การประยุกต์ใช้ดาต้าไมนิ่งในทางธุรกิจ

DATA MINING APPLICATION FOR BUSINESS

นายณัฐพงษ์ สวิบูลย์

นายอินทกะ พิริยะกุล

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2546

ปริญญานิพนธ์ ปีการศึกษา 2546

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง การประยุกต์ใช้ดาต้าไมนิ่งในทางธุรกิจ

DATA MINING APPLICATION FOR BUSINESS

คณะผู้จัดทำ นายณัฐพงษ์ สววิบูลย์ รหัส 43010131

นายอินทกะ พิริยะกุล รหัส 43010547

..... อาจารย์ที่ปรึกษา

(รศ. ประทีป บัญญัตินพรัตน์)

การประยุกต์ใช้ดาต้าไมนิ่งในทางธุรกิจ

นายณัฐพงษ์ สววิบูลย์ 43010131
นายอินทกะ พิริยะกุล 43010547
รศ. ประทีป บัญญัตินพรัตน์ อาจารย์ที่ปรึกษา
ปีการศึกษา 2546

บทคัดย่อ

โครงงานนี้เป็นการพัฒนาโปรแกรมโดยการนำเอาเทคนิคของดาต้าไมนิ่งมาใช้ในการมาวิเคราะห์ปัญหาทางธุรกิจในลักษณะต่างๆ ด้วยเหตุเนื่องมาจากว่าการเก็บข้อมูลของธุรกิจต่างๆนั้น มีขนาดใหญ่ การสืบค้นและเสาะหาประโยชน์จากฐานข้อมูลเหล่านั้นอาจทำได้ลำบาก การนำเทคนิคของดาต้าไมนิ่งมาใช้จะทำให้ช่วยทำให้สามารถดึงส่วนสำคัญที่เราจะนำไปใช้ในการวางแผนทางธุรกิจได้ โดยความสามารถของโปรแกรมที่จัดทำเช่นเดียวกับส่วนหนึ่งของโปรแกรมที่วางขายกันในท้องตลาดปัจจุบัน โดยใช้ฐานข้อมูลใน 3 ลักษณะงาน คือการทำนายค่า, การวาดแผนภาพต้นไม้ และการวิเคราะห์ตะกร้าตลาด

DATA MINING APLLICATION FOR BUSINESS

Mr.Nattapong Savapibool

Mr.Intaka Piriyakul

Assoc.Prof. Prateep Banyatnopparat Advisor

Academic Year 2003

ABSTRACT

This thesis is to develop an application by using data mining techniques for analyze business problem. It's hard to looking for information in the huge database. By using data mining techniques, the solution is possible. We implement an application in the 3 tasks contained prediction task, classification task, and association analysis task.

สารบัญ

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1. ความสำคัญและเป็นมา

เนื่องจากปัจจุบัน การเก็บข้อมูลในฐานข้อมูลจะมีขนาดใหญ่มาก อันเป็นผลสืบเนื่องมาจากสื่อบันทึกข้อมูลขนาดใหญ่ที่มีราคาถูกลง (Inexpensive of Mass Storage) รวมทั้งสมรรถนะของระบบเครื่องที่มีหน่วยความจำมาก รวมทั้งความเร็วสูงในการประมวลผลและมีการเชื่อมโยงในลักษณะของเครือข่าย (Network) จึงเป็นการเปิดโอกาสให้มีการสืบค้นและเสาะหาประโยชน์จากฐานข้อมูลขนาดใหญ่ให้ได้มากขึ้น แนวทางหนึ่งในการพัฒนาทางด้านนี้คือการพัฒนาทางดาต้าไมนิ่ง (Data Mining :Knowledge Discovery in Database :KDD) ซึ่งมีเจตนาที่จะเสาะหาความรู้และรูปแบบพฤติกรรมจากข้อมูลดิบที่สะสมอยู่ในฐานข้อมูล (Unknown Pattern in Database : Hidden Predictive Information) โดยที่ผลประโยชน์ที่ได้นั้นจะนำไปส่งเสริมในส่วนของระบบสนับสนุนการตัดสินใจ (Decision Support System) ในแต่ละสาขา โดยเฉพาะอย่างยิ่งงานในทางธุรกิจ ซึ่งจำเป็นจะต้องสู้กับคู่แข่งขันทางการค้า (Competitor) ในตัวของสารสนเทศเองนั้นลักษณะจะคล้ายกับ ทรัพยากรในเหมืองแร่ที่เราจะขุดออกมาเพื่อใช้งาน ทั้งนี้เพราะการดำเนินงานในส่วนที่เป็นการประมวลผลธุรกรรม (Transaction Processing) แต่พียงอย่างเดียวนั้นไม่เพียงพอในการนำไปใช้งาน เพราะการประมวลผลธุรกรรมเปรียบเสมือนการนำแค่เปลือกนอกของสารสนเทศไปใช้เท่านั้น โดยไม่สามารถดึงส่วนที่สำคัญที่อยู่ภายในออกมาใช้ประโยชน์ได้

2. ขอบเขตของการศึกษาและพัฒนา

โครงงานนี้กำหนดขอบเขตในการพัฒนาโปรแกรมต้นแบบ จุดประสงค์หลักเพื่อให้ผู้ใช้นำไปใช้ในการประยุกต์ใช้ได้จริงในงานทางธุรกิจในงานด้านต่างๆ เป็นงานพัฒนาโปรแกรมเพื่อวิเคราะห์ฐานข้อมูลเพื่อใช้งานด้านการพาณิชย์ดังเช่นเดียวกับโปรแกรมที่นิยมใช้กัน เช่น SPSS , PolyAnalyst พร้อมทั้งทดสอบกับฐานข้อมูลตัวอย่าง สำหรับการบทประยุกต์ของดาต้าไมนิ่งจะดำเนินโครงงานโดยศึกษาจัดทำเครื่องมือ 3 ประเภท สำหรับ 3 ลักษณะงาน อันประกอบด้วย

- การทำนายค่า โดยใช้สมการถดถอยเชิงเส้นหลายตัวแปร ซึ่งเป็นวิธีการในทางสถิติ สามารถใช้ประยุกต์ในการพยากรณ์ค่าเชิงตัวเลขเพื่อคาดการณ์สิ่งที่เราสนใจ ตัวอย่างเช่น นำไปประเมินตัวเลขโดยการคำนวณจากข้อมูลที่มีอยู่ โดยโครงงานนี้จะใช้ข้อมูลทดสอบมาจากฐานข้อมูล ซึ่งมีจำนวนรถยนต์ทั้งสิ้น 406 คัน โดยการเก็บข้อมูลรายละเอียดของรถยนต์แต่ละคันที่อาจมีผลต่อระยะทางใน

การวิ่งต่อปริมาณการใช้น้ำมัน 1 แกลลอน โดยที่ รูปแบบของการ ตาดเดาระยะทางในการวิ่งของรถ จาก ปัจจัยต่างๆ เช่น ขนาดเครื่องยนต์ จำนวนแรงม้า น้ำหนักรถ อัตราเร่ง ปีที่จดทะเบียน บริษัทที่สร้าง จำนวนลูกสูบ

- การแยกประเภทโดยใช้แผนภาพต้นไม้ สามารถใช้ในการแบ่งกลุ่มสินค้าหรือลูกค้าตามคุณสมบัติต่างๆ โดยสามารถนำไปประยุกต์ใช้ทางธุรกิจได้ ตัวอย่างเช่น แบ่งแยกลูกค้าส่วนกลุ่มเป้าหมายของสินค้า เพื่อช่วยในการวางนโยบายบริหารการตลาด

- การวิเคราะห์ความสัมพันธ์โดยโครงงานนี้จะใช้ข้อมูลทดสอบมาจากฐานข้อมูลธุรกรรมจากร้านขายของชำซึ่งมี จำนวน ทั้งสิ้น 2167 ระเบียนข้อมูล เพื่อนำมาจำแนกว่า ลูกค้ามักจะซื้อสินค้ารายการใดควบกันไปโดยการศึกษาความสัมพันธ์ของการซื้อสินค้า 2 ประเภท – 3 ประเภท 4 ประเภทเรื่อยไป

3. ประโยชน์ที่คาดว่าจะได้รับ

- เรียนรู้และเข้าใจวิธีการและอัลกอริทึมต่างๆในการดาต้าไมนิ่ง
- รู้จักการประยุกต์นำเอาดาต้าไมนิ่งไปใช้ในธุรกิจต่างๆ
- เรียนรู้การพัฒนาโปรแกรมในการคำนวณกับฐานข้อมูลขนาดใหญ่

4. ขั้นตอนการดำเนินงาน

- ศึกษาและทำความเข้าใจกระบวนการทางดาต้าไมนิ่ง
- ศึกษาลักษณะงานที่นำเอาดาต้าไมนิ่งไปประยุกต์ใช้
- การศึกษาเทคนิคของดาต้าไมนิ่งที่เหมาะสมกับงานต่างๆ
- ศึกษาความเป็นไปได้จริงและความเหมาะสมที่จะนำดาต้าไมนิ่งมาประยุกต์ใช้กับธุรกิจทั่วๆไป ประเภทต่างๆ
- กำหนดขอบเขตและความสามารถของโปรแกรมต้นแบบและดำเนินการพัฒนา
- หาฐานข้อมูลเพื่อนำมาทดสอบ

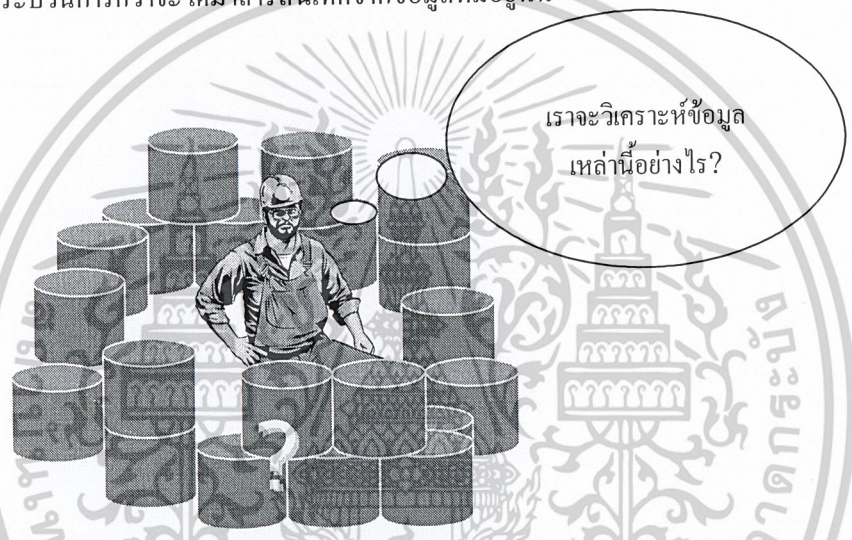
เดือน

กิจกรรม	1	2	3	4	5	6	7	8	9	10	11
ศึกษาทฤษฎี	←	→									
สำรวจข้อมูล			←	→							
จัดเก็บข้อมูล				←	→						
พัฒนา Software						←	→	→			
ทดสอบ									←	→	
เขียนรายงาน										←	→
นำเสนองาน											↔

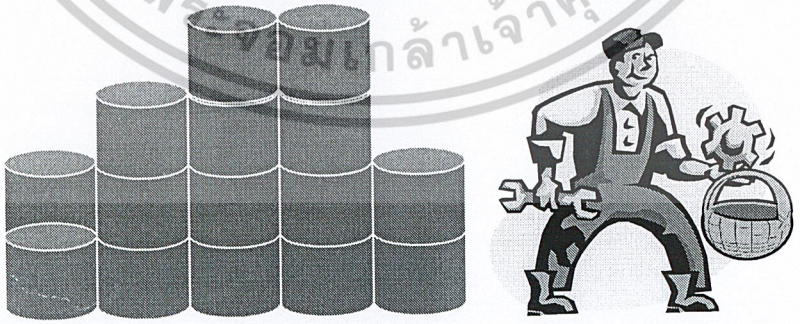
บทที่ 2
ความรู้พื้นฐานเกี่ยวกับดาต้าไมนิ่ง

1. ความหมายของดาต้าไมนิ่ง

ดาต้าไมนิ่ง เป็น กระบวนการดำเนินการกับข้อมูลเพื่อเสาะหาความรู้ที่จะนำไปใช้หรือประโยชน์อื่นใดที่มิได้แสดงโดยตรงในตัวข้อมูลเองนั้น เป็นการสำรวจและการวิเคราะห์ข้อมูลจำนวนมากเพื่อที่จะได้มาซึ่งรูปแบบและกฎที่มีประโยชน์ ดังจะเรียกได้อีกอย่างหนึ่งว่า “เหมืองข้อมูล” ซึ่งดาต้าไมนิ่งนั้นเป็นกระบวนการส่วนหนึ่งของกระบวนการกว่าจะได้มาสารสนเทศจากข้อมูลที่มีอยู่นั้น



รูปที่ 2-1 : การที่มีข้อมูลมากมาย แต่ขาดซึ่งสารสนเทศหรือข่าวสารที่เป็นประโยชน์ไปใช้



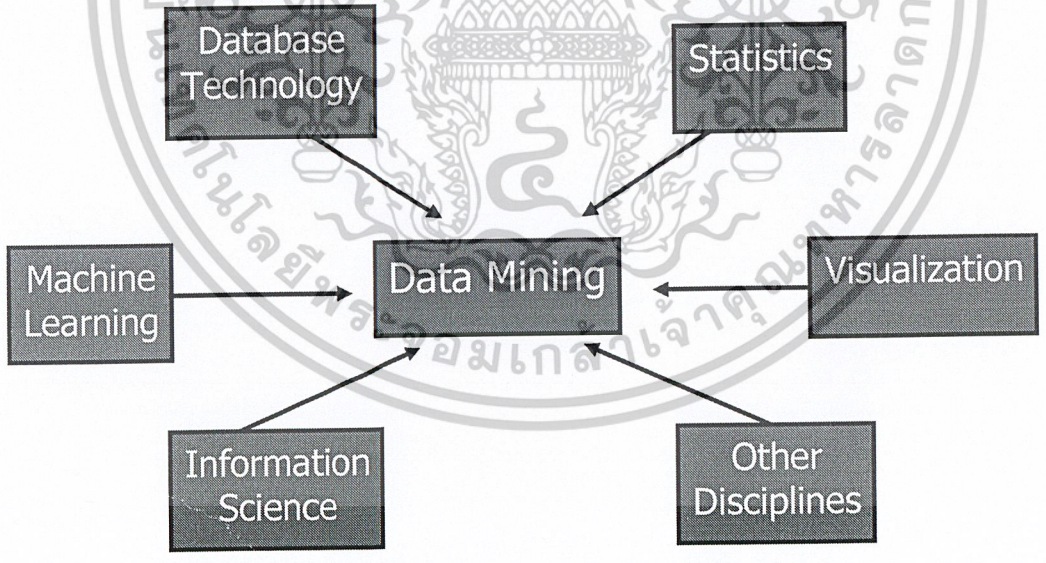
รูปที่ 2-2 : การหาองค์ความรู้จากข้อมูลที่มี

ในปัจจุบันองค์กรส่วนใหญ่จะเผชิญกับปัญหาของการที่มีข้อมูลดิบมีจำนวนมากแต่สารสนเทศหรือข้อมูลที่ประยุกต์ใช้ได้นั้นมีน้อย (Data Rich But Information Poor) เนื่องจากในโลกปัจจุบัน นับวันฐานข้อมูลยิ่งจะมีขนาดใหญ่มากขึ้นโดยเฉพาะองค์กรทางธุรกิจ ดังนั้นการดึงความรู้ออกมาจากข้อมูลจำนวนมากจึงมีความจำเป็นโดยผลลัพธ์ที่ได้จะนำไปใช้ประโยชน์ต่างๆ ในการวางแผนดำเนินงาน หรือวางกลยุทธ์ขององค์กร

พื้นฐานของดาต้าไมนิ่งจะประกอบด้วย

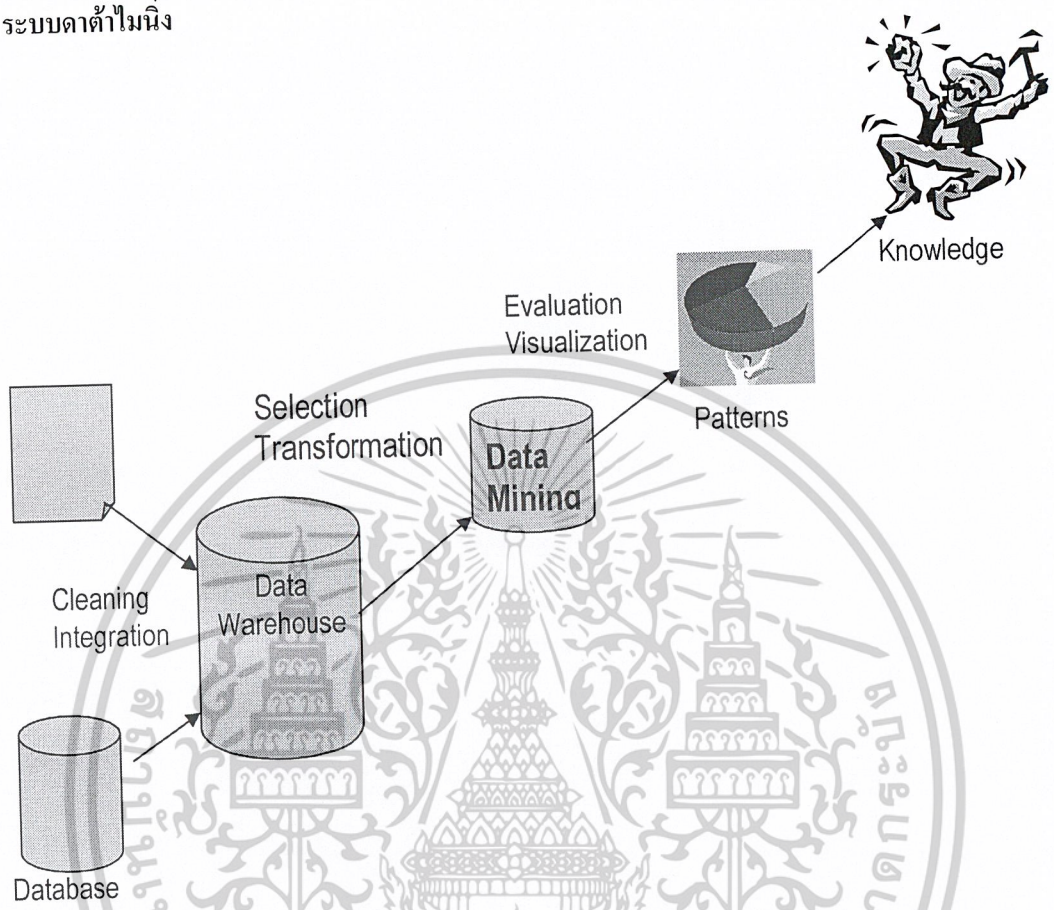
1. การจัดเก็บข้อมูลขนาดมหาศาล
2. การประมวลผลที่สมรรถนะสูงในลักษณะขนาน (Parallel Processing)
3. การเลือกใช้อัลกอริทึมที่มีประสิทธิภาพ

ดังนั้นจึงต้องใช้ความรู้หลายๆสาขาประกอบกัน องค์ประกอบต่างๆที่ใช้ในการทำงานของดาต้าไมนิ่งจึงมีเทคนิคและวิธีการหลากหลาย



รูปที่ 2-3 : วิธีการและเทคนิคในด้านต่างๆที่นำมาใช้ในดาต้าไมนิ่ง

2. ระบบดาต้าไมนิ่ง



รูปที่ 2-4 : ดาต้าไมนิ่งนั้นเป็นกระบวนการส่วนหนึ่งของกระบวนการได้มาซึ่งสารสนเทศจากข้อมูล

ส่วนประกอบของระบบดาต้าไมนิ่ง

2.1 ฐานข้อมูลหรือโกดังข้อมูล (Database / Data warehouse)

เป็นส่วนข้อมูลที่จะนำมาวิเคราะห์ โดยได้มาจากฐานข้อมูลต่างๆ โดยผ่านการคลีน (Data cleaning) จากระบบโกดังข้อมูลเพื่อให้อยู่ในรูปแบบที่พร้อมสำหรับกระบวนการดาต้าไมนิ่ง

2.2 เครื่องให้บริการฐานข้อมูลหรือโกดังข้อมูล (Server)

2.3 ความรู้พื้นฐาน (Knowledge base)

2.4 กลไกทางดาต้าไมนิ่ง (Data mining Engine)

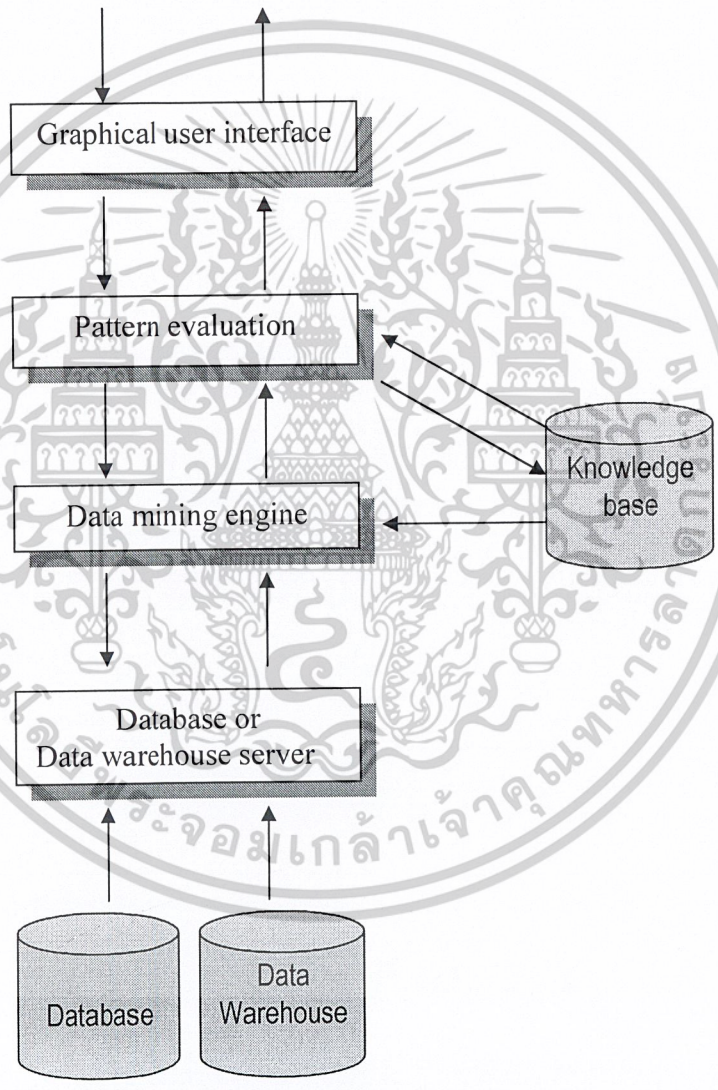
ระบบดาต้าไมนิ่ง ฟังก์ชันต่างๆที่จะเป็นตัววิเคราะห์ข้อมูลออกมาเป็นผลลัพธ์ต่างๆ

2.5 การคัดเลือกรูปแบบ

เพื่อที่จะกลั่นกรองรูปแบบที่ได้มา เราจะนำเอาเฉพาะรูปแบบที่เราสนใจไปใช้ประโยชน์

2.6 ส่วนติดต่อกับผู้ใช้ (Graphical user interface)

เพื่อให้ผู้ใช้ได้ติดต่อสื่อสารกับระบบ การระบุงาน การเลือกฐานข้อมูลที่ต้องการ



รูปที่ 2-5 : สถาปัตยกรรมของระบบดาต้าไมนิ่ง

3. งานของดาต้าไมนิ่ง (Tasks of data mining)

3.1 การทำนายค่า (Prediction) และการประเมินค่า (Estimation)

