การหาค่าความเสี่ยงของโรคเบาหวานโดยการใช้ทฤษฎีการทำเหมืองข้อมูล

ภาวิณี ธรรมเกตุ

รายงานนี้เป็นส่วนหนึ่งของการศึกษาวิชา

โครงงานวิทยาการคอมพิวเตอร์ 2 (2025499 -1)

ภาคเรียนที่ 1 ปีการศึกษา 2558

มหาวิทยาลัยราชภัฏมหาสารคาม

RISK FACTOR ANALYSIS OF DIABETES MELLITUS DIAGNOSIS BY DATA MINING

PAWINEE THAMMAKET

A Report Presented In Computer Science Project 2 Of The Requirement

Of The Course : 2025499 -1

Rajabhat Maha Sarakham University

1st Semester 2015

คณะกรรมการที่ปรึกษาประจำตัวนักศึกษา และคณะกรรมการสอบ ได้พิจารณารายงานการศึกษาโครงงานด้านโปรแกรมคอมพิวเตอร์ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาหลักสูตร ปริญญาวิทยาศาสตรบัณฑิตของมหาวิทยาลัยราชภัฏมหาสารคามได้

คณะกรรมการสอบ

..... **ประธานกรรมการ**
(อาจารย์รักถิ่น เหลาหา)

..... **กรรมการ**
(ผู้ช่วยศาสตราจารย์กาญจนา คำสมบัติ)

..... **กรรมการ**
(อาจารย์กิตติพงษ์ ชินสุข)

กลุ่มโปรแกรมวิชาคอมพิวเตอร์และเทคโนโลยีสารสนเทศ อนุมัติให้รับรายงานการศึกษาโครงงานของนักศึกษาฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร ปริญญาวิทยาศาสตรบัณฑิตของ มหาวิทยาลัยราชภัฏมหาสารคาม

..... **หัวหน้ากลุ่มโปรแกรมวิชาคอมพิวเตอร์**
(รองศาสตรตราจารย์ ดร. สิทธิชัย บุษหมั่น)

วันที่ 16 เดือน พฤศจิกายน พ.ศ. 2558

ชื่อเรื่อง การหาค่าความเสี่ยงของโรคเบาหวานโดยการใช้ทฤษฎีการทำเหมืองข้อมูล
นักศึกษา ภาวิณี ธรรมเกตุ
อาจารย์ที่ปรึกษา รักถิ่น เหลาหา
ปริญญา วิทยาศาสตรบัณฑิต(วท.บ.) **สาขาวิชา** วิทยาการคอมพิวเตอร์
มหาวิทยาลัย มหาวิทยาลัยราชภัฏมหาสารคาม **ปีที่พิมพ์** 2558

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษา ปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวานจาก ปัจจัยเสี่ยงด้านพฤติกรรมและโรคแทรกซ้อนของผู้ป่วยโรคเบาหวานโรงพยาบาลมหาสารคามจังหวัดมหาสารคาม เป็นการศึกษาเชิงวิเคราะห์และสรุปผลโดยการนำข้อมูลที่ได้จากโรงพยาบาลมหาสารคามมาทำการแปลงค่าแล้วกำหนดตัวทำการจัดกลุ่มปัจจัยเสี่ยงที่ทำให้เกิดโรคเบาหวาน จากนั้นนำค่าปัจจัยเสี่ยงมวิเคราะห์และคำนวณทางสถิติเชิงพรรณนา ได้แก่ แจกแจงความถี่ ร้อยละ ค่าเฉลี่ยส่วนเบี่ยงมาตรฐานในการวิเคราะห์ความสัมพันธ์ระหว่างปัจจัยต่างๆ

งานวิจัยนี้จัดทำขึ้นเพื่อทำการทดลองการความเสี่ยงของการเกิดโรคเบาหวาน โดยใช้โปรแกรมสำเร็จรูป Microsoft Excel เป็นเครื่องมือที่ช่วยในการวิเคราะห์โดยการนำเข้าซึ่งปัจจัยต่างๆ ที่กำหนดได้แก่ อายุ เพศ ประวัติการสูบบุหรี่ ประวัติการดื่มแอลกอฮอล์ ประวัติด้านพันธุกรรม โดยการวิเคราะห์สัมประสิทธิ์การจัดกลุ่มหาได้จากการนำตัวแปรเข้าไปในสมการแล้วพิจารณาสัมประสิทธิ์ของแต่ละตัวแปรและแบ่งตาม cluster Model จะแสดงค่า cluster Model โดยใช้ทฤษฎี K-Means ซึ่งได้ผ่านกระบวนการจัดกลุ่ม แบ่งออกได้ดังนี้ cluster 0 มีจำนวน 1,891 คน cluster 1 มีจำนวน 2,456 และ cluster 2 มีจำนวน 996 คน โดยมีจำนวนผู้ป่วยโรคเบาหวานทั้งหมด 5,343 คน

TITLE Risk Factor Analysis of Diabetes Mellitus Diagnosis by Data mining

AUTHOR Pawinee Thammaket

ADVISOR Rukthin laoha

EGREE Bachelor of Science (B.Sc.) **MAJOR** Computer Science

UNIVERSITY Rajabhat Maha Sarakham University **YEAR** 2015

ABSTRACT

This study aimed to investigate. Risk factors affecting the incidence of diabetes. Behavioral risk factors and complications of diabetes Mahasarakham Hospital. Mahasarakham province The study analyzed and summarized. The data were obtained from the conversion to the University Hospital. Then Define The clustering of risk factors that cause diabetes, then take on the risk analysis and calculation of descriptive statistics including frequency, percentage, mean, standard. In analyzing the relationship between different factors.

This research was conducted to test the risk of developing diabetes. UsingFinished Microsoft Excel is a tool that helps to analyze the import of various factors.The age, sex, smoking history. Alcohol History Genetic History By analyzing the clustering coefficient obtained from the variable into the equation. Then the coefficients of each variable and share-based cluster Model shows the cluster Model theory using K-Means clustering process, which has gone through. Classified as cluster 0 1891 amounted to a total of 2456 people cluster 1 and cluster 2 has a number of 996 people, with a total number of 5343 people with diabetes.

กิตติกรรมประกาศ

การศึกษานี้สำเร็จลุล่วงได้ด้วยดีเพราะได้รับความกรุณาอย่างสูงจาก อาจารย์รักถิ่น เหลาหา ที่ได้กรุณาให้คำปรึกษาแนะแนวทางการศึกษาการทำงานและเสียสละเวลาให้สำเร็จตามเป็นไปตามเป้าหมายที่วางไว้ด้วยความเอาใจใส่ท่านเป็นแบบอย่างในการวางแผนการดำเนินงานทั้งด้านการทำงานการตรงต่อเวลา ความรับผิดชอบต่องาน วิธีการสอน ผู้เขียนมี ความประทับใจ และขอกราบขอบพระคุณเป้นอย่างสูงไว้ ณ ที่นี้ด้วย ขอขอบพระคุณคณาจารย์ประจำหลักสูตรคอมพิวเตอร์สาขาวิทยาการคอมพิวเตอร์คณะวิทยาศาสตร์และเทคโนโลยีมหาวิทยาลัยราชภัฏมหาสารคามทุกท่าน ที่ประสิทธิ์สารทความรู้ อบรมสั่งสอนแนะแนวทางรวมถึงเจ้าหน้าที่ทุกคนที่คอยให้ความช่วยเหลือติดต่อประสานงานและขอบคุณเจ้าหน้าที่ศูนย์ข้อมูลโรงพยาบาลมหาสารคามที่ได้ให้ความอนุเคราะห์ข้อมูลมาเพื่อทำการวิจัยในครั้งนี้

ขอขอบคุณเพื่อนๆพี่ๆนักศึกษาหลักสูตรคอมพิวเตอร์ สาขาวิทยาการวิทยาการคอมพิวเตอร์ที่คอยเป็นกำลังใจให้คำปรึกษา และให้ความช่วยเหลือต่างๆแก่ผู้ทำการศึกษา

ภาวิณี ธรรมเกตุ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เบาหวาน เกิดจากความผิดปกติของร่างกายที่มีการผลิตฮอร์โมนอินซูลินไม่เพียงพอ อันส่งผลทำให้ระดับน้ำตาลในกระแสเลือดสูงเกินโรคเบาหวานจะมีอาการเกิดขึ้นเนื่องมาจากการที่ร่างกายไม่สามารถใช้น้ำตาลได้อย่างเหมาะสม ซึ่งโดยปกตินำ้ตาลจะเข้าสู่เซลล์ร่างกายเพื่อใช้เป็นพลังงานภายใต้การควบคุมของฮอร์โมนอินซูลินซึ่งผู้ที่เป็นโรคเบาหวานร่างกายจะไม่สามารถนำน้ำตาลไปใช้งานได้อย่างมีประสิทธิภาพ ผลที่เกิดขึ้นทำให้ระดับน้ำตาลในเลือดสูงขึ้นในระยะยาวจะมีผลในการทำลายหลอดเลือด ถ้าหากไม่ได้รับการรักษาอย่างเหมาะสม อาจนำไปสู่สภาวะแทรกซ้อนที่รุนแรงได้

เบาหวานเป็นโรคที่พบบ่อยและมีแนวโน้มเพิ่มมากขึ้น ในประชาชนไทยการค้นหาประชากรกลุ่มเสี่ยงต่อเบาหวาน จึงเป็นเรื่องสำคัญทั้งในด้านการรักษาและปรับเปลี่ยนพฤติกรรมการวิเคราะห์ปัจจัยเสี่ยงในผู้ป่วยโรคเบาหวานได้จากการนำข้อมูลพื้นฐานรายบุคคลมาเป็นเครื่องมือในการวิเคราะห์หาความเสี่ยง ประเมินความเสี่ยงซึ่งเป็นเครื่องมือที่ใช้ได้ง่ายในระดับปฐมภูมิและในประชากรทั่วไปในการประเมินตนเองเมื่อทราบปัจจัยที่ทำให้เกิดความเสียงจะช่วยกระตุ้นให้ผู้ที่มีความเสียงสูงเปลี่ยนแปลงพฤติกรรมได้อย่างตรงเป้าหมายแต่การวิเคราะห์ความเสี่ยงเป็นเรื่องที่ซับซ้อน และยากต่อการจำแนกปัจจัยต่างๆออกมาอย่างชัดเจนได้ นักวิจัยหลายท่านทั้งคนไทยและชาวต่างชาติได้พยายามที่จะหาค่าดัชนีความเสี่ยงต่อโรคเบาหวานเพื่อนำมาประเมินการเป็นเบาหวานในอนาคต ซึ่งส่วนใหญ่จะใช้วิธีการถดถอยโลจิสติก (Logistic Regression) เป็นเครื่องมือในการวิเคราะห์ซึ่งเป็นการวิเคราะห์ข้อมูลในลักษณะการแจกแจงเพื่อให้สอดคล้องกับสมมุติฐานของตัวแบบเท่านั้นซึ่งไม่สามารถนำไปสร้างแบบประเมินความเสี่ยงเบาหวาน

งานวิจัยฉบับนี้จึงได้มุ่งเน้นที่จะศึกษาและสร้างระบบที่สามรถจัดกลุ่มของประชากรที่เป็นโรคเบาหวานเพื่อที่จะให้ทราบถึงปัจจัยเสี่ยงใดบ้างที่ทำให้เกิดโรคเบาหวานและนำค่าของปัจจัยที่เสี่ยงได้มาวิเคราะห์และจัดกลุ่มผู้ป่วยว่ามีค่าความเสี่ยงที่จะเป็นโรคเบาหวานหรือไม่ ทั้งนี้หากพบว่าผู้ป่วยใดมีโอกาสที่จะเป็นโรคเบาหวานแพทย์ผู้รักษาสามารถตรวจคนไข้อย่างละเอียดอีกครั้งเพื่อให้ทันท่วงทีในการรักษา อีกทั้งยังสามารถช่วยผู้บริหารในการตัดสินใจกำหนดนโยบายในการที่จะป้องกันดูและดูแลประชากรให้รวดเร็วและถูกต้องขึ้นและยังสามารถเอื้อประโยชน์อย่างมากต่อประชาชนทั่วไปที่ไม่มีความรู้ความชำนาญด้านสาธารณสุขให้ตระหนักถึงการรักษาสุขภาพมากยิ่งขึ้น

1.2 วัตถุประสงค์ของการศึกษา

เพื่อวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวานโดยนำปัจจัยความเสี่ยงที่มีผลต่อประชากรจังหวัดมหาสารคามมาวิเคราะห์โดยใช้หลักการทำเหมืองข้อมูล

1.3 **ขอบเขตการศึกษา**

1.3.1 เพื่อวิเคราะห์ปัจจัยเสี่ยงที่ส่งผลต่อประชากรจังหวัดมหาสารคาม จึงนำข้อมูลของประชากรจังหวัดมหาสารคามมาวิเคราะห์ โดยนำปัจจัยดังต่อไปนี้ใช้ในการวิจัย

1.3.1.1 อายุ

1.3.1.2 เพศ

1.3.1.3 ความอ้วน

1.3.1.4 ความดันเลือด

1.3.1.5 พันธุกรรม

1.3.1.6 ประวัติเบาหวานในครอบครัว พ่อแม่ พี่น้อง

1.3.2 เพื่อสร้างตัวแบบที่เหมาะสมที่สุดในการวิเคราะห์ปัจจัยเสี่ยงโรคเบาหวานซึ่งตัวแบบที่นำมาใช้ในการจักกลุ่มของโรคเบาหวาน มีทั้ง หมด 2 ตัวแบบ คือ

1.3.2.1 การทำเหมืองข้อมูล (Data Mining)

1.3.2.2 วิธีการจัดกลุ่ม (Clustering)

1.4 สถานที่ทำวิจัย

1.4.1 โรงพยาบาลมหาสารคาม จังหวัดมหาสารคาม

1.4.2 ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏมหาสารคาม

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 สามารถประยุกต์ด้านเทคนิคการทำเหมืองข้อมูล เพื่อจำแนกผู้ป่วยโรงเบาหวานและการจัดกลุ่ม ข้อมูลโรคเบาหวาน

1.5.2 ศึกษาผลจากการแบ่งกลุ่มผู้ป่วยโรคเบาหวานเพื่อนำไปประยุกต์ใช้กับผู้ป่วยโรคอื่น

1.5.3 ผลจากการค้นคว้าทำให้สามารถทราบปัจจัยเสี่ยงต่อการเป็นโรคเบาหวาน อย่างมีนัยสำคัญทางสถิติ

1.6 เครื่องมือที่ใช้

1.6.1 ทะเบียนประวัติผู้ป่วยทั่วไปและเป็นโรคเบาหวาน

1.6.2 อุปกรณ์คอมพิวเตอร์

1.6.2.1 ฮาร์ดแวร์ (Hardware) ที่ใช้

1) HP Notebook , Genuine Intel (R) CPU T2250 @1.73 GHz,

2) 14”WXGA wide TFT LCD ,

3) 1.25 GB DDR

1.6.2.2 ซอฟต์แวร์ (Software) ที่ใช้

1) Microsoft Excel

2) Data Mining

3) Clustering

1.7 โครงสร้างการศึกษา

ในการศึกษานี้จะแบ่งเนื้อหาออกเป็น 5 บทซึ่งแต่ละบทจะมีรายละเอียดดังนี้

บทที่ 1 บทนำ เป็นการนำเสนอความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ ขอบเขตของการศึกษา สถานที่ทำวิจัย ประโยชน์ที่คาดว่าจะได้รับ รวมทั้งเครื่องมือที่ใช้ในการทดลอง

บทที่ 2 วรรณกรรมและผลงานวิจัยที่เกี่ยวข้อง จะอธิบายถึงความรู้พื้นฐานวรรณกรรมที่นำมาใช้ในการศึกษา และงานวิจัยที่นำมาอ้างอิงในการศึกษา ซึ่งในส่วนแรกจะกล่าวถึงแบบการเรียนรู้ของตัวแบบต่างๆและอธิบายถึงคุณลักษณะและวิธีการทำงานของแต่ละตัวแบบที่เกี่ยวข้องส่วนที่สองจะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับการศึกษาที่น่าสนใจทั้งของไทยและต่างประเทศซึ่งเป็นแรงบันดาลใจในการทำวิจัยครั้งนี้

บทที่ 3 การดำเนินงานวิจัย เป็นการนำเสนอที่เกี่ยวกับวิธีการและขั้นตอนในการดำเนินการ

บทที่ 4 ผลการวิจัยและอภิปรายผลเป็นการนำเสนอผลการวิเคราะห์และจัดลำดับความสำคัญของปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวานด้วยตัวแบบต่างๆ

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะเป็นส่วนที่แสดงให้เห็นถึงคุณค่าของงานวิจัยเพื่อให้เป็นฐานความรู้ในการนำไปวิเคราะห์และการจัดลำดับความสำคัญของปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวานและนำเสนอแนวทางในการปรับปรุงแก้ไขปัญหาที่พบจากการทดลองให้สมารถนำไปพัฒนาเป็นระบบผู้เชียวชาญสำหรับการวินิจฉัยโรคเบาหวานต่อไป

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวของ

การวิจัยครั้งนี้เน้นการศึกษาและทำนายความเสี่ยงการเกิดโรคเบาหวานด้วยวิธีการเหมืองข้อมูล โดยใช้สองเทคนิค คือ การจัดกลุ่มข้อมูลซึ่งเป็นกระบวนการจัดการข้อมูลผู้ป่วยโรคเบาหวานให้อยู่ในกลุ่มที่ต้องการจากนั้นทำการสร้างกฎเพื่อช่วยในการตัดสินใจจากข้อมูลที่มีอยู่ เพื่อใช้แนวโน้มการเกิดขึ้นของโรคเบาหวานในผู้ป่วย ดังนั้นผู้วิจัยได้ศึกษาทฤษฎีและวรรณกรรมที่เกี่ยวข้องเพื่อใช้เปนแนวทางในการศึกษาและอภิปรายดังนี้

2.1 วรรณกรรมที่เกี่ยวข้อง

2.1.1 การวิเคราะห์ทางสถิติ

2.1.2 ทำเหมืองข้อมูล (Data Mining)

2.1.3 เทคนิคการทำเหมืองข้อมูล (Data Mining Technique)

2.1.4 เคมีน (K-Means Clustering)

2.2. งานวิจัยที่เกี่ยวข้อง

2.2.1 การรับรู้และพฤติกรรมการดูแลตนเองของผู้ป่วยโรคเบาหวาน

2.2.2 การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล (Data Mining Technique)

2.2.3 จัดกลุ่มคนไข้ที่ป่วยโรคเบาหวาน

2.2.3 ขั้นตอนวิธีการจัดกลุ่ม (k-means)

2.3 วรรณกรรมที่เกี่ยวข้อง

2.3.1 การวิเคราะห์ข้อมูลและทางสถิติ

สถิติ หมายถึง ตัวเลขที่แสดงข้อเท็จจริงเกี่ยวกับเรื่องใดเรื่องหนึ่ง เช่นสถิติที่แสดงปริมาณน้ำฝนสถิติอุบัติเหตุสถิตินักศึกษาจำนวนผู้ป่วยเป็นโรคเบาหวานของจังหวัดมหาสารคามในรอบปีสถิติจำนวนผู้ป่วยที่มารับการรักษาในโรงพยาบาลมหาสารคามเป็นต้นหมายถึงศาสตร์ที่ว่าด้วยวิธีการจัดเก็บรวบรวมความจริงและตัวเลขที่แสดงข้อเท็จจริงซึ่งเรียกว่าการเก็บรวบรวมข้อมูลการนำเสนอข้อมูลการวิเคราะห์ข้อมูลและการตีความหมายข้อมูลเพื่อใช้เป็นประโยชน์ในการตัดสินใจที่มีเหตุผลซึ่งระเบียบวิธีทางสถิติประกอบด้วย

2.3.1.1 การเก็บรวบรวมข้อมูล จากแหล่งข้อมูลตามที่ได้มีการวางแผนซึ่งเป็นได้ทั้งข้อมูลปฐมภูมิหรือทุติยภูมิ

