2.1.2.1 Selection ขั้นตอนการดึงข้อมูลเพื่อการวิเคราะห์จากแหล่งที่บันทึกไว้

2.1.2.2 Preprocessing ขั้นตอนการเตรียมข้อมูลสำหรับการทำ Data Mining

2.1.2.3 Transformation ขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน

2.1.2.4 Data mining ขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่

2.1.2.5 Interpretation/Evaluation ขั้นตอนการประเมินรูปแบบและวัดผลประสิทธิภาพที่ได้จากการทำเหมืองข้อมูล

## **2.1.3 เทคนิคในการทำเหมืองข้อมูล** (วิกิพีเดีย สารานุกรมเสรี, 2559)

2.1.3.1 เทคนิคการค้นหาความสัมพันธ์ของข้อมูล (Association)

เทคนิคการค้นหาความสัมพันธ์ของข้อมูลเป็นเทคนิคที่แสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุที่เกิดขึ้นพร้อมกันตัวอย่างของการประยุกต์ใช้กฎเชื่อมโยง เช่น การวิเคราะห์ข้อมูลการขายสินค้า โดยเก็บข้อมูลจากระบบ ณ จุดขาย (POS) หรือร้านค้าออนไลน์แล้วพิจารณาสินค้าที่ผู้ซื้อมักจะซื้อพร้อมกันเช่น ถ้าพบว่าคนที่ซื้อเทปวิดีโอมักจะซื้อเทปกาวด้วยร้านค้าก็อาจจะจัดร้านให้สินค้าสองอย่างอยู่ใกล้กันเพื่อเพิ่มยอดขายหรืออาจจะพบว่าหลังจากคนซื้อหนังสือ ก แล้ว มักจะซื้อหนังสือ ข ด้วยก็สามารถนำความรู้นี้ไปแนะนำผู้ที่กำลังจะซื้อหนังสือ ก ได้

2.1.3.2 เทคนิคการจำแนกประเภทข้อมูล (Classification)

เทคนิคการจำแนกประเภทข้อมูลเป็นเทคนิคที่หากฎเพื่อระบุประเภทของวัตถุจากคุณสมบัติของวัตถุ เช่น หาความสัมพันธ์ระหว่างผลการตรวจร่างกายต่าง ๆ กับการเกิดโรคโดยใช้ข้อมูลผู้ป่วยและการวินิจฉัยของแพทย์ที่เก็บไว้เพื่อนำมาช่วยวินิจฉัยโรคของผู้ป่วยหรือการวิจัยทางการแพทย์ ในทางธุรกิจจะใช้เพื่อดูคุณสมบัติของผู้ที่จะก่อหนี้ดีหรือหนี้เสียเพื่อประกอบการพิจารณาการอนุมัติเงินกู้

2.1.3.3 เทคนิคการแบ่งกลุ่มของข้อมูล (Clustering)

เทคนิคการแบ่งกลุ่มของข้อมูลเป็นเทคนิคแบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่ม แบ่งกลุ่มผู้ป่วยที่เป็นโรคเดียวกันตามลักษณะอาการเพื่อนำไปใช้ประโยชน์ในการวิเคราะห์หาสาเหตุของโรคโดยพิจารณาจากผู้ป่วยที่มีอาการคล้ายคลึงกัน

**2.1.4 ขั้นตอนวิธีในการจัดกลุ่ม**

## การจัดกลุ่ม (Clustering) เป็นวิธีการที่พิจารณาข้อมูลแต่ละแถวเสมือนเป็นวัตถุ (object) ซึ่งจะมีหลักการเหมือนกับการจำแนกประเภทข้อมูล คือจะทำการแบ่งข้อมูลออกเป็นกลุ่ม (คลัสเตอร์) โดยจะจัดให้ข้อมูลที่มีความคล้ายคลึงกันอยู่ในคลัสเตอร์เดียวกันและข้อมูลที่อยู่ต่างคลัสเตอร์กันจะมีความคล้ายคลึงกันน้อยที่สุด

