

dependency network diagram:

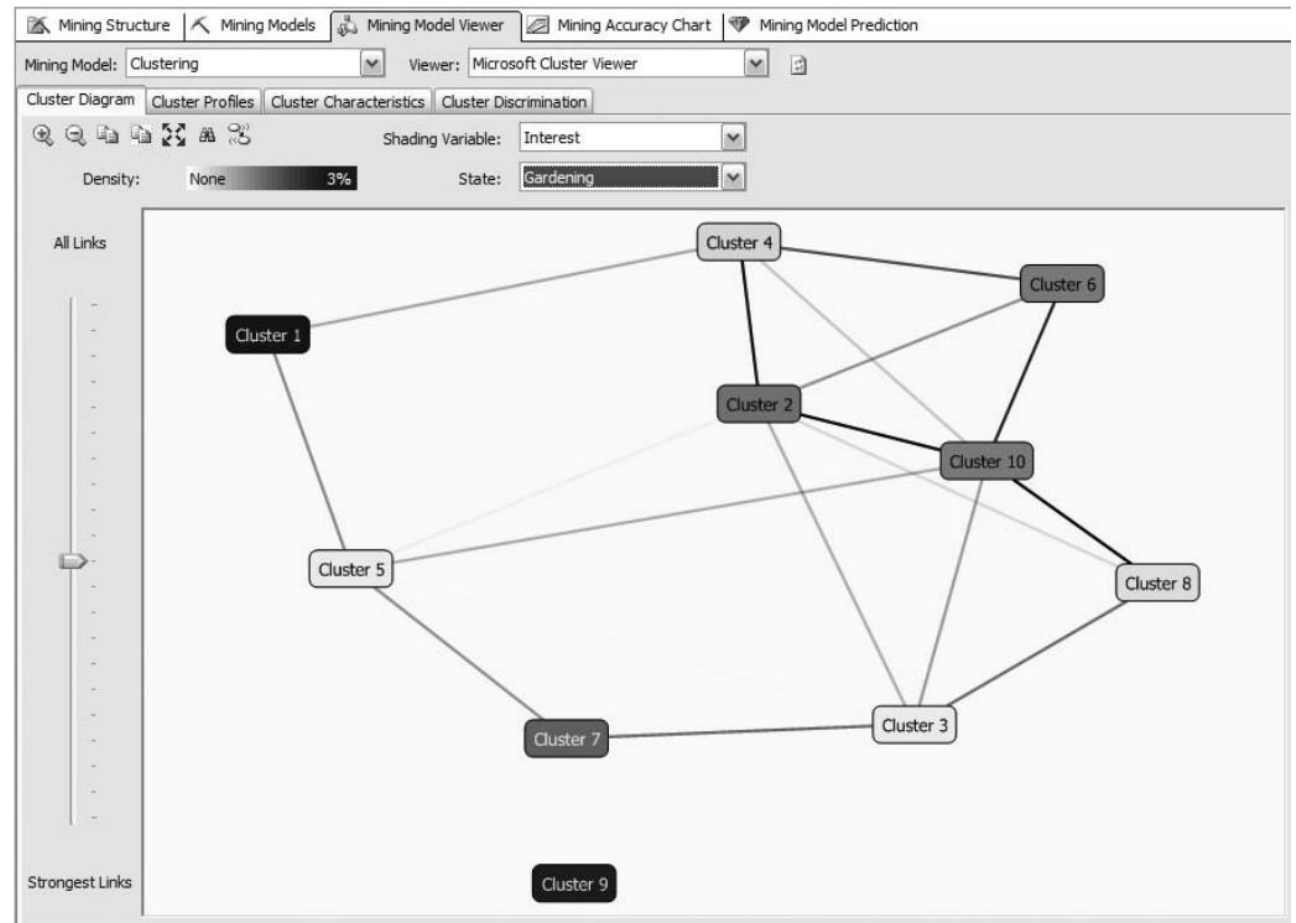
Thể hiện mối liên hệ / phụ thuộc giữa input column và predictable column



demographic factor has stronger influence on the music purchase: occupation or interest

