แอพพลิเคชันธุรกิจอัจฉริยะสำหรับการวิเคราะห์ข้อมูลธุรกิจ โดยใช้เทคนิคการทำเหมืองข้อมูล

Business Intelligence Application for Analyzing Business Data using Data Mining Techniques

นางสาวภาวดี โพพูงา

โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2556

แอพพลิเคชันธุรกิจอัจฉริยะสำหรับการวิเคราะห์ข้อมูลธุรกิจ โดยใช้เทคนิคการทำเหมืองข้อมูล

Business Intelligence Application for Analyzing Business Data using Data Mining Techniques

นางสาวภาวดี โพพูงา

โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2556

Business Intelligence Application for Analyzing Business Data using Data Mining Techniques

MISS PHAWADEE PHOPHOONGA

A SPECIAL PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIRMENT FOR THE DEGREE OF BACHELOR OF SCIENCE

IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2013

หัวข้อโครงงานพิเศษ แอพพลิเคชันธุรกิจอัจฉริยะสำหรับการวิเคราะห์ข้อมูลธุรกิจ โดยใช้เทคนิคการทำเหมืองข้อมูล
Business Intelligence Application for Analyzing Business Data using Data Mining Techniques

ชื่อนักศึกษา นางสาวภาวดี โพพูงา 53051049

ปริญญา วิทยาศาสตรบัณฑิต

สาขาวิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษา ดร.สายชล ใจเย็น

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ประจำปีการศึกษา 2556

คณะกรรมการสอบ	ลายมือชื่อ
ดร.สุวรรณ จันทิวาสารกิจ ประธานกรรมการ	
ผศ.ธีระ ศิริธีรากุล กรรมการ	
ดร.สายชล ใจเย็น กรรมการและอาจารย์ที่ปรึกษา	

หัวข้อโครงงานพิเศษ	แอพพลิเคชันธุรกิจอัจฉริยะสำหรับการวิเคราะห์ข้อมูลธุรกิจ โดยใช้เทคนิคการทำเหมืองข้อมูล
ชื่อนักศึกษา	นางสาวภาวดี โพพูงา 53051049
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
ปีการศึกษา	2556
อาจารย์ที่ปรึกษา	ดร.สายชล ใจเย็น

บทคัดย่อ

ในปัจจุบันข้อมูลมีผลต่อการประกอบธุรกิจเป็นอย่างมาก เนื่องจากสามารถนำไปสร้างรายงานเพื่อประกอบการตัดสินใจ และวางแผนกลยุทธ์ภายในองค์กรได้ แต่ในขณะนี้แอพพลิเคชันที่ใช้ในการวิเคราะห์ข้อมูลทางธุรกิจ ยังไม่สามารถสร้างรายงานและทำนายผลข้อมูลให้ผู้ใช้ได้ทันที ปัญหาพิเศษนี้มีจุดประสงค์เพื่อแก้ปัญหาดังกล่าว โดยพัฒนาเว็บแอพพลิเคชันธุรกิจอัจฉริยะสำหรับการวิเคราะห์ข้อมูลธุรกิจ โดยใช้เทคนิคการทำเหมืองข้อมูล ซึ่งในงานวิจัยนี้ได้ใช้เทคนิคพื้นฐานการทำเหมืองข้อมูล คือขั้นตอนวิธีที่เรียกว่า ID3 และ Naïve Bayes ที่ใช้ในการจำแนกประเภทข้อมูล ผู้ใช้สามารถเปรียบเทียบค่าความถูกต้องระหว่างขั้นตอนวิธี ID3 และ Naïve Bayes ได้ ยิ่งไปกว่านั้นแอพพลิเคชันนี้ยังสามารถสร้างรายงาน เพื่อให้ผู้ใช้ดาวน์โหลดออกไปใช้งานได้ทันที

คำสำคัญ : ธุรกิจอัจฉริยะ, การทำเหมืองข้อมูล

Title	Business Intelligence Application for Analyzing Business Data using Data Mining Techniques
Students	Miss Phawadee Phophoonga 53051049
Degree	Bachelor of Science
Major Program	Computer Science
Academic Year	2013
Advisor	Dr.Saichon Jaiyen

ABSTRACT

Currently, the information is the key to business success. Every company use information to create reports that support their strategic planning and decision making but there is no application can make the report and predict the result at the same time. So, our business intelligent web application is developed for solving these problems. In our application, ID3 algorithm and Naive Bayes algorithm are used for classifying the data. Furthermore, the user can compare the accuracy between ID3 algorithm and Naïve Bayes algorithm and create the report from the selected algorithms.

Keywords: Business Intelligence, Data Mining

กิตติกรรมประกาศ

ในการจัดทำคู่มือการทำปัญหาพิเศษครั้งนี้สามารถสำเร็จลุล่วงไปได้ด้วยดี เนื่องจากได้รับความช่วยเหลือ และสนับสนุนจาก ดร. สายชล ใจเย็น กรรมการและอาจารย์ที่ปรึกษา ซึ่งเป็นผู้เสียสละเวลาในการให้ความรู้และแนะแนวทางในการพัฒนา ชี้ให้เห็นถึงปัญหา และคอยตรวจสอบแนะแนวทางในการแก้ปัญหามาโดยตลอด ดร.สุวรรณ จันทิวาสารกิจ และ ผศ.ธีระ ศิริธีรากุล ประธานกรรมการ และกรรมการ ซึ่งเป็นผู้คอยให้คำแนะนำ ชี้จุดบกพร่องที่ควรแก้ไข ผู้จัดทำจึงขอกราบขอบพระคุณในความกรุณาของทุกท่านเป็นอย่างยิ่ง ไว้ ณ ที่นี้

สุดท้ายนี้ขอขอบคุณคณาจารย์ในภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ ซึ่งได้ให้ความรู้ทางวิชาการ จนทำให้ผู้จัดทำพอมีสามารถที่จะดำเนินการทำปัญหาพิเศษให้สำเร็จลุล่วงได้เช่นนี้ ขอบพระคุณทุกท่านจากใจจริง

สารบัญ

สารบัญ(ต่อ)

สารบัญ(ต่อ)

สารบัญ(ต่อ)

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ(ต่อ)

สารบัญภาพ(ต่อ)

สารบัญภาพ(ต่อ)

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

ในยุคปัจจุบันที่เทคโนโลยีมีการเปลี่ยนแปลงอย่างรวดเร็ว และตลอดเวลา เช่นเดียวกันกับระบบธุรกิจ ที่มีการแข่งขันกันมากขึ้น จึงเป็นสิ่งที่ปฏิเสธไม่ได้เลย ว่าการที่องค์กรจะอยู่รอดได้นั้นจะต้องมีการใช้ข้อมูลสารสนเทศที่ทันสมัยและทันท่วงที เพื่อสนับสนุนการตัดสินใจอย่างรวดเร็วและสามารถนำไปวางแผน หรือ โต้ตอบปัญหา เชิงธุรกิจได้ทันต่อเหตุการณ์ ให้กับผู้บริหารระดับสูงขององค์กร การที่จะได้มาซึ่งข้อมูล สารสนเทศเหล่านั้น จำเป็นต้องมีการแสวงหาหนทางในการเก็บรวบรวมข้อมูลให้ได้จำนวนมาก เพราะว่าข้อมูลเหล่านั้นมิใช่เพียงข้อมูลภายในองค์กรเท่านั้น ซึ่งอาจรวมไปถึงข้อมูลขององค์กรที่เป็นคู่แข่งหรือเป็นข้อมูลของ องค์กรอื่นๆ ที่อยู่ในธุรกิจเดียวกันก็เป็นได้ การเลือกสรรข้อมูลสารสนเทศที่มีคุณค่าจากกองข้อมูลที่มีขนาดใหญ่ เพื่อให้แน่ใจว่าระบบข้อมูลสารสนเทศที่พัฒนาขึ้นมานั้น เป็นข้อมูลสารสนเทศที่สามารถตอบสนองต่อความต้องการของผู้บริหารระดับสูงขององค์กรได้ เพื่อเอาชนะอุปสรรคเหล่านี้จึงจำเป็นต้องมีระบบที่สามารถช่วยเตรียมข้อมูลที่มีคุณภาพ และมีคุณค่าทางกิจกรรมทางธุรกิจให้แก่องค์กรได้

แต่ในปัจจุบัน แอพพลิเคชั่นที่ใช้ในการทำเหมืองข้อมูล(Data Mining)เพื่อใช้วิเคราะห์ข้อมูลทางด้านธุรกิจ ยังไม่สามารถเปรียบเทียบค่าความถูกต้องของโมเดลในการจำแนกประเภทแต่ละแบบได้พร้อมกัน เมื่อผู้ใช้ต้องการจำแนกข้อมูล (Classification) ทำให้ผู้ใช้เกิดความยุ่งยากที่จะเลือกวิธีการจำแนกข้อมูลที่ให้ค่าความถูกต้องที่เหมาะสมที่สุดในการจำแนกข้อมูล อีกทั้งโปรแกรมแอพพลิเคชันนี้ยังไม่สามารถสร้างรายงานสำหรับผู้บริหารออกมาได้ทันที นอกจากนี้ผู้ใช้ต้องทำการติดตั้งแอพพลิเคชั่นดังกล่าวก่อนจึงจะสามารถทำการวิเคราะห์ข้อมูลได้ ทำให้เกิดความลำบากในการใช้งาน หากมีเทคโนโลยีที่ช่วยในการลดความซับซ้อนในการใช้งาน โดยผู้ใช้สามารถใช้งานได้ทันทีผ่านอินเตอร์เน็ต ซึ่งจะสามารถสร้างรายงานเปรียบเทียบค่าความถูกต้องของโมเดลในแต่ละขั้นตอนวิธีที่ใช้ในการจำแนกข้อมูล โดยสามารถเลือกขั้นตอนวิธีที่เหมาะสมที่สุดในการจำแนกข้อมูลได้ และสามารถสร้างรายงานสำหรับผู้บริหารได้ คงจะมีประโยชน์ต่อการดำเนินธุรกิจเป็นอย่างมาก

1.2 วัตถุประสงค์ของปัญหาพิเศษ

1. เพื่อพัฒนาเว็บแอพพลิเคชัน สำหรับวิเคราะห์ข้อมูลทางธุรกิจโดยใช้เทคนิคการทำเหมืองข้อมูล
2. เพื่อช่วยให้ผู้ใช้สามารถสร้างรายงานการวิเคราะห์ข้อมูลธุรกิจสำหรับผู้บริหาร เพื่อช่วยในการตัดสินใจได้

1.3 ข้อจำกัดและขอบเขตของปัญหา

1. จัดทำเว็บแอพพลิเคชันเพื่อให้ผู้ใช้ได้ทำการวิเคราะห์ข้อมูลเชิงธุรกิจ
2. สามารถสร้างรายงานเปรียบเทียบค่าความถูกต้องของโมเดลในวิธีการจำแนกข้อมูลแต่ละวิธี เพื่อให้ผู้ใช้สามารถเลือกใช้วิธีการที่มีค่าความถูกต้องที่เหมาะสมที่สุดได้ในการรันเพียงครั้งเดียว
3. สามารถสร้างรายงานเพื่อนำเสนอผู้บริหาร เพื่อช่วยในการวิเคราะห์และวางแผนกลยุทธ์ของผู้บริหาร
4. นำไปใช้ได้กับผู้ที่ต้องการวิเคราะห์ข้อมูลทางธุรกิจ
5. ข้อมูลที่จะนำเข้าระบบต้องอยู่ในรูปแบบเอกสาร Excel
6. ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลทางธุรกิจ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ช่วยให้ผู้ใช้สามารถวิเคราะห์ข้อมูลโดยที่ไม่ตองทำการติดตั้งโปรแกรม ทำให้สะดวกต่อการใช้งาน สามารถใช้งานแอพพลิเคชันได้ทุกที่ที่มีอินเตอร์เน็ต
2. ช่วยให้ผู้ใช้สะดวกในการวิเคราะห์ข้อมูล เนื่องจากสามารถเลือกขั้นตอนวิธีในการจำแนกข้อมูลได้หลายวิธี และยังสามารถเลือกขั้นตอนวิธีที่ดีที่สุดจากรายงานการเปรียบเทียบค่าความถูกต้องของการจำแนกข้อมูลของแต่ละขั้นตอนวิธีที่ได้เลือกไว้ในตอนแรก
3. ช่วยให้สามารถสร้างรายงานนำเสนอผู้บริหารได้ง่ายและสะดวกมากยิ่งขึ้น

1.5 ขั้นตอนการดำเนินงาน

1. ศึกษาเอกสาร ทฤษฎีที่เกี่ยวข้อง

เป็นขั้นตอนในการศึกษาทฤษฎีที่ใช้ในการออกแบบระบบงาน รวมไปถึงการศึกษาซอร์ฟแวร์ต่างๆที่เกี่ยวข้อง

2. ออกแบบขั้นตอนการทำงานของระบบงาน

เป็นขั้นตอนที่นำเอาทฤษฎี และวิธีการด้านการออกแบบขั้นตอนการทำงานข้างต้นมาวิเคราะห์ เพื่อออกแบบระบบงาน โดยแบ่งออกเป็นส่วนๆ เช่น ส่วนรับข้อมูล ส่วนจัดการข้อมูล ส่วนแสดงผลลัพธ์ ส่วนประมวลผล เป็นต้น

3. พัฒนาระบบ

เป็นขั้นตอนการเขียนโปรแกรมให้ครอบคลุม ตามขั้นตอนการทำงานที่ได้ออกแบบไว้

4. ทดสอบและติดตั้งระบบ

เป็นการทดสอบโปรแกรมที่ได้พัฒนาขึ้น อีกทั้งยังสามารถทราบได้ถึงข้อจำกัดหรือปัญหาที่เกิดขึ้นในระบบได้

5. ทดลองใช้งานกับกลุ่มตัวอย่างขนาดเล็ก

เป็นขั้นตอนที่นำเอาระบบงานที่ได้พัฒนาขึ้นมา ให้คนประมาณ 5-6 คนได้ทดลองใช้งานจริง เพื่อที่จะนำจุดบกพร่องมาแก้ไข

6. แก้ไขข้อบกพร่องและปรับปรุงระบบงาน

เป็นขั้นตอนการแก้ไขข้อบกพร่องต่างๆ ที่ได้รับมาจากกลุ่มตัวอย่าง เพื่อให้ระบบงานสามารถตอบโจทย์กับความต้องการของผู้ใช้ได้

7. สรุปและอภิปรายผล

เป็นขั้นตอนการสร้างเอกสารเพื่อสรุปผลของระบบงานทั้งหมด รวมทั้งเอกสารการใช้งาน และเอกสารอ้างอิง

1.6 อุปกรณ์ที่ใช้ในการทำปัญหาพิเศษ

1. เครื่องคอมพิวเตอร์ และโน้ตบุ๊ก
2. ฮาร์ดดิสก์ และอุปกรณ์ต่อพ่วง
3. ซอร์ฟแวร์ที่เกี่ยวข้อง ได้แก่

- ระบบปฏิบัติการ Windows
- Eclipse for JAVA EE Developer คือ เครื่องมือ IDE สำหรับการพัฒนาโปรแกรมด้วยจาวาแบบ Enterprise
- Apache Server คือ เครื่อง Server จำลองที่ใช้ติดตั้งในเครื่องเพื่อการรันโปรแกรม

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

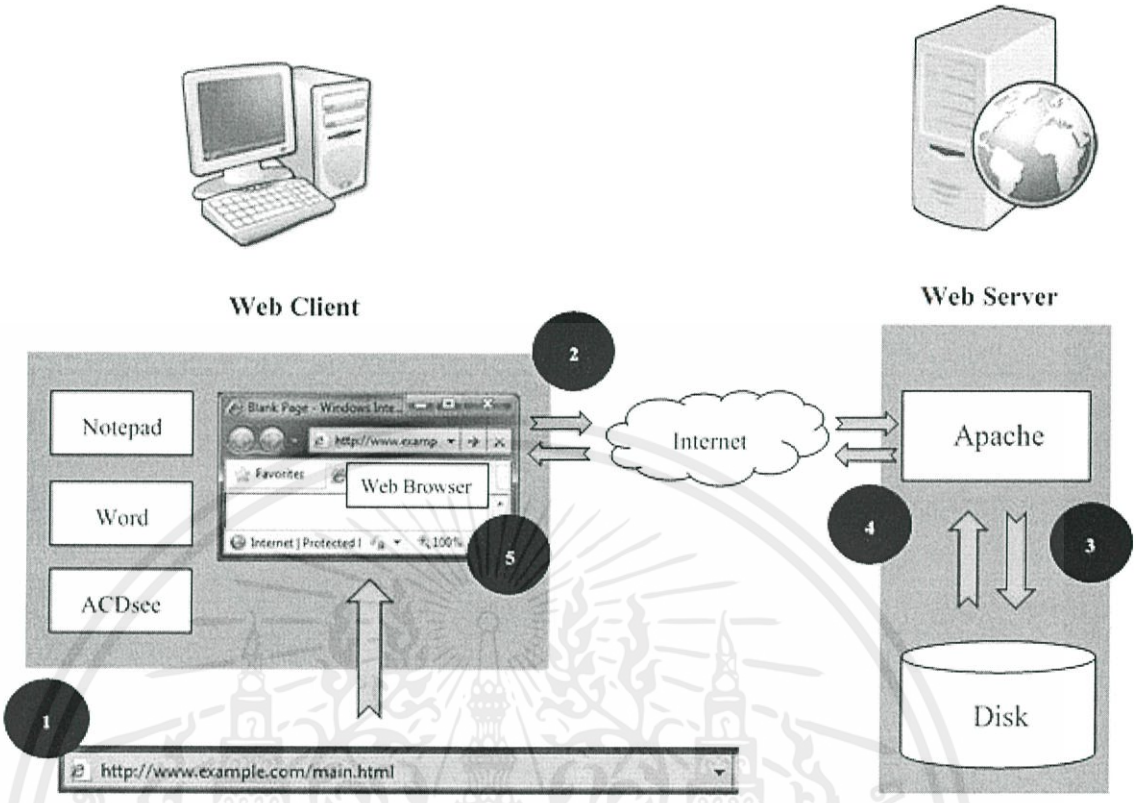
2.1 World Wide Web (WWW)

เว็บ หรือ (World Wide Web-WWW) คือชื่อบริการชนิดหนึ่งในอินเตอร์เน็ตลักษณะของบริการนี้มุมมองของผู้ใช้จะเป็นหน้าเว็บที่เชื่อมโยงกันเหมือนใยแมงมุม โดยที่หน้าเว็บเหล่านั้นอาจเก็บอยู่ในเครื่องคอมพิวเตอร์เครื่องเดียวกันหรือคนละเครื่องที่ห่างออกไปอีกมุมหนึ่งของโลกก็ได้

การใช้งานเว็บนั้น ผู้ใช้โปรแกรมเว็บเบราว์เซอร์ในเครื่องของตนเอง เรียกไปยังหน้าเว็บที่เกี่ยวข้องไว้ในเครื่องคอมพิวเตอร์ต่างๆทั่วโลก และสามารถคลิกไฮเปอร์ลิงค์ (Hyperlink) ที่แสดงอยู่ภายในหน้าเว็บหนึ่งเชื่อมไปยังหน้าเว็บหนึ่งได้อย่างง่ายดาย

การทำงานของ World Wide Web จะอาศัยหลักการ Client/Server (ผู้ขอใช้บริการ/ผู้ให้บริการ) เช่นเดียวกับบริการชนิดอื่นๆในอินเตอร์เน็ต โดย Client หรือผู้ขอใช้บริการก็คือโปรแกรมเว็บเบราว์เซอร์ในเครื่องผู้ใช้ เช่น โปรแกรม Internet Explorer และ Mozilla Firefox ส่วน Server หรือผู้ให้บริการ ก็คือโปรแกรมที่เรียกว่า Web Server ในเครื่องคอมพิวเตอร์ที่ทำหน้าที่เป็นผู้ให้บริการ WWW โปรแกรมเหล่านี้ก็ เช่น Apache Web Server และ IIS เป็นต้น

เมื่อเราใช้โปรแกรมเว็บเบราว์เซอร์เรียกไปยังหน้าเว็บหนึ่งๆในอินเตอร์เน็ต ก็จะเกิดการ “พูดคุย” กันระหว่างโปรแกรมเว็บเบราว์เซอร์ในเครื่องของเรา กับโปรแกรมเว็บเซิร์ฟเวอร์ในเครื่องที่เก็บหน้าเว็บนั้นไว้ ซึ่งขั้นตอนทั้งหมดที่เกิดขึ้นสามารถอธิบายได้ดังรูป



รูปที่ 2.1 แผนภาพแสดงขั้นตอนการติดต่อระหว่าง Client กับ Server

รายละเอียดและขั้นตอนวิธีการติดต่อระหว่างไคลเอนท์กับเซิร์ฟเวอร์

1. ผู้ใช้พิมพ์ http://www.example.com/main.html ลงในช่อง address หรือช่อง URL ของโปรแกรมเว็บเบราว์เซอร์
2. เว็บเบราว์เซอร์ส่ง message ผ่านเน็ตเวิร์คไปยังเครื่องคอมพิวเตอร์ที่มีชื่อว่า www.example.com เพื่อร้องขอ (request) หน้า /main.html
3. โปรแกรมเว็บเซิร์ฟเวอร์ที่ทำงานอยู่ในเครื่อง www.example.com (ในรูปนี้สมมติว่าเป็นโปรแกรม Apache) เมื่อได้รับ message นั้นก็จะอ่านเนื้อหาของไฟล์ main.html ขึ้นมาจากที่เก็บข้อมูล
4. โปรแกรมเว็บเซิร์ฟเวอร์ส่งเนื้อหาของไฟล์ main.html กลับไปให้โปรแกรมเว็บบราวเซอร์ในเครื่องของผู้ใช้ เพื่อเป็นการตอบสนอง (response) ต่อคำร้องขอของเว็บเบราว์เซอร์
5. เว็บเบราว์เซอร์แสดงหน้า main.html ออกมาบนหน้าจอ ตามคำสั่งภาษา HTML (HTML page) ที่กำหนดไว้ในหน้านั้น

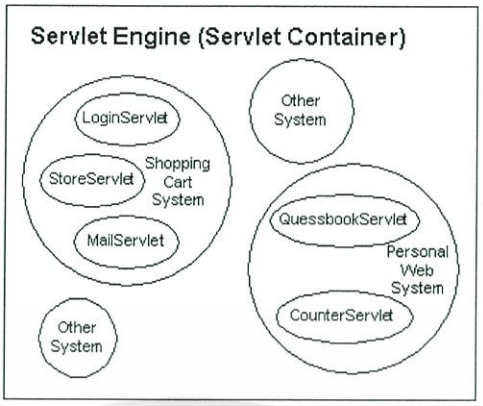
2.2 Java Servlet

Servlet เป็น Server Side Application แบบหนึ่งซึ่งอ้างอิงหลักการมาจาก CGI(Common Gateway Interface) โดย CGI เป็นมาตรฐานสำหรับ Web Server ในการส่งผ่านคำขอเว็บของผู้ใช้ไปยังโปรแกรมประยุกต์และนำข้อมูลส่งต่อไปยังผู้ใช้ เมื่อผู้ใช้ร้องขอหน้าเว็บ ข้อดีของ Servlet ที่มากกว่า CGI คือ Servlet ใช้ภาษาจาวา ซึ่งภาษาจาวาเป็นภาษาที่มีหลักการแบบ Object Oriented ซึ่งมีคุณสมบัติของการ Reusable ที่จะสามารถนำโค้ดที่เขียนไว้มาใช้ใหม่ได้ นอกจากนี้จาวายังเป็นภาษาที่มีลักษณะแบบ Platform Independent ซึ่งจะช่วยให้ผู้พัฒนาโปรแกรมสามารถที่จะทำการพัฒนาระบบโดยใช้ Environment ใดก็ได้

นอกจากนี้ Servlet ยังมีความเร็วที่สูงกว่า CGI เพราะ Servlet ใช้หลักการของ Thread โดยจะทำการสร้าง 1 thread ต่อหนึ่ง Request ที่มาจาก Client ซึ่งในทางกลับกัน CGI จะทำการสร้าง 1 process ต่อหนึ่ง request ซึ่งจะทำให้โปรแกรมใช้ทรัพยากรมากกว่า และ ทำงานได้ช้ากว่าด้วย จุดเด่นที่สำคัญของ Servlet คือ API (Application Programming Interface) โดยระบบที่ทำการพัฒนาโดยใช้หลักการของ Servlet จะสามารถเรียกใช้ API ที่ทางจาวามีมาให้ได้ (javax.servlet.*, javax.servlet.http.*) ซึ่งจะช่วยทำให้การพัฒนาระบบดังกล่าวได้ง่ายและรวดเร็วยิ่งขึ้น

2.3 Servlet Engine (Servlet Container)

ในการรันระบบที่เขียนขึ้นโดยใช้หลักการของ Servlet นั้น เราจะต้องนำระบบดังกล่าวมาบรรจุอยู่ในสิ่งที่เรียกว่า Servlet Engine หรือ Servlet Container โดยทั่วไป Server Side Application ที่ถูกเขียนขึ้นโดยใช้ Servlet API จะถูกเรียกว่า Servlet ในหนึ่งระบบอาจประกอบด้วย Servlet หลายอัน ยกตัวอย่างเช่น ระบบที่เกี่ยวกับ Shopping Cart อาจจะประกอบด้วย Servlet ที่ทำหน้าที่ในการตรวจสอบการเข้าสู่ระบบ, Servlet ที่ทำหน้าที่ในการเก็บข้อมูลสินค้า, Servlet ที่ทำหน้าที่ในการส่งเมล์กลับไปยังลูกค้าเพื่อแจ้งสถานะของการส่งสินค้า เป็นต้น ดังนั้นหากมองโดยรวมแล้ว Servlet Engine คือสถานที่รวบรวมของระบบตั้งแต่ระบบเดียวไปจนถึงหลายระบบ โดยแต่ละระบบจะประกอบด้วย Servlet หนึ่งอันหรือมากกว่า ดังรูปที่ 2.2

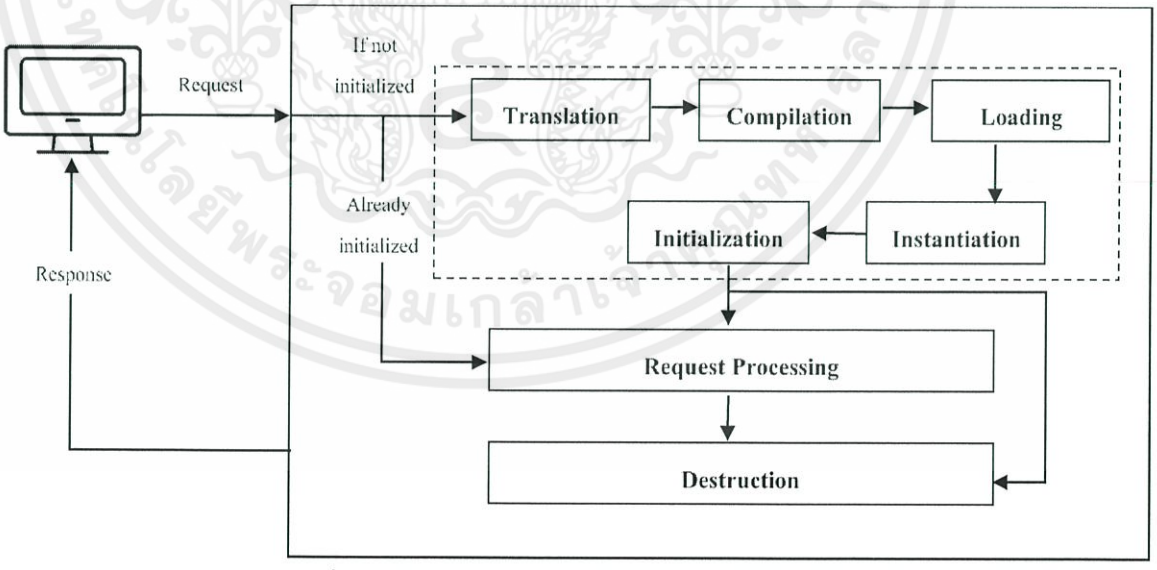


รูปที่ 2.2 แผนภาพแสดงตัวอย่าง Servlet และ Servlet Engine

2.4 Java Server Page

Java Server Page หรือ JSP เป็นเทคโนโลยีจาวา ซึ่งมีการทำงานอยู่บนฝั่ง Server หรือเรียกว่าการทำงานแบบ Server Side โดยขั้นตอนการทำงานจะเริ่มตั้งแต่การร้องขอ หรือ การ Request จาก Browser หรือ Client มาที่ JSP บนฝั่ง Server จากนั้น Server ก็ทำการประมวลผล JSP เป็น Servlet จากนั้น ทาง Server จะส่ง Response กลับไปให้ Client ในรูปของ HTML

JSP จะมีลักษณะเป็น Web-Scripting เทคโนโลยีคล้ายกับ Netscape server-side JavaScript (SSJS) หรือ Microsoft Active Server Pages (ASP) แต่ JSP จะมีพื้นฐานจาก JAVA ซึ่งทำให้มีคุณสมบัติในการทำงานได้ทุก Platform เช่นเดียวกับ JAVA



รูปที่ 2.3 แผนภาพแสดงการทำงานของ JSP

2.5 JSP Containers

JSP Pages (ไฟล์ที่เขียนขึ้นโดยใช้ JSP script และลงท้ายด้วย .jsp) จะถูกรันโดย JSP Container ซึ่งมักจะเป็นส่วนประกอบที่อยู่ใน Webserver หรือ เป็นตัว Add-on ใน Application Server โดยทั่วไป JSP Container จะเป็นตัวรับ Request จาก Client ส่งผ่านไปยัง JSP Page และส่งค่าที่ได้จากการประมวลผลโดย JSP Page กลับไปยัง Client โดย JSP Container ที่ใช้กันอยู่มีมาหลายค่าย ยกตัวอย่างเช่น GNU JSP, Expresso, Tomcat Jakarta, Resin, Weblogic เป็นต้น

2.6 Business Intelligence

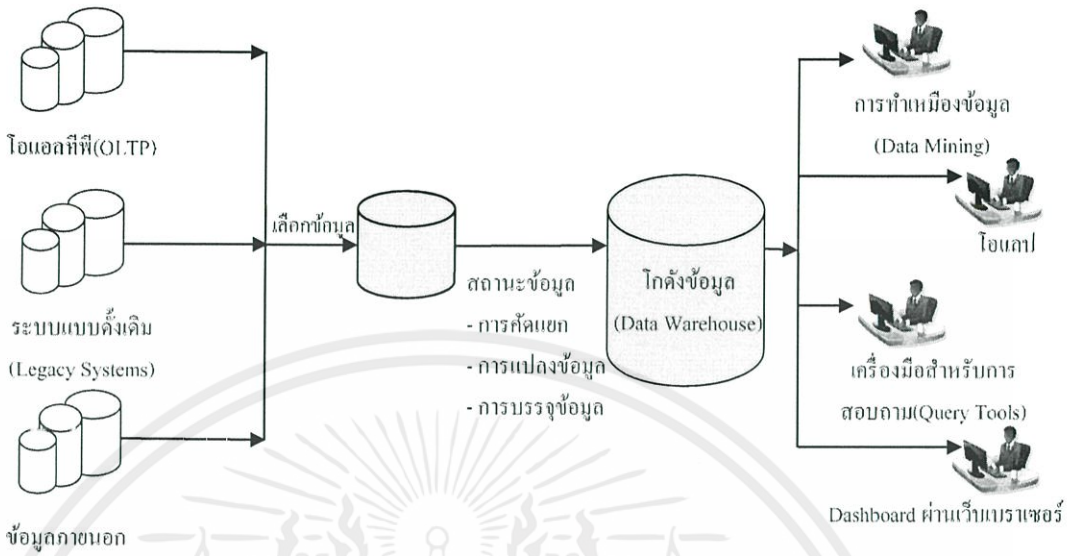
Business Intelligence (BI) เป็นเครื่องมือทางด้านเทคโนโลยีสารสนเทศที่ผู้ใช้สามารถนำไปประมวลผล วิเคราะห์ ข้อมูลจำนวนมากที่มาจากแหล่งข้อมูลหลายแหล่ง ที่มีทั้งรูปแบบโครงสร้างข้อมูล ที่มีความแตกต่างกัน เพื่อให้สามารถสร้างสารสนเทศในรูปแบบที่ผู้ใช้ต้องการได้อย่างมีประสิทธิภาพ

BI เป็นเครื่องมือสนับสนุนการบริหาร การตัดสินใจได้ทุกงานไม่ว่าจะเป็นฝ่าย การตลาด การเงิน การผลิต ฯลฯ และตอบสนองบุคลากรทุกระดับในองค์กร เครื่องมือของ BI มีตั้งแต่ระดับปฏิบัติการไปจนถึงระดับวางแผนกลยุทธ์

BI ยังเป็นระบบสารสนเทศที่สามารถช่วยรองรับความต้องการ สารสนเทศที่ดีและรวดเร็ว เป็นระบบสารสนเทศแบบ Decision Support System หรือระบบสารสนเทศที่สนับสนุนการตัดสินใจ รูปแบบสารสนเทศที่ได้จะมีความง่ายในการใช้งาน และสามารถปรับเปลี่ยนมุมมองของข้อมูลได้อย่างหลากหลายตามเงื่อนไขและความต้องการของผู้ใช้ ในการพัฒนาระบบ BI นั้นผู้ใช้ต้องมีการจัดเตรียมข้อมูลและสารสนเทศที่มีคุณภาพสำหรับจะเป็นแหล่งข้อมูลตั้งต้นในการทำระบบ และการจัดเตรียมข้อมูลและสารสนเทศที่มีคุณภาพ จำเป็นต้องใช้วิธีทำคลังข้อมูล การทำคลังข้อมูลนั้นจะต้องมีกระบวนการและขั้นตอนของการเตรียมข้อมูลดิบที่จะนำเข้ามาในคลังข้อมูลให้อยู่ในรูปของข้อมูลที่มีคุณภาพ

2.6.1 องค์ประกอบของ BI

BI เป็นชุดของเครื่องมือทางด้านเทคโนโลยีสารสนเทศและชุดคำสั่งงานเพื่อใช้ในการรวบรวมข้อมูลจากแหล่งข้อมูล ที่มาจากระบบสารสนเทศต่างๆ นำมาวิเคราะห์ด้วยชุดคำสั่งงานให้เป็นสารสนเทศที่ผู้ใช้ประสงค์ตามที่กล่าวมาแล้วนั้น เพื่อให้การ ทำงานบรรลุตามเป้าหมายการทำงานของ BI จึงประกอบไปด้วยองค์ประกอบ 3 ส่วนหลักดังรูปที่ 2.4



รูปที่ 2.4 แผนภาพแสดงองค์ประกอบของ Business Intelligence

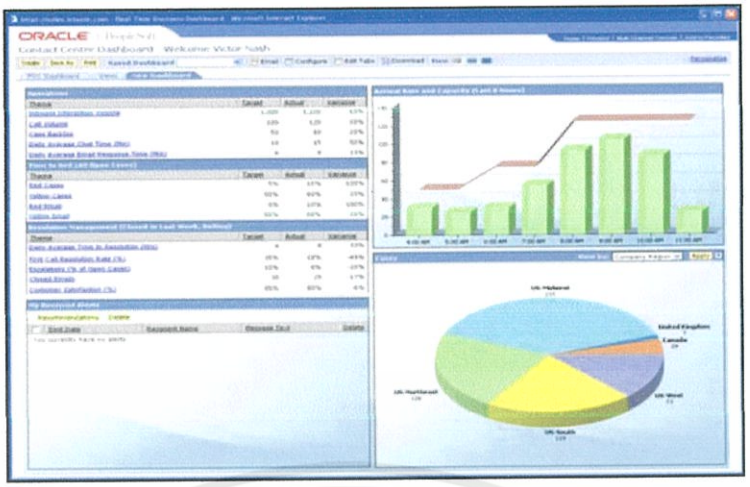
1. ชุดเครื่องมือในการคัดแยก (Extract) เปลี่ยนแปลง (Transform) และบรรจุ (Load) ในที่จัดเก็บ เครื่องมือชุดนี้เป็นที่รู้จักกันในชื่อที่เรียกว่า อีทีแอล (ETL) เนื่องจากข้อมูลในแหล่งกำเนิดข้อมูลมีทั้งจำนวนและปริมาณที่สูงมาก ในการวิเคราะห์ข้อมูลผู้ใช้ข้อมูลจะมีความต้องการข้อมูลเฉพาะอย่างไม่ใช่ข้อมูลทั้งหมดและที่สำคัญคือข้อมูลที่ต้องการนั้นไม่ได้อยู่ในแหล่งข้อมูลเดียวกันทั้งหมด เครื่องมือชุดนี้จะช่วยทำหน้าที่คัดแยกข้อมูลเฉพาะที่ผู้ใช้ต้องจากทุกแหล่งข้อมูลมารวมกัน เมื่อข้อมูลเข้ามาจากแหล่งข้อมูลที่ต่างกันทำให้เกิดความแตกต่างในเรื่องต่างๆ เช่น ขนาดของข้อมูล ลักษณะ รูปแบบ ดังนั้นเครื่องมืออีทีแอลจะทำการทำความสะอาดข้อมูล (Data Cleansing) เพื่อให้ข้อมูลมีความสม่ำเสมอ สอดคล้องกันทั้งหมด ก่อนจะนำบรรจุลงที่เก็บที่เรียกว่าคลังข้อมูล (Data Warehouse)
2. คลังข้อมูล (Data Warehouse) เป็นที่จัดเก็บข้อมูลนำมาจากแหล่งข้อมูลภายในองค์กร ซึ่งก็คือระบบสารสนเทศ ในระดับปฏิบัติการ แหล่งข้อมูลภายนอกที่ผู้บริหารเห็นว่ามีความจำเป็นต้องใช้ในการทำงานการตัดสินใจของผู้บริหาร และข้อมูล ส่วนบุคคล(Personnel Data) เช่น ข้อมูลที่ผู้บริหารบันทึกไว้สำหรับในการทำงานของตนเองข้อมูลเหล่านั้นจะถูกนำมาจัดเตรียม ให้อยู่ในรูปแบบที่พร้อมจะทำงานเชิงวิเคราะห์(Analytical Data) ตามที่ผู้บริหารต้องการได้คลังข้อมูลจะเป็นฐานข้อมูลสำหรับ การวิเคราะห์ด้วยชุดคำสั่งงานต่าง ๆ เช่น การประมวลผลเชิง