ซึ่งความเหมือนหรือต่างกันสามารถเปรียบเทียบได้กับความใกล้ชิดกันของวัตถุใดๆโดยใช้ระยะทางเป็นตัวชี้วัดคุณภาพของแต่ละคลัสเตอร์สามารถอธิบายได้จากเส้นผ่านศูนย์กลางของคลัสเตอร์ (diameter) ระยะทางเป็นตัวชี้วัดคุณภาพของแต่ละคลัสเตอร์สามารถอธิบายได้จากเส้นผ่านศูนย์กลางของคลัสเตอร์ (diameter) ซึ่งแสดงระยะห่างมากสุดของวัตถุสองชิ้นที่อยู่ในคลัสเตอร์เดียวกันแต่ละคลัสเตอร์จะมีตัวแทนที่สามารถแทนวัตถุทุกชิ้นของคลัสเตอร์นั้นได้เช่นการใช้จุดศูนย์กลางคลัสเตอร์ (centroid) แทนคลัสเตอร์นั้น สำหรับบางเทคนิคตัวแทนของคลัสเตอร์อาจมีได้หลายตัวแทน ทั้งนี้ขึ้นอยู่กับความเหมาะสมของแต่ละเทคนิคที่เลือกใช้

## **2.1.5 ขั้นตอนวิธีในการค้นหาความสัมพันธ์** (อดุลย์, 2559)

## Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูล โดยค้นหาความสัมพันธ์ของข้อมูลทั้งสองชุดหรือมากกว่าสองชุดขึ้นไปไว้ด้วยกัน ความสําคัญของกฎทําการวัดโดยใช้ข้อมูลสองตัวด้วยกันคือค่าสนับสนุน (Support) ซึ่งเป็นเปอร์เซ็นต์ของการดําเนินการที่กฎสามารถนําไปใช้หรือเป็นเปอร์เซ็นต์ของการดําเนินการที่กฎที่ใช้มีความถูกต้องและข้อมูลตัวที่สองที่นํามาใช้วัดคือค่าความมั่นใจ (Confidence) ซึ่งเป็นจํานวนของกรณีที่กฎถูกต้องโดยสัมพันธ์กับจํานวนของกรณีที่กฎสามารถ นําไปใช้ได้ ในการหากฎความสัมพันธ์นั้นจะมีขั้นตอนวิธีการหาหลายวิธีด้วยกัน แต่ขั้นตอนวิธีที่เป็นที่รู้จักและใช้อย่างแพร่หลายคือ อัลกอริทึม Apriori (เอกสิทธิ์, 2559)

การหากฎความสัมพันธ์ด้วยวิธี Apriori ซึ่งจะมีอยู่ 2 ขั้นตอนใหญ่ๆ คือ

1. การหา frequent itemset เป็นการหารูปแบบของข้อมูลที่เกิดขึ้นร่วมกันบ่อยๆในฐานข้อมูลหรือมากกว่าค่า minimum support ที่ผู้ใช้กำหนด ในขั้นตอนนี้จะแบ่งได้อีกเป็น 2 ขั้นตอนย่อย

1.1 การสร้างรูปแบบของ itemset (join) จะใช้รูปแบบของ itemset ที่มีค่ามากกว่าminimum support มาทำการสร้างรูปแบบของ itemset ที่มีขนาดยาวมากขึ้นทีละหนึ่งขั้นไปเรื่อยๆ

1.2 การนับค่า support (count) หลังจากที่สร้างรูปแบบของ itemset ได้แล้ว ขั้นถัดมาจะทำการคำนวณค่า support ที่เกิดขึ้น โดยที่ support คือจำนวนเปอร์เซ็นต์ที่พบ itemset ในฐานข้อมูล

2. การสร้าง association rule หลังจากที่หา frequent itemset ได้แล้วจะนำรูปแบบที่หาได้มาสร้างเป็นกฎความสัมพันธ์โดยเช่น Apple => Cereal หมายความว่าเมื่อลูกค้าซื้อ Apple แล้วลูกค้าจะซื้อ Cereal ร่วมไปด้วย

**ตารางที่ 2-1** แสดง transaction database ที่จะใช้หากฎความสัมพันธ์

|  |  |
| --- | --- |
| **Transaction ID** | **Items** |
| 1 | Apple, Cereal, Diapers |
| 2 | Beer, Cereal, Eggs |
| 3 | Apple, Beer, Cereal, Eggs |
| 4 | Beer, Eggs |

ตัวอย่างข้อมูลที่แสดงในตารางที่ 2-1 ซึ่งเรียกว่า transaction database ในฐานข้อมูลนี้จะประกอบด้วย 4 transaction และมีสินค้าที่ลูกค้าซื้อทั้งหมด 5 ประเภทคือ Apple, Beer, Cereal, Diapers และ Eggs กำหนดค่า minimum support ไว้ที่ 50% นั่นคือ สินค้าที่มีการซื้อมากกว่าหรือเท่ากับ 50% จะถือว่าเป็น frequent itemset ขั้นตอนการทำงานมีดังต่อไปนี้