2.3.1.2 การนำเสนอข้อมูล เป็นการจัดทำข้อมูลที่รวบรวมได้ให้อยู่ในรูปแบบที่กะทัดรัด เช่น ตาราง กราฟ แผนภูมิ ข้อความเป็นต้นเพื่อสะดวกในการอ่านข้อมูลให้เข้าใจง่ายและเพื่อประโยชน์ในการวิเคราะห์ต่อไป

2.3.1.3 การวิเคราะห์ข้อมูล เป็นขั้นตอนการประมวลผลซึ่งในการวิเคราะห์จำเป็นต้องใช้สูตรทางสถิติต่างๆหรือใช้การอ้างอิงทางสถิติขึ้นอยู่กับวัตถุประสงค์ของงานนั้นๆเช่นการวิเคราะห์แนวโน้มเข้าสู่ส่วนกลาง การวัดการกระจาย การทดสอบสมมติฐานการประมาณค่าเป็นต้น

2.3.1.4. การตีความหมายข้อมูล เป็นขั้นตอนการนำผลผลิตการวิเคราะห์มาอธิบายให้บุคคลทั่วไปเข้าใจอาจจำเป็นต้องมีการขยายความในการอธิบายเพื่อให้งานที่ศึกษามีประสิทธิภาพ

2.4 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล การสืบค้นหรือสกัดความรู้ที่เป็นประโยชน์และที่น่าสนใจบนฐานข้อมูลขนาดใหญ่ หรือที่เรียกว่า เหมืองข้อมูลเป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้หรือสิ่งสำคัญออกมาเพื่อใช้ในการวิเคราะห์หรือทำนายสิ่งต่างๆที่จะเกิดขึ้นซึ่งการค้าหาข้อมูลความจริงที่แฝงอยู่ในข้อมูลแลเป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในกองข้อมูลที่เรามีอยู่ซึ่งต่างจากระบบฐานข้อมูลตรงที่ไม่ต้องเป็นคนกำหนดคำสั่งเพื่อค้นหาข้อมูลที่เราต้องการแต่ระบบการทำเหมืองข้อมูลจะมีขบวนการ/วิธีการแค่บอกว่าต้องการอะไรแต่ไม่จำเป็นต้องระบุทำอย่างไรระบบฐานข้อมูลทั่วไปจะบังคับให้เราต้องทำทั้งสองหน้าที่นี้คือคิดก่อนว่าจะค้นอะไรแล้วก็ไปประดิษฐ์คำสั่ง SQL เพื่อค้นหาข้อมูลนั้น ดังนั้นถ้าคิดไม่รอบคอบหรือคิดดีแล้วแต่แปลเป็นคำสั่งผิดก็จะได้ข้อมูลผิดๆหรือไม่ตรงกับความต้องการการทำเหมืองข้อมูลมีประโยชน์มากโดยเฉพาะการค้นหาข้อมูลซึ่งข้อมูลที่ได้จะเป็น

การทำเหมืองข้อมูลเป็นกระบวนการทำงานที่เรียกว่ากระบวนการที่สกัดข้อมูลเพื่อให้ได้สารสนเทศที่เรายังไม่รู้โดยเป็นสารสนเทศที่มีเหตุผลและสามารถนำไปใช้ได้ซึ่งเป็นสิ่งสำคัญในการที่จะช่วยการตัดสินใจในการทำธุรกิจและเป็นกระบวนการที่สำคัญในการการสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ที่เราเรียกสั้นๆว่า KDD (Knowledge Discovery in Database) ส่วนเหมืองเรียกสั้นๆว่า DM โดยเหมืองคือชุดโปรแกรมวิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อระบบสนับสนุนการตัดสินใจของผู้ใช้เป็นโปรแกรมที่สมบูรณ์ทั้งเรื่องการค้นหาการทำรายงานและโปรแกรมในการจัดการ ซึ่งเราคุ้นเคยดีกับคำว่าระบบสนับสนุนการตัดสินใจของผู้บริหารระดับสูง (Executive Information System : EIS) หรือระบบข้อมูลสำหรับการตัดสินใจในการบริหารซึ่งเป็นเครื่องมือชิ้นใหม่ที่สามารถค้นหาข้อมูลในฐานข้อมูลขนาดใหญ่หรือข้อมูลที่เป็นประโยชน์ในการบริหารซึ่งเป็นการเพิ่มคุณค่าให้กับฐานข้อมูลที่มีอยู่

เหมืองข้อมูล คือ ชุดโปรแกรมวิเคราะห์ข้อมูลที่ถูกออกแบบมาเพื่อสนับสนุนการตัดสินใจของผู้ใช้โดยเป็นโปรแกรมสมบูรณ์ทั้งเรื่องการค้นหาการทำรายงานและโปรแกรมการจัดการซึ่งคุ้นเคยดีในคำว่าระบบสารสนเทศเพื่อผู้บริหารหรือระบบข้อมูลตัดสินใจในการบริหารซึ่งเป็นเครื่องมือชิ้นใหม่ที่สามารถค้นหาข้อมูลในฐานข้อมูลขนาดใหญ่หรือข้อมูลที่เป็นประโยชน์ในการบริหารซึ่งเป็นการเพิ่มคุณค่าให้กับฐานข้อมูลที่มีอยู่

ขั้นตอนการทำเหมืองข้อมูล

ประกอบด้วยขั้นตอนการทำงานย่อยที่จะเปลี่ยนข้อมูลดิบให้กลายเป็นความรู้ ประกอบด้วยขั้นตอนดังนี้

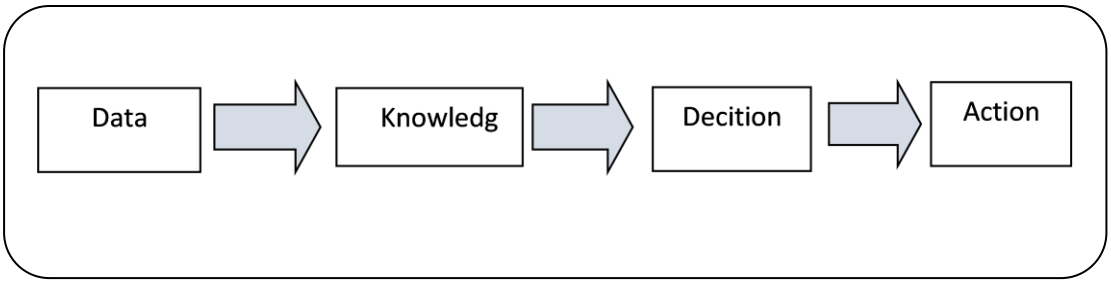
- Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป
- Data Integration เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน
- Data Selection เป็นขั้นตอนการดึงข้อมูลสำหรับการวิเคราะห์จากแหล่งที่บันทึกไว้
- Data Transformation เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน
- Data Mining เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่
- Pattern Evaluation เป็นขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล
- Knowledge Representation เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ

วัฏจักรขั้นตอนการทำงานของ Data Mining ประกอบด้วย 4 ขั้นตอนหลักๆ ดังนี้

1. การระบุโอกาสหรือการระบุปัญหาที่เกิดขึ้นเป็นการระบุขอบเขตของข้อมูลที่นำมาทำการวิเคราะห์เพื่อนำมาทางการแก้ไขปัญหา
2. ส่วนของ Data Mining เป็นการนำเทคนิคของ Data Mining ไปใช้ถ่ายทอดหรือทำการเปลี่ยนแปลงข้อมูลดิบให้อยู่ในรูปของข้อมูลที่จะนำไปใช้ได้จริง
3. การปฏิบัติตามข้อมูลคือการนำเอาข้อมูลที่เป็นผลลัพธ์ของ Data Mining มาลองปฏิบัติจริง
4. การวัดประสิทธิภาพจากผลลัพธ์การวัดประสิทธิภาพของเทคนิคของ Data Mining ที่จะนำมาใช้จากผลลัพธ์ซึ่งสามารถตรวจสอบได้หลายทาง เช่น วัดจากส่วนแบ่งของตลาด,วัดจากปริมาณลูกค้าหรือวัดจากกำไรสุทธิเป็นต้น

จากทั้ง 4 ขั้นตอนที่กล่าวมาข้างต้น คือ การนำเอา Data Mining ไปใช้กับระบบทางธุรกิจ โดยแต่ล่ะขั้นตอนจะพึ่งพาอาศัยกันผลลัพธ์จากขั้นตอนนี้จะกลายมาเป็น Input จากอีกขั้นต่อไป ซึ่ง Data Mining จะเปลี่ยนข้อมูลดิบให้เป็นข้อมูลประยุกต์หรือสารสนเทศ ดังนั้นการระบุแหล่งข้อมูลที่ถูกต้องจึงเป็นสิ่งที่สำคัญอย่างยิ่งต่อผลลัพธ์ที่ได้จากการวิเคราะห์

ระบบสนับสนุนการตัดสินใจ (Decition Support System) คือทำอย่างไรให้ข้อมูลเรามีอยู่กลายเป็นความรู้อันมีค่าได้สร้างคำตอบของอนาคตได้ดังภาพที่ 1



ภาพที่ 2-1 รูปภาพของระบบสนับสนุนการตัดสินใจ

จากภาพที่ 2-1 แสดงขั้นตอนการทำงานของระบบการสนับสนุนการตัดสินใจโดยเริ่มจากข้อมูลนำมาสกัด วิเคราะห์เพื่อให้ความรู้เพื่อใช้ในการประกอบการตัดสินใจกระทำกิจกรรมใดๆเช่น ผู้ตัดสินใจขยายสาขา โดยนำข้อมูลต่างๆทั้งอดีตปัจจุบันมาวิเคราะห์ให้ได้คำตอบว่าควรขยายสาขาหรือไม่

ปัจจุบันระบบสนับสนุนข้อมูลในการตัดสินใจได้เข้ามามีอิทธิพลในการรวบรวมข้อมูลและปรับค่าข้อมูลในคลังสินค้า ซึ่งฐานข้อมูลขนาดใหญ่นี้จะประกอบไปด้วยข้อมูลเป็นพันๆล้านไบต์ ยากแก่การค้นหาได้อย่างทันการด้วยวิธีระบบการจัดการฐานข้อมูลโดยทั่วไปที่เป็นที่น่าสนใจของผู้บริหารธุรกิจวันนี้ค้นหาได้ง่ายขึ้นแล้ว ซึ่งจะเป็นประโยชน์อย่างยิ่งในการค้นหาข้อมูลที่ต้องการในมหาสมุทรข้อมูลเพื่อนำมาเทียบเคียงและดูแนวโน้มและนำข้อมูลที่จำเป็นของบริษัทส่งกลับให้ผู้บริหารตัดสินใจได้เร็วทันเวลา

นี้คือจุดประสงค์ของการทำเหมืองข้อมูลที่จะมาช่วยในเรื่องของเทคนิคการจัดการข้อมูลและปรับค่าข้อมูล ซึ่งได้พยายามทดสอบแล้วและข้อมูลสนับสนุนที่มีอาจย้อนหลังไปถึง 30 ปีด้วยเทคนิคเดียวกันนี้สามารถใช้ค้นข้อมูลสำคัญที่ปะปนกับข้อมูลอื่นๆในฐานข้อมูลที่ไม่ใช่การสุ่มหาบางคนเรียกการสืบค้นความรู้บนพื้นฐานข้อมูล หรือ การค้นหาข้อมูลด้วยความรู้และนั่นก็คือเหมืองข้อมูล

2.4.1 ประเภทข้อมูลที่สามารถทำเหมืองข้อมูล

2.4.1.1 ฐานข้อมูลเชิงสัมพันธ์ เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตารางโดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย ER model

2.4.1.2 คลังข้อมูล เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ๆเดียวกัน

2.4.1.3 ฐานข้อมูลธุรกิจ ประกอบด้วยข้อมูลที่แต่รายการแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่งเช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูปชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

2.4.1.4 ฐานข้อมูลชั้นสูง เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่นๆเช่น ข้อมูลเชิงวัตถุ ข้อมูลที่เป็นตัวหนังสือ ข้อมูลมัลติมีเดีย ข้อมูลในรูปของเว็บ

2.4.2 ประเภทของการทำเหมืองข้อมูล

อัลกอลิทึมของการทำเหมืองข้อมูลสามารถแบ่งออกเป็น 2 ประเภท

2.4.2.1 การสร้างตัวแบบในการทำงาน (Predictive modeling) หรือเรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised learning) คือ การนำข้อมูลในอดีตมาสร้างตัวแบบเพื่อทำการทำนายอนาคตโดยมีการใช้ข้อมูลฝึกหัด ซึ่งข้อมูลทุกตัวจะมีคุณสมบัติที่ใช้ในการทำงานอัลกอริทึมประเภทนี้จะมุ่งเน้นในการแบ่งแยกข้อมูลออกเป็นกลุ่มตามค่าคุฯสมบัติของข้อมูลซึ่งถ้าค่าคุณสมบัติของข้อมูลมีค่าไม่ต่อเนื่องจะเรียกกระบวนการต่อเนื่องว่า การจำแนกประเภท (Classification) แต่ถ้าคุณสมบัติมีค่าไม่ต่อเนื่อง จะเรียกกระบวนการที่ใช้แบ่งแยกว่าการถดถอย (regression)

2.4.2.2 การสร้างตัวแบบในการพรรณนาหรือบรรยาย (Descriptive modeling) หรือเรียกว่า การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) คือการนำข้อมูลที่มีอยู่มาศึกษาเพื่อหากฎความสัมพันธ์ (association) หรือการจัดกลุ่ม (Clustering) ซึ่งไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย

โครงการในการทำเหมืองข้อมูลของ CRISP-DM มีวงจรประกอบด้วย 6 ขั้นตอน จะสั่งเกตได้ว่าลำดับขั้นตอนเป็นแบบปรับ (adaptive phase sequence) นั้นคือลำดับในขั้นตอนถัดไปบ่อยครั้งขึ้นอยู่กับผลลัพธ์ซึ่งมีความสัมพันธ์กับขั้นตอนก่อนหน้านี้ การขึ้นอยู่ต่อกันในระหว่างขั้นตอนกำหนดได้โดยลูกศร ตัวอย่างเช่น สมมติว่าเราอยู่ในขั้นตอนการสร้างตัวแบบขึ้นอยู่กับพฤติกรรมและคุณลักษณะของตัวแบบ เราอาจจะต้องกลับไปตรวจสอบในขั้นตอนการเตรียมข้อมูลก่อนที่จะไปขั้นตอนการประเมินตัวแบบ

การวนซ้ำ CRISP แทนด้วยสัญลักษณ์วงกลมด้านนอก การแก้ไขปัญหาการวิจัยหรือธุรกิจที่เฉพาะหนึ่งๆอาจจะไปสู่คำถามที่น่าสนใจได้ บทเรียนที่ได้เรียนจากโครงการในระหว่างขั้นตอนการประเมินตัวแบบสามารถส่งกลับไปวิเคราะห์ในขั้นตอนการสร้างตัวแบบสำหรับการปรับปรุงดีขึ้น

CRISP-DM ประกอบด้วย 6 ขั้นตอน

1. ขั้นตอนการทำความเข้าใจงานวิจัยและธุรกิจ (Business/research understanding phase) ขั้นตอนแรกในกระวนการที่เป็นมาตรฐาน CRISP-DM คือขั้นตอนการทำความเข้าใจการวิจัยและธุรกิจ เขียนวัตถุประสงค์และความจำเป็นของโครงการอย่างชัดเจนในรูปของงานวิจัยและธุรกิจ

    1.1 ถ่ายทอดเป้าหมายและข้อจำกัดลงในกฎเกณฑ์ของการนิยามปัญหาการทำเหมืองข้อมูล

    1.2 จัดเตรียมกลยุทธ์เบื้องต้นสำหรับบรรลุวัตถุประสงค์

2. ขั้นตอนการทำความเข้าใจข้อมูล (Data understanding phase)

    2.1 เก็บรวบรวมข้อมูล

    2.2 ใช้การวิเคราะห์ข้อมูลที่ได้จากการสำรวจเพื่อสร้างความคุ้นเคยกับข้อมูลให้แก่ตัวเราเองและค้นหาความเข้าในเชิงลึกในเบื้องต้น

    2.3 ประเมินคุณภาพของข้อมูล

    2.4 เลือกข้อมูลที่น่าสนใจซึ่งสามารถนำไปใช้ในทางปฏิบัติ

3. ขั้นตอนการจัดเตรียมข้อมูล (Data preparation phase)

    3.1 จัดเตรียมข้อมูลจากข้อมูลดิบเบื้องต้นเพื่อใช้สำหรับขั้นตอนต่อไป ขั้นตอนเป็นขั้นตอนที่ต้องใช้ความอุตสาหะเป็นอย่างมาก

    3.2 เลือกการกรณี ข้อมูล หรือตัวแปรที่ต้องการวิเคราะห์และอยู่มนรูปที่เหมาะสมสำหรับการวิเคราะห์

    3.3 ถ้ามีความจำเป็นต้องแปลงข้อมูลหรือตัวแปรให้ทำการแปลงข้อมูลหรือตัวแปร

    3.4 จัดเตรียมข้อมูลดิบให้อยู่ในรูปที่พร้อมสำหรับเป็นเครื่องมือในการสร้างตัวแบบ

4. ขั้นตอนการสร้างตัวแบบ (Modeling phase)

    4.1 เลือกและใช้เทคนิคการสร้างตัวแบบที่เหมาะสม

    4.2 จัดทำตัวแบบให้เป็นมาตรฐานเพื่อผลลัพธ์ที่ดีที่สุด

    4.3 ใช้เทคนิคที่แตกต่างกันสำหรับปัญหาการทำเหมืองข้อมูลที่เหมือนกันได้

    4.4 ถ้ามีความจำเป็นให้กลับไปขั้นตอนการจัดเตรียมข้อมูลเพื่อนำรูปแบบของข้อมูลใส่ไว้ในความจำเป็นที่กำหนดของเทคนิคการทำเหมืองข้อมูล

5. ขั้นตอนการประเมินตัวแบบ (Evaluation phase)

5.1 ประเมินตัวแบบที่ได้จากขั้นตอนการสร้างตัวแบบสำหรับคุณภาพและทำการประสบความสำเร็จก่อนการใช้งาน