งานการประเมินค่าจะเกี่ยวข้องกับผลลัพธ์แบบต่อเนื่อง (continuous) เพื่อต้องการจะทำนายหรือประมาณค่าสำหรับตัวแปรที่เราไม่สามารถทราบค่าที่แน่นอนจากตัวแปรต่างๆที่เราทราบจากฐานข้อมูลที่มีนั่นเอง สำหรับการทำนายค่าจะเป็นในลักษณะการทำนายไปถึงอนาคต

3.2 การแยกประเภท (Classification)

เรียกอีกอย่างหนึ่งว่าเป็นการจัดหมวดหมู่ ใช้ในการจำแนกประเภทของข้อมูลที่เราต้องการ โดยการหาคุณลักษณะและนำมาตัดสินใจแบ่งหมวดหมู่อาศัยหลักการแตกเป็นแผนภูมิต้นไม้ (Decision Tree) ดังนั้นข้อมูลจะมีลักษณะเป็นค่าที่ไม่ต่อเนื่อง (discrete)

3.3 การแบ่งกลุ่ม (Clustering)

เป็นการแบ่งกลุ่มของสิ่งที่แตกต่างกันออกเป็นกลุ่มย่อยหรือคลัสเตอร์ซึ่งมีความคล้ายคลึงกันภายในกลุ่ม ซึ่งไม่ต้องกำหนดหมวดหมู่ล่วงหน้า นั่นคือลักษณะการแบ่งแยกข้อมูลเป็นกลุ่มบนพื้นฐานของความคล้ายคลึงกันในตัวเอง

3.4 การวิเคราะห์ความสัมพันธ์ (Association Analysis)

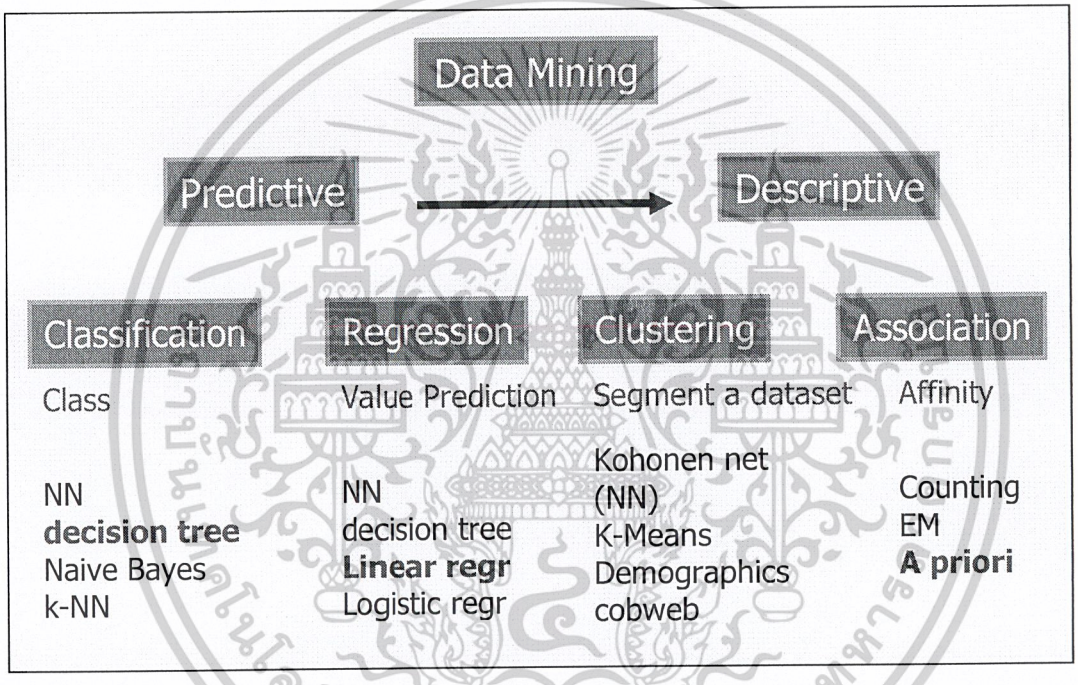
การวิเคราะห์หาความสัมพันธ์ของแต่ละสิ่งเพื่อหากฎความสัมพันธ์หรือความเกี่ยวข้องต่อกัน หรือมีความน่าจะเป็นที่จะเกิดขึ้นพร้อมกัน

4. เทคนิคของดาต้าไมนิ่ง (Data Mining Techniques)

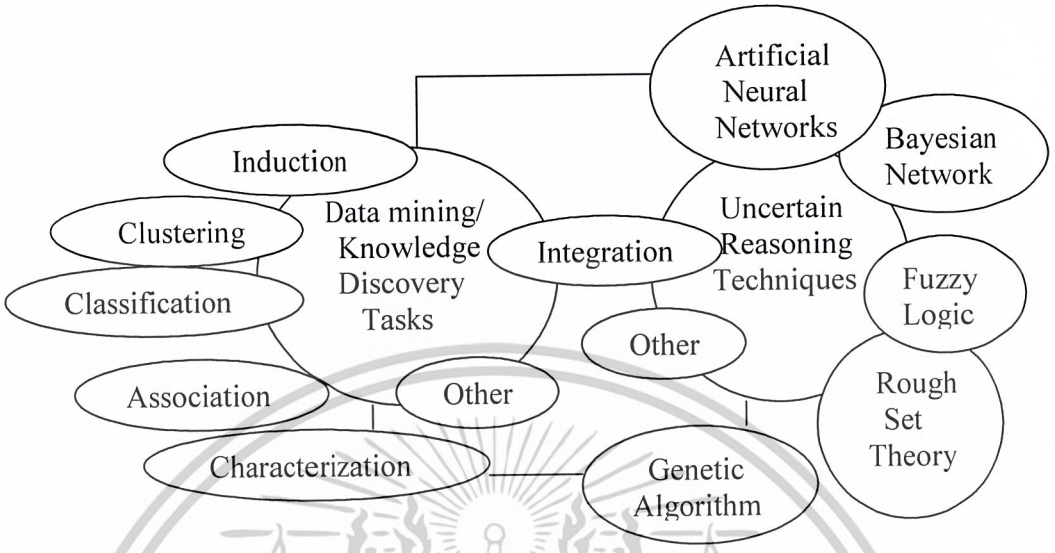
เทคนิคหรืออัลกอริทึมต่างๆที่นำมาใช้ในเรื่องดาต้าไมนิ่งตามลักษณะการใช้งาน ยกตัวอย่างเช่น

- สมการถดถอยเชิงเส้น (Regression) เป็นวิธีการคำนวณทางสถิติ
- นิวรอนเน็ตเวิร์ก (Artificial Neural Network) ซึ่งนับเป็นตัวแบบคาดการแบบไม่เชิงเส้น (Non linear predictive models) ซึ่งมีลักษณะคล้ายกับกลไกเส้นประสาททางชีววิทยา
- แผนภาพต้นไม้สำหรับการตัดสินใจ (Decision Tree : Classification And Regression Tree (CART) , Chi Square And Interaction Detection (CHAID))
- จีเนติกอัลกอริทึม (Genetic Algorithms : Optimization Technique) โดยใช้แนวคิดเลียนแบบพันธุกรรมทางชีววิทยา

- เนียเรสเนเบอร์ (Nearest Neighbor Method) ซึ่งเป็นเทคนิคทำการจำแนก แต่ละระเบียนข้อมูล (record) ในแต่ละกลุ่ม (data set) โดยให้ภายใน กลุ่มเดียวกันมีลักษณะคล้ายกัน และในต่างกลุ่มจะมีลักษณะแตกต่างกัน
- รูลอินดัคชัน (Rule Induction) เป็นการจำแนกข้อมูล โดยใช้กฎเกณฑ์ (If Then Rule) โดยใช้พื้นฐานของวิชาสถิติในการวิเคราะห์เพื่อหานัยความแตกต่างในการนำมาสร้างเป็นกฎ



รูปที่ 2-6 : การจำแนกวิธีการและเทคนิคต่างๆออกเป็นกลุ่มย่อยๆตามลักษณะการประมวลและลักษณะการใช้งานต่างๆในงานดาต้าไมนิ่ง



รูปที่ 2-7 แสดงถึงเทคนิคต่างๆของดาต้าไมนิ่ง

บทที่ 3
ความรู้พื้นฐานเกี่ยวกับสมการถดถอยเชิงเส้น

1. การคำนวณสมการถดถอยเชิงเส้น (Regression Equation)

สมการจะอยู่ในรูปแบบ

$$Y_1 = a_0 + a_1X_{11} + a_2X_{12} + \ldots + a_kX_{1k}$$
$$Y_2 = a_0 + a_1X_{21} + a_2X_{22} + \ldots + a_kX_{2k}$$
$$\vdots$$
$$Y_k = a_0 + a_1X_{k1} + a_2X_{k2} + \ldots + a_kX_{kk}$$

ซึ่งสามารถเขียนเป็นเมตริกได้ดังนี้

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ X_{21} & \cdots & X_{2k} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kk} \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_k \end{bmatrix}$$

นั่นคือ

$$\mathbf{Y = Xa}$$
$$\mathbf{X^{-1}Y = X^{-1}Xa = a}$$

จากเมตริกข้างต้นเมื่อคำนวณอินเวิร์ดได้เราก็สามารถคำนวณค่า a แทนกลับในสมการได้ เราจะสามารถหาค่า y เมื่อระบุค่า $x_1 \ldots x_k$ ได้

$$\mathbf{Y_k = a_0 + a_1X_{x1} + a_2X_{x2} + \ldots + a_kX_{xk}}$$

วิธีการประยุกต์ สมการถดถอยเชิงเส้นแบบหลายตัวแปรคือการเขียนโปรแกรมเพื่อรับค่าเซตของข้อมูล แล้วนำไปเขียนในรูปแบบ เมตริก แล้วนำค่าที่ได้มาคำนวณหาอินเวิร์ด เมื่อคำนวณได้ค่าสมการดังกล่าวก็จะสามารถหาค่าตัวแปรตามที่ต้องการได้เมื่อระบุปัจจัยตัวแปรต้น

วิธีการที่นำไปใช้จริงสำหรับฐานข้อมูลที่ เมตริกไม่เป็นสมมาตรจะต้องใช้วิธีการที่ต่างไปจากวิธีการข้างต้น โดยต้องใช้สมการดังนี้ในการหาค่า a

$$a = \begin{bmatrix} a_0 \\ a_1 \\ .. \\ \vdots \\ a_n \end{bmatrix} = (x'x)^{-1}x'y$$

โดยสิ่งที่จะต้องทำการคำนวณเพื่อหาค่า คือ การทำ ทรานโพส การหาอินเวิร์ด และการคูณกัน ของเมตริก

2. การพิจารณาความถูกต้องของข้อมูล

S.E.E. (standard error of estimated)

ค่า s

โดย $s^2 = \dfrac{1}{n-(k-1)}(y'y - a'x'y)$

ใช้ในการวัดความแม่นยำของข้อมูลมีค่า $[0,\infty)$ ค่าน้อยแสดงว่าความแม่นยำสูงกว่า หากค่าเป็น 0 แสดงว่าความแม่นยำสูงสุด

ค่า R-square

โดย $R^2 = \left[a'x'y - n\bar{y}^2\right] \Big/ \left[y'y - n\bar{y}^2\right]$

ใช้เพื่อพิจารณาว่าโมเดลถูกต้องหรือไม่ คือใช้วัดว่าตัวแปรต้นทั้งหมดที่เรากำหนดนั้นสามารถอธิบายตัวแปรตาม (y)ได้ดีเพียงใด โดยมีค่า $(0,1.00)$ ค่ามากกว่าจะถูกต้องมากกว่าโดยอาจคูณ 100 เพื่อวัดออกมาเป็นร้อยละได้

3. วิธีการทดสอบสมมุติฐาน

โดยตั้งสมมุติฐานว่า a ในแต่ละตัวมีค่าเท่ากับ 0 หรือไม่ซึ่งคือการที่ข้อมูลตัวแปรต้นนั้นไม่มีผลในการทำนายค่าของตัวแปรตามที่กำหนด

F-test

F-test ใช้เพื่อพิสูจน์ สมมุติฐานว่าค่าตัวแปรต้นทุกตัวที่กำหนดขึ้นมีค่าสัมประสิทธ์เป็น 0 โดยใช้ตาราง ANOVA ดังนี้

Source	df	SumSquare	MeanSquare	F-ratio
intercept	1	$n\bar{y}^2$	$n\bar{y}^2$	$\dfrac{n\bar{y}^2}{\left(y'y - a'x'y\right)/n-k}$ ***
Regression	k-1	$a'x'y - n\bar{y}^2$	$\left(a'x'y - n\bar{y}^2\right)/k-1$	$\dfrac{\left(a'x'y - n\bar{y}^2\right)/k-1}{\left(y'y - a'x'y\right)/n-k}$ ***
Residual	n-k	$y'y - a'x'y$	$\left(y'y - a'x'y\right)/n-k$	
Total (adj)	n-1	$y'y - n\bar{y}^2$		

ตารางที่ 3-1 ตาราง ANOVA

โดยค่า F-ratio ที่ได้จะเป็นค่าบวก นำไปเทียบกับค่า F standard ในตารางเพื่อเปรียบเทียบว่าค่า F นั้นยืนยันสมมุติฐานได้หรือไม่

t-test

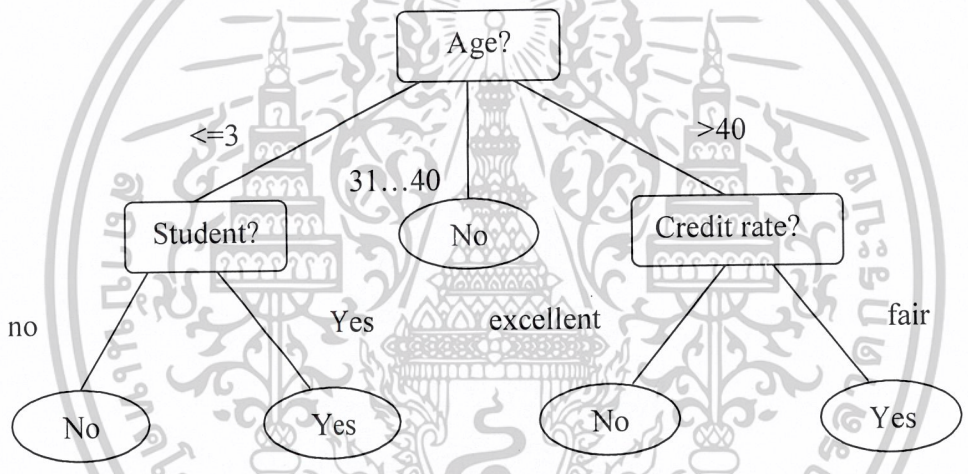
t-test ใช้เพื่อพิสูจน์ สมมุติฐานว่าค่าตัวแปรต้นตัวนั้นๆ ที่กำหนดขึ้นมีค่าสัมประสิทธิ์เป็น 0 หรือไม่

$$t_j = \frac{a_j}{\sqrt{s^2 (x'x)^{-1}_{jj}}}$$

โดยค่า t ที่ได้จะนำไปเทียบกับค่า t standard ในตารางเพื่อเปรียบเทียบว่าค่า t นั้นยืนยันสมมุติฐานได้หรือไม่

บทที่ 4
แผนภาพต้นไม้สำหรับการตัดสินใจ

แผนภาพต้นไม้สำหรับการตัดสินใจคือแผนภาพที่มีลักษณะแผนผังรูปต้นไม้ โดยที่โหนดภายใน (internal node) แต่ละโหนดจะเป็นแอทริบิวส์ (Attribute) ของตัวข้อมูลที่กำลังทำการทดสอบ และแต่ละกิ่ง (branch) จะแสดงผลของข้อมูลที่ได้ทดสอบ และโหนดที่ปลายกิ่ง (leaf node) แต่ละโหนดนั้นจะแสดงคลาส (classes) หรือคลาสรวม(class distribution) โดยที่บนสุดของโหนดบนแผนภาพคือ โหนดราก(root node) ลักษณะของแผนภาพรูปต้นไม้มีตัวอย่างดังรูปที่ที่ 4-1



รูปที่ 4-1 ตัวอย่างแผนภาพรูปต้นไม้สำหรับการตัดสินใจ

จากรูปที่ 4-1 แสดงให้เห็นแผนภาพรูปต้นไม้แสดงข้อมูลกรณีตัวอย่างลูกค้าภายในร้านอุปกรณ์ใช้ไฟฟ้าโดยโหนดภายในจะแสดงด้วยกล่องสี่เหลี่ยมและแต่ละกิ่งใช้ภาพวงกลม

สำหรับการคัดเลือกประเภทข้อมูลจากข้อมูลที่ไม่ทราบชนิด แอทริบิวส์ของข้อมูลนั้นจำเป็นจะต้องนำมาทดสอบเสียก่อนจึงจะนำผลที่ได้นั้นมาวาดลงบนแผนภูมิตั้งแต่รากไปจนถึงปลายของแผนภูมิก่อนเมื่อได้แผนภูมิมาแล้วแผนภูมิที่ได้นั้นสามารถจะนำไปสร้างแผนภูมิสำหรับการจำแนกข้อมูลได้ต่อไป

1. ขั้นตอนการสร้างแผนภาพต้นไม้สำหรับการตัดสินใจ

รูปประกอบถัดไปนี้คือขั้นตอนและวิธีการในการสร้างแผนภูมิจากข้อมูลตัวอย่าง

Algorithm: Generate_decision_tree
Input: The training samples ,*samples*, represented by discrete-valued attributes ;the set of candidate attributes,*attribure-list*
Output: A decision tree.

Method:

```
1.  create a node N;
2.  if samples are all of the same class, C then
3.      return N as a leaf node labeled with the class C;
4.  if attribute-list empty then
5.      return N as a leaf node labeled with the most common class in sample;//majority voting
6.  select test-attribute ,the attribute among attribute-list with the highest information gain;
7.  label node N with test-attribute;
8.  for each known value a_i of test-attribute // partition the samples
9.      grow a branch from node N for the condition test-attribute=a_i;
10.     let s_i be the set of sample in samples for which test-attribute=a_i;//partition
11.     if s_i is empty then
12.         attach a leaf labeled with the most common class in samples;
13.     else attach the node returned by generate_decision_tree( s_i, attribute-list-test-attribute);
```

รูปที่ 4-2 อัลกอริทึมในการสร้างแผนภาพต้นไม้สำหรับการตัดสินใจ

อัลกอริทึมที่แสดงในภาพข้างต้นนี้เป็นอัลกอริทึมพื้นฐานในการสร้างแผนภาพการตัดสินใจในลักษณะบนลงล่าง โดยใช้วิธีการแบบรีเคอชั่น(recursions) ในการทำงาน อัลกอริทึมนี้เป็นรุ่น ID3 ที่เป็นที่รู้จักในการดำเนินการแบบแผนผังการตัดสินใจรูปต้นไม้ โดยอธิบายวิธีการทำงานของอัลกอริทึม ดังนี้

- ต้นไม้เริ่มที่การสร้างโหนดเดี่ยวที่บ่งบอกถึงลักษณะข้อมูล(ขั้นที่ 1)
- ถ้าตัวข้อมูลทั้งหมดเป็นคลาสเดียวกัน ให้โหนดเปลี่ยนมาเป็นปลายโหนด(leaf) แล้วตั้งชื่อด้วยคลาสนั้น (ขั้นที่ 2 และ 3)
- ในขั้นตอนถัดไปเป็นการใช้ กรรมวิธีการวัดค่า(entropy-based measure) ในการวัดค่าหรือเรียกอีกอย่างว่า อินฟอเมชันเกน(information gain) ในการประเมินแอทริบิวส์ที่เหมาะสมที่สุดในการที่จะแยกออกเป็นคลาสอิสระ(ขั้น 6) แอทริบิวส์ที่ได้นี้จะนำมาเป็นค่าแอทริบิวส์ที่ใช้ทดสอบ หรือตัดสินใจที่โหนด(ขั้น 7) ดังจะกล่าวรายละเอียดต่อไป โดยในรุ่นของอัลกอริทึมนี้ค่าของตัวข้อมูลต่างๆนั้นต้องเป็นลักษณะ ค่าไม่ต่อเนื่อง ค่าที่เป็นค่าต่อเนื่อง จำเป็นต้องปรับเปลี่ยนหรือแปลงเป็นค่าที่ไม่ต่อเนื่องเสียก่อน

- ทำการสร้างกิ่งในแต่ละข้อมูลที่ทราบค่าของข้อมูลแอทริบิวส์ที่ใช้ทดสอบ จากนี้ตัวข้อมูลที่กำลังทดสอบอยู่จะถูกแบ่งส่วนออกจากกันแล้ว(ขั้นตอนที่ 8-10)

2. กรรมวิธีการวัดค่าในการประเมินแอทริบิวส์

ดังที่ได้กล่าวไว้ข้างต้น ค่าอินฟอเมชันเกนของแต่ละแอทริบิวส์จะถูกใช้ในการเปรียบเทียบกันเพื่อดูว่าแอทริบิวส์ใดมีความเหมาะสมที่สุดที่จะเป็นแอทริบิวส์ที่ใช้ทดสอบหรือตัดสินใจสำหรับโหนดนั้นๆ โดยค่าอินฟอเมชันเกนที่คาดหวัง(expected) คือ

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

S คือ set ของข้อมูลที่ยกมา (Sample Data)

m คือจำนวนของคลาสแอทริบิวส์

p คือ ความน่าจะเป็นที่ข้อมูลอยู่ในคลาส

ถ้าให้แอทริบิวส์ A มี v ค่าที่เป็นไปได้ $\{a_1, a_2, ..., a_v\}$ แอทริบิวส์ A ทำให้เกิดการแบ่งพาร์ทิชั่น (partition) เป็น v เซตย่อย $\{S_1, S_2, ..., S_v\}$ ซึ่ง S_j จะประกอบด้วยข้อมูลใน S ที่มีค่า a_j หาก A ถูกเลือกเป็นแอทริบิวส์ที่ใช้ทดสอบ ค่าเอนโทรปี (entropy) หรือค่าอินฟอเมชันเกนที่คาดหวังของเซตย่อย A

$$E(A) = \sum_{j=1}^{v} \frac{Sij + ... + Smj}{S} I(S_{1j}, ..., S_{mj})$$

เทอม $\frac{Sij + ... + Smj}{S}$ จะเป็นน้ำหนักของเซตย่อยที่ j ซึ่ง ค่าจะเท่ากับจำนวนข้อมูลที่ยกมาในเซตย่อยหารด้วยจำนวนข้อมูลทั้งหมด (S) ค่าเอนโทรปียิ่งน้อยส่วนของพาร์ทิชั่นจะยิ่งมีความเหมือนกันของข้อมูลในคลาสมากขึ้น

สำหรับเซตย่อยที่ j

$$I(s_{1j}, s_{2j}, ..., s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij})$$

ซึ่ง P หรือความน่าจะเป็นที่ข้อมูลที่ยกมาจะอยู่ในคลาส C_i คือ $P_{ij} = \frac{S_{ij}}{|S_j|}$

และค่าอินฟอเมชันเกนหาได้จาก

$$Gain(A) = I(S_1, S_2, ..., S_m) - E(A)$$

ตัวอย่างเช่น

ในมีคลาสแอทริบิวส์ชื่อ “buys_computer” มีค่าที่เป็นไปอยู่ 2 ค่าคือ “yes” กับ “no” {yes, no} นั่นคือ m = 2 ให้คลาส C_1 สำหรับ “yes” C_2 สำหรับ “no” จะได้ว่า มี C_1 9 ข้อมูล มี C_2 5 ข้อมูล เราจะหาค่าอินฟอเมชันเกนที่คาดหวังได้ $I(s_1, s_2) = I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$

ต่อไปเราทำการหาค่าเอนโทรปีของแต่ละ Attribute เริ่มจาก age พิจารณา “yes” กับ “no” ในแต่ละค่าของ age

RID	age	income	student	credit_rating	Class:buys_computer
1	<=30	high	no	Fair	no
2	<=30	high	no	excellent	no
3	31…40	high	no	Fair	Yes
4	>40	medium	no	Fair	Yes
5	>40	low	yes	Fair	Yes
6	>40	low	yes	excellent	no
7	31…40	low	yes	excellent	Yes
8	<=30	medium	no	Fair	no
9	<=30	low	yes	Fair	Yes
10	>40	medium	yes	Fair	Yes
11	<=30	medium	yes	excellent	Yes
12	31…40	medium	no	excellent	Yes
13	31…40	high	yes	Fair	Yes
14	>40	medium	no	excellent	no