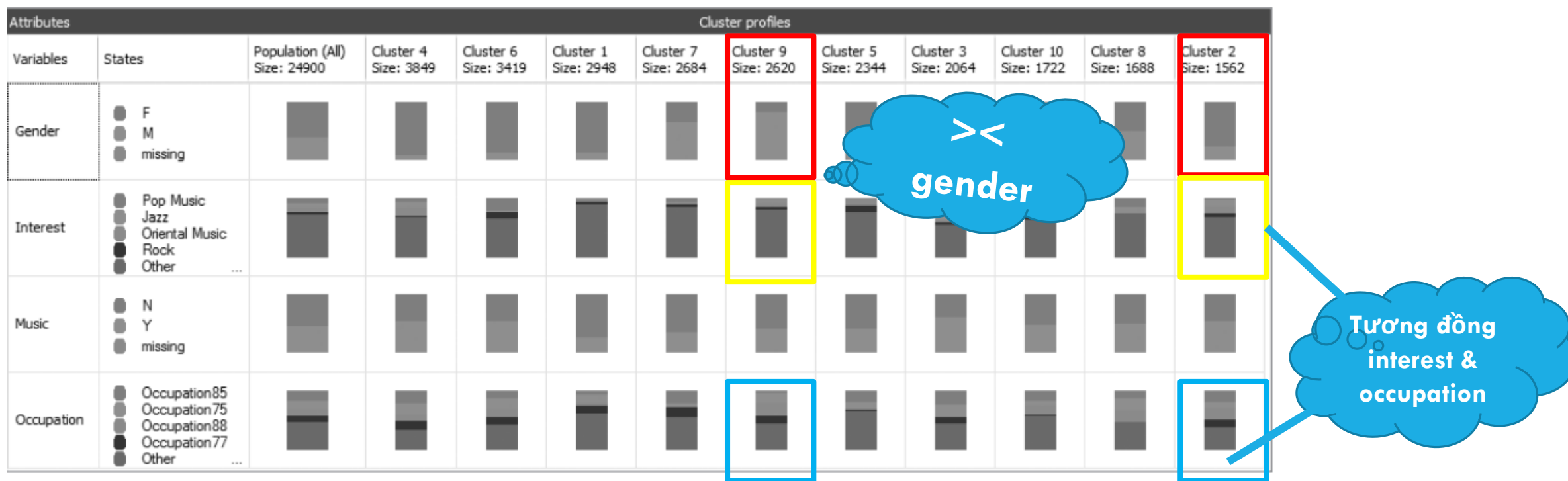
BƯỚC 4.1 – XÂY DỰNG MÔ HÌNH

- Mô hình gom cụm khách hàng
- Để xem chi tiết thông tin cụm → click vào tab thứ hai “**cluster profile**”



BƯỚC 4.1 – XÂY DỰNG MÔ HÌNH

Thông tin cụm → dùng để tìm ra thuộc tính nào ảnh hưởng việc mua sp music?



BƯỚC 4.1 – XÂY DỰNG MÔ HÌNH

So sánh hai clusters 2 and 9:

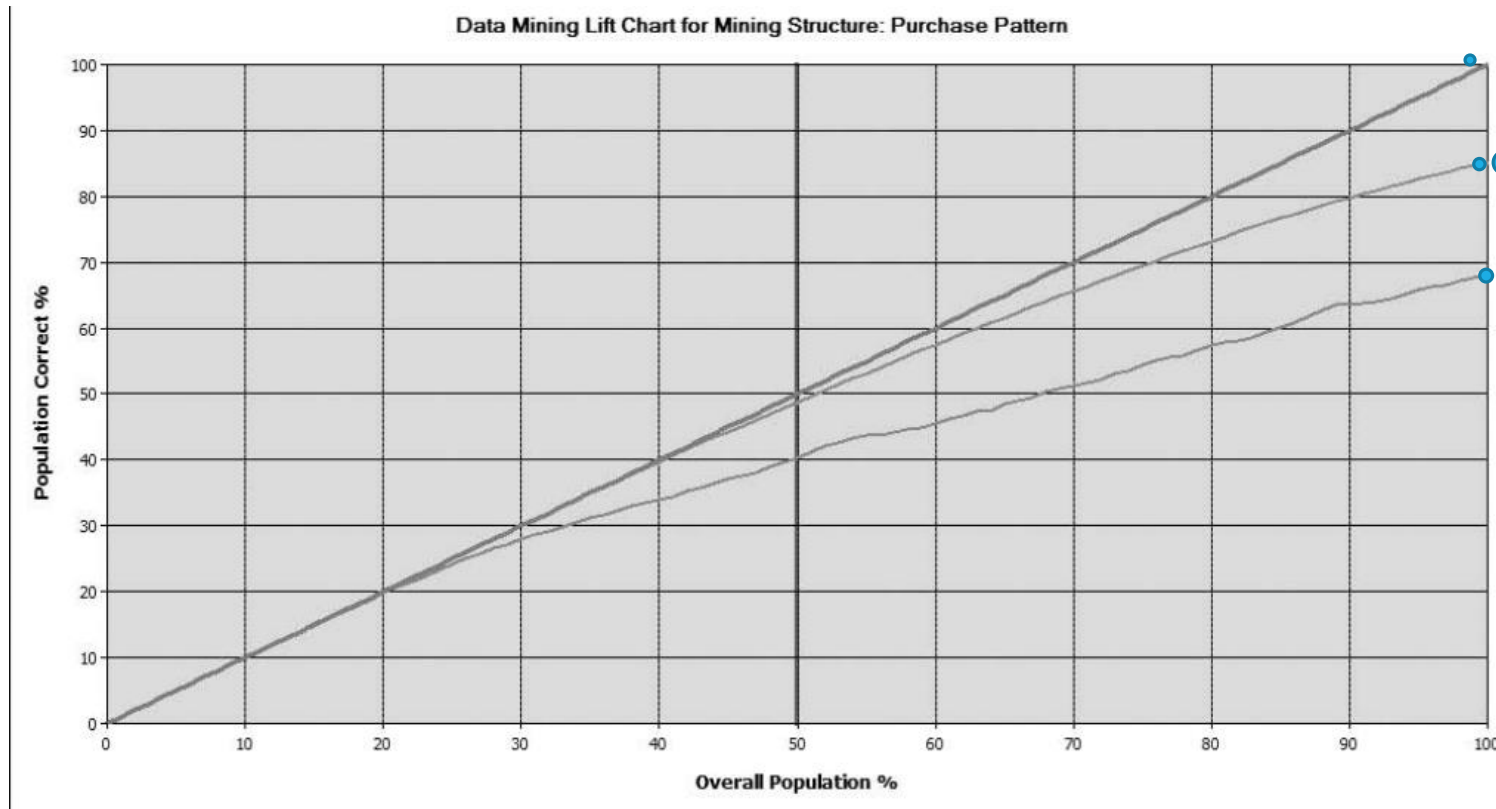
- Có interest, occupations tương đồng
- Rất khác về gender
- Nhưng lại cùng tương đồng việc mua sp music → gender không ảnh hưởng việc mua sp music

ĐÁNH GIÁ

Kiểm tra độ chính xác của mô hình - how well the models perform the predictions

- Mining Accuracy Chart tab
- Accuracy is the percentage of the predictions that are correct

ĐÁNH GIÁ



ideal scenario

Decision tree
(85%)

Clustering
(69%)

the prediction is whether a customer will purchase music products

ĐÁNH GIÁ

1. Tại sao Decision Tree model chính xác hơn Clustering model?
2. Có phải luôn như vậy?
3. Khi nào nên dùng Decision Tree và khi nào dùng Clustering?

BƯỚC 4.2 – TRIỂN KHAI

Sử dụng để dự đoán:

Liệu 1 khách hàng tiềm năng sẽ quan tâm đến việc mua sản phẩm music hay không?