1. คำนวณหาค่า support ของแต่ละ item ดังแสดงในตารางที่ 2-1 แต่ละแถวคือข้อมูลสินค้าแต่ละชนิดและแต่ละคอลัมน์มีความหมายดังนี้

- คอลัมน์ที่ 1 คือชื่อสินค้าหรือ item หรือ itemset

- คอลัมน์ที่ 2-5 คือหมายเลขของ transaction ถ้ามีการซื้อสินค้าใน transaction ใดที่คอลัมน์ของ transaction นั้นจะมีหมายเลข 1 แต่ถ้าไม่มีจะเป็นเลข 0

- คอลัมน์ที่ 6 คือ ค่า support

**ตารางที่ 2-2** แสดงการคำนวณค่า support ของสินค้าแต่ละชนิด

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Items** | **Transaction ID** | | | | **Support** |
| **1** | **2** | **3** | **4** |
| Apple | 1 | 0 | 1 | 0 | 2/4 = 50% |
| Beer | 0 | 1 | 1 | 1 | 3/4 = 75% |
| Cereal | 1 | 1 | 1 | 0 | 3/4 = 75% |
| Diapers | 1 | 0 | 0 | 0 | 1/4 = 25% |
| Eggs | 0 | 1 | 1 | 1 | 3/4 = 75% |

จากค่า support ที่คำนวณได้ในตารางที่ 2-2 จะเห็นว่า Diapers มีค่า support ต่ำกว่าค่า minimum support (25% < 50%) ดังนั้น Diapers จะถูกตัดออกและไม่นำไปพิจารณาสร้างเป็น itemset ที่มีความยาว 2 ต่อไป ดังตารางที่ 2-3 และเรียก Apple, Beer, Cereal และ Eggs ว่า frequent itemset

**ตารางที่ 2-3** Diapers ถูกตัดทิ้งเนื่องจากมีค่า support ต่ำกว่าค่า minimum support ที่กำหนดไว้

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Items** | **Transaction ID** | | | | **Support** |
| **1** | **2** | **3** | **4** |
| Apple | 1 | 0 | 1 | 0 | 2/4 = 50% |
| Beer | 0 | 1 | 1 | 1 | 3/4 = 75% |
| Cereal | 1 | 1 | 1 | 0 | 3/4 = 75% |
| ~~Diapers~~ | ~~1~~ | ~~0~~ | ~~0~~ | ~~0~~ | ~~1/4 = 25%~~ |
| Eggs | 0 | 1 | 1 | 1 | 3/4 = 75% |

2. นำ frequent itemset ที่ได้จากตารางที่ 2-3 มาสร้างเป็น itemset ที่มีความยาว 2 โดยใช้การ join ได้เป็น {Apple, Beer}, {Apple, Cereal}, {Apple, Eggs}, {Beer, Cereal}, {Cereal, Eggs} และเนื่องจากเป็นเซตลำดับของข้อมูลไม่มีผล นั่นคือ {Apple, Beer} = {Beer, Apple} ส่วนการคำนวณค่า support ก็นำข้อมูล transaction ในตารางที่ 2-3 มาทำ intersect กัน (หรือใช้ bit operation AND ก็ได้) เช่น การหา support ของ {Apple, Beer} จะเกิดจาก

- Transaction (Apple) = {1, 0, 1, 0}

- Transaction (Beer) = {0, 1, 1, 1}

- Transaction (Apple, Beer) = {0, 0, 1, 0}

ค่า support ของ itemset ที่มีความยาว 2 ทั้งหมดซึ่งจะแสดงในตารางที่ 2-4 และ itemset ที่มีค่า support น้อยกว่า minimum support จะถูกตัดทิ้งไปทำให้เหลือ frequent itemset ที่มีความยาว 2 คือ {Apple, Cereal}, {Beer, Cereal}, {Beer, Eggs}, {Cereal, Eggs} ดังแสดงในตารางที่ 2-5