ตารางที่ 4-1 แสดงฐานข้อมูลตัวอย่าง

สำหรับ age ที่ “<=30”;

S_{11}=2 S_{21}=3 $I(S_{11},S_{21})$=0.917

สำหรับ age ที่ “30…40”;

S_{12}=4 S_{22}=0 $I(S_{12},S_{22})$=0

สำหรับ age ที่ “>40”;

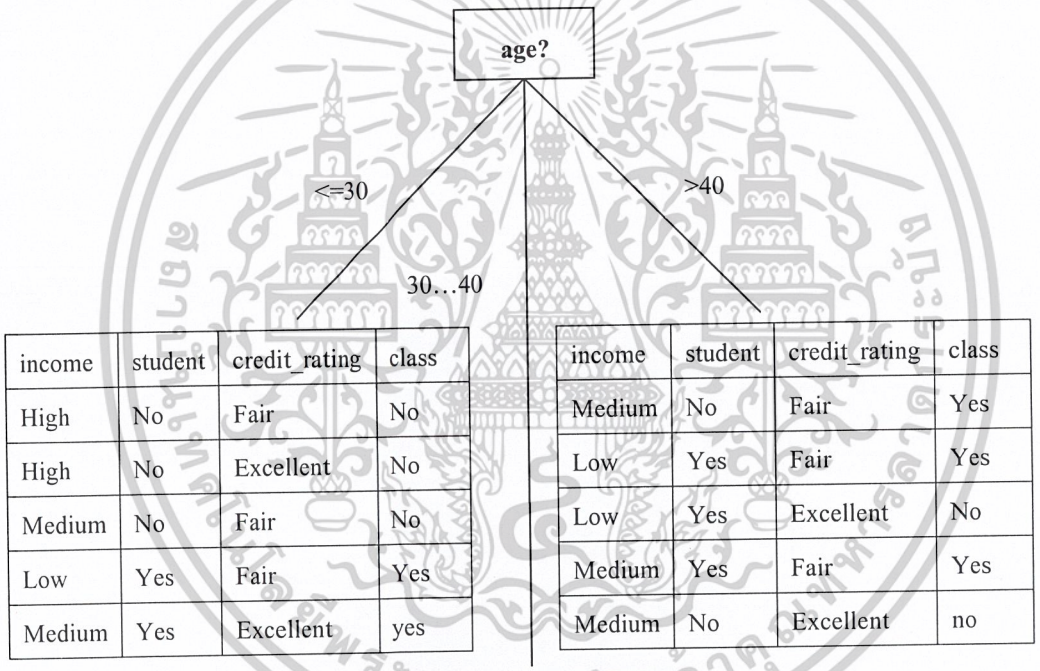
S_{13}=3 S_{23}=2 $I(S_{13},S_{23})$=0.917

$$E(age)=\frac{5}{14}I(S_{11},S_{21})+\frac{4}{14}I(S_{12},S_{22})+\frac{5}{14}I(S_{13},S_{23})=0.694$$

ดังนั้น $Gain(age)=I(S_1,S_2)-E(age)=0.246$

ในทำนองเดียวกันเราหา Gain(income)=0.029, Gain(student)=0.151 และ Gain(credit_rating)=0.048

เมื่อเปรียบเทียบกันจะเห็นว่าค่าของ Gain(age) มีค่ามากที่สุดในบรรดาค่า Gain ของแอทริบิวส์ต่างๆ เราจึงเลือกเป็นแอทริบิวส์ที่ใช้ทดสอบ ทำการสร้างโหนดชื่อ "age" และแตกกิ่งสาขาไปในแต่ละค่าของมัน ข้อมูลที่ยกมาก็จะถูกแบ่งเป็นพาร์ทิชั่นตามกิ่งสาขานี้



income	student	credit_rating	class
High	No	Fair	No
High	No	Excellent	No
Medium	No	Fair	No
Low	Yes	Fair	Yes
Medium	Yes	Excellent	yes

income	student	credit_rating	class
High	No	Fair	No
High	Yes	Excellent	Yes
Medium	No	Excellent	Yes
high	Yes	Fair	Yes

income	student	credit_rating	class
Medium	No	Fair	Yes
Low	Yes	Fair	Yes
Low	Yes	Excellent	No
Medium	Yes	Fair	Yes
Medium	No	Excellent	no

รูปที่ 4-3 แสดงตัวอย่างของการแบ่งข้อมูลในแต่ละกิ่ง

บทที่ 5

กฎของความสัมพันธ์ (Association Rules)

กฎของความสัมพันธ์เป็นหนึ่งในวิธีการศึกษาด้านดาต้าไมนิ่งซึ่งทำการศึกษาความสัมพันธ์ ความเกี่ยวเนื่องกันของตัวข้อมูลที่ได้ทำการจัดเก็บเอาไว้ ในฐานข้อมูล ซึ่งในระบบอุตสาหกรรม และระบบเศรษฐกิจในยุคปัจจุบันการเติบโตของปริมาณข้อมูลที่เก็บเอาไว้นั้นมีปริมาณมากขึ้น ทั้งนี้เพราะการพัฒนาของเทคโนโลยีที่ช่วยในการจัดเก็บข้อมูล เช่น บาร์โค้ด (barcode) ที่เกิดขึ้นจึงทำให้มีโอกาสที่จะได้รับทราบข้อมูลผลิตภัณฑ์ หรือ สินค้าต่างๆถึงระดับรายละเอียดได้ หากเราสามารถสร้างประโยชน์ขึ้นมาได้จากข้อมูลที่มีอยู่เหล่านี้เราจะสามารถสร้างประโยชน์ทางธุรกิจ และประโยชน์ด้านอื่นๆโดยไม่ปล่อยให้ข้อมูลที่มีอยู่นั้นไร้ค่า

กฎของความสัมพันธ์นั้นมีวิธีการศึกษาและใช้งานแตกต่างกันไปในหลากหลายรูปแบบ โดยจะเป็นการตั้งกฎขึ้นมาเพื่อใช้ในการพิจารณาข้อมูลที่ต้องการศึกษา ยกตัวอย่างเช่น การวิเคราะห์ตะกร้าตลาด(Market basket analysis) จะเป็นการศึกษาที่ทำการตั้งกฎที่น่าสนใจ 2 กฎหลักๆคือ กฎของซัพพอร์ต (support) และกฎของ คอนฟิเดนส์ (confidence) ที่ใช้ในการนิยามสิ่งที่ต้องการจะศึกษา ดังจะกล่าวในรายละเอียดต่อไปในบทข้อถัดไป

ซึ่งการตั้งและใช้กฎต่างๆนั้นมีหลากหลายวิธีการ ไม่ได้บังคับอยู่เพียงวิธีการใดวิธีหนึ่งซึ่งสามารถจำแนกวิธีการออกได้เป็น

1. การใช้พื้นฐานตามชนิดของตัวข้อมูลในการสร้างกฎ

ยกตัวอย่างการที่พิจารณาข้อมูล ณ จุดที่ว่าข้อมูลนั้น มี (Presence) หรือไม่มี (absence) นั่นคือกฎความสัมพันธ์แบบตรรกะ (Boolean association rule) หรือหากการอธิบายความสัมพันธ์นั้นมีปริมาณ (quantitative) หรือแอทริบิวส์ ของข้อมูลเข้ามาเกี่ยวข้องนั่นคือ กฎความสัมพันธ์เชิงปริมาณ (quantitative association rule) โดยยกตัวอย่าง

Age(X,"30...39")^income(X,"42K...48K)

=>buys(X,high resolution TV)

เป็นการยกตัวอย่างกฎที่ตั้งขึ้นเพื่อพิจารณาสินค้าที่ลูกค้าในกลุ่มอายุและกลุ่มรายได้นั้นจะซื้อเช่นตามตัวอย่างชี้ให้เห็นว่าหากลูกค้าที่มีอายุ ตั้งแต่ 30 ถึง 39 ปีและมีรายได้ตั้งแต่ 4หมื่นสองพันถึงสี่หมื่นแปดพันบาทต่อเดือนนั้นจะซื้อสินค้าคือโทรทัศน์ความละเอียดสูง จะเห็นว่าเป็นการตั้งกฎจากปริมาณรายได้ และอายุของลูกค้า

2. การใช้พื้นฐานของมิติของตัวข้อมูลในการสร้างกฎ

หากไอเทมหรือแอทริบิวส์ ที่ใช้อยู่ในกฎของความสัมพันธ์นั้นขึ้นกับตัวแปรต้นและตัวแปรตามพียงตัวเดียวนั่นคือ กฎความสัมพันธ์แบบมิติเดียว (Single dimension association rule) ยกตัวอย่างเช่น

Buys(X,"computer")=>buys(X,"financial_management_software")

แต่หากมีการใช้ปัจจัยในการสร้างกฎด้วยแอทริบิวส์ที่มากกว่านั้น คือมีตัวแปรในการตั้งกฎตั้งแต่ 3 ตัวขึ้นนั้นจะเรียกว่า กฎความสัมพันธ์แบบหลายมิติ(multidimension association rule) ซึ่งจะอาศัยตัวแปรหลักและเทคนิคการศึกษาและการใช้งานที่แตกต่างกันไป

3. การใช้พื้นฐานของระดับของสาระที่ใช้ในการตั้งกฎ

ในบางวิธีการสำหรับงานการค้นหาข้อมูลจากกฎของความสัมพันธ์นั้นอาจมีการใช้กฎในการแบ่งแยกข้อมูลมีหลากหลายระดับในตัวกฎเองยกตัวอย่างเช่นมีกฎสองข้อนี้อยู่ในการทำงานดาต้าไมนิ่ง

Age(X."30...39") => buys(X,"laptop computer")

Age(X."30...39") => buys(X,"computer")

จะเห็นว่ากฎข้อหลังนั้นจะมีระดับที่สูงกว่ากล่าวคือ laptop computer นั้น ก็ถือเป็นคอมพิวเตอร์ประเภทหนึ่งดังนั้นการตั้งกฎขึ้นมาใช้งานทั้งสองข้อแบบนี้จึงใช้กับการแบ่งงานออกเป็นระดับตามรายละเอียดที่ใช้พิจารณา

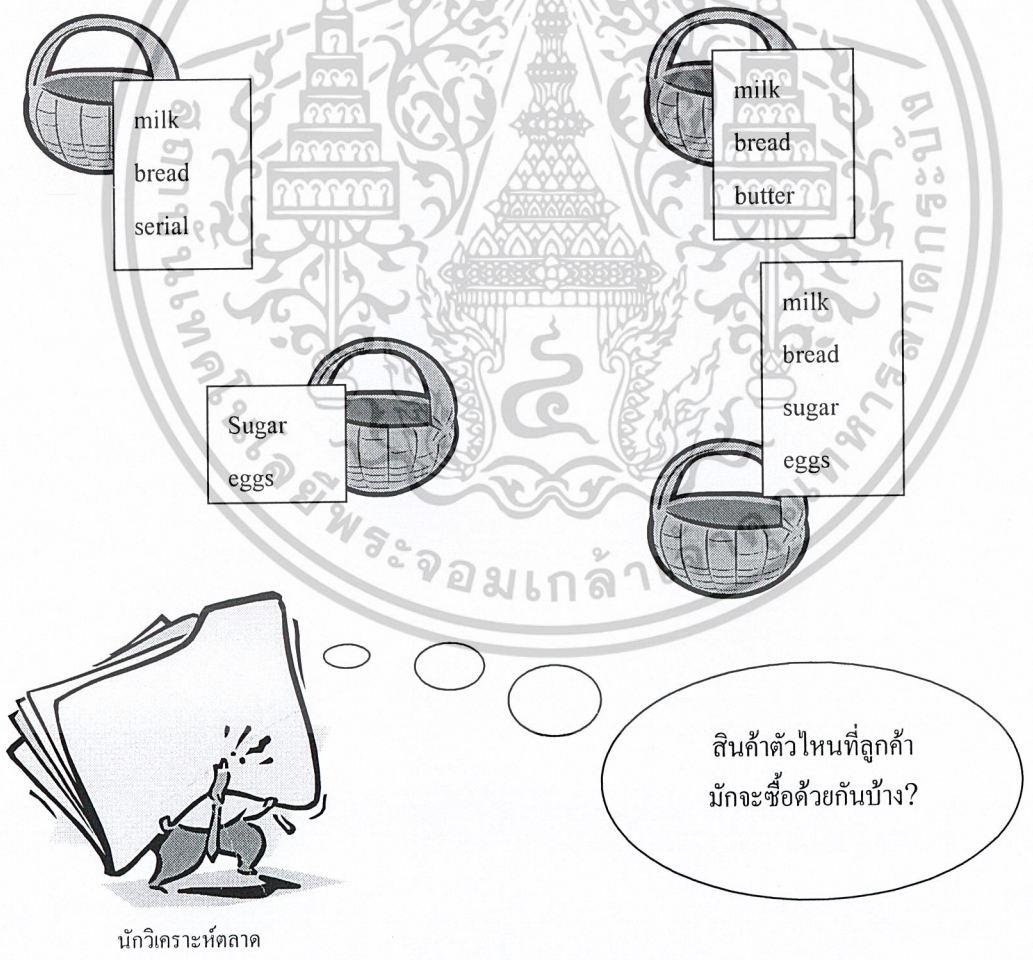
4. การใช้พื้นฐานบนส่วนขยายหลายๆแบบ

การค้นหาความสัมพันธ์ของข้อมูลนั้นสามารถที่จะใช้ความสัมพันธ์จากความเกี่ยวข้องของข้อมูลอื่นๆมาช่วยเพิ่มเติมในการปรับปรุงวิธีการค้นหาได้โดยไม่จำเป็นต้องใช้ข้อมูลแบบตรงๆนั่นคือจะใช้ในรูปแบบเพิ่มเติมจากข้อมูลจริง โดยอาศัยหลักการ แม็กซ์แพทเทิล (maxpattern) ซึ่งเป็นฟรีเควนส์ของแพทเทิลหรือ ฟรีเควนส์แพทเลแบบปิด(frequent closed itemset) ซึ่งทั้งสองแบบจะใช้หลักการคือการที่เราทราบจำนวนที่แท้จริงของรูปแบบทั้งหมดที่ข้อมูลส่วนนั้นๆมีโอกาสที่จะเกิดขึ้น ฉะนั้น หากข้อมูลมีรูปแบบใดรูปแบบหนึ่งแล้วเราจะทราบด้วยว่าข้อมูลจุดนั้นจะไม่เป็นรูปแบบอื่นในส่วนที่ไม่ได้ระบุอยู่การนำไปใช้โดยอาศัยลักษณะพื้นฐานแบบหลังสุดนี้มีการนำไปใช้น้อยกว่าวิธีอื่นเพราะความยุ่งยากในการทำความเข้าใจตัวข้อมูลนั้นมีมากกว่า

บทที่ 6

การวิเคราะห์ตะกร้าตลาด (Market Basket Analysis: MBA)

การวิเคราะห์ตะกร้าตลาดเป็นวิธีการหนึ่งในการศึกษาเหมืองข้อมูลในด้านหนึ่งของวิธีการแบบ กฎของความสัมพันธ์ โดยเป็นการศึกษาในข้อมูลประเภทหนึ่งมิติ ซึ่งเป็นวิธีการพื้นฐานที่ใช้ในการทำความเข้าใจพฤติกรรมของผู้บริโภค โดยเป็นการศึกษาว่าหากลูกค้าซื้อสินค้าชนิดใดชนิดหนึ่งแล้วจะมีผลต่อการที่ลูกค้าจะซื้อสินค้าชนิดอื่นๆอีกหรือไม่ เป็นการศึกษาพฤติกรรมความเชื่อมโยงของสินค้าในแต่ละตะกร้าหนึ่งๆที่ลูกค้าแต่ละรายได้ทำรายการเอาไว้ โดยการปฏิบัติจริงนั้นการทำการศึกษาจะต้องทำการศึกษาจากข้อมูลใบเสร็จการทำรายการ (Transaction) การศึกษาในวิธีการวิเคราะห์ตะกร้าตลาดนั้นจึงอาจเรียกในอีกชื่อหนึ่งว่า การวิเคราะห์ใบเสร็จ (Transaction analysis) ซึ่งเป็นหนึ่งในวิธีการในการวิเคราะห์ทางด้านธุรกิจด้านหนึ่งที่ใช้กรรมวิธีการทางด้านเหมืองข้อมูลโดยตรง



รูปที่ 6-1 : การวิเคราะห์ตะกร้าตลาด

1. การนำเอาการวิเคราะห์ตะกร้าตลาดมาวางแผนกลยุทธ์ทางธุรกิจ

ในฐานะผู้จัดการสาขาของร้านจำหน่ายอุปกรณ์อิเล็กทรอนิกส์จำเป็นต้องมีการตั้งคำถามว่า "กลุ่มของสินค้าตัวใดที่ลูกค้ามักจะมีความต้องการที่จะซื้อพร้อมกันในการมาซื้อในแต่ละครั้งนั้นคืออะไร" คำตอบของคำถามข้างต้นนั้นสามารถหาได้จากการวิเคราะห์ตะกร้าตลาดโดยนำข้อมูลการซื้อขายของร้านมาวิเคราะห์ ผลลัพธ์ที่ได้จากการดำเนินการข้างต้นสามารถนำมาใช้ในการวางแผนการดำเนินการทางการตลาด การวางกลยุทธ์ทางการตลาดได้ ยกตัวอย่างเช่น การออกแบบลักษณะการจัดวางร้านการจัดวางสินค้าในร้าน

กลยุทธ์หนึ่งคือการจัดวางสินค้าที่พบว่ามีการซื้อพร้อมกันบ่อยๆนำมาจัดวางไว้ใกล้ๆกันเพื่อที่จะส่งเสริมการขายได้ เช่นลูกค้าที่ต้องการซื้อคอมพิวเตอร์มีแนวโน้มที่จะซื้อซอร์ฟแวร์ไปด้วย ดังนั้นการจัดวางฮาร์ดแวร์และซอร์ฟแวร์ใกล้ๆกันนี้อาจจะช่วยให้เพิ่มยอดขายของสินค้าทั้งสอง

อีกกลยุทธ์หนึ่ง จัดวางฮาร์ดแวร์และซอร์ฟแวร์ไว้คนละฟากของร้านก็จะช่วยล่อลูกค้าให้เดินชมสินค้าหรือตัดสินใจซื้อสินค้าตัวอื่นๆระหว่างทาง เช่นระหว่างที่ตัดสินใจเลือกซื้อคอมพิวเตอร์ ระหว่างทางที่จะซื้อซอร์ฟแวร์ลูกค้าอาจจะสังเกตเห็นระบบรักษาความปลอดภัยในบ้านที่วางขายอยู่เลยตัดสินใจซื้อไปด้วย ผลจากการวิเคราะห์ตะกร้าตลาดยังสามารถนำมาใช้ในการวางแผนการขายอื่นๆเช่นการตั้งราคาสินค้า ว่าสินค้าตัวใดควรลดราคาขายพิเศษ เช่น ถ้าผลการวิเคราะห์ออกมาว่าลูกค้าที่ซื้อคอมพิวเตอร์มักจะซื้อเครื่องพิมพ์ด้วย การที่เราลดราคาเครื่องพิมพ์ก็จะช่วยส่งเสริมการขายทั้งเครื่องพิมพ์เองและคอมพิวเตอร์ไปด้วย

โดยการนิยามจากเซตว่าสินค้าทุกชนิดนั้นมีอยู่ภายในร้าน นิยามว่ายูนิเวิร์ส ทั้งหมดของสินค้าจะมีอยู่ภายในร้านฉะนั้นจะสามารถแสดงข้อมูลของสินค้าแต่ละชนิดด้วยเวคเตอร์ได้ทั้งหมด ระหว่างมี หรือ ไม่มี ตัวใดตัวหนึ่งเพื่อแสดงสถานะของสินค้าซึ่งจะนำไปใช้ในกฎของความสัมพันธ์โดยนิยามกฎของความสัมพันธ์ดังตัวอย่าง

Computer => financial_management_software

[*support*=2%, *confidence*=60%]

โดยใช้กฎของซัพพอร์ต และ คอนฟิเดนส์ ดังที่ได้กล่าวไว้ในบทก่อนหน้า ซึ่งค่าปริมาณดังกล่าวนั้นจะสะท้อนให้เห็นความสำคัญและความน่าสนใจของตัวข้อมูลที่กำลังศึกษาอยู่ ตัวอย่างค่า ซัพพอร์ต 2% นั้นหมายถึงค่าปริมาณซึ่งคือมีจำนวนข้อมูลธุรกรรมที่กำลังสนใจทำการศึกษาอยู่ ในกรณีนี้คือปริมาณเครื่องคอมพิวเตอร์ที่ซื้อพร้อมกับซอร์ฟแวร์การจัดการด้านการเงิน ส่วนค่าคอนฟิเดนส์ 60% นั้นหมายถึง 60% ของลูกค้าที่สั่งซื้อเครื่องคอมพิวเตอร์นั้นจะสั่งซื้อซอร์ฟแวร์การจัดการด้านการเงินด้วย โดยทั่วไปแล้วการวัดความน่าสนใจของข้อมูลที่ต้องการจะทำการวิเคราะห์นั้นจะใช้ค่าทั้งสองนี้ในการวัด มักใช้การกำหนดค่า ซัพพอร์ตต่ำสุด ที่รับได้ (minimum support threshold) และค่าคอนฟิเดนส์ต่ำสุด (minimum confidence threshold) ที่ต้องการจะศึกษาซึ่งอาจใช้ค่านี้พร้อมกันมากกว่าหนึ่งค่าสำหรับข้อมูลต่างๆในผู้ใช้ที่มีความชำนาญในการวิเคราะห์สูง

2. แนวความคิดพื้นฐาน

นิยาม ให้ $\tau = \{i_1,i_2,\ldots,i_m\}$ เป็นเซตของสินค้าหรือ ของ(item) ให้ D เป็นเซตของฐานข้อมูลธุรกรรม ที่ ธุรกรรม T แต่ละตัวนั้นเป็นเซตของของ นั่นคือ $T \subseteq \tau$ โดยแต่ละธุรกรรม นั้นจะมีตัวแสดงการแบ่งแยกแต่ละตัวออกจากกันโดยเรียกในที่นี้ว่า TID

ให้ A เป็นเซตของของ(set of item)

A transaction T หมายถึง A เป็นส่วนหนึ่งของ T เมื่อ $A \subseteq T$ เท่านั้น

โดยกฎของความสัมพันธ์หนึ่งคือ รูปแบบความเกี่ยวเนื่อง(Implication of the form) $A \Rightarrow B$ เมื่อ $A \subset \tau$, $B \subset \tau$ และ $A \cap B = \varnothing$ จะได้ว่า

เมื่อ ซัพพอร์ต s ในเงื่อนไข $A \Rightarrow B$ แทนค่าภายในเซตของธุรกรรม D แล้ว s คือปริมาณเป็นร้อยละของจำนวนธุรกรรมใน D ที่ครอบคลุม $A \cup B$ (จะนับรวมทั้งหมดที่มีทั้ง A และทั้ง Bหรืออย่างใดอย่างหนึ่ง) โดยเขียนแทนในรูปแบบความน่าจะเป็นคือ $P(A \cup B)$