5.2 หาตัวแบบที่ได้บรรลุวัตถุประสงค์สำหรับในขั้นตอนแรก

5.3 สร้างกฎเกณฑ์ที่สำคัญบางอย่างของปัญหาการวิจัยหรือธุรกิจ

5.4 ตัดสินใจโดยพิจารณาถึงผลประโยชน์ที่จะได้จากการทำเหมืองข้อมูล

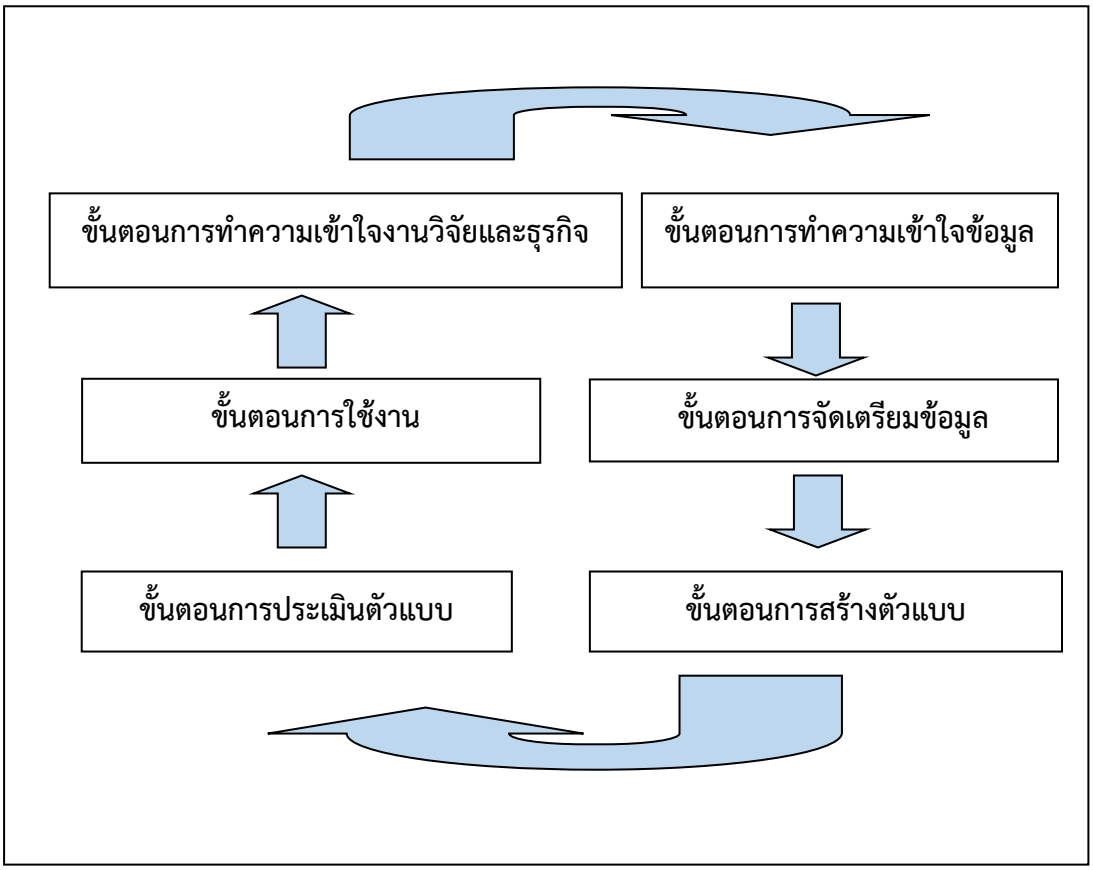
6. ขั้นตอนการใช้งาน (Deployment phase)

6.1 ใช้ตัวแบบที่สร้างขึ้น การสร้างตัวแบบยังไม่สามารถทำให้โครงการหนึ่งๆ สมบูรณ์แบบได้

6.2 ตัวอย่างของการใช้งานอย่างง่าย เช่น การสร้างงานวิจัยชิ้นหนึ่ง

6.3 ตัวอย่างการใช้งานที่มีความซับซ้อนมากขึ้น เช่น ส่งเสริมกระบวนการทำงานข้อมูลควบคู่ไปอีกแผนกหนึ่ง

6.4 สำหรับธุรกิจ บ่อยครั้งลูกค้าประสบความสำเร็จในการใช้งานโดยใช้ตัวแบบที่สร้างขึ้นให้เราสามารถหาข้อมูลเพิ่มเติมเกี่ยวกับกระบวนการที่เป็นมาตราฐาน CRISP-DM



ภาพที่ 2-2 CRISP-DM

2.4.2.3 การทำเหมืองข้อมูลในสารสนเทศทางการแพทย์

มีการทำเหมืองข้อมูลของสาขาด้านการตลาดหรือในเรื่องของการบริหารความสัมพันธ์ลูกค้ามากขึ้นแต่สำหรับการทำเหมืองข้อมูลกับสารสนเทศทางสุขภาพหรือทางการแพทย์ยังคงเป็นเรื่องใหม่มากๆ นักวิจัยในการทำเหมืองข้อมูลเริ่มให้ความสนใจมากขึ้นโดยเหตุผลสำคัญคือแนวโน้มการแพทย์ที่เป็นส่วนบุคคลมากขึ้นที่มุ่งไปที่การค้นหาการักษาและมาตราป้องกันส่วนบุคคลแหล่งข้อมูลที่สำคัญในการทำเหมืองข้อมูลยังคงมาจากความก้าวหน้าทางเทคโนโลยี ชีวภาพ เช่น Single Nucleotide Polymorphisms (SNP) chips การทำ DNA profiling จาก comparative GenomicHybridiztion array (CGH array) และการทำ mass spectrometry โดยทั้งนี้สามารถแบ่งงานทำเหมืองข้อมูลทางการแพทย์ออกได้เป็น 4 ด้านดังนี้

1) งานวินิจฉัยโรค เพื่อวิเคราะห์ว่าผู้ป่วยเจ็บป่วยจากภาวะทางการแพทย์อะไรบ้าง เช่น การวินิจฉัยผู้ป่วยโรคเบาหวานและช่องปากระยะแรกเป็นเรื่องยากหากกระทำโดยวิธีทางการแพทย์ตามปกติ การใช้ข้อมูลทางพันธุศาสตร์ช่วยได้มากทางการวินิจฉัยที่รวดเร็วขึ้นรวมทั้งความแม่นยำในการวินิจฉัยด้วย

2) การพยากรณ์โรค เพื่อคาดคะเนว่าผู้ป่วยจะหายเจ็บป่วยได้ดีขนาดไหนและโรคดำเนินไปอย่างไรตามระยะเวลา เช่น การใช้ biomarker ในการคาดคะเนว่า อวัยวะที่ปลูกถ่ายสามารถทนอยู่ในร่างกายของผู้รับได้นานเท่าใด

3) การรักษาที่เหมาะสม เพื่อคาดคะเนผลการรักษาโรค เช่น การใช้ biomarker ในการคาดคะเนว่าการรักษาทางเคมีบำบัดได้ผลอย่างไร

4) การทำความเข้าใจเกี่ยวกับกลไกโรค เพื่อให้เกิดแนวคิดหรือความรู้สึกใหม่เกี่ยวกับสาเหตุการเกิดโรค เช่น การวิจัยเกี่ยวกับ Signaling pathway ในระหว่างการติดเชื้อไวรัส

2.4.2.4 ความผิดพลาดของการทำเหมืองข้อมูล (Fallacies of Data Mining)

อธิบายความผิดพลาดที่เกิดขึ้นในการทำเหมืองข้อมูลมี 6 ประการ คือ

ความผิดพลาดที่ประการที่ 1 มีเครื่องมือในการทำเหมืองข้อมูลที่ใช้ในการแก้ไขปัญหาเพื่อหาคำตอบ

ความผิดพลาดประการที่ 2 กระบวนการในการทำเหมืองข้อมูลเป็นอิสระกัน ต้องการการดูแลเอาใจใส่จากผู้เกี่ยวข้องเพียงเล็กน้อยหรืออาจจะไม่จะเป็นต้องมีผู้เกี่ยวข้องคอยดูแลเลยก็ได้ ความจริงแล้วกระบวนการในการทำเหมืองข้อมูลจำเป็นต้องใช้ผู้มีความชำนาญทำวนซ้ำในแต่ล่ะขั้นตอน แม้ว่าหลังจากการใช้ตัวแบบแล้วคำแนะนำเกี่ยวกับข้อมูลใหม่ๆบ่อยครั้งจำเป็นต้องปรับปรุงตัวแบบให้ทันสมัยยุเสมอการเตือนคุณภาพอย่างต่อเนื่องและการประเมินการวัดต้องได้รับการประเมินโดยผู้ชำนาญเป็นผู้วิเคราะห์

ความผิดพลาดประการที่ 3 การทำเหมืองข้อมูลให้ผลตอบแทนอย่างรวดเร็วความจริงแล้วอัตราการตอบแทนมีการแปรผันไปขึ้นอยู่กับค่าใช้จ่ายเบื้องต้นค่าใช้จ่ายการวิเคราะห์ส่วนบุคคล ค่าใช้จ่ายในการจัดเตรียมคลังสินค้า และอื่นๆ

ความผิดพลาดประการที่ 4 ซอฟแวร์การทำเหมืองข้อมูลง่ายในการใช้งานความจริงแล้วความง่ายในการใช้งานมีการแปรผันไปอย่างไรก็ตามผู้วิเคราะห์ข้อมูลต้องรวบรวมสาระความรู้เกี่ยวกับเชิงวิเคราะห์และสร้างความคุ้นเคยตัวแบบการวิจัยหรือธุรกิจ

ความผิดพลาดประการที่ 5 การทำเหมืองข้อมูลจะระบุสาเหตุของปัญหาการวิจัยหรือธุรกิจความจริงแล้วกระบวนการค้นหาความรู้จะช่วยเปิดเผยรูปแบบของพฤติกรรมมนุษย์เป็นผู้ระบุสาเหตุของปัญหาการวิจัยหรือธุรกิจ

ความผิดพลาดประการที่ 6 การทำเหมืองข้อมูลจะทำให้ฐานข้อมูลทำงานอย่างเป็นอัตโนมัติความจริงแล้วการทำงานไม่ได้เป็นไปอย่างอัตโนมัติขั้นตอนเบื้องต้นของกระบวนการในการทำเหมืองข้อมูลบ่อยครั้งไม่ได้ทำการตรวจสอบจัดเตรียมข้อมูลหรือไม่ได้ใช้งานเป็นเวลาหลายปีดังนั้น องค์กรที่เริ่มดำเนินการในการทำเหมืองข้อมูลใหม่ๆบ่อยครั้งจะเผชิญหน้ากับปัญหาด้วนข้อมูลซึ่งมีการบิดเบือนข้อมูลเป็นเวลานานหลายปีและมีความจำเป็นที่จะต้องมีการปรับปรุงข้อมูลเป็นอย่างมาก

2.5 เทคนิคการทำเหมืองข้อมูล (Data Mining Technique)

การจัดกลุ่มข้อมูล (Clustering) หมายถึงการจัดกลุ่มของเรคคอร์ดและประมวลผลให้อยู่ในกลุ่มของกลุ่มวัตถุที่คล้ายๆกันการจัดกลุ่มต่างกันการจำแนกตรงที่ไม่มีกลุ่มที่แน่นอนในการจัดกลุ่มอัลกอริทึมจะพยายามมองหาส่วนของข้อมูลทั้งหมดแบ่งเป็นกลุ่มย่อยซึ่งมีความสัมพันธ์กันหรือเหมือนกันโดยที่เรคคอร์ดที่คล้ายกันจัดให้อยู่ในกลุ่มเดียวกันให้ได้มากที่สุดการจัดกลุ่มอาศัยแนวคิดที่ว่าสมาชิกที่มีลักษณะทั่วไปเหมือนกันหรือคล้ายกันก็ควรจัดอยู่ในกลุ่มเดียวกันความเหมือนหรือความคล้ายคลึงกันก็อาศัยการวัดระยะทางหรือความห่างกันเป็นหลักในการคำนวณหน่วยวิเคราะห์หรือรายการข้อมูลใดให้ห่างกันน้อยที่สุดก็แสดงว่าหน่อยวิเคราะห์คู่นั้นมีความคล้ายคลึงกันมากที่สุดซึ่งในงานวิจัยนี้จะใช้หลักการจัดกลุ่มโดย IDA (Intellingent Data Analysis) ซึ่งเป็นโปรแกรมที่ช่วยการจัดกลุ่มข้อมูลและอธิบายวิธีการในการจัดกลุ่มต่อไปนี้

2.5.1 หลักการของการจัดกลุ่ม

การจัดกลุ่มเป็นเทคนิคการจำแนกสมาชิกออกเป็นกลุ่มย่อยๆจะใช้เมื่อมีสมาชิกจำนวนมากโดยจะต้องกำหนดจำนวนกลุ่มหรือจำนวนกลุ่มที่ต้องการเทคนิคการจัดกลุ่มจะมีการทำงานหลายๆรอบโดยแต่ละรอบจะมีการรวมสมาชิกให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่งโดยเลือกกลุ่มที่สมาชิกนั้นมีระยะห่างค่ากลางของกลุ่มนั้นน้อยลงที่สุดแล้วคำนวณค่ากลางของกลุ่มใหม่จะทำเช่นนี้จนค่ากลางไม่เปลี่ยนแปลงหรือครบตามจำนวนที่กำหนดไว้โดยอัลกอริทึมของจะเริ่มจาก

2.5.1.1 กำหนดจำนวนของกลุ่ม เป็นการที่กำหนดว่าเราต้องการแบ่งประชากรออกเป็นกี่กลุ่ม

2.5.1.2 ทำการกำหนดค่าของค่ากลางแต่ลำกลุ่มโดยสุ่มมาจากประชากรที่มีอยู่

2.5.1.3 ทำการนำประชากรที่มีอยู่จับเข้าแต่ละกลุ่มโดยที่ต้องใกล้กับค่ากลางของกลุ่มนั้นมากที่สุดในแต่ละครั้งที่มีการจับเข้ากลุ่มต้องคิดค่าค่ากลางของกลุ่มนั้นๆใหม่เสมอ โดยหาจากค่าเฉลี่ยของประชากรในกลุ่มทำไปจนกว่าจำนวนประชากรจะถูกจัดเข้าทุกกลุ่มจนครบ

2.5.1.3 ทำหารตรวจสอบว่ากรณีที่ได้มานี้ผลของฟังก์ชันจุดประสงค์นั้นมีค่าต่ำสุดแทบจะไม่เปลี่ยนแปลงค่าหรือวนครบจำนวนรอบที่กำหนดไว้ใช่หรือไม่

2.5.1.4 ได้ผลลัพธ์เป็นข้อมูลที่แบ่งเป็นกลุ่มๆทั้งหมดทุกกลุ่มและมีค่าน้อยที่สุด

2.5.2 การจำแนกประเภทข้อมูล (Data Classification Model)

(Data Classification Model) คือกระบวนการสร้างโมเดลจำแนกประเภทข้อมูลเพื่อทำนายกลุ่มของข้อมูลใหม่ (Unseen data) ตัวอย่างของกลุ่มเช่น กลุ่มของลูกค้าที่ซื้อคอมพิวเตอร์-ไม่ซื้อคอมพิวเตอร์กลุ่มของลูกค้าที่ฐานะดี-ปานกลาง-แย่ กลุ่มของการผลิตสินค้า ผ่านเกณฑ์-ไม่ผ่านเกณฑ์ ในที่นี้คำว่ากลุ่มจะเรียกว่า class ของข้อมูล ซึ่งใน class เดียวกันนั้นจะต้องมีข้อมูลที่มีความเหมือนหรือคล้ายคลึงกันมากกว่าข้อมูลที่อยู่ใน class ที่แตกต่างกัน

การสร้างโมเดลจำแนกประเภทข้อมูล จะเกิดขึ้นมาจากการหาความสัมพันธ์ของข้อมูลในฐานข้อมูลขนาดใหญ่ โดยข้อมูลทั้งหมดจะมีการแบ่งออกเป็น 2 กลุ่มคือกลุ่มข้อมูลเรียนรู้ (Training set) เป็นชุดข้อมูลที่มีบทบาทในการสร้างโมเดลจำแนกประเภทข้อมูลขึ้นมา และมีกลุ่มข้อมูลทดสอบ (Test set) เป็นชุดข้อมูลประเมินความถูกต้องของโมเดลจำแนกประเภทข้อมูล

โมเดลจำแนกประเภทข้อมูลได้ถูกนำมาประยุกต์ใช้งานหลายๆด้านไม่ว่าจะเป็นการวิเคราะห์หุ้นเพื่อหาว่าหุ้นแต่ละบริษัทมีคุณภาพเป็นอย่างไรเมื่อมีปัจจัยที่เกี่ยวข้องไม่ว่าจะเป็น การเติบโตของรายได้ความสามารถในการควบคุมต้นทุนความผันผวนของรายได้และกำไรและผู้บริหาร หรือจะเป็นการพยากรณ์อากาศ การจัดสรรกฎหมายที่เหมาะสมในการพิจารณาคดีความการจัดการความสัมพันธ์ของลูกค้า (Customer relationship management) และอื่นๆ

กระบวนการจำแนกประเภทข้อมูลแบ่งออกเป็น 3 ขั้นตอน คือ

1) Model Construction (Learning) เป็นขั้นตอนการสร้างโมเดลจำแนกประเภทโดยอาศัยการเรียนรู้จากข้อมูลที่ได้กำหนด class ไว้เรียบร้อยแล้วหรือเรียกว่าข้อมูลเรียนรู้ (Training data) ซึ่งโมเดลจำแนกประเภทที่ได้จะแสดงด้วยวิธีการพื้นฐานทางเหมืองข้อมูล (Data mining) ยกตัวอย่างเช่น ต้นไม้ตัดสินใจ (Decision Tree) โมเดลจำแนกประเภทที่ได้จะมีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดรากอยู่ด้านบนสุดและโหนดใบอยู่ล่างสุดของต้นไม้ แต่ละโหนดบนต้นไม้จะมี

คุณลักษณะ (attribute) เป็นตัวเลือกทดสอบ ซึ่งจะมีกิ่งซึ่งเป็นค่าที่เป็นไปได้ของคุณลักษณะ (attribute value) ที่ถูกเลือกทดสอบไว้ และมีโหนดใบแสดง class ที่กำหนดไว้

2) Model Evaluation (Accuracy) เป็นขั้นตอนตรวจสอบความถูกต้องโดยอาศัยข้อมูลที่ใช้สำหรับทดสอบเรียกว่าข้อมูลทดสอบ (Testing data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบจะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลจำแนกประเภทเพื่อทดสอบว่าโมเดลจำแนกประเภทนี้สามารถจัดกลุ่มประเภทข้อมูลได้อย่างถูกต้องมากน้อยเพียงใดและมีการปรับปรุงโมเดลจำแนกประเภทจนกว่าจะได้ค่าความถูกต้องในระดับที่ยอมรับได้

3) Model Usage (Classification) เป็นขั้นตอนการนำโมเดลจำแนกประเภทที่สร้างขึ้นมาใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) เพื่อทำนายและกำหนดกลุ่มให้กับข้อมูลนั้น

2.5.3 ความสัมพันธ์ (Association)

เป็นเทคนิคหนึ่งของ Data Mining ที่สำคัญและสามารถนำไปประยุกต์ใช้ได้จริงกับงานต่างๆหลักการทำงานวิเคราะห์ของวิธีนี้ คือ การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการหรือทำนายปรากฏการณ์ต่างๆหรือมากจากการวิเคราะห์การซื้อสินค้าของลูกค้าเรียกว่า “ Market Basket Analysis ” ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหาซึ่งการวิเคราะห์แบบนี้เป็นการใช้“กฎความสัมพันธ์” (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริงได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติ ของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆมักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน

2.6 เคมีน (K-Means Clustering)

การจัดกลุ่มข้อมูลแบบเคมีน (K-Means Clustering) ขั้นตอนวิธีการจัดกลุ่ม k-means เป็นเทคนิคหนึ่งที่จัดอยู่ในประเภท Partition Method มีการใช้ค่าเฉลี่ยของข้อมูลที่ถูกจัดให้อยู่ในคลัสเตอร์เดียวกันเป็นตัวแทนของทุกข้อมูลในคลัสเตอร์นั้น

ขั้นตอนวิธีเริ่มต้นจากการรับค่าพารามิเตอร์ k ซึ่งค่านี้คือจำนวนคลัสเตอร์ที่ต้องการค้นหา จากนั้นขั้นตอนวิธีจะทำการสุ่มเลือกข้อมูลเริ่มต้นจำนวน k ชุด ซึ่งแต่ละชุดที่ได้มานั้นจะเป็นจุดศูนย์กลางเริ่มต้นของแต่ละคลัสเตอร์ (centroid) จากนั้นทำการจัดกลุ่มให้กับข้อมูลที่เหลือข้อมูลจะถูกจัดให้อยู่ในคลัสเตอร์เดียวกันเมื่อข้อมูลนั้นมีความคล้ายกับตัวแทนของคลัสเตอร์นั้นมากที่สุด จากนั้นจึงทำการคำนวณหาค่าเฉลี่ยของคลัสเตอร์ใหม่และดำเนินกระบวนการเดียวกันกับข้อมูลที่เหลือต่อไปจนกระทั่งทุกข้อมูลถูกจัดกลุ่มอย่างสมบูรณ์และข้อมูลไม่มีการเปลี่ยนกลุ่มอีกต่อไป