**ตารางที่ 2-4** แสดงค่า support ของ itemset ที่มีความยาว 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Items** | **Transaction ID** | | | | **Support** |
| **1** | **2** | **3** | **4** |
| {Apple, Beer} | 0 | 0 | 1 | 0 | 1/4 = 25% |
| {Apple, Cereal} | 1 | 0 | 1 | 0 | 2/4 = 50% |
| {Apple, Eggs} | 0 | 0 | 1 | 0 | 1/4 = 25% |
| {Beer, Cereal} | 0 | 1 | 1 | 0 | 2/4 = 50% |
| {Beer, Eggs} | 0 | 1 | 1 | 1 | 3/4 = 75% |
| {Cereal, Eggs} | 0 | 1 | 1 | 0 | 2/4 = 50% |

**ตารางที่ 2-5** แสดง frequent itemset ที่มีความยาว 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Items** | **Transaction ID** | | | | **Support** |
| **1** | **2** | **3** | **4** |
| ~~{Apple, Beer}~~ | ~~0~~ | ~~0~~ | ~~1~~ | ~~0~~ | ~~1/4 = 25%~~ |
| {Apple, Cereal} | 1 | 0 | 1 | 0 | 2/4 = 50% |
| ~~{Apple, Eggs}~~ | ~~0~~ | ~~0~~ | ~~1~~ | ~~0~~ | ~~1/4 = 25%~~ |
| {Beer, Cereal} | 0 | 1 | 1 | 0 | 2/4 = 50% |
| {Beer, Eggs} | 0 | 1 | 1 | 1 | 3/4 = 75% |
| {Cereal, Eggs} | 0 | 1 | 1 | 0 | 2/4 = 50% |

3. ถัดมาทำการสร้าง itemset ที่มีความยาวเท่ากับ 3 โดยมีเงื่อนไขว่าข้อมูล item แรกจะต้องมีค่าเหมือนกันจึงสามารถทำการ join กันได้ เช่น {Beer, Cereal, Eggs} เกิดจากการ join {Beer, Cereal} และ {Beer, Eggs} แต่ไม่สามารถ join {Apple, Cereal} และ {Beer, Cereal} ได้

ดังนั้นในขั้นตอนนี้จะมีเพียงแค่ itemset เดียวคือ {Beer, Cereal, Eggs} และเป็น frequent itemset ดังแสดงในตารางที่ 2-6

**ตารางที่ 2-6** แสดง itemset ที่มีความยาวเท่ากับ 3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Items** | **Transaction ID** | | | | **Support** |
| **1** | **2** | **3** | **4** |
| {Beer, Cereal, Eggs} | 0 | 1 | 1 | 0 | 2/4 = 50% |

จากตารางที่ 2-6 เราไม่สามารถสร้าง itemset ที่มีความยาวมากขึ้นกว่านี้ได้อีก ขั้นตอนการหา frequent itemset จึงหยุด และได้ frequent itemset ทั้งหมดดังแสดงในตารางที่ 2-7

**ตารางที่ 2-7** แสดง frequent itemset ทั้งหมดที่หาได้

|  |  |  |
| --- | --- | --- |
| **Frequent itemset** | **Support** | **Size** |
| {Apple} | 1/4 = 25% | 1 |
| {Beer} | 3/4 = 75% | 1 |
| {Cereal} | 3/4 = 75% | 1 |
| {Eggs} | 3/4 = 75% | 1 |
| {Apple, Cereal} | 2/4 = 50% | 2 |
| {Beer, Cereal} | 2/4 = 50% | 2 |
| {Beer, Eggs} | 3/4 = 75% | 2 |
| {Cereal, Eggs} | 2/4 = 50% | 2 |
| {Beer, Cereal, Eggs} | 2/4 = 50% | 3 |

4. ขั้นตอนถัดมาคือการสร้างกฎความสัมพันธ์จาก frequent itemset ที่หาได้โดยจะพิจารณา frequent itemset ที่มีความยาวมากกว่า 2 item ขึ้นไปมาสร้างเป็นกฎความสัมพันธ์เช่น {Apple, Cereal} จะสร้างกฎความสัมพันธ์ได้เป็น Apple => Cereal เป็นต้น โดยกฎความสัมพันธ์ทั่วไปจะแสดงเป็น

โดยที่ LHS (Left Hand Side) แสดงรูปแบบของ itemset ด้านซ้ายของกฎความสัมพันธ์และ RHS (Right Hand Side) แสดงรูปแบบของ itemset ด้านขวาของกฎความสัมพันธ์

ในการพิจารณาว่ากฎความสัมพันธ์ที่สร้างได้ดีหรือไม่จำเป็นจะต้องมีตัววัดประสิทธิภาพของกฎ 2 ตัววัด คือ confidence และ lift