วิเคราะห์แบบออนไลน์หรือโอแลบ (On-Line Analytical Processing, OLAP) การทำเหมืองข้อมูล(Data Mining) และระบบสารสนเทศอื่น ๆเป็นต้น

3. ชุดคำสั่งงานเพื่อการวิเคราะห์จะประกอบไปด้วยชุดคำสั่งงานหลายชุดคำสั่งที่จะทำการวิเคราะห์ในประเด็นที่แตกต่าง หลากหลายกันไปผู้ใช้จะเลือกชุดคำสั่งงานตามที่ต้องการมาใช้ อันได้แก่

   3.1. ชุดคำสั่งงานในการจัดทำรายงาน รวมถึงการนำเสนอรายงานจากการสอบถามที่ไม่ได้มีการคาดการณ์ไว้ก่อน (Ad Hoc Query) รายงานที่นำเสนอมักจะเป็นผลการดำเนินงานตามตัวบ่งชี้การดำเนินงานต่างๆของหน่วยงานหรือการติดตาม ค่าเป้าหมายของการดำเนินงานที่สำคัญ การนำเสนอรายงานมักจะอยู่ในรูปแบบของกราฟเพื่อทำให้เกิดความเข้าใจได้ง่าย ผ่าน Dashboard ที่ผู้ใช้สามารถเข้าถึงผ่านหน้าเว็บไซต์ที่จัดทำไว้ดังรูปที่ 2.5 และรูปที่ 2.6

รูปที่ 2.5 แผนภาพแสดงตัวอย่างของรายงานที่นำเสนอผลการดำเนินงานบน Dashboard (ข้อมูลจาก https://www.jaspersoft.com/de/dashboards-DE)



รูปที่ 2.6 แผนภาพตัวอย่างของรายงานที่นำเสนอผลการดำเนินงานบน dashboard (ข้อมูลจาก http://www.perceptualedge.com/blog/?p=154)

3.2. การประมวลผลเชิงวิเคราะห์แบบออนไลน์หรือโอแลบ (Online Analytical Processing, OLAP) เป็นชุดคำสั่งงาน ที่ช่วยให้ผู้ใช้งานวิเคราะห์ข้อมูลที่มาจากคลังข้อมูล การวิเคราะห์ข้อมูลที่เกิดขึ้นบ่อยจะเป็นการวิเคราะห์ข้อมูลหลายมิติ (Multidimensionality) เพื่อช่วยให้ผู้วิเคราะห์ได้มองเห็นข้อมูลในเชิงลึกในมิติต่าง ๆเป็นการเสริมความเข้าใจในสถานการณ์ ให้มากขึ้น

3.3. การทำเหมืองข้อมูล(Data Miming) เป็นชุดคำสั่งงานที่ใช้ในการวิเคราะห์ข้อมูลเพื่อค้นหาความสัมพันธ์ในระหว่าง ข้อมูลที่ไม่เคยมีการค้นพบมาก่อนหรือคาดการณ์กันมาก่อน การได้ค้นพบสิ่งใหม่ก่อนผู้อื่นอาจจะสร้างความได้เปรียบในการแข่งขัน ผลการวิเคราะห์ที่นำเสนอจากการทำเหมืองข้อมูล เช่น การวิเคราะห์เพื่อจัดประเภทลูกค้าการค้นหากลุ่มของลูกค้าการค้นหาลักษณะหรือพฤติกรรมของลูกค้าในแต่ละกลุ่ม การพยากรณ์พฤติกรรมของลูกค้าที่อาจจะพาไปสู่การกระทำที่ไม่ดีเช่น การฉ้อโกงองค์กร เป็นต้น

BI เป็นความหวังของผู้ประกอบการที่จะได้เข้าถึงข้อมูลในเชิงลึกเพื่อทำความเข้าใจในผลประกอบการ สภาพการแข่งขัน เพื่อแสวงหาวิธีการกลยุทธ์มาใช้ช่วยในการสร้างความได้เปรียบในการแข่งขัน การสร้างผลิตภัณฑ์หรือบริการใหม่ๆ แต่การจะบรรลุเป้าหมายดังกล่าวได้ไม่ได้ขึ้นอยู่กับการมีบีไอใช้ในองค์กรอย่างเดียวแต่ขึ้นอยู่กับปัจจัยอื่นๆ ดังต่อไปนี้ด้วย

1. ผู้บริหารเป็นผู้ที่ใช้ข้อมูลสารสนเทศเป็นฐานในการบริหารและการตัดสินใจ
2. ผู้บริหารเป็นผู้ที่ตั้งใจและตระหนักว่าตนเองต้องการข้อมูลหรือสารสนเทศอะไร

3. องค์กรหรือหน่วยงานต้องมีความพร้อมในข้อมูลจากการปฏิบัติงานเพื่อนำเข้าสู่คลังข้อมูล

มิฉะนั้นองค์กรอาจจะประสบความล้มเหลวของการนำบีไอมาใช้งาน ซึ่งหมายถึงเงินจำนวนมากที่สูญเสียไป เนื่องจากการนำ BI มาใช้มีค่าใช้จ่ายสูง

2.7 การจำแนกประเภทข้อมูล

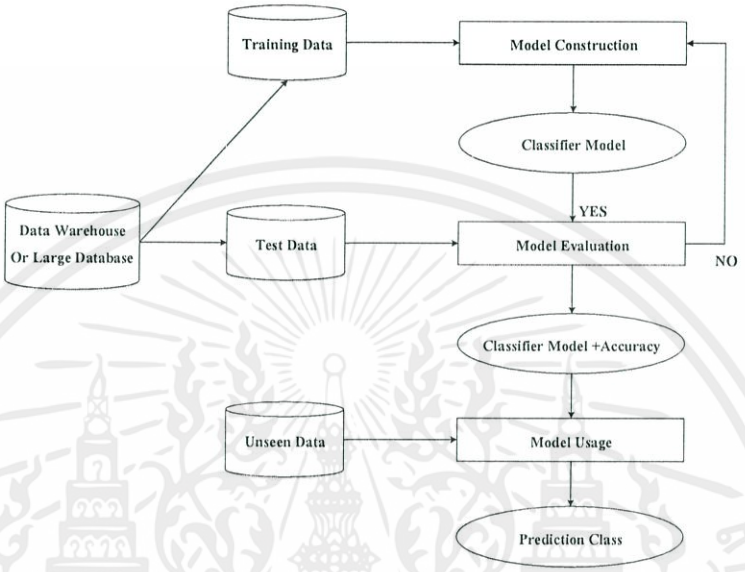
การจำแนกข้อมูลคือการคือกระบวนการสร้างโมเดลจำแนกประเภทข้อมูล (Data Classification Model) เพื่อทำนายกลุ่มของข้อมูลใหม่ (Unseen Data) ตัวอย่างของกลุ่มข้อมูลเช่น กลุ่มของบุคคลที่โกงบัตรเครดิต-ไม่โกงบัตรเครดิต กลุ่มของการผลิตสินค้าผ่านเกณฑ์-ไม่ผ่านเกณฑ์ ในที่นี้คำว่ากลุ่มจะเรียกว่า class ของข้อมูล ซึ่งใน class เดียวกันนั้นจะต้องมีข้อมูลที่มีความหมายเหมือนหรือคล้ายคลึงกันมากว่าข้อมูลที่อยู่ใน class ที่แตกต่างกัน

การสร้างโมเดลจำแนกประเภทข้อมูลจะเกิดจากการหาความสัมพันธ์ของข้อมูลในฐานข้อมูลขนาดใหญ่โดยข้อมูลทั้งหมดจะถูกแบ่งออกเป็น 2 กลุ่ม คือกลุ่มข้อมูลเรียนรู้ (Training Set) เป็นชุดข้อมูลที่มีบทบาทในการสร้างโมเดลจำแนกประเภทข้อมูลขึ้นมา และมีกลุ่มข้อมูลทดสอบ (Test Set) เป็นชุดข้อมูลประเมินความถูกต้องของโมเดลจำแนกประเภทข้อมูล

โมเดลจำแนกประเภทข้อมูลได้ถูกนำมาประยุกต์ใช้ในงานหลายๆด้าน ไม่ว่าจะเป็นการวิเคราะห์หุ้น เพื่อหาว่าหุ้นแต่ละบริษัทมีคุณภาพเป็นอย่างไร เมื่อมีปัจจัยที่เกี่ยวข้อง ไม่ว่าจะเป็นการเติบโตของรายได้ ความสามารถในการควบคุมต้นทุน ความผันผวนของรายได้ กำไร และผู้บริหาร หรือจะเป็นการพยากรณ์อากาศการจัดสรรกฎหมายที่เหมาะสมในการพิจารณาคดี การจัดการความสัมพันธ์ของลูกค้า (CMR) และอื่นๆ

2.7.1 กระบวนการสร้างโมเดลจำแนกประเภทข้อมูล

แบ่งออกเป็น 3 ขั้นตอน ซึ่งภาพรวมของกระบวนการสร้างโมเดลจำแนกประเภทข้อมูล แสดงได้ดังรูป



รูปที่ 2.7 กระบวนการสร้างโมเดลจำแนกประเภทข้อมูล

กระบวนการของแต่ละขั้นตอนมีดังนี้

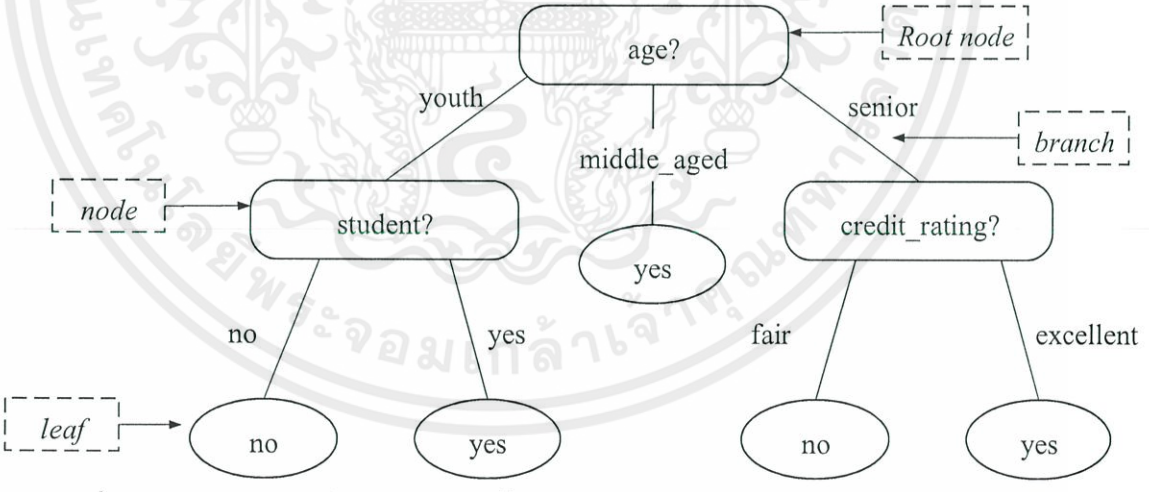
1) Model Construction (Learning) เป็นขั้นตอนการสร้างโมเดลจำแนกประเภทโดยอาศัยการเรียนรู้จากข้อมูลที่ได้กำหนด class ไว้เรียบร้อยแล้วหรือเรียกว่า ข้อมูลเรียนรู้ (Training Data) ซึ่งโมเดลจำแนกประเภทที่ได้แสดงด้วยวิธีการพื้นฐานทางเหมืองข้อมูล (Data Mining) ยกตัวอย่างเช่น ต้นไม้ตัดสินใจ (Decision Tree) โมเดลจำแนกประเภทที่ได้จะมีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดรากอยู่ด้านบนสุด และโหนดใบอยู่ล่างสุดของต้นไม้ แต่ละโหนดบนต้นไม้จะมีลักษณะ (attribute) เป็นตัวเลือกทดสอบ ซึ่งจะมีกิ่งซึ่งเป็นค่าที่เป็นไปได้ของคุณลักษณะ (attribute value) ที่ถูกเลือกทดสอบไว้และมีโหนดใบแสดง class ที่กำหนดไว้

2) Model Evaluation (Accuracy) เป็นขั้นตอนตรวจสอบความถูกต้อง โดยอาศัยข้อมูลที่ใช้สำหรับทดสอบเรียกว่าข้อมูลทดสอบ (Test Data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบจะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลจำแนกประเภท เพื่อทดสอบว่าโมเดลจำแนกประเภทนี้สามรถจัดกลุ่มประเภทข้อมูลได้อย่างถูกต้องมากน้อยเพียงใด และมีการปรับปรุงโมเดลจำแนกประเภทจนกว่าจะได้ค่าความถูกต้องในระดับที่ยอมรับได้

3) Model Usage (Classification) เป็นขั้นตอนการนำโมเดลจำแนกประเภทที่สร้างขึ้นมาใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen Data) เพื่อทำนายและกำหนดกลุ่มให้กับข้อมูลนั้น

2.7.2 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree) เป็นโครงสร้างข้อมูลชนิดเป็นลำดับชั้น (hierarchy) ใช้สนับสนุนการตัดสินใจ โดยมีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดรากอยู่ด้านบนสุดและโหนกในใบอยู่ล่างสุดของต้นไม้ โดยที่ภายในต้นไม้จะประกอบด้วยโหนด (node) ซึ่งแต่ละโหนดจะมีคุณลักษณะ (attribute) เป็นตัวทดสอบ กิ่งของต้นไม้ (branch) แสดงถึงค่าที่เป็นไปได้ของลักษณะที่ถูกเลือกทดสอบ และใบ (leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (class) หรือนั่นคือผลลัพธ์ที่ได้จากการทำนาย โหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดราก (root node) ดังแสดงโครงสร้างของต้นไม้ตัดสินใจดังรูปที่ 2.8 ซึ่งเป็นต้นไม้ตัดสินใจที่จะเลือกซื้อคอมพิวเตอร์หรือไม่ (Quinlan, 1986) มีคุณลักษณะที่พิจารณาคืออายุ (age) นักศึกษา (student) อัตราเครดิต (credit_rating) โดยที่โหนดสี่เหลี่ยมมุมโค้งจะเป็นการทดสอบคุณลักษณะของข้อมูล ท้ายสุดจะได้ผลลัพธ์ของการทำนายว่าจะซื้อคอมพิวเตอร์ (yes) หรือไม่ซื้อคอมพิวเตอร์ (no) จากการทดสอบตามเส้นทางของต้นไม้ตัดสินใจตั้งแต่โหนดรากไปถึงใบ



รูปที่ 2.8 ต้นไม้ตัดสินใจที่ใช้ในการเลือกซื้อคอมพิวเตอร์ (Han and Kamber, 2006, p.291)

การสร้างต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจจะสร้างลักษณะจากบนลงล่าง (top-down) นั้นก็คือเริ่มจากการหาคุณลักษณะที่เหมาะสมที่สุดเพื่อนำมาเป็นรากของต้นไม้แล้วจึงแตกกิ่งไปจนถึงใบ โดยขั้นตอนการสร้างต้นไม้ตัดสินใจจะมีดังนี้ (Han and Kamber, 2006, p.293)

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition *D*.

Input: - Data partition, *D,* which is a set of training tuples and their associated class labels;

- *attribute_list,* the set of candidate attributes;

- Attribute_selection_mathod, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset.*

Output: A decision tree

Method:

(1) Create a node *N*;
(2) **if** tuples in *D* are all of the same class, *C* **then**
(3) return *N* as a leaf node labeled with the class *C*;
(4) **if** *attribute_list* is empty **then**
(5) return *N* as a leaf node labeled with the majority class in *D*;// majority voting
(6) apply **Attribute_selection_method**(*D*, *attribute_list*) to find the "best" *splitting_criterion*;
(7) label node *N* with *splitting_criterion*;
(8) **if** *splitting_attribute* is discrete –valued **and**
multiway splits allowed **then** // not restricted to binary trees
(9) *attribute_list* ← *attribute_list-splitting_attibute*; // remove spriting_attribute
(10) **for each** outcome j of *spritting_criterion*
// partition the tuples and grow subtrees for each partition
(11) let *Dj* be the set of data tuples in *D* satisfying outcome j; // a partition
(12) **if** *Dj* is empty **then**
(13) attach a leaf labeled with the majority class in *D* to node *N*;
(14) **else** attach the node return by Generate_decision_tree(*Dj*, attribute_list) to node *N*;
end for
(15) return *N*;

รูปที่ 2.9 ขั้นตอนวิธีพื้นฐานในการสร้างต้นไม้ตัดสินใจด้วยข้อมูลเรียนรู้

(Han and Kamber, 2006, p.293)

การสร้างต้นไม้ตัดสินใจในลักษณะจากบนลงล่าง (top-down) นั้นก็คือเริ่มจากการหาคุณลักษณะที่เหมาะสมที่สุดเพื่อนำมาเป็นรากของต้นไม้แล้วจึงต้องแตกกิ่งไปจนถึงใบ โดยขั้นตอนการสร้างต้นไม้ตัดสินใจจะมีดังนี้

1) เริ่มต้นสร้างโหนดขึ้นมาหนึ่งโหนด
2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดที่สร้างขึ้นเป็นโหนดใบละกำหนดค่าด้วยกลุ่มของข้อมูลนั้น
3) ถ้าข้อมูลไม่มีคุณลักษณะที่เหมาะสมในการแบ่งกลุ่ม ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและกำหนดค่าด้วยกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด
4) ถ้าข้อมูลมีหลายกลุ่มปะปนกัน จะทำการเลือกคุณลักษณะที่มีความเหมาะสมมากที่สุดเป็นตัวทดสอบการตัดสินใจ โดยการวัดค่าเกน (gain) ของแต่ละคุณลักษณะ และกำหนดค่าให้โหนดที่สร้างขึ้นด้วยตัวทดสอบการตัดสินใจที่ได้
5) เมื่อได้ตัวทดสอบการตัดสินใจ หลังจากนั้นให้สร้างกิ่งของต้นไม้ด้วยค่าต่างๆที่เป็นไปได้ของตัวทดสอบ และแบ่งข้อมูลตามกิ่งต่างๆที่สร้างขึ้น
6) พิจารณาข้อมูลแต่ละกิ่ง หากพบว่าข้อมูลทั้งหมดอยู่ในกิ่งเดียวกัน ให้ต่อกิ่งด้วยโหนดใบและกำหนดค่าด้วยกลุ่มของข้อมูลนั้น แต่ถ้าพบว่าข้อมูลมีหลายกลุ่มปะปนกัน ให้ทำการวนซ้ำการหาตัวทดสอบการตัดสินใจที่เหมาะสมต่อไป
7) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแต่งกิ่งของต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้เป็นจริง
    a. ข้อมูลทั้งหมดในโหนดอยู่ในกลุ่มเดียวกัน ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบและกำหนดค่าด้วยกลุ่มของข้อมูลนั้น
    b. ไม่มีคุณลักษณะใดที่เหมาะสมในการแบ่งกลุ่ม ให้โหนดที่สร้างขึ้นนั้นเป็นโหนดใบ และกำหนดค่าด้วยกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

2.7.2.1 ตัวอย่างการสร้างต้นไม้ตัดสินใจด้วยวิธี ID3

ตารางที่ 2.1 ข้อมูลเรียนรู้ที่ใช้ประกอบการตัดสินใจซื้อคอมพิวเตอร์ (Han and Kamber, 2006, p.299)

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

เนื่องจากข้อมูลที่มีนั้นประกอบด้วยข้อมูลหลากหลายกลุ่มปะปนกัน ฉะนั้นจะต้องวัดมาตรฐานเกน (Gain) ของแต่ละคุณลักษณะ (attribute) ซึ่งมีค่ามาตรฐานเกน (Gain) ที่คำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ (information gain) ซึ่งค่าสารสนเทศของข้อมูลจะขึ้นอยู่กับความน่าจะเป็นของข้อมูล สามารถเขียนในรูปสมการที่ 2.1 (Han and Kamber, 2006, p.297) ได้ดังนี้

$$Info(D) = -\sum_{i=1}^{m} pi \log_2(pi) \tag{2.1}$$

โดยที่ pi เป็นความน่าจะเป็นที่ข้อมูลในฐานข้อมูล D อยู่ในกลุ่ม Ci ซึ่ง $\frac{|Ci,D|}{|D|}$

m เป็นจำนวนกลุ่มทั้งหมดที่แตกต่างกันของข้อมูลชุดนั้น

Ci เป็นกลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง m

$|Ci, D|$ เป็นจำนวนข้อมูลในฐานข้อมูล D ที่อยู่ใน Ci

$|D|$ เป็นจำนวนข้อมูลในฐานข้อมูล D

ค่า $Info(D)$ ที่ได้สามารถเรียกอีกชื่อหนึ่งว่า Entropy ของฐานข้อมูล D

ค่าความรู้จากทฤษฎีสารสนเทศจะช่วยในการแยกแยะข้อมูลทำให้ลดจำนวนครั้งได้อีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจที่ได้จะไม่มีความซับซ้อนมากจนเกินไป

เมื่อทำการพิจารณาเลือกคุณลักษณะเป็นตัวเลือกทดสอบ จะใช้ความรู้จากทฤษฎีสารสนเทศของคุณลักษณะ สามารถเขียนในรูปสมการที่ 2.2 (Han and Kamber, 2006, p.298) ได้ดังนี้

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{2.2}$$

โดยที่ v เป็นจำนวนค่าที่เป็นไปได้ของคุณลักษณะ

$|D|$ เป็นจำนวนข้อมูลในฐานข้อมูล D

$|D_j|$ เป็นจำนวนข้อมูลในฐานข้อมูล D ที่มีค่าคงที่ j ของคุณลักษณะ A

ค่ามาตรฐานเกนที่จะพิจารณาคุณลักษณะ A มาเป็นโหนดของต้นไม้เท่ากับผลต่างของความรู้จากทฤษฎีสารสนเทศ กับ ความรู้จากทฤษฎีสารสนเทศคุณลักษณะ สามารถเขียนในรูปสมการที่ 2.3 (Han and Kamber, 2006, p.298) ได้ดังนี้

$$Gain(A) = Info(D) - Info_A(D) \tag{2.3}$$

เริ่มต้นสร้างต้นไม้ตัดสินใจเราต้องพิจารณาคุณลักษณะที่มีของข้อมูลเรียนรู้เพื่อเลือกเป็นโหนดราก โดยการคำนวณค่ามาตรฐานเกนของคุณลักษณะที่มีทั้งหมด แล้วจังตัดสินใจเลือกคุณลักษณะที่มีค่ามาตรฐานเกนสูงที่สุด ในที่นี้จะสังเกตได้ว่าจะมีคุณลักษณะที่สามารถนำมาใช้ตัดสินใจคือ age, income, student และ credit_rating ส่วนคุณลักษณะ RID มีลักษณะเป็นค่าไม่ซ้ำ (unique value) จึงไม่เหมาะสมในการนำมาใช้ตัดสินใจ และ Class ก็เป็นกลุ่มของข้อมูล ก็ไม่เหมาะสมเช่นกัน

จากตัวอย่างข้อมูลคุณลักษณะประกอบการตัดสินใจซื้อคอมพิวเตอร์ในตารางที่ 2.1 เซตของข้อมูลเรียนรู้ T ประกอบด้วยข้อมูลจำนวน 14 แถว แบ่งออกเป็น 2 กลุ่มคือ ข้อมูลที่ตัดสินใจซื้อคอมพิวเตอร์ (Class = yes) จำนวน 9 แถว และตัดสินใจไม่ซื้อคอมพิวเตอร์ (Class = no) จำนวน 5 แถว

$$Info(T) = -\left(\frac{9}{14}\right) \times \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \times \log_2\left(\frac{5}{14}\right)$$

$$= 0.940$$

พิจารณาแต่ละคุณลักษณะโดยการหาความรู้จากทฤษฎีสารสนเทศของคุณลักษณะ และค่ามาตรฐานเกนออกมาโดยใช้สมการที่ 2.2 ละ 2.3 ตามลำดับดังนี้

$$\begin{aligned} Info_{age}(T) = & \left(\frac{5}{14}\right) \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\ & + \left(\frac{4}{14}\right) \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) \\ & + \left(\frac{5}{14}\right) \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \\ = & \ 0.693 \end{aligned}$$

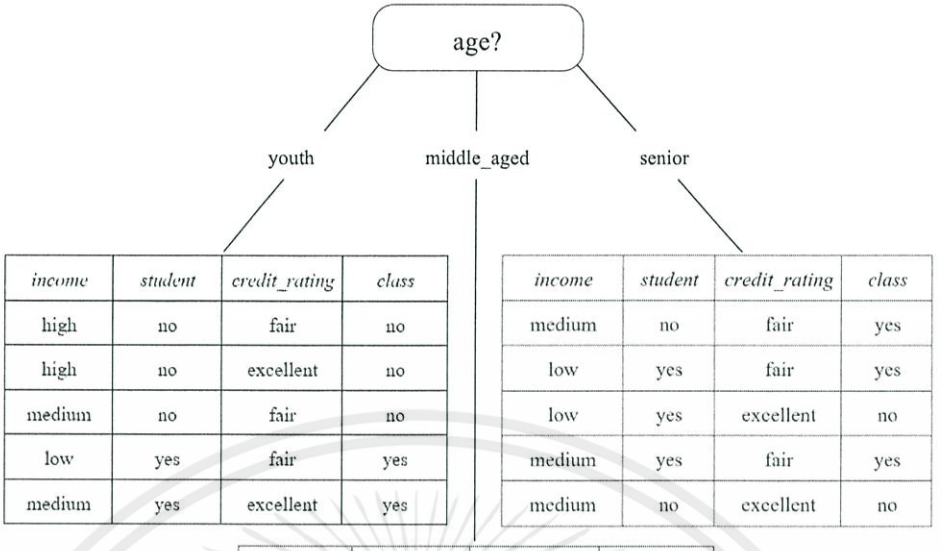
$$\begin{aligned} Gain(age) &= Info(T) - Info_{age}(T) \\ &= 0.940 - 0.693 \\ &= 0.247 \end{aligned}$$

$$\begin{aligned} Gain(income) &= Info(T) - Info_{income}(T) \\ &= 0.940 - 0.911 \\ &= 0.029 \end{aligned}$$

$$\begin{aligned} Gain(student) &= Info(T) - Info_{student}(T) \\ &= 0.940 - 0.788 \\ &= 0.152 \end{aligned}$$

$$\begin{aligned} Gain(credit\_rating) &= Info(T) - Info_{credit\_rating}(T) \\ &= 0.940 - 0.892 \\ &= 0.048 \end{aligned}$$

ซึ่งจะเห็นได้ว่าคุณลักษณะที่ให้มาตรฐานเกนสูงที่สุดคือ age ดังนั้นคุณลักษณะ age จึงถูกเลือกเป็น โหนดรากของต้นไม้ตัดสินใจ แต่เนื่องจากข้อมูลทั้งหมดของแต่ละกิ่งไม่ได้อยู่ในกลุ่มเดียวกันทั้งหมด จึงต้องมีการเลือกโหนดสร้างต้นไม้ตัดสินใจต่อไปบนกิ่งของโหนดราก ยกเว้นกรณี age = middle_aged จะไม่ต้องสร้างต้นไม้ตัดสินใจเพิ่มเนื่องจากสามารถจัดกลุ่มของข้อมูลที่เป็นกลุ่ม yes ได้ทั้งหมดแล้ว ซึ่งจะแสดงได้ดังรูปที่ 2.10



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*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

รูปที่ 2.10 ต้นไม้ตัดสินใจที่ได้จากการเลือกคุณลักษณะ age เป็นโหนดราก

(Han and Kamber,2006, p.300)

การเลือกโหนกระดับที่ 2 จะมีเพียง income, student, และ credit_rating เท่านั้นที่สามรถเป็นตัวทดสอบการตัดสินใจได้ การสร้างต้นไม้ระดับที่ 2 จะแบ่งพิจารณาเป็นทีละส่วนคือ age = youth และ age = senior ในที่นี้เราจะพิจารณาต้นไม้ age = youth ก่อน โดยใช้วิธีหาคุณลักษณะที่เหมาะสมซึ่งจะมีการใช้มาตรฐานเกนดังนี้

$$Info(age = youth) = -\left(\frac{2}{5}\right) \times \log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right) \times \log_2 \frac{3}{5}$$
$$= 0.971$$

พิจารณาแต่ละคุณลักษณะโดยการหาความรู้จากทฤษฎีสารสนเทศของคุณลักษณะ และค่าเกนออกมาโดยใช้สมการที่ 2.2 และ 2.3 ตามลำดับ ดังนี้

$$Info_{income}(age = youth) = \left(\frac{2}{5}\right) \times \left(-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) + \left(\frac{2}{5}\right) \times \left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \left(\frac{1}{5}\right) \times \left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right)$$

$$= 0.4$$

$$Gain(income) = Info(age = youth) - Info_{income}(age = youth)$$

$$= 0.971 - 0.4$$

$$= 0.571$$

$$Gain(student) = Info(age = youth) - Info_{student}(age = youth)$$

$$= 0.971 - 0$$

$$= 0.971$$

$$Gain(credit_{rating}) = Info(age = youth) - Info_{credit\_rating}(age = youth)$$

$$= 0.971 - 0.951$$

$$= 0.020$$

จะเห็นว่าคุณสมบัติที่ให้ค่ามาตรฐานเกนสูงที่สุดคือ *student* ดังนั้นจึงเลือกคุณสมบัตินี้เป็นโหนดระดับที่ 2 ต่อจาก *age = senior* ซึ่งทำให้ข้อมูลทั้งหมดของแต่ละกิ่งอยู่ในกลุ่มเดียวดันทั้งหมด ดังนั้นจึงไม่ต้องสร้างต้นไม้ตัดสินใจต่อไป แต่ยังเหลือโหนดระดับสองทางขวา (*age = senior*) ที่ต้องพิจารณาเลือกคุณสมบัติด้วยวิธีการเดียวกับที่ผ่านมา ซึ่งคุณสมบัติ *credit_rating* มีค่าเกนสูงที่สุด จึงถูกเลือกเป็นโหนดระดับที่2 ต่อจาก *age = senior* ซึ่งทำให้ข้อมูลทั้งหมดของแต่ละกิ่งอยู่ในกลุ่มเดียวกันทั้งหมด และจะได้โครงสร้างเป็นดังรูปที่ 2.8

2.7.2.2 ขั้นตอนวิธีต้นไม้ตัดสินใจ C4.5

ขั้นตอนวิธี C4.5 เป็นขั้นตอนวิธีที่มีชื่อเสียงและเป็นที่รู้จักอย่างแพร่หลาย พัฒนาโดย Ross Quinlan(1993) โยพัฒนาต่อมาจากขั้นตอนวิธี ID3 ที่เขาได้พัฒนาขึ้น (Ross Quinlan, 1986) ซึ่งขั้นตอนวิธีนี้ใช้เพื่อสร้างต้นไม้ตัดสินใจสำหรับจัดแบ่งกลุ่มข้อมูล และมีการใช้หลักการของ

information gain เช่นเดียวกับ ID3 แต่จะมีส่วนเพิ่มเติมจาก ID3 เข้ามา ซึ่งสามารถแก้ไขจุดด้วยของ ID3 ได้เป็นอย่างดี ดังนี้

1) สามารถใช้งานได้ทั้งข้อมูลแบบต่อเนื่อง (Continuous data) และแบบไม่ต่อเนื่อง (Discrete data) โดยในส่วนของข้อมูลแบบต่อเนื่องนั้น C4.5 จะสร้างจุดแบ่ง (Threshold) แยะคุณลักษณะนั้นออกเป็น 2 ส่วน คือส่วนส่วนที่มีค่ามากกว่ากับน้อยกว่าเท่ากับค่าที่ใช้ในการสร้างจุดเริ่มต้น
2) สามารถใช้กับชุดข้อมูลทดสอบ ที่มีค่าข้อมูลขาดหายได้ (missing data) โดยจะแทนค่าด้วย "?" และไม่นำค่านั้นมาคำนวณในกฎของความรู้จากทฤษฎีสารสนเทศ
3) สามารถใช้กับชุดข้อมูลทดสอบที่มีค่าผิดปกติหรือมีความเสียหายได้
4) สามารถทำการตัดกิ่งต้นไม้ตัดสินใจในขณะสร้างได้ โดยไม่ทำให้ความถูกต้องลดลง

การเลือกคุณลักษณะที่ใช้เป็นโหนดรากหรือโหนดบนต้นไม้ตัดสินใจนั้นขั้นตอนวิธี ID3 จะใช้ค่าเกนในการเลือก แต่ขั้นตอนวิธี C4.5 นั้นได้เพิ่มการใช้ค่ามาตรฐานอัตราส่วนเกน (Gain ratio criterion) ในการตัดสินใจเลือกคุณลักษณะ เนื่องจากค่าเกนจะมีการเอนเอียง (Bias) อย่างมากกับข้อมูลที่ประกอบด้วยคุณลักษณะที่มีค่าที่เป็นไปได้จำนวนมากๆ

การแก้ไขความเอนเอียงของค่าเกนสามารถทำได้โดยปรับค่ามาตรฐานเกนให้ถูกต้องโดยใช้ค่าสารสนเทศของการแบ่งแยก (split information) ของคุณลักษณะแต่ละตัว ค่าสารสนเทศของการแบ่งแยกสามารถเขียนในรูปสมการที่ 2.4 (Han Kamber, 2006, p.301) ได้ดังนี้

$$SplitInfo(A) = \sum_{i=1}^{v} \left|\frac{Dj}{D}\right| \times \log_2\left(\frac{Dj}{D}\right) \qquad (2.4)$$

ค่าสารสนเทศของการแบ่งแยกนี้จะแสดงถึงระดับการกระจายของข้อมูล เมื่อนำค่านี้ไปหารค่าเกนจะได้ค่ามาตรฐานอัตราส่วนเกน สามารถเขียนในรูปสมการที่ 2.5 (Han and Kamber, 2006, p.301) ได้ดังนี้

$$GainRatio(D) = \frac{Gain(A)}{SplitInfo(A)} \qquad (2.5)$$

ค่ามาตรฐานอัตราส่วนเกนช่วยแก้ไขความเอนเอียงของค่าเกนได้ โดยทำให้ค่ามาตรฐานอัตราส่วนในการแบ่งด้วยคุณลักษณะที่มีการกระจายตัวสูงถูกปรับลดลง ดังนั้นค่ามาตรฐานอัตราส่วนเกนในคุณลักษณะที่มีการกระจายตัวของข้อมูลสูงดังที่กล่าวมาจึงไม่มีค่าสูงที่สุดเสมอ

2.7.2.3 การตัดกิ่งต้นไม้ตัดสินใจ (Pruning)

ในขณะที่กำลังสร้างต้นไม้ตัดสินใจ ในแต่ละกิ่งอาจเกิดการสร้างอย่างผิดปกติ เนื่องมาจากข้อมูลฝึกที่อาจมีข้อมูลรบกวน (noise) ซึ่งเกิดจากการบันทึกข้อมูลผิดพลาดหรือความผิดพลาดที่เกิดจากระบบเอง หรือในชุดข้อมูลอาจมีข้อมูลที่ผิดปกติ จากข้อมูลส่วนใหญ่ (outlier) ปะปนมาด้วย การตัดกิ่งต้นไม้ตัดสินใจเป็นเทคนิคที่ใช้สำหรับการแก้ปัญหานี้ และจะช่วยลดการเกิดปัญหาการเจาะจงโมเดลกับข้อมูลมากเกินไป (overfitting) ได้โดยปัญหานี้ทำให้ได้โครงสร้างต้นไม้ที่สามารถจำแนกได้ดีกับชุดข้อมูลที่ใช้โครงสร้างต้นไม้ตัดสินใจเท่านั้น แต่เมื่อนำไปใช้กับข้อมูลใหม่ประสิทธิภาพในการจำแนกข้อมูลจะลดลง การตัดกิ่งต้นไม้ตัดสินใจจะใช้ค่าทางสถิติในการตัดกิ่งที่มีความน่าเชื่อถือน้อยที่สุดออกไป เพื่อทำให้ต้นไม้ใหม่ที่ได้ทำงานได้รวดเร็วมากขึ้น และยังเป็นการปรับปรุงขีดความสามารถของต้นไม้ในการทำนายข้อมูลใหม่ๆ ได้แม่นยำมากขึ้นอีกด้วย โดยการตัดกิ่งต้นไม้ตัดสินใจที่เป็นที่นิยมมีอยู่ 2 ประเภท

1) การตัดกิ่งขณะเรียนรู้ (pre-pruning) เป็นการแตกกิ่งต้นไม้ หรือหยุดการแตกกิ่งในขั้นตอนการสร้างต้นไม้ตัดสินใจ โดยการทำให้โหนดที่ถูกตัดนั้นเปลี่ยนเป็นใบ และให้ใบนั้นแสดงกลุ่มที่มีจำนวนข้อมูลสนับสนุนหรือมีความน่าจะเป็นที่ข้อมูลอยู่ในกลุ่มนั้นมากที่สุด ในขณะที่การสร้างต้นไม้ตัดสินใจนั้น จะต้องมีการคำนวณหรือวดค่าทางสถิติที่สำคัญต่างๆ เช่นไคสแควร์ ค่าสารสนเทศข้อมูล เพื่อใช้ประเมินว่าควรที่จะสร้างหรือแตกกิ่งของต้นไม้อย่างไร ถ้าค่าที่วัดได้ไม่ถึงจุดที่กำหนดไว้จะถือว่าโหนดนั้นไม่สมควรที่จะทำการแตกกิ่งต่อไป ซึ่งเป็นการยากที่จะกำหนดว่าค่าที่ใช้เป็นเกณฑ์เหล่านั้นควรจะมีค่าเป็นเท่าไร ถ้ากำหนดค่านั้นสูงไปก็จะทำให้ต้นไม้ที่มีความซับซ้อน แต่ถ้ากำหนดต่ำไปก็จะทำให้ได้ต้นไม้ที่มีขนาดเล็กจนไม่สามารถนำไปใช้งานได้