การทำงานของ k-means จะมีประสิทธิภาพสูงก็ต่อเมื่อข้อมูลเกาะกลุ่มกันหนาแน่นแต่ละกลุ่มแยกจากกันอย่างชัดเจนและความหนาแน่นของข้อมูลในแต่ละกลุ่มใกล้เคียงกันจุดเด่นของ k-means คือง่ายและสามารถใช้ได้กับข้อมูลหลายประเภทและยังมีประสิทธิภาพในด้านความเร็ว แต่จุดด้อยของ k-means ก็พบว่ายังไม่เหมาะสมกับข้อมูลทุกประเภทและไม่สามารถจัดการกลุ่มที่มีรูปร่างไม่เป็นรูปทรงกลมหรือกลุ่มที่มีขนาดหรือความหนาแน่นแตกต่างกันได้นอกจากนี้ k-means ยังถูกจำกัดสำหรับข้อมูลที่มีตัวแทนข้อมูลที่คลุมเครือหรือไม่ชัดเจน

2.7 งานวิจัยที่เกี่ยวข้อง

2.7.1 การรับรู้และพฤติกรรมการดูแลตนเองของผู้ป่วยโรคเบาหวาน

โรคเบาหวาน (DIABETES MELLITUS) คือโรคเรื้อรังชนิดหนึ่งเกิดจากความผิดปกติของอินซูลินทำให้รางกายมีการเผาผลาญผิดปกติทั้งคาร์โบไฮเดรต ไขมัน และโปรตีนมีลักษณะเดนชัดคือมีระดับน้ำตาลในเลือดสูงกว่าปกติการเผาผลาญที่ผิดปกติมากที่สุดในโรคเบาหวานคือคาร์โบไฮเดรตและฮอรโมนที่ควบคุมการเผาผลาญของคาร์โบไฮเดรต คือ อินซูลิน ซึ่งหลั่งจากเบตาเซลลในกลุมเซลลแลนเกอรแฮน (Islet of Langerhan) ของตับออนโรคเบาหวานคือโรคเรื้อรังชนิดหนึ่งเกิดจากความผิดปกติของอินซูลินทำให้รางกายมีการเผาผลาญผิดปกติทั้งคาร์โบไฮเดรต ไขมัน และโปรตีนมีลักษณะเดนชัดคือ มีระดับน้ำตาลในเลือดสูงกวาปกติการเผาผลาญที่ผิดปกติมากที่สุดในโรคเบาหวาน คือ คาร์โบไฮเดรตและฮอรโมนที่ควบคุมการเผาผลาญของคาร์โบไฮเดรต คือ อินซูลิน ซึ่งหลั่งจากเบตาเซลล ในกลุมเซลลแลนเกอรแฮน (Islet of Langerhan) ของตับออน

อาการของโรคเบาหวานคนปกติกอนรับประทานอาหารเช้าจะมีระดับน้ำตาลในเลือด 70-110 มก.ดล. หลังรับประทานอาหารแลว 2 ชม. จะมีระดับน้ำตาลไมเกิน 140 มก.% ผูที่ระดับน้ำตาลสูงไมมากอาจจะไมมีอาการอะไรอาการที่พบไดบอย

1) ปสสาวะบอยและมากโดยเฉพาะในเวลากลางคืน และอาจจะพบวาปสสาวะมีมดตอม เนื่องจากเมื่อน้ำตาลในกระแสเลือดมากกวา 180 มก.ดล. น้ำตาลจะถูกขับออกทางปสสาวะทำใหน้ำถูกขับออกมากขึ้นจึงมีอาการปสสาวะบอยและเกิดการสูญเสียน้ำ

2) ผูปวยจะหิวน้ำบอยเนื่องจากตองทดแทนน้ำที่ถูกขับออกทางปสสาวะ

3) ออนเพลีย น้ำหนักลดเกิดเนื่องจากรางกายไมสามารถใชน้ำตาลจึงยอยสลายสวนที่เปนโปรตีนและไขมันออกมา

4) ผูปวยจะกินเกงหิวเกงแตน้ำหนักจะลดลงเนื่องจากรางกายนำน้ำตาลไปใชเปนพลังงานไมไดจึงมีการสลายพลังงานจากไขมันและโปรตีนจากกลามเนื้อ

5) อาการอื่นๆที่อาจเกิดไดแก การติดเชื้อ แผลหายชา คัน

6) คันตามผิวหนังมีการติดเชื้อราโดยเฉพาะอยางยิ่งบริเวณชองคลอดของผูหญิง

7) เห็นภาพไมชัดตาพรามัวตองเปลี่ยนแวนบอยทั้งนี้อาจจะเปนเพราะมีการ

เปลี่ยนแปลงสายตาเชน สายตาสั้น ตอกระจก

8) ชาไมมีความรูสึกเจ็บตามแขนขาหยอนสมรรถภาพทางเพศเนื่องจากน้ำตาลสูง นานๆทำใหเสนประสาทเสื่อมเกิดแผลที่เทาไดงายเพราะไมรูสึก

9) อาเจียน

สาเหตุของโรคเบาหวาน

ในประเทศไทยพบผูเปนเบาหวานตั้งแตชวงอายุ 15 ปขึ้นไปประมาณทั้งสิ้นถึงเกาแสนคน นอกจากนั้นยังพบวาเมื่ออายุสูงขึ้นมีโอกาสเปนเบาหวานไดงายขึ้น ไดแกประชากรอายุระหวาง 40 – 60 ปจะพบประมาณรอยละ 4-7 และอายุ 60 ปขึ้นไปอาจพบสูงถึงรอยละ 5-10

ปจจัยที่อาจเปนตนเหตุของการเกิดโรคดังนี้

1) น้ำหนักเกินความอวนและขาดการเคลื่อนไหวออกกำลังกายที่เพียงพอ

2) กรรมพันธุมักพบโรคนี้ในผูที่มีบิดา มารดา เปนเบาหวานลูกมีโอกาสเปนเบาหวาน 6-10 เทาของคนที่เปนเบาหวาน

3) จากเชื้อโรคหรือยาบางชนิดไปทำลายเซลลของตับออนทำใหตับออนไมสามารถหลั่งฮอรโมนอินซุลินไดเพียงพอ

4) เกิดรวมกับโรคตอมไรทอบางชนิด เชน โรคเนื้องอกของตอมใตสมองหรือตอมหมวกไต

2.8 การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล (Data Mining Technique)

งานวิจัยชิ้นนี้ได้แบ่งเทคนิคการทำเหมืองข้อมูลออกเป็น 7 กลุ่ม คือ การวิเคราะห์ความสัมพันธ์ การจำแนกกลุ่ม การแบ่งกลุ่ม การทำนาย คาดเดาทางสถิติการสืบค้นแบบเรียงลำดับ การแสดงด้วยภาพ ในส่วนของ Association มีเป้าหมายเพื่อสร้างความสัมพันธ์ระหว่างสิ่งต่างๆที่มีอยู่ร่วมกัน เครื่องมือที่ใช้ในการสร้างรูปแบบความสัมพันธ์คือสถิติการจำแนกกลุ่มเป็นหนึ่งในรูปแบบการเรียนรู้ในการทำเหมืองข้อมูลที่เป็นที่รู้จักมากที่สุดมีจุดมุ่งหมายเพื่อสร้างรูปแบบการพยากรณ์พฤติกรรมของลูกค้าในอนาคตเครื่องมือที่ใช้ได้แก่โครงข่ายประสาทเทียม ต้นไม่ตัดสินใจ เป็นต้น สำหรับการแบ่งกลุ่มจะเป็นการแยกกลุ่มประชากรที่มีลักษณะที่แตกต่างกันให้กลายเป็นกลุ่มย่อยๆที่มีลักษณะคล้ายกันซึ่งสามารถกำหนดจำนวนของกลุ่มประชากรที่ต้องการได้การทำนายเป็นตัวคาดการณ์ในอนาคต

2.9 จัดกลุ่มคนไข้ที่ป่วยโรคเบาหวาน

นักวิจัยได้ศึกษาวิธีการทำนาย อัตราการมีชีวิตรอดของคนป่วยโรคเบาหวาน โดยอาศัยหลักการจักกลุ่มคนไข้ เดิมมีการพยากรณ์โดยใช้ระบบ TNM ที่เกี่ยวข้องกับ 3 ปัจจัย คือ ปสสาวะบอยและมากโดยเฉพาะในเวลากลางคืนการติดเชื้อ แผลหายชา คัน เห็นภาพไมชัดตาพรามัวตองเปลี่ยนแวนบอยเนื่องจากองค์ประกอบในการทำนายที่มีผลตัวอื่นไม่ได้ถูกนำมาใช้ในระบบจากระดับ

ความสามารถของ Data set ของคนไข้มีขนาดใหญ่มีความเป็นไปได้ที่จะจัดทำระบบการพยากรณ์ที่มีประสิทธิภาพ โดยใช้กระบวนการการเรียนรู้ และวิธีการทางสถิติในการวิจัยผู้วิจัยได้นำเสนอพื้นฐานวิธีการจัดกลุ่มเพื่อพัฒนาระบบการทำนายคนไข้โรคเบาหวานวิธีการเริ่มจากด้วยการจัดกลุ่มความสัมพันธ์ที่ถูกบันทึกในฐานข้อมูล จากนั้นวัดความแตกต่างระหว่างความสัมพันธ์เชื่อมโยงที่ได้มาจากการเรียงลำดับของ Data Partition ที่เกิดจาก Multiple Cluatering ต่อจากนั้นการวัดความแตกต่างถูกใช้กับวิธีการจัดกลุ่มแบบ hierarchical เพื่อค้นหากลุ่มของความเชื่อมโยงการพยากรณ์การรอดชีวิตของผู้ป่วยนี้ ได้มาจาก ฟังค์ชั่นการรอดชีวิตที่ได้มาจากแต่ล่ะกลุ่ม ทฤษฎีนี้ใช้ ตัวแปรร่วมและทำให้เกิดเครื่องมือที่เป็นประโยชน์และสะดวกในการพยากรณ์ผลลัพธ์ของคนไข้โรคเบาหวาน

2.10 ขั้นตอนวิธีการจัดกลุ่ม (k-means)

การจัดกลุ่มแบบ เค-มีน คัสเตอร์ริ่ง คือ หนึ่งในอัลกอริทึมเทคนิคการเรียนรู้โดยไม่มีผู้สอนที่ง่ายที่สุด เพราะเป็นการแก้ปัญหาการจัดกลุ่มที่รู้จักกันทั่วไป โดยอัลกอริธึม เค-มีน จะตัดแบ่ง (Partition) วัตถุออกเป็น K กลุ่ม โดยแทนแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่มซึ่งใช้เป็นจุดศูนย์กลาง (centroid)ของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกันในขั้นแรกของการจัดกลุ่มโดยการหาค่าเฉลี่ยแบบ K ต้องกำหนดจำนวนกลุ่ม (K) ที่ต้องการและกำหนดจุดศูนย์กลางเริ่มต้นจำนวน K จุด สิ่งสำคัญในการกำหนดจุดศูนย์กลางเริ่มต้นของแต่ละกลุ่มนี้ควรจะถูกกำหนดด้วยวิธีที่เหมาะสม เพราะตำแหน่งจุดศูนย์กลางเริ่มต้นที่แตกต่างกันทำให้ได้ผลลัพธ์สุดท้ายแตกต่างกันดังนั้นในทางที่ดีควรจะกำหนดจุดศูนย์กลางนี้ให้ห่างจากจุดศูนย์กลางอื่นๆขั้นตอนต่อไปนี้คือสร้างกลุ่มข้อมูลและความสัมพันธ์กับจุดศูนย์กลางที่ใกล้มากที่สุดโดยแต่ละจุดจะถูกกำหนดไปยังจุดศูนย์กลางที่ใกล้เคียงที่สุดจนครบหมดทุกจุดและคำนวณจุดศูนย์กลางใหม่โดยการหาค่าเฉลี่ยทุกวัตถุที่อยู่ในกลุ่มหากจุดศูนย์กลางในแต่ละกลุ่มถูกเปลี่ยนตำแหน่งจะได้จุดมีความสัมพันธ์กับกลุ่มใหม่และใกล้กับจุดศูนย์กลางใหม่ทำซ้ำแบนี้ไปเรื่อยๆจะสังเกตเห็นว่าผลลัพธ์จากการทำซ้ำแบบนี้ทำให้จุดศูนย์กลางเปลี่ยนตำแหน่งทุกรอบ จนกระทั่งจุดศูนย์กลางจำนวน K จุดไม่มีการเปลี่ยนแปลงจึงจะสิ้นสุดกระบวนการ

บทที่3
วิธีการดำเนินงานวิจัย

จากวรรณกรรมและงานวิจัยที่เกี่ยวข้องในบทที่ 2 สามารถนำมาใช้ในการวิเคราะห์และจัดลำดับความสำคัญของปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวานโดยได้อาศัยแนวคิดและวิธีดำเนินการวิจัยตามหลักการและทฤษฎีพื้นฐานที่ได้กล่าวไว้ในบทที่ 2 จุดประสงค์ของงานวิจัยครั้งนี้มุ่งเน้นเพื่อศึกษาการใช้เทคนิคของการทำเหมืองข้อมูล คือ การจัดกลุ่มโดยการจักกลุ่มใช้ Inteligent Data Analysis (IDA) เพื่อหาความถี่ของปัจจัยเสี่ยงของผู้ป่วยจากนั้นมาวิเคราะห์ความเสี่ยงของการเกิดโรคเบาหวานสร้างกฎเพื่อช่วยในการตัดสินใจเพื่อทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดซึ่งจากความไม่แน่นอนของสาเหตุการเกิดโรคเบาหวานที่เกิดจากหลากหลายปัจจัยได้โดยบทที่ 3 นี้จะอธิบายถึงขั้นตอนการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวานการคัดเลือกตัวแบบที่เหมาะสมในการวิเคราะห์ปัจจัยเสี่ยงและการพยากรณ์ความเสี่ยงการเกิดโรคเบาหวาน

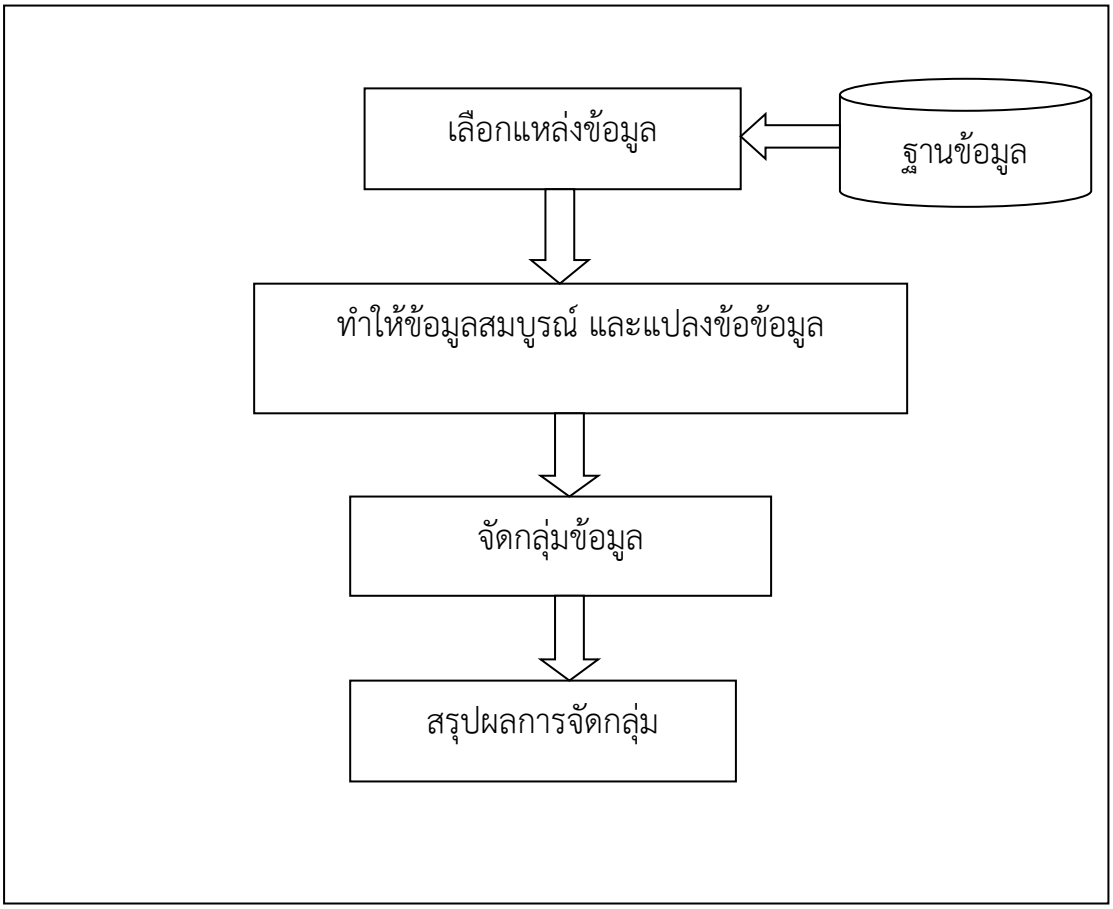
โดยบทนี้จะกล่าวถึงการดำเนินการค้นคว้าอิสระ โดยอ้างอิงจากทฤษฎีในบทที่ 2 มาประยุกต์ใช้ ซึ่งจะดำเนินงานตามขั้นตอนดังต่อไปนี้

3.1 แผนภาพกระบวนการดำเนินงาน

3.2 ขั้นตอนการดำเนินงาน

โดยกระบวนการทำงานงานวิจัยทำการแบ่งการทำงานออกเป็นขั้นตอนย่อยๆซึ่งประกอบด้วยเลือกแหล่งของข้อมูล ทำข้อมูลให้สมบูรณ์ และแปลงข้องข้อมูล สรุปผลพยากรณ์ ทั้งนี้เพื่อให้สะดวกและง่ายต่อการทดลองเนื่องจากได้แบ่งขอบเขตการทำงานในแต่ละขั้นตอนอย่างชัดเจน

3.1 แผนภาพกระบวนการดำเนินงาน



ภาพที่ 3-1 ขั้นตอนการดำเนินงาน

จากภาพแสดงถึงขั้นตอนในการดำเนินงานวิจัยซึ้งประกอบไปด้วยขั้นตอนสำคัญๆคือ ขั้นตอน การจัดกลุ่มข้อมูลเพื่อหาค่าของต่างๆ ของผู้ป่วยโรคเบาหวานและการพยากรณ์ความน่าจะเป็นของการเกิดโรคเบาหวานซึ่งอธิบายได้โดยละเอียดดังนี้

3.2 ขั้นตอนการดำเนินงาน

3.2.1 การเลือกข้อมูล

จุดประสงค์ คือการระบุแหล่งของข้อมูลที่มี และทำการดึงเอาข้อมูลออกมาใช้สำหรับกาวิเคราะห์เบื้องต้นในการเตรียมตัวสำหรับการที่จะทำการทำเหมืองข้อมูลในขั้นต่อๆไป การเลือกข้อมูลนั้นจะแตกต่างไปตามวัตถุประสงค์ของแต่ละธุรกิจ ที่ได้กำหนดไว้ตั้งแต่ต้นและการเลือกข้อมูลก็ยังถูกำหนดโดยลักษณะงานที่จะถูกนำมาใช้อีกด้วย

ตัวแปรที่ถูกเลือกมาแต่ละตัวนั้นจะต้องถูกทำความเข้าใจว่าตัวแปรแต่ละตัวหมายความว่าอะไร ประกอบด้วยอะไรไม่เพียงแต่คำจำกัดความทางธุรกิจเท่านั้นแต่จะต้องมีคำอธิบายอย่างชัดเจนเกี่ยวกับชนิดของข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบของข้อมูล และลักษณะอื่นๆ

3.2.1.1 รูปแบบตัวแปร

3.2.1.2 Nominal Variabel กล่าวถึงชนิดนี้ของ Object ที่มาอ้างถึงแต่ไม่มีลำดับในค่าที่เปลี่ยนไปได้ Possibie Value ตัวอย่างเช่น สถานะ เพศ (ชาย หญิง)