- ค่า confidence แสดงความเชื่อมั่นของกฎความสัมพันธ์ที่เมื่อรูปแบบ LHS เกิดขึ้นแล้วรูปแบบ RHS จะเกิดขึ้นด้วยเป็นจำนวนกี่เปอร์เซ็นต์ การคำนวณค่า confidence หาได้จาก

โดยที่ support (LHS, RHS) คือ ค่า support ที่รูปแบบ LHS และ RHS เกิดขึ้นพร้อมๆ กันตัวอย่างการหา confidence ของกฎ Apple => Cereal แสดงได้ดังนี้

- ค่า lift คือค่าที่บ่งบอกว่าการเกิดรูปแบบ LHS และ RHS มีความสัมพันธ์กันแค่ไหน โดยถ้าค่า lift เป็น 1 แสดงว่ารูปแบบ LHS และ RHS ไม่ขึ้นต่อกัน ค่า lift คำนวณได้จาก

ตัวอย่างของการหาค่า lift ของกฎ Apple => Cereal แสดงได้ดังนี้

ค่า confidence และ lift ของกฎความสัมพันธ์ทั้งหมดที่สร้างได้แสดงในตารางที่

2-8 โดยทำการเรียงตามค่า confidence และ lift จากมากไปหาน้อย

**ตารางที่ 2-8** แสดงกฎความสัมพันธ์ทั้งหมดที่สร้างได้พร้อมทั้งค่า confidence และ lift

|  |  |  |  |
| --- | --- | --- | --- |
| **Rule No.** | **Frequent itemset** | **Confidence** | **Lift** |
| 1 | Apple Cereal | 100% | 1.33 |
| 2 | Beer Eggs | 100% | 1.33 |
| 3 | Eggs Beer | 100% | 1.33 |
| 4 | Beer, Cereal Eggs | 100% | 1.33 |

**ตารางที่ 2-8 (ต่อ)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rule No.** | **Frequent itemset** | **Confidence** | **Lift** |
| 5 | Cereal, Eggs Beer | 100% | 1.33 |
| 6 | Cereal Apple | 67% | 1.33 |
| 7 | Beer Cereal, Eggs | 67% | 1.33 |
| 8 | Eggs Beer, Cereal | 67% | 1.33 |
| 9 | Beer Cereal | 67% | 0.89 |
| 10 | Cereal Beer | 67% | 0.89 |
| 11 | Cereal Eggs | 67% | 0.89 |
| 12 | Eggs Cereal | 67% | 0.89 |
| 13 | Cereal Beer, Eggs | 67% | 0.89 |
| 14 | Beer, Eggs Cereal | 67% | 0.89 |