2) การตัดกิ่งหลังการเรียนรู้ (post-pruning) เป็นการตัดกิ่งของต้นไม้ตัดสินใจที่ถูกสร้างขึ้นสมบูรณ์แล้ว โดยใช้ค่าวัดความซับซ้อนของแต่ละโหนด หลังจากที่ทำการตัดกิ่งของต้นไม้แล้ว โหนกที่อยู่ล่างสุดที่ไม่ได้ถูกตัดจะถูกเปลี่ยนไปเป็นใบและแสดงกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด (Bleslow and Aha, 1997) สำหรับทุกๆโหนดที่ไม่ใช่ใบของต้นไม้ จะมีการคำนวณค่าอัตราความผิดพลาดที่คาดหวังไว้ ซึ่งเป็นค่าที่แสดงถึงความผิดพลาดที่จะเกิดขึ้นถ้าโหนดของต้นไม้ย่อยนั้นถูกตัดออกไปโดยที่ค่าความผิดพลาดของโหนดที่ไม่ถูกตัดจะถูกคำนวณโดยใช้ค่าผลรวมความผิดพลาดของแต่ละกิ่ง และให้ค่าน้ำหนักตามความผิดพลาดของกิ่งนั้นๆ ถ้าการตัดโหนดนั้นนำไปสู่การเกิดความผิดพลาดที่สูงขึ้น โหนดของต้นไม้ย่อยนั้นก็จะต้องยังคงไว้ แต่ถ้าการตัดนั้นทำให้ได้

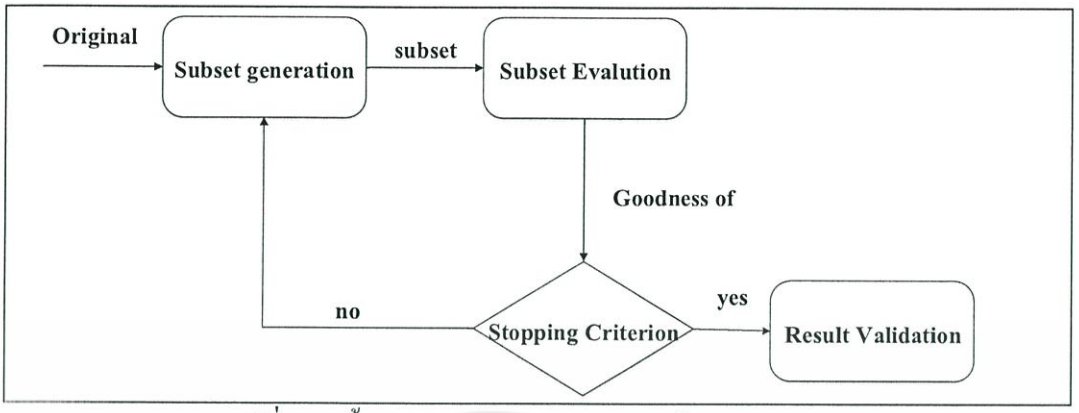
ค่าความผิดพลาดที่เป็นที่ยอมรับได้โหนดนั้นจะต้องถูกตัดออกไป หลังจากทำการตัดกิ่งต้นไม้ตัดสินใจแล้วจะต้องวัดค่าความแม่นยำ (accuracy) ของต้นไม้ที่ทำการตัดกิ่งแล้วด้วย โดยที่ต้นไม้ที่ให้ค่าความผิดพลาดน้อยสุดจะถูกเลือก

นอกจากการตัดกิ่งต้นไม้ตัดสินใจโดยอาศัยการวัดค่าความผิดพลาดที่เกิดขึ้นแล้วยังมีเทคนิคอื่นๆอีก เช่น การเข้ารหัส (encode) ในการพิจารณาตัดกิ่งของต้นไม้โดยใช้หลักการของ Minimum Description Length (MDL) ด้วย (Quinlan and Rivest, 1989) เป็นต้น

วิธีการตัดกิ่งต้นไม้ตัดสินใจที่ศึกษาค้นคว้าในงานวิจัยนี้เป็นชนิด post-pruning ที่ขั้นตอนการตัดกิ่งเริ่มทำงานหลังจากต้นไม้ตัดสินใจได้สร้างขึ้นสมบูรณ์แล้ว โดยที่ Breiman, Friedman, Olshen, and Stone (1984) ได้ศึกษาพบว่าการตัดกิ่งประเภทนี้มีความเสถียรมากกว่า และให้ประสิทธิภาพสูงกว่าการตัดกิ่งขณะที่เรียนรู้ (pre-pruning) เพราะสามารถเลือกตัดโหนดที่ไม่เกิดประโยชน์จากต้นไม้ตัดสินใจที่สร้างขึ้นอย่างดีจากชุดข้อมูลแล้ว และใช้วิธีต่างๆในการวัดค่าความผิดพลาดของโหนดเพื่อพิจารณาว่าจะตัดกิ่งของต้นไม้หรือไม่

2.7.2.4 การคัดเลือกคุณลักษณะ (Attribute Selection)

Liu และ Motoda (อ้างถึงใน Borges and Nievola, 2005) กล่าวว่าการคัดเลือกคุณลักษณะคือกระบวนการเลือกเซตย่อยจำนวน M แอททริบิวต์ จากจำนวนข้อมูล N แอททริบิวต์ เพื่อลดขนาดของคุณลักษณะที่เกี่ยวข้องและทำให้มั่นใจว่าข้อมูลที่ได้รับเพื่อนำไปใช้งานต่อนั้นจะมีคุณภาพที่ดี ซึ่งสอดคล้องกับ สุคนธ์ทิพย์ วงศ์พันธ์ และ อนงค์นาฏ ศรีวิหค (2551) ที่กล่าวว่าการคัดเลือกคุณลักษณะที่เหมาะสมเป็นการคัดเลือกตัวแปรหรือคุณลักษณะสำคัญที่อยู่ในชุดข้อมูลออกมาจากคุณลักษณะทั้งหมดที่มีอยู่ โดยการคัดเลือกคุณลักษณะที่เหมาะสมนั้นจะช่วยลดตัวแปรที่ไม่มีความเกี่ยวข้องกับการทำนายรวมไปถึงลดความซ้ำซ้อนของการเก็บข้อมูล ทั้งนี้สามารถสรุปได้ว่าการคัดเลือกคุณลักษณะที่เหมาะสมนั้นมีขั้นตอนการทำงานที่สำคัญ 2 ขั้นตอนคือ การค้นหาชุดของคุณลักษณะที่เหมาะสม และการประเมินค่าชุดลักษณะที่ได้จากการค้นหา ซึ่งขั้นตอนในการคัดเลือกคุณลักษณะที่เหมาะสมตามแนวคิดของ Liu และ Motoda รวมถึงแนวคิดของสุคนธ์ทิพย์ วงศ์พันธ์และอนงค์นาฏ ศรีวิหค นั้นแสดงได้ดังรูปที่ 2.11



รูปที่ 2.11 ขั้นตอนการทำงานของการคัดเลือกคุณลักษณะ

การค้นหาชุดของคุณลักษณะที่เหมาะสม

เป็นขั้นตอนของการค้นหาชุดของคุณลักษณะเพื่อส่งต่อไปให้อัลกอริทึมในการประเมินค่า โดยจะประเมินว่าควรจะใช้ชุดของคุณลักษณะที่ถูกเลือกหรือไม่ โดยอัลกอริทึมในการค้นหาจะสามารถแบ่งได้เป็น 3 ประเภทได้แก่

1) **Exponential Algorithms** เป็นการค้นหาชุดของแอททริวบิวต์โดยค้นหาให้ครบทุกชุดที่เป็นไปได้ ซึ่งเทคนิคในการค้นหาแบบนี้มีข้อเสียในเรื่องเวลาของการคำนวณ ทำให้เสียเวลาในการค้นหาชุดของคุณลักษณะซึ่งอาจใช้เวลานานมากถ้าข้อมูลมีขนาดใหญ่ ดังนั้น เทคนิคนี้จึงเหมาะกับข้อมูลที่มีขนาดเล็ก ซึ่งตัวอย่างของเทคนิคนี้ได้แก่ Exhaustive Search

2) **Random Algorithms** เป็นการค้นหาชุดของแอททริวบิวต์แบบสุ่มค่าไปเรื่อยๆ จนกว่าจะพบชุดของแอททริวบิวต์ที่เมื่อนำไปผ่านการประเมินค่าแล้วให้ค่าความถูกต้องที่ดีที่สุด ตัวอย่างของเทคนิคแบบนี้ได้แก่ Genetic Search

3) **Sequential Algorithms** การค้นหาชุดของแอททริวบิวต์ด้วยเทคนิคนี้จะทำได้สองทางคือ Forward Selection และ Backward Elimination โดยการทำงานของ Forward Selection จะเริ่มต้นการค้นหาชุดของคุณลักษณะโดยกำหนดให้เริ่มค้นหาจากเซตว่างแล้วค่อยๆเพิ่มคุณลักษณะเข้ามาแต่ถ้าเป็นการทำงานของ Backward Elimination จะเริ่มต้นการค้นหาโดยมีชุดของคุณลักษณะที่เหมาะสมเริ่มต้นก่อน จากนั้นจึงค่อยตัดเอาคุณลักษณะที่ไม่เหมาะสมออกในภายหลัง

การประเมินค่าชุดคุณลักษณะที่ได้จากการค้นหา

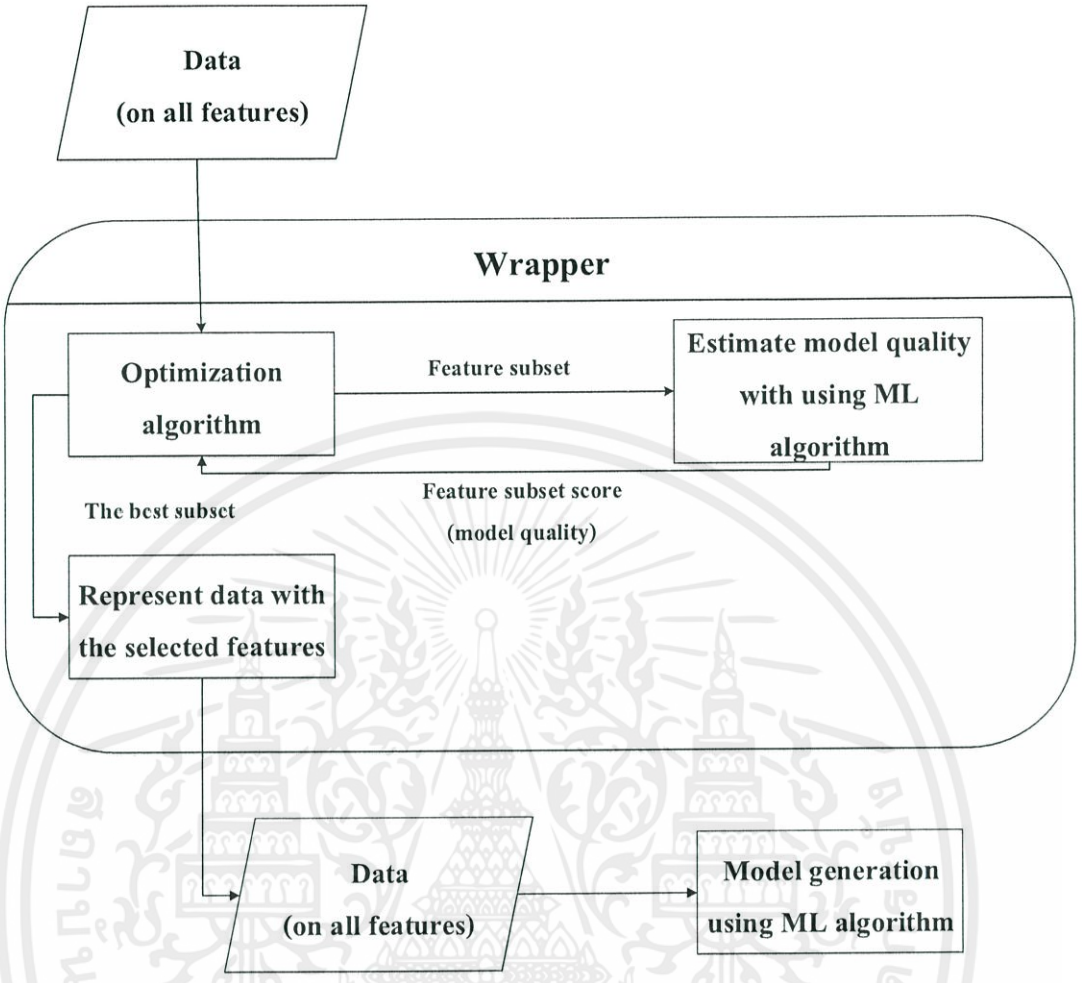
สำหรับขั้นตอนของการประเมินค่าชุดของคุณลักษณะที่ผ่านการค้นหานั้นเป็นขั้นตอนที่ทำต่อจากอัลกอริทึมในการค้นหา ซึ่งขั้นตอนนี้ทำเพื่อประเมินค่าชุดของคุณลักษณะที่ได้มาว่าเหมาะสมหรือไม่ หากได้ชุดของคุณลักษณะที่เหมาะสมแล้วจึงหยุดการคัดเลือกคุณลักษณะ โดยอัลกอริทึมที่ใช้ในการประเมินค่าค่าชุดของคุณลักษณะมี 2 ประเภท คือ

1) **Filter Approach** เป็นการประเมินคุณลักษณะทั่วไปของข้อมูลและคัดเลือกชุดของคุณลักษณะโดยไม่มีการนำอัลกอริทึมทางด้านการทำเหมืองข้อมูลมาใช้ร่วมด้วย ซึ่งอัลกอริทึมในการประเมินค่าแบบ Filter Approach ได้แก่ Correlation-based Feature Selection หรือ CFS และ Consistency-based Subset Evaluation

2) **Wrapper Approach** จะเป็นการประเมินคุณลักษณะของข้อมูลโดยมีการนำอัลกอริทึมในการทำเหมืองข้อมูลมาทำงานร่วมในการค้นหา ซึ่งจะสามารถค้นหาได้ดีกว่าแบบ Filter Approach แต่จะใช้เวลาในการคำนวณมากกว่า

2.7.2.5 การประเมินผลย่อยแบบแรปเปอร์ (Wrapper Subset Evaluation)

แรปเปอร์[4] เป็นการเลือกเซตคุณลักษณะใหม่โดยใช้วิธีการค้นหา เพิ่มหรือลดคุณลักษณะเมื่อเซตใหม่ถูกสร้างขึ้น สร้างตัวจำแนกทำการวัดว่าดีที่สุดหรือยัง ถ้ายัง ให้วนทำซ้ำจนกว่าจะได้เซตคุณลักษณะใหม่ที่ดีที่สุด ตัวอย่างอัลกอริทึมที่ใช้เช่น Hill Climbing หรือ Simulated Annealing และนำเซตใหม่มาทำการสร้างโมเดลโดยใช้ตัวจำแนกเดียวกันกับที่เป็นตัวประเมิน (Mladenic and Grobelnik, 2003; Forman, 2007) ดังรูปที่ 2.12 โดยทั่วไป แนวทางนี้มีความซับซ้อนในการนำไปใช้งาน โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีขนาดใหญ่มากๆ เช่น มีคุณลักษณะทั้งหมด m ตัว จะมีเซตใหม่ถูกสร้างขึ้นที่เป็นไปได้ 2m



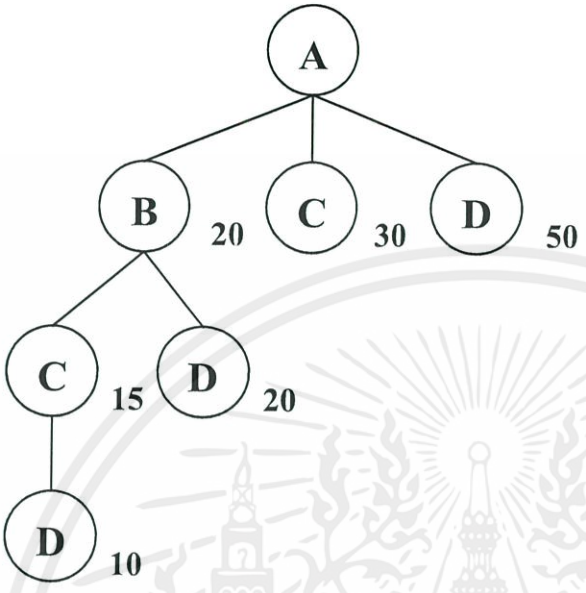
รูปที่ 2.12 แนวทาง Wrapper

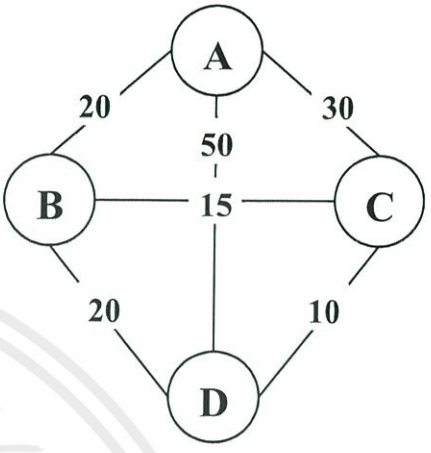
2.7.2.6 การค้นหาแบบขั้นตอนวิธีละโมบ (Greedy Algorithm)

ขั้นตอนวิธีละโมบ[11] เป็นการค้นหาแบบดีที่สุดก่อน(Best first search) ที่ง่านที่สุด หลักการของการค้นหาแบบนี้คือ การเลือกโหนดที่ดีที่สุดตลอดเวลา ซึ่งมีขั้นตอนดังนี้

1. เลือกโหนดเริ่มต้นมาหนึ่งโหนด
2. ให้โหนดที่เลือกมานี้เป็นสถานะปัจจุบัน
3. ให้ทำตามขบวนการต่อไปนี้จนกว่าจะไม่สามรถสร้างโหนดลูกได้อีก
   3.1. สร้างสถานะใหม่ที่เป็นโหนดลูกที่เป็นไปได้ทั้งหมดจากสถานะปัจจุบัน
   3.2. จากสถานะใหม่ที่สร้างขึ้นมาทั้งหมด ให้เลือกสถานะ หรือ โหนดลูกที่ดีที่สุด ออกมาเพียงโหนดเดียว
4. กลับไปที่ขั้นตอนที่ 2

ตัวอย่าง จากเรื่องการเดินทางของเซลแมนที่จะเดินทางไปยังเมือง A B C D ซึ่งมีระยะทางตามตารางที่ 2.2 สามารถแก้ปัญหาได้ดังนี้

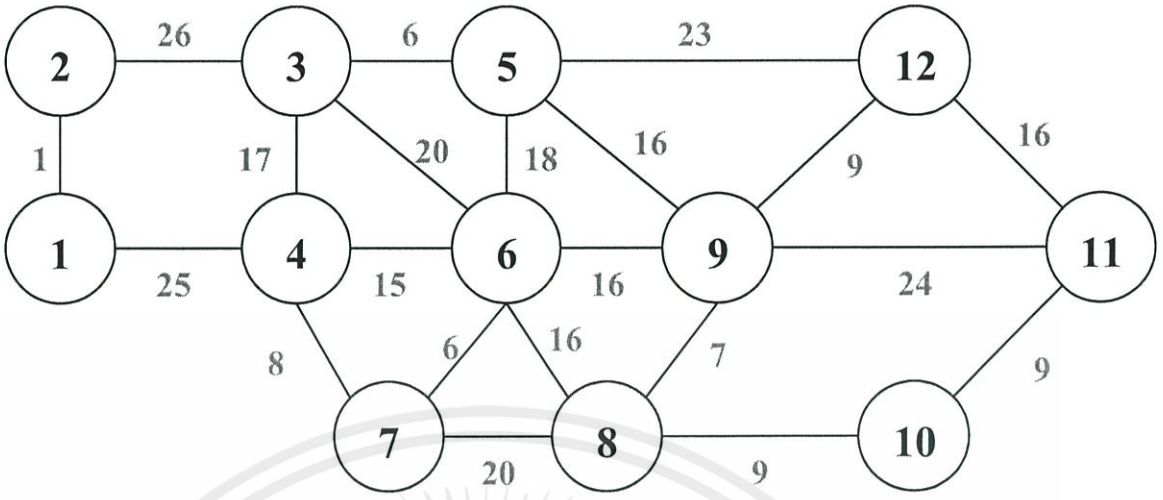




ระยะทางระหว่างเมืองต่างๆ

รูปที่ 2.13 การแก้ปัญหาการเดินทางของเซลแมนด้วยกรีดีอัลกอริทึม

จากรูปที่ 2.13 การแก้ปัญหาเริ่มจาก การเลือก A เป็นเมืองเริ่มแรก จากนั้นทำการสร้างโหนดลูก B C และ D หาระยะทางระหว่าง A ถึงเมืองเหล่านี้ ได้ 20 30 และ 50 ตามลำดับ เลือก B เป็นเมืองที่จะเดินทางต่อมา จากนั้นสร้างโหนดลูกของ B ได้ C และ D และได้ระยะทางเท่ากับ 15 และ 20 ตามลำดับ เลือก C เป็นเมืองที่จะเดินทางต่อไป จากนั้นสร้างโหนดลูกให้ C ได้ D มีค่าเท่ากับ 10 เลือกเดินมาที่ D เป็นเมืองสุดท้ายก่อนกลับไป A รวมระยะทางเท่ากับ 20 + 15 + 10 + 50 = 95



รูปที่ 2.14 ข้อมูลในรูปแบบกราฟ

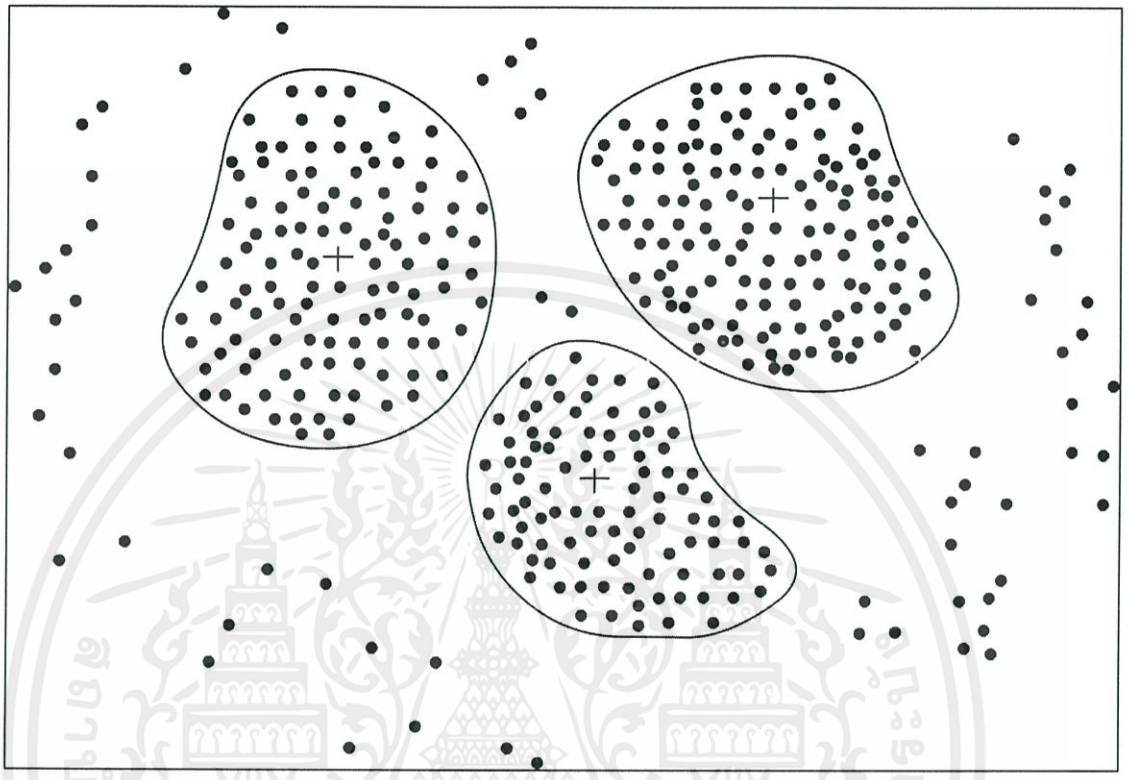
ตารางที่ 2.2 การค้นหาแบบขั้นตอนวิธีละโมบ

ขั้นตอนที่	ตัวเลือก	กิ่งที่เลือก	น้ำหนักรวม
1	(1,2)	(1,2)	1
2	(1,4),(2,3)	(1,4)	26
3	(2,3),(3,4),(4,6),(4,7)	(4,7)	34
4	(2,3),(3,4),(4,6),(6,7),(7,8)	(6,7)	40

2.8 การจัดกลุ่ม (Clustering)

การจัดกลุ่ม (Clustering) เป็นวิธีการที่พิจารณาข้อมูลแต่ละแถวเสมือนเป็นวัตถุ (object) ซึ่งจะมีหลักการเหมือนกับการจำแนกประเภทข้อมูล คือจะทำการแบ่งข้อมูลออกเป็นกลุ่ม (Cluster) โดยจะจัดให้ข้อมูลที่มีความคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน และข้อมูลที่อยู่ต่างกลุ่มกันจะมีความคล้ายคลึงกันน้อยที่สุด ซึ่งความเหมือนหรือต่างกันสามารถเปรียบเทียบได้กับความใกล้ชิดกันของวัตถุใดๆ โดยใช้ระยะทางเป็นตัวชี้วัด คุณภาพของแต่ละกลุ่มสามารถอธิบายได้จากเส้นผ่านศูนย์กลางของกลุ่ม(diameter) ซึ่งแสดงระยะห่างมากที่สุดของวัตถุสองชิ้นที่อยู่ในกลุ่มเดียวกัน แต่ละกลุ่มจะมีตัวแทนที่สามารถแทนวัตถุทุกชิ้นของกลุ่มนั้นได้ เช่น การใช้จุดศูนย์กลางกลุ่ม

(centroid) แทนกลุ่มนั้น สำหรับบางเทคนิคตัวแทนของกลุ่มอาจมีได้หลายตัวแทน ทั้งนี้ขึ้นอยู่กับความเหมาะสมของแต่ละเทคนิคที่เลือกใช้

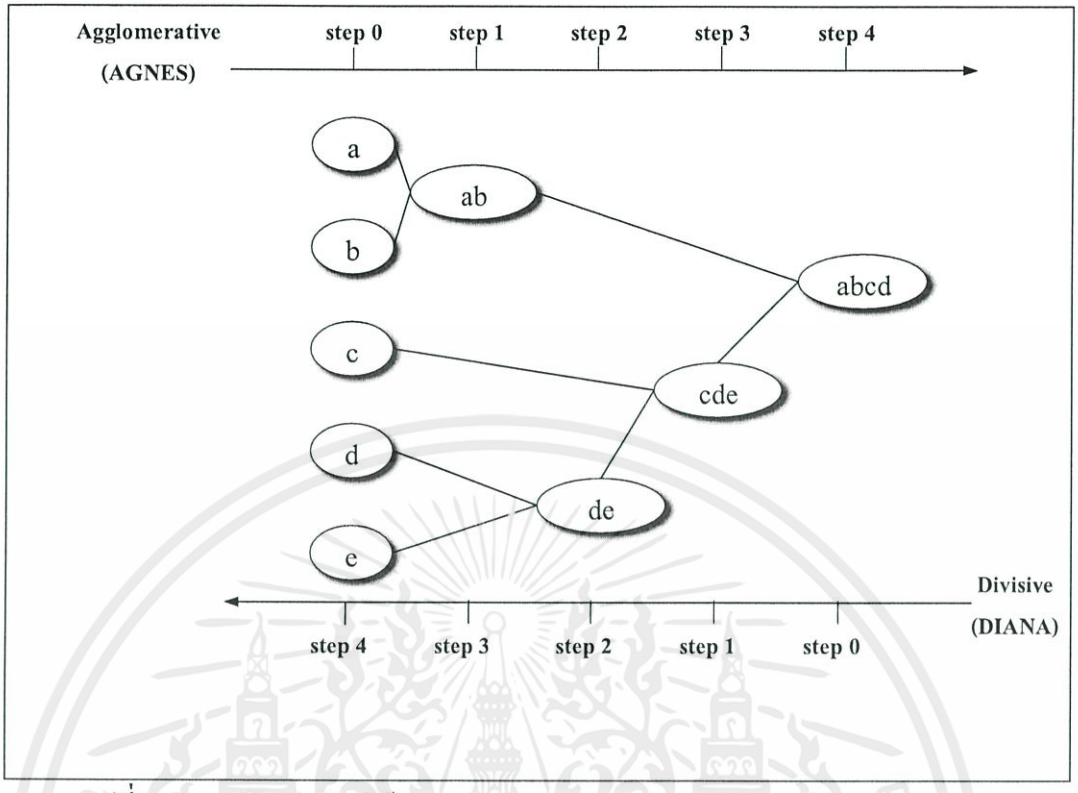


รูปที่ 2.15 ข้อมูลตัวอย่างประกอบด้วย 3 คลัสเตอร์

2.8.1 ประเภทขั้นตอนวิธีการจัดแบ่งกลุ่มข้อมูล

1) Partition Method การจัดกลุ่มข้อมูลประเภทนี้จะทำการสร้าง k พาร์ทิชันบนฐานข้อมูลจำนวน n เรคคอร์ด โดยแต่ละพาร์ทิชันจะแสดงถึงข้อมูลที่ถูกแบ่งออกเป็นกลุ่ม ในแต่ละกลุ่มจะประกอบไปด้วยข้อมูลอย่างน้อยที่สุด 1 แถว และข้อมูลแต่ละแถวจะต้องถูกจัดให้อยู่ในกลุ่มข้อมูลเพียงกลุ่มเดียวเท่านั้น (สำหรับบางเทคนิคอาจอนุโลมให้ข้อมูลใดๆสามารถถูกจัดอยู่ในกลุ่มข้อมูลได้มากกว่า 1 กลุ่ม) ตัวอย่างเทคนิคของการจัดกลุ่มแบบนี้ได้แก่ k-means algorithm, k-modoids algorithm, CLARA (Clustering LARge Application), CLARANS (Clustering LARge Application based upon RANdomized Search)

2) Hierarchical Method การจัดกลุ่มข้อมูลประเภทนี้จะอาศัยหลักการแบ่งข้อมูลออกเป็นลำดับชั้นคล้ายต้นไม้ ซึ่งวิธีการแบ่งกลุ่มข้อมูลแบบนี้สามารถแบ่งออกเป็น 2 แนวทางตามลักษณะการสร้างลำดับชั้น คือ Agglomerative approach กับ Divisive approach



รูปที่ 2.16 การจัดกลุ่มโดยใช้ AGNES และ DIANA (Han and Kamber, 2006, p.409)

ตัวอย่างเทคนิคของการจัดกลุ่มแบบนี้ได้แก่ AGNES (Agglomerative NESting) ซึ่งจะเป็น agglomerative hierarchical clustering, DIANA (Divisive ANAlysis) ซึ่งจะเป็น divisive hierarchical clustering, BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies), CURE (Clustering Using REpresentatives)

3) Density-Based Method การจัดกลุ่มข้อมูลประเภทนี้จะพิจารณาความหนาแน่นของข้อมูลเป็นเกณฑ์ในการค้นหาคลัสเตอร์ หลักการทั่วไปของเทคนิคนี้คือการแผ่ขยายขอบเขตของคลัสเตอร์ไปเรื่อยๆ ตราบใดที่ความหนาแน่นของข้อมูลยังมีค่าน้อยกว่าหรือเท่ากับค่าที่ผู้ใช้กำหนด นั่นคือแต่ละข้อมูลของคลัสเตอร์ใดๆ จะต้องประกอบด้วยข้อมูลซึ่งอยู่ใกล้กันภายในรัศมีที่กำหนด (neighborhood) ด้วยเทคนิคนี้สามารถใช้ในการกรองข้อมูลรบกวน (noisy) ซึ่งเป็นข้อมูลที่มีความหนาแน่นเบาบางได้ และยังสามารถค้นหาคลัสเตอร์ที่รูปทรงซับซ้อนได้อีกด้วย ตัวอย่างเทคนิคของการจัดกลุ่มแบบนี้ได้แก่ DBSCAN

2.8.2 ขั้นตอนวิธีการจัดกลุ่ม k-means

ขั้นตอนวิธีการจัดกลุ่ม k-means เป็นเทคนิคหนึ่งที่จัดอยู่ในประเภท Partition Method มีการใช้ค่าเฉลี่ยของข้อมูลที่ถูกจัดให้อยู่ในคลัสเตอร์เดียวกันเป็นตัวแทนของทุกข้อมูลในคลัสเตอร์นั้น ขั้นตอนวิธีเริ่มต้นจากการรับค่าพารามิเตอร์ k ซึ่งค่านี้คือจำนวนคลัสเตอร์ที่ต้องการค้นหา จากนั้นขั้นตอนวิธีจะทำการสุ่มเลือกข้อมูลเริ่มต้นจำนวน k ชุด ซึ่งแต่ละชุดที่ได้มานั้นจะเป็นศูนย์กลางเริ่มต้นของแต่ละคลัสเตอร์ (centroid) จากนั้นทำการจัดกลุ่มให้กับข้อมูลที่เหลือ ข้อมูลจะถูกจัดให้อยู่ในคลัสเตอร์เดียวกันเมื่อข้อมูลนั้นมีความคล้ายกับตัวแทนของคลัสเตอร์นั้นมากที่สุด จากนั้นจึงทำการคำนวณหาค่าเฉลี่ยของคลัสเตอร์ใหม่ และดำเนินกระบวนการเดียวกันกับข้อมูลที่เหลือต่อไป จนกระทั่งทุกข้อมูลถูกจัดกลุ่มอย่างสมบูรณ์และข้อมูลไม่มีการเปลี่ยนกลุ่มอีกต่อไป

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the object in the cluster.

Input: - k, the number of cluster;
- D, a data set containing n objects;

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects form D as the initial cluster center;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) **until** no change;

รูปที่ 2.17 ขั้นตอนวิธี k-means (Han and Kamber, 2006, p.403)

การทำงานของ k-means จะมีประสิทธิภาพสูงก็ต่อเมื่อข้อมูลเกาะกลุ่มกันหนาแน่น แต่ละกลุ่มแยกจากกันอย่างชัดเจน และความหนาแน่นของข้อมูลแต่ละกลุ่มใกล้เคียงกัน

จุดเด่นของ k-means คือง่ายและสามารถใช้ได้กับข้อมูลหลายประเภท และยังมีประสิทธิภาพในด้านความเร็ว แต่จุดด้อยของ k-means คือยังไม่เหมาะสมกับข้อมูลทุกประเภท และไม่สามารถจัดการกลุ่มที่มีรูปร่างไม่เป็นรูปทรงกลมหรือกลุ่มที่มีขนาดหรือความหนาแน่นต่างกันได้ นอกจากนี้ k-means ยังถูกจำกัดสำหรับข้อมูลที่มีตัวแทนข้อมูลคลุมเครือหรือไม่ชัดเจน

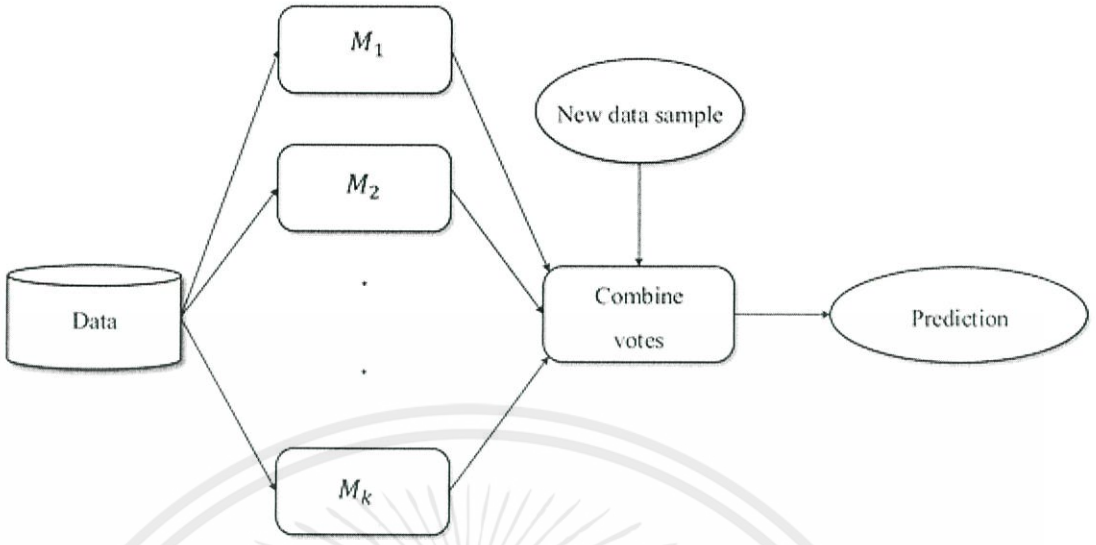
2.9 การรวมตัวกันของตัวจำแนกประเภท (Combining classifier)

ตัวจำแนกประเภท (Classifier) ทั่วไปนั้นจะให้ผลลัพธ์โดยการนำข้อมูลมาผ่านขั้นตอนวิธีการตัดสินใจของตัวจำแนกประเภทนั้นๆ ซึ่งแน่นอนว่าคำตอบนั้นจะได้มาด้วนการทำงานเพียงกระบวนการเดียว ทำให้ความถูกต้องและแม่นยำโดยเฉลี่ยมีค่าไม่สูงมากนัก (ยกเว้นกรณีที่ข้อมูลมีรูปแบบที่เรียบง่าย และแต่ละกลุ่มแยกจากกันอย่างชัดเจน) ด้วยเหตุนี้จึงมีงานวิจัยจำนวนมากที่ได้คิดค้นวิธีการที่จะปรับปรุงขั้นตอนวิธีใหม่แทนที่จะใช้ตัวจำแนกประเภทเพียงเทคนิคเดียวก็จะหันมาใช้หลายเทคนิค หรือใช้หลายโมเดลจำแนกประเภทเข้ามาทำงานร่วมกันแทน หรือเรียกว่า "การรวมกันของตัวจำแนกประเภท"

การรวมตัวกันของตัวจำแนกประเภท (combining classifier) สามารถแบ่งออกเป็นกลุ่มใหญ่ได้ 2 ประเภท ดังนี้

2.9.1 การรวมตัวกันของตัวจำแนกประเภทเดียวกัน (Combining homogeneous classifier)

วิธีการนี้จะใช้โมเดลจำแนกประเภทหลายๆ โมเดล ซึ่งแต่ละโมเดลจะใช้ขั้นตอนวิธีเดียวกันในการสร้างโมเดล เช่นถ้าเลือกใช้ขั้นตอนวิธีต้นไม้ตัดสินใจ C4.5 ตัวจำแนกประเภททุกตัวก็จะใช้ขั้นตอนวิธีตัดต้นไม้ตัดสินใจ C4.5 ทั้งหมดส่วนที่แตกต่างกันคือ ข้อมูลเรียนรู้ (Training Data) ที่ถูกแบ่งให้กับตัวจำแนกประเภทแต่ละแบบ ในท้ายที่สุดเราจะได้โมเดลจำแนกประเภททั้งหมด k แบบ $(M_1, M_2, \dots, M_k)$ ดังรูปที่ 2.18



รูปที่ 2.18 โครงสร้างการรวมตัวกันของตัวจำแนกประเภทเดียวกัน (Han and Kamber, 2006, p.366)

ในการทำนายข้อมูลใหม่นั้น (New data sample) จำใช้การโหวตเสียงข้างมาก เพื่อทำนายกลุ่มออกมา ตัวอย่างของวิธีการรวมกันของตัวจำแนกประเภทเดียวกัน ได้แก่ขั้นตอนวิธี Bagging และ Boosting

2.9.1.1 ขั้นตอนวิธี Bagging

คือการสร้างโมเดลจำแนกประเภทหลายโมเดล ด้วยชุดข้อมูลเรียนรูที่แตกต่างกัน แต่จะใช้เทคนิคในการสร้างโมเดลด้วยอัลกอริทึมเดียวกัน ซึ่งจะช่วยปรับปรุงประสิทธิภาพในการทำนายข้อมูลทั้งในปัญหาการจำแนกประเภท และการประมาณค่าได้ ขั้นตอนวิธีของ Bagging แสดงดังรูปที่ 2.19

แต่ละโมเดลจำแนกประเภท M_i จะถูกสอนด้วยชุดข้อมูลเรียนรู้ D_i

คำตอบสุดท้ายของการทำนายนั้น วิธี Bagging จะนับผลโหวตที่ได้จากโมเดลจำแนกประเภทที่มีหมด และกำหนดกลุ่มผลโหวตที่มากที่สุดให้กับข้อมูลใหม่

Algorithm: Bagging. The bagging algorithm—create an ensemble of models (classifiers or predictors) for a learning scheme where each model gives an equally-weight prediction

Input:
- D, a set of training tuple;
- k, the number of models in the ensemble;
- A learning scheme (e.g., decision tree algorithm, backpropagation, etc.)