3.2.1.3 Oedinal Variable มีลำดับสำหรับค่าที่เป็นไปได้ ตัวอย่างเช่น ลำดับของผู้ป่วย (เป็น มีโอกาสเป็น ไม่เป็น)

3.2.2 ตัวแปรแบบ Quantitative ซึ่งมีการวัดความแตกต่างระหว่างค่าที่เป็นไปได้

3.2.2.1 ค่าต่อเนื่อง (Continuous) เช่น รายได้ เฉลี่ยจำนวนครั้งที่ซื้อ

3.2.2.2 ค่าไม่ต่อเนื่อง (Discrete) เช่น จำนวนผู้ป่วย เวลา ปี เดือน

ตัวแปรข้อมูลมีหลายตัวมากแต่ตัวแปรที่ถูกเลือกสำหรับการทำเหมืองข้อมูลนั้น ถูกเรียกว่า “Active Variable” เพราะว่ามันจะถูกใช้ความแตกต่างของกลุ่มย่อยต่างๆและสามารถนำมาทำนายผลไม้ เมื่อคุณทำการเลือกข้อมูลจะต้องพิจารณาอายุของข้อมูลด้วยเพราะว่าสถานการณ์ภายนอกเปลี่ยนแปลงตลอดเวลาซึ่งจะทำให้ประสิทธิภาพของการทำเหมืองข้อมูลลดลง ตัวอย่างเช่น รสนิยมการใช้ชีวิต การเปลี่ยนงาน

3.3 การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน

ในโครงสร้างการวินิจฉัยโรคเบาหวานโดยปกติจะทำการวินิจฉัยด้วย 3 วิธี คืออาการแสดงทางโรงพยาบาล พฤติกรรมเสี่ยง และผลจากการตรวจเลือด ซึ่งหากมีผลการตรวจเลือดแล้วอาจนำมาใช้ในการวินิจฉัยเพียงอย่างเดียวก็ได้ แต่เนื่องจากการวินิจฉัยด้วยการตรวจเลือดต้องใช้ค่าใช้จ่ายสูงและไม่สามารถทำได้ครอบคลุมส่วนพฤติกรรมเสี่ยงจะเก็บข้อมูลของปัจจัยเสี่ยงได้ยากเนื่องจากพฤติกรรมของแต่ละบุคคลมีความแตกต่างกันทำให้ยากต่อการมาสร้างเป็นฐานความรู้ ผู้ศึกษาจึงสนใจในส่วนของการแสดงทางโรงพยาบาลมาใช้ในการวิเคราะห์เพียงอย่างเดียวโดยแพทย์ผู้เชี่ยวชาญได้สรุปปัจจัยและสภาวะความเสี่ยงสำคัญต่อการเกิดโรคเบาหวานที่มีนัยสำคัญทางสถิติ

3.4 การเก็บรวบรวมข้อมูล

ผู้วิจัยได้ประสานความร่วมมือชี้แจงถึงวัตถุประสงค์ความสำเร็จของการศึกษาสิทธิของผู้ป่วยในฐานะผู้มีส่วนร่วมในการศึกษานี้เพื่อขอเก็บข้อมูลจากการคัดกรองความเสี่ยงต่อโรคเบาหวานรายบุคคล โดยผู้วินิจฉัยได้นำข้อมูลปัจจัยเสี่ยงของโรคเบาหวาน จากการศึกษางานวิจัยที่กล่าวถึงปัจจัยเสี่ยงต่อโรคเบาหวาน โดยมีปัจจัยเสี่ยงคือ อายุ เพศ และประวัติโรคเบาหวานใน พ่อแม่ พี่น้อง เป็นต้น และเพื่อให้ได้ข้อมูลครอบคลุมและเจาะลึกการเก็บรวบรวมข้อมูลทั้งหมดโดยเป็นประชากรและกลุ่มข้อมูลตัวอย่างได้จากผู้ป่วยที่มารับบริการที่โรงพยาบาลมหาสารคาม จังหวัดมหาสารคาม

3.5 การทำให้ข้อมูลให้สมบูรณ์และการแปลงข้อมูล

เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่นำมาใช้วิเคราะห์ว่ามีความถูกต้อง เหมาะสม และครบถ้วน เพื่อลดความผิดพลาดที่จะเกิดขึ้นเนื่องจากการเก็บข้อมูลของผู้ป่วย โดยอาจมีการผิดพลาดของการเก็บสะสมข้อมูลเกิดขึ้นในขณะที่ทำการรวบรวมข้อมูลจากหลักฐานข้อมูลหลายๆ แหล่งเข้ามาเป็นหนึ่งเดียวเพื่อใช้การวิเคราะห์ในบางปีไม่ได้มีการเก็บข้อมูลผลที่ตามมาก็คือทำให้เกิดค่าสูญหายของข้อมูล (Missing Values)

ดังนั้น การทำให้ข้อมูลสมบูรณ์เป็นขั้นตอนที่กระทำก็คือการตรวจสอบข้อมูลสาเหตุที่ต้องตรวจสอบข้อมูลก็เพราะว่าอาจมีความผิดพลาดของการเก็บรวบรวมข้อมูลเกิดขึ้นในขณะที่ทำการรวบรวมข้อมูลจากฐานข้อมูลหลายๆแหล่งเข้ามาเป็นหนึ่งเดียวเพื่อใช้การวิเคราะห์ข้อมูลให้ถูกต้องสมบูรณ์ ได้แก่ การแก้ไขค่าว่างของข้อมูลซึ่งสามารถแก้ไขได้หลายวิธี เช่น แก้ไขโดยกำจัดข้อมูลที่มนแนวค่าว่าง (NULL)

การแปลงข้อมูลเป็นส่วนของการแปลงข้อมูลสำหรับขั้นตอนการทำเหมืองข้อมูลเมื่อกำหนดข้อมูลที่จะใช้ในการทำเหมืองข้อมูลได้แล้วบางครั้งข้อมูลที่กำหนดไว้นั้นรูปแบบของข้อมูลไม่เหมาะสมที่จะทำกระบวนการประมวลผลอาจต้องทำการแปลงหรือปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับวิธีการที่จะใช้ในการทำเหมืองข้อมูลซึ่งโดยปกติแล้วการแปลงข้อมูลจะถูกโดยเงื่อนไขของการปฏิบัติงานและวิธีการทำเหมืองข้อมูล

3.5.1ขั้นตอนการแปลงข้อมูล

ค่าตัวแปรก่อนการเรียนการรู้ด้วยกับแบบต่างๆและแทนค่าดังตารางที่ 2

ตารางที่ 3-1 การนำข้อมูลปัจจัยและสภาวะความเสี่ยงของโรคเบาหวานแทนค่าก่อนทำการแบ่งกลุ่ม

ปัจจัยเสี่ยง	ค่าของปัจจัยเสี่ยง	ค่าที่แปลงข้อมูลแล้ว
อายุ (ปี)	ตามอายุของผู้ป่วย	ตามอายุของผู้ป่วย
เพศ	ชาย หญิง	0 1
ประวัติการสูบบุหรี่	สูบบุหรี่ ไม่สูบบุหรี่	1 0
ประวัติการดื่มสุรา	ดื่มสุรา ไม่ดื่มสุรา	1 0
ญาติสายตรงเป็นโรคเบาหวาน	ไม่มี ญาติสายตรงเป็นโรคเบาหวาน ไม่ทราบ	0 1 2
ญาติสายตรงเป็นโรคความดันโลหิต	ไม่มี ญาติสายตรงเป็นโรคความดันโลหิต ไม่ทราบ	0 1 2
BMI	อ้วน ผอม ปกติ	1 0 2

ตารางแสดงผลของการแปลงข้อมูลเพื่อเตรียมข้อมูลก่อนทำการจัดกลุ่มข้อมูลโดยวิธีการ IDA โดย Microsoft Excel

3.5.2 ผลการวินิจฉัยทางการแพทย์

จากจำนวนผู้มารับการคัดกรองโรคเบาหวานทั้งหมด 5,344 คน ผลการวินิจฉัยจากแพทย์คือเป็นโรคเบาหวาน จำนวน 5,344 คน และข้อมูลที่เลือกมาใช้โดยได้แปลงชื่อปัจจัยเสี่ยงต่างๆเพื่อความสะดวกในการทำงานซึ่งมีรายละเอียดดังต่อไปนี้

SEX	เพศ
AGE	อายุ
SM	ประวัติการสูบบุหรี่
AL	ประวัติการดื่มสุรา
Relatives with diabetes	ญาติสายตรงเป็นโรคเบาหวาน
Cousin's blood pressure	ญาติสายตรงเป็นโรคความดันโลหิต
BMI	ดัชนีมวลร่างกาย

ตารางที่ 3-2 ตัวอย่างข้อมูลหลังจากการแทนค่าและแปลงข้อมูลสำหรับการจัดกลุ่มข้อมูล

SEX	AGE	SM	AL	RD	CP	BMI
2	44	0	0	2	0	1
2	62	0	0	1	0	1
2	60	0	0	1	0	1
1	72	0	0	1	0	1
1	82	0	0	1	0	0
1	44	0	0	1	0	1
1	64	0	0	2	1	1
1	67	0	0	1	0	1
1	56	0	0	1	0	1
1	66	0	0	1	0	1

ตารางจะแสดงตัวอย่างข้อมูลที่ทำการเตรียมพร้อมสำหรับการใช้งานในขั้นตอน จัดกลุ่ม โดยข้อมูลที่เตรียมไว้ได้ผ่านการตรวจสอบความถูกต้อง และ การแปลงข้อมูลเรียบร้อยแล้ว

3.5.3 การจัดกลุ่ม

เมื่อได้ข้อมูลที่ถูกต้องพร้อมแล้วในขั้นตอนนี้จะเป็นการแบ่งกลุ่มของผู้ป่วยโดยผู้วิจัยได้แบ่งกลุ่มของผู้ป่วยออกเป็นสองกลุ่มคือ กลุ่มที่เป็นโรคเบาหวาน และกลุ่มที่ไม่เป็นโรคเบาหวาน **โดย**มุ่งเน้นที่จะพิจารณาความถี่ของข้อมูลในแต่ล่ะกลุ่มเพื่อที่จะได้ความสำคัญของแต่ละปัจจัยโดยขั้นตอนการทำงานในเทคนิคของการจัดกลุ่มโดยใช้โปรแกรม Excel โดยเพิ่มโปรแกรมเสริม IDA ในการแบ่งกลุ่ม ซึ่งมีขั้นตอนดังนี้

3.5.3.1 กำหนดว่าเราจะแบ่งประชากรออกเป็นกี่กลุ่ม

3.5.3.2 ทำการกำหนดค่าของกลางแต่ละกลุ่มโดยสุ่มมาจากประชากรที่มีอยู่

3.5.3.3 ทำการนำประชากรที่มีอยู่จับเข้าแต่ละกลุ่มโดยแต่ละค่าต้องใกล้กับค่ากลางของกลุ่มนั้นมากที่สุดในแต่ละครั้งมีการจับเข้ากลุ่มต้องคิดค่ากลางของกลุ่มนั้นๆใหม่เสมอโดยหาจากค่าเฉลี่ยของประชากรในกลุ่มทำไปจนกว่าจำนวนประชากรจะถูกจัดเข้าทุกกลุ่มจนครบ

3.5.3.4 ทำการตรวจสอบว่ากรณีที่ได้มานี้ผลของฟังก์ชันจุดประสงค์นั้นมีค่าต่ำสุดหรือค่าไม่เปลี่ยนแปลงหรือวนครบจำนวนรอบที่กำหนดไว้

บทที่ 4
ผลการวิจัยและอภิปรายผล

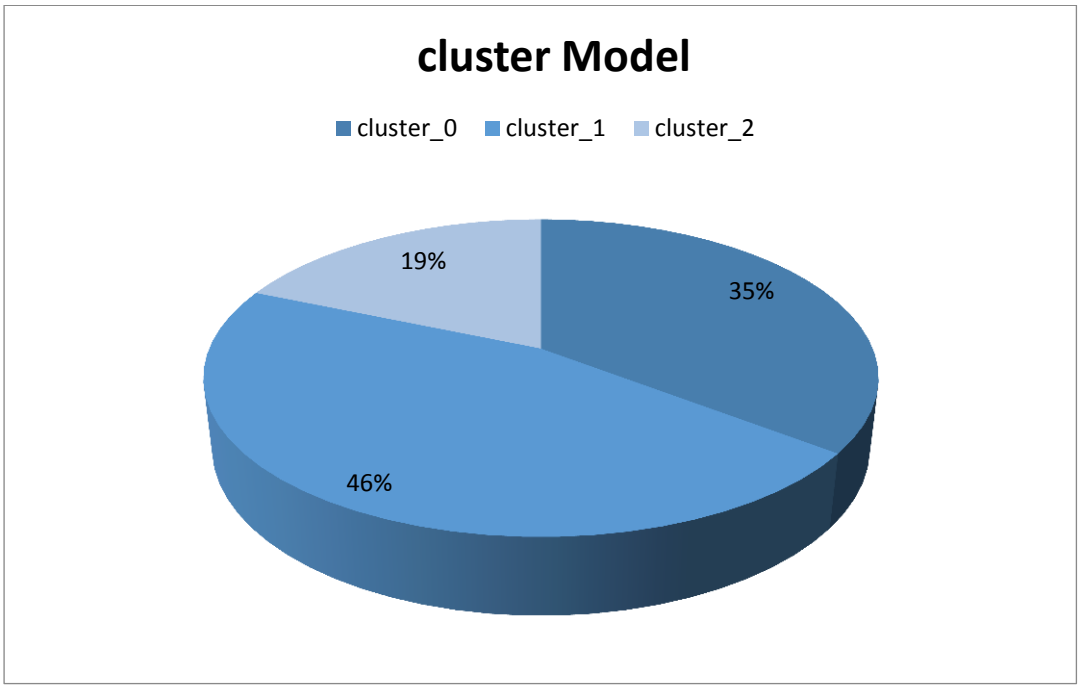
เนื้อหาในบทนี้กล่าวถึงการวิเคราะห์ผลการวิจัยที่ได้ทดลองการจัดกลุ่มของผู้ป่วยโรคเบาหวานจากนั้นได้ทำการวิเคราะห์ปัจจัยการเกิดโรคเบาหวานและสุดท้ายได้ทำการพยากรณ์ความเสี่ยงของโรคเบาหวานโดยวิธีการจัดกลุ่ม (clustering) โดยที่กล่าวมาทั้งหมดจะถูกพัฒนาตามขั้นตอนการดำเนินการวิจัยที่ได้อธิบายในบทที่ 3 ซึ่งในการอธิบายผลการพัฒนาในบทนี้จะอภิปรายถึงการวิจัยการพัฒนาการใช้ทฤษฎีการทำเหมืองข้อมูลในการเกิดโรคเบาหวานประกอบด้วย

4.1 ผลการจัดกลุ่มผู้ป่วยโรคเบาหวาน

งานวิจัยนี้จัดทำขึ้นเพื่อทำการทดลองการความเสี่ยงของการเกิดโรคเบาหวาน โดยใช้โปรแกรมสำเร็จรูป Microsoft Excel เป็นเครื่องมือที่ช่วยในการวิเคราะห์โดยการนำเข้าซึ่งปัจจัยต่างๆ ที่กำหนดได้แก่ อายุ เพศ ประวัติการสูบบุหรี่ ประวัติการดื่มแอลกอฮอล์ ประวัติด้านพันธุกรรม โดยการวิเคราะห์สัมประสิทธิ์การจัดกลุ่มหาได้จากการนำตัวแปรเข้าไปในสมการแล้วพิจารณาสัมประสิทธิ์ของแต่ละตัวแปรและแบ่งตาม cluster Model ผลปรากฏดังตารางที่ 4 ดังต่อไปนี้

ตารางที่ 4-1 cluster Model

Cluster 0 (Items)	Cluster 1 (items)	Cluster 2 (items)	Total number of items
1,891	2,456	996	5,343



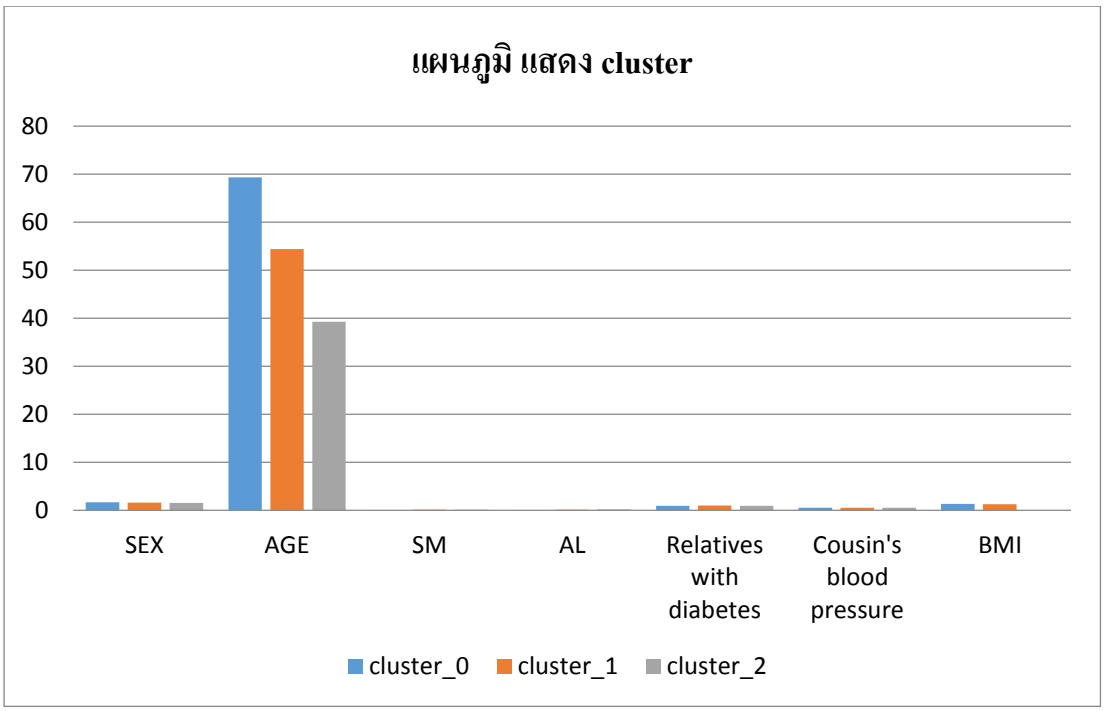
ภาพที่ 4–1 Charts แผนภาพสถิติ

จากแผนภาพสถิติจะแสดงค่า cluster Model โดยใช้ทฤษฎี K-Means ซึ่งได้ผ่านกระบวนการจัดกลุ่ม แบ่งออกได้ดังนี้ cluster 0 มีจำนวน 1,891 คน cluster 1 มีจำนวน 2,456 และ cluster 2 มีจำนวน 996 คน โดยมีจำนวนผู้ป่วยโรคเบาหวานทั้งหมด 5,343 คน และจะเห็นได้ว่า cluster_1 จะมีความเสี่ยงมากที่สุดถึง 46% cluster_0 มีระดับความเสียง 35% และ cluster_2 มีค่าความเสี่ยงน้องที่สุด 19%

ตารางที่ 4-2 ค่าของปัจจัยต่างๆ ที่ได้จากการจัดกลุ่ม

Attribute	Cluster_0	Cluster_1	Cluster_2
SEX	1	1	1
AGE	69	54	39
SM	0	0	0
AL	0	0	0
Relatives with diabetes	1	1	1
Cousin's blood pressure	0	1	1
BMI	1	1	1

จากตารางจะแสดงถึงการจัดกลุ่มของข้อมูลผู้ป่วย ซึ่งสามารถจัดกลุ่มได้ดังนี้ Attribute แสดงให้เห็นถึงค่าปัจจัยต่างๆ เช่น อายุ เพศ ประวัติการสูบบุหรี่ ประวัติการดื่มสุรา ญาติสายตรงเป็นโรคเบาหวาน ญาติสายตรงเป็นโรคความดันโลหิต และค่าดัชนีมวลกาย ซึ่งค่าของจำนวนสมาชิกที่ได้มาใช้ในการคำนวณหาค่าความเสี่ยงของการเกิดโรคเบาหวานต่อไป