เมื่อคอนฟิเดนส์ c ในเงื่อนไข $A \Rightarrow B$ แทนค่าภายในเซตของธุรกรรม D แล้ว c คือปริมาณเป็นร้อยละของจำนวนธุรกรรมใน D ที่ครอบคลุมทั้งAและB(จะนับเฉพาะที่มีทั้งAและB ทั้งคู่) โดยเขียนในรูปแบบความน่าจะเป็นได้คือ $P(B|A)$

$$\textit{support}\ (A \Rightarrow B) = P(A \cup B)$$

$$\textit{confidence}\ (A \Rightarrow B) = P(B|A)$$

โดยการกำหนดกฎที่ใช้นั้นจะทำโดยกำหนดค่าซัพพอร์ตต่ำสุด หรือ ค่าคอนฟิเดนส์ต่ำสุด ในการใช้พิจารณาหากกำหนดทั้งสองค่าจะเรียกว่าการตั้งกฎแบบเข้มงวด(strong rule) โดยปกติในทางปฏิบัตินั้นมักจะนิยมใช้การกำหนดค่าเหล่านี้เป็นร้อยละ ที่มีค่า 0%ถึง100% มากกว่าการกำหนดเป็นทศนิยมที่มีค่าระหว่าง 0.0 ถึง 1.0

นิยามเซตของของ ในชื่อใหม่ว่า ไอเทมเซต(itemset) ไอเทมเซตที่ครอบคลุม K ไอเทม(item) นั้นเรียกว่า K-ไอเทมเซต เช่น เซต {computer,financial_management_software} ถือเป็น 2-ไอเทมเซต จำนวนความถี่ที่ใช้ในการระบุจำนวนของไอเทมเซตนั้นเรียกชื่อว่า ความถี่(frequency) ซัพพอร์ตเคาท์(support count) หรือ จำนวนของไอเททมเซต(count of itemset) อย่างใดก็ได้ โดยการนำมาใช้งานของค่าต่างๆเหล่านี้จะเริ่มต้นโดยการกำหนดค่า ซัพพอร์ตเคาท์ต่ำสุด(minimum support count) เพื่อใช้เป็นตัววัดปริมาณของข้อมูลที่จะใช้เป็นเกณฑ์ตัดสินใจในการวิเคราะห์ข้อมูลต่างๆ

3. ขั้นตอนและวิธีการในการสร้างกฎจากฐานข้อมูลขนาดใหญ่

1.หาความถี่ทั้งหมดของไอเทมเซตโดยนิยามว่าไอเทมเซตทั้งหมดที่จะหานี้ต้องมีค่าความถี่อย่างน้อยเท่ากับค่าซัพพอร์ตเคาท์ต่ำสุดที่ได้กำหนดไว้

2.สร้างกฎแบบเข้มงวดจากความถี่ของไอเทมเซตโดยนิยามว่ากฎที่ตั้งขึ้นต้องระบุทั้งค่าซัพพอร์ตต่ำสุด และ ค่าคอนฟิเดนส์ต่ำสุด ทั้งสองค่า

เนื่องจากขั้นตอนและวิธีการข้างต้นนั้นเป็นการวางหลักการและวิธีการกว้างๆ ในการปฏิบัติจริงนั้นตัวข้อมูลที่จะทำการวิเคราะห์นั้นมักจะมีขนาดที่ใหญ่โตมากเกินกว่าจะจัดการด้วยวิธีการทั่วๆไปจึงจำเป็นต้องมีเทคนิคในการประมวลผลกับข้อมูลขนาดใหญ่รองรับการทำงานดังกล่าว โดยในปริญญานิพนธฉบับนี้จะอาศัยเทคนิคและแนวคิดแบบอัพไพรออรี อัลกอริทึม ในการแก้ไขปัญหาโดยจะกล่าวถึงรายละเอียดในหัวข้อถัดไป

4. การหากฎความสัมพันธ์โดยใช้เทคนิคแบบอัพไพรออรีอัลกอริทึม

แนวคิดพื้นฐานของอัพไพรออรีอัลกอริทึมนั้นจะใช้วิธีการสร้าง 1.แคนดิเดตเซต(candidate set)ของไอเทมเซตจำนวนมาก 2.จะทำการนับจำนวนที่เกิดขึ้นของแคนดิเดตเซต จากนั้นจึงจะทำการประเมินและวิเคราะห์ข้อมูลจากทั้งสองส่วนที่ได้กับค่าซัพพอร์ตต่ำสุดเพื่อนใช้เป็นเกณฑ์ในการตัดสินใจ โดยเทคนิคพิเศษหลักๆที่เป็นคุณสมบัติหลักของ อัพไพรออรีก็คือการที่ซับเซตทุกตัวของ ฟรีเควนส์ไอเทมเซต(frequent itemset) นั้นจะต้องเป็นไอเทมเซตที่มีความถี่พอในการพิจารณาด้วยเสมอ โดยสิ่งที่ใช้ในการพิจารณาด้วยเทคนิคอัลไพรออรีนั้นจะมีส่วนประกอบดังต่อไปนี้

- ไอเทมเซต -เซตของไอเทม
- K-ไอเทมเซต -ไอเทมเซตที่มีจำนวนไอเทมเป็น K ตัว
- ฟรีเควนส์ไอเทมเซต -ไอเทมเซตพร้อมทั้ง มีค่าซัพพอร์ตต่ำสุด หรือ อาจเรียกว่าไอเทมเซตขนาดใหญ่(large itemset)
- L_K -เซตของไอเทมเซต K ขนาดใหญ่
- C_K -เซตของแคนดิเดต K ไอเทมเซต

ภาพประกอบถัดไปคือตัวอย่างการประมวลผลจากไอเทมเซตขนาดใหญ่ โดยในที่นี้ให้เข้าใจว่าค่าซัพพอร์ตต่ำสุดที่ใช้พิจารณานั้นมีค่าเท่ากับ 2 โดยขั้นตอนในวิธีการอพไพรออรีส่วนที่จะกล่าวจากนี้มีชื่อเรียกว่า "ส่วนการตัดเลือกข้อมูล" (Prune phase)

จากตาราง แสดงให้เห็นข้อมูลในฐานข้อมูล Dที่มีข้อมูลธุรกรรมต่างๆอยู่

TID	Items
1000	A C D
2000	B C E
3000	A B C E
4000	B E

ตารางที่ 6-1: แสดงลักษณะข้อมูลในฐานข้อมูล D

วิธีการขั้นแรกของอัพไพรออรีคือการแสกนข้อมูลในฐานข้อมูล D เพื่อหาความถี่ของข้อมูลแต่ละตัวที่เกิดขึ้นจริงๆ

C_1	
Itemset	Support
A	2
B	3
C	3
D	2
E	3

ตารางที่ 6-2 : แสดงลักษณะC_1ที่ได้จากการแสกนฐานข้อมูล D

หลังจากได้ความถี่ของข้อมูลแต่ละตัวแล้วก็จะนำข้อมูลมาเปรียบเทียบกับค่าซัพพอร์ตต่ำสุดหากมีค่ามากกว่าหรือเท่ากับค่าซัพพอร์ตต่ำสุดก็จะ ฟรีเควนส์ไอเทมเซตนั้นไว้ในตาราง L_1

L_1	
Itemset	Support
A	2
B	3
C	3
D	3

ตารางที่ 6-3 : แสดงลักษณะข้อมูล L_1 ที่ได้จากการเลือกจากข้อมูลใน C_1

การพิจารณาถัดไปคือการสร้าง C_2 โดยใช้ข้อมูลที่ได้ใน L_1 มาใช้ในการจับคู่โดยจะจับคู่ทุกคู่ทุกกรณีภายใน L_1 แล้วทำการแสกนหาข้อมูลจริงใน D เพื่อสร้าง C_2

C_2	
Itemset	Support
AB	1
AC	2
AE	1
BC	2
BE	3
CE	2

ตารางที่ 6-4 : แสดงลักษณะข้อมูล C_2 ได้จากการแสกนข้อมูลใน D โดยใช้ L_1 ตัดสินใจ

หลังจากได้ C2 ก็จะทำการสร้าง L_2 โดยพิจารณาจาก ค่าซัพพอร์ตต่ำสุดเช่นเดียวกับการสร้าง L_1 โดยจะเลือกไว้เฉพาะส่วนที่มีค่าเกินกว่าหรือเท่ากับค่าซัพพอร์ตต่ำสุด

L_2	
Itemset	Support
AC	2
BC	2
BE	3
CE	2

ตารางที่ 6-5 : แสดงลักษณะข้อมูล L_2 ที่ได้จากการเลือกจากข้อมูลใน C_2

จากนั้นทำการสร้าง C_3 ขึ้นมาจากข้อมูลใน L_2 ซึ่งจะได้ข้อมูลดังตาราง เหตุที่ไม่พิจารณากรณี ABC และ ABE มาในตารางนี้เพราะว่า ซับเซตของทั้งคู่นั้น คือ AB นั้นมีค่าซัพพอร์ตต่ำเกินไปจากแนวคิดพื้นฐานที่ว่าจะพิจารณาก็ต่อเมื่อซับเซตทุกตัวของกรณีการจับกลุ่มนั้นๆต้องมีค่ามากกว่าค่าซัพพอร์ตต่ำสุดจึงได้ตาราง C_3 ดังที่แสดง

C_3	
Itemset	Support
BCE	2

ตารางที่ 6-6 : แสดงลักษณะข้อมูล C_3 ได้จากการแสกนข้อมูลใน D โดยใช้ L_2 ตัดสินใจ

พิจารณา L_3 เช่นเดียวกับ L_1 และ L_2 จะได้ตารางดังที่แสดงไว้

L_3	
Itemset	Support
BCE	2

ตารางที่ 6-7: แสดงลักษณะข้อมูล L_3 ที่ได้จากการเลือกจากข้อมูลใน C_3

5. ข้อสังเกตของอัพไพรออรีอัลกอริทึม

จากข้อมูลในเนื้อหาข้างต้นจะแสดงให้เห็นว่าการใช้งานการวิเคราะห์ด้วยวิธีการอัพไพรออรีนั้นเราจะทำการสร้างกฎของความสัมพันธ์ในฐานข้อมูลที่ตัวข้อมูลมีค่ามากว่าค่าซัพพอร์ตต่ำสุด ซึ่งสดงให้เห็นถึงกฎของความถี่ และค่าคอนฟิเดนส์ต่ำสุดต้องมีค่ามากพอ โดยการใช้ทั้งสองค่านี้ในการพิจารณาเราจะได้กฎแบบเข้มข้น ในการวิเคราะห์ข้อมูลแต่การใช้งานจริงนั้นเราจะต้องใช้ขั้นตอนการค้นหาปริมาณข้อมูลในไอเทมเซตที่มีค่ามากกว่าค่าซัพพอร์ตต่ำสุดซึ่งจะใช้พื้นที่ในการค้นหาทั้งหมดเท่ากับ 2^m ซึ่ง m คือจำนวนของไอเทม โดยจะมีลักษณะเป็นเอกโพเนนเชียล โดยที่เราให้จำกัดว่าค่าจำนวนของไอเทมนั้นมีจำกัดเราก็จะสามารถประเมินเวลาและพื้นที่ที่ใช้ในการค้นหาข้อมูลได้

บทที่ 7
การนำไปประยุกต์ใช้ทางธุรกิจ

ในระบบธุรกิจในปัจจุบันเทคโนโลยีการจัดเก็บข้อมูลนั้นได้เข้ามามีบทบาทมากขึ้นตามลำดับ โดยเฉพาะในปัจจุบันที่ข้อมูลทางธุรกิจ หรือข้อมูลอื่นใดนั้นล้วนแต่เก็บไว้ในฐานข้อมูลเป็นส่วนใหญ่ ทั้งนี้เพื่อความสะดวกในการจัดการข้อมูลและเป็นการเพิ่มบริการที่จำเป็นในการแข่งขันทางธุรกิจต่างๆนั้นล้วนแต่ใช้คอมพิวเตอร์เป็นสื่อในการเก็บรักษา รวบรวมและนำไปใช้งานจนเกือบจะเข้ามาแทนระบบเอกสารแบบเดิมๆที่เคยใช้กันอยู่แทบทั้งหมด

โดยรูปแบบและวิธีการของข้อมูลที่จัดเก็บอยู่ในปัจจุบันนั้นมีอยู่หลากหลายรูปแบบและวิธีการ เช่น การใช้ระบบฐานข้อมูล (database system) ระบบแฟ้มข้อมูล(file base system) ไม่ว่าจะใช้วิธีการจัดเก็บและจัดการเช่นไร ซึ่งสิ่งที่เราจะได้เก็บ บันทึกไว้จริงๆนั้นก็คือ ตัวข้อมูล (data) ซึ่งข้อมูลที่ทำการจัดเก็บนั้นก็มีหลากหลายประเภทเช่น ข้อมูลส่วนตัวของบุคคล ข้อมูลการทำธุรกรรม ข้อมูลประวัติ ข้อมูลสินค้าต่างๆ ซึ่งพบว่ามีอยู่บ่อยมากที่ตัวข้อมูลที่เก็บเอาไว้นั้นไม่ได้มีการนำไปใช้ดำเนินการใดๆ และมีแต่จะมากขึ้นทุกวันโดยไม่ก่อให้เกิดประโยชน์และความคุ้มค่าทางเศรษฐกิจอะไรเลย ยกตัวอย่างเช่น ข้อมูลใบเสร็จค่าไฟฟ้า ค่าน้ำประปา ข้อมูล รายชื่อประวัตินักศึกษาเก่า ซึ่งข้อมูลเหล่านี้เป็นสิ่งที่จำเป็นต้องเก็บเอาไว้ตลอด ไม่สามารถนำไปทำลายทิ้งได้ นี่คือจุดเริ่มต้นของแนวความคิดในการที่จะนำข้อมูลเหล่านี้มาสร้างประโยชน์คือการค้นหาสารสนเทศ (information) ให้แยกออกมาจากข้อมูลจำนวนมากที่เมื่อมองผิวๆแล้วเป็นสิ่งที่ไม่มีค่า

การจะนำงานดาต้าไมนิ่งไปประยุกต์ใช้ในทางธุรกิจนั้นเป็นสิ่งที่ทำได้ง่ายและตรงตัวมากกว่าการนำไปใช้ในทางวิทยาศาสตร์ อนึ่ง เพราะเทคนิคที่นำมาใช้นั้นจะสามารถใช้วิธีการแบบมีเหตุผล (reasoning) ได้มากกว่าการนำไปใช้ทางวิทยาศาสตร์และสังคมศาสตร์ที่มักจะใช้วิธีการแบบไร้เหตุผล (unreasoning) ในการวิเคราะห์ตัวข้อมูลยกตัวอย่างวิธีการแบบใช้เหตุผลเช่นการวิเคราะห์โดยใช้หลักสถิติ การวิเคราะห์ด้วยกฎของความสัมพันธ์เช่นการวิเคราะห์ตะกร้าตลาด ซึ่งจะเป็นการวิเคราะห์แบบอาศัยหลักการและเหตุผลที่อธิบายที่มาที่ไปได้ ส่วนวิธีการแบบไม่อาศัยเหตุผล เช่น วิธีการแบบนิวรอนเนตเวิร์ค วิธีการจีเนติกอัลกอริทึม ซึ่งวิธีการเหล่านี้นั้นจะเป็นวิธีการที่หาเหตุผลมาอธิบายละตอบคำตามจากสิ่งที่ได้เกิดขึ้นไม่ได้ซึ่งไม่สมควรจะนำไปใช้ในทางธุรกิจประเภทที่ต้องการเหตุผลรองรับการตัดสินใจ แต่ในกรณีการศึกษาข้อมูลประเภทที่ไม่สนใจเหตุผลรองรับนั้นก็สามารถใช้วิธีการพวกดังกล่าวศึกษาได้เช่นกัน

โปรแกรมที่มีความสามารถหลากหลายทางดาต้าไมนิ่งนั้นสามารถช่วยเราในการนำมาใช้วิเคราะห์ในงานธุรกิจด้านต่างๆ ยกตัวอย่างเช่น

Users	Applications
การตลาด	การทำนายการตอบสนองของตลาด การแบ่งกลุ่มตลาด การประเมินลูกค้า
ธนาคาร	การวิเคราะห์ความเสี่ยงเครดิต การตรวจหาการทุจริต การวางกลยุทธ์ดึงดูดลูกค้า
บริษัทกลุ่มสินเชื่อ	การตรวจหาการทุจริต รูปแบบพฤติกรรมการซื้อ
บริษัทประกัน	ใช้ในการประเมินประเภทสัญญากรมธรรม์ ประเมินพฤติกรรมการฉ้อโกงประกัน ใช้ในการตัดสินใจกรณีพิเศษต่างๆของการทำประกัน ใช้ในการประเมินเบี้ยประกันที่สมควรของลูกค้าในแต่ละราย
ร้านค้าปลีก	การวิเคราะห์ข้อมูลที่จุดขาย การวิเคราะห์ตะกร้าตลาด การวางแผนดำเนินงานจัดซื้อ การจัดวางร้าน
ด้านความปลอดภัยและงานเกี่ยวกับหลักทรัพย์และเงินทุน	ใช้ในการทำนายและประเมินความปลอดภัยของกรณีต่างๆ ซึ่งรวมถึงการประเมินพฤติกรรมของตลาดหลักทรัพย์ และช่วยได้การปรับปรุงประสิทธิภาพของการวางกลยุทธ์ต่างๆ
บริษัทสื่อสารและอุปกรณ์ด้านการสื่อสาร	เพิ่มประสิทธิภาพของบริการ ช่วยทำนายเวลาที่มีผู้ใช้เครือข่ายสูงสุดในแต่ละวันเพื่อตรวจสอบและบำรุงเครือข่ายและบริการ
ผู้ตรวจสอบด้านภาษี	ใช้ในการตรวจหาการทุจริตและช่วยในการประเมินรายได้และสินทรัพย์ต่างๆที่จะเกิดขึ้นในอนาคต
บริษัทผู้ผลิตยา	ทำนายผลการทดสอบในอนาคตและโปรแกรมการทดสอบ
โรงพยาบาล การแพทย์	การวินิจฉัยโรคอัตโนมัติ
ผู้จัดการโรงงานอุตสาหกรรม	การควบคุมคุณภาพ การเพิ่มประสิทธิภาพ

ตารางที่ 7-1 แสดงการนำดาต้าไมนิ่งไปประยุกต์ใช้งานในด้านต่างๆ

รายละเอียดของโปรแกรมและตัวอย่างการประยุกต์

1. โปรแกรมคำนวณสมการถดถอยเชิงเส้น

อินพุท :ค่าแอทริบิวส์ของค่าตัวแปรต้น และค่าตัวแปรตามที่ต้องการจะใช้ให้เป็นปัจจัยในการคำนวณ ค่าข้อมูลที่ใช้เป็นตัวแปรต้นในการคำนวณ

เอาท์พุท :ค่าตัวแปรตาม ณ ตัวแปรต้นที่กำหนดให้

ลักษณะข้อมูลที่ใช้ในการวิเคราะห์

Factor 1	Factor 2	Factor 3	Factor 4	…	Factor N
Value 1	Value 1	Value 1	Value 1	…	Value 1
Value 2	Value 2	Value 2	Value 2	…	Value 2
Value 3	Value 3	Value 3	Value 3	…	Value 3
…	…	…	…	…	…
Value M	Value M	Value M	Value M	…	Value M

ตารางที่ 7-2 แสดงลักษณะข้อมูลที่ใช้กับโปรแกรมสมการถดถอย

โดยค่าที่ใช้จะต้องเป็นค่าเชิงตัวเลขหากต้องการจะวิเคราะห์ข้อมูลที่มิได้เป็นตัวเลขให้ทำการแปลกลักษณะของข้อมูลนั้นให้อยู่ใช่รูปตัวเลขเสียก่อน เช่น

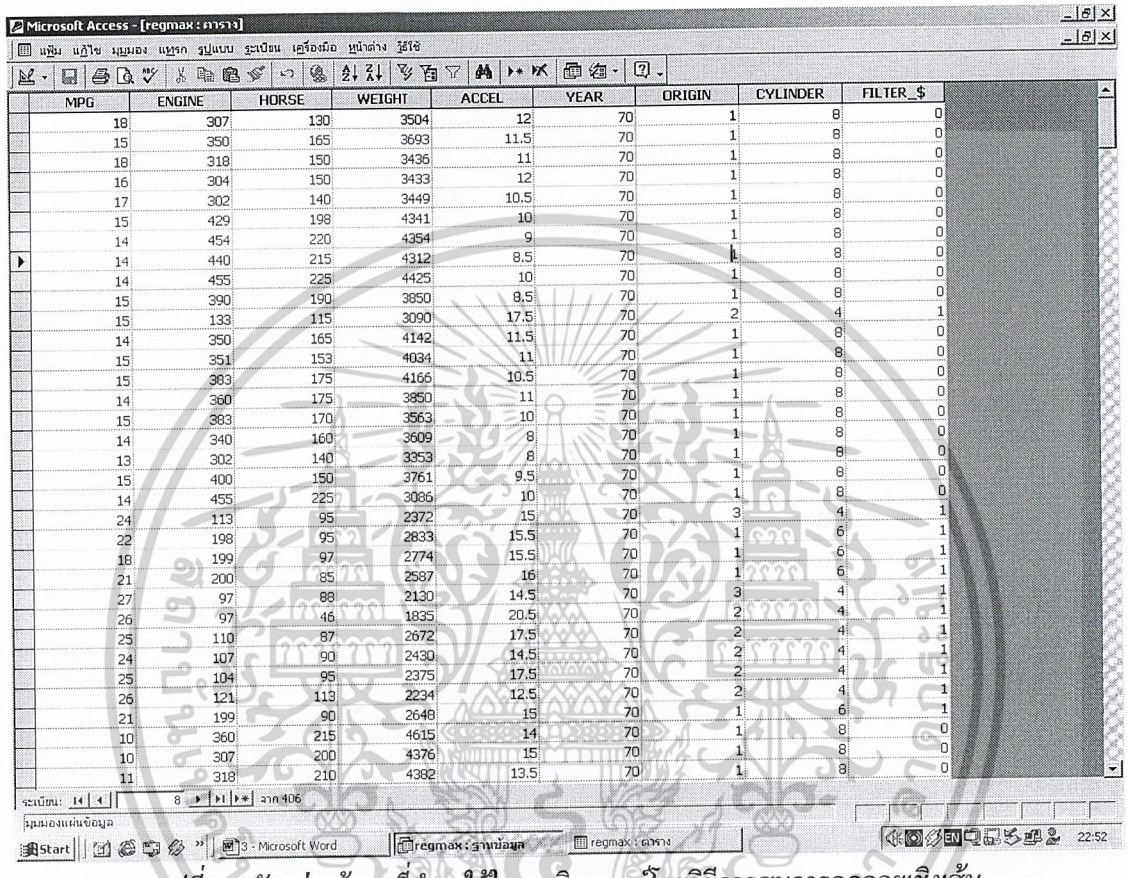
เพศ	เพศ
ชาย	0
หญิง	1
ไม่ระบุ	2

ตารางที่ 7-3 แสดงวิธีการแปลงข้อมูลที่ไม่ใช่ข้อมูลในรูปตัวเลข

เป็นการใช้ตัวเลขในการเป็นตัวแทนข้อมูลที่มิใช่ตัวเลขเพื่อสามารถนำไปใช้คำนวณได้

ตัวอย่างการประยุกต์ใช้งานโปรแกรมคำนวณสมการถดถอยเชิงเส้น

การประมาณการบริษัทผลิตรถยนต์รายหนึ่งต้องการทำการวิเคราะห์ข้อมูลเกี่ยวกับรถที่ผลิตจากปัจจัยต่างๆ เท่าที่เราทราบจากฐานข้อมูลของบริษัท