Output: A composition model, M^*.

Method:

(1) **for** $i = 1$ to k **do** // create k models:

(2) create boosting sampling, D_i, by sampling D with replacement

(3) use D_i to derive a model, M_i;

(4) **end for**

To use composite model on a tuple, X:

(1) **if** classification **then**

(2) let each of the k models classify X and return the majority vote;

(3) **If** prediction **then**

(4) let each of the k models predict a value for X and return the average predicted value;

รูปที่ 2.19 ขั้นตอนวิธี Bagging (Han and Kamber, 2006, p.367)

2.9.1.2 ขั้นตอนวิธี Boosting

คือการใช้ขั้นตอนวิธีการเรียนรู้หลายๆโมเดลจำแนกประเภทในการตัดสินใจ โดยจะใช้การถ่วงน้ำหนักเข้ามาให้แต่ละโมเดลจำแนกประเภท ซึ่งค่าน้ำหนักนั้นจะได้มาจากความแม่นยำบนข้อมูลเรียนรู้ และสำหรับคำตอบสุดท้ายของการทำนายนั้น วิธี Boosting จะใช้การโหวตแบบถ่วงน้ำหนัก เพื่อกำหนดกลุ่มให้กับข้อมูลใหม่

ขั้นตอนที่ได้รับความนิยมอย่างมากของ Boosting คือ "Adaboost" ขั้นตอนวิธีของ Adaboost แสดงดังรูปที่ 2.20

Algorithm: Adaboost. A boosting algorithm—create an ensemble of classifiers. Each one gives a weighted vote

Input:
- D, a set of d class-labeled training tuples;
- k, the number of rounds (one classifier is generated per round);
- A classification learning scheme.

Output: A composite model.

Method:

(1) initialize the weight of each tuple in D to $1/d$;
(2) **for** i = 1 to k **do** // for each round:
(3) sample D with replacement according to the tuple weights to obtain D_i;
(4) use training set D_i to derive a model, M_i;
(5) compute error (M_i) , the error rate of M_i (Equation 6.66)
(6) **if** error $(M_i) > 0.5$ **then**
(7) reinitialize the weights to $1/d$
(8) go back to step 3 and try again;
(9) **end if**
(10) for each tuple in D_i that was correctly classified do
(11) Multiply the weight of the tuple by $error(M_i)/(1\text{-}error(M_i))$; // update weights
(12) normalize the weight of each tuple;
(13) **end for**

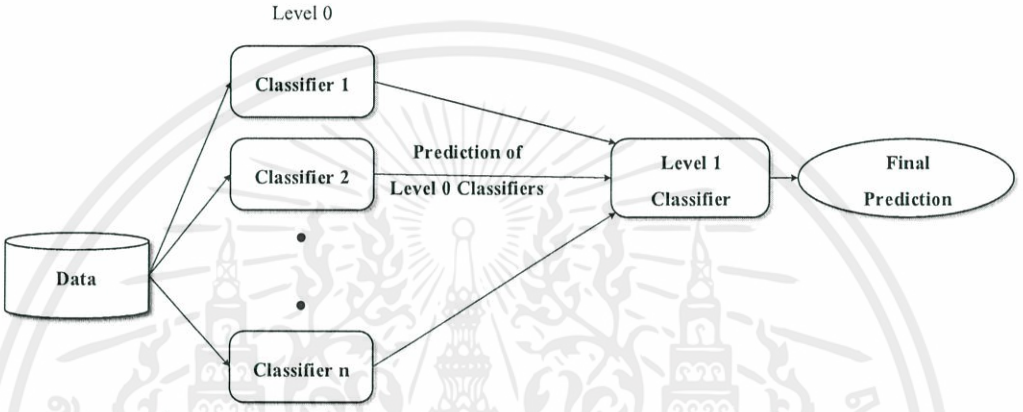
To use the composite model on a tuple, X:

(1) initialize weight of each class to 0;
(2) **for** i = 1 to k **do** // for each classifier:
(3) $w_i = \log \frac{1-error(M_i)}{error(M_i)}$ // weight of the classifier's vote
(4) C = $M_i(X)$; // get class prediction for X from M_i
(5) Add w_i to weight for class c
(6) **end for**
(7) return the class with the largest weight;

รูปที่ 2.20 ขั้นตอนวิธี Adaboost (Han and Kamber, 2006, p.369)

2.9.2 การรวมตัวกันของตัวจำแนกประเภทที่แตกต่างกัน (Combining heterogeneous classifier)

วิธีนี้จะใช้ตัวจำแนกประเภทหลายตัว ซึ่งแต่ละตัวจะใช้ขั้นตอนวิธีแตกต่างกันในการสร้างโมเดล การทำงานนั้นตัวโมเดลจำแนกประเภทจะถูกแบ่งออกเป็นระดับ (level) โดยที่ระดับสูงสุดจะเป็นตัวตัดสินใจผลลัพธ์สุดท้ายให้กับข้อมูลใหม่ ซึ่งการตัดสินใจนั้นจะอาศัยผลจากการทำนายของโมเดลจำแนกประเภทก่อนหน้าทั้งหมด แสดงได้ดังรูปที่ 2.21



รูปที่ 2.21 การทำงานของการรวมกันของตัวจำแนกประเภทที่แตกต่างกัน

ขั้นตอนการทำงานจะแบ่งออกเป็น 2 ขั้นตอน

ขั้นตอนการเรียนรู้ (Training Phase) ในระยะนี้นั้นจะมีการให้แต่ละตัวจำแนกประเภทระดับ 0 เรียนรู้โดยใช้เทคนิค (leave-one-out cross validation) แล้วจึงสร้างเมทริกซ์ เพื่อจัดเก็บผลทำนาย จากตัวจำแนกประเภทระดับ 0 ซึ่งมีขนาด n+1 แถว i คอลัมน์ โดย n คือจำนวนโมเดลจำแนกประเภทในระดับ 0 และอีก 1 ค่าคือกลุ่มที่แท้จริง (actual class) ของข้อมูล ส่วน i คือจำนวนข้อมูลที่ใช้ในการเรียนรู้ต่อมาจึงให้ตัวจำแนกประเภทระดับ 1 เรียนรู้โดยใช้เมทริกซ์ที่ได้มาจากก่อนหน้าสุดท้ายให้แต่ละตัวจำแนกประเภทระดับ 0 เรียนรู้อีกครั้งโดยใช้ข้อมูลเรียนรู้ทั้งหมด

ขั้นตอนประยุกต์ใช้งาน (Application Phase) ในระยะนี้จะใช้การจำแนกกลุ่มตัวอย่างใหม่ โดย เริ่มต้นจะใช้ตัวจำแนกระดับ 0 ทั้งหมด แล้วจึงเก็บผลลัพธ์สุดท้ายให้กับตัวอย่างใหม่

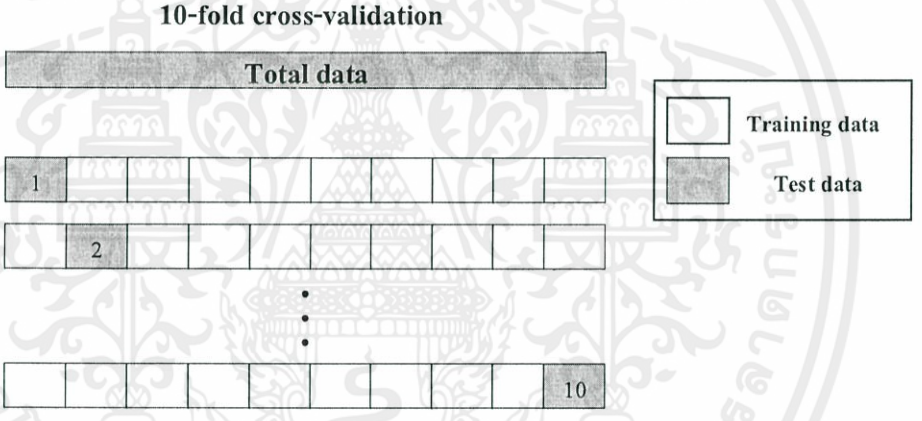
การรวมตัวจำแนกประเภทที่ต่างกันสามารถนำมาใช้ในกรณีที่จำนวนตัวอย่างในชุดข้อมูลมีจำนวนน้อย ซึ่งแตกต่างกับการทำงานของการรวมตัวกันของจำนวนตัวจำแนกประเภทเดียวกันที่ต้องการความหลากหลายของตัวอย่างมาช่วยในการสร้างโมเดล

2.10 การวัดประสิทธิภาพ (Performance Evaluation Measurement)

2.10.1 k-fold cross-validation

เทคนิค k-fold cross-validation (Ron, 1995) เป็นวิธีการวัดประสิทธิภาพในการทำนายตัวอย่างของโมเดล โดยพื้นฐานของเทคนิคนี้คือการสุ่มตัวอย่าง (resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆ หรือเรียกว่า fold และนำบางส่วนจากชุดข้อมูลนั้นมาทดสอบของลัพธ์จากการทำนายข้อมูลทดสอบของโมเดล

กรณีการเลือกสุ่มข้อมูลแบบความเที่ยงตรง k กลุ่ม เราจะแบ่งกลุ่มข้อมูลออกเป็น k ชุดเท่าๆกัน และทำการคำนวณค่าความแม่นยำจากการทำนาย k รอบ โดยแต่ละรอบจะมีการสร้างโมเดลจำแนกประเภทหนึ่งตัว จากข้อมูลเรียนรู้ k-1 ชุด และใช้ข้อมูลทดสอบ 1 ชุด (ชุดที่ไม่ได้นำมาเรียนรู้)



รูปที่ 2.22 10-fold cross-validation

จากรูปที่ 2.22 ในการทำงานรอบแรก ข้อมูลในชุดที่ 1 จะใช้เป็นข้อมูลทดสอบ ส่วนข้อมูลในชุดที่ 2 ถึง 10 จะนำมาใช้เป็นชุดข้อมูลสำหรับการเรียนรู้ ซึ่งจะได้โมเดลจำแนกประเภท 1 ตัว ต่อมารอบที่สอง ก็จะใช้ข้อมูลในชุดที่ 2 เป็นข้อมูลทดสอบ ส่วนข้อมูลในชุดที่ 1 และ 3 ถึง 10จะนำมาใช้เป็นชุดข้อมูลสำหรับเรียนรู้ ซึ่งจะได้โมเดลจำแนกประเภทอีก 1 ตัว จะมีการทำงานลักษณะนี้ไปเรื่อยๆ จนถึงรอบที่ 10 เป็นชุดข้อมูลทดสอบ ส่วนข้อมูลในชุดที่ 1 ถึง 9 จะนำมาใช้เป็นชุดข้อมูลสำหรับการเรียนรู้ และจะได้โมเดลจำแนกประเภทอีก 1 ตัว

2.10.2 มาตราวัดประสิทธิภาพของโมเดล

ในการวัดประสิทธิภาพของโมเดลจำแนกประเภทข้อมูลนั้น จะอาศัย Confusion Matrix ในการเก็บข้อมูลจำนวนแถวที่จำแนกจากกลุ่มข้อมูลจริงและกลุ่มข้อมูลจากการทำนาย โดยที่ตารางนั้นจะมีขนาด m × m โดยที่ m คือจำนวนของกลุ่ม

ตารางที่ 2.3 Confusion Matrix

		Predicted class	
Actual class	Class	C_1	C_2
	C_1	TP	FN
	C_2	FP	TN

ค่าต่างๆ ภายใน Confusion Matrix มีความหมายดังนี้

- True positive (TP) คือจำนวนข้อมูลที่โมเดลจำแนกกลุ่มเป็น C_1 และคำตอบเป็น C_1
- True negative (TN) คือจำนวนข้อมูลที่โมเดลจำแนกกลุ่มเป็น C_2 และคำตอบเป็น C_2
- False positive (FP) คือจำนวนข้อมูลที่โมเดลจำแนกกลุ่มเป็น C_1 และคำตอบเป็น C_2
- False negative (FN) คือจำนวนข้อมูลที่โมเดลจำแนกกลุ่มเป็น C_2 และคำตอบเป็น C_1

ตารางที่ 2.4 Confusion Matrix ของการจำแนกประเภทการสั่งซื้อคอมพิวเตอร์(Han and Kamber, 2006, p.360)

Classes	buy_computer = yes	buy_computer = no	Total	Accuracy (%)
buy_computer = yes	6,954	46	7,000	99.34
buy_computer = no	412	2,588	3,000	86.27
Total	7,366	2,634	10,000	95.37

จากตารางที่ 2.4 ชุดข้อมูลการซื้อคอมพิวเตอร์จำนวน 10000 ตัวอย่าง จำแนกออกเป็นกลุ่ม คือ buy_computer = yes จำนวน 7,000 ตัวอย่าง และ buy_computer = no จำนวน 3,000 ตัวอย่าง จากการจำแนกประเภทพบว่ามีจำนวนข้อมูลโมเดลจำแนกกลุ่มเป็น buy_computer = yes และคำตอบ buy_computer = yes เท่ากับ 6,954 ตัวอย่าง (True positive = 6,954) มีจำนวนข้อมูลที่โมเดลจำแนกเป็น buy_computer = no เท่ากับ 2,588 ตัวอย่าง (True negative = 2,588) มีจำนวนข้อมูลที่โมเดลจำแนกกลุ่มเป็น buy_computer = yes แต่คำตอบเป็น buy_computer = no เท่ากับ 412 ตัวอย่าง (False positive = 412) มีจำนวนข้อมูลที่โมเดลจำแนกกลุ่มเป็น buy_computer = no แต่คำตอบเป็น buy_computer = yes เท่ากับ 46 ตัวอย่าง(False negative = 46) คำนวณค่าความแม่นยำการจำแนก

ประเภทกลุ่ม buy_computer = yes ได้เท่ากับ 99.34% ค่าความแม่นยำการจำแนกประเภทกลุ่ม buy_computer = no ได้เท่ากับ 86.27% และค่าความแม่นยำการจำแนกประเภททั้งหมดเท่ากับ 95.37%

ค่าความแม่นยำ (Accuracy) คืออัตราการทำนายถูกต้อง มีสูตรในการคำนวณ คือ

$$acuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

บทที่ 3
การวิเคราะห์และออกแบบระบบ

Business Intelligence Application for Analyzing Business Data เป็นเว็บแอพพลิเคชั่นที่พัฒนาขึ้นมาเพื่อผู้ใช้ที่ต้องการวิเคราะห์ข้อมูลทางธุรกิจ ซึ่งมีการทำงานที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลทางธุรกิจโดยใช้วิธีการพื้นฐานของการทำเหมืองข้อมูล

3.1 ขอบเขตความสามารถของระบบ

1. ผู้ใช้สามารถเลือก Classifier ได้ 2 Classifier คือ
    1) ID3
    2) Naïve Bayes
2. ผู้ใช้สามารถสร้างรายงานที่ได้จากการวิเคราะห์ข้อมูล โดยประกอบด้วยข้อมูลความถูกต้องของโมเดล, ผลสรุปของการวิเคราะห์ข้อมูล และ ตารางผลลัพธ์ที่ได้จากการทำนายหรือจัดกลุ่มข้อมูล
3. ผู้ใช้สามารถนำออกรายงานผลการวิเคราะห์ข้อมูลในรูปแบบเอกสารประเภท PDF

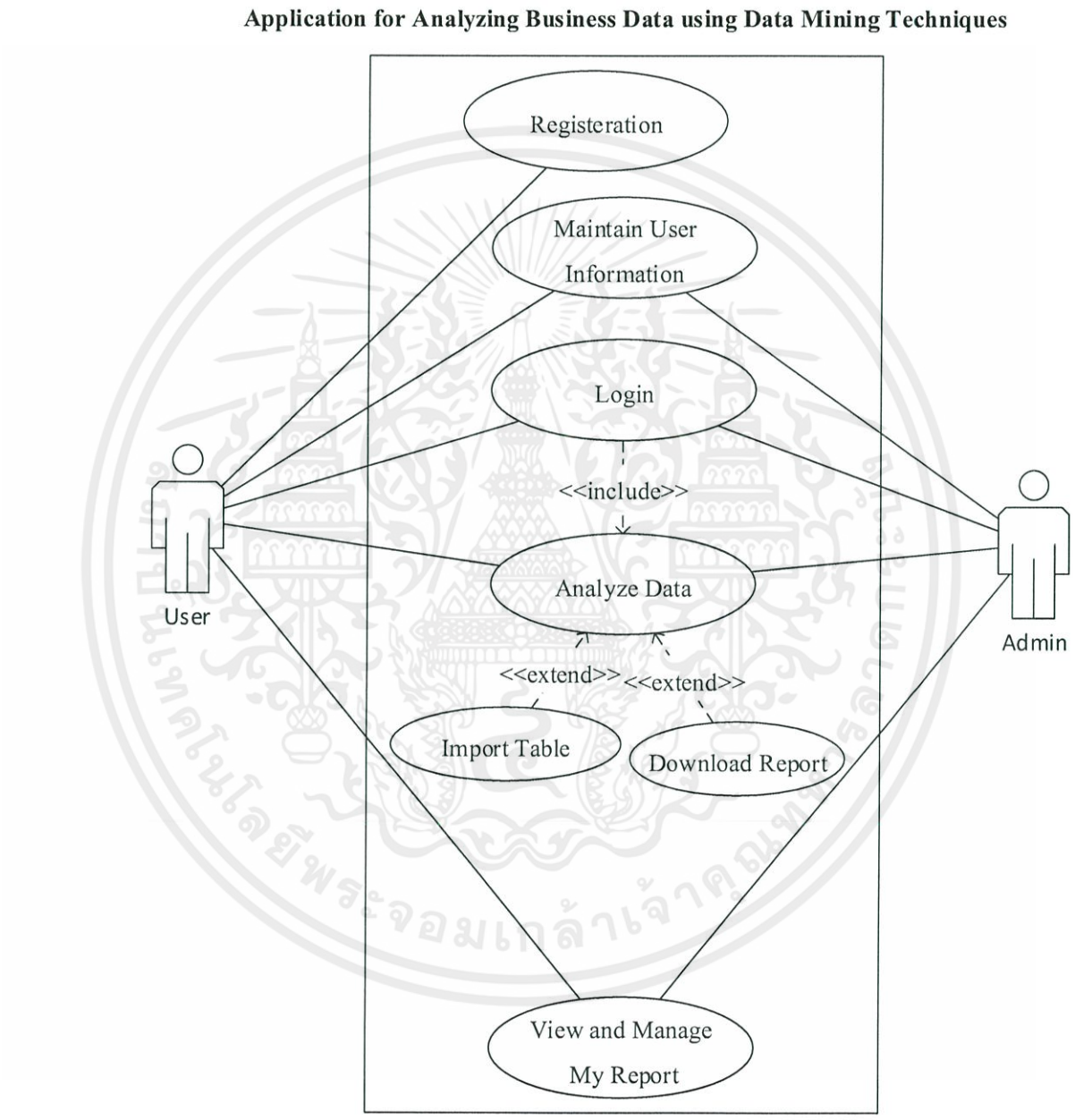
3.2 การออกแบบระบบ

ในการออกแบบระบบโครงการพิเศษฉบับนี้ใช้หลักการออกแบบเชิงวัตถุ (Object Oriented Approach) ซึ่งประกอบด้วยแผนภาพ Use Case และแผนภาพ Sequence แต่เนื่องจากฐานข้อมูลจะใช้ฐานข้อมูลประเภท Relational Database จึงต้องมีการใช้แผนภาพอีอาร์ (E-R Diagram) ในการอธิบายฐานข้อมูล โดย Use Case จะอธิบายเกี่ยวกับบทบาทในการใช้งานระบบของผู้ใช้แต่ละประเภทและแผนภาพ Sequence จะอธิบายจังหวะในการทำงานของ Object ในระบบ

3.2.1 การออกแบบโปรแกรมประยุกต์

3.2.1.1 Use Case Diagram

Use Case Diagram จะเป็น Diagram ที่แสดงให้เห็นถึงบทบาทการทำงานของ User, Admin



รูปที่ 3.1 แผนภาพ Use Case

ผู้ใช้งาน (User) : เมื่อผู้ใช้ต้องการเข้าใช้งานระบบ จำเป็นจะต้องทำการลงทะเบียนกับระบบก่อน จึงจะสามารถเข้ามาใช้งานในระบบได้ จากนั้นนำชื่อผู้ใช้และรหัสผ่าน ที่ได้ทำการลงทะเบียนไว้ มาเข้าใช้งานในระบบ เมื่อผู้ใช้เข้ามาในระบบแล้ว ระบบจะให้ผู้ใช้ทำการนำเข้าตารางข้อมูล จากนั้นผู้ใช้จึงทำการวิเคราะห์ข้อมูลด้วยวิธีการต่างๆ จากนั้น เมื่อได้ผลลัพธ์ที่ต้องการ ผู้ใช้งานจะสารมารถนำออกข้อมูลด้วยการดาวน์โหลดได้ นอกจากนี้ ผู้ใช้งานระบบยังสามารถเปลี่ยนแปลงและแก้ไขข้อมูลของตนเองเพื่อความถูกต้องได้

ผู้ดูแลระบบ (Admin) : ผู้ดูแลระบบมีความสามารถในการใช้งานระบบในการวิเคราะห์ข้อมูลต่างๆดังเช่นผู้ใช้ทั่วไป แตกต่างกันที่ผู้ดูแลระบบนั้น มีสิทธิ์ในการจัดการ เปลี่ยนแปลง แก้ไขข้อมูลของผู้ใช้อื่นได้ด้วย

3.2.1.2 คำอธิบาย Use Case

ตารางที่ 3.1 คำอธิบาย Use Case การลงทะเบียน

Use Case Name :	การลงทะเบียน	
Scenario :	ผู้ใช้งานสมัครเป็นสมาชิกของระบบ	
Trigger Event :	ผู้ใช้ต้องการสมัครสมาชิก	
Brief Description :	ก่อนที่ผู้ใช้จะเข้าใช้งานระบบได้ ผู้ใช้ต้องมีชื่อผู้ใช้และรหัสผ่านที่ได้จากการสมัครสมาชิก	
Actor :	ผู้ใช้งาน	
Related Use Case :	-	
Stakeholders :	ผู้ใช้งาน	
Precondition :	-	
Post condition :	ผู้เล่นได้รับชื่อผู้ใช้และรหัสผ่านในการเข้าระบบ	
Flow of Activity :	Actor	System
	1. เข้ามายังเว็บไซต์ 2. เลือกเมนูลงทะเบียน 3. ป้อนข้อมูลของผู้ใช้	4. บันทึกข้อมูลผู้ใช้
Exception Condition	-	

ผู้ใช้งานจำเป็นจะต้องทำการลงทะเบียนเป็นสมาชิกของระบบก่อน เพื่อจะนำชื่อผู้ใช้และรหัสผ่านเหล่านี้ ไปใช้ในการเข้าสู่ระบบ จากนั้นจึงสามารถใช้งานโปรแกรมได้

ตารางที่ 3.2 คำอธิบาย Use Case การเข้าสู่ระบบ

Use Case Name :	การเข้าสู่ระบบ	
Scenario :	ผู้ใช้งานใส่ชื่อผู้ใช้และรหัสผ่านเพื่อเข้าสู่ระบบ	
Trigger Event :	ผู้ใช้ต้องการเข้าสู่ระบบ	
Brief Description :	เมื่อผู้ใช้ได้ทำการสมัครเป็นสมาชิกแล้ว และต้องการเข้าใช้ระบบ	
Actor :	ผู้ใช้งาน	
Related Use Case :	-	
Stakeholders :	ผู้ใช้งาน	
Precondition :	การลงทะเบียน	
Post condition :	ผู้ใช้เข้ามาในระบบด้วยชื่อผู้ใช้ที่สมัครไว้	
Flow of Activity :	Actor	System
	1. เข้ามายังเว็บไซต์ 2. เลือกเมนูเข้าสู่ระบบ 3. ป้อนชื่อผู้ใช้และรหัสผ่านสำหรับเข้าระบบ	4. ตรวจสอบชื่อผู้ใช้และรหัสผ่านและเข้าไปยังหน้าแรก
Exception Condition	หากชื่อผู้ใช้หรือรหัสผ่านไม่ถูกต้อง ให้กลับไปหน้ากรอกข้อมูลเพื่อให้ใส่ข้อมูลอีกครั้ง	

เมื่อผู้ใช้ต้องการเข้าสู่ระบบ เพื่อที่จะใช้งานโปรแกรมวิเคราะห์ ผู้ใช้จะนำชื่อผู้ใช้และรหัสผ่านที่ได้ทำการลงทะเบียนกับระบบ มาใช้เพื่อการเข้าสู่ระบบ หลังจากใส่ชื่อผู้ใช้และรหัสผ่านเรียบร้อยแล้ว จากนั้นกดปุ่มเข้าสู่ระบบ ผู้ใช้ก็จะสามารถเข้าสู่ระบบได้

ตารางที่ 3.3 คำอธิบาย Use Case การจัดการข้อมูลผู้ใช้

Use Case Name :	การจัดการข้อมูลผู้ใช้	
Scenario :	แก้ไขหรือเปลี่ยนแปลงข้อมูลผู้ใช้	
Trigger Event :	ผู้ใช้ต้องการเปลี่ยนแปลงหรือแก้ไขข้อมูลผู้ใช้	
Brief Description :	เมื่อผู้ใช้ต้องการแก้ไขข้อมูลผู้ใช้ ผู้ใช้งานจะต้องกดที่เมนู แก้ไขข้อมูลส่วนตัวหรือผู้ดูแลเข้าสู่เมนูจัดการผู้ใช้	
Actor :	ผู้ใช้งาน, ผู้ดูแลระบบ	
Related Use Case :	-	
Stakeholders :	ผู้ใช้งาน	
Precondition :	การเข้าสู่ระบบ	
Post condition :	ข้อมูลของผู้ใช้เปลี่ยนแปลงไปตามที่ต้องการ	
Flow of Activity :	Actor	System
	1. เลือกเมนูจัดการข้อมูลผู้ใช้ 2. เลือกผู้ใช้ที่ต้องการแก้ไข 3. แก้ไขข้อมูลผู้ใช้แล้วบันทึก	4. บันทึกข้อมูล
Exception Condition		

เมื่อผู้ใช้งานต้องการแก้ไขข้อมูลส่วนตัว ผู้ใช้จะสามารถเข้าไปแก้ไขได้ในเมนู แก้ไขข้อมูลส่วนตัว หากเป็นผู้ดูแลระบบ จะสามารถเข้าไปแก้ไขข้อมูลผู้อื่นได้ในเมนูจัดการข้อมูลผู้ใช้

ตารางที่ 3.4 คำอธิบาย Use Case การนำเข้าตารางข้อมูล

Use Case Name :	การนำเข้าตารางข้อมูล	
Scenario :	ผู้ใช้ทำการนำเข้าตารางข้อมูลของผู้ใช้เข้าสู่ระบบ	
Trigger Event :	ผู้ใช้ต้องการนำเข้าตารางข้อมูล	
Brief Description :	เมื่อจะทำการวิเคราะห์ข้อมูล ผู้ใช้จำเป็นต้องนำเข้าตารางข้อมูลเข้าสู่ระบบ เพื่อใช้ในการวิเคราะห์	
Actor :	ผู้ใช้งาน	
Related Use Case :	-	
Stakeholders :	ผู้ใช้งาน	
Precondition :	การเข้าสู่ระบบ	
Post condition :	ระบบแสดงข้อความนำเข้าข้อมูลเสร็จสิ้น	
Flow of Activity :	Actor	System
	1. เลือกเมนูวิเคราะห์ข้อมูล 2. กดปุ่มนำเข้าข้อมูล 3. เลือกข้อมูลที่ต้องการนำเข้าและกดตกลง	4. บันทึกข้อมูลที่นำเข้ามา พร้อมวิเคราะห์ข้อมูลดังกล่าว
Exception Condition	-	

เมื่อผู้ใช้จะทำการวิเคราะห์ข้อมูล ผู้ใช้จำเป็นต้องนำเข้าตารางข้อมูลเข้าสู่ระบบก่อน เพื่อให้ระบบทำการวิเคราะห์ข้อมูลเหล่านั้น และเมื่อการนำเข้าข้อมูลเสร็จสิ้น ระบบจะแสดงข้อความแสดงสถานะของการนำเข้าเสร็จสมบูรณ์ จากนั้นจึงเริ่มเข้าสู่ขั้นตอนการวิเคราะห์ต่อไป

ตารางที่ 3.5 คำอธิบาย Use Case การวิเคราะห์ข้อมูล

Use Case Name :	การวิเคราะห์ข้อมูล	
Scenario :	ระบบจะทำการวิเคราะห์ข้อมูลที่นำเข้ามาด้วยวิธีที่ผู้ใช้กำหนด	
Trigger Event :	ผู้ใช้ต้องการวิเคราะห์ข้อมูลใดๆ	
Brief Description :	เมื่อผู้ใช้ต้องการวิเคราะห์ข้อมูล ผู้ใช้จะกดที่เมนูวิเคราะห์ข้อมูล จากนั้น ทำการนำเข้าตารางข้อมูล และเลือกวิธีการวิเคราะห์ต่อไป	
Actor :	ผู้ใช้งาน	
Related Use Case :	-	
Stakeholders :	ผู้ใช้งาน	
Precondition :	การเข้าสู่ระบบและการนำเข้าข้อมูล	
Post condition :	ระบบแสดงผลการวิเคราะห์ให้ผู้ใช้ทราบ	
Flow of Activity :	Actor	System
	1. เลือกเมนูวิเคราะห์ข้อมูล 2. นำเข้าข้อมูล 3. ใส่รายละเอียดและตั้งค่าพารามิเตอร์	4. ประมวลผลและแสดงรายงาน
Exception Condition	หากข้อมูลหรือพารามิเตอร์ไม่ถูกต้องจะกลับไปหน้ากรอกข้อมูลเพื่อให้ใส่ข้อมูลอีกครั้ง	

การวิเคราะห์ข้อมูลจะเริ่มต้นด้วยการที่ผู้ใช้นำเข้าตารางข้อมูลผู้ใช้ต้องการวิเคราะห์ จากนั้น ระบบจะให้ผู้ใช้เลือกวิธีการวิเคราะห์ที่ต้องการ จากนั้น เมื่อการวิเคราะห์เสร็จสิ้น ระบบจะแสดงผลการวิเคราะห์ในหน้าจอแก่ผู้ใช้งาน

ตารางที่ 3.6 คำอธิบาย Use Case การนำออกผลการวิเคราะห์

Use Case Name :	การนำออกผลการวิเคราะห์	
Scenario :	ผู้ใช้งานนำออกข้อมูลผลการวิเคราะห์	
Trigger Event :	ผู้ใช้ต้องการข้อมูลการวิเคราะห์	
Brief Description :	เมื่อผู้ใช้ต้องการผลการวิเคราะห์ ผู้ใช้สามารถดาวน์โหลดผลการวิเคราะห์ได้	
Actor :	ผู้ใช้งาน	
Related Use Case :	-	
Stakeholders :	ผู้ใช้งาน	
Precondition :	การวิเคราะห์ข้อมูล	
Post condition :	แสดงหน้าจอยืนยันการดาวน์โหลด	
Flow of Activity :	Actor	System
	1. กดปุ่มนำออกข้อมูล	2. แสดงหน้าต่างยืนยันการดาวน์โหลด
Exception Condition	-	

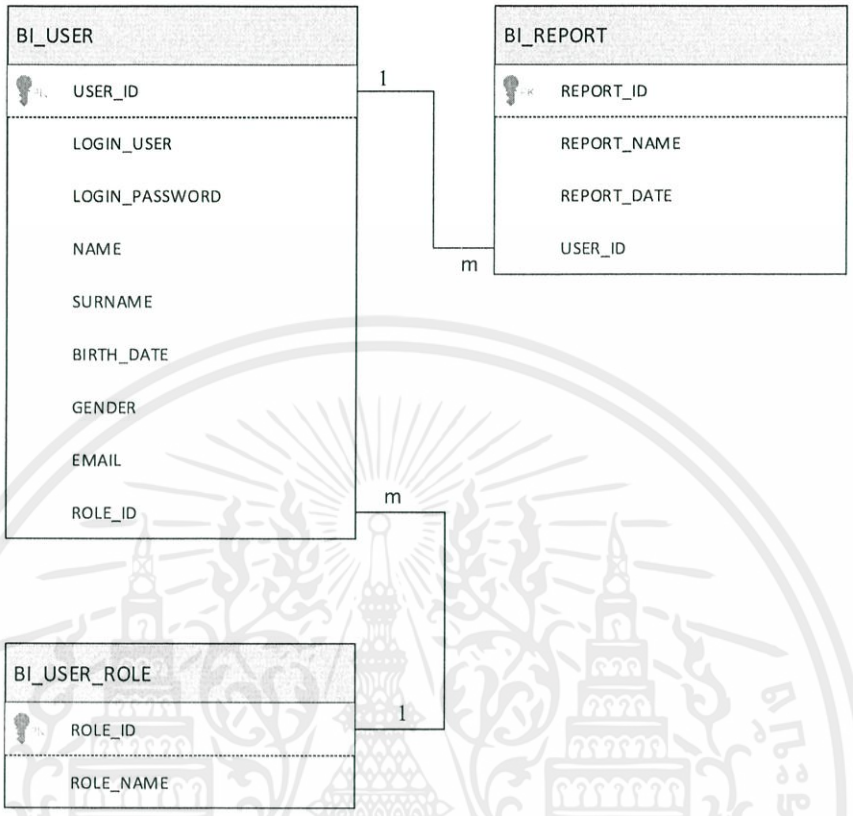
เมื่อผู้ใช้ได้ทำการวิเคราะห์เสร็จแล้ว ผู้ใช้สามารถนำออกรายงานผลการวิเคราะห์ออกมาเป็นไฟล์ประเภท PDF ได้

ตารางที่ 3.7 คำอธิบาย Use Case การดูรายงานและจัดการรายงานการประมวลผลย้อนหลัง

Use Case Name :	การดูรายงานและจัดการรายงานการประมวลผลย้อนหลัง	
Scenario :	ผู้ใช้เปิดดูรายงานหรือจัดการรายงานผลการประมวลผลย้อนหลังทั้งหมด	
Trigger Event :	ผู้ใช้ต้องการดูรายงานหรือจัดการรายงานผลการประมวลผลในครั้งที่ผ่านมา	
Brief Description :	เมื่อผู้ใช้ต้องการดูรายงานหรือจัดการรายงานการประมวลผลย้อนหลังที่ผู้ใช้เคยทำการประมวลผลไปแล้ว	
Actor :	ผู้ใช้งาน	
Related Use Case :	-	
Stakeholders :	ผู้ใช้งาน	
Precondition :	การวิเคราะห์ข้อมูล	
Post condition :	แสดงหน้าจอยืนยันการดาวน์โหลด	
Flow of Activity :	Actor	System
	1. เข้ามายังเมนูรายงานของฉัน 2. เลือกรายงานที่ต้องการดู หรือต้องการลบ	3. แสดงหน้าต่างยืนยันการทำงาน
Exception Condition	-	