**2.1.6 งานวิจัยที่เกี่ยวข้อง**

จากการสำรวจพบว่ามีงานวิจัยที่มีการใช้เทคนิคการทำเหมืองข้อมูลในการเลือกภาควิชาที่เหมาะสมให้กับนิสิต (ชิดชนก ส่งศิริ, ธนาวินท์ รักธรรมานนท์ และกฤษณะ ไวยมัย, 2544) นั้นสามารถทำได้หลายแนวทางแนวทางหนึ่งคือการใช้ classification โดยสร้างโมเดลกลาง (รวมข้อมูลจากทุกภาควิชา) เพื่อทำนายหาภาควิชาที่เหมาะสมที่สุดสำหรับนิสิตโดยแต่ละกิ่งของ tree แสดงปัจจัยต่างๆที่มีผลต่อการเลือกภาควิชาของนิสิตและลีฟโหนด(คลาสปลายทาง) แทนภาควิชาต่างๆเมื่อมีนิสิตคนหนึ่งต้องการทราบว่าตนเองเหมาะสมกับภาควิชาใดก็จะนำลักษณะต่างๆของตนไปเทียบกับในโมเดลนี้ตามเส้นทางใน tree จนกระทั่งได้คลาสปลายทางซึ่งก็คือภาควิชาที่ตรงกับลักษณะตามต้องการวิธีการสร้างโมเดลของแต่ละภาควิชาสร้างโดยพิจารณาถึงปัจจัยที่มีผลต่อการเลือกภาควิชาของนิสิตมากที่สุดตามลำดับแล้วนำมาสร้างเป็นเส้นทางเพื่อไปสู่คลาสปลายทางว่าเป็นนิสิตที่เหมาะสมกับภาควิชานั้นหรือไม่โดยในแต่ละกิ่งแทนปัจจัยที่มีผลต่อการเลือกภาควิชาของนิสิตดังที่ได้กล่าวมาแล้วและลีฟโหนด (คลาสปลายทาง) มี 2 คลาสคือ GOOD และ BAD โดยมีเกณฑ์ว่า GOOD คือนิสิตที่มีเกรดเฉลี่ยสะสมอยู่ในอันดับ 40% แรกของแต่ละภาควิชาและ BAD คือนิสิตที่มีเกรดเฉลี่ยสะสมอยู่ในอันดับ 40% สุดท้ายในแต่ละภาควิชาโดยเปอร์เซ็นต์นี้สามารถปรับเปลี่ยนให้เหมาะสมได้ตามลักษณะของข้อมูล ในการวิจัยนี้พบว่าที่ 40% เป็นอัตราส่วนที่เหมาะสมที่สุดเพราะจะได้ข้อมูลที่นำมาเป็น training data มากและมีการเว้นช่วงผิดพลาดที่สามารถเกิดได้ในช่วงกลางของนิสิต(ช่วงระหว่างเกรด GOOD และ BAD) ดังนั้นในงานวิจัยนี้จะต้องสร้างโมเดล classification ให้มีจำนวนโมเดลเท่ากับจำนวนภาควิชาโดยสร้างโมเดลจาก training data แบ่งมาจากข้อมูลจำนวน 60% จากข้อมูลนิสิตที่มีอยูทั้งหมดในภาควิชานั้นๆ (ประมาณ 4800 คน) และทดสอบด้วย testing data ที่ได้มาจากข้อมูลนิสิตอีก 40% หลังจากการแบ่งข้อมูลไปทำ training data แล้ว (ประมาณ 3,200 คน) มาทดสอบในโมเดลที่ได้สร้างไว้แล้ววัดความถูกต้องของโมเดลจากการนำผลลัพธ์ที่ได้จากการนำ testing data มาผ่านโมเดลนี้เปรียบเทียบผลลัพธ์ที่ทราบผลอยู่แล้วใน testing data ว่าเป็นนิสิตประเภท GOOD หรือ BAD แล้วนำมาวัดเป็นเปอร์เซ็นต์ความถูกต้อง

ดังนั้นเมื่อเราต้องการทราบภาควิชาที่เหมาะสมของนิสิตคนหนึ่งเราจะพิจารณาลักษณะต่างๆของนิสิตกับโหนดนั้นใน tree แล้วลงมาตามทางของ decision tree ที่ตรงกับลักษณะของนิสิตจนกระทั่งถึงคลาสปลายทางสำหรับบางคลาสปลายทางนั้นโมเดลที่ได้อาจแบ่งแยกความถูกต้องไม่ได้100% ดังนั้นถ้าคลาสปลายทางที่ได้ประกอบด้วยอัตราส่วน GOOD มากกว่า BAD มากๆย่อมแสดงว่านิสิตที่มีลักษณะเช่นนี้เข้าภาควิชาวิศวกรรมไฟฟ้าแล้วจะมีผลการเรียนดี (GOOD) มากกว่าผลการเรียนไม่ดี (BAD) ดังนั้นภาควิชาไฟฟ้านี้เป็นภาควิชาหนึ่งที่นิสิตคนนี้ควรพิจารณาเราจะทดสอบข้อมูลนิสิตรายหนึ่งๆกับโมเดลของทุกภาควิชาและเลือกเฉพาะภาควิชาที่คลาสปลายทางมีอัตราส่วนของ GOOD มากกว่า BAD ถ้าผลออกมาว่ามีหลายภาควิชาที่เหมาะสมเราสามารถเลือกภาควิชาที่ดีที่สุดได้โดยพิจารณา จากอัตราส่วนของ GOOD ในโมเดลที่มากกว่าเป็นหลักหลังการทดสอบพบว่าโมเดลการจัดประเภทนิสิตในทุกภาควิชามีประสิทธิภาพค่อนข้างดีโดยทุกโมเดลมีความถูกต้องมากกว่า 80% ซึ่งจากวิธีการและผลลัพธ์ดังกล่าวสามารถนำไปใช้งานได้จริง

สำหรับในงานวิจัยนี้จะทําการหาปัจจัยที่ส่งผลต่อผลการเรียนของนักศึกษาเรียนอ่อน (ณัฐธิดา สุวรรณโน และอันธิกา สิงห์เอี่ยม, 2554) โดยใช้เทคนิคกฎความสัมพันธ์ซึ่งเป็นเทคนิคในการทําเหมืองข้อมูลเพื่อนํามาทํานายความเสี่ยงของการเรียนอ่อนของนักศึกษาชั้นปีที่ 1 ในคณะต่างๆ ของมหาวิทยาลัยสงขลานครินทร์ งานวิจัยนี้ใช้ข้อมูลของนักศึกษาที่เก็บไว้แล้วในฐานข้อมูลของมหาวิทยาลัย ตั้งแต่ปี พ.ศ. 2550-2552 มาศึกษาโดยผ่านกระบวนการต่างๆของการทําเหมืองข้อมูลและใช้เทคนิคกฎความสัมพันธ์เพื่อหารูปแบบของข้อมูล จากผลการวิจัยพบว่าปัจจัยที่ส่งผลต่อผลการศึกษาของนักศึกษาในแต่ละคณะแตกต่างกันและความถูกต้องของผลลัพธ์ก็แตกต่างกันอันเนื่องมาจากข้อมูลนําเข้าของนักศึกษาในแต่ละคณะต่างกัน ดังนั้นผลลัพธ์และความถูกต้องของกฎความสัมพันธ์จึงขึ้นอยู่กับข้อมูลที่ใช้ในการเรียนรู้เมื่อมองในภาพรวมแล้วพบว่าปัจจัยที่ส่งผลต่อผลการศึกษาของนักศึกษาเรียนอ่อนคือ วิธีการเข้าศึกษา คะแนนสอบเข้าศึกษาของวิชาต่างๆ และเพศของนักศึกษา

จากผลของงานวิจัยได้มีการพัฒนาโปรแกรมต้นแบบเพื่อทํานายความเสี่ยงต่อการเรียนอ่อนของนักศึกษาในแต่ละคณะโดยใช้ปัจจัยที่ส่งผลต่อผลการเรียนของนักศึกษาแต่ละคณะมาทําการพัฒนาผลของงานวิจัยฉบับนี้สามารถนําไปใช้ในการทํานายความเสี่ยงที่จะเรียนอ่อนของนักศึกษาเพื่อช่วยป้องกันปัญหาทางผลการเรียนได้จากการประมวลผลด้วยเทคนิคกฎความสัมพันธ์ด้วยโปรแกรม Weka พบว่าผลลัพธ์ของกฎความสัมพันธ์สามารถมีได้หลายกฎและผลลัพธ์ของกฎนั้นๆ สามารถนําไปเป็นข้อมูลนําเข้าของอีกกฎหนึ่งได้แต่ในงานวิจัยนี้มุ่งเน้นที่จะหาความสัมพันธ์ของข้อมูลที่ส่งผลต่อผลการเรียนของนักศึกษาเรียนอ่อนจึงคัดเลือกผลลัพธ์ของกฎที่มีสถานภาพของนักศึกษาเป็นเรียนอ่อนเท่านั้น โดยผลลัพธ์ที่ได้เกิดกฎความสัมพันธ์เป็นจํานวนมาก ดังนั้นเพื่อเป็นการลดจํานวนกฎความสัมพันธ์จึงกําาหนดค่าความเชื่อมั่นต่ำสุด และค่าสนับสนุนต่ำสุดนั่นหมายความว่าจะสนใจเฉพาะกฎความสัมพันธ์ที่มีค่าความ เชื่อมั่นและค่าสนับสนุนมากกว่าหรือเท่ากับความเชื่อมั่นต่ำสุด และค่าสนับสนุนต่ำสุดและกําหนดว่ากฎความสัมพันธ์ที่เกิดขึ้นนั้นจะต้องไม่ซ้ำกันดังนั้นแต่ละเงื่อนไขสามารถมีผลลัพธ์ได้เพียงเงื่อนไขเดียวเท่านั้น

จากการวิจัยพบว่าสามารถใช้เทคนิคกฎความสัมพันธ์ซึ่งเป็นเทคนิคในการทําเหมืองข้อมูลมาประยุกต์ใช้เพื่อค้นหารูปแบบของข้อมูลนักศึกษาที่มีความเสี่ยงที่จะประสบปัญหาการเรียนอ่อนได้ โดยจากการนําข้อมูลของนักศึกษาชั้นปีที่ 1 มหาวิทยาลัยสงขลานครินทร์ในคณะต่างๆทั้งหมด 14หลักสูตรมาเป็นข้อมูลนําเข้านั้นพบว่าปัจจัยที่มีผลต่อการเรียนอ่อนของนักศึกษาแต่ละคณะมีปัจจัยที่แตกต่างกันแต่เมื่อนํามาเปรียบเทียบในภาพรวมพบว่าปัจจัยส่วนใหญ่ที่ส่งผลให้นักศึกษามีความเสี่ยงที่จะเรียนอ่อนคือ วิธีการเข้าศึกษา คะแนนการสอบเข้าศึกษาของวิชาต่างๆ และเพศของนักศึกษา ดังนั้นหากผู้ที่เกี่ยวข้องนําผลจากการวิจัยชิ้นนี้ไปใช้เพื่อป้องกันไม่ให้เกิดการเรียนอ่อนด้วยการหาความเสี่ยงของนักศึกษาชั้นปีที่ 1 แล้วให้คําแนะนําและดูแลนักศึกษากลุ่มที่มีค่าความเสี่ยงที่จะเรียนอ่อนสูงก็จะช่วยให้นักศึกษาวางแผนการศึกษาและลดความเสี่ยงที่จะเรียนอ่อนได้ อย่างไรก็ตามผลจากงานวิจัยชิ้นนี้ใช้ข้อมูลของนักศึกษาในปีการศึกษา 2550 - 2552 แต่ในปีการศึกษา 2553 เป็นต้นมาระบบการรับเข้ามหาวิทยาลัยได้มีการเปลี่ยนแปลงการสอบเข้า เช่น มีการสอบวิชา GAT, PAT เพิ่มขึ้น ดังนั้นผลของงานวิจัยนี้อาจจะมีการเปลี่ยนแปลงหากใช้กับนักศึกษาในรุ่นตั้งแต่ปีการศึกษา 2553 ดังนั้นหากมีการศึกษาเพิ่มเติมจากงานวิจัยนี้ก็จะเป็นประโยชน์ต่อนักศึกษาในรุ่นถัดๆไป

และสำหรับการวิจัยเรื่องการค้นหาความรู้คุณลักษณะสําคัญของนักศึกษาที่มีต่อผลสัมฤทธิ์ทางการเรียนในกลุ่มโปรแกรมคอมพิวเตอร์ มหาวิทยาลัยราชภัฏนครปฐมด้วยเทคนิคเหมืองข้อมูล (ไก้รุ่ง เฮงพระพรหม, สุพจน์ เฮงพระพรหม และสุวิมล มรรควิบูลย์, 2557) นี้มีวัตถุประสงค์เพื่อค้นหาความรู้คุณลักษณะสําคัญของนักศึกษาที่มีต่อผลสัมฤทธิทางการเรียนในกลุ่มโปรแกรมคอมพิวเตอร์ มหาวิทยาลัยราชภัฏนครปฐมด้วยเทคนิคเหมืองข้อมูล เพื่อสร้างโมเดลหรือต้นแบบเพื่อการพัฒนาระบบสนับสนุนการตัดสินใจการเข้าศึกษาสําหรับผู้สนใจเรียนในสาขาวิชาของกลุ่มโปรแกรมคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม โดยได้ทําการศึกษาข้อมูลจากนักศึกษาที่ศึกษาอยู่ในสาขาคอมพิวเตอร์ชั้นปีที่ 3 และ 4 ในปัจจุบัน โดยใช้ระดับผลการเรียนเฉลี่ย ณ ปัจจุบันเป็นเกณฑ์ในการจําแนกกลุ่มนักศึกษา

ในการศึกษาผู้วิจัยได้ทําการทดลองกับเทคนิคทางเหมืองข้อมูล 2 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ J48 (C4.5) และ เทคนิคโครงข่ายเบย์ (BayesNet) โดยแบ่งการทดลองออกเป็น 2 ลักษณะ คือ การทดลองกับชุดข้อมูลทั้งหมดทุกแอททริบิวท์และการทดลองด้วยการเลือกคุณลักษณะด้วยเทคนิค SNR (Signal to Noise Ratio) เพื่อเพิ่มประสิทธิภาพในการทํางานของอัลกอริทึมซึ่งจากผลการทดลองทั้งสองแบบยืนยันว่าเทคนิคต้นไม้ตัดสินใจ J48 (C4.5) ให้ประสิทธิภาพในการทํางานที่ดีกว่าเมื่อเทียบกับเทคนิคโครงข่ายเบย์ (BayesNet)