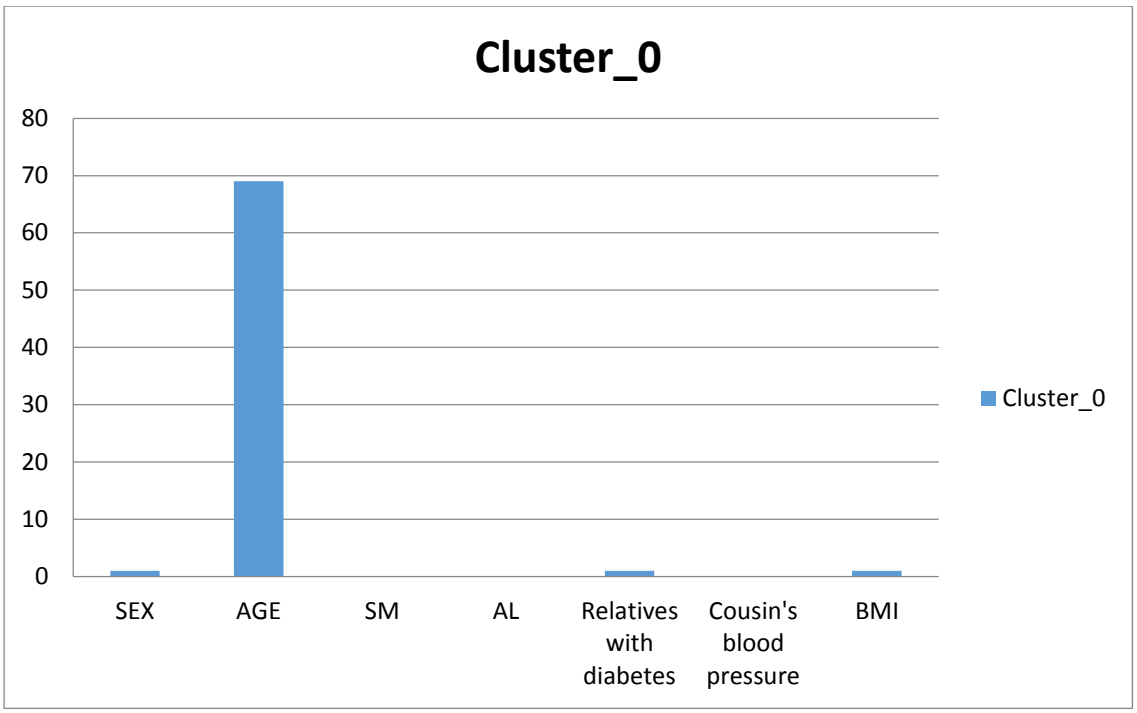


ภาพที่ 4-2 แผนภูมิ แสดง cluster_0 , cluster_1 , cluster_2

ตารางที่ 4-3 ผลการวัดค่าความเสี่ยง **Cluster_0**

Attribute	Cluster_0
SEX	1
AGE	69
SM	0
AL	0
Relatives with diabetes	1
Cousin's blood pressure	0
BMI	1

จากตารางที่ 4-3 แสดงค่าสถิติทดสอบใน Cluster_0 ซึ่งเป็นสถิติอยู่ในเกณฑ์ที่ผอม การเกิดโรควัดได้จากการคำนวณจากค่าดัชนีมวลกาย (BMI) เข้าใกล้ 1 มีค่าเป็น อ้วน , เพศ (Sex) เข้าใกล้ 1 มีค่าเป็น ผู้หญิง , อายุ (AGE) จะอยู่ในช่วงอายุ 69 ปี , ประวัติการสูบบุหรี่เฉลี่ย (SM) เข้าใกล้ 0 มีค่าเป็น ไม่สูบบุหรี่, ประวัติการดื่มแอลกอฮอล์เฉลี่ย (AL) เข้าใกล้ 0 มีค่าเป็น ไม่ดื่มแอลกอฮอร์ , ประวัติของญาติที่เป็นโรคเบาหวาน (Relatives with diabetes) เข้าใกล้ 1 มีค่าเป็น มีญาติที่เป็นโรคเบาหวานโดยตรง และ ญาติสายตรงเป็นโรคความดันโลหิต (Cousin's blood pressure) เข้าใกล้ 0 มีค่าเป็น ไม่มี นั่นแสดงว่าค่าเฉลี่ยใน Cluster_0 จัดอยู่ในเกณฑ์ที่มีความเสี่ยงน้อยที่จะเป็นโรคเบาหวาน

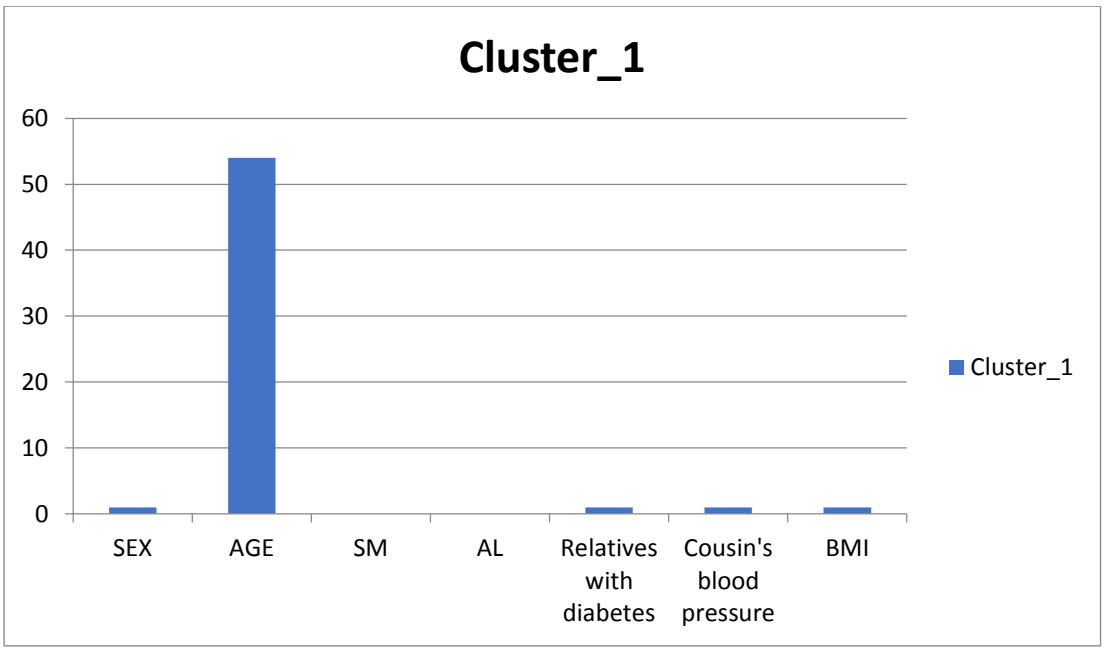


ภาพที่ 4-3 แผนภูมิ แสดง cluster_0

ตารางที่ 4-4 ผลการวัดค่าความเสี่ยง **Cluster_1**

Attribute	Cluster_1
SEX	1
AGE	54
SM	0
AL	0
Relatives with diabetes	1
Cousin's blood pressure	1
BMI	1

จากตารางแสดงค่าสถิติทดสอบใน Cluster_1 ซึ่งเป็นสถิติอยู่ในเกณฑ์ที่อ้วน การเกิดโรควัดได้จากการคำนวณจากค่าดัชนีมวลกาย (BMI) เข้าใกล้ 1 มีค่าเป็น อ้วน , เพศ (Sex) เข้าใกล้ 1 มีค่าเป็น ผู้หญิง , อายุ (AGE) จะอยู่ในช่วงอายุ 54 ปี , ประวัติการสูบบุหรี่เฉลี่ย (SM) เข้าใกล้ 0 มีค่าเป็น ไม่สูบบุหรี่, ประวัติการดื่มแอลกอฮอล์เฉลี่ย (AL) เข้าใกล้ 0 มีค่าเป็น ไม่ดื่มแอลกอฮอร์ , ประวัติของญาติที่เป็นโรคเบาหวาน (Relatives with diabetes) เข้าใกล้ 1 มีค่าเป็น มีญาติที่เป็นโรคเบาหวานโดยตรง และ ญาติสายตรงเป็นโรคความดันโลหิต (Cousin's blood pressure) เข้าใกล้ 1 มีค่าเป็น มีญาติสายตรงเป็นโรคความดันโลหิต นั่นแสดงว่าค่าเฉลี่ยใน Cluster_1 จัดอยู่ในเกณฑ์ที่มีความเสี่ยงมากที่สุดที่จะเป็นโรคเบาหวาน

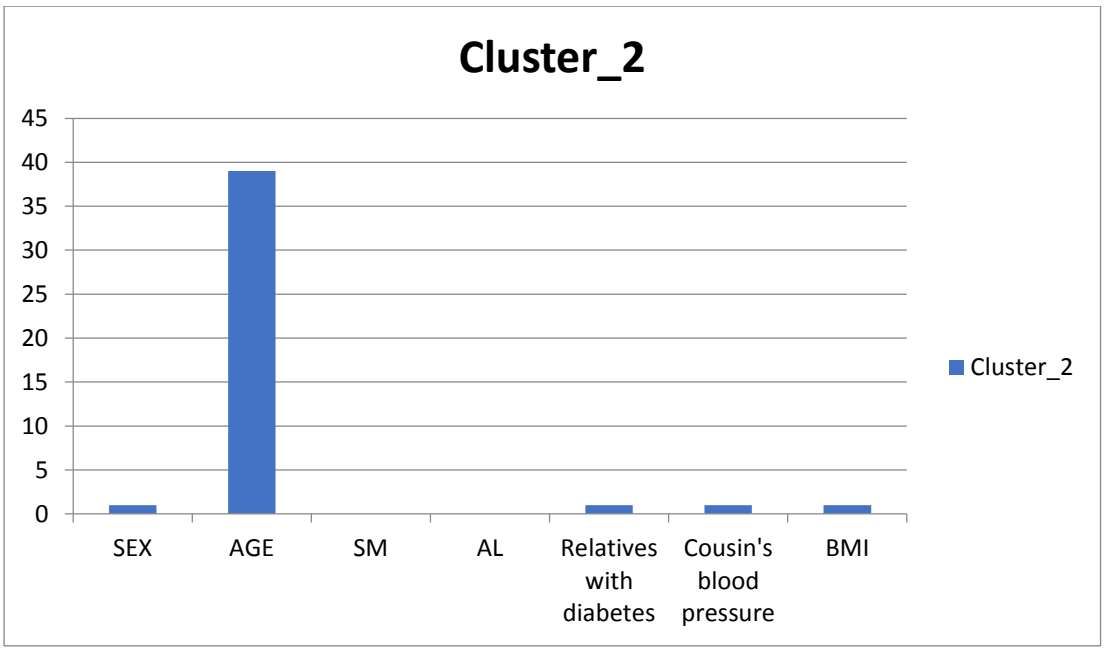


ภาพที่ 4-4 แผนภูมิ แสดง cluster_1

ตารางที่ 4-5 ผลการวัดค่าความเสี่ยง Cluster_2

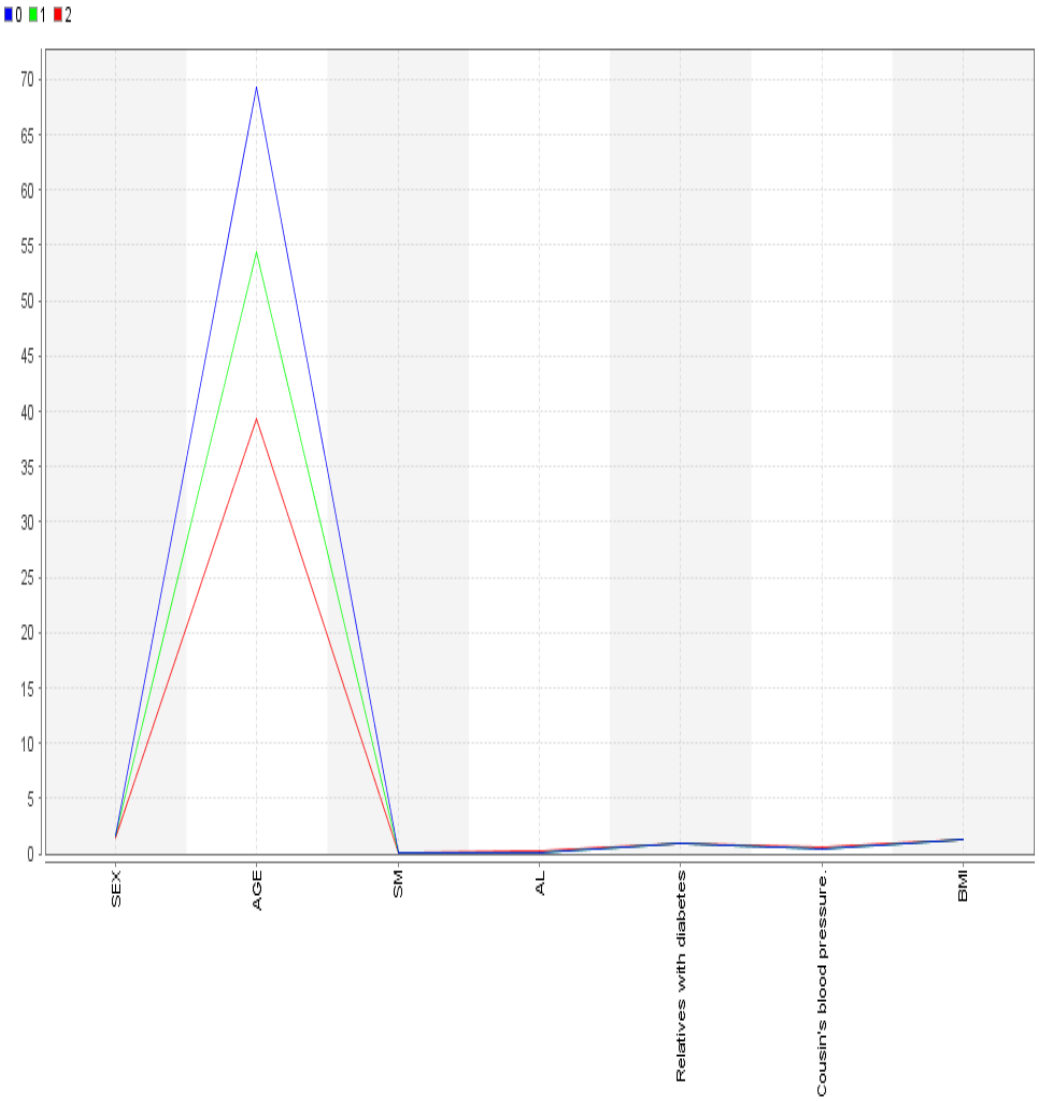
Attribute	Cluster_2
SEX	1
AGE	39
SM	0
AL	0
Relatives with diabetes	1
Cousin's blood pressure	1
BMI	1

จากตารางแสดงค่าสถิติทดสอบใน Cluster_2 ซึ่งเป็นสถิติอยู่ในเกณฑ์ที่ผอม การเกิดโรควัดได้จากการคำนวณจากค่าดัชนีมวลกาย (BMI) เข้าใกล้ 1 มีค่าเป็น อ้วน , เพศ (Sex) เข้าใกล้ 1 มีค่าเป็น ผู้หญิง **, อายุ (** AGE) จะอยู่ในช่วงอายุ 39 ปี , ประวัติการสูบบุหรี่เฉลี่ย (SM) เข้าใกล้ 0 มีค่าเป็น ไม่สูบบุหรี่, ประวัติการดื่มแอลกอฮอล์เฉลี่ย (AL) เข้าใกล้ 0 มีค่าเป็น ไม่ดื่มแอลกอฮอร์ , ประวัติของญาติที่เป็นโรคเบาหวาน (Relatives with diabetes) เข้าใกล้ 1 มีค่าเป็น มีญาติที่เป็นโรคเบาหวานโดยตรง และ ญาติสายตรงเป็นโรคความดันโลหิต (Cousin's blood pressure) เข้าใกล้ 1 มีค่าเป็น มีญาติสายตรงเป็นโรคความดันโลหิต นั่นแสดงว่าค่าเฉลี่ยใน **Cluster_2** จัดอยู่ในเกณฑ์ที่มีความเสี่ยงน้อยที่สุดที่จะเป็นโรคเบาหวาน



ภาพที่ 4-5 แผนภูมิ แสดง cluster_2

ภาพที่ 1 แสดง Cluster Model ให้เห็นถึง กราฟ (Plot) ของการเกิดโรคเบาหวาน



ภาพที่ 4-1 กราฟแสดง Cluste (อ้วน) , (ผอม) , (ปกติ)

แสดงให้เห็นถึงค่าที่จะเกิดความเสี่ยงต่อการเป็นโรคเบาหวานซึ่งดูได้จากค่าที่แสดงที่อยู่ในกราฟ

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยฉบับนี้มีวัตถุประสงค์ที่จะศึกษาและสร้างระบบที่สามารถจัดกลุ่มของประชากรที่เป็นโรคเบาหวาน เพื่อที่จะทราบถึงปัจจัยใดบ้างที่ทำให้เกิดโรคเบาหวานและนำค่าของปัจจัยเสี่ยงมาวิเคราะห์ว่ามีค่าความเสี่ยงที่จะเป็นโรคเบาหวานหรือไม่ทั้งนี้หากพบว่าผู้ป่วยคนใดมีโอกาสที่จะเป็นโรคเบาหวาน แพทย์ผู้รักษาสามารถตรวจคนไข้อย่างละเอียดอีกครั้งเพื่อให้ทันท่วงทีในการรักษา อีกทั้งยังสารถช่วยผู้บริหารในการตัดสินใจกำหนดนโยบายที่จะป้องกันและดูแลประชากรให้รวดเร็วและถูกต้องขึ้น

โดยข้อมูลที่นำมาใช้ในทดลอง คือ ข้อมูลประวัติส่วนตัวของผู้ป่วยจากโรงพยาบาลมหาสารคาม ข้อมูลผู้ป่วยโรคเบาหวาน ตั้งแต่ปี 2556 -2558 จำนวน 5,343 คน โดยเป็นผู้ป่วยโรคเบาหวาน 5,343 คน จากนั้นได้ซึ่งศึกษาและวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน โดยใช้ปัจจัยเฉพาะด้าน ได้แก่ เพศ อายุ ประวัติการดื่มแอลกอฮอร์ ประวัติการสูบบุหรี่ ประวัติด้านพันธุกรรม และ ค่าดัชนีมวลกาย (BMI) ซึ่งในการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวา คือ Clustering โดยใช้ K-Means

โดยข้อมูลที่นำมาใช้ในการทดลอง คือข้อมูลประวัติส่วนตัวของผู้ป่วยโรคเบาหวานจากโรงพยาบาลมหาสารคาม จำนวน 5,343 คน โดยเป็นผู้ป่วยโรคเบาหวานทำการแบ่งกลุ่มออกเป็น 3 cluster คือ cluster_0 มีจำนวน 1,891 คน โดยพบว่ามีอายุเฉลี่ย 69 ปี เป็นเพศหญิง ประวัติการสูบบุหรี่ไม่สูบบุหรี่ , ประวัติการดื่มแอลกอฮอล์ไม่ดื่มแอลกอฮอร์ , มีญาติที่เป็นโรคเบาหวานโดยตรง และ มีญาติสายตรงเป็นโรคความดันโลหิต นั่นแสดงว่าค่าเฉลี่ยใน **Cluster_0** จัดอยู่ในเกณฑ์ที่มีความเสี่ยงปกติที่จะเป็นโรคเบาหวาน

cluster_1 มีจำนวน 2,456 คน โดยพบว่ามีอายุเฉลี่ย 54 ปี เป็นเพศหญิง ประวัติการสูบบุหรี่ไม่สูบบุหรี่ , ประวัติการดื่มแอลกอฮอล์ไม่ดื่มแอลกอฮอร์ , มีญาติที่เป็นโรคเบาหวานโดยตรง และ มีญาติสายตรงเป็นโรคความดันโลหิต นั่นแสดงว่าค่าเฉลี่ยใน **Cluster_1** จัดอยู่ในเกณฑ์ที่มีความเสี่ยงมากที่สุดที่จะเป็นโรคเบาหวาน

cluster_2 มีจำนวน 996คน โดยพบว่ามีอายุเฉลี่ย 39 ปี เป็นเพศหญิง ประวัติการสูบบุหรี่ไม่สูบบุหรี่ , ประวัติการดื่มแอลกอฮอล์ไม่ดื่มแอลกอฮอร์ , มีญาติที่เป็นโรคเบาหวานโดยตรง และ มีญาติสายตรงเป็นโรคความดันโลหิต นั่นแสดงว่าค่าเฉลี่ยใน **Cluster_2** จัดอยู่ในเกณฑ์ที่มีความเสี่ยงน้อยที่สุดที่จะเป็นโรคเบาหวาน