เมื่อผู้ใช้เคยทำการวิเคราะห์ข้อมูลมาก่อนแล้ว ระบบจะเก็บผมการประมวลผลไว้ ในรูปของเอกสาร PDF และเมื่อผู้ใช้ต้องการดูผลย้อนหลัง ก็สามารถเข้ามาดูและดาวน์โหลดเอกสารเหล่านี้ออกไปได้ นอกจากนั้นยังสามารถลบรายงานเหล่านี้ได้ด้วย

3.2.2 แผนภาพอีอาร์ (The Entity Relationship Diagram)



ภาพที่ 3.2 แผนภาพอีอาร์

ตาราง BI_USER

ตารางนี้จะเก็บข้อมูลของผู้ใช้งาน ทั้งข้อมูลการเข้าสู่ระบบ เช่น ชื่อผู้ใช้ หรือรหัสผ่าน รวมทั้งเก็บข้อมูลเบื้องต้นของผู้ใช้ เช่น ชื่อ นามสกุล อีเมล์ เป็นต้น

ตาราง BI_USER_POSITION

ตารางนี้จะเก็บตำแหน่งของผู้ใช้งานในระบบ โดยในระบบนี้จะมีเพียง ผู้ใช้งาน กับ ผู้ดูแลระบบ เท่านั้น

ตาราง BI_REPORT

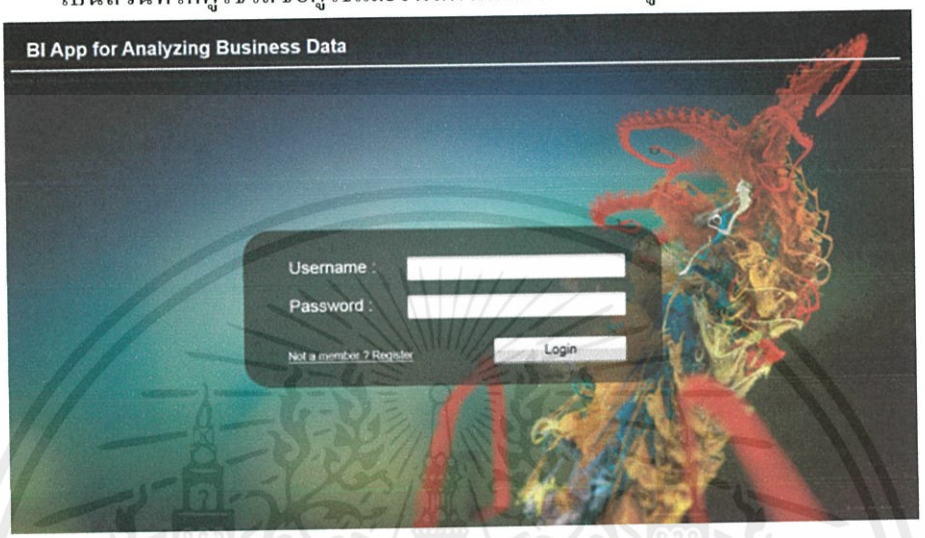
ตางรางนี้จะเก็บข้อมูลของรายงาน ไม่ว่าจะเป็นรหัสรายงาน, ชื่อรายงาน, ค่าความถูกต้องของขั้นตอนวิธีที่ใช้ในการวิเคราะห์รายงานแต่ละแบบ, ชื่อของข้อมูลเรียนรู้ และชื่อของข้อมูลทดสอบเป็นต้น

3.3 การออกแบบส่วนติดต่อผู้ใช้

ส่วนติดต่อผู้ใช้ประกอบไปด้วยหน้าเว็บไซต์ดังต่อไปนี้

3.3.1 หน้าการเข้าสู่ระบบ

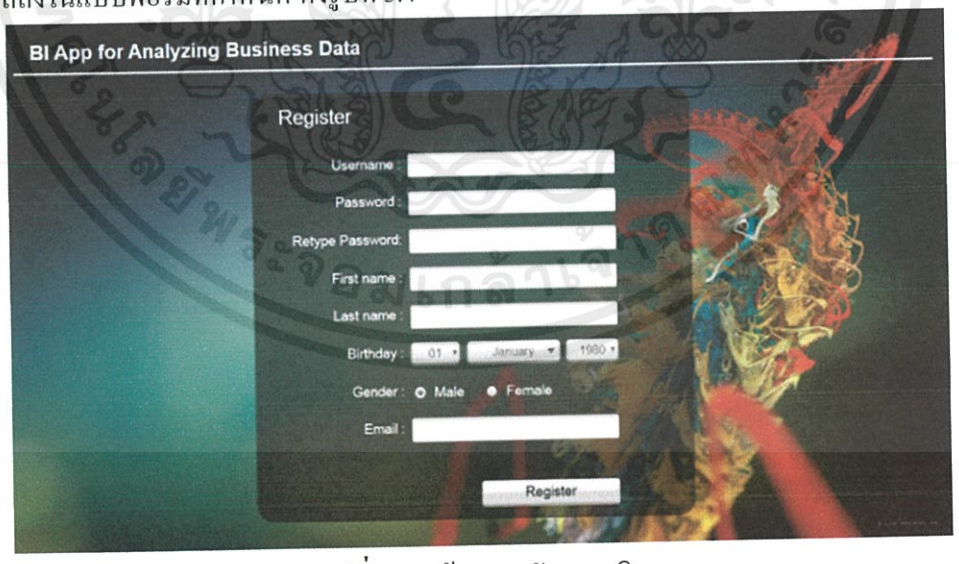
เป็นส่วนที่ให้ผู้ใช้ใส่ชื่อผู้ใช้และรหัสผ่านเพื่อทำการเข้าสู่ระบบ



รูปที่ 3.3 หน้าการเข้าสู่ระบบ

3.3.2 หน้าสมัครสมาชิก

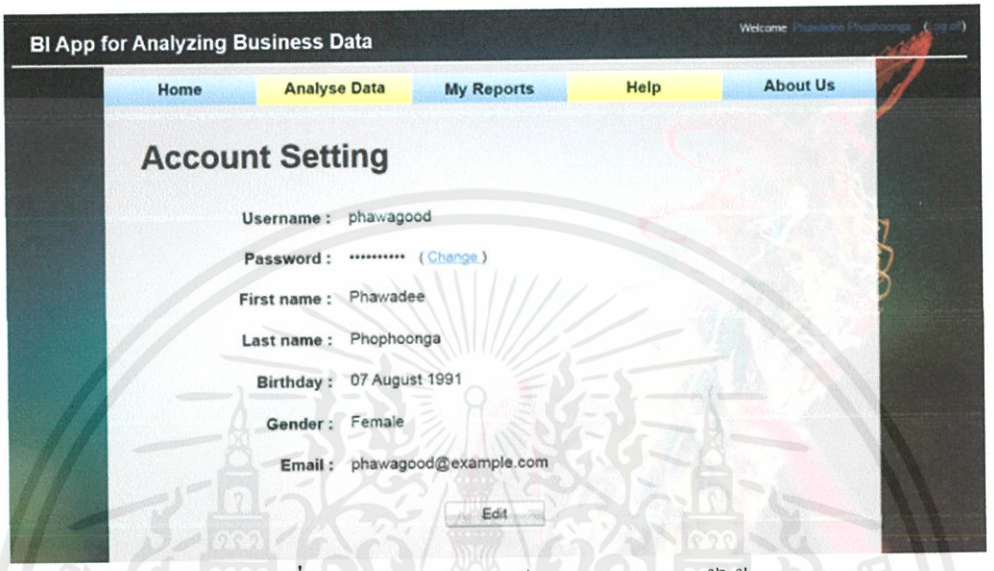
เป็นหน้าเว็บไซต์สำหรับการลงทะเบียนเป็นสมาชิกภายในระบบ โดยให้ผู้ใช้กรอกข้อมูลลงในแบบฟอร์มที่กำหนด ดังรูปที่ 3.4



รูปที่ 3.4 หน้าการสมัครสมาชิก

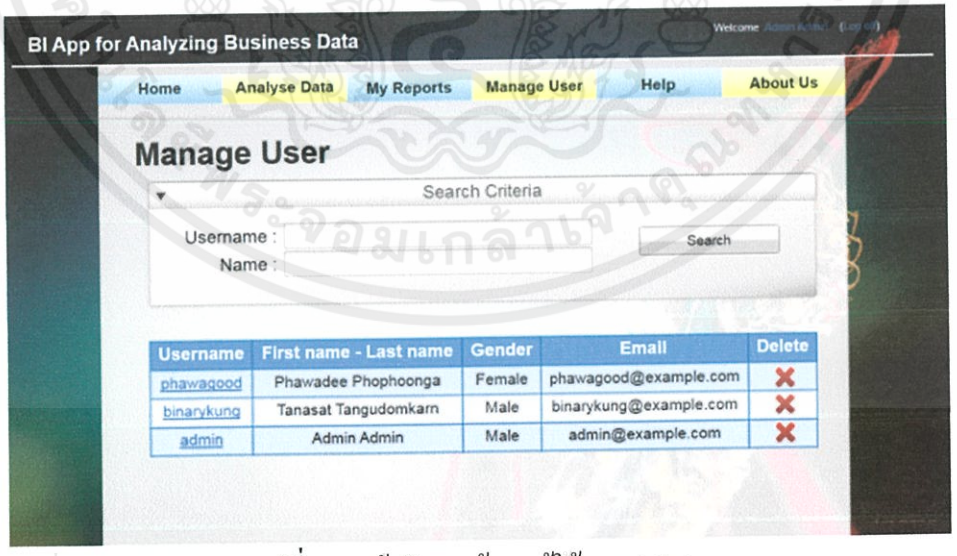
3.3.3 หน้าจัดการข้อมูลของผู้ใช้

เป็นหน้าที่ให้ผู้ใช้สามารถจัดการกับข้อมูลของผู้ใช้เองได้ ไม่ว่าจะเป็นการเปลี่ยนรหัสผ่าน หรือแก้ไขข้อมูลส่วนตัวสามารถใช้งานได้โดยการกดลิงค์ที่เป็นชื่อผู้ใช้ทางด้านขวาบนของหน้าจอ หน้าที่ใช้จัดการข้อมูลส่วนตัวของผู้ใช้จะแสดงดังรูปที่ 3.5



รูปที่ 3.5 หน้าการจัดการข้อมูลส่วนตัวของผู้ใช้

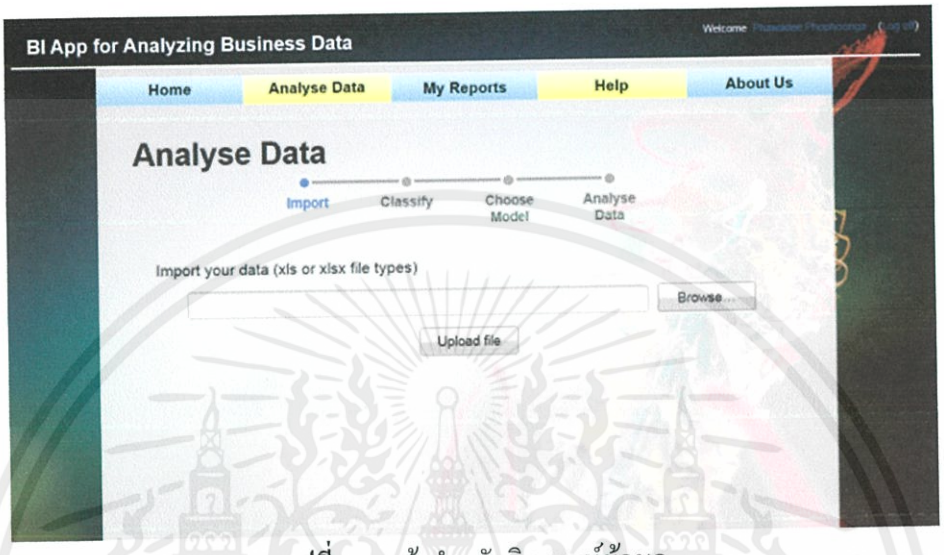
ส่วนต่อมาจะเป็นหน้าเว็บไซต์สำหรับจัดการข้อมูลผู้ใช้ของ Admin โดยที่ Admin ของระบบ สามารถที่จะทำการแก้ไข หรือลบข้อมูลของผู้ใช้อื่นๆได้ โดยกดที่แถบเมนู Manage User ดังรูปที่ 3.6



รูปที่ 3.6 หน้าจัดการข้อมูลผู้ใช้ของ Admin

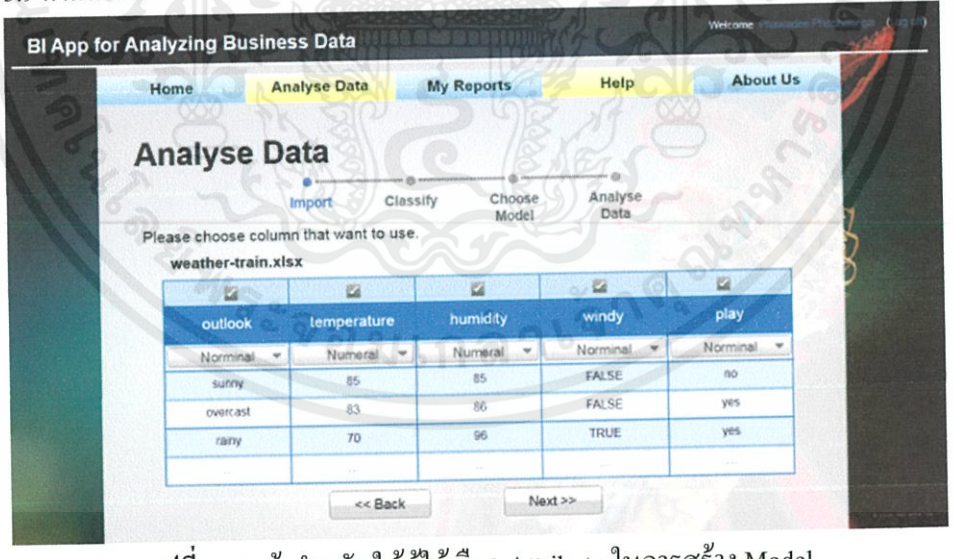
3.3.4 หน้าสำหรับวิเคราะห์ข้อมูล

เป็นหน้าสำหรับให้ผู้ใช้นำข้อมูลทางธุรกิจเข้ามาวิเคราะห์และทำนาย ภายในระบบผู้ใช้สามารถเลือกใช้งานได้ที่แถบเมนู Analyse Data ดังรูปที่ 3.7 จะเป็นขั้นตอนให้ผู้ใช้นำเข้าเอกสารประเภท Excel ที่จะนำมาเป็นตัว Training Data

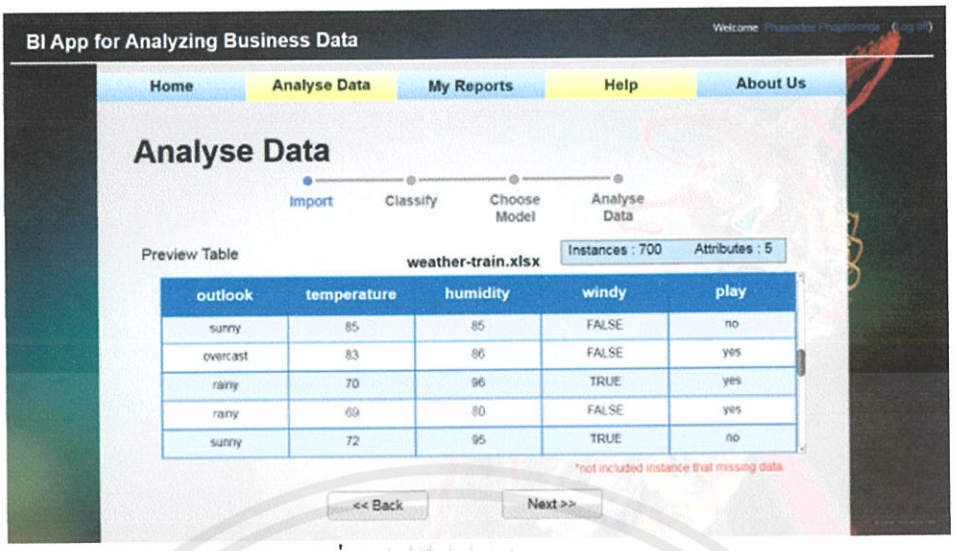


รูปที่ 3.7 หน้าสำหรับวิเคราะห์ข้อมูล

เมื่อนำเข้าเอกสารแล้วผู้ใช้ต้องทำการเลือก Attribute ที่จะใช้ในการสร้าง Model หลังจากที่เลือกแล้วระบบจะ Preview Train Table ที่มี attribute ตามที่ User ได้เลือกไว้ ดังรูปที่ 3.8 และ 3.9 ตามลำดับ

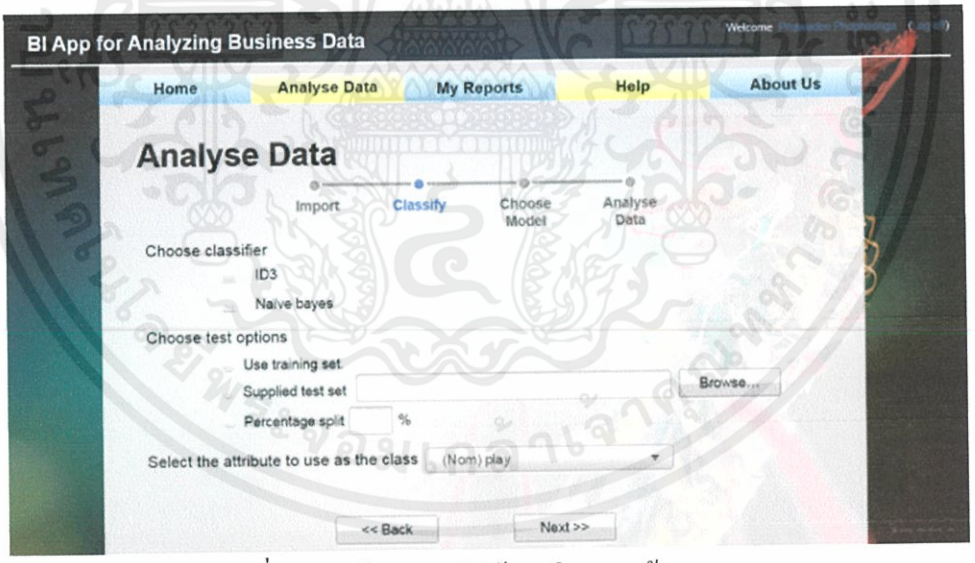


รูปที่ 3.8 หน้าสำหรับให้ผู้ใช้เลือก Attribute ในการสร้าง Model



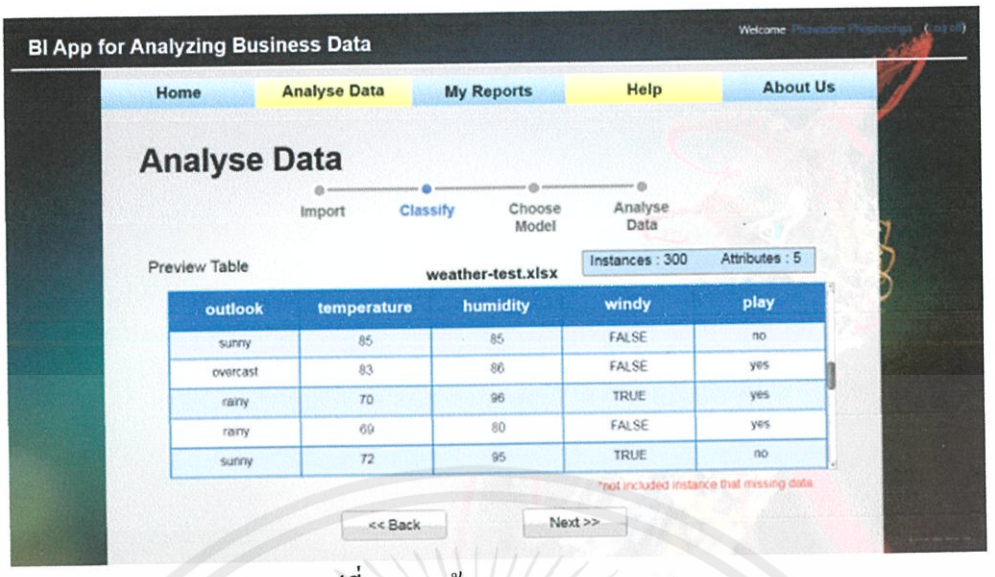
รูปที่ 3.9 หน้า Preview Train Table

จากนั้นในหน้าถัดไปจะเป็นหน้าเว็บไซต์สำหรับให้ผู้ใส่ข้อมูลสำหรับการสร้าง Model ได้แก่ข้อมูล Classifier ในส่วนนี้ ผู้ใช่สามารถเลือก Classifier 1 หรือเลือกทั้งหมดเลยก็ได้ จากนั้นเลือก Option ในการทดสอบ Model มีให้เลือก 3 แบบด้วยกัน คือ Use training set, Supplied test set หรือ Percentage split เมื่อผู้ใช้เลือกเสร็จแล้ว ผู้ใช้จะต้องระบุ Attribute ที่จะนำไปใช้เป็น Class ในการทำนาย ดังรูปที่ 3.10



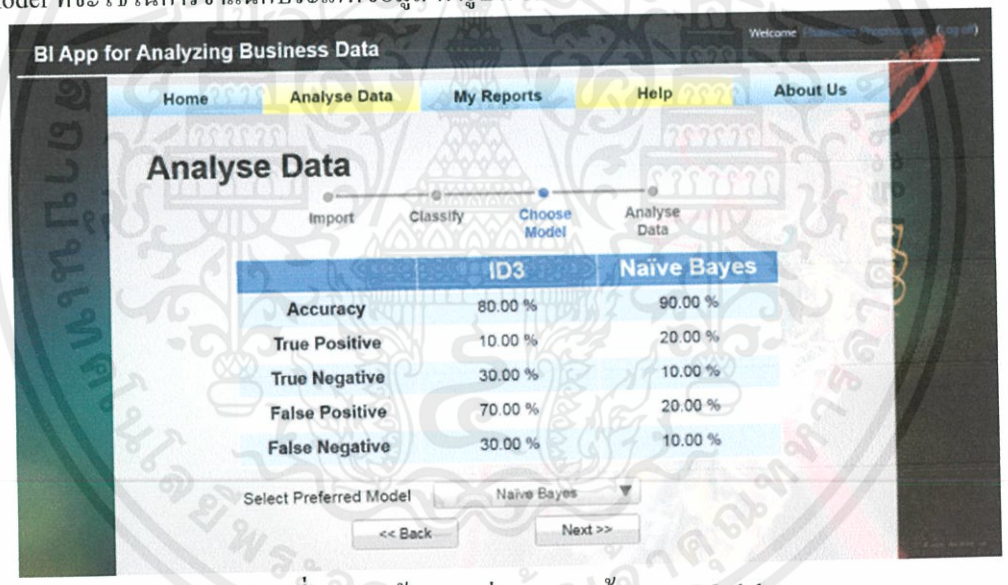
รูปที่ 3.10 หน้าสำหรับใส่ข้อมูลในการสร้าง Model

ระบบจะทำการ Preview test table ที่ผู้ใช้ได้เลือกไว้จากรูปที่ 3.10 ดังรูปที่ 3.11



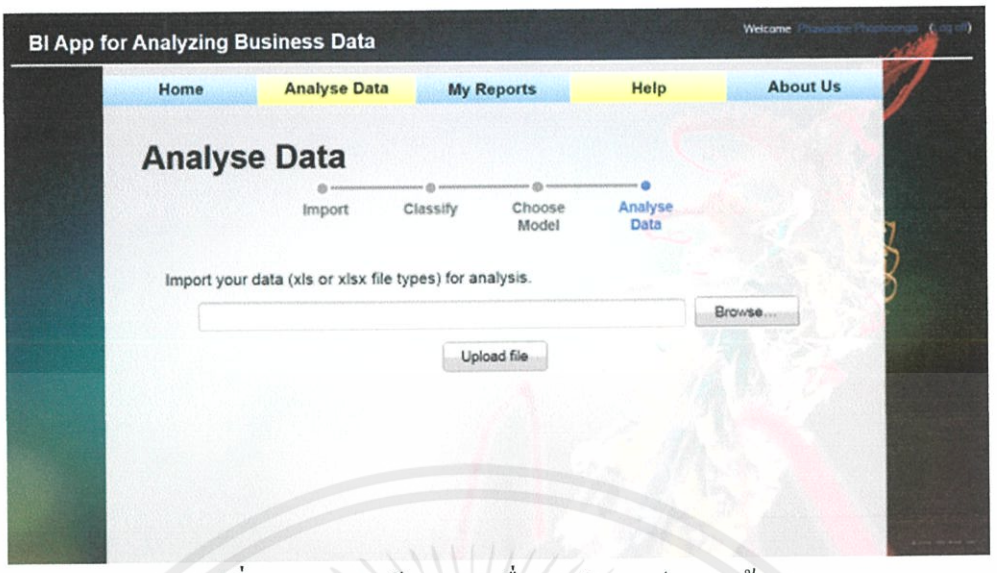
รูปที่ 3.11 หน้า Preview Test Table

ต่อมาระบบจะแสดงตารางเปรียบเทียบค่าความแม่นยำ ที่ได้จากการสร้างโมเดลแต่ละแบบที่ผู้ใช้ได้เลือกไว้ เมื่อดูรายงานเปรียบเทียบค่าความแม่นยำของ Model แล้ว ผู้ใช้ต้องเลือก Model ที่จะใช้ในการจำแนกประเภทข้อมูล ดังรูปที่ 3.12

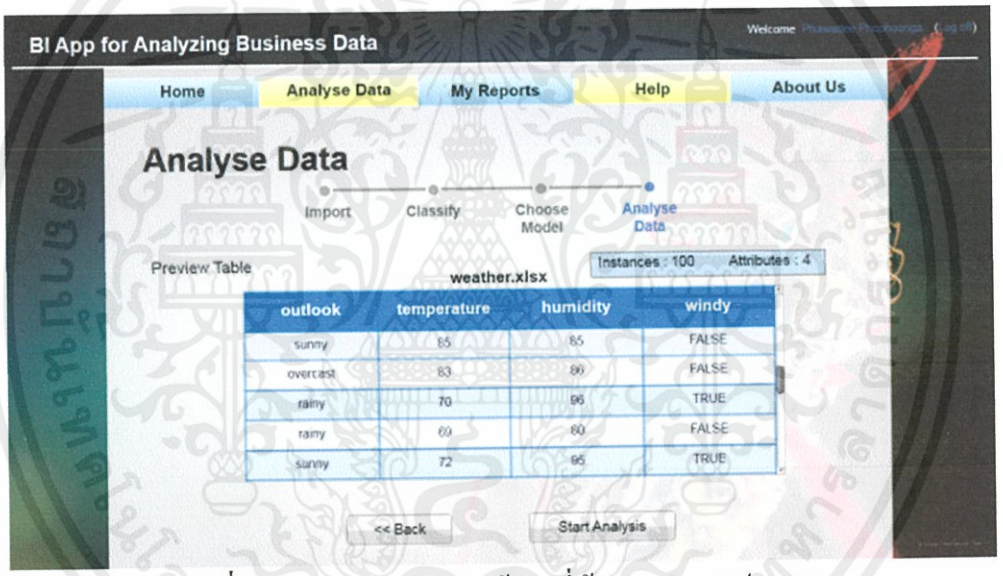


รูปที่ 3.12 หน้าแสดงค่าความถูกต้องของ Model

หลังจากที่เลือก Model ที่จะใช้ในการจำแนกประเภทข้อมูลแล้ว ผู้ใช้สามารถนำเข้าเอกสารข้อมูล ที่ต้องการจำแนกประเภทข้อมูล และระบบจะทำการ Preview table ของข้อมูลที่ต้องการจำแนกประเภท ดังรูปที่ 3.13 และ 3.14 ตามลำดับ



รูปที่ 3.13 การนำเข้าเอกสารเพื่อการจำแนกประเภทข้อมูล



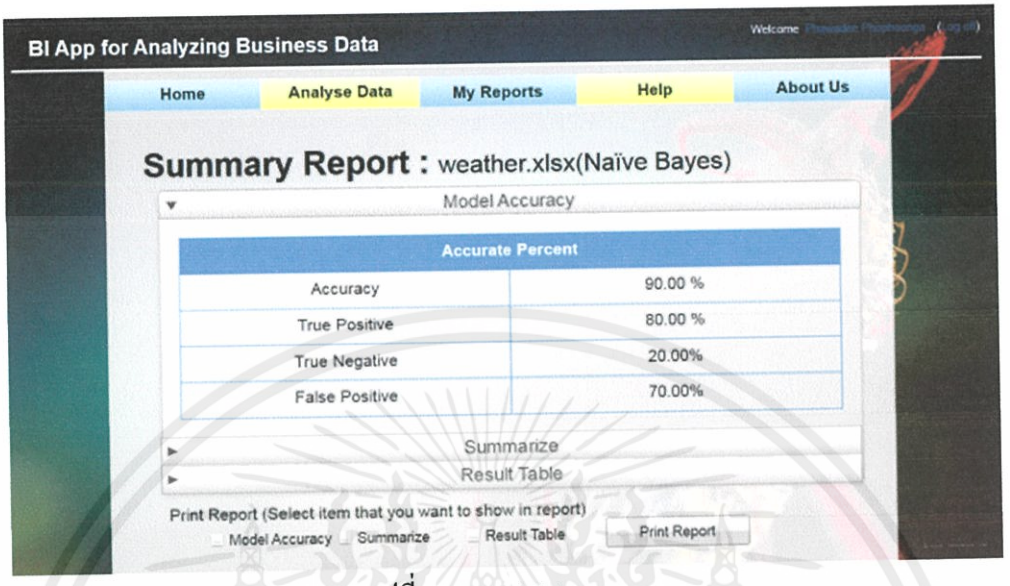
รูปที่ 3.14 Preview table ของข้อมูลที่ต้องการจำแนกประเภท

และในขั้นตอนสุดท้ายระบบจะแสดง report โดยผู้ใช้สามารถเลือกข้อมูลที่จะแสดงใน Report ของผู้ใช้เองได้ ว่าจะให้ Report ประกอบด้วยข้อมูลอะไรบ้าง ข้อมูลของ Report จะมีให้เลือก 3 อย่างด้วยกันคือ

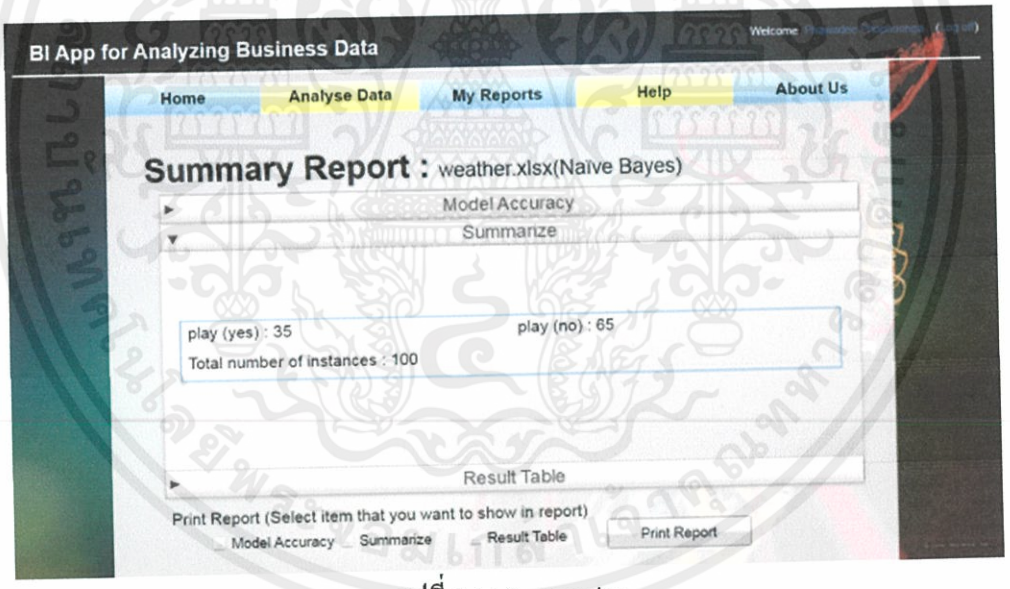
1. Model Accuracy คือรายงานที่จะแสดงค่าความถูกต้องของ Model ที่ผู้ใช้เลือกใช้ในการจำแนกประเภทข้อมูล

2. Summarize คือรายงานที่แสดงข้อมูลเกี่ยวกับ Class ในการทำนาย ว่ามีจำนวนแถวของข้อมูลที่ตรงกับ Class แต่ละ Class เป็นจำนวนเท่าไหร่บ้าง พร้อมทั้งแสดงจำนวนข้อมูลทั้งหมดที่ได้ทำการจำแนกประเภทไว้

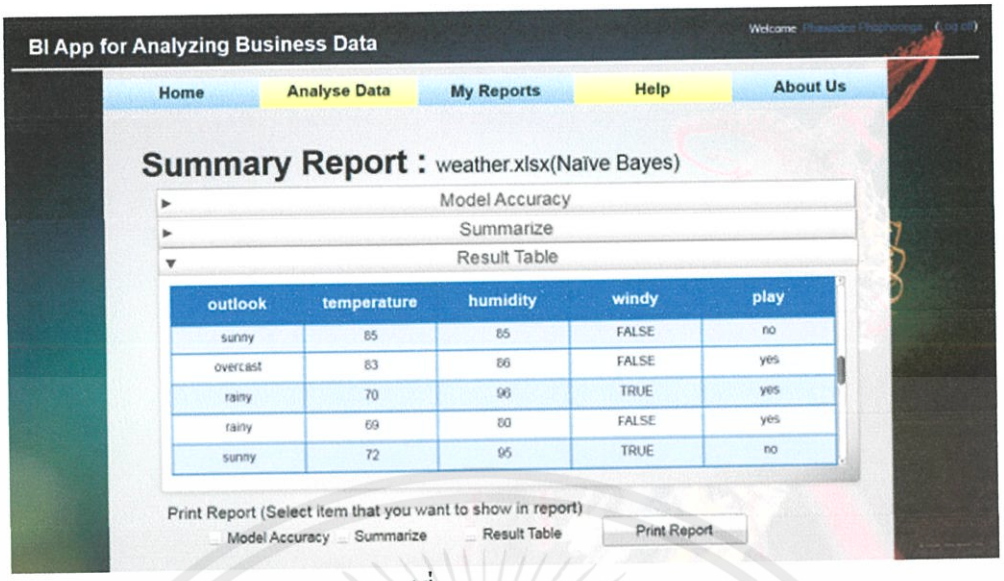
3. Result Table จะแสดงตารางข้อมูลที่ได้ผ่านการจำแนกประเภทข้อมูลเรียบร้อยแล้ว จะแสดงตัวอย่างของ ข้อมูลทั้งหมด ในรูปที่ 3.15, 3.16 และ 3.17 ตามลำดับ



รูปที่ 3.15 Model Accuracy



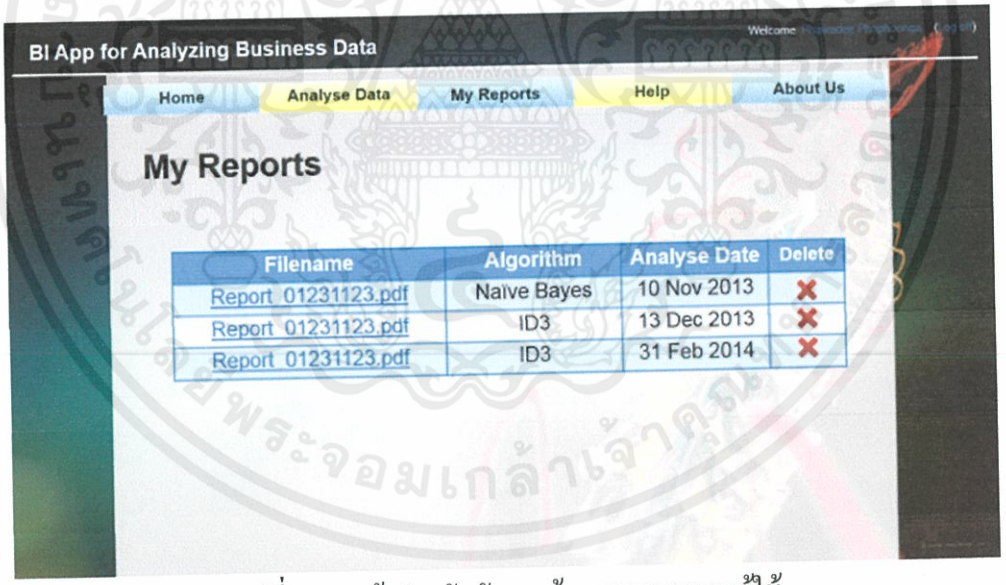
รูปที่ 3.16 Summarize



รูปที่ 3.17 Result Table

3.3.5 หน้าจัดการรายงานของผู้ใช้

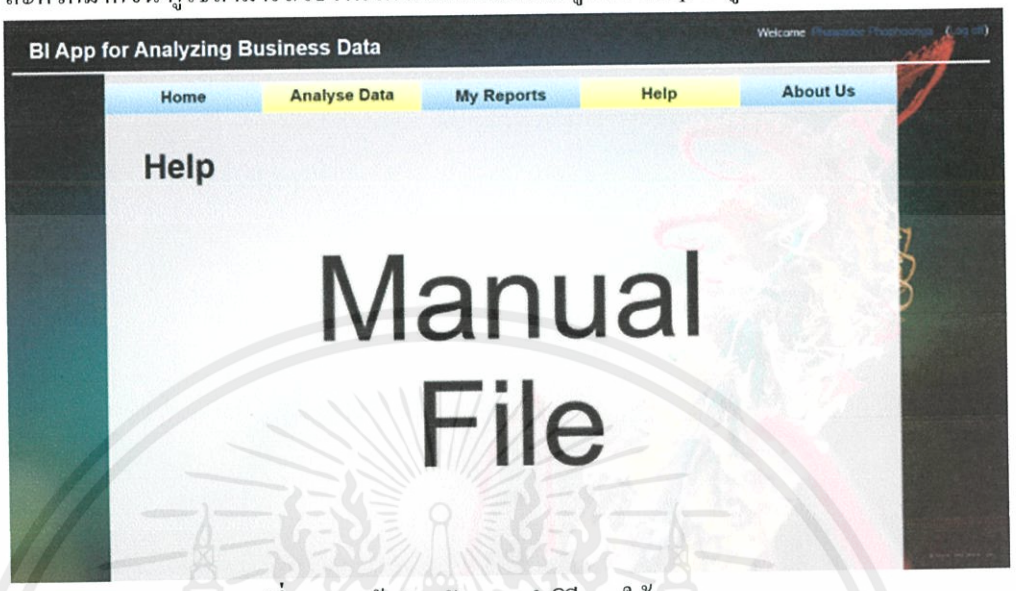
เป็นหน้าเว็บไซต์ที่ผู้ใช้สามารถเข้ามาดาวน์โหลดรายงานที่ผู้ใช้เคยสร้างไว้ หรือเข้ามาลบรายงานที่เคยสร้างไว้ได้ โดยกดที่แถบเมนู My Report ดังรูปที่ 3.18