BƯỚC 4.2 – TRIỂN KHAI

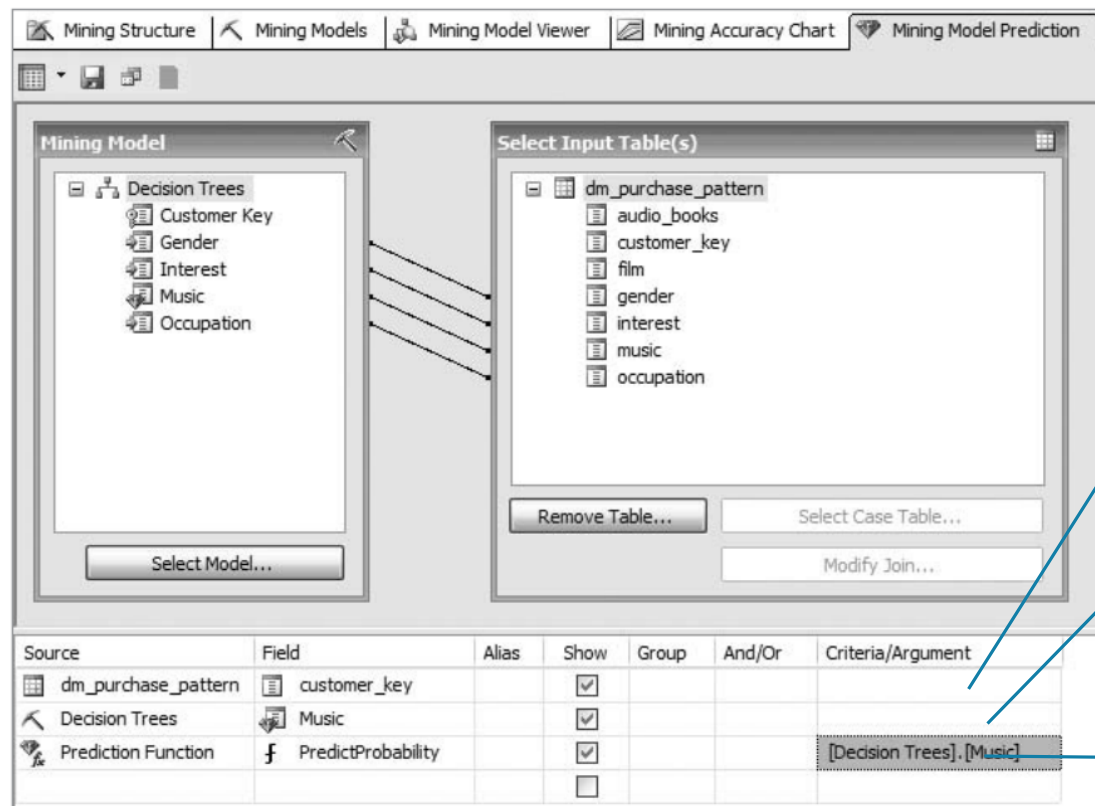


Figure 13-17. Using the model to predict future values

Mining Model Prediction tab

→ click Select Case Table

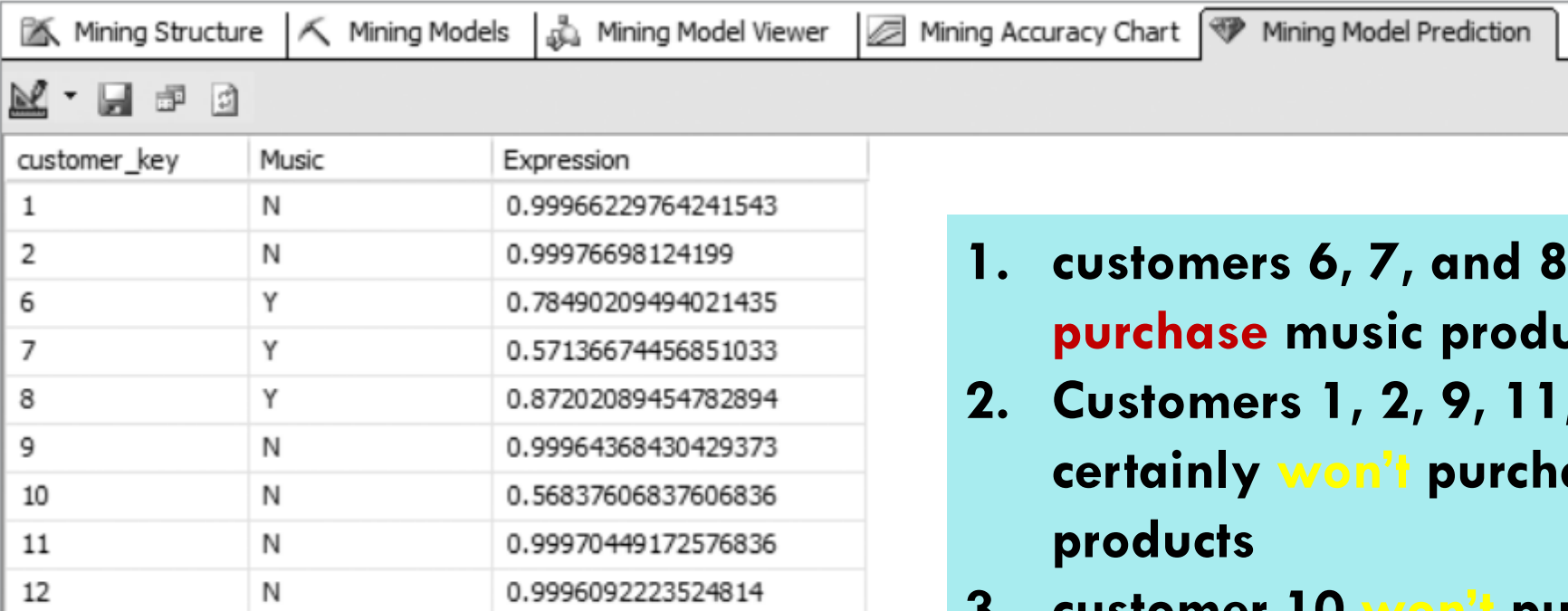
→ choose dm_purchase_pattern

to display the customer_keycolumn

to display music purchases from the Decision Trees model.

to display the probability of the music purchase prediction using the PredictProbability function

BƯỚC 4.2 – TRIỂN KHAI



The screenshot shows a software window titled 'Mining Model Prediction'. It contains a table with three columns: 'customer_key', 'Music', and 'Expression'. The table lists predictions for 12 customers. Customers 6, 7, and 8 are predicted to purchase music (Music=Y), while others are predicted not to (Music=N). The 'Expression' column contains numerical values representing the model's confidence or score for each prediction.

customer_key	Music	Expression
1	N	0.99966229764241543
2	N	0.99976698124199