รูปที่ 7-1 ตัวอย่างข้อมูลที่นำมาใช้ในการวิเคราะห์โดยวิธีการสมการถดถอยเชิงเส้น

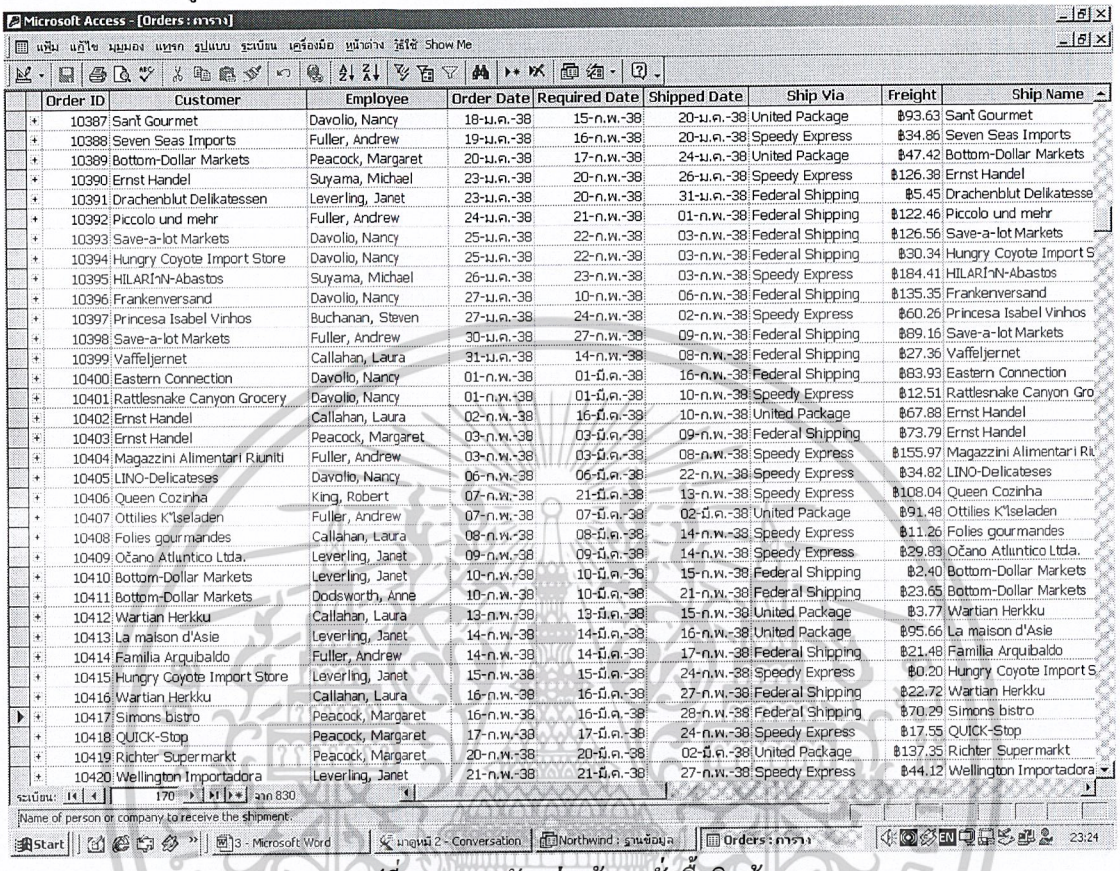
ตัวอย่างเป็นข้อมูลเกี่ยวกับข้อมูลต่างๆของรถยนต์เช่น เครื่องยนต์ แรงม้า น้ำหนัก ความเร่ง ปริมาณน้ำมันที่ใช้ในการวิ่ง ซึ่งปัจจัยที่มีนั้นสามารถนำมาคำนวณเพื่อหาความสัมพันธ์ และ สามารถคำนวณค่าตัวแปรตามจากปัจจัยที่กำหนดได้

จากโปรแกรมทำการเลือกปัจจัยที่เป็นตัวแปรต้นและตัวแปรตาม จากนั้นกำหนดค่าตัวแปรต้นทั้งหมด โปรแกรมจะทำการคำนวณค่าตัวแปรตามออกมาให้ ตัวอย่างการประยุกต์จากข้อมูลตัวอย่างนี้ เช่น การคำนวณปริมาณน้ำมันที่ใช้ต่อตัวแปรต้นที่เปลี่ยนไปเช่น ขนาดและแรงม้าของเครื่องยนต์

ตัวอย่างการประยุกต์อื่น

การประเมินราคายอดสั่งซื้อสินค้า

ตัวอย่าง ข้อมูลยอดการสั่งซื้อประกอบไปด้วยข้อมูลต่างๆในใบสั่งซื้อซึ่งมีแอทริบิวส์จำนวนมากที่เป็นข้อมูลที่ไม่ใช่เชิงตัวเลขดังรูป



รูปที่ 7-2 แสดงตัวอย่างข้อมูลสั่งซื้อสินค้า

การเตรียมข้อมูลเพื่อนำมาใช้ในการวิเคราะห์จะทำได้โดยเริ่มจากการเปลี่ยนข้อมูลที่ไม่ใช่ข้อมูลเชิงตัวเลขเป็นข้อมูลเชิงตัวเลข

เช่น การแปลงวันที่

Order_date	Order_day	Order_month	Order_year
01-กพ-38	1	2	38
10-กพ-38	10	2	38
18-มีค-38	18	3	38

ตารางที่ 7-4 แสดงการแบ่งและปรับเปลี่ยนข้อมูลวันที่

การแปลงชื่อเป็นเลขประจำตัว

Ship Via	Ship Via
United Package	12
Speedy Express	13
Federal Shipping	14
Federal Shipping	14
Federal Shipping	14

ตารางที่ 7-5 แสดงการเปลี่ยนข้อมูลชื่อเป็นตัวเลข

จากตัวอย่างการปรับเปลี่ยนข้อมูลดังกล่าวจากตัวอย่างข้างต้นเมื่อกระทำการปรับเปลี่ยนแล้วนำไปทำการคำนวณจากตัวโปรแกรม สิ่งที่เราจะทำการวิเคราะห์ได้จากข้อมูลที่ป้อนเข้าไปยกตัวอย่างเช่น

- วันเวลาที่ผู้ซื้อมักจะซื้อมากที่สุดคือช่วงใดของเดือน จะคำนวณได้โดยใช้ยอดการซื้อและวันที่ในการสั่งเป็นตัวแปรในการคำนวณ
- ปริมาณการสั่งซื้อล่วงหน้าในเดือนถัดไปที่น่าจะเป็น จะคำนวณได้โดยใช้ยอดการสั่งซื้อ เดือนที่สั่งซื้อเป็นตัวแปรในการคำนวณ
- ประเมินยอดการสั่งซื้อ จะใช้ข้อมูล ชื่อลูกค้า ประเทศที่สั่ง เวลาในการสั่ง วิธีในการจัดส่งในการประเมินว่ายอดสั่งควรจะเป็นเท่าไร

การประเมินราคาหุ้น

การประเมินราคาหุ้นโดยใช้โปรแกรมนี้เข้าช่วยจำเป็นที่ผู้ใช้จะต้องตัดสินใจเลือกและหาปัจจัยที่มีผลต่อราคาหุ้น ณ เวลานั้นด้วยตัวผู้ใช้เอง ตัวอย่างการใช้ทำได้โดยทำการเก็บข้อมูลของหุ้นในแต่ละตัวที่คาดว่าจะเป็นหุ้นและข้อมูลที่มีผลกระทบถึงกัน เช่น เก็บราคาปัจจุบัน ราคาเสนอซื้อ ราคาเสนอขาย ปริมาณเสนอซื้อ เสนอขาย เวลาที่เก็บข้อมูล ราคาหุ้นในกลุ่มเดียวกันราคาหุ้นต่างกลุ่ม ปัจจัยภายนอก โดยตัวอย่างการทำรูปแบบข้อมูลที่จะนำมาใช้ในการประเมินราคาหุ้น

Stock_ID	Sell_offer_price	Sell_volume	Bid_offer_price	Bid_Volume	day	month	year	time	Group_id	EX_factor
012	10.5	100000	9.5	2000	3	3	2004	12	1	-10
014	25.8	120000	24.5	120000	3	3	2004	36	1	5
121	13.6	1000000	13	1000	3	3	2004	28	6	0

ตารางที่ 7-6 ตัวอย่างการเก็บข้อมูลเพื่อทำนายราคาหุ้น

การจะประเมินราคาหลักทรัพย์หรือหุ้นนั้นเป็นเรื่องที่เป็นไปได้ที่จะทำโดยการใช้หลักสถิติในการประเมินแต่จำนวนปัจจัยที่มีผลต่อค่าความเปลี่ยนแปลงที่จะเกิดขึ้นนั้นมีมากน้อยเพียงใดก็ขึ้นอยู่กับประสบการณ์ของผู้ใช้ในการที่จะหาข้อมูลที่คาดว่าจะมีผลกระทบต่อตัวข้อมูลที่ต้องการจะทราบ จากตัวอย่างเป็นการเก็บข้อมูลของหุ้นหรือหลักทรัพย์ซึ่งเป็นตารางส่วนหนึ่งของจำนวนปัจจัยทั้งหมดที่น่าจะมี ในการจะประเมินราคาของ

หลักทรัพย์ใดๆนั้นจำเป็นที่จะต้องเก็บ ราคาและปริมาณการเสนอซื้อ/ขาย โดยข้อมูลดังกล่าวนั้นเป็นข้อมูลในรูปแบบตัวเลขอยู่แล้ว แต่ข้อมูลบางประเภท เช่น เวลานั้นสมควรที่จะแบ่งย่อยเวลาจริงของแต่ละวันที่ทำการซื้อขายออกเป็นช่วงเช่นแบ่งเวลาการซื้อขายของวันหนึ่งๆที่ตลาดเปิดทำการออกเป็น 200 ช่วง แล้วให้แต่ละช่วงแทนแต่ละเวลา อีกทั้งค่าต่างๆที่อาจมีผลกับความเปลี่ยนแปลงของราคา เช่นปัจจัยภายนอกก็ควรนำมาใช้ในการเก็บข้อมูลด้วย ซึ่งความแม่นยำในการทำนายค่าต่างๆนั้นขึ้นกับประสบการณ์ และความสามารถในการสะสมข้อมูลของผู้ใช้ในแต่ละรายไป โดยจะอาศัยข้อมูลพื้นฐานที่ผู้ใช้มีอยู่ในการพิจารณาค่าที่ควรจะเป็นเป็นหลักในการทำงาน

2.โปรแกรมวิเคราะห์ตะกร้าตลาด

หลักการคร่าวๆของการวิเคราะห์ตะกร้าตลาดนั้นก็คือการมองหาสินค้าที่ผู้ซื้อมักจะทำการซื้อคู่กัน หรือ ซื้อด้วยกันเป็นหัวใจ จากมุมมองของการนำไปใช้ประโยชน์นั้นผู้ใช้จำเป็นที่จะต้องมีรูปแบบข้อมูลที่เข้ากับโปรแกรมก่อนแล้วผู้ใช้ต้องเข้าใจหลักการของ ซัพพอร์ต และคอนฟิเดนส์ก่อนจึงจะใช้ประโยชน์จากโปรแกรมได้

อินพุท :ฐานข้อมูลธุรกรรมหรือทรานแซคชัน ค่าซัพพอร์ตต่ำสุด
เอาท์พุท :ค่าของซัพพอร์ต ค่าของคอนฟิเดนส์ ในการจับกลุ่มของข้อมูลประเภทต่างๆ

ตัวอย่างการประยุกต์ใช้งานโปรแกรมวิเคราะห์ตระกร้าตลาด

การใช้งานจะเริ่มที่การป้อนข้อมูลฐานข้อมูลและค่าซัพพอร์ตต่ำสุดเมื่อโปรแกรมวิเคราะห์และคำนวณแล้วเสร็จก็จะได้ค่าและกลุ่มของสินค้าที่มีค่าซัพพอร์ตและคอนฟิเดนส์รายงานไว้ ในการประเมินค่าความเป็นไปได้ในการใช้ประโยชน์จากค่าซัพพอร์ตและคอนฟิเดนส์นั้นผู้ใช้จำเป็นที่จะต้องพิจารณาทั้งสองค่าเพราะซัพพอร์ตคือปริมาณธุรกรรมจริงที่เกิดขึ้นและคอนฟิเดนส์คือความน่าเชื่อมั่นในการเลือกข้อมูลกลุ่มนั้นว่าเป็นไปได้กี่เปอร์เซ็นที่การเกิดขึ้นนั้นจะเป็นจริง หากการวิเคราะห์พบค่าคอนฟิเดนส์ต่ำก็หมายความว่าการเกิดขึ้นของกรณีการจับกลุ่มแบบนั้นๆอาจจะไม่เป็นจริงก็ได้ แต่หากคอนฟิเดนส์สูงแต่ซัพพอร์ตต่ำก็อาจจะไม่เกิดประโยชน์เพราะเป็นปริมาณที่เกิดขึ้นน้อย ฉะนั้นค่าทั้งสองจำเป็นที่จะต้องพิจารณาควบคู่กันไป ในการจัดการส่งเสริมการขายสินค้าหรือการจัดวางหน้าร้านหากใช้ปริมาณทั้งสองเป็นตัวบ่ง ก็จะเกิดแนวทางที่มีข้อมูลเป็นพื้นฐานในการตัดสินใจ

Order ID	Product	Unit Price	Quantity	Discount
10248	Queso Cabrales	$14.00	12	0.00%
10248	Singaporean Hokkien Fried Mee	$9.80	10	0.00%
10248	Mozzarella di Giovanni	$34.80	5	0.00%
10249	Tofu	$18.60	9	0.00%
10249	Manjimup Dried Apples	$42.40	40	0.00%
10250	Jack's New England Clam Chowder	$7.70	10	0.00%
10250	Manjimup Dried Apples	$42.40	35	15.00%
10250	Louisiana Fiery Hot Pepper Sauce	$16.80	15	15.00%
10251	Gustaf's Knäckebröd	$16.80	6	5.00%
10251	Ravioli Angelo	$15.60	15	5.00%
10251	Louisiana Fiery Hot Pepper Sauce	$16.80	20	0.00%
10252	Sir Rodney's Marmalade	$64.80	40	5.00%
10252	Geitost	$2.00	25	5.00%
10252	Camembert Pierrot	$27.20	40	0.00%
10253	Gorgonzola Telino	$10.00	20	0.00%
10253	Chartreuse verte	$14.40	42	0.00%
10253	Maxilaku	$16.00	40	0.00%
10254	Guaraná Fantástica	$3.60	15	15.00%
10254	Pâté chinois	$19.20	21	15.00%
10254	Longlife Tofu	$8.00	21	0.00%
10255	Chang	$15.20	20	0.00%
10255	Pavlova	$13.90	35	0.00%
10255	Inlagd Sill	$15.20	25	0.00%
10255	Raclette Courdavault	$44.00	30	0.00%
10256	Perth Pasties	$26.20	15	0.00%
10256	Original Frankfurter grüne Soße	$10.40	12	0.00%
10257	Schoggi Schokolade	$35.10	25	0.00%
10257	Chartreuse verte	$14.40	6	0.00%
10257	Original Frankfurter grüne Soße	$10.40	15	0.00%
10258	Chang	$15.20	50	20.00%
10258	Chef Anton's Gumbo Mix	$17.00	65	20.00%
10258	Mascarpone Fabioli	$25.60	6	20.00%
10259	Sir Rodney's Scones	$8.00	10	0.00%
10259	Gravad lax	$20.80	1	0.00%
10260	Jack's New England Clam Chowder	$7.70	16	25.00%
10260	Ravioli Angelo	$15.60	50	0.00%
10260	Tarte au sucre	$39.40	15	25.00%
10260	Outback Lager	$12.00	21	25.00%
10261	Sir Rodney's Scones	$8.00	20	0.00%
10261	Steeleye Stout	$14.40	20	0.00%

Order ID	Product	Unit Price	Quantity	Discount
10262	Chef Anton's Gumbo Mix	$17.00	12	20.00%
10262	Uncle Bob's Organic Dried Pears	$24.00	15	0.00%
10262	Gnocchi di nonna Alice	$30.40	2	0.00%
10263	Pavlova	$13.90	60	25.00%
10263	Guaraná Fantástica	$3.60	28	0.00%
10263	Nord-Ost Matjeshering	$20.70	60	25.00%
10263	Longlife Tofu	$8.00	36	25.00%
10264	Chang	$15.20	35	0.00%

ตารางที่ 7-7 แสดงข้อมูลตัวอย่างที่นำมาใช้ในงานวิเคราะห์ตะกร้าตลาด

ข้อมูลในตารางคือตัวอย่างของข้อมูลที่ใช้ในการวิเคราะห์ตะกร้าตลาดโดยข้อมูลที่ใช้งานจริงๆคือเฉพาะข้อมูล 2 แอทริบิวส์แรกเท่านั้นคือค่า TID และค่ารายชื่อสินค้าในการวิเคราะห์ตามหลักของ การวิเคราะห์ตะกร้าตลาด

3. โปรแกรมแผนภาพต้นไม้เพื่อการตัดสินใจ

อินพุท :ฐานข้อมูลทั่วๆไป

เอาท์พุท :แผนภาพต้นไม้เพื่อการตัดสินใจ

ในการจำแนกข้อมูลโดยการใช้แผนภาพเพื่อการตัดสินใจนั้นหากสร้างแผนภาพขึ้นมาโดยไม่มีหลักในการสร้างแล้วข้อมูลในแผนภาพที่ได้นั้นอาจจะเป็นแผนภาพที่มีลักษณะการแบ่งส่วนข้อมูลได้ไม่เหมาะสม กล่าวคือข้อมูลที่กระจายออกมาตามแอทริบิวส์ที่ใช้แบ่งนั้นอาจได้ข้อมูลที่อยู่ในคลาสเดียวกันไม่ได้มากเท่าที่ควรข้อมูลบางส่วนอาจจะปนกันคนละคลาสมากจนไม่เกิดประโยชน์จากการใช้แผนภาพในการแบ่ง ฉะนั้นการแบ่งส่วนข้อมูลควรจะมีหลักการรองรับซึ่งทำได้โดยใช้โปรแกรมสร้างแผนภาพต้นไม้ขึ้นมาก่อนแล้วจึงนำลักษณะแผนภาพที่ได้ไปทำการใช้งานจริงอีกที

ตัวอย่างการประยุกต์ใช้งานโปรแกรมสร้างแผนภาพต้นไม้

ข้อมูลจากฐานข้อมูลลูกค้าของบริษัทขายอุปกรณ์ไฟฟ้าแห่งหนึ่ง เป้าหมายเราอยากจะวิเคราะห์ดูว่าลูกค้าในกลุ่มใดที่เป็นกลุ่มที่ซื้อสินค้าคอมพิวเตอร์ จากตัวฐานข้อมูลเองนั้นจะเห็นได้ว่าลูกค้าแต่ละรายต่างก็มีแอทริบิวส์หรือคุณสมบัติต่างๆหลากหลายกันไป เช่น อายุ รายได้ เครดิต เป็นต้น การนำโปรแกรมสร้างแผนภาพต้นไม้เข้าช่วยจะทำให้เรามองเห็นภาพของธุรกิจเราได้ง่ายขึ้น ว่ากลุ่มคนที่ซื้อสินค้าเรานั้นจริงๆน่าจะมีคุณลักษณะอย่างไรบ้าง โดยฐานข้อมูลเป็นดังนี้

RID	age	income	student	credit_rating	buys_computer
1	20	high	no	Fair	No
2	20	high	no	excellent	No
3	35	high	no	Fair	Yes
4	50	medium	no	Fair	Yes
5	50	low	yes	Fair	Yes
6	50	low	yes	excellent	No
7	35	low	yes	excellent	Yes
8	20	medium	no	Fair	No
9	20	low	yes	Fair	Yes
10	50	medium	yes	Fair	Yes
11	20	medium	yes	excellent	Yes
12	35	medium	no	excellent	Yes
13	35	high	yes	Fair	Yes
14	50	medium	no	excellent	no

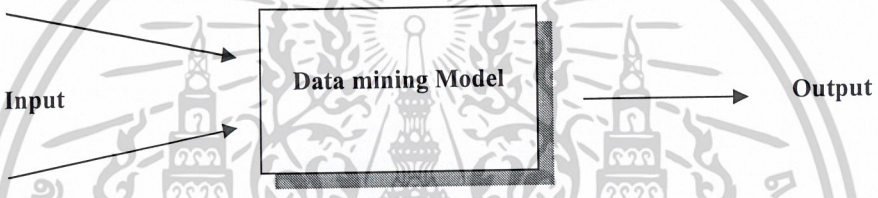
ตารางที่ 7-8 แสดงข้อมูลตัวอย่างสำหรับใช้ในโปรแกรมสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ

ผลลัพธ์สุดท้ายก็จะได้แผนภาพต้นไม้ที่แตกสาขาโดยความเหมาะสมตามทฤษฎี โดยผลที่ได้จะเป็นแผนภาพที่สามารถนำไปประยุกต์ใช้ในการจำแนกกลุ่มข้อมูลออกจากกันได้โดยการใช้ประโยชน์ยกตัวอย่าง เช่น การพิจารณาสินเชื่อของบุคคลโดยอาศัยหลักเกณฑ์ของแผนผังในการไล่ลำดับ โดยอาจไล่จาก อายุ รายได้ หรือข้อมูลอื่นๆที่มีใช้ในการประกอบการตัดสินใจอนุมัติสินเชื่อต่างๆ หรืออีกตัวอย่างหนึ่งคือการทำการสำรวจกลุ่มเป้าหมายในการพิจารณาจำเป็นที่จะต้องเลือกลักษณะกลุ่มเป้าหมายที่ต้องการโดยมีเงื่อนไขต่างๆกำหนดเอาไว้ ส่วนการไล่ว่าข้อมูลใดควรนำขึ้นมาพิจารณาก่อน ข้อมูลใดควรนำมาพิจารณาทีหลังนั้นจะเลือกทำตามแผนผังที่โปรแกรมสร้างออกมาให้จะช่วยให้ได้ผลลัพธ์ที่ดีที่สุด

บทที่ 8
โปรแกรมต้นแบบ

1. แนวความคิดในการออกแบบ

เทคนิคของการทำดาต้าไมนิ่งซึ่งสามารถแยกย่อยออกได้เป็นหลายวิธีแล้วแต่ประเภทของงาน ยกตัวอย่างเช่น แผนภาพต้นไม้ การใช้วิธีการทางสถิติเช่นการใช้สมการถอยเชิงเส้น การวิเคราะห์ความสัมพันธ์โดยใช้อัพไพรออรีอัลกอริทึม การใช้ทฤษฎีนิวรอนเน็ตเวิร์ก การใช้จีเนติกอัลกอริทึม เป็นต้น ขึ้นอยู่กับลักษณะงาน ซึ่งเราควรเลือกให้เหมาะสม



รูปที่ 8-1 โมเดลกับอินพุทและเอาท์พุท

การเลือกใช้โมเดลมีสิ่งที่ควรพิจารณาคำนึงถึง

1. ความอ่อนไหวต่อรูปแบบของอินพุท
2. ความสามารถในการอธิบายที่มาของเอาท์พุท
3. ความยากง่ายในการประยุกต์ใช้งาน

แต่กระบวนการหรือวิธีการหลายอย่างนั้นเป็นวิธีการที่ยากที่จะนำไปทำความเข้าใจในระดับผู้ที่จะนำผลลัพธ์นั้นไปใช้งาน ในทางธุรกิจนั้นผู้ที่ใช้งานจริงนั้นมักเป็นกลุ่มคนทางธุรกิจ เช่นฝ่ายบริหารการตลาด หรือฝ่ายจัดการสินค้าคงคลัง ซึ่งหากเราจะให้กลุ่มคนเหล่านี้ใช้งานกระบวนการหรือเทคนิคที่เลือกมานั้นต้องเป็นวิธีที่ทำความเข้าใจได้ง่ายมีความซับซ้อนในการใช้งานน้อย ดังนั้นการเลือกใช้โมเดลที่ค่อนข้างซับซ้อนและเข้าใจยากบางครั้งจึงไม่เหมาะสมเพราะจะทำให้ผู้ใช้เกิดความยุ่งยากในการทำความเข้าใจดังที่กล่าวมาแล้ว ดังนั้นเทคนิคที่นำมาใช้ในโครงงานจึงเป็นเทคนิคแบบพื้นฐานที่ง่ายต่อการอธิบายหรือฝึกอบรมต่อเจ้าหน้าที่ในการเรียนรู้และใช้งาน สามารถอธิบายความเป็นไปของตัวแปรที่มีส่วนเกี่ยวข้องอินพุทและเอาท์พุทได้และมีประสิทธิภาพใช้งานกันในธุรกิจทั่วไป

สำหรับวิธีการที่เลือกใช้ซึ่งจะครอบคลุมลักษณะการใช้งานได้จริงในงานทางธุรกิจ 3 ประเภท โดยแบ่งโปรแกรมเป็น 3 ส่วนแยกการทำงานจากกันคือ