รูปที่ 3.18 หน้าสำหรับจัดการข้อมูลรายงานของผู้ใช้

3.3.6 หน้าแนะนำวิธีการใช้งาน

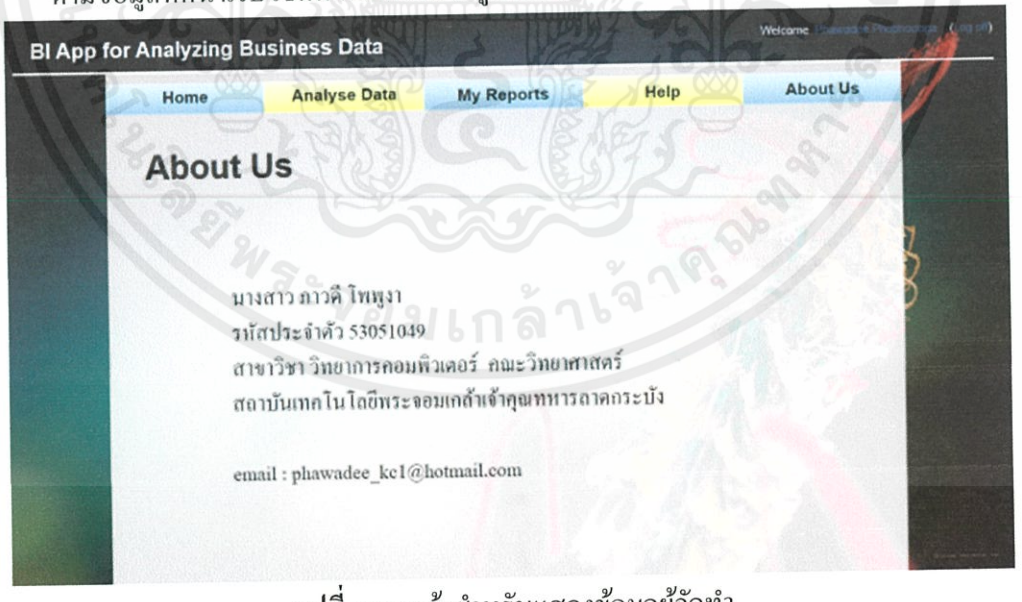
เป็นหน้าเว็บไซต์ที่มีไว้สำหรับแนะนำวิธีการใช้ เพื่อให้ผู้ใช้สามารถใช้งานเว็บไซต์ได้สะดวกมากขึ้น ผู้ใช้สามารถใช้งานได้โดยกดที่แถบเมนูที่ชื่อ Help ดังรูปที่ 3.19



รูปที่ 3.19 หน้าสำหรับแนะนำวิธีการใช้งาน

หน้าแสดงข้อมูลผู้จัดทำ

เป็นหน้าเว็บไซต์ที่จะแสดงข้อมูลของผู้จัดทำเว็บไซต์ ผู้ใช้สามารถติดต่อผู้จัดทำได้ตามข้อมูลที่หน้าเว็บไซต์นี้ ดังที่แสดงในรูปที่ 3.20



รูปที่ 3.20 หน้าสำหรับแสดงข้อมูลผู้จัดทำ

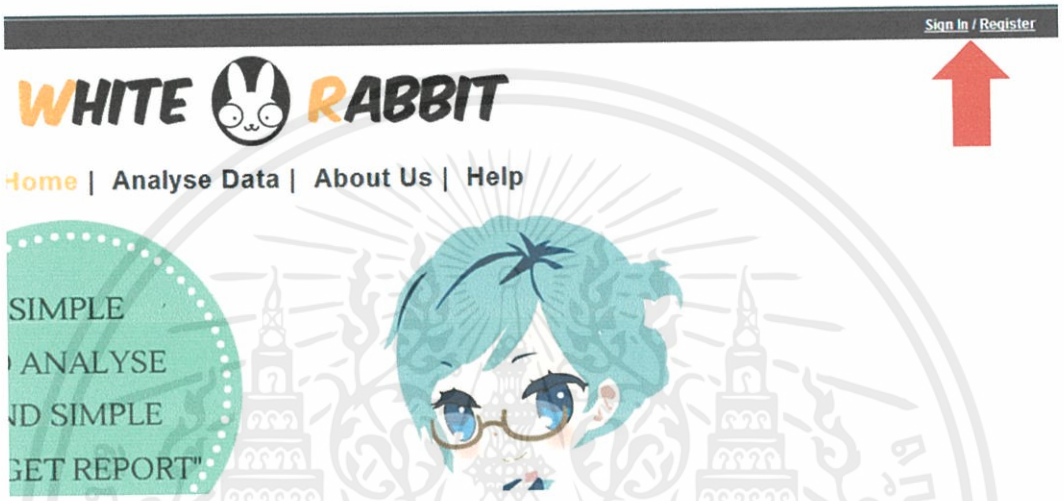
บทที่ 4

ผลการดำเนินงาน

4.1 ความสามารถของระบบ

4.1.1 ส่วนของการสมัครสมาชิก

การใช้งานนั้น ผู้ใช้จำเป็นต้องทำการสมัครสมาชิกก่อน ได้โดยการกดที่ปุ่ม Register

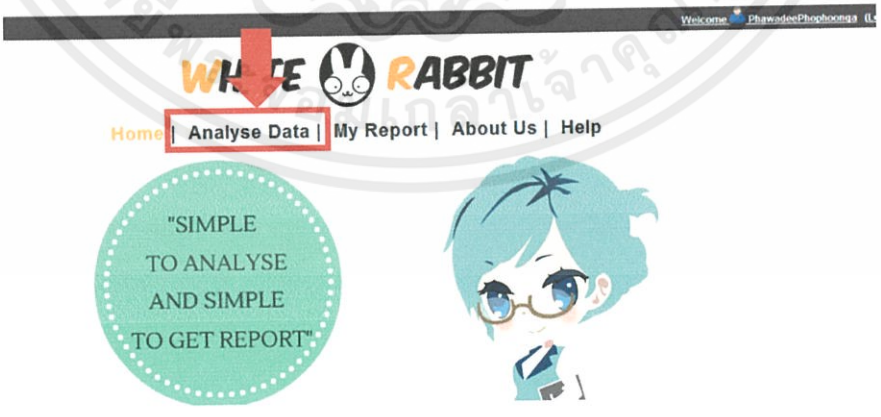


รูปที่ 4.1 แสดงส่วนลงทะเบียน

ระบบจะแสดงหน้าดังภาพ ผู้ใช้ต้องกรอกข้อมูลในการสมัครสมาชิก เมื่อกรอกข้อมูลเสร็จเรียบร้อย กด Submit จากนั้นก็จะสามารถนำ Username กับ Password ไป Log in เข้าระบบได้

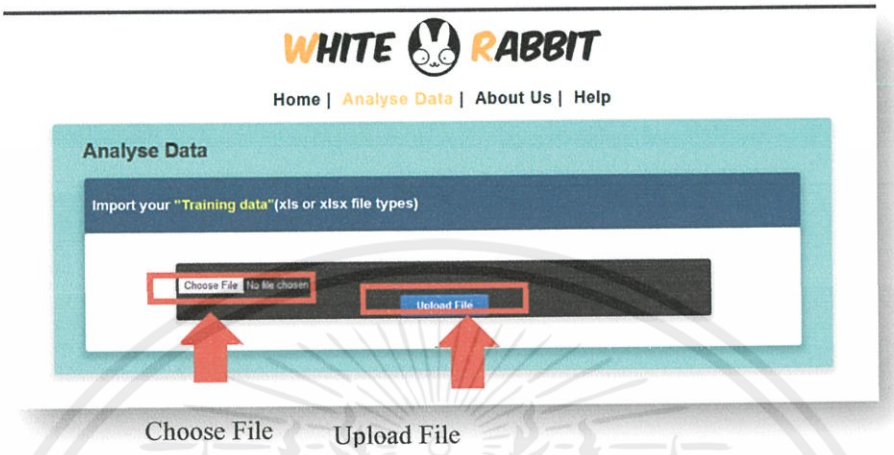
4.1.2 ส่วนของการวิเคราะห์ข้อมูลทางธุรกิจ

เมื่อผู้ใช้ต้องการวิเคราะห์ข้อมูลทางธุรกิจ กดที่เมนู Analyse Data ที่แถบเมนูในหน้าแรก



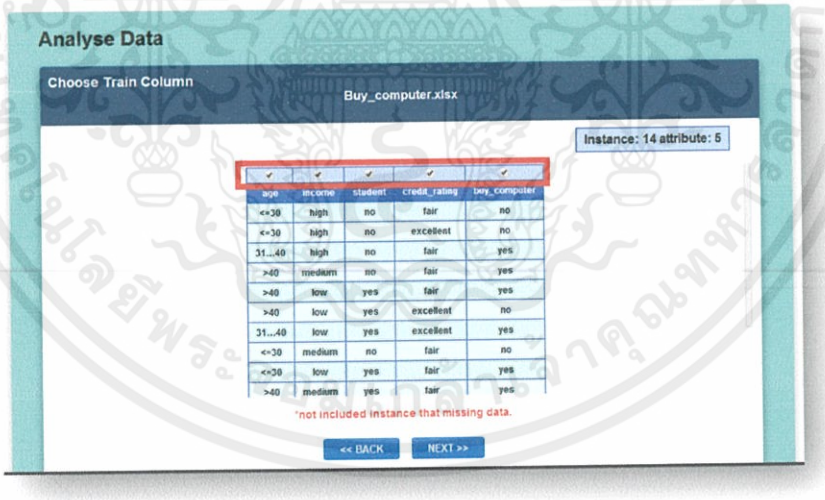
รูปที่ 4.2 แสดงเมนู Analyse Data

ระบบจะแสดงหน้า Analyse Data ดังภาพ ผู้ใช้สามารถ Import ข้อมูล Train ในรูปแบบไฟล์ xls หรือ xlsx(ลักษณะ file ที่จะนำเข้าระบบมีอธิบายในหัวข้อ “ลักษณะ file ที่นำเข้าระบบ”) จากนั้นกด Upload



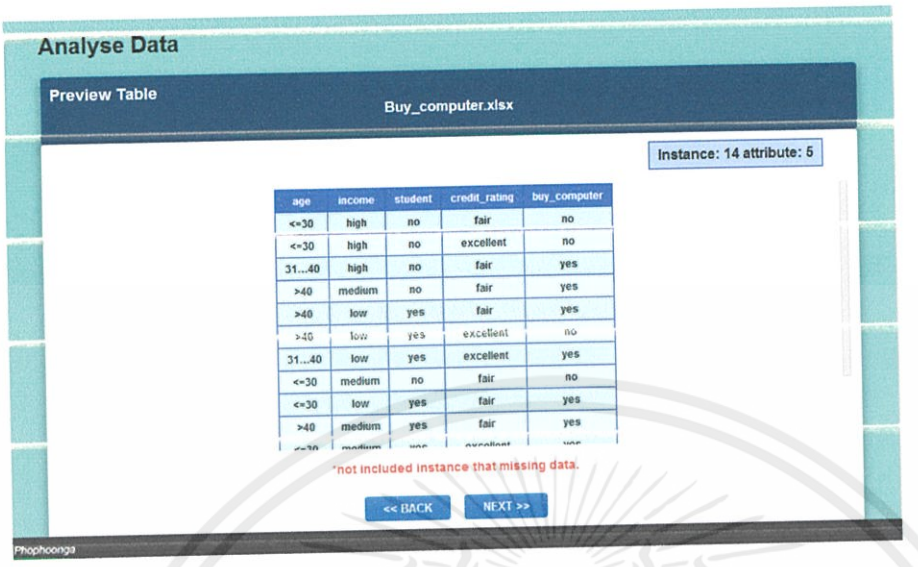
รูปที่ 4.3 แสดงส่วน Import Train Data

เมื่อ Upload แล้วระบบจะแสดงผลข้อมูลที่ผู้ใช้นำเข้าระบบ จากนั้นผู้ใช้สามารถกำหนด Attribute ที่จะนำเข้าไปสร้าง โมเดลในการจำแนกประเภทข้อมูลได้ (ในบริเวณกรอบสีแดง) จากนั้น กด NEXT



รูปที่ 4.4 แสดงส่วนเลือก Attribute ในการสร้างโมเดล

ระบบจะทำการ Preview ตารางข้อมูลที่ที่ผู้ใช้ได้เลือกไว้ดังภาพด้านล่าง จากนั้นกด NEXT

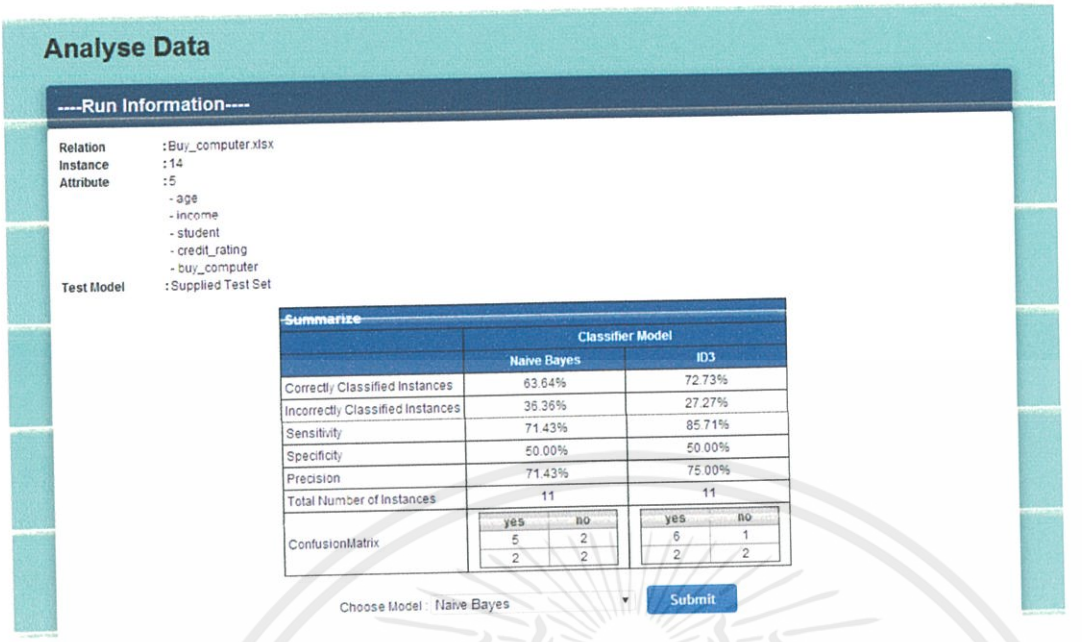


รูปที่ 4.5 แสดงส่วน Preview Train Data

ระบบจะแสดงหน้าให้ผู้ใช้เลือกอัลกอริทึม ในการจำแนกประเภทข้อมูล ผู้ใช้สามารถเลือกอัลกอริทึม อย่างเดียว หรือทั้งสองอย่าง เพื่อเปรียบเทียบค่าความแม่นยำของโมเดลได้ จากนั้นเลือก Option ในการ Test แล้วกด Start Classification

รูปที่ 4.6 แสดงส่วนเลือกอัลกอริทึม ในการจำแนกประเภทข้อมูล

จากนั้น ระบบจะทำการหาความแม่นยำของ อัลกอริทึม ที่ผู้ใช้ได้เลือกไว้ และแสดงผลตามภาพด้านล่าง ผู้ใช้สามารถเลือก Model ที่เหมาะสมกับข้อมูลผู้ใช้ เพื่อที่จะนำไปทำนายต่อไป จากนั้น กด Submit

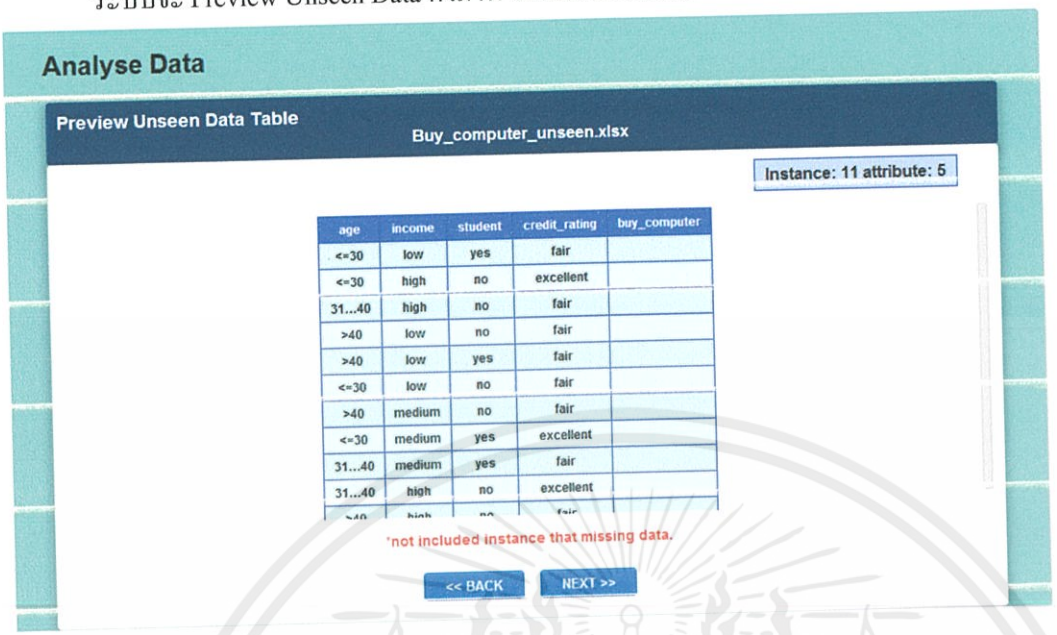


รูปที่ 4.7 แสดงค่าความแม่นยำของโมเดล

และระบบจะแสดงหน้าที่ให้ผู้ใช้ Import Unseen Data ที่ต้องการทำนาย เข้ามาในระบบ

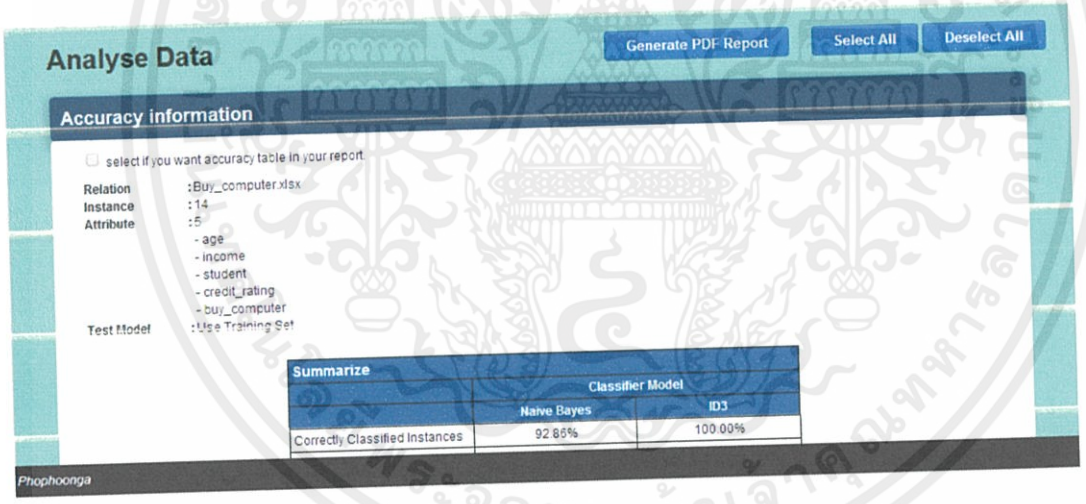
รูปที่ 4.8 แสดงส่วน Import Unseen Data

ระบบจะ Preview Unseen Data ดังภาพ จากนั้นกด NEXT



รูปที่ 4.9 แสดงส่วน Preview Unseen Data

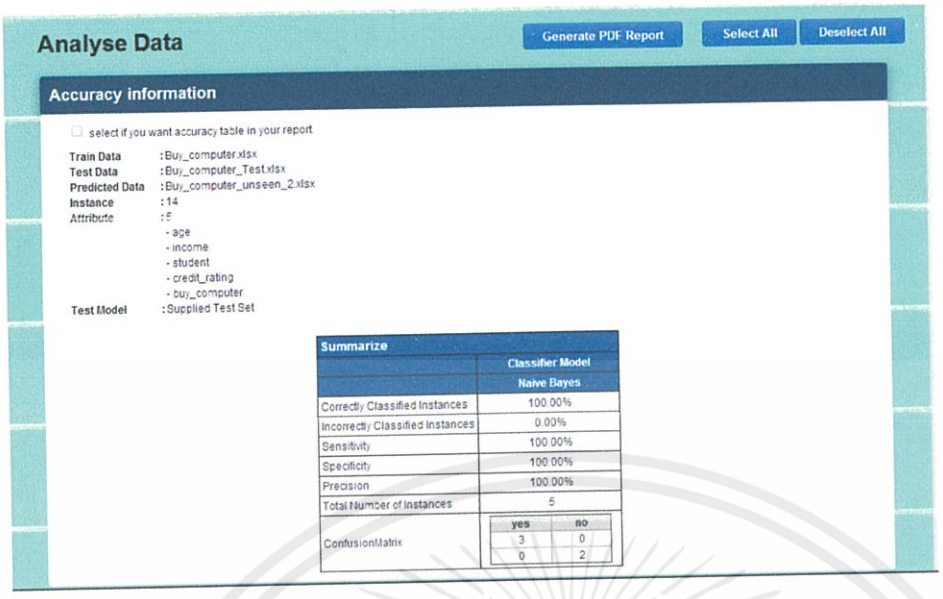
เมื่อกด NEXT แล้วระบบจะทำการนำข้อมูล ไปทำนายหาคำตอบที่ผู้ใช้ต้องการ ดังภาพ



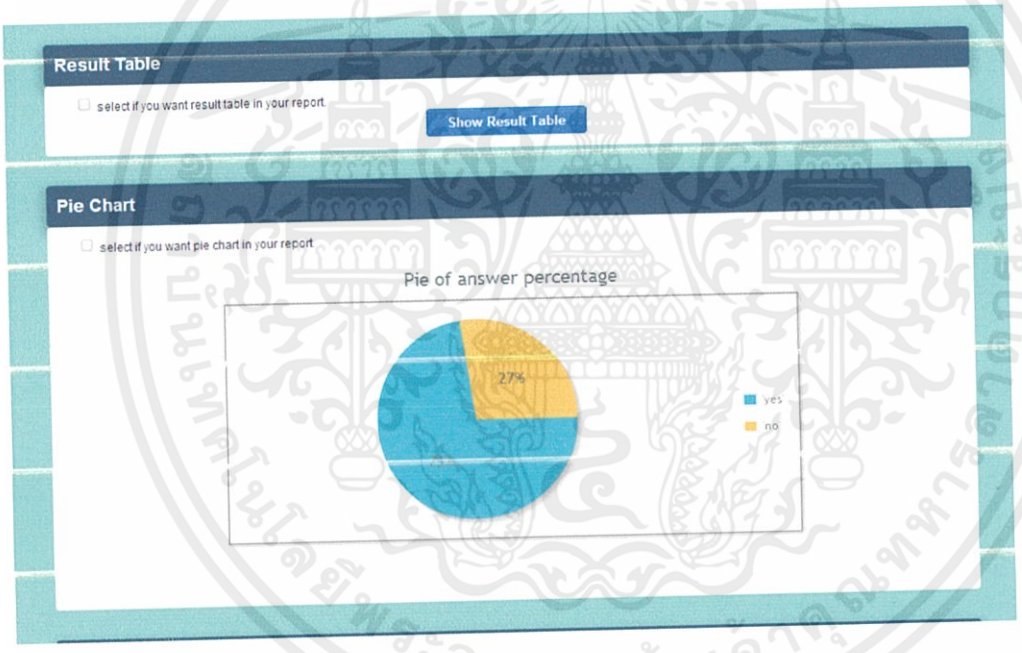
รูปที่ 4.10 แสดงผลคำตอบ

4.1.3 ส่วนรายงาน

เมื่อผู้ใช้ทำนายผลข้อมูลแล้วระบบจะแสดงหน้าเว็บ ดังภาพด้านล่าง ที่จะมีทั้ง ตารางข้อมูล ค่าความแม่นยำของโมเดลที่ผู้ใช้ได้ใช้ในการทำนายผลข้อมูล, ตารางข้อมูลที่ผ่านการทำนาย เรียบร้อยแล้ว, แผนภูมิวงกลม(แสดงเปอร์เซ็นของข้อมูลที่ตรงกับคำตอบแต่ละคำตอบ), และ แผนภูมิแท่งของแต่ละแอททริบิว ตามลำดับ



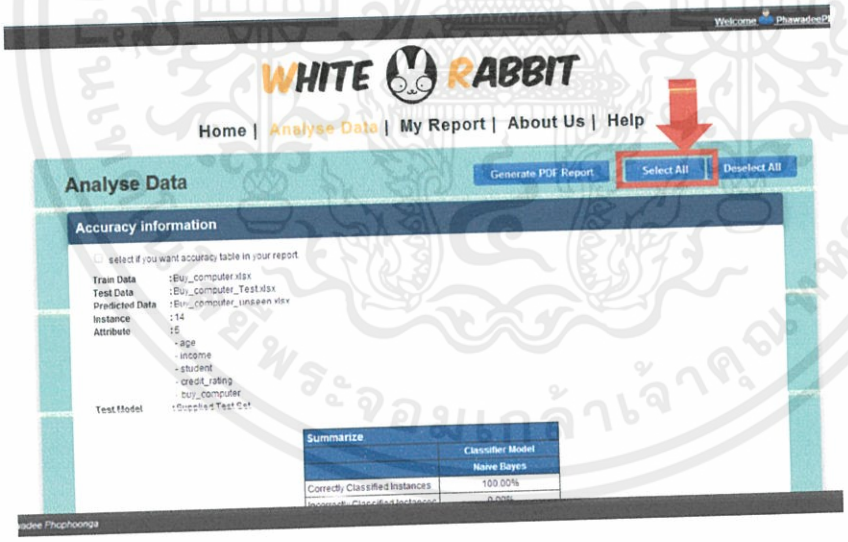
รูปที่ 4.11 แสดงส่วนการสร้างรายงาน



รูปที่ 4.12 แสดงส่วน Pie Chart

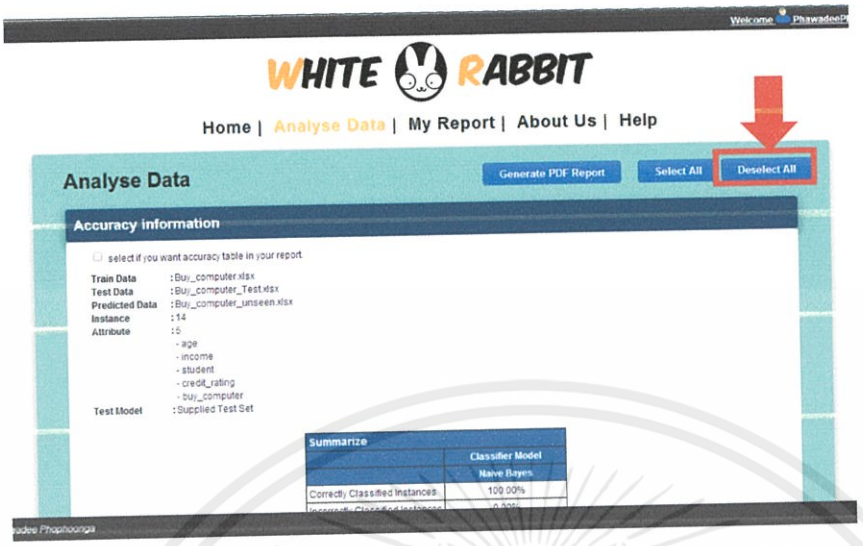
รูปที่ 4.13 แสดงส่วน Bar Chart

ผู้ใช้สามารถสร้างรายงานในหน้านี้ได้ โดยการเลือกที่ Check box หรือสามารถเลือกที่จะสร้างรายงานด้วยข้อมูลทั้งหมด โดยการกดที่ Select All (ในบริเวณกรอบสีแดง)



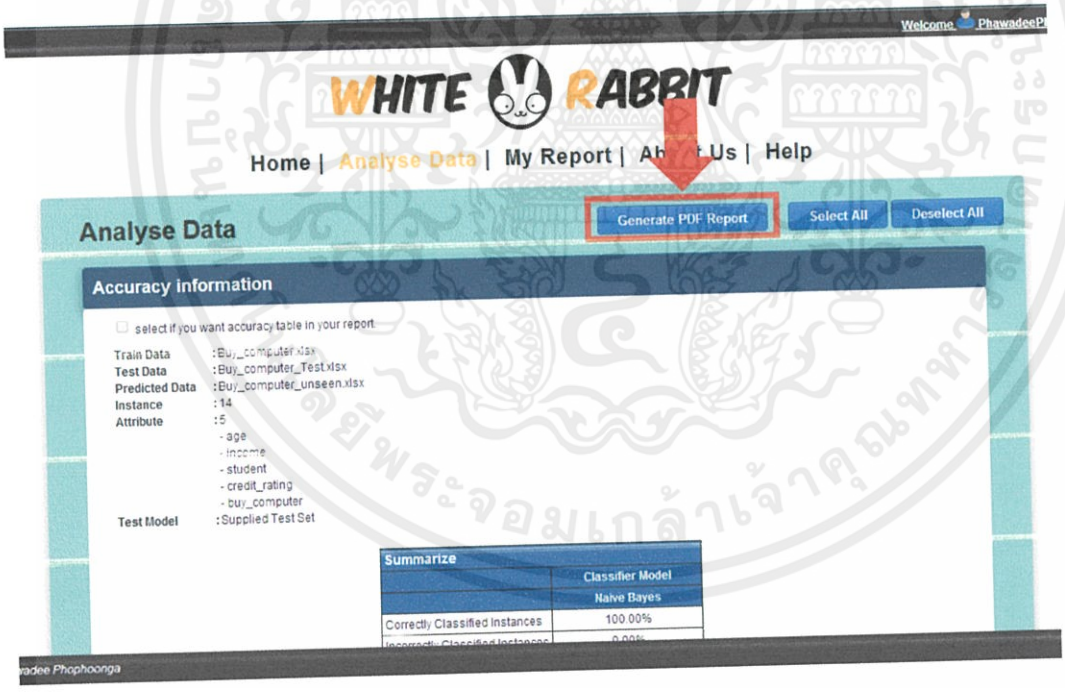
รูปที่ 4.14 แสดงปุ่มSelect All

หรือยกเลิกการเลือกทั้งหมดได้ที่ปุ่ม Deselect All



รูปที่ 4.15 แสดงปุ่มDeselect All

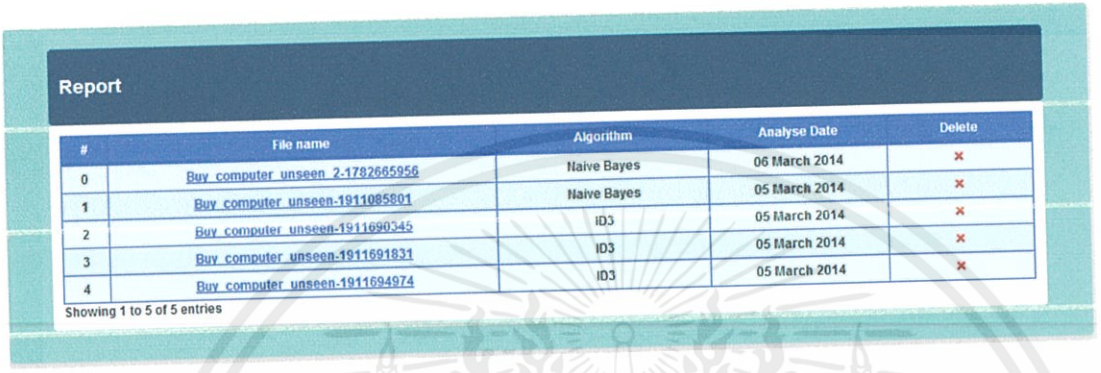
หลังจากเลือกข้อมูลที่ต้องการสร้างรายงานได้แล้ว ผู้ใช้สามารถดาวน์โหลดรายงานออกไปใช้งานได้ โดยการกดที่ปุ่ม Generate PDF Report ดังภาพด้านล่าง



รูปที่ 4.16 แสดงส่วนGenerate PDF Report

ถ้าหากผู้ใช้ต้องการดาวน์โหลดรายงานที่เคยสร้างไว้ สามารถเข้าไปดาวน์โหลดได้โดยการกดที่ My Report ในหน้า Home

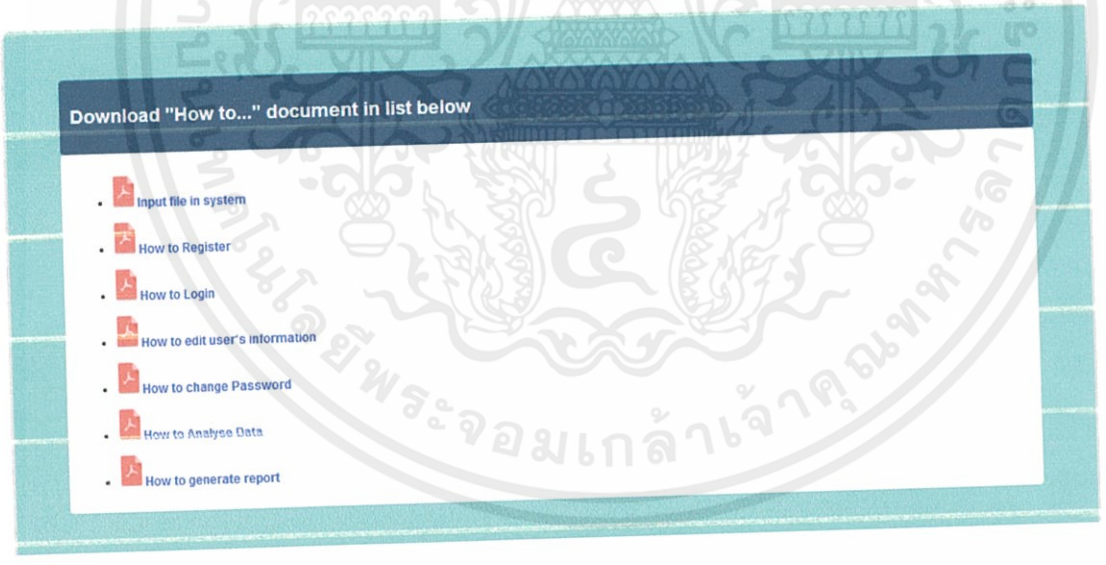
จากนั้นระบบจะแสดงหน้า Report ของผู้ใช้ หน้านี้จะเก็บไฟล์ของรายงานที่ผู้ใช้เคยสร้าง ที่จะประกอบไปด้วยข้อมูลชื่อไฟล์รายงาน, Algorithm ที่ใช้ในการทำนายผลข้อมูล, และวันที่สร้างรายงาน ผู้ใช้สามารถ ดาวน์โหลดรายงาน หรือ ลบรายงานที่ไม่ต้องการได้



รูปที่ 4.17 ส่วนแสดงรายงานของผู้ใช้

4.1.4 ส่วนช่วยเหลือ

เป็นส่วนแสดงวิธีใช้งานระบบทั้งหมด ผู้ใช้สามารถเข้าไปดูรายละเอียดได้ที่หน้า Help



รูปที่ 4.18 แสดงส่วนช่วยเหลือ

4.2 แหล่งที่มาและรายละเอียดชุดข้อมูล

ข้อมูลที่ใช้ในการวิจัยเพื่อใช้ทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำมาใช้ในการจำแนกประเภทข้อมูลจะคัดเลือกข้อมูลจากแหล่งข้อมูลของมหาลัยแห่งรัฐแคลิฟอร์เนีย เมืองเออร์ไวย์ ประเทศสหรัฐอเมริกา (http://archive.ics.uci.edu/ml/) ซึ่งได้รวบรวมข้อมูลสำหรับใช้เป็นเกณฑ์สำหรับทดสอบประสิทธิภาพการทำงานของอัลกอริทึมต่างๆ ของการทำเหมืองข้อมูล ซึ่งมีรายละเอียดข้อมูลดังนี้