6	Y	0.78490209494021435
7	Y	0.57136674456851033
8	Y	0.87202089454782894
9	N	0.99964368430429373
10	N	0.56837606837606836
11	N	0.99970449172576836
12	N	0.9996092223524814

1. customers 6, 7, and 8 are **likely to purchase** music products
2. Customers 1, 2, 9, 11, and 12 almost certainly **won't** purchase music products
3. customer 10 **won't** purchase music either, but it's **not so certain**

Figure 13-18. *The prediction output*

BÀI TẬP 4 - NHÓM

Yêu cầu: hãy dự đoán liệu 1 customer sẽ mua sản phẩm “bike” không?

- ☐ Xác định kỹ thuật
- ☐ Xác định dữ liệu input, minh họa

BÀI TẬP - 4 NHÓM

☐ **Kỹ thuật:** cây quyết định

- Liệu person X có thể là người mua sản phẩm “bike”?

☐ **Làm sao xác định được dữ liệu input cho mô hình khai thác?**

- Salary?
- Number of cars?
- Email?
- Address?
-

BÀI TẬP 4 - NHÓM

Có nhiều dữ liệu input có thể hữu ích cho việc dự đoán liệu 1 khách hàng có mua “bike” không. Vậy làm thế nào xác định được column nào là quan trọng và ngược lại?