ผลการทดลองพบว่าจากการจัดกลุ่มข้อมูล ทำให้สามารถรู้ถึงค่าความเสี่ยงของการเกิดโรคเบาหวานในแต่ละปัจจัย เมื่อวิเคราะห์หาค่าความเสี่ยงแล้วจะทำให้สามารถตัดปัจจัยที่ไม่มีความสำคัญออกไปได้ โดยผลที่ได้สามารถแบ่งกลุ่มออกเป็น 3 กลุ่มคือ Class 0 , Class 1

และ Class 2 โดยพบว่า Class 1 เป็นกลุ่มของผู้ป่วยที่เป็นโรคเบาหวานที่มีความเสี่ยงมากที่สุดถึง 2,456 คน โดยปัจจัยมีค่าความเสี่ยงมากที่สุด คือ ค่าดัชนีมวลกาย (BMI) ที่มีค่าอ้วน และประวัติด้านพันธุกรรม ส่วนค่าปัจจัยที่มีความเสี่ยงน้อยที่สุด คือ ประวัติการดื่มแอลกอฮอร์ และประวัติการสูบบุหรี่

ปัจจัยเสี่ยงของโรคเบาหวาน

1.) อายุของผู้ป่วยโรคเบาหวานอายุเฉลี่ย 54 ปี

2.) เพศของผู้ป่วยโรคเบาหวานเพศหญิง

3.) ประวัติการสูบบุหรี่ไม่สูบ

4.) ประวัติกาดื่มแอลกอฮอร์ไม่ดื่ม

5.) มีญาติที่เป็นโรคเบาหวานโดยตรง

6.) มีญาติสายตรงเป็นโรคความดันโลหิต

7.) ค่าดัชนีมวลกาย (BMI) อ้วน

5.2 ข้อเสนอแนะ

ผลการวิเคราะห์ความเสี่ยงการเกิดโรคเบาหวานที่ได้ เป็นเพียงเครื่องมือในการคัด

กรองหรือประเมินความเสี่ยงเบื้องต้นแต่ปัจจัยที่ทำให้เกิดโรคเบาหวานแท้จริงแล้วอาจมาจากหลายปัจจัยตามแต่ละพื้นที่ของข้อมูลชุดที่สร้างข้น จากแต่ละตัวแปรที่ค้นพบจึงเป็นเรื่องสำคัญในการหาค่าความเสี่ยง ซึ่งมีการทดสอบตัวแปรเหล่านี้ในจำนวนประชากรที่มากขึ้นและทดลองกับชุดข้อมูลต่างสถานที่ และควรศึกษาติดตามประชากรกลุ่มเสี่ยงที่หลากหลายในระยะยาวต่อไป

บรรณานุกรม

นายสิทธิชัย คำคง.การทาเหมืองข้อมูล (Data Mining) [โรคงงานพิเศษเทคโนโลยีสารสนเทศและการสื่อสารทางเทคนิคศึกษาภาควิชาครุศาสตร์เทคโนโลยีคณะครุศาสตร์อุตสาหกรรม “เทคนิคการจัดกลุ่มแบบ K-means (K-means Clustering:การจัดกลุ่มแบบ K-meansClustering http://datamining-techniques.blogspot.com/2012/09/k-means-k-meansclustering.html)

รักถิ่น เหลาหา: 2553 การพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอด โดยใช้ทฤษฎีของการทำเหมืองข้อมูล กัลยา วานิชย์บัญชา. (2552). การวิเคราะห์ข้อมูลหลายตัวแปร. กรุงเทพมหานคร ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย พิมพ์ครั้งที่ 4

อัจฉราพร สี่หิรัญวงศ์ และรณชัย คงสกนธ์. (2541). แบบวัด Hamilton Rating Scale for Depression : การวิเคราะห์การรวมกลุ่ม วารสารสมาคมจิตแพทย์ แห่งประเทศไทย

อำนวย มณีศรีวงศ์กูล. (2541). **Cluster Analysis** . วารสารวิจัย มหาวิทยาลัยขอนแก่น

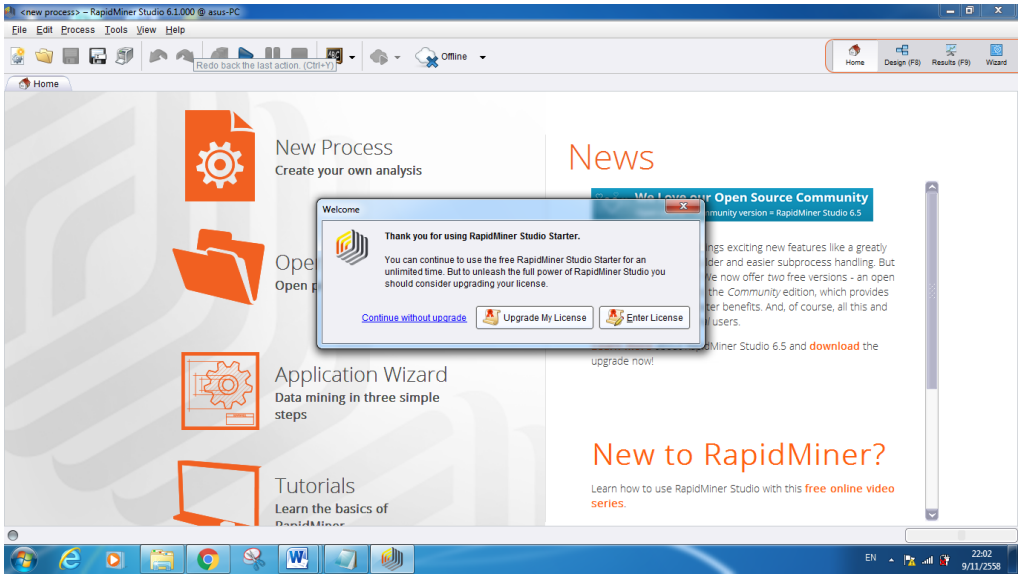
อัจฉราพร สี่หิรัญวงศ์ และรณชัย คงสกนธ์. (2541). แบบวัด Hamilton Rating Scale for Depression : การวิเคราะห์การรวมกลุ่ม. วารสารสมาคมจิตแพทย์แห่งประเทศไทย

ภาคผนวก

คู่มือการใช้งานโปรแกรม Repidminer Studio 6

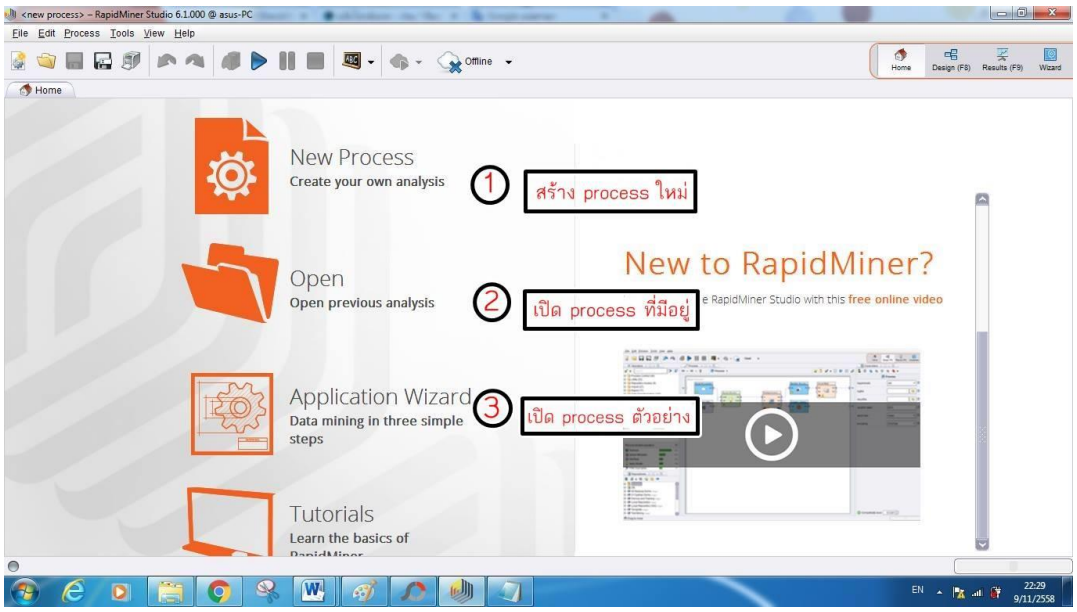
ภาคผนวก

คู่มือการใช้งานโปรแกรม Repidminer Studio 6

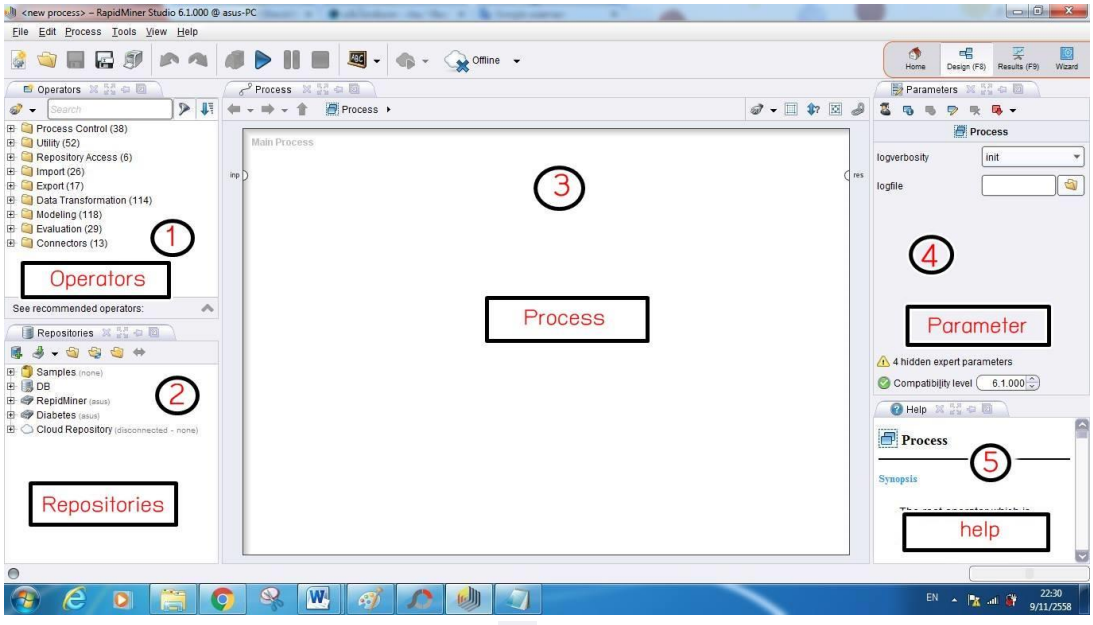


ภาพที่ ผ – 1 เข้าสู่หน้า Downloads และ Licenses

ภาพที่ ผ – 2 Copy Key ของ License และไปใส่ใน RepidMiner Studio 6 กดปุ่ม Install License



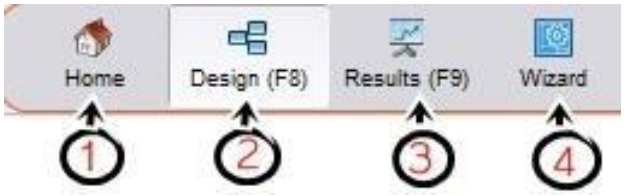
ภาพที่ ผ – 3 หน้าต่าง Home Scessn



ภาพที่ ผ – 4 องค์ประกอบของ RepidMiner Studio 6

ภาพที่ ผ – 5 เมนูใน RepidMiner Studio 6

	ใช้สร้าง Process ใหม่		undo หรือ redo
	โหลด Process เดิม		สั่งให้ Process ทำงาน (run)
	บันทึก Process		หยุด Process ชั่วคราว (pause)
	บันทึก Process ที่เป็นชื่อใหม่		ยกเลิก Process (stop)
	พิมพ์ Process ออกทาง printer		เรียกดู tutorial



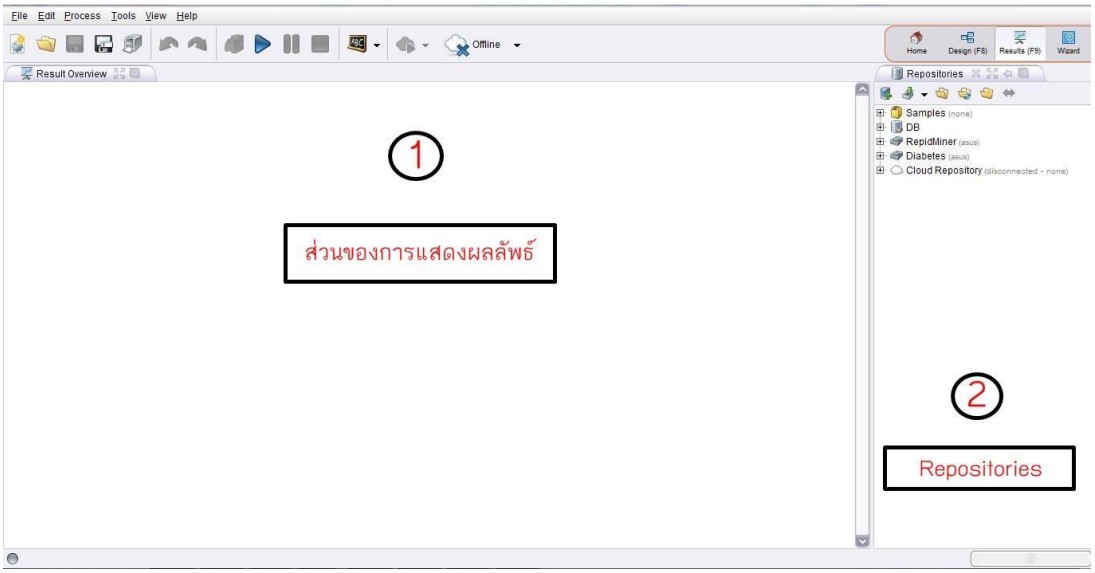
ภาพที่ ผ – 6 หน้าจอ (perspective)

(1) หน้า Home กลับไปหน้าเริ่มต้นของ RepidMiner Studio 6

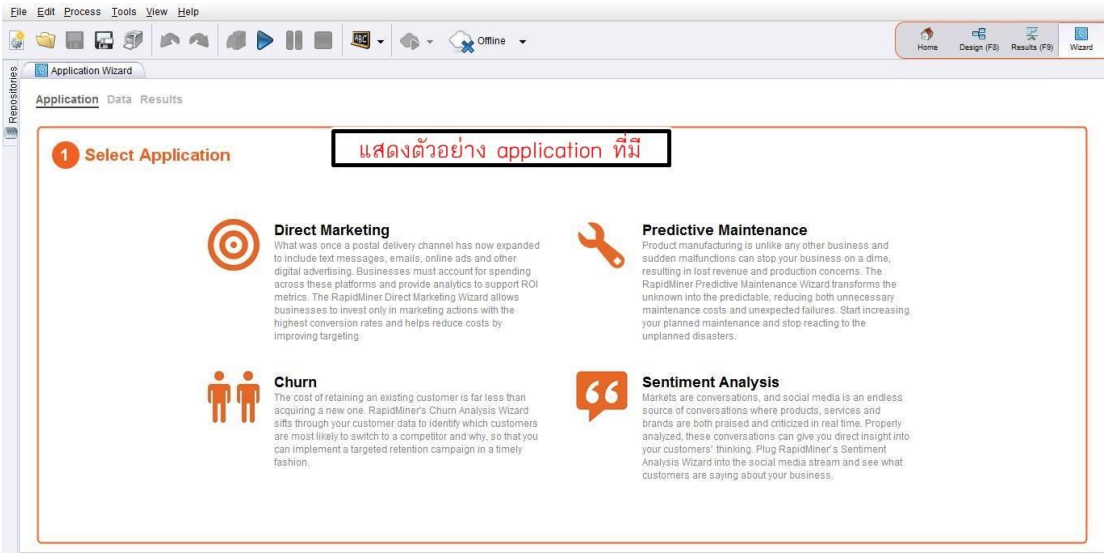
(2) หน้า Desing แสดงหน้าสำหรับการสร้าง Process

(3) หน้า Resulis แสดงหน้าผลลัพธ์การทำงาน

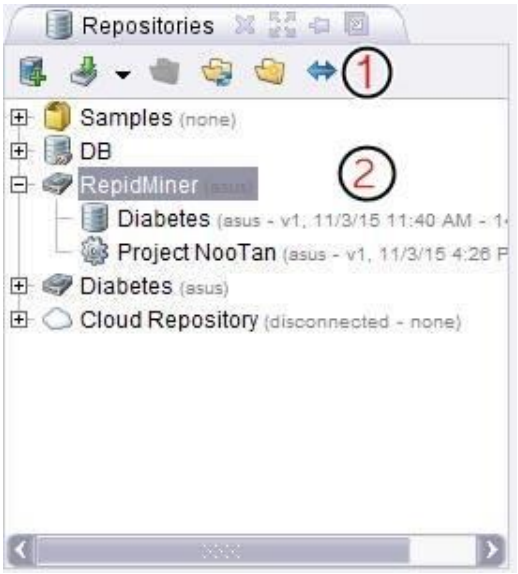
(4) หน้า Wizard แสดงตัวอย่างระบบที่เตรียมไว้ให้



ภาพที่ ผ – 7 หน้า Results



ภาพที่ ผ – 8 หน้า Wizard



ภาพที่ ผ – 9 Repositories

- Repositories เป็นที่เก็บข้อมูลและ process เพื่อใช้งานใน RepidMiner Studio 6
- ทำให้ไม่ต้องโหลดข้อมูลจากไฟล์ใหม่ทุกครั้ง

ส่วนที่ 1.