ตารางที่ 4.1 รายละเอียดชุดข้อมูลโดยสรุปที่ใช้ในการทดสอบประสิทธิภาพ

ชื่อชุดข้อมูล	จำนวนตัวอย่าง	คุณลักษณะ	จำนวนกลุ่ม
1.	German Credit fraud data	1000	21
2.	Adult data	1000	14

ข้อมูลแต่ละชุดมีรายละเอียดและการจำแนกกลุ่มของข้อมูลดังต่อไปนี้

1. German Credit fraud data เป็นข้อมูลที่ประกอบด้วยการตรวจสอบการโกงบัตรเครดิต มีจำนวนข้อมูล 1000 ตัวอย่าง จำแนกออกเป็น 2 กลุ่ม คือ
   - Good (จำนวน 300 ตัวอย่าง)
   - Bad (จำนวน 700 ตัวอย่าง)
2. Adult data เป็นข้อมูลที่ใช้ทำนายเกี่ยวกับรายได้ของประชากร ที่สามารถทำเงินได้ >50K ต่อปี มีจำนวนข้อมูล 1000 ตัวอย่าง จำแนกเป็น 2 กลุ่ม คือ
   - >50K (จำนวน 756 ตัวอย่าง)
   - <=50K (จำนวน 244 ตัวอย่าง)

4.3 ผลการทดลอง

การวิเคราะห์เปรียบเทียบประสิทธิภาพของการจำแนกประเภทนั้น มีการเปรียบเทียบขั้นตอนวิธี 2 แบบด้วยกัน คือ Naïve Bayes และ ID3 โดยประสิทธิภาพของการจำแนกประเภทนั้นจะนำไปเปรียบเทียบกับโปรแกรม WEKA ผลดังตารางที่ 4.2

ตารางที่ 4.2 ผลเปรียบเทียบค่าความแม่นยำ

ชื่อชุดข้อมูล	WEKA		BI Application	
	Naïve Bayes	ID3	Naïve Bayes	ID3
German Credit fraud data	77.20%	100%	77.40%	100.%
Adult data	82.90%	97.60%	83.10%	97.60%

เมื่อเปรียบเทียบความแม่นยำในการวิเคราะห์ข้อมูลทางธุรกิจ ระหว่าง โปรแกรม WEKA ที่เป็นที่ยอมรับในด้านการจำแนกข้อมูล กับ เว็บแอพพลิเคชั่นที่พัฒนาขึ้นมาเอง จะพบว่า ค่าความแม่นยำในการวิเคราะห์ข้อมูลของเว็บแอพพลิเคชั่นมีค่าที่มากกว่า ซึ่งสามารถสรุปได้ว่า การวิเคราะห์ข้อมูลด้วยเว็บแอพพลิเคชั่น มีความน่าเชื่อถือ

บทที่ 5

สรุปผลวิจัยและข้อเสนอแนะ

5.1 สรุปผลงานวิจัย

งานวิจัยนี้มีจุดมุ่งหมายเพื่อที่จะพัฒนาเว็บแอพพลิเคชั่นสำหรับวิเคราะห์ข้อมูลทางธุรกิจ เพื่อให้สามารถทำนายผลข้อมูลและสร้างรายงาน ได้อย่างมีประสิทธิภาพ โดยขั้นตอนการวิจัยได้เริ่มจากการศึกษาค้นคว้าเกี่ยวกับขั้นตอนในการจำแนกประเภทข้อมูลต่างๆ ที่น่าสนใจ ได้แก่ Naïve Bayes และ ID3 เพื่อนำมาประยุกต์ใช้ในเว็บแอพพลิเคชั่นเพื่อใช้ในการวิเคราะห์และทำนายผลข้อมูลทางธุรกิจ โดยที่ได้พัฒนาให้เว็บแอพพลิเคชั่นสามารถเปรียบเทียบค่าความถูกต้องของโมเดลที่ใช้ในการจำแนกประเภทข้อมูลให้ผู้ใช้ ก่อนทำการทำนายผลข้อมูล ทำให้ผู้ใช้สามารถเลือกโมเดลที่มีความแม่นยำ และเหมาะสมกับข้อมูลของผู้ใช้มากที่สุด ทั้งนี้ผู้พัฒนาได้ออกแบบแอพพลิเคชั่นให้ผู้ใช้สามารถเลือกสร้างรายงานตามข้อมูลที่ผู้ใช้ต้องการไม่ว่าจะเป็น ข้อมูลเกี่ยวกับความแม่นยำของโมเดล, ตารางข้อมูลที่ผ่านการทำนายผล, และผลการทำนายในรูปแบบของกราฟ

5.2 ปัญหาที่พบและการพัฒนาโครงงาน

5.2.1 ปัญหาที่พบ

เมื่อนำข้อมูลจำนวนมากเข้าสู่ระบบ ถ้าหากภายในข้อมูล มี Missing Data จะไม่สามารถจำแนกประเภทและทำนายผลข้อมูลได้ เนื่องจากยังไม่มีส่วนที่ใช้ในการปรับแต่งข้อมูล เมื่อผู้ใช้ต้องการวิเคราะห์ข้อมูลต้องทำการปรับแต่งข้อมูลเองก่อนนำเข้าระบบ

5.2.2 ข้อเสนอแนะ

การพัฒนาแอพพลิเคชันธุรกิจอัจฉริยะ สำหรับการวิเคราะห์ข้อมูลธุรกิจโดยใช้เทคนิคการทำเหมืองข้อมูล สามารถปรับปรุงดังนี้

1) ควรพัฒนาให้มีส่วนที่ใช้สำหรับปรับแต่งข้อมูล ก่อนนำไปจำแนกประเภทข้อมูลได้
2) ผู้พัฒนาควรพัฒนาให้มีอัลกอรึทึมในการจำแนกประเภทข้อมูลมากกว่านี้ เพื่อให้ผู้ใช้สามารถเลือกอัลกอรึทึมที่สามารถให้ความถูกต้องแม่นยำในการจำแนกประเภทข้อมูล

เอกสารอ้างอิง

ชินพัฒน์ แก้วชินพร และณัฐกานต์ วงศ์สุโชโต. "การจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ และการจัดกลุ่ม." วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิตสาขาวิชาวิทยาการคอมพิวเตอร์ คณะ-วิทยาศาสตร์, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2553.

จักรกฤษณ์ สีน้ำเงิน, ณัฏฐ์ชานนท์ สุขขดำรงรักษ์, และต้า เกียรติไกรวัลศิริ. "การวิเคราะห์การสกัด-คุณลักษณะสำหรับระบบรู้จำตัวอักษรภาษาไทย." วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2554.

ธนิศร์ ชื่นอุระจิตร, วสุ คูดิษฐาเลิศ, และศิขริน จันพละ. "โปรแกรมช่วยสร้างการติดต่อ API ของเฟซบุค." วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิตสาขาวิชาวิทยาการคอมพิวเตอร์ คณะ-วิทยาศาสตร์, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2554.

[Online]. Available : www.gotoknow.org/posts/52660/. 2007.

[Online]. Available : http://searchnetworking.techtarget.com/definition/client-server. 2008.

[Online]. Available : www.jba.tbs.tu.ac.th/files/Jba137/Column/JBA137SrisomrukC.pdf. 2013.

[Online]. Available : www.jarticles.com/tutorials/servlet/intro_servlet.html (accessed August 15, 2013)

[Online]. Available : www.itmelody.com/tu/introjsp.htm (accessed August 25, 2013)

[Online]. Available : www.jarticles.com/jsp/ (accessed August 25, 2013)

[Online]. Available : http://stackoverflow.com/ (accessed August 25, 2013)

[Online]. Available : http://www.deepakgaikwad.net/wp-content/uploads/2009/04/jsplifecycle.jpg (accessed October 13, 2013)

[Online]. Available : http://www.jqplot.com/ (accessed October 13, 2013)

[Online]. Available : http://weka.8497.n7.nabble.com/file/n23121/credit_fruad.arff (accessed October 13, 2013)

[Online]. Available : http://archive.ics.uci.edu/ml/index.html (accessed January 12, 2013)

ภาคผนวก ก

คู่มือการใช้งานระบบ

ก.1 ขั้นตอนการใช้งานเว็บแอพพลิเคชัน

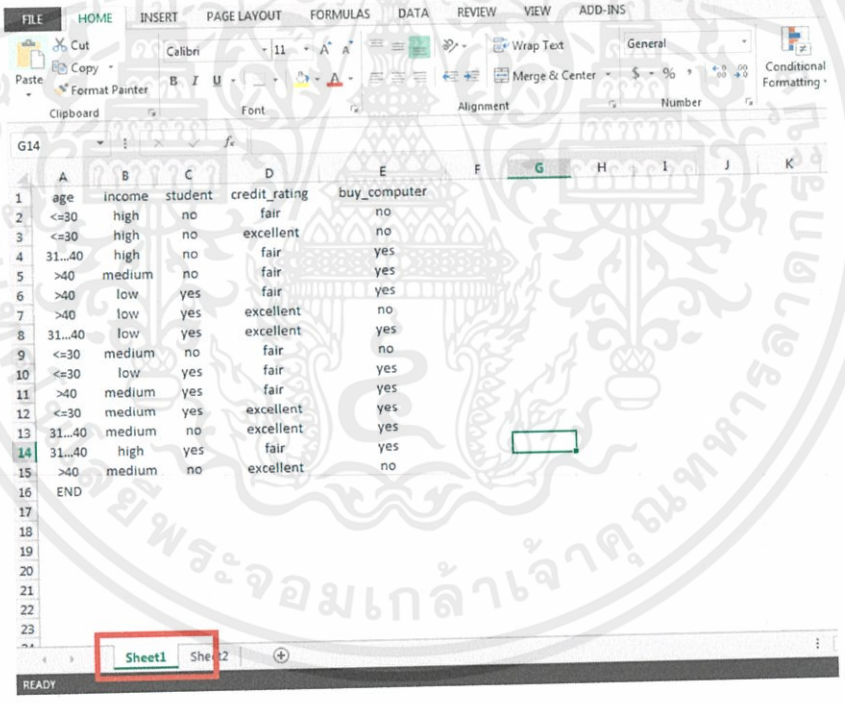
เว็บแอพพลิเคชัน ได้แบ่งการทำงานออกเป็น 2 ส่วนหลัก ดังนี้

ก.1.1 ส่วนของผู้ใช้ทั่วไป

1) ลักษณะ File ที่นำเข้าระบบ

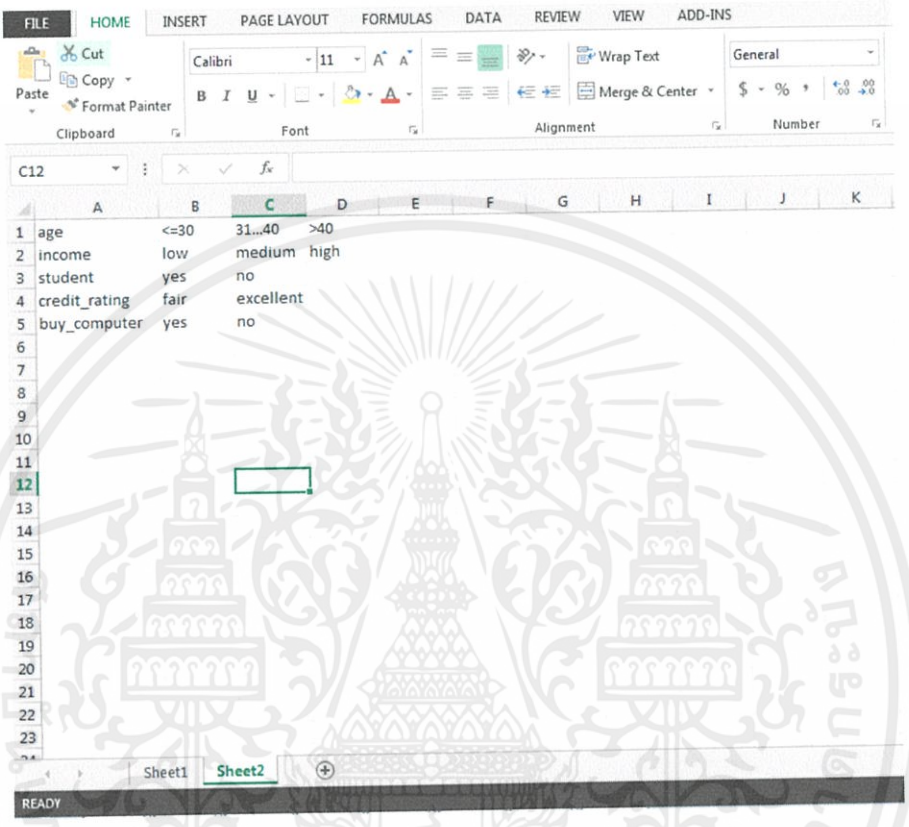
File ที่จะ Import สู่ระบบต้องอยู่ในรูปแบบ xls หรือ xlsx ซึ่งไฟล์ที่จะนำเข้าประมวลผลในระบบมีทั้งหมด 3 File คือไฟล์ Train Data, Test Data และ Unseen Data ซึ่งแต่ละไฟล์จะต้องมีลักษณะ ดังนี้

Train Data File คือไฟล์ข้อมูลที่มีไว้สำหรับสร้างโมเดลก่อนนำไป ทดสอบและทำนายผลข้อมูล มีทั้งหมด 2 Sheet โดยที่ Sheet ที่ 1 จะเป็น Sheet ข้อมูล Train ซึ่งแถวที่ 1 โดยเริ่มจากคอลัมน์ A จะเป็นชื่อ Attribute จากนั้นแถวที่ 2 เป็นต้นไปจะเป็นข้อมูลภายในแต่ละ Attribute เมื่อใส่ข้อมูลครบทั้งหมด ต้องใส่ “END” คอลัมน์ที่ 1 ดังภาพ



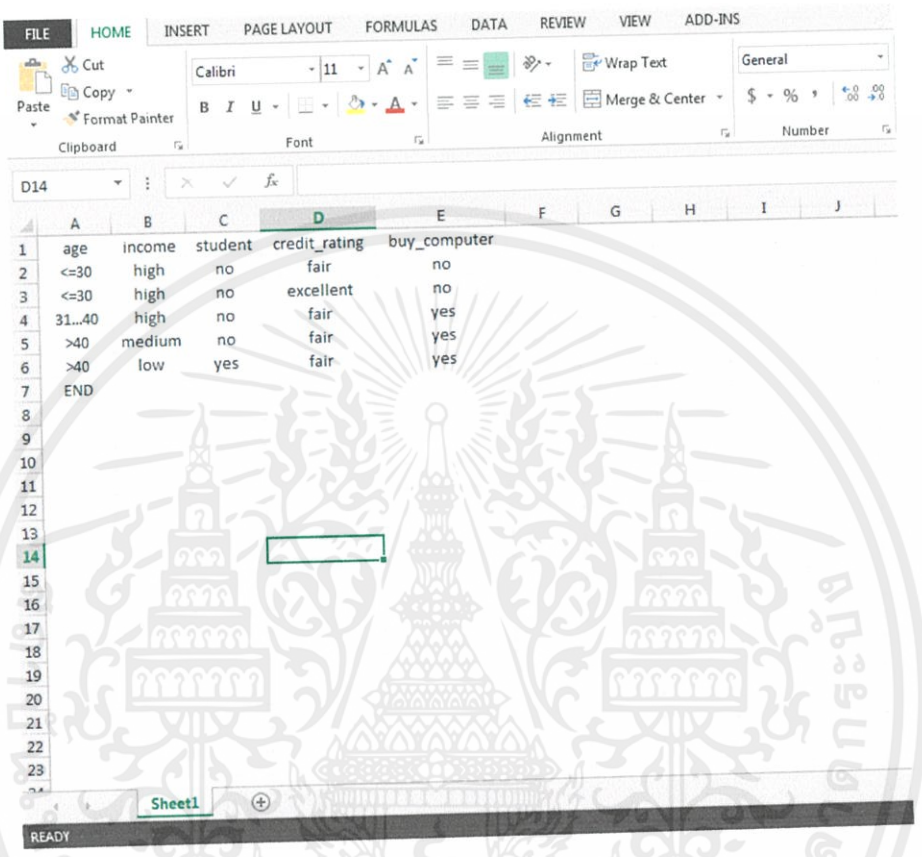
รูปที่ ก.1 แสดงลักษณะไฟล์ Train Data Sheet 1

ใน Sheet ที่ 2 จะแสดงโครงสร้างของ Attribute โดยที่คอลัมน์ที่ 1 จะเป็นชื่อของ Attribute จากรูปที่ ก.2 จะมีทั้งหมด 5 Attribute คือ age, income, student, credit_rating, buy_computer คอลัมน์ถัดจาก ชื่อ Attribute จะเป็นค่าภายในแต่ละ Attribute จากตัวอย่าง Attribute age มีค่าภายใน Attribute ทั้งหมด 3 ค่า คือ <=30, 31…40, >40 เป็นต้น



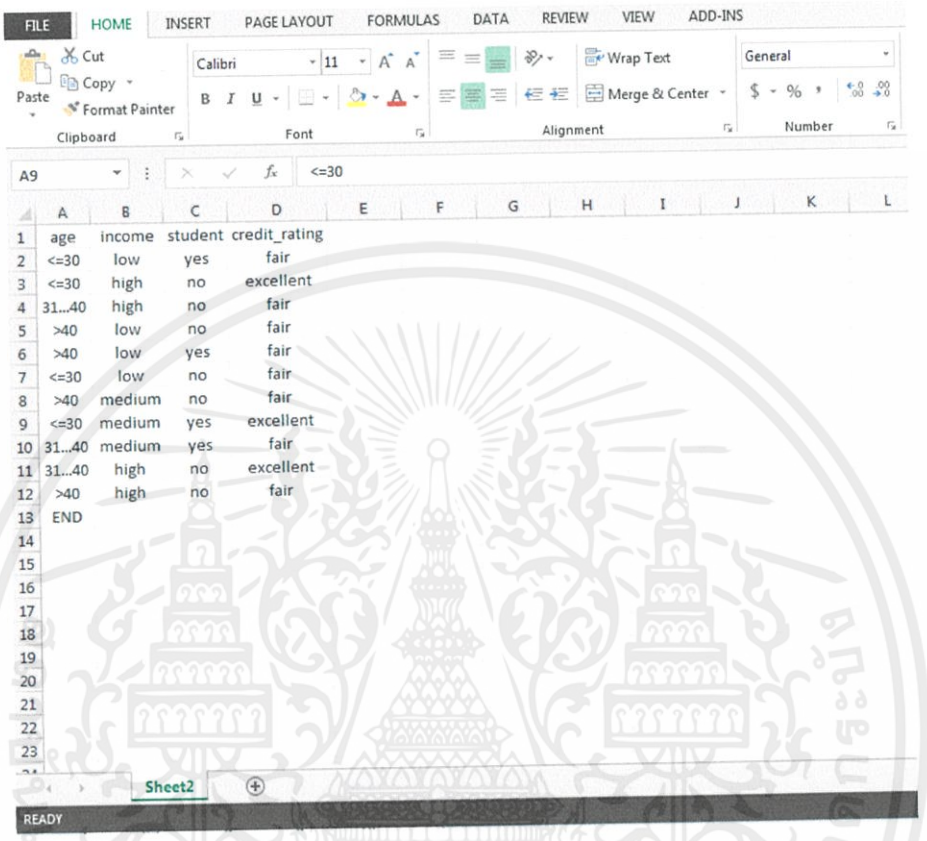
รูปที่ ก.2 แสดงลักษณะไฟล์ Train Data Sheet 2

Test Data File เป็นไฟล์ข้อมูลที่มีไว้สำหรับนำไปตรวจค่าความถูกต้องของ Model ที่สร้างขึ้นจาก ข้อมูลเรียนรู้ (Training Data) จะมี Sheet เดียว โดยแถวที่ 1 เริ่มจากคอลัมน์ A จะเป็นชื่อ Attribute จากนั้นแถวที่ 2 เป็นต้นไปจะเป็นข้อมูลภายในแต่ละ Attribute เมื่อใส่ข้อมูลครบทั้งหมดต้องใส่ “END” ในแถวถัดมา ดังรูปที่ ก.3



รูปที่ ก.3 แสดงลักษณะไฟล์ Test Data

Unseen Data คือข้อมูลที่ผู้ใช้ต้องการทำนายผล ซึ่งจะมีลักษณะดังภาพ ด้านล่างซึ่งจะมีลักษณะเหมือนกับ Test Data File แต่จะไม่มี Column ที่เป็นคำตอบ

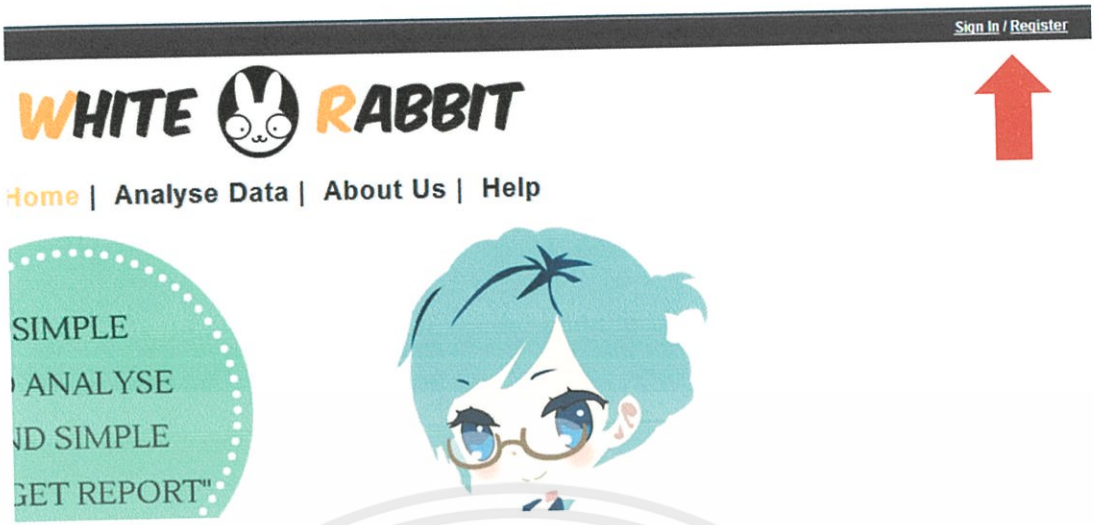


	A	B	C	D
1	age	income	student	credit_rating
2	<=30	low	yes	fair
3	<=30	high	no	excellent
4	31...40	high	no	fair
5	>40	low	no	fair
6	>40	low	yes	fair
7	<=30	low	no	fair
8	>40	medium	no	fair
9	<=30	medium	yes	excellent
10	31...40	medium	yes	fair
11	31...40	high	no	excellent
12	>40	high	no	fair
13	END			

รูปที่ ก.4 แสดงลักษณะไฟล์ Unseen Data

2) ส่วนของการสมัครสมาชิก

การใช้งานนั้น ผู้ใช้จำเป็นต้องทำการสมัครสมาชิกก่อน ได้โดยการกดที่ปุ่ม Register



รูปที่ ก.5 แสดงส่วนลงทะเบียน

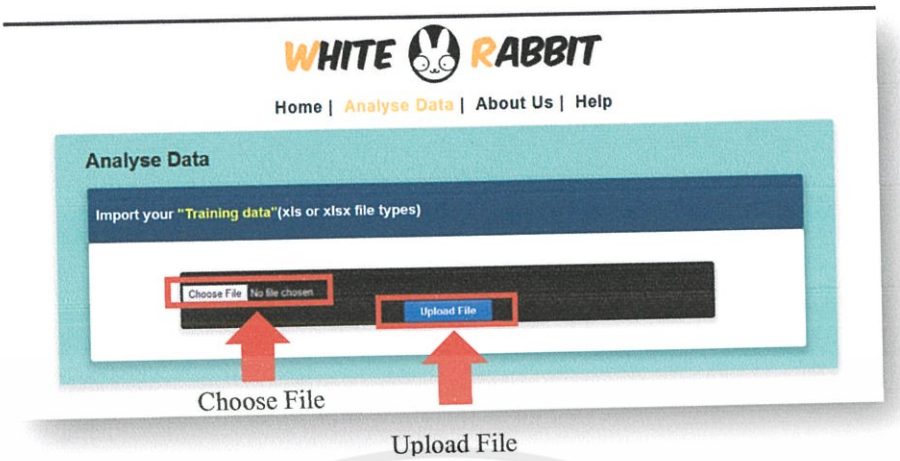
ระบบจะแสดงหน้าดังภาพ ผู้ใช้ต้องกรอกข้อมูลในการสมัครสมาชิก เมื่อกรอกข้อมูลเสร็จเรียบร้อย กด Submit จากนั้นก็จะสามารถนำ Username กับ Password ไป Log in เข้าระบบได้

3) ส่วนของการวิเคราะห์ข้อมูลทางธุรกิจ

เมื่อผู้ใช้ต้องการวิเคราะห์ข้อมูลทางธุรกิจ กดที่เมนู Analyse Data ที่แถบเมนูในหน้าแรก

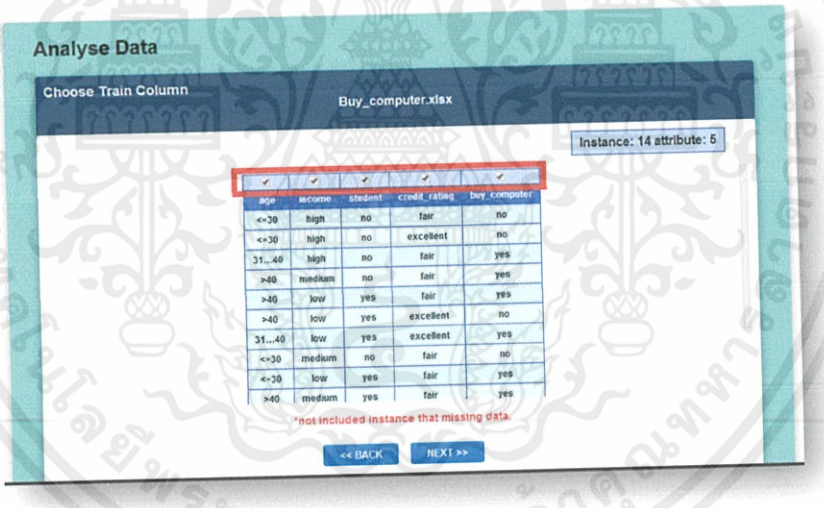
รูปที่ ก.6 แสดงเมนู Analyse Data

ระบบจะแสดงหน้า Analyse Data ดังภาพ ผู้ใช้สามารถ Import ข้อมูล Train ในรูปแบบไฟล์ xls หรือ xlsx(ลักษณะ file ที่จะนำเข้าระบบมีอธิบายในหัวข้อ “ลักษณะ file ที่นำเข้าระบบ”) จากนั้นกด Upload



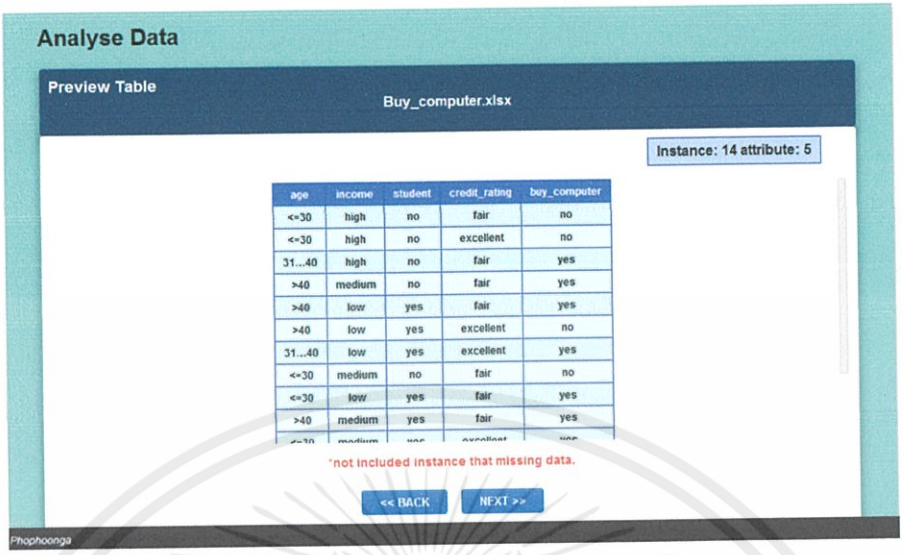
รูปที่ ก.7 แสดงส่วน Import Train Data

เมื่อ Upload แล้วระบบจะแสดงผลข้อมูลที่ผู้ใช้นำเข้าระบบ จากนั้นผู้ใช้สามารถกำหนด Attribute ที่จะนำเข้าไปสร้าง โมเดลในการจำแนกประเภทข้อมูลได้ (ในบริเวณกรอบสีแดง) จากนั้น กด NEXT



รูปที่ ก.8 แสดงส่วนเลือก Attribute ในการสร้างโมเดล

ระบบจะทำการ Preview ตารางข้อมูลที่ที่ผู้ใช้ได้เลือกไว้ดังภาพด้านล่าง จากนั้นกด NEXT

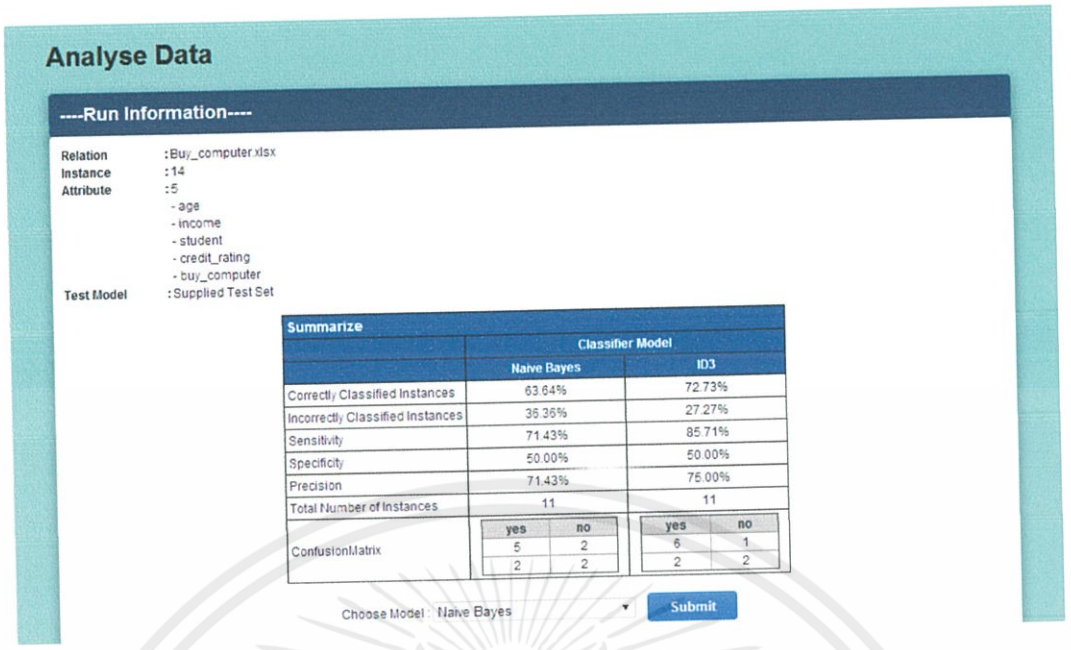


รูปที่ ก.9 แสดงส่วน Preview Train Data

ระบบจะแสดงหน้าให้ผู้ใช้เลือกอัลกอริทึม ในการจำแนกประเภทข้อมูล ผู้ใช้สามารถเลือกอัลกอริทึม อย่างเดียว หรือทั้งสองอย่าง เพื่อเปรียบเทียบค่าความแม่นยำของโมเดลได้ จากนั้นเลือก Option ในการ Test แล้วกด Start Classification

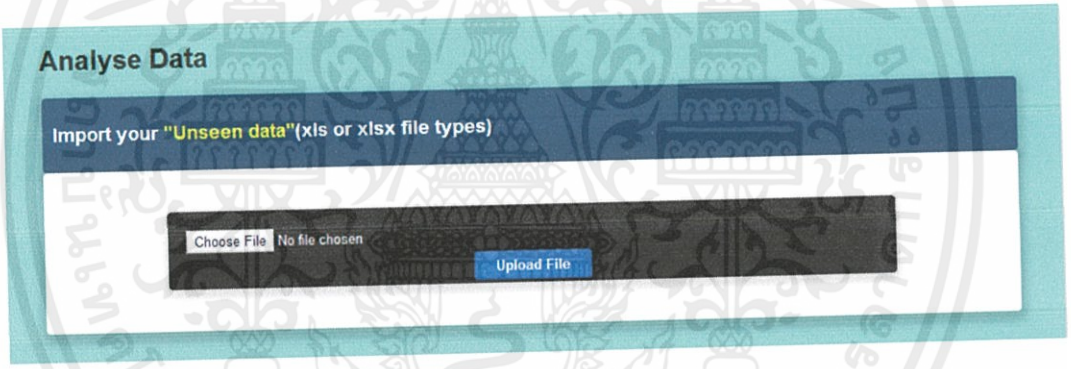
รูปที่ ก.10 แสดงส่วนเลือกอัลกอริทึม ในการจำแนกประเภทข้อมูล

จากนั้น ระบบจะทำการหาความแม่นยำของ อัลกอริทึม ที่ผู้ใช้ได้เลือกไว้ และแสดงผลตามภาพด้านล่าง ผู้ใช้สามารถเลือก Model ที่เหมาะสมกับข้อมูลผู้ใช้ เพื่อที่จะนำไปทำนายต่อไป จากนั้น กด Submit



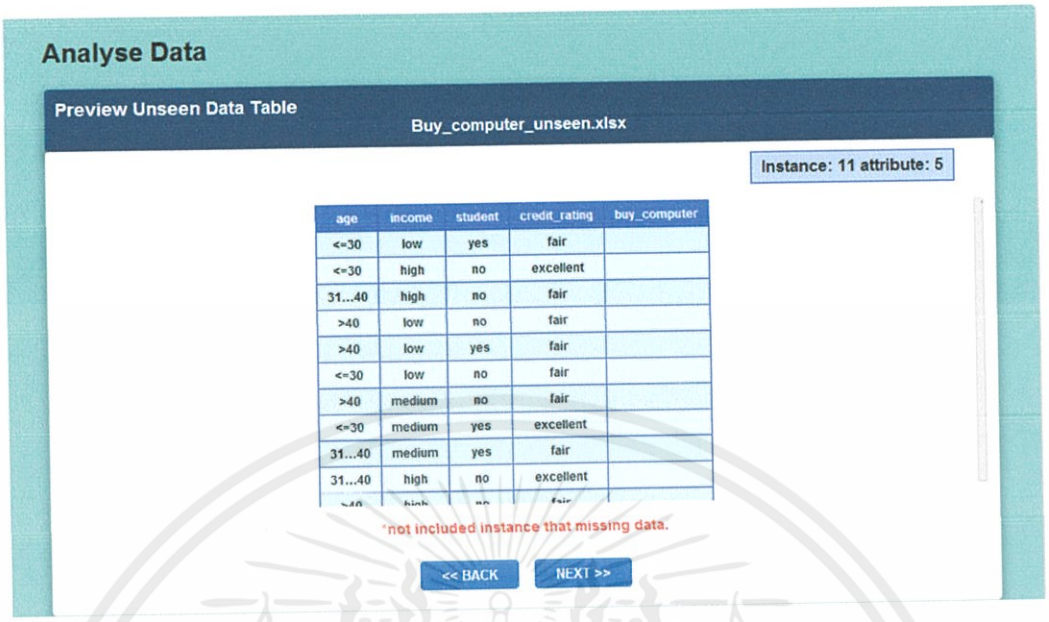
รูปที่ ก.11 แสดงค่าความแม่นยำของโมเดล

และระบบจะแสดงหน้าที่ให้ผู้ใช้ Import Unseen Data ที่ต้องการทำนาย เข้ามาในระบบ



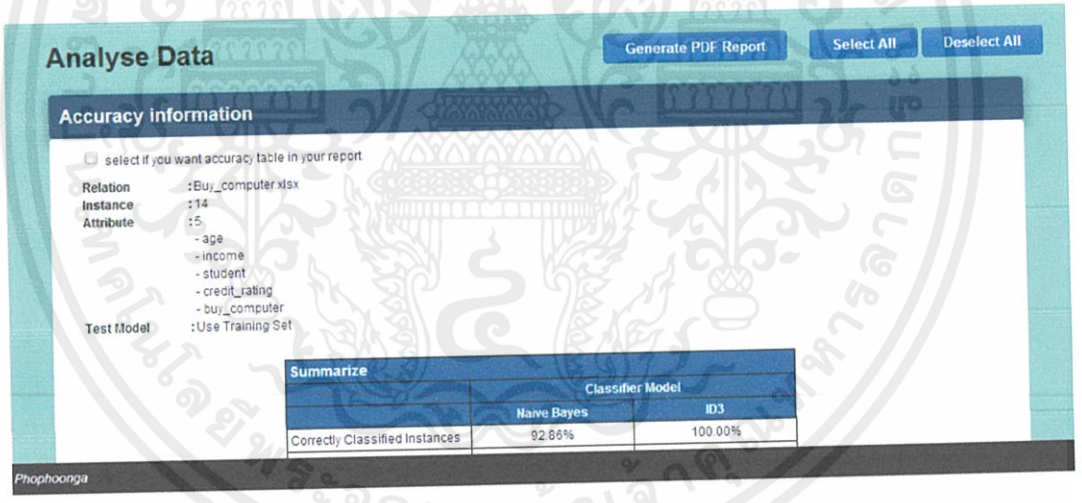
รูปที่ ก.12 แสดงส่วน Import Unseen Data

ระบบจะ Preview Unseen Data ดังภาพ จากนั้นกด NEXT



รูปที่ ก.13 แสดงส่วน Preview Unseen Data

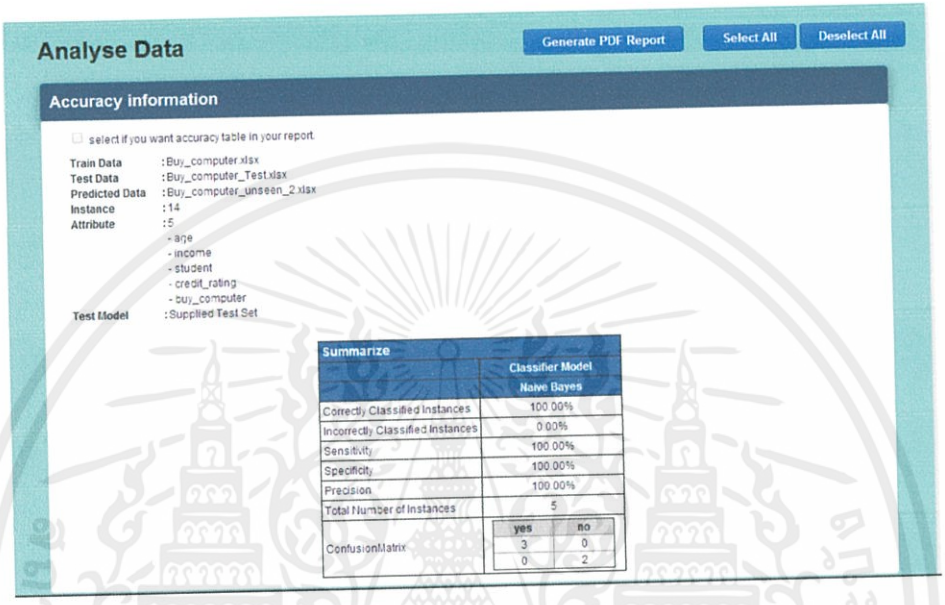
เมื่อกด NEXT แล้วระบบจะทำการนำข้อมูล ไปทำนายหาคำตอบที่ผู้ใช้ต้องการ ดังภาพ