- Không có tiền / thu nhập tốt thì có thể mua không?.
- Nếu đã có 5 xe hơi thì có mua thêm bike không?

Có nhiều column input data chúng ta có thể loại bỏ. Tuy nhiên, mô hình mining cho phép chúng ta chọn lựa đặc trưng nào ảnh hưởng và không

- email có ảnh hưởng việc mua không? Liệu gmail và Hotmail có ảnh hưởng gì quyết định mua hay không?

BÀI TẬP 4 - NHÓM

Input có thể là: The number of Children, Yearly Income, Region, occupation, number car owned, education, age, commute distance, marital status, house owned flag

→ sử dụng: **dbo.Prospectivebuyers** table

BÀI TẬP 1

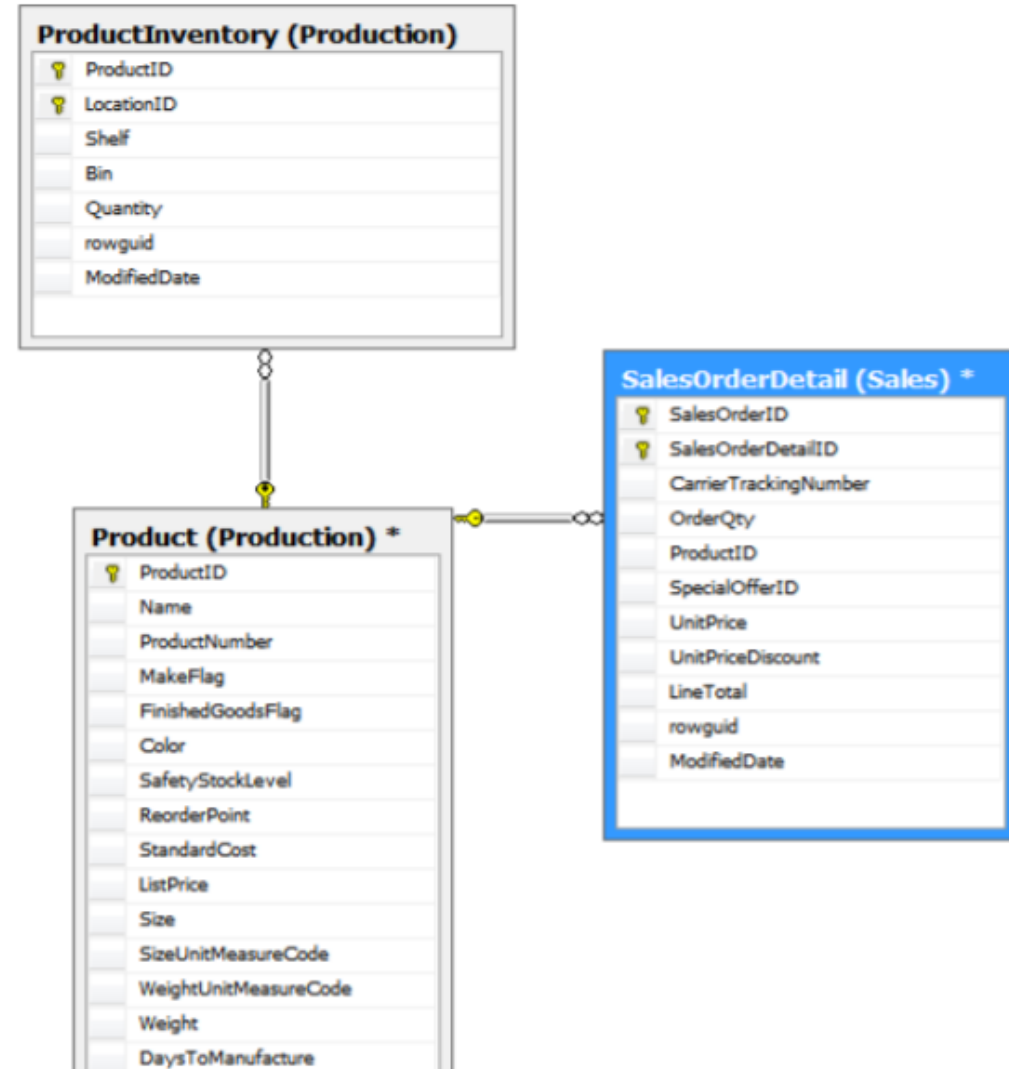
Tình huống: cần gợi ý sản phẩm có số lượng tồn kho cao khi khách hàng mua một sản phẩm liên quan (product with high inventory)