1. การทำนายค่าโดยใช้ข้อมูลที่มีอยู่ โดยใช้สมการถดถอยเชิงเส้นหลายตัวแปร ซึ่งเป็นวิธีการพื้นฐานทางสถิติซึ่งเป็นวิธีการพื้นฐานที่น่าเชื่อถือ ซึ่งใช้ในการประมาณหรือทำนายค่า
2. การแยกประเภท โดยใช้แผนภาพต้นไม้ ซึ่งสามารถนำไปใช้แบ่งกลุ่มข้อมูลที่ต้องการเป็นกลุ่มๆ
3. การวิเคราะห์ความสัมพันธ์ของข้อมูล โดยใช้อัพไพรออรีอัลกอริทึม หรือการวิเคราะห์ตะกร้าตลาด ซึ่งจะอธิบายความสัมพันธ์กันในตัวข้อมูล

โดยทำการพัฒนาบน Visual Basic 6.0 กับฐานข้อมูล Microsoft Access

2. ขั้นตอนและเทคนิคการประมวลผลของโปรแกรม

2.1 โปรแกรมสมการถดถอยเชิงเส้น

ลักษณะการประมวลตามทฤษฎีจะต้องเป็นไปตามสมการ

$$
\begin{aligned}
y_1 &= a_0 + a_1x_{11} + a_2x_{21} + \ldots + a_kx_{k1} \\
y_2 &= a_0 + a_1x_{12} + a_2x_{22} + \ldots + a_kx_{k2} \\
y_3 &= a_0 + a_1x_{13} + a_2x_{23} + \ldots + a_kx_{k3} \\
&\vdots \\
y_n &= a_0 + a_1x_{1n} + a_2x_{2n} + \ldots + a_kx_{kn}
\end{aligned}
$$

ซึ่งเขียนเป็นตัวแบบได้ดังนี้

$$
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \vdots & x_{k1} \\ 1 & x_{12} & x_{22} & \vdots & x_{k2} \\ 1 & x_{13} & x_{23} & \vdots & x_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \vdots & x_{kn} \end{bmatrix} \times \begin{bmatrix} a_0 & a_1 & a_2 & ..a_k \end{bmatrix}
$$

ซึ่งเขียนเป็นรูปแบบของเมทริกได้ดังนี้

$$Y = X.A$$

ดังนั้น เราจะคำนวณหาค่า A = $\begin{bmatrix} a_0 & a_1 & a_2 & ..a_k \end{bmatrix}$ ได้ดังนี้

$$A = \left[X'X\right]^{-1}X'Y$$

แต่เนื่องจากการที่เราจะนำข้อมูลมาจากฐานข้อมูลขนาดใหญ่มาเก็บไว้เป็นเมตริกซ์ในการดำเนินงานด้วยโปรแกรมแบบนี้โดยตรงนั้น การคำนวณโดยใช้ตัวแบบที่ปรากฏนี้จะมีปัญหา เนื่องจากขนาดของ

เมตริกซ์ X และ Y จะใหญ่มาก (ขนาดเท่ากับจำนวน Transaction จากฐานข้อมูล) ดังนั้นจึงจำเป็นจะต้องปรับสูตรในการคำนวณใหม่เพื่อให้เหมาะสมกับการคำนวณจากฐานข้อมูลดังนี้คือ

$$[X'X] = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & : & \sum x_{ki} \\ \sum x_{1i} & \sum x^2_{1i} & \sum x_{1i}x_{2i} & : & \sum x_{1i}x_{ki} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x^2_{2i} & \sum x_{2i}x_{3i} & : \\ \sum x_{2i} & : & : & : & : \\ : & : & : & : & \sum x^2_{ki} \end{bmatrix}$$

และ

$$[X''Y] = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \\ \sum x_{3i}y_i \\ : \\ \sum x_{ki}y_i \end{bmatrix}$$

ซึ่งการคำนวณโดยเทคนิคดังกล่าวจะสามารถดำเนินการได้ภายในทรัพยากรของเครื่องที่มีจำกัดได้

2.2 โปรแกรมการสร้างแผนภาพเพื่อการตัดสินใจ

- เขียนโปรแกรมสร้างโหนดเดี่ยวเริ่มต้นไว้
- ถ้าตัวข้อมูลทั้งหมดเป็นคลาสเดียวกัน ให้โหนดเปลี่ยนมาเป็นปลายโหนดแล้วตั้งชื่อด้วยคลาสนั้น
- หาค่าจากฟังก์ชั่นในการวัดค่าซึ่งรีเทิร์นค่าอินฟอเมชันเกนเพื่อนำมาประเมินแอทริบิวส์ที่เหมาะสมที่สุด (ค่าที่มากที่สุด) มาใช้ในการเป็นตัวแยกข้อมูล
- ทำการสร้างกิ่งในแต่ละข้อมูลที่ทราบค่าของข้อมูลแอทริบิวส์ที่ใช้ทดสอบ
- จากนี้ตัวข้อมูลที่ทดสอบจะถูกแบ่งส่วนออกจากกัน แยกไปตามกิ่งของต้นไม้
- ทำลักษณะเดียวกันเช่นนี้ รีเคอซีฟไปจนกระทั่งหยุดเมื่อ

  1. ถ้าตัวข้อมูลทั้งหมดที่โหนดนั้นเป็นคลาสเดียวกัน
  2. หมดแอทริบิวส์ที่จะนำมาพิจารณาใช้แล้ว
  3. ถ้าข้อมูลที่ทดสอบหมดลง

2.3 โปรแกรมวิเคราะห์ตะกร้าตลาด

หลักการสร้างโปรแกรมดำเนินการในการประมวลผลข้อมูล แบ่งขั้นตอนการประมวลผลออกเป็น

- ทำการแสกนข้อมูลในฐานข้อมูลในลักษณะความถี่ของไอเทมแต่ละตัว
- เก็บค่าที่แสกนได้ลงบนอาเรย์สำหรับเก็บค่าความถี่
- ตรวจเทียบค่าความถี่ที่แสกนได้กับความถี่ต่ำสุดที่ผู้ใช้กำหนด
- ตรวจนับเฉพาะค่าที่มากกว่าค่าความถี่ต่ำสุดที่ผู้ใช้กำหนด
- เก็บค่าที่ผ่านเกณฑ์ข้างต้นลงอาเรย์
- ทำการจับคู่ทุกกรณีเฉพาะข้อมูลในข้อก่อนหน้า
- ทำการแสกนหาความถี่ของข้อมูลทุกคู่ที่จับไว้ในข้อก่อนหน้าบนฐานข้อมูลและบันทึกความถี่ไว้บนอาเรย์
- ข้อมูลทั้งหมดที่ได้ ณ จุดนี้จะสามารถนำมาคำนวณและแสดงผลค่าซัพพอร์ต คอนฟิเดนส์ของไอเทมประเภท 1 ไอเทม และ 2 ไอเทม ที่มีค่ามากกว่าค่าซัพพอร์ตต่ำสุดได้แล้ว
- ดำเนินการในการหาข้อมูลประเภท 3 ไอเทมและ 4 ไอเทมด้วยวิธีการเช่นเดียวกันกับข้างต้น
- ลักษณะการเก็บข้อมูลเพื่อการประมวลผลภายในโปรแกรมนั้นจะใช้อาเรย์ที่มีมิติเท่ากับจำนวนไอเทมที่จับกลุ่มในการเก็บเมื่อทำการแสกนเพื่อให้ลำดับก่อนหลังของไอเทมในรายการที่แสกนไม่มีผลกระทบต่อค่าของจำนวนไอเทม เมื่อทำการแสกนแล้วเสร็จจะต้องทำการนำข้อมูลในอาเรย์หลายมิตินั้นมารวมกันในส่วนที่มีลักษณะการจับคู่เหมือนกัน เช่น (A,B) กับ (B,A) จะต้องนำเอาผลของความถี่ทั้งสองกรณีมารวมกัน

	1,1	1,2	1,3	1,4
1,1		6	2	4
2,1	3		0	1
3,1	4	9		4
4,1	0	1	2	

ตาราง 8-1 แสดงลักษณะอาเรย์ที่ใช้ในการจัดเก็บการจับคู่แบบ 2 ไอเทมมีลักษณะเป็นอาเรย์ 2 มิติ

ข้อมูลของการจับคู่จะยกตัวอย่างเช่น (1,2) นั้นในกระบวนการอะไพรออรีจะไม่สนใจลำดับฉะนั้นค่า (2,1) ก็ให้ถือเป็นค่าเดียวกัน เมื่อทำการแสกนครบแล้วต้องนำมารวมกัน จากตัวอย่างจะได้ค่าเท่ากับ 6+3 =9 คือความถี่ที่ถือว่าแสกนได้จากกรณี (1,2) โดยการแสกนในการจับกลุ่มที่มีค่ามากกว่า 2 ไอเทมเช่น 3 ไอเทม นั้นก็ต้องใช้อาเรย์ที่มีมิติเท่ากับจำนวนไอเทมเช่น 3 และ 4 มิติในการจัดการกับตัวข้อมูลด้วย

ความสามารถต่างๆของโปรแกรมอื่นที่ขายกันในเชิงพาณิชย์

Find Laws (SKAT algorithm)

Find Dependencies (n-dimensional distributions)

Linear Regression (Stepwise and rule-enriched)

PolyNet Predictor (GMDH-Neural Net hybrid)

Cluster (Localization of anomalies)

Classify (Fuzzy logic modeling)

Decision Tree (Information Gain criterion)

Decision Forest

Market Basket Analysis (Association rules)

Memory Based Reasoning (k-NN + GA)

Discriminate (Unsupervised classification)

Summary Statistics (Data summarization)

Link Analysis (Visual correlation analysis)

Link Terms

Taxonomy Categorizer

Text Analysis (Semantic text analysis)

Text Categorizer

Text OLAP

บทที่ 9
ผลทดลองเปรียบเทียบ

ผลลัพธ์จากการประมวลผลข้อมูล (ชุดที่ใช้ในการทดสอบ กับโปรแกรมที่สร้าง) โดยใช้โปรแกรมสำเร็จรูป SPSS/PC เพื่อเปรียบเทียบกับ ผลลัพธ์ที่ได้จากการปฏิบัติงานของโปรแกรม

1. ผลลัพธ์จากการทำงานโดยใช้โปรแกรมสำเร็จรูป SPSS/PC

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.904[a]	.817	.813	3.39

a. Predictors: (Constant), FILTER_$, YEAR, ORIGIN, ACCEL, WEIGHT, CYLINDER, HORSE, ENGINE

รูปที่ 9-1 ผลจากโปรแกรม สำเร็จรูป SPSS/PC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20287.991	8	2535.999	221.183	.000[a]
	Residual	4551.857	397	11.466		
	Total	24839.848	405			

a. Predictors: (Constant), FILTER_$, YEAR, ORIGIN, ACCEL, WEIGHT, CYLINDER, HORSE, ENGINE

b. Dependent Variable: MPG

รูปที่ 9-2 ผลจากโปรแกรม สำเร็จรูป SPSS/PC

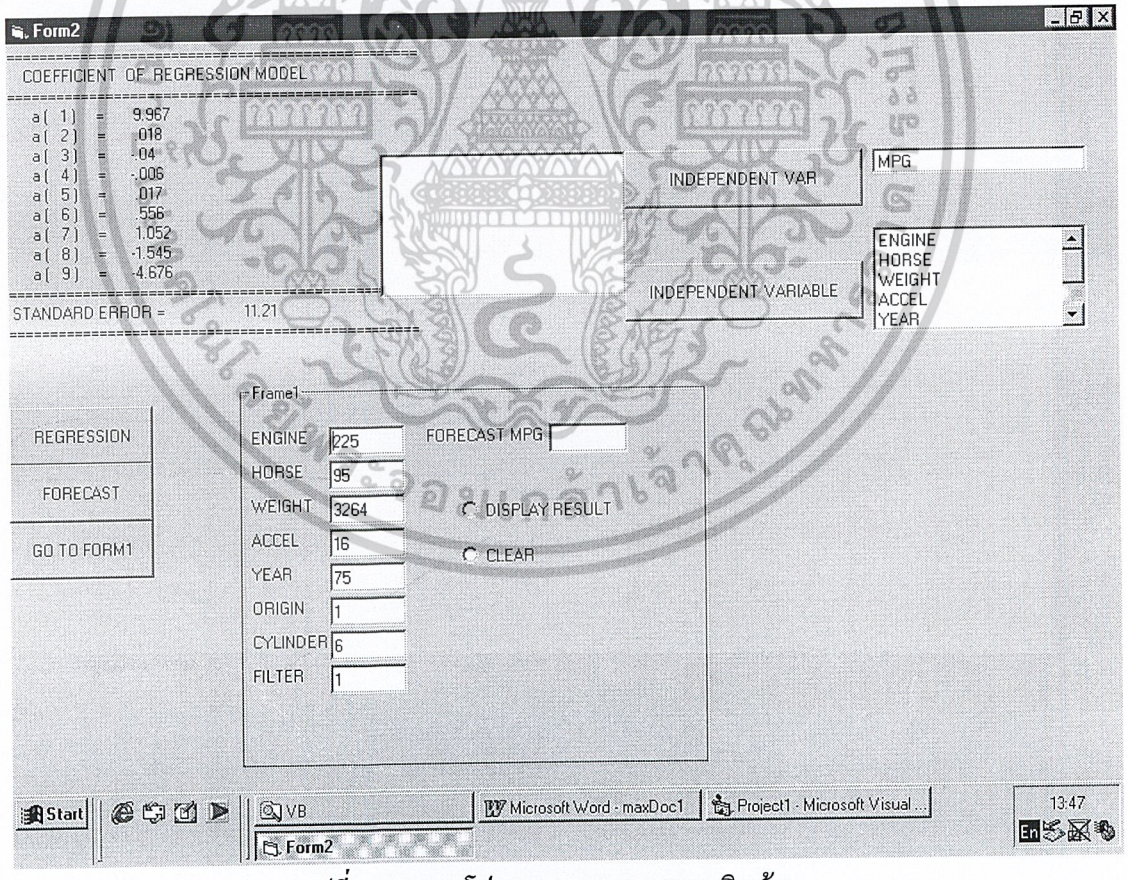
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.967	3.630		2.746	.006
	ENGINE	1.846E-02	.008	.248	2.406	.017
	HORSE	-4.020E-02	.014	-.197	-2.847	.005
	WEIGHT	-6.101E-03	.001	-.662	-9.359	.000
	ACCEL	1.721E-02	.098	.006	.175	.861
	YEAR	.556	.035	.377	16.066	.000
	ORIGIN	1.052	.281	.107	3.748	.000
	CYLINDER	-1.545	.367	-.340	-4.209	.000
	FILTER_$	-4.676	.888	-.264	-5.264	.000

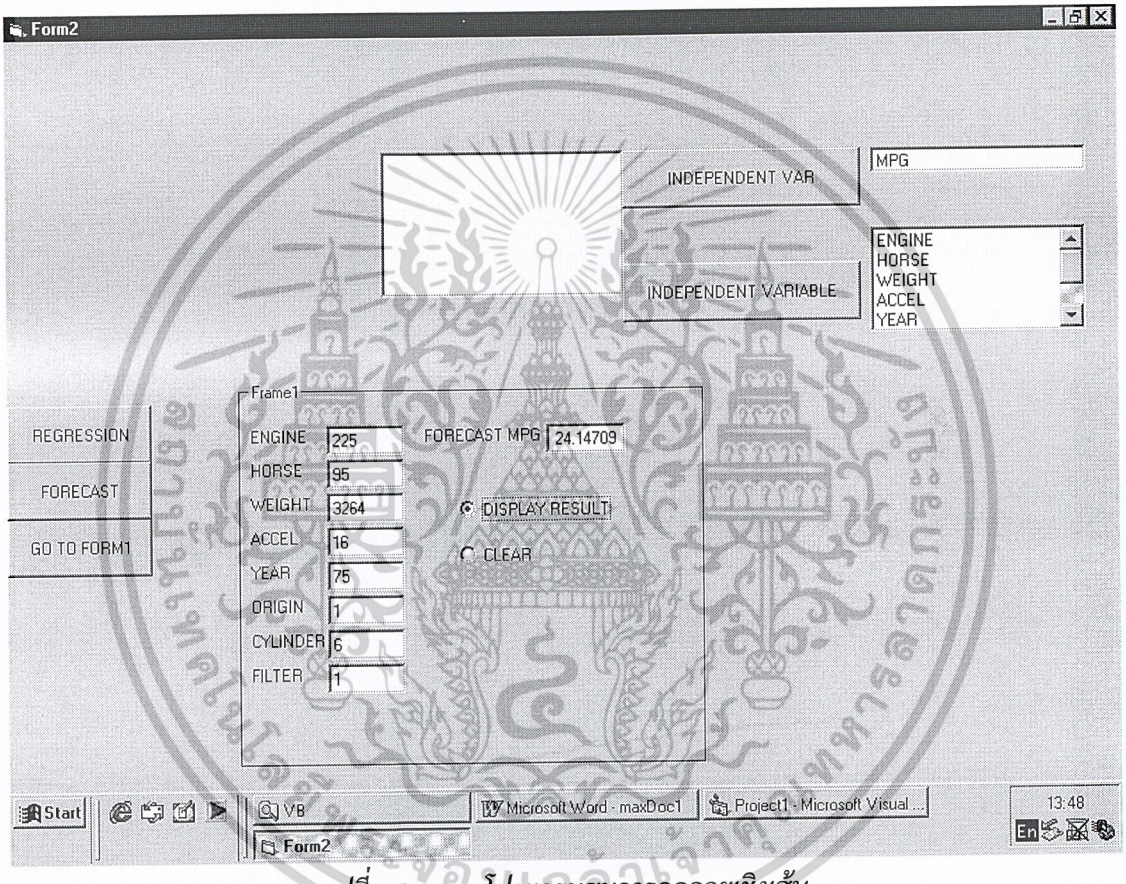
a. Dependent Variable: MPG

รูปที่ 9-3 ผลจากโปรแกรม สำเร็จรูป SPSS/PC

2. ผลลัพธ์จากการทำงานโดยใช้โปรแกรมสมการถดถอยเชิงเส้น



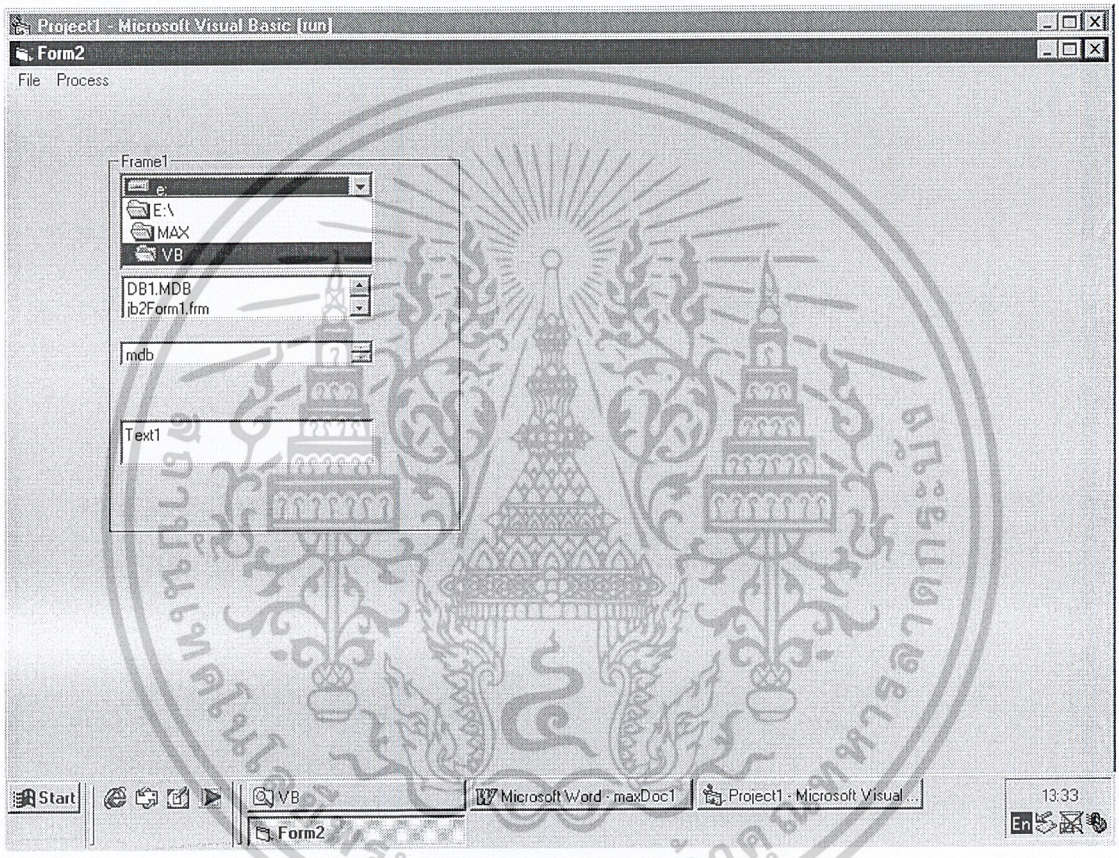
รูปที่ 9-4 ผลจากโปรแกรมสมการถดถอยเชิงเส้น



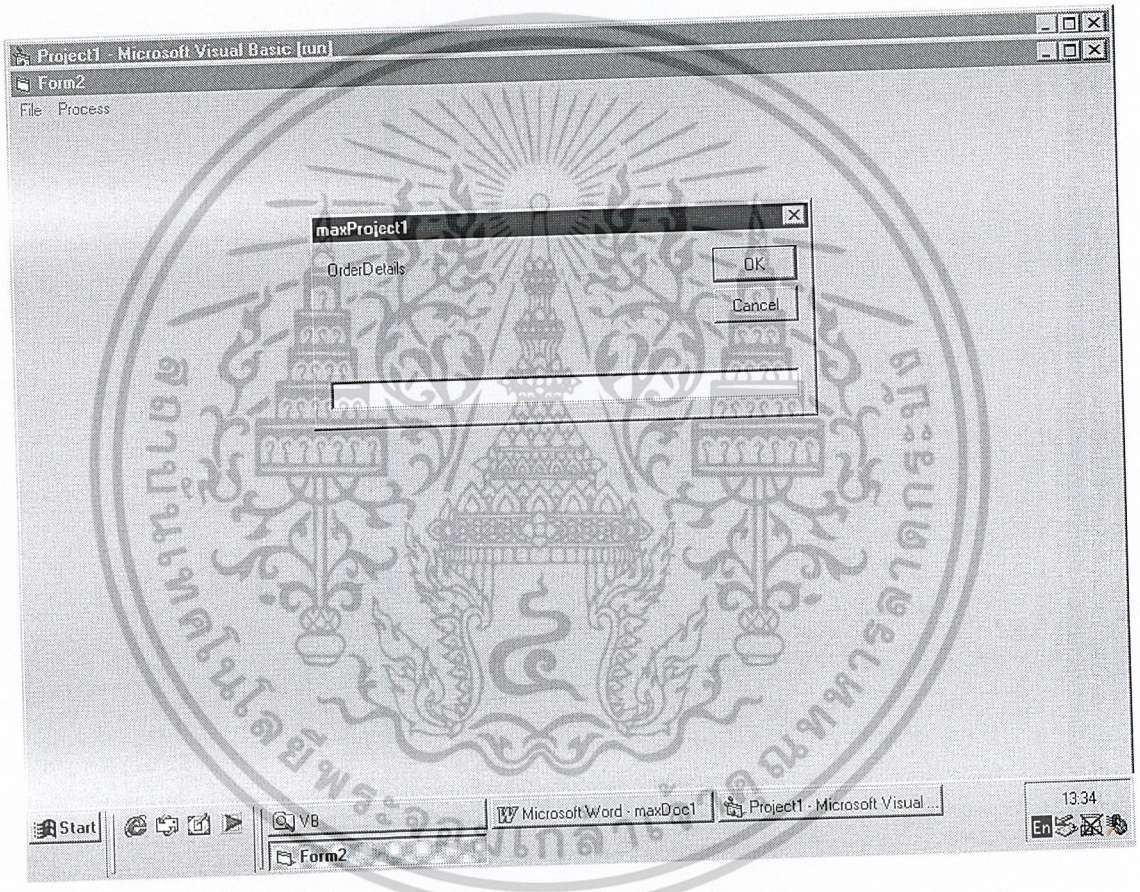
รูปที่ 9-5 ผลจากโปรแกรมสมการถดถอยเชิงเส้น

