# การพัฒนากฎการทำนายผลการเรียนของนักศึกษาชั้นปีที่ 1 โดยใช้เทคนิคการทำเหมืองข้อมูล

# Development of Rules for Predicting Academic Learning Outcomes of Freshmen Students using Data Mining Techniques

พรพิมล ชัยวุฒิศักดิ์[1] และ ยุวดี กล่อมวิเศษ[1]

## บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำความรู้การทำเหมืองข้อมูลมาวิเคราะห์ผลการเรียนของนักศึกษาในรายวิชาต่างๆ ของแผนการศึกษาชั้นปีที่ 1 ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง และสร้างกฎความสัมพันธ์ระหว่างผลการเรียนของรายวิชาและเกรดเฉลี่ยสะสมต่ำกว่า 2.00 โดยใช้กฎความสัมพันธ์ (Association Rules) ด้วยอัลกอริทึมอพริโอริ (Apriori algorithm) และกฎการตัดสินใจสำหรับจำแนกข้อมูล (Data Classification) ด้วยเทคนิค J48 เพื่อจะได้นำมาวางแผนการเรียนของนักศึกษา จากการศึกษาพบว่ากฎที่ใช้ในการจำแนกผลการเรียนของนักศึกษาชั้นปีที่ 1 กลุ่มที่เกรดเฉลี่ยสะสมต่ำกว่า 2.00 และ กลุ่มที่ได้เกรดเฉลี่ยสูงกว่า 2.00 ด้วยเทคนิค J48 ให้ค่าความถูกต้องสูงถึง 91% และจำนวนกฎความสัมพันธ์ของรายวิชาที่มีผลต่อเกรดเฉลี่ยสะสมต่ำกว่า 2.00 ของนักศึกษาชั้นปีที่ 1 มีจำนวนเท่ากับ 5 ด้วยความเชื่อมั่นที่ 1.00 และ ค่าสหสัมพันธ์มากกว่า 1.00

**คำสำคัญ:** ผลการเรียน กฎความสัมพันธ์ การจำแนกข้อมูล การทำเหมืองข้อมูล การทำนาย

## ABSTRACT

In this research investigation, the researchers apply the knowledge of data mining to analyze the academic learning outcomes of the students in various courses in the first year study plan of the Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang. The researchers also establish the relationship rules between the learning outcomes of the courses taken with a grade point average (GPA) lower than 2.00 using the association rules with a priori algorithm and the decision rules for data classification with the technique of J48. The results could be used to make a plan for the study of the students. The study found that the rules used for the classification of the learning outcomes of the students under study with a GPA lower than 2.00 and those with a GPA higher than 2.00 using the J48 technique yielded the accuracy of 91 percent. The number of the association rules of the courses that affected a GPA lower than 2.00 of the students under study was 5 with the reliability at 1.00 and Lift>1.00.

**Keywords:** academic learning outcome, association rules, data classification prediction, data mining

[1] อาจารย์ประจำภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

## บทนำ

การทำเหมืองข้อมูล (Data Mining) เป็นการวิเคราะห์และสืบค้นองค์ความรู้ หรือสิ่งที่สำคัญออกมาจากข้อมูลจำนวนมาก โดยได้มีการนำเทคนิคการทำเหมืองข้อมูลไปประยุกต์ใช้ในงานด้านต่างๆ มากขึ้น เช่น การวิเคราะห์เครดิตลูกค้าของบริษัทสำหรับการให้สินเชื่อ (ทิพย์ธิดา, 2556) การวิเคราะห์ข้อมูลทางการพยาบาล (ศุภามณ, 2561) การจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ (จิราภาและคณะ, 2558) การวิเคราะห์รายวิชาที่มีผลต่อการพ้นสภาพของนักศึกษา (บุษราภรณ์และคณะ, 2559) รวมทั้งการนำมาใช้ในการเกิดประโยชน์ในการสืบค้นสิ่งที่น่าสนใจออกมาจากข้อมูลการศึกษาของนักศึกษา

การหากฎความสัมพันธ์เป็นกระบวนการหนึ่งในการทำเหมืองข้อมูลที่ได้รับความนิยมมาก โดยจะใช้กฎความสัมพันธ์ในการหาความสัมพันธ์ของข้อมูลสองชุดหรือมากกว่าสองชุดขึ้นไปภายในฐานข้อมูลที่มีขนาดใหญ่ โดยที่อัลกอริทึมอพริโอริ (Apriori Algorithm) เป็นการหากฎความสัมพันธ์ (Association Rule) พื้นฐานโดยการสร้างรูปแบบของเซตไอเทม (Itemset) ที่มีค่ามากกว่าค่าสนับสนุนต่ำสุด (Minimum Support) ที่ได้กำหนดไว้ โดยทำการสร้างรูปแบบของเซตไอเทม (Itemset) ที่มีขนาดยาวมากขึ้นทีละหนึ่งขั้นไปเรื่อยๆ จนกระทั่งไม่เหลือเซตไอเทม (Itemset) ที่จะสร้างอีกต่อไป ทำให้ได้กลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยทั้งหมด (Agrawal, et al., 1993)

ในขณะที่ต้นไม้ตัดสินใจเป็นเทคนิคการจำแนกประเภท (Classification) ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning) โดยอัลกอริทึม J48 หรือ C4.5 พัฒนาโดย Quinlan ในปี ค.