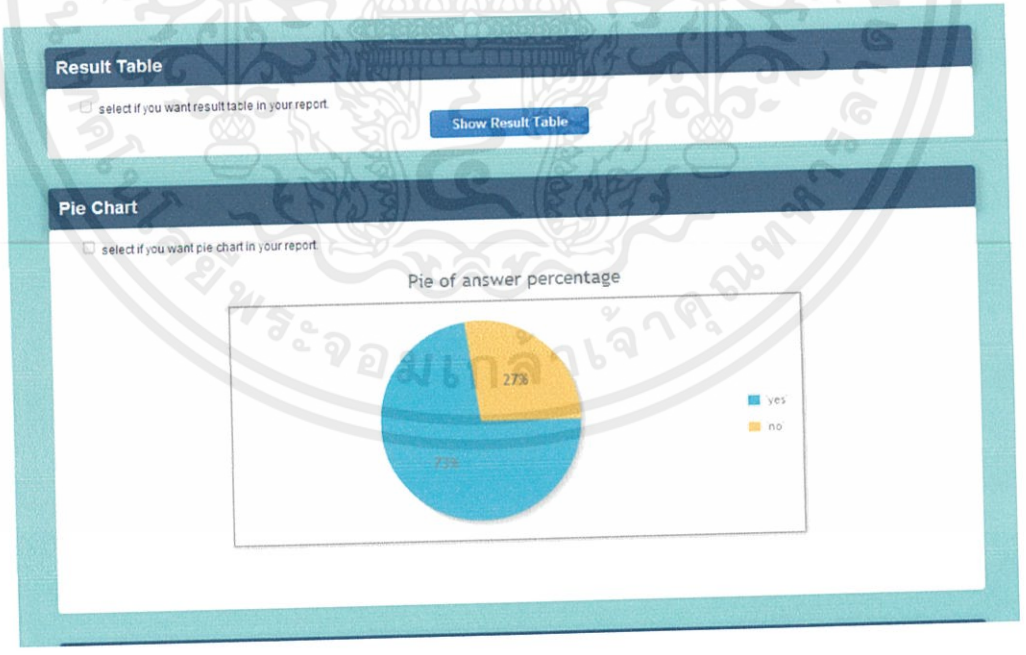
รูปที่ ก.14 แสดงผลคำตอบ

4) ส่วนรายงาน

เมื่อผู้ใช้ทำนายผลข้อมูลแล้วระบบจะแสดงหน้าเว็บ ดังภาพด้านล่าง ที่จะมีทั้ง ตารางข้อมูลค่าความแม่นยำของโมเดลที่ผู้ใช้ได้ใช้ในการทำนายผลข้อมูล, ตารางข้อมูลที่ผ่านการทำนายเรียบร้อยแล้ว, แผนภูมิวงกลม(แสดงเปอร์เซ็นของข้อมูลที่ตรงกับคำตอบแต่ละคำตอบ), และแผนภูมิแท่งของแต่ละแอททริบิว ตามลำดับ



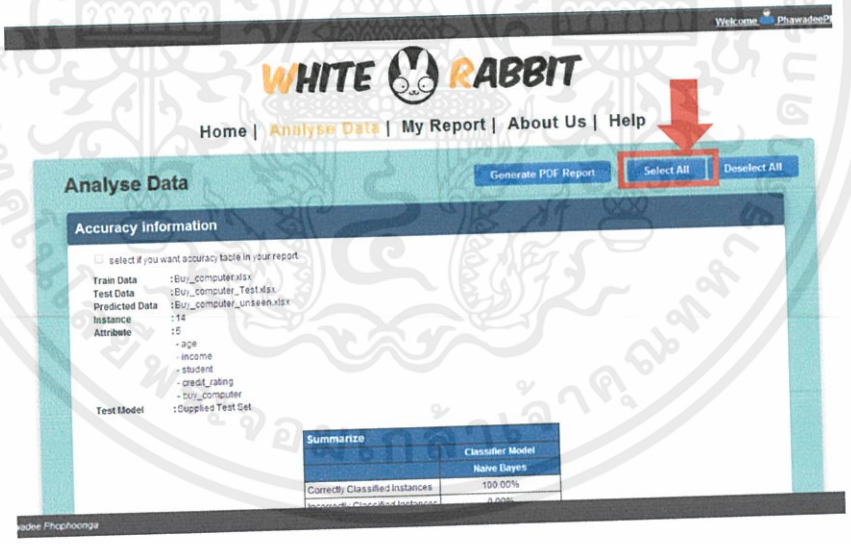
รูปที่ ก.15 แสดงส่วนการสร้างรายงาน



รูปที่ ก.16 แสดงส่วน Pie Chart

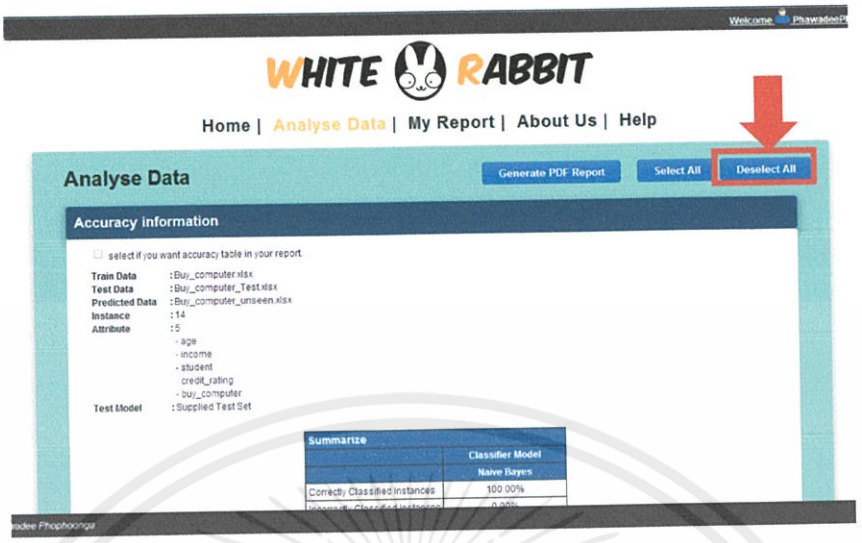
รูปที่ ก.17 แสดงส่วน Bar Chart

ผู้ใช้สามารถสร้างรายงานในหน้านี้ได้ โดยการเลือกที่ Check box หรือสามารถเลือกที่จะสร้างรายงานด้วยข้อมูลทั้งหมด โดยการกดที่ Select All (ในบริเวณกรอบสีแดง)



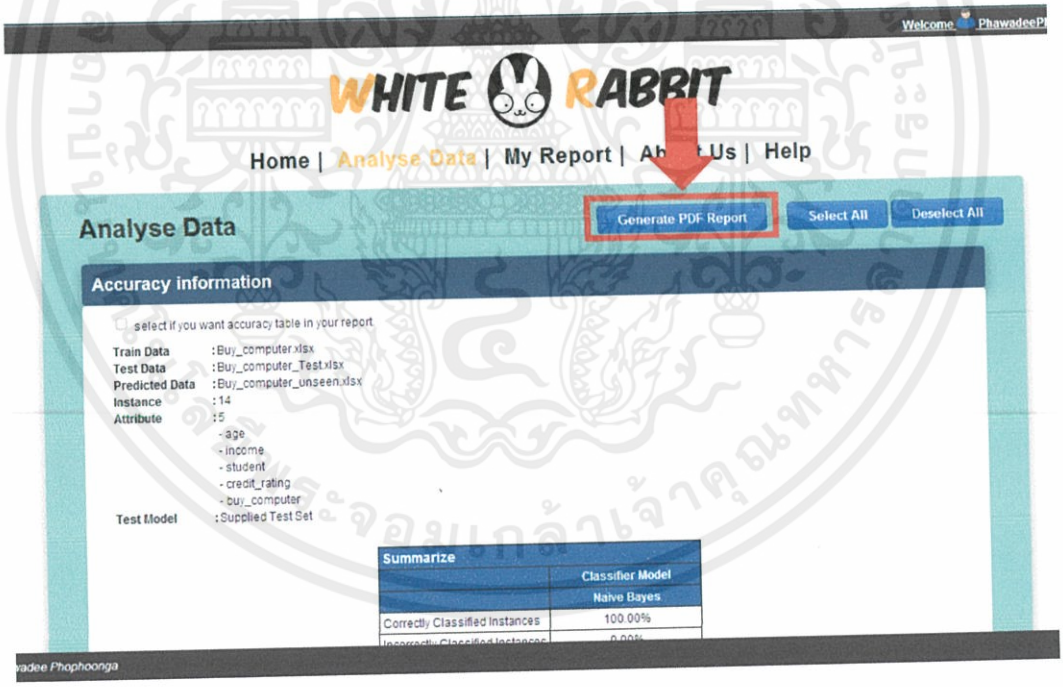
รูปที่ ก.18 แสดงปุ่มSelect All

หรือยกเลิกการเลือกทั้งหมดได้ที่ปุ่ม Deselect All



รูปที่ ก.19 แสดงปุ่มDeselect All

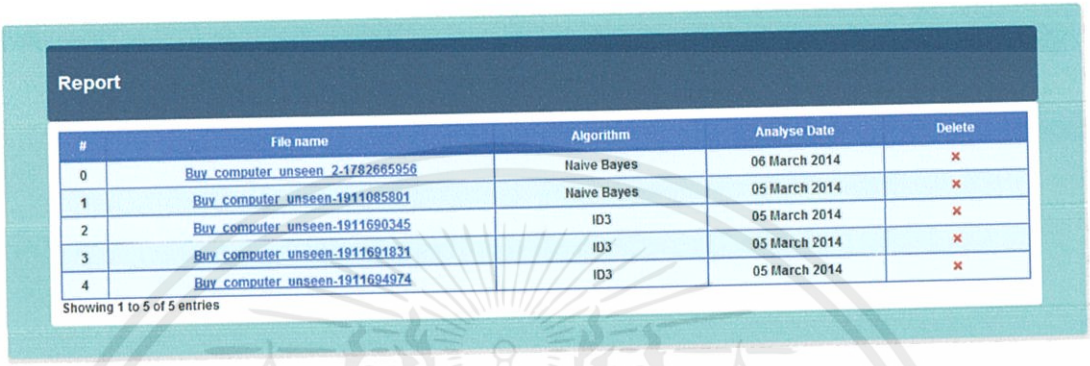
หลังจากเลือกข้อมูลที่ต้องการสร้างรายงานได้แล้ว ผู้ใช้สามารถดาวน์โหลดรายงานออกไปใช้งานได้ โดยการกดที่ปุ่ม Generate PDF Report ดังภาพด้านล่าง



รูปที่ ก.20 แสดงส่วนGenerate PDF Report

ถ้าหากผู้ใช้ต้องการดาวน์โหลดรายงานที่เคยสร้างไว้ สามารถเข้าไปดาวน์โหลดได้โดยการกดที่ My Report ในหน้า Home

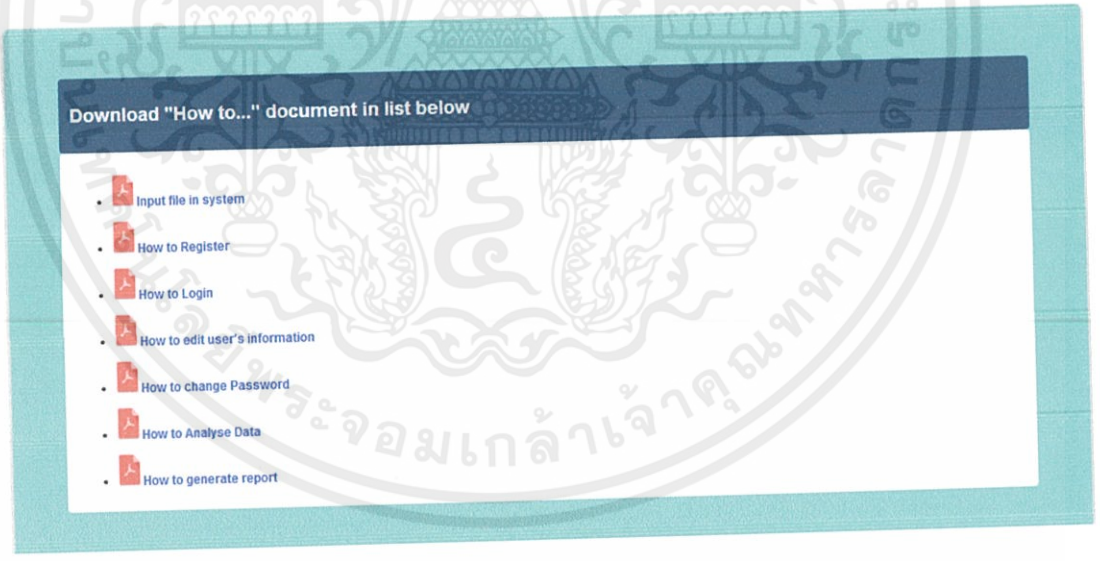
จากนั้นระบบจะแสดงหน้า Report ของผู้ใช้ หน้านี้จะเก็บไฟล์ของรายงานที่ผู้ใช้เคยสร้าง ที่จะประกอบไปด้วยข้อมูลชื่อไฟล์รายงาน, Algorithm ที่ใช้ในการทำนายผลข้อมูล, และวันที่สร้างรายงาน ผู้ใช้สามารถ ดาวน์โหลดรายงาน หรือ ลบรายงานที่ไม่ต้องการได้



รูปที่ ก.21 ส่วนแสดงรายงานของผู้ใช้

5) ส่วนช่วยเหลือ

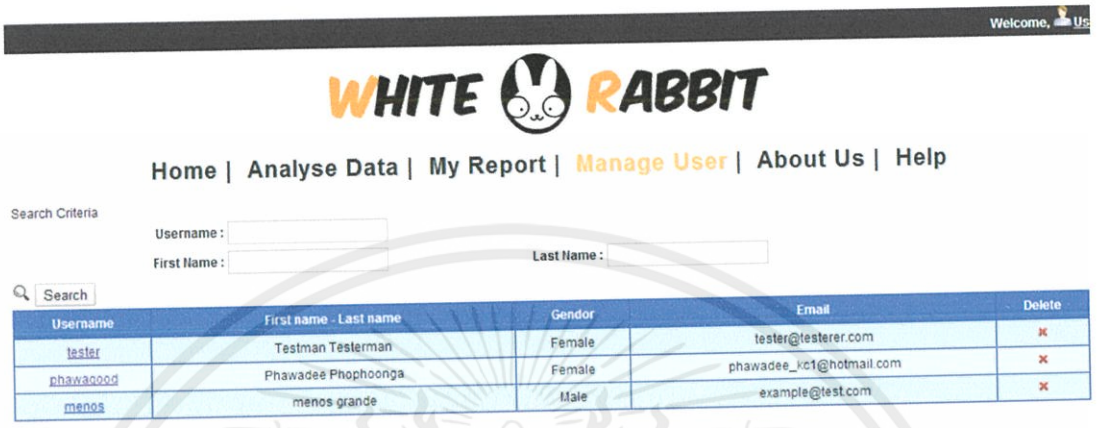
เป็นส่วนแสดงวิธีใช้งานระบบทั้งหมด ผู้ใช้สามารถเข้าไปดูรายละเอียดได้ที่หน้า Help



รูปที่ ก.22 แสดงส่วนช่วยเหลือ

ก.1.2 ส่วนของแอดมิน

ส่วนของแอดมิน ในเรื่องของการสมัครสมาชิก วิเคราะห์ข้อมูล หรือการสร้าง Report จะมีลักษณะเดียวกับ ผู้ใช้อื่นๆ แต่แอดมินสามารถจัดการกับข้อมูลของผู้ใช้ได้ โดยการกดที่ปุ่มเมนู Manage User จากนั้นแอดมินสามารถเสิร์จข้อมูลของผู้ใช้ได้



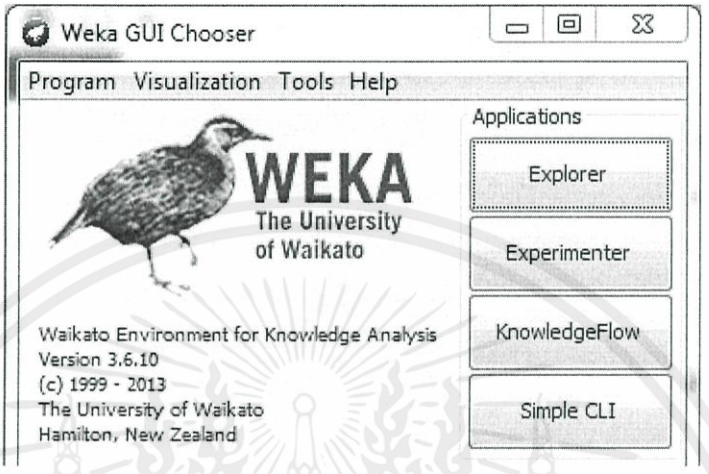
รูปที่ ก.23 แสดงส่วนจัดการข้อมูลผู้ใช้

ภาคผนวก ข

คู่มือการใช้งานระบบ

ข.1 ข้อมูลเบื้องต้น

GUI Chooster ประกอบด้วยปุ่ม 4 ปุ่ม ซึ่งหน้าต่างหน้าเดียวกันจะมีแอพพลิเคชั่นของ WEKA อยู่ด้วยกัน 4 ตัว และมี 4 เมนู ดังรูป



รูปที่ ข.1 GUI ของ WEKA

ปุ่มต่างๆที่สามารถใช้เริ่มแอพพลิเคชั่นดังนี้

- **Explorer** เป็นสภาพเพื่อใช้ค้นหาข้อมูลด้วย WEKA
- **Experimenter** เป็นส่วนที่มีไว้ทำการทดลอง และสร้างการทดสอบทางสถิติระหว่างโครงสร้างการเรียนรู้
- **KnowledgeFlow** ส่วนนี้จะสนับสนุนฟังก์ชั่นที่จำเป็นที่จำเป็นเหมือนกับ Explorer แต่มีการใช้ อินเตอร์เฟสแบบ drag and drop อีกหนึ่งความได้เปรียบของส่วนนี้คือรองรับการเรียนรู้แบบเพิ่มขึ้น
- **Simple CLI** ประกอบด้วยอินเตอร์เฟสแบบ command line ซึ่งอนุญาตให้มีการประมวลผลโดยตรงผ่านชุดคำสั่งของ WEKA ในระบบปฏิบัติการที่ไม่มีอินเตอร์เฟสแบบ command line เป็นของตนเอง

ข.2 หน้าที่การใช้งาน

มีส่วนการทำงานดังนี้

1) Preprocess เลือกและปรับแต่งข้อมูลก่อนนำไปใช้

2) Classify ฝึกและทดสอบโครงสร้างที่เรียนรู้ซึ่งทำการแบ่งประเภทไว้

3) Cluster เป็นการจัดกลุ่มข้อมูล

4) Associate การหากฎการสัมพันธ์กันของข้อมูล

5) Select attributes เลือกคุณลักษณะที่มีความสัมพันธ์กับข้อมูลมากที่สุด

6) Visualize ดูการพลอตกราฟแบบสองมิติของข้อมูลที่เราทำการโต้ตอบด้วย

ข.3 กระบวนการก่อนประมวลผล (Preprocessing)

ข.3.1 การโหลดข้อมูล (Loading Data)

1) **Open file…** ประกอบด้วยกล่องซึ่งให้คุณค้นหาตำแหน่งสำหรับตัวข้อมูลที่จะใช้
2) **Open URL…** ใช้ Uniform Resource Locator address สำหรับหาว่าข้อมูลถูกเก็บไว้ที่ไหน
3) **Open DB…** ใช้ในการอ่านข้อมูลจากฐานข้อมูล
4) **Generate…** ทำให้สามารถสร้างข้อมูลเทียมจากจากหลากหลาย DataGenerators

ข.3.2 ความสัมพันธ์ปัจจุบัน (Current Relation)

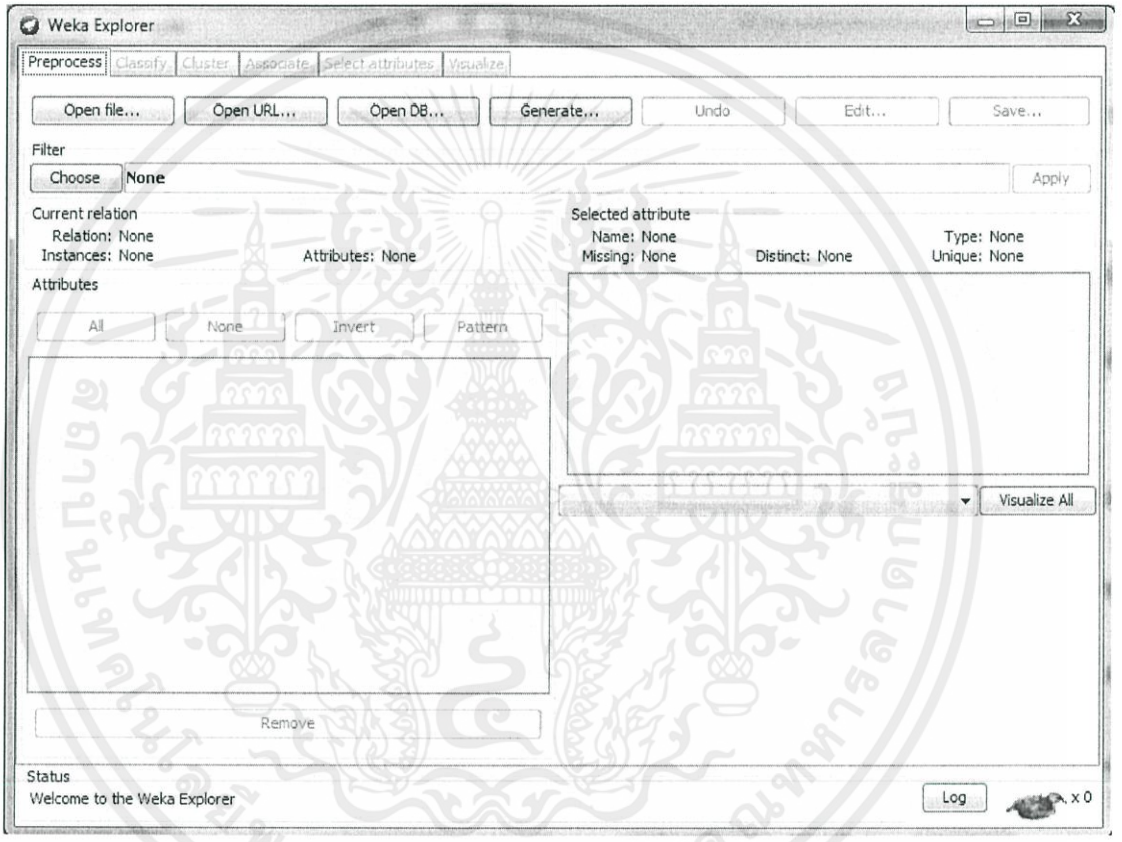
หลังจากโหลดข้อมูลมาแล้วส่วน Preprocess panel จะแสดงข้อมูลมากมายกล่อง Current Relation เป็นส่วนแสดงข้อมูลที่โหลดมาซึ่งสามารถถูกแปลงเป็นตารางแสดงความสัมพันธ์ หนึ่งตารางในฐานข้อมูลได้ ประกอบด้วย 3 ส่วนย่อยคือ

1) **Relation** เป็นชื่อความสัมพันธ์ที่ให้ไว้กับไฟล์ที่โหลดมา
2) **Instances** จำนวนตัวอย่างทั้งหมดที่มี
3) **Attributes** จำนวนคุณลักษณะ ของข้อมูล

ข.3.3 การคัดเลือกคุณลักษณะ (Selected Attribute)

ข้างล่างกล่อง Current relation คือกล่องที่มีชื่อว่า Attributes ซึ่งมีอยู่สี่ปุ่ม และลงไปอีกเป็นรายการคุณลักษณะในความสัมพันธ์ที่มีอยู่ประกอบด้วย 3 คอลัมน์

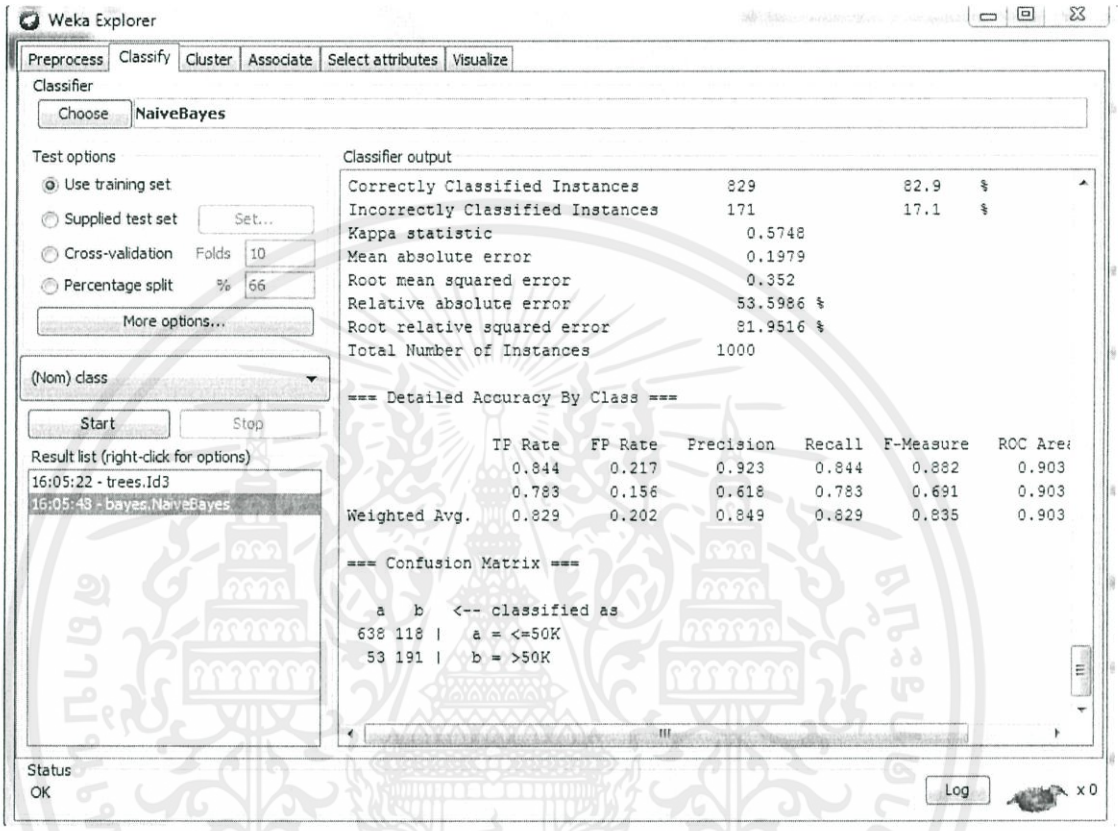
1) **No.** หมายเลขที่กำหนดไว้กับคุณลักษณะตามลำดับที่อยู่ในไฟล์ข้อมูล
2) **Selection tick boxes** เป็นส่วนที่เลือกว่าคุณลักษณะใดให้แสดงบนความสัมพันธ์นี้บ้าง
3) **Name** ชื่อคุณลักษณะที่ประกาศไว้ในไฟล์ข้อมูล



รูปที่ ข.2 พื้นที่ทำงานของ Preprocessing

ข.4 การจำแนกประเภท

มีพื้นทำงาน 6 ส่วน คือ การเลือกการจัดหมวดหมู่ ตัวเลือกการทดสอบ ประเภทคุณลักษณะ การฝึกฝนตัวจำแนกประเภท ข้อความที่ได้ออกมาจากตัวจำแนกประเภท และรายการผลลัพธ์ ดังรูป



รูปที่ ข.3 พื้นที่ทำงานของ Classification

ข.4.1 การเลือกตัวจัดหมวดหมู่

ด้านบนสุดในส่วนนี้เป็นกล่อง Classifier หรือตัวจำแนกประเภทซึ่งมีกล่องข้อความที่มีชื่อของตัวจัดหมวดหมู่ที่ใช้อยู่ และตัวเลือกต่างๆของมัน การคลิกซ้ายบนกล่องข้อความจะเป็นการเปิดให้ขึ้นมา มีชื่อว่า GenericObjectEditor ซึ่งสามารถตั้งค่าตัวเลือกต่างๆของตัวจำแนกประเภทนี้ได้ การคลิกขวาจะเป็นการคัดลอกข้อความที่เราตั้งใจใส่ไว้บนคลิปบอร์ด

ข.4.2 ตัวเลือกการทดสอบ

หลังจากเลือกตัวจำแนกประเภทข้อมูลนำไปสู่การตั้งค่าตัวเลือกซึ่งอยู่ในกล่อง Test options มีวิธีการทดสอบ 4 แบบด้วยกันคือ

1) **Use training set** ตัวจำแนกประเภทข้อมูลถูกประเมินโดยการทำนายตัวอย่างที่เราทำการฝึกเข้าไปว่าได้ผลลัพธ์ดีแค่ไหน

2) **Supplied test set** ตัวจำแนกประเภทข้อมูลจะถูกประเมินโดยการทำนายตัวอย่างที่เอามาจากไฟล์หนึ่งๆ ว่าได้ผลลัพธ์ดีแค่ไหน โดยการคลิก set จะเป็นการเปิดกล่องซึ่งทำให้เลือกไฟล์ที่จะนำมาทดสอบได้

3) **Cross validation** ตัวจำแนกประเภทจะถูกประเมินจากการตรวจสอบแบบไขว้ ซึ่งเป็นการแบ่งข้อมูลเป็นก้อนๆ ตามจำนวนโฟลเดอร์ในกล่องข้อความ

4) **Percentage split** ตัวจำแนกประเภทจะถูกประเมินว่าทำนายได้ดีแค่ไหนจากข้อมูลที่ถูกดึงออกมาเป็นตัวทดสอบ โดยการกำหนดเปอร์เซ็นต์ที่จะแบ่งออกมา

ข.4.3 ประเภทของคุณลักษณะ

ตัวจำแนกประเภทใน WEKA ถูกออกแบบมาเพื่อฝึกแล้วใช้ทำนายคลาสของลักษณะหนึ่งออกมา ซึ่งเป็นเป้าหมายของการทำนายตัวจำแนกประเภทบางตัวทำได้แค่การเรียนรู้คลาสแบบ nominal (เป็นค่าที่ไม่ใช่ตัวเลข) บางตัวเรียนได้แต่คลาสแบบ Numeric(มีค่าเป็นตัวเลข) มีแค่บางตัวที่เรียนรู้ได้ทั้งสองแบบ โดยทั่วไปคุณลักษณะที่จะนำมาจัดเป็นคลาสจะเป็นคุณลักษณะตัวสุดท้ายที่อยู่ในข้อมูล

ข.4.4 การฝึกฝนตัวจำแนกประเภท

ในตัวจำแนกระเภทหนึ่งที่ค่า Test Option และคลาสได้ถูกตั้งค่าไว้แล้วทั้งหมดกระบวนการเรียนรู้จะถูกเริ่มโดยการคลิกที่ปุ่ม Start ในขณะที่ตัวจำแนกประเภทกำลังยุ่งอยู่กับการฝึกฝน คุณสามารถหยุดการฝึกเมื่อไหร่ก็ได้โดยการกดที่ปุ่ม Stop เมื่อการฝึกฝนเสร็จสิ้นจะมีบางอย่างเปิดขึ้นบนพื้นที่ของ Classifier output ทางด้านขวาจะแสดงข้อความผลลัพธ์ของการฝึกและการทดสอบ จะมีตัวเลือกใหม่ปรากฏใน result list เราจะมองเห็นผลลัพธ์ได้จากด้านล่างแต่อย่างแรกเราจะดูคือส่วนข้อความที่ได้ออกมา

ข.4.5 ข้อความที่ได้ออกมาจากตัวจำแนกประเภท

พื้นที่ Classifier Output เป็นแบบสกอล์บาร์ ซึ่งเราสามารถเลื่อนหาดูส่วนต่างๆของผลลัพธ์ได้ ผลลัพธ์ได้ถูกแบ่งเป็นส่วนต่างๆ ดังนี้

1) **Run information** มีรายการข้อมูลต่างๆได้แก่ส่วนประกอบต่างๆของโครงสร้างเรียนรู้ ชื่อความสัมพันธ์ จำนวนข้อมูล คุณลักษณะต่างๆและ การทดสอบที่ใช้

2) **Classifier model** เป็นการแสดงแบบข้อความของโมเดลการจำแนกประเภทซึ่งถูกสร้างไว้ด้วยข้อมูลฝึกฝนทั้งหมด

3) **ผลลัพธ์จากการทดสอบ** แบ่งออกเป็น

Summary รายการของผลสรุปทางสถิติบอกถึงความแม่นยำของตัวจำแนกประเภท ว่าสามารถทำนายคลาสได้อย่างถูกต้องจากตัวอย่างภายใต้การทดสอบที่กำหนดไว้

Detail Accuracy by Class ความแม่นยำการทำนายของตัวจำแนกประเภท โดยแบ่งเป็นรายละเอียดแต่ละคลาส

Confusion Matrix แสดงว่ามีตัวอย่างเท่าใดถูแบ่งลงไปในแต่ละคลาส ประกอบด้วยการแสดงจำนวนของตัวอย่างที่ใช้ ซึ่งแถวจะแสดงถึงคลาสที่ถูกต้อง คอลัมน์คือคลาสที่ทำนายออกมา

ข.4.6 รายงานผลลัพธ์

หลังจากผ่านกระบวนการฝึกฝนตัวจำแนกประเภทต่างๆแล้ว รายการผลลัพธ์ถูกเก็บไว้หลายตัวสามารถคลิกซ้ายเลือกแต่ละรายการไปมา และแสดงผลระหว่างผลลัพธ์ที่ถูกสร้างไว้ได้ กด delete เพื่อลบผลลัพธ์ที่เลือกไว้ได้ คลิกขวาบนรายการจะเป็นการเรียกเมนูขึ้นมาซึ่งประกอบไปด้วย

1) **View in main window** แสดงผลลัพธ์ในหน้าต่างหลัก
2) **View in separate window** เปิดหน้าต่างลอยขึ้นมาเพื่อดูผล
3) **Save result buffer** เปิดกล่องซึ่งสามารถบันทึกผลลัพธ์เป็นไฟล์ข้อความได้
4) **Load model** โหลดโมเดลที่ยังไม่ได้ฝึกจากไฟล์ไบนารี่
5) **Save model** บันทึกโมเดลเป็นไฟล์ไบนารี่ ใน Java'serialized object' form
6) **Re-evaluate model on current test** นำโมเดลที่สร้างไว้แล้วมาทดสอบประสิทธิภาพด้วยดาต้าเซต ซึ่งกำหนดไว้ด้วยปุ่ม set ใน supplied test set option

7) **Visual classifier** errors เปิดหน้าต่างจำลองขึ้นมาเพื่อพล็อตผลลัพธ์ของการจัดหมวดหมู่ตัวอย่างไว้อย่างถูกต้อง จะถูแสดงเป็นรูปกากบาท ส่วนที่ถูกจัดผิดจะเป็นรูปสี่เหลี่ยม

8) **Visualize tree or Visualize** graph เปิดการแสดงแบบภาพของโครงสร้างของโมเดลจัดแสดงประเภทหากสามารถแสดงได้

9) **Visualize margin curve** สร้างพล็อตเป็นรูปขอบเขตการทำนาย ขอบเขตจะถูกกำหนดเป็น ค่าความแตกต่างความน่าจะเป็นในการทำนายทุกคลาส และโอกาสสูงสุดของการทำนายในคลาสอื่นๆ

10) **Visualize threshold curve** สร้างพล็อตในรูปในการทำนายเกินขอบเขตที่กำหนดค่าโดยใช้ Threshold ในระหว่างคลาสต่างๆ เช่น default threshold มีค่า 0.5 ความน่าจะเป็นที่จะทำนายเป็น+ จะต้องมากกว่า 0.5 เพื่อเป็นการรับประกันการทำนายเป็น + นี้

11) **Visualize cost curve** สร้างพล็อตซึ่งให้การแสดงที่ชัดเจนของค่าคาดหวัง

12) **Plugin** เมนูนี้จะพบเมื่อมีการอนุญาตให้ใช้ปลั๊กอินของตัวจำลอง