ภาคผนวก ก.
การทำงานโดยใช้โปรแกรมต้นแบบ

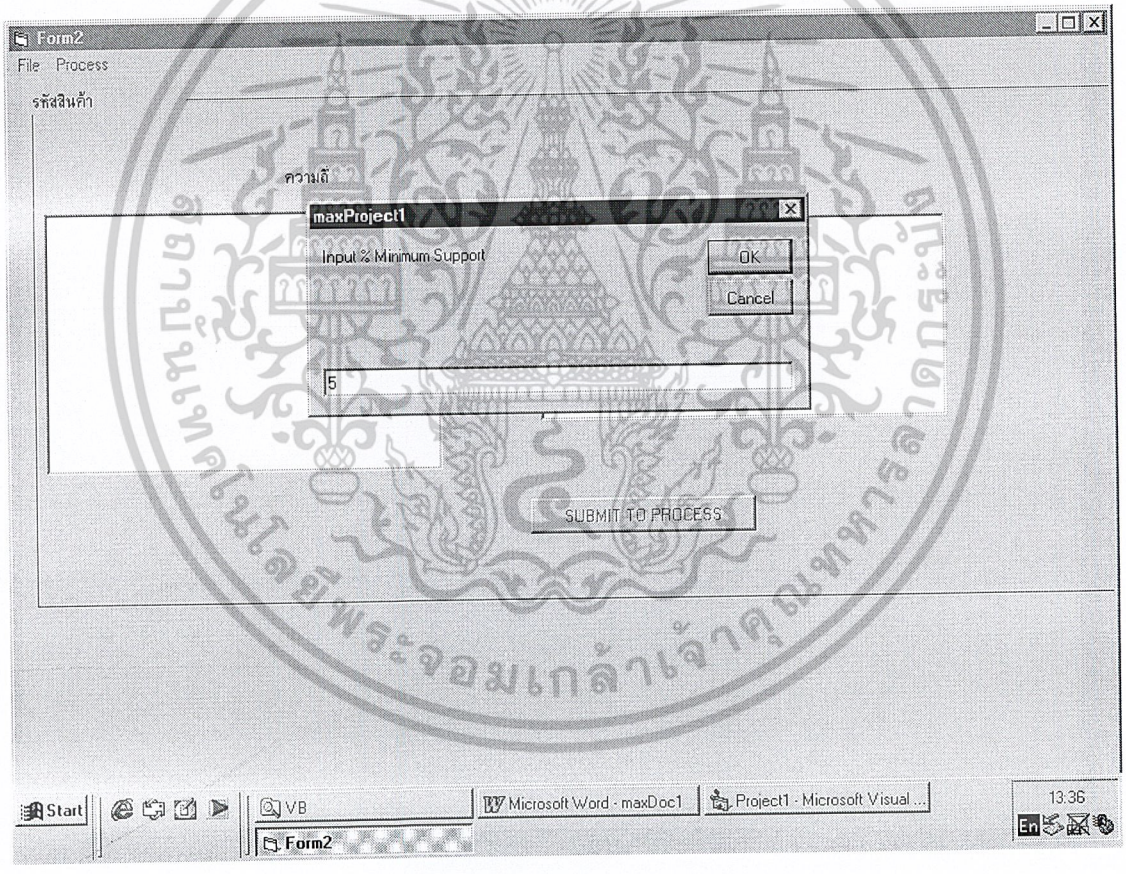
1. การทำงานโดยใช้โปรแกรมวิเคราะห์ตะกร้าตลาด



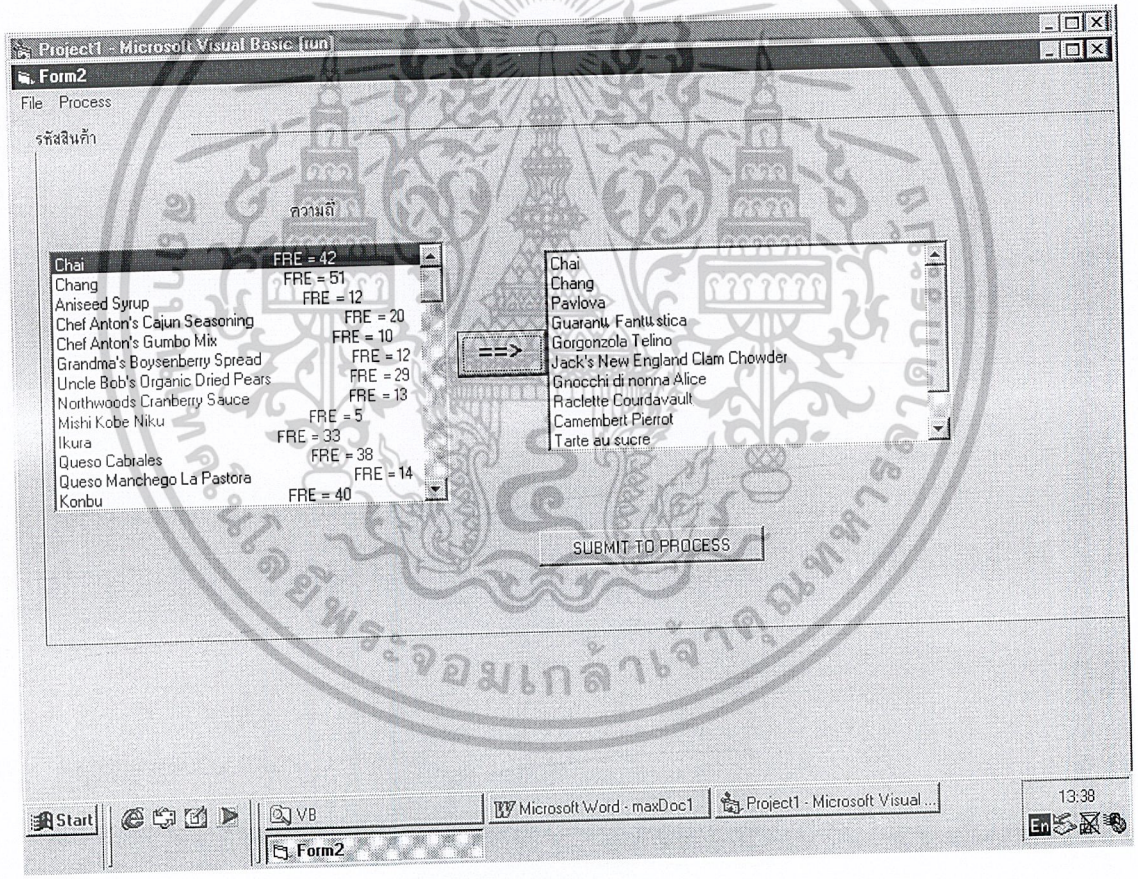
รูปที่ 10-1 เปิดไฟล์ฐานข้อมูล



รูปที่ 10-2 ป้อนอินพุทชื่อตาราง

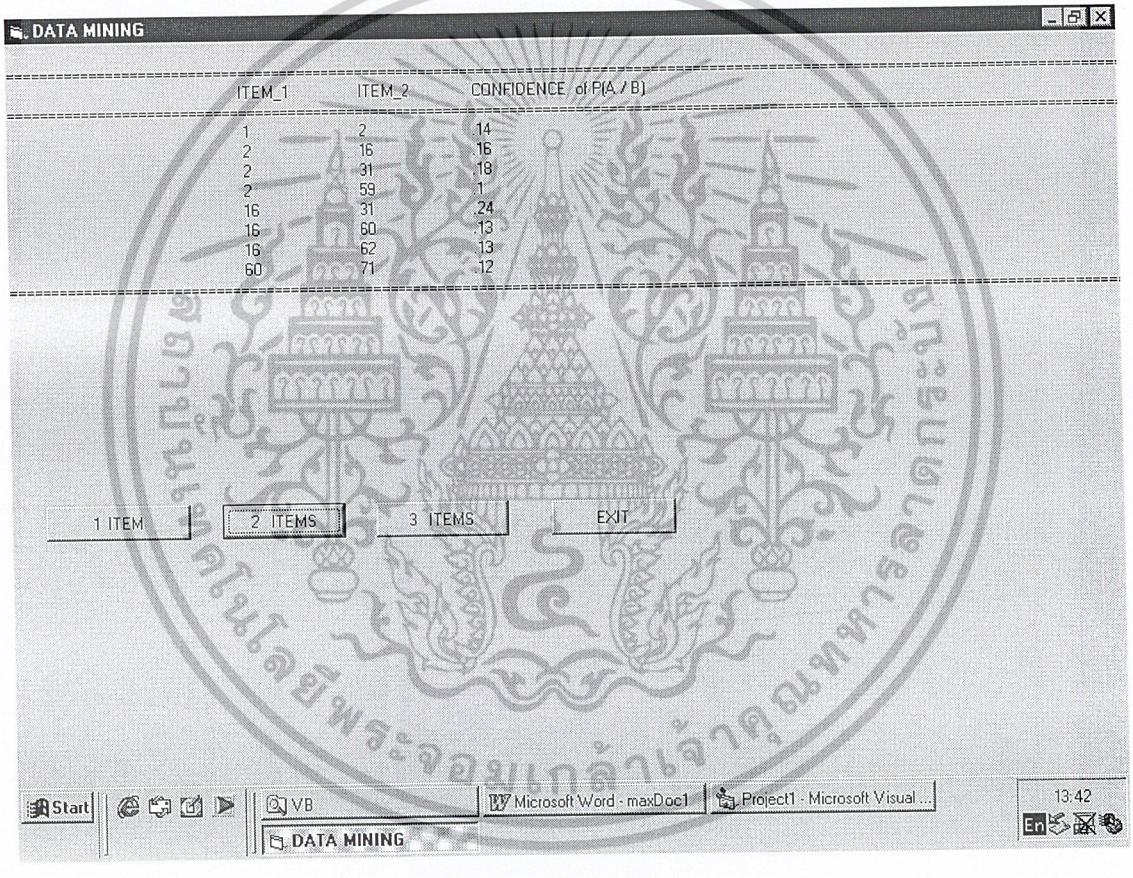


รูปที่ 10-3 ป้อนอินพุทค่าซัพพอร์ต



รูปที่ 10-4 แสดงค่าความถี่ของไอเทม

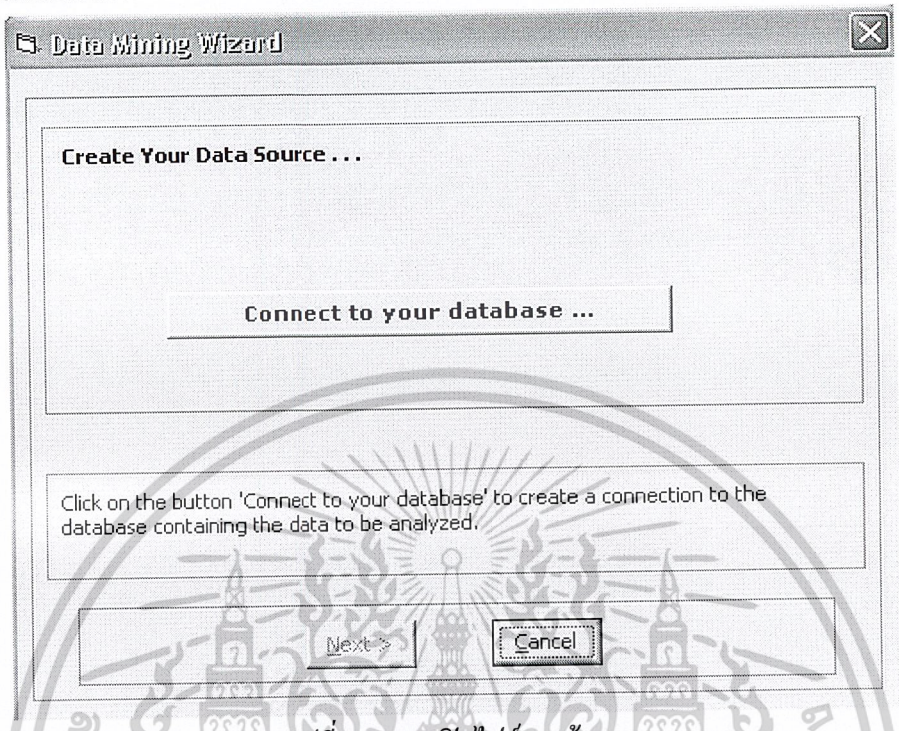
รูปที่ 10-5 การทำงานในการคำนวณ 1 ไอเทม



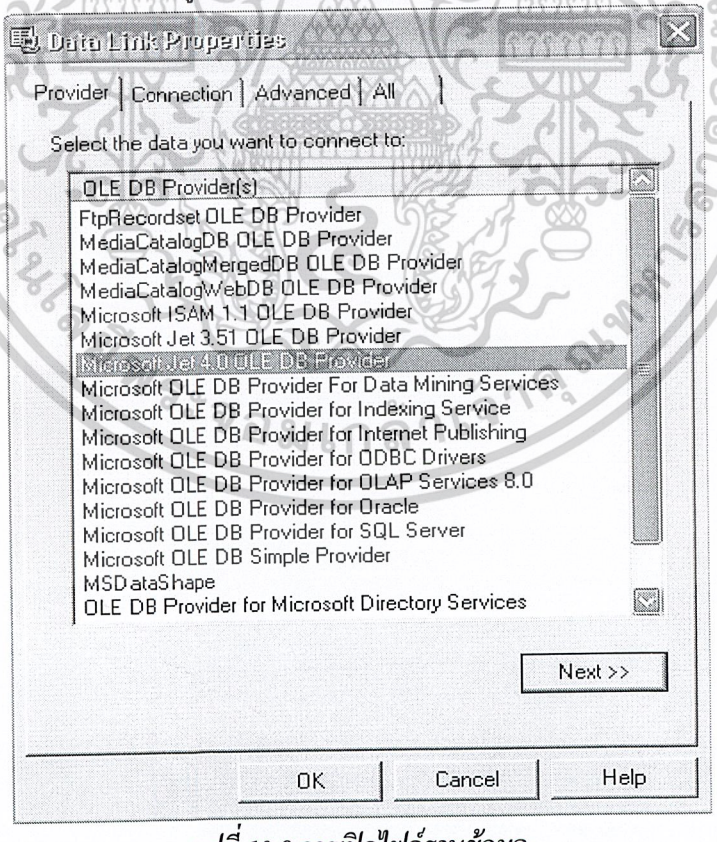
รูปที่ 10-6 การทำงานในการคำนวณ 2 ไอเทม

รูปที่ 10-7 การทำงานในการคำนวณ 3 ไอเทม

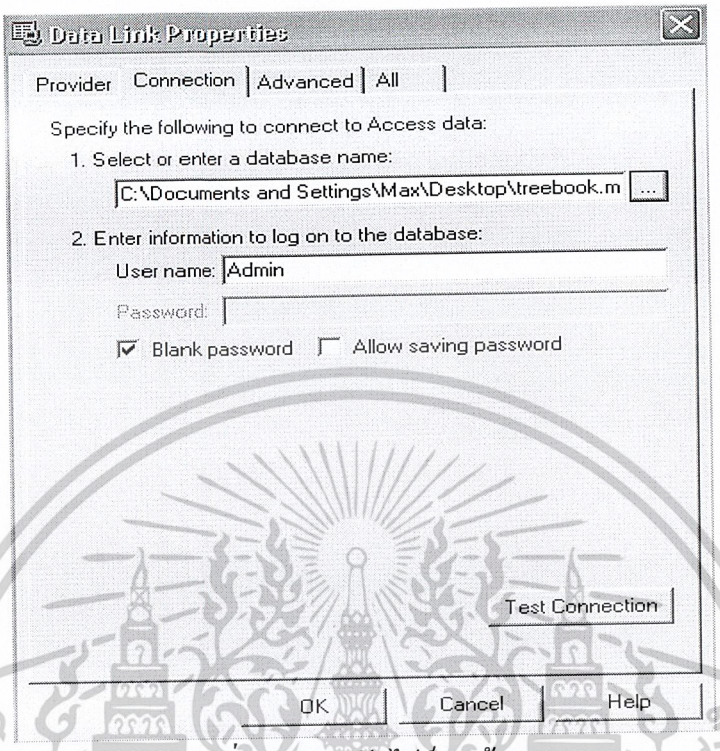
2. การทำงานโดยใช้โปรแกรมสร้างแผนภาพต้นไม้



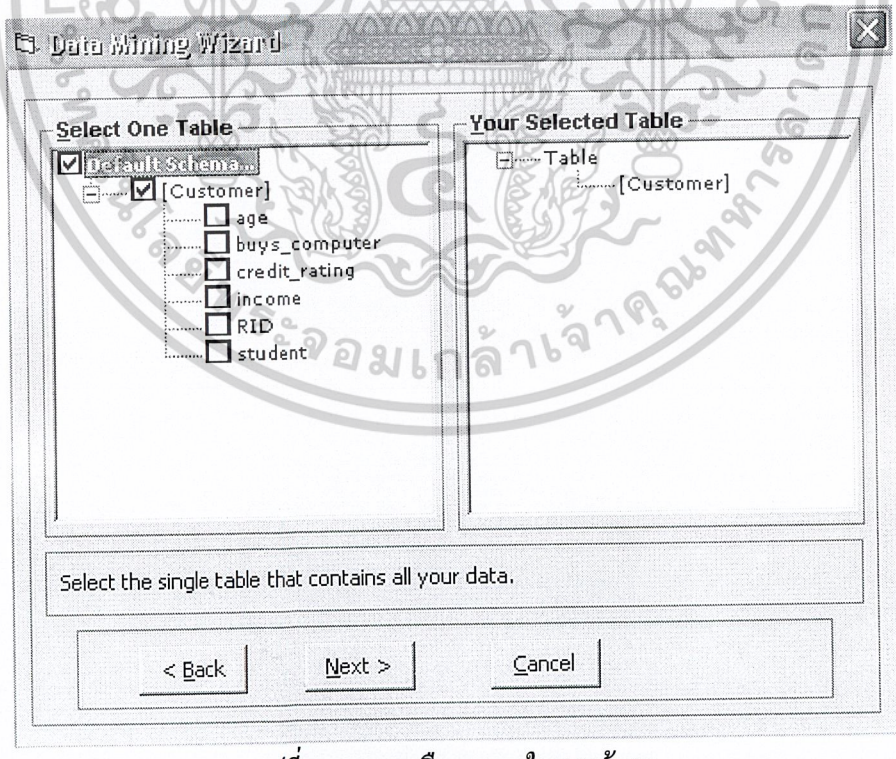
รูปที่ 10-8 การเปิดไฟล์ฐานข้อมูล



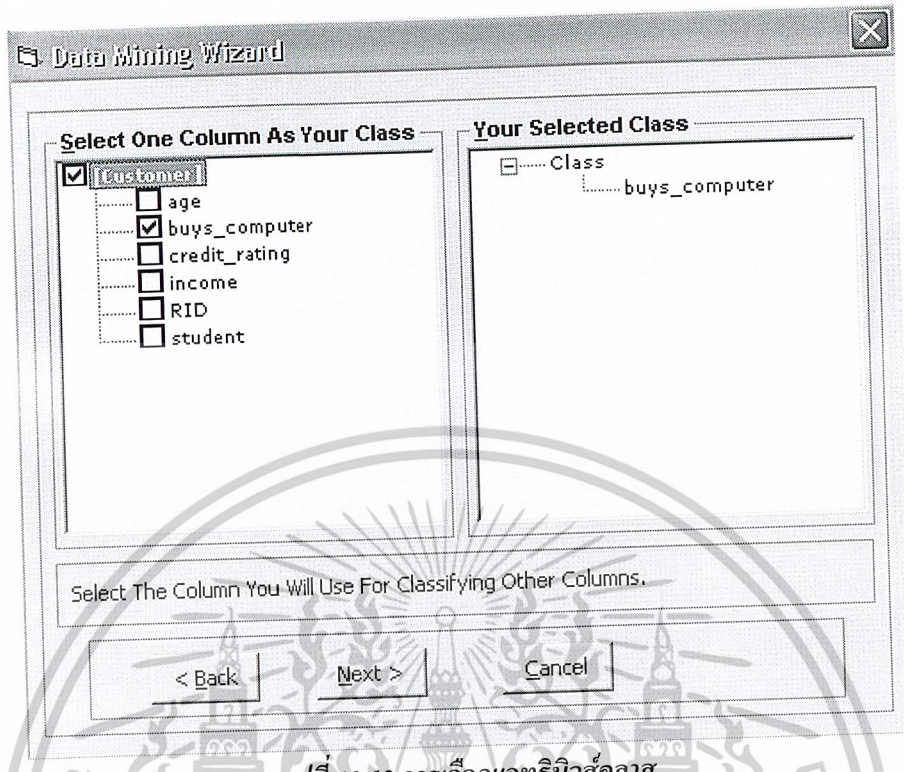
รูปที่ 10-9 การเปิดไฟล์ฐานข้อมูล



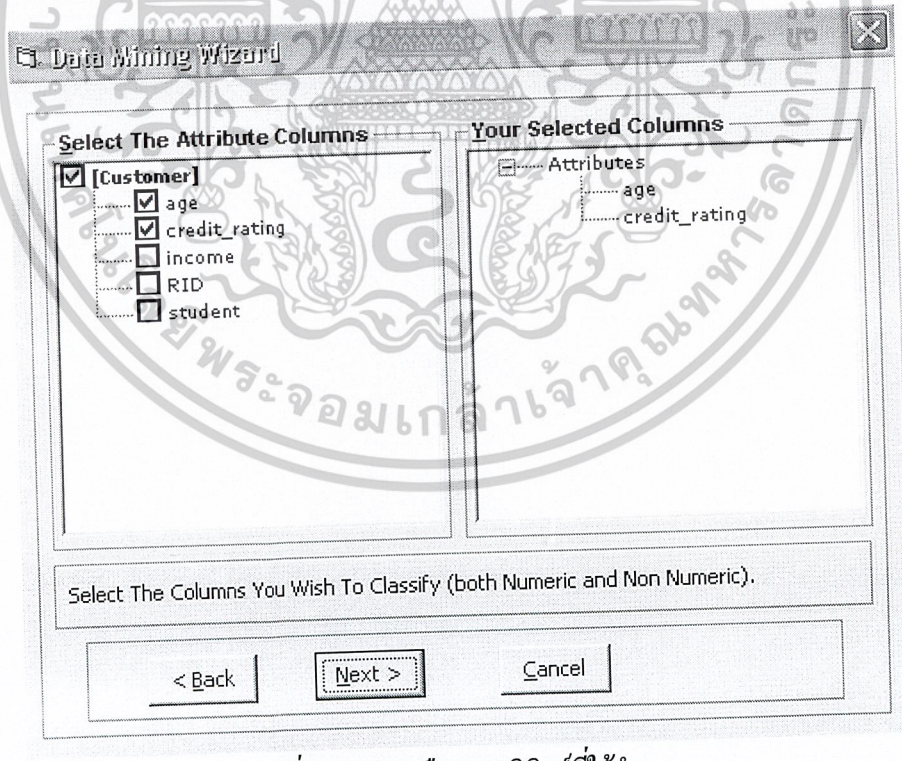
รูปที่ 10-10 การเปิดไฟล์ฐานข้อมูล



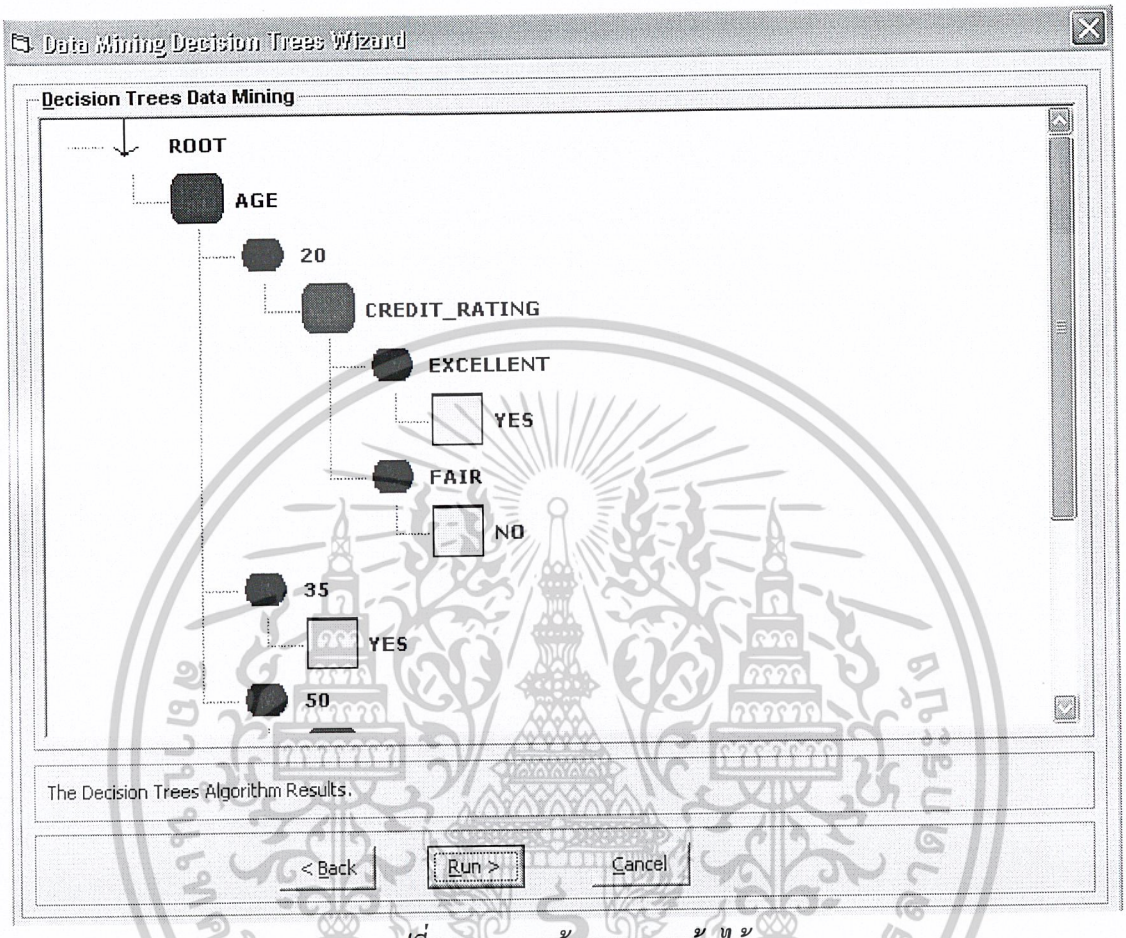
รูปที่ 10-11 การเลือกตารางในฐานข้อมูล



รูปที่ 10-12 การเลือกแอทริบิวส์คลาส

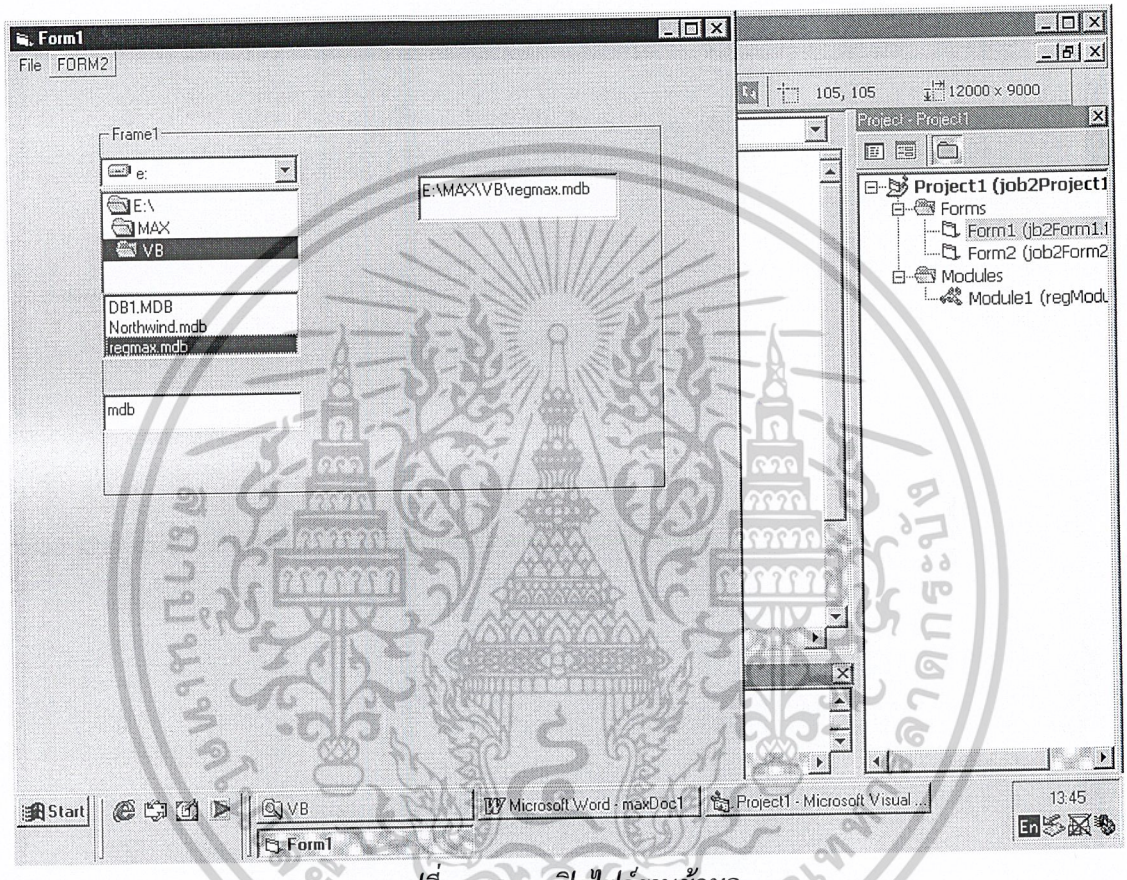


รูปที่ 10-13 การเลือกแอทริบิวส์ที่ใช้จำแนก

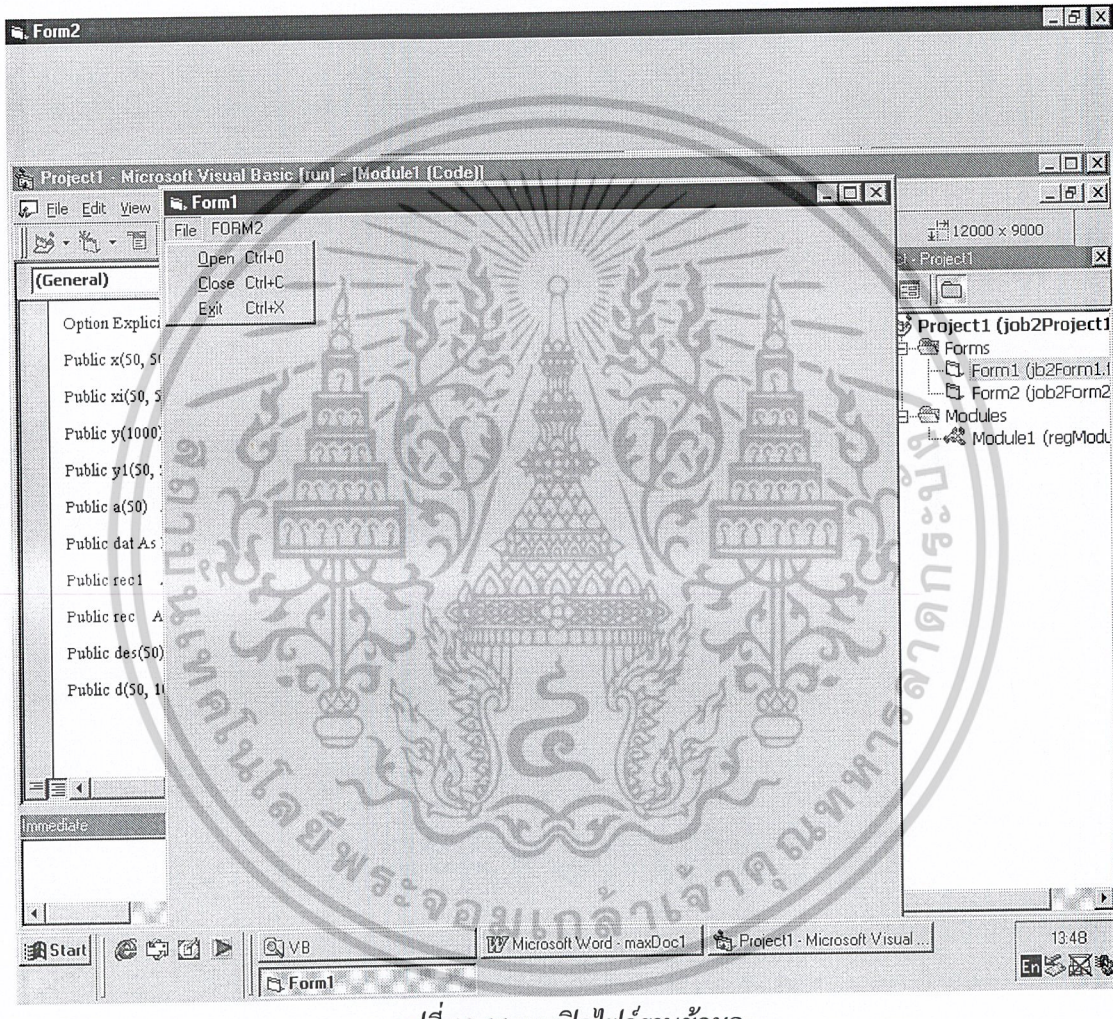


รูปที่ 10-14 การสร้างแผนภาพต้นไม้

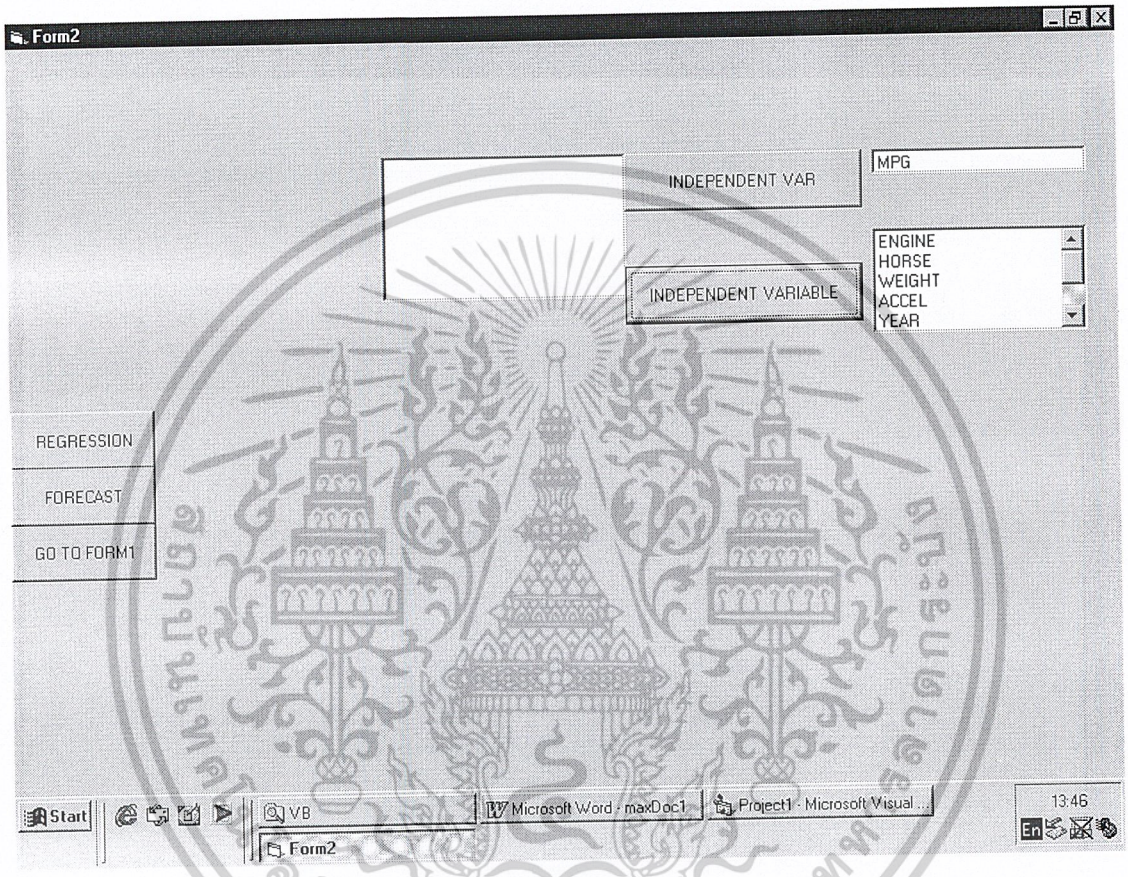
3. การทำงานโดยใช้โปรแกรมสมการถดถอยเชิงเส้น



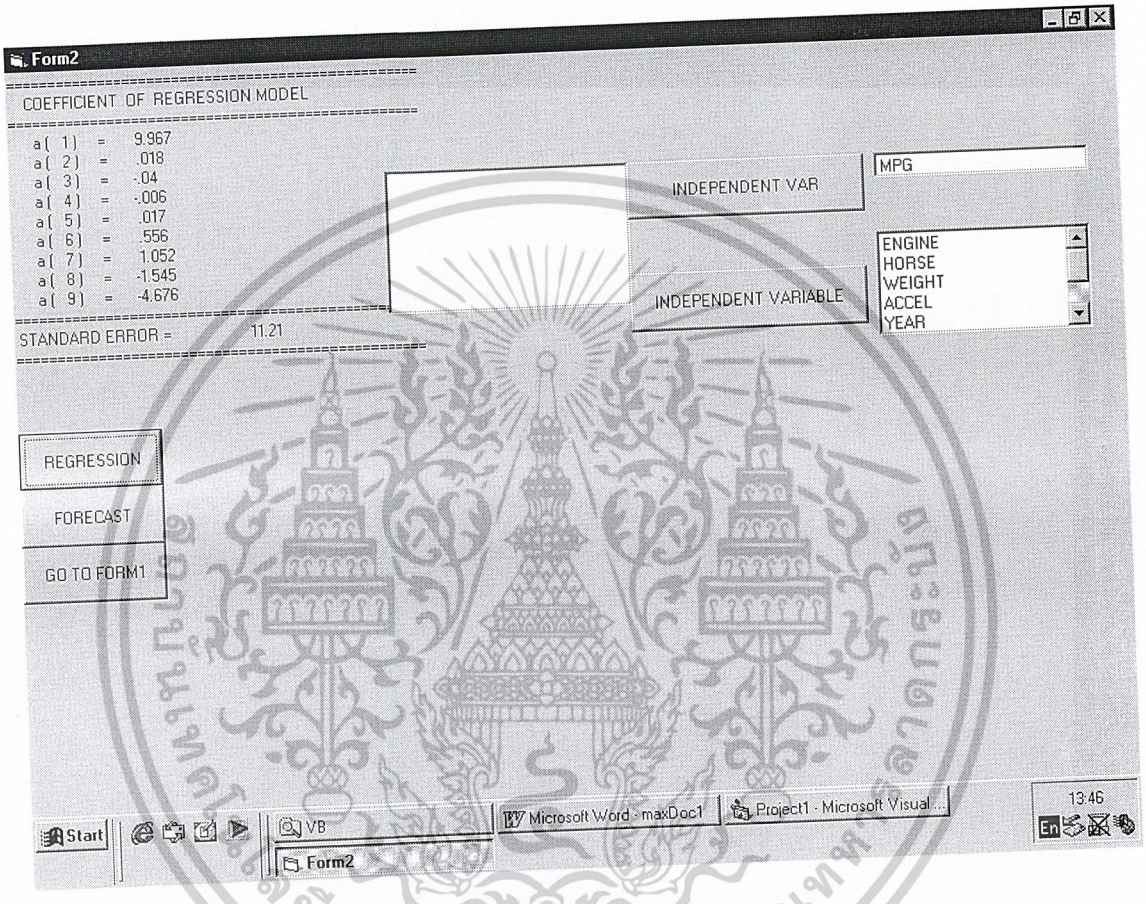
รูปที่ 10-15 การเปิดไฟล์ฐานข้อมูล



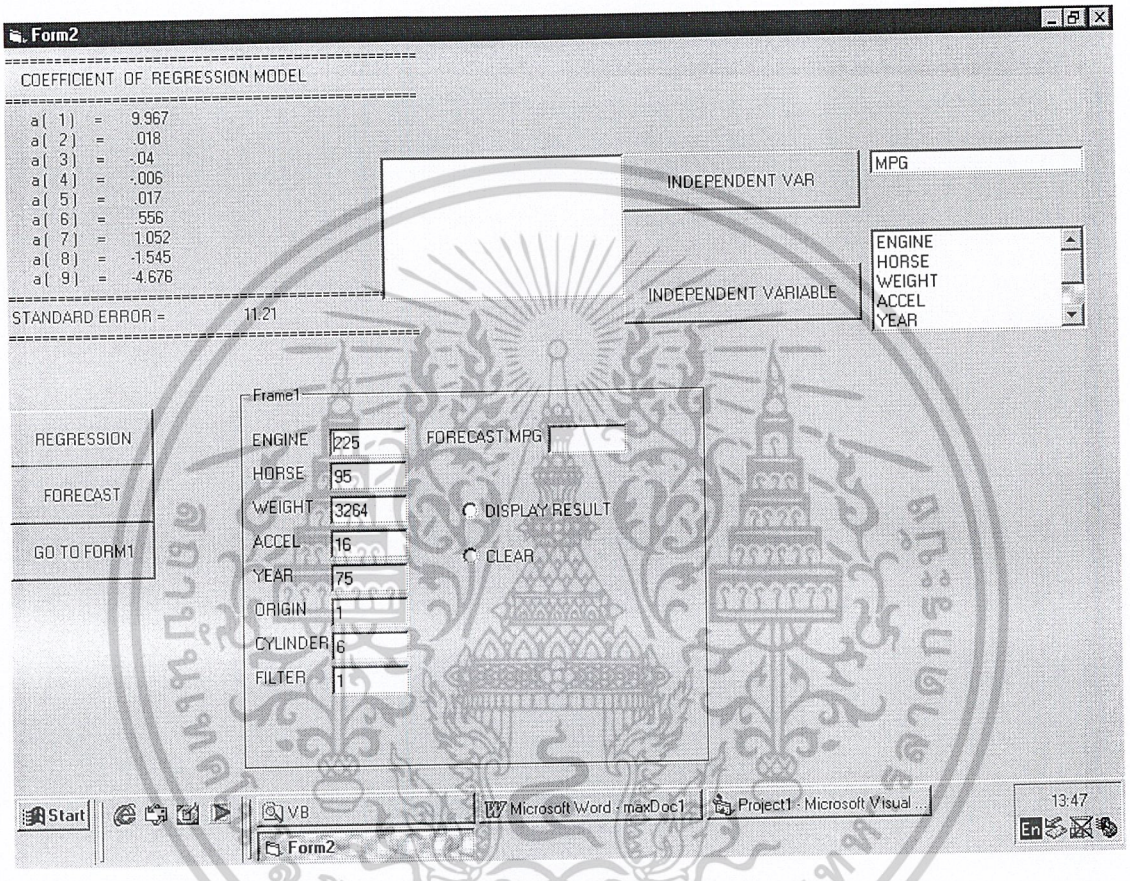
รูปที่ 10-16 การเปิดไฟล์ฐานข้อมูล



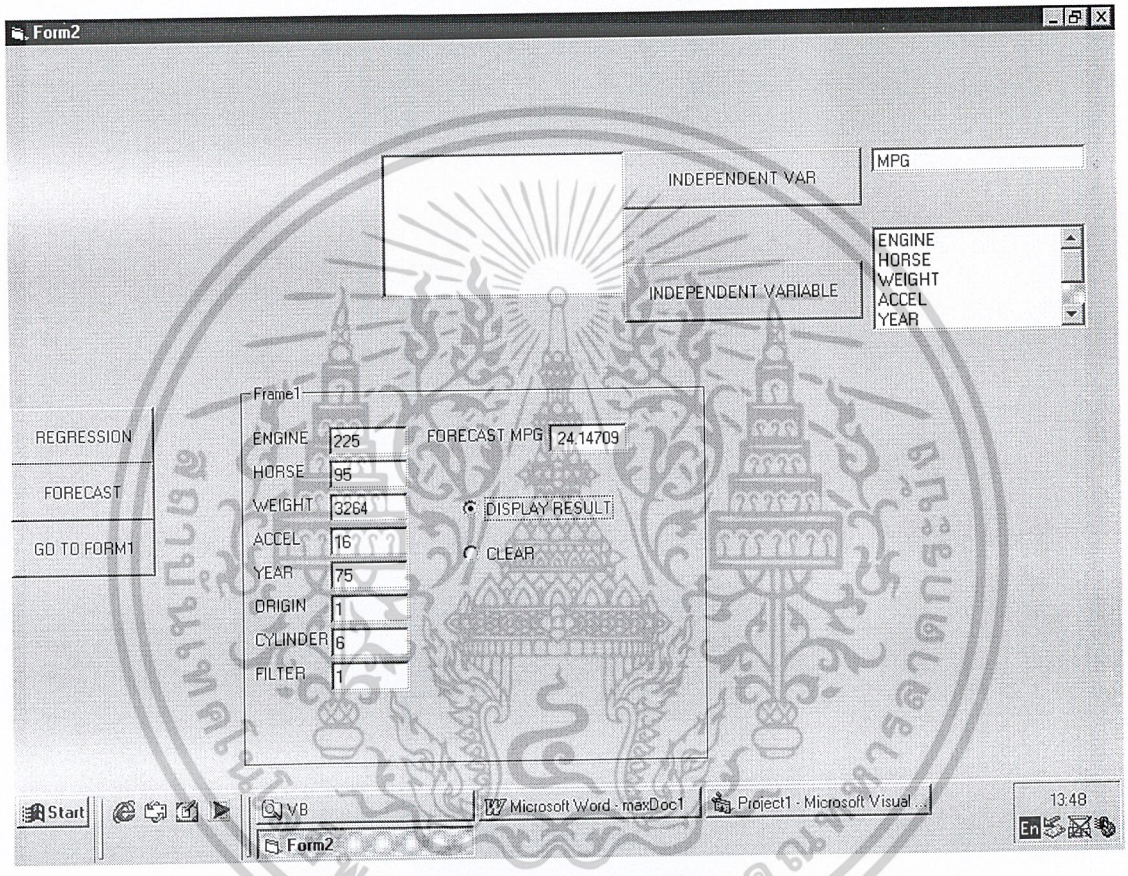
รูปที่ 10-17 การเลือกตัวแปรอิสระและตัวแปรอินดีเพนเดนท์



รูปที่ 10-18 แสดงผลค่าตัวแปร a และค่าแสตนดาร์ดเออเรอร์



รูปที่ 10-19 ป้อนอินพุทในการทำนายค่าที่ต้องการ



รูปที่ 10-20 แสดงผลลัพธ์ค่าที่คำนวณได้

บรรณานุกรม

[1] Zhengxin Chen (2001): "DATA MINING AND UNCERTAIN REASONING An Integrated Approach" , John Wiley & Sons, Inc.
[2] Jiawei Han, Micheline Kamber (2001): "Data Mining Concepts and Techniques" , Morgan Kaufmann Publishers
[3] James L. Buchanan, Peter R. Turner (1992): "NUMERICAL METHODS AND ANALYSIS" , McGraw-Hill, Inc.
[4] Terrence J. Akai (1994): "APPLIED NUMERICAL METHODS FOR ENGINEERS" , John Wiley & Sons, Inc.