1. Các thông tin cần phân tích? Các bảng liên quan
2. Xác định kỹ thuật và dữ liệu input?

BÀI TẬP 1

1. các thông tin cần phân tích cho nhu cầu trên:

- product information,
- product inventory information,
- product sales order information



BÀI TẬP 2

Hãy xây dựng 1 mô hình dự đoán doanh số bán hàng (sale ammounts) cho 1 loại sản phẩm (product category) cụ thể trong các tháng tới.

BÀI TẬP 2

Thông tin liên quan?

Mô tả dữ liệu liên quan phục vụ cho mining từ source view?

Mô tả dữ liệu input?

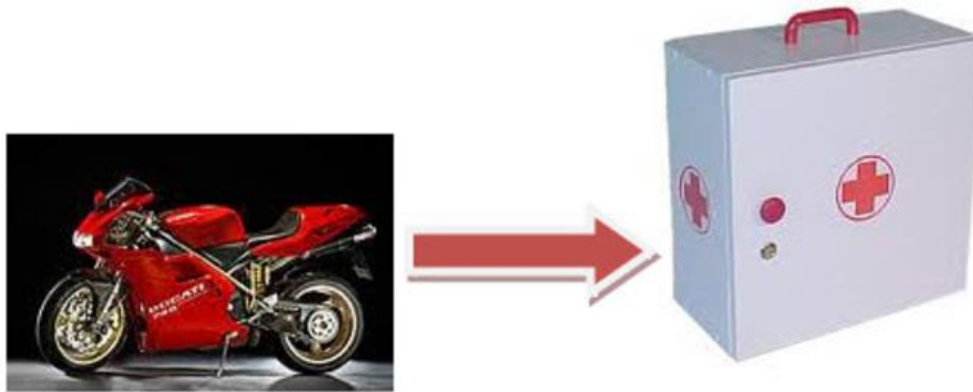
BÀI TẬP 3

Phân tích thú tự của sản phẩm và tối ưu vị trí của các sản phẩm nhằm tăng doanh số bán hàng.

- ☐ Hãy xác định dữ liệu cần cho phân tích → các table liên quan ?
- ☐ Hãy chọn kỹ thuật thực hiện
- ☐ Hãy cho dữ liệu minh họa đầu vào (input và predict data)

BÀI TẬP 5

Dự đoán nhu cầu của khách hàng dựa vào thông tin mua hàng của khách hàng. Từ đó giúp đề nghị thêm sản phẩm, giúp việc bán hàng đạt hiệu quả hơn.

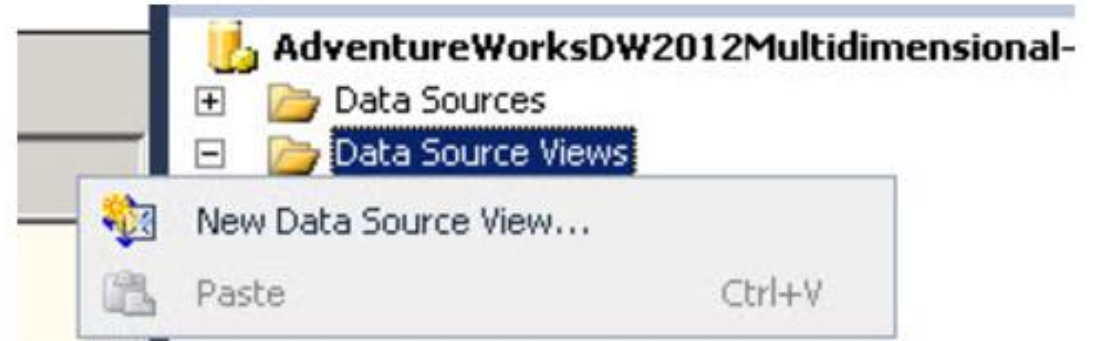


<http://www.sqlservercentral.com/articles/Microsoft+Association+algorithm/101807/>

BÀI TẬP 5

Chọn dữ liệu → data source view

1. vAsscSeqLineItems
2. vAssocSeqOrders



BÀI TẬP 5

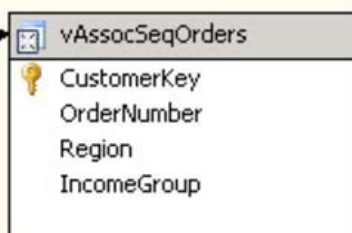
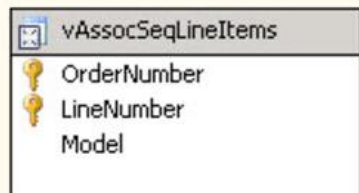
Dữ liệu sử dụng để mining

vAssocSecOrder

Table			
OrderNumber	CustomerKey	Region	IncomeGroup
SO51176	18239	Pacific	High
SO51177	27873	Pacific	Low
SO51178	11245	Europe	High
SO51179	22430	Europe	Moderate
SO51180	16313	Europe	Low
SO51181	12132	Europe	High
SO51182	22998		
SO51183	20662		
SO51184	11263		
SO51185	27767		
SO51186	24339		
SO51187	22261		

vAssocLineItems

Table		
OrderNumber	LineNumber	Model
SO70821	1	Touring-3000
SO70821	2	Water Bottle
SO70821	3	Road Bottle Cage
SO70821	4	Hydration Pack
SO59831	1	Mountain-400-W
		Cycling Cap
		Women's Mountain Shorts
		Mountain Bottle Cage
		Water Bottle
		LL Mountain Tire
		LL Mountain Tire



BÀI TẬP 5

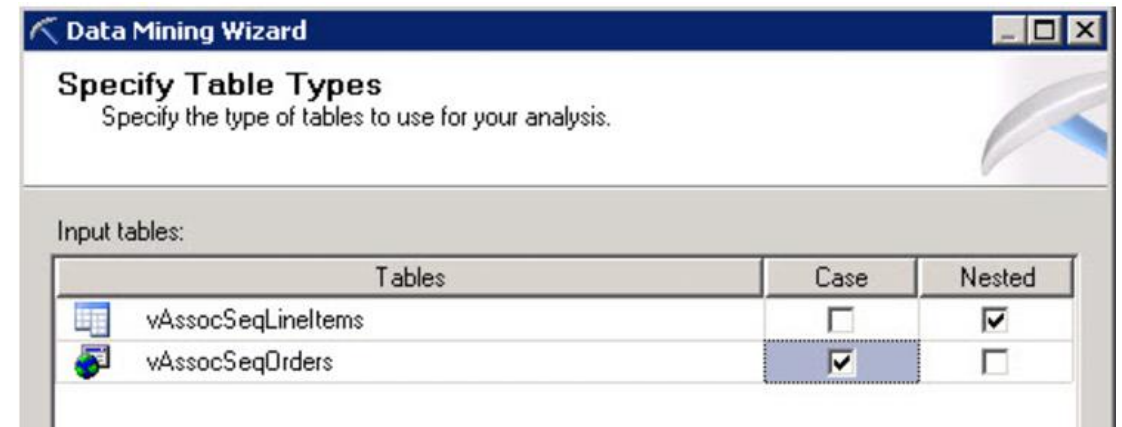
Xây dựng mô hình với kỹ thuật luật kết hợp

vAssocSeqOrders chứa thông tin Orders

vAssocSeqLineItems chứa nhiều dòng thông tin của các Order này

➤ vAssocSeqLineItems is nested

➤ vAssocSeqOrders is a Case



BÀI TẬP 5

Làm tiếp và phân tích kết quả