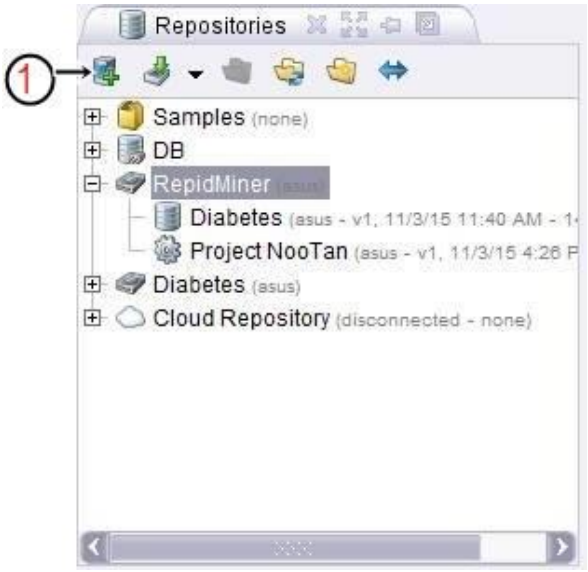
สำหรับสร้าง Repository ใหม่

โหลดไฟล์ประเภทต่างๆ เข้าไปไว้ใน Repository

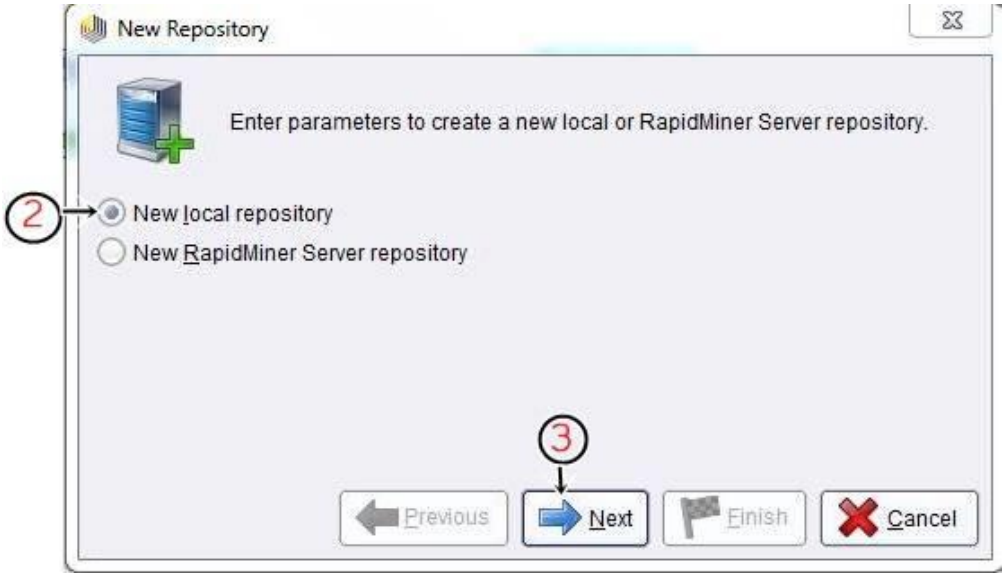
สร้างโฟลเดอร์ใหม่

ส่วนที่ 2.

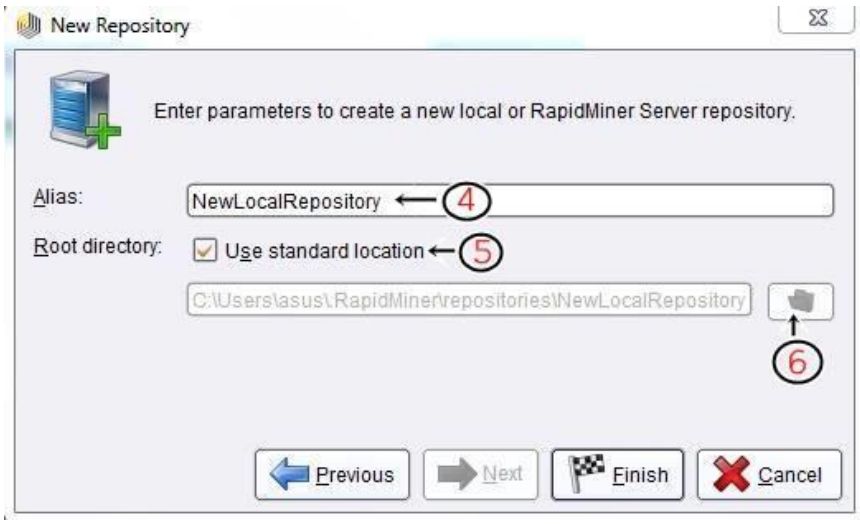
- ข้อมูลและ processSample ที่ RepidMiner Studio 6 ที่เตรียมไว้ให้
- ข้อมูลที่เก็บในแต่ล่ะ Repository



ภาพที่ ผ – 10 สร้าง Repository ใหม่ คลิก ที่

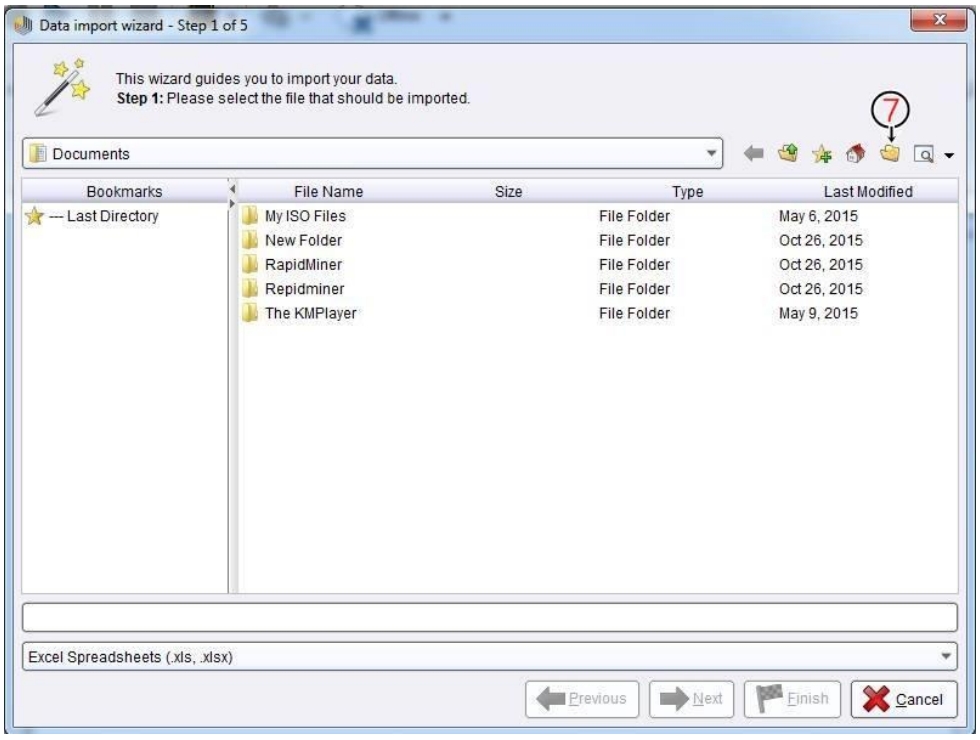


ภาพที่ ผ – 11 เลือก New local Repository กดปุ่ม Next

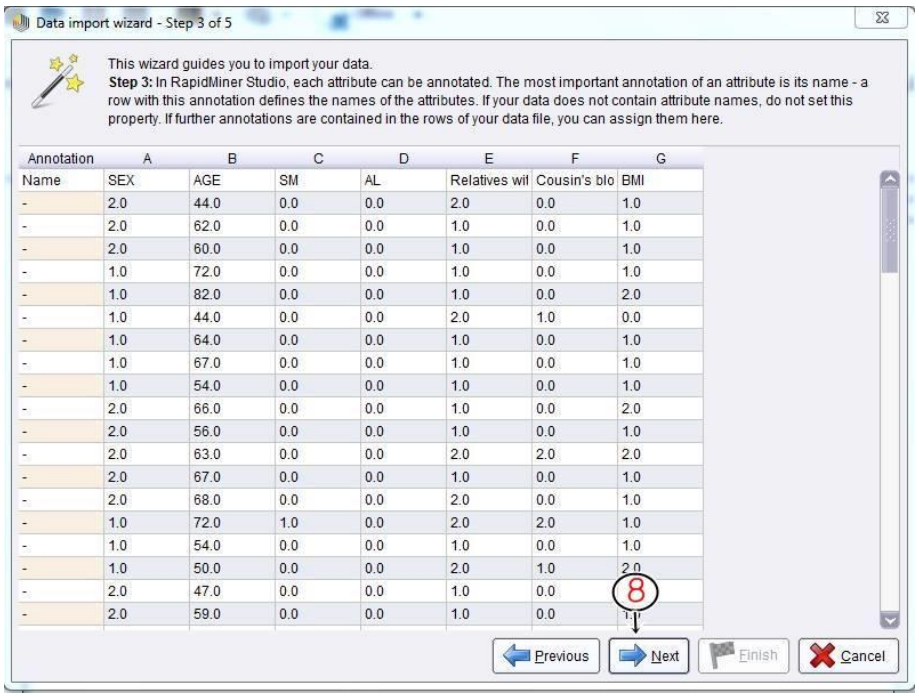


ภาพที่ ผ – 12 สร้าง Repository ใหม่ (ต่อ)

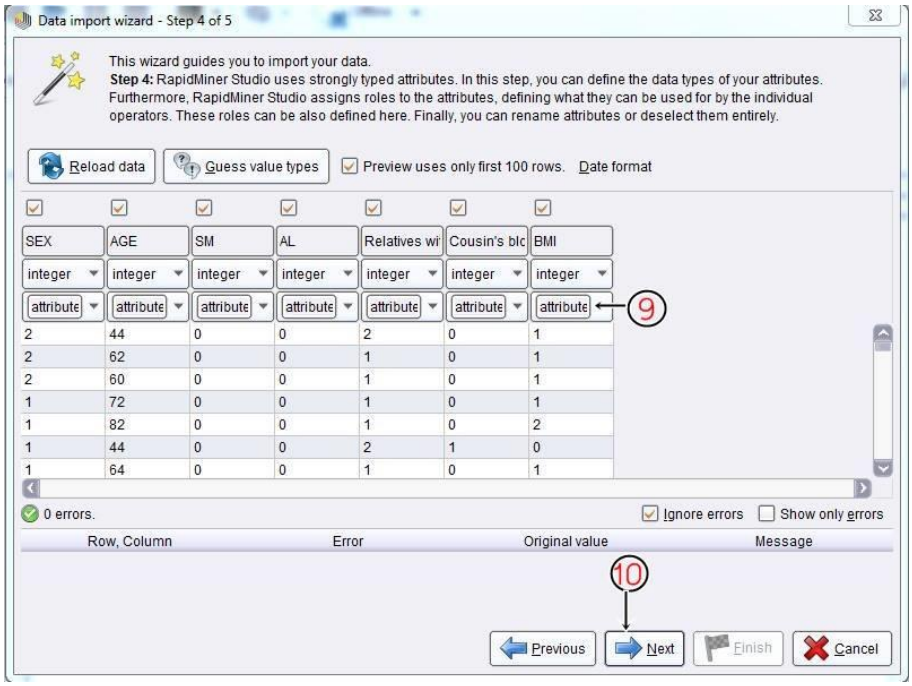
- เปลี่ยนชื่อ Alias เป็น RepidMinerTraining
- คลิกที่ Use stsndard location เพื่อเอาออก
- คลิกที่ไอคอน Folder เพื่อเลือก Root directory



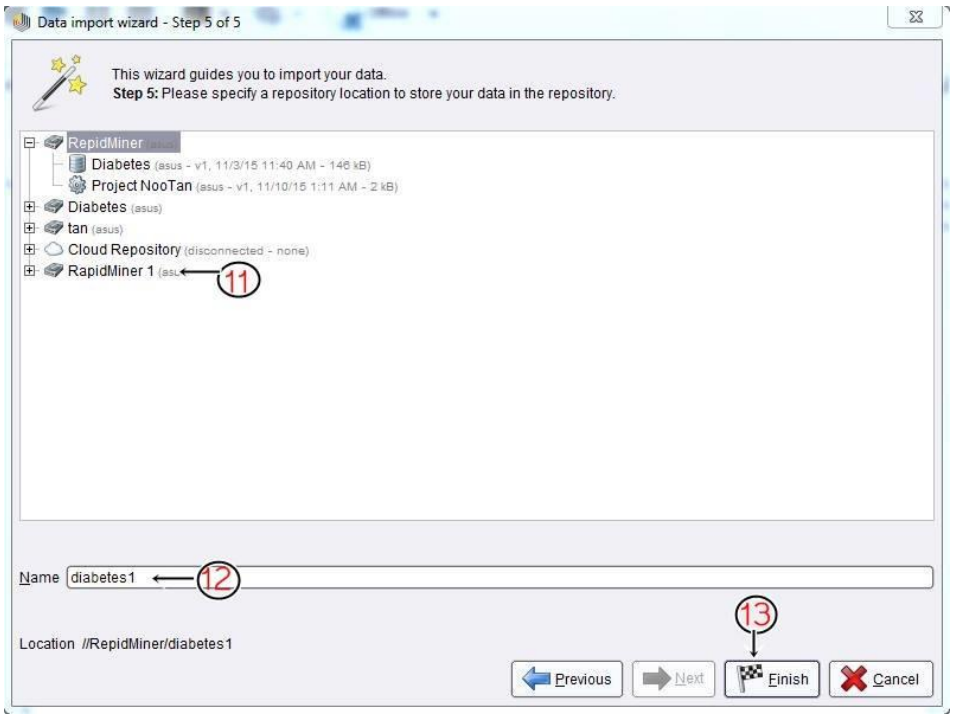
ภาพที่ ผ – 13 สร้าง Repository ใหม่ (ต่อ) คลิกที่ไอคอน  เพื่อสร้างโฟล์เดอร์ใหม่



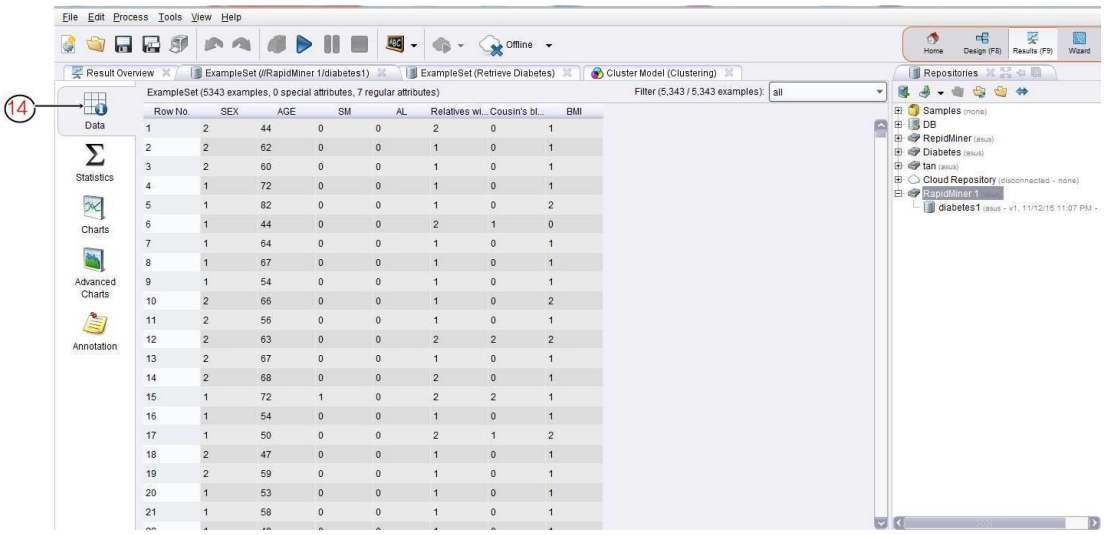
ภาพที่ ผ – 14 โหลดไฟล์เข้าไปไว้ใน Repository แล้วคลิกปุ่ม Next



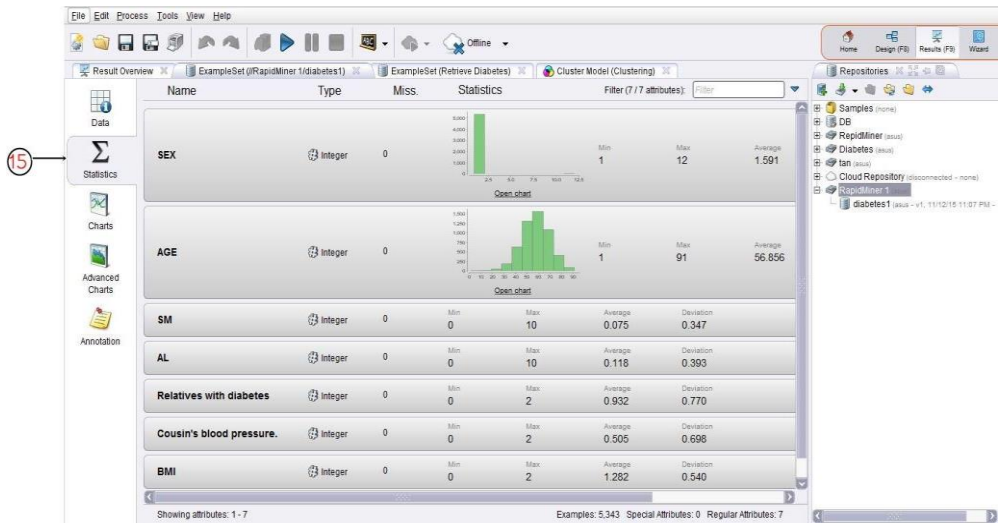
ภาพที่ ผ – 15 โหลดไฟล์เข้าไปไว้ใน Repository แล้วคลิกปุ่ม Next



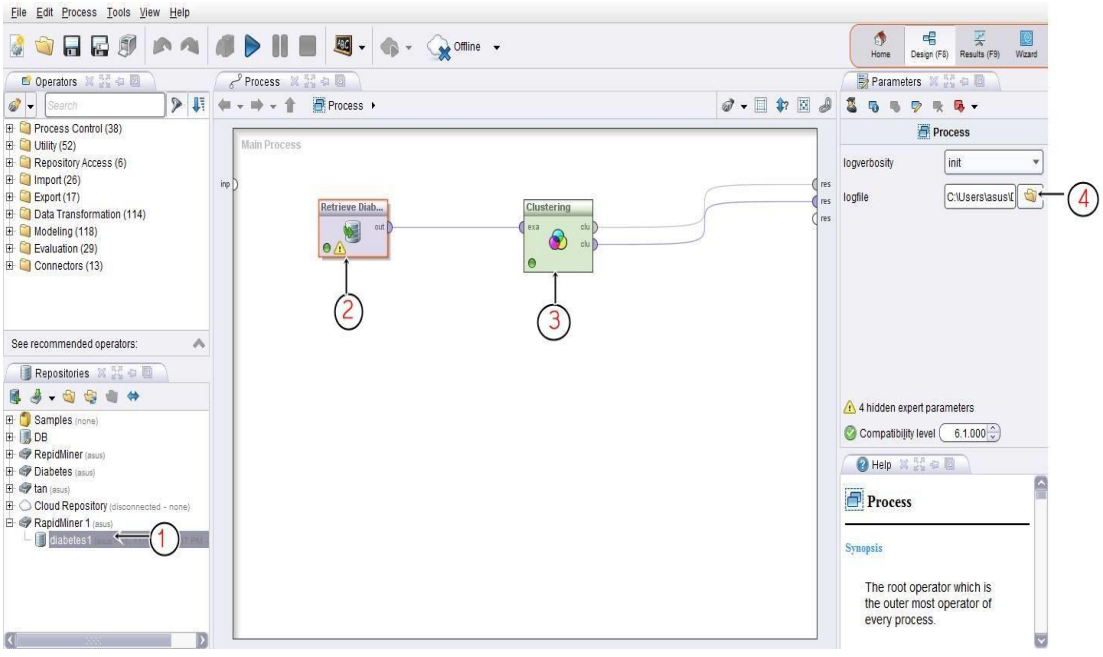
ภาพที่ ผ – 16 โหลดไฟล์เข้าไปไว้ใน Repository save ชื่อว่า diabetes ไว้ที่ RapidMiner 1 แล้วกด Finish



ภาพที่ ผ – 17 ข้อมูลที่โหลดเข้าไปแสดงในรูปแบบของตาราง



ภาพที่ ผ – 18 ข้อมูลที่โหลดเข้าไปแสดงในรูปแบบของค่าสถิติ



ภาพที่ ผ – 19 เขียนข้อมูลลงในหน้า Process ใช้ข้อมูลจาก Repositories

ประวัติผู้เขียน

ชื่อ นางสาวภาวิณี ธรรมเกตุ

รหัสนักศึกษา 553120100308

วันเดือนปีเกิด 16 พฤศจิกายน 2536

ภูมลำเนา อ.อาจสามารถ จังหวัดร้อยเอ็ด

ประวัติการศึกษา สำเร็จการศึกษามัธยมตอนต้น โรงเรียนหนองหมื่นถ่านวิทยา จังหวัดร้อยเอ็ด ปีการศึกษา 2551

สำเร็จการศึกษามัธยมตอนปลาย โรงเรียนหนองหมื่นถ่านวิทยา จังหวัดร้อยเอ็ด ปีการศึกษา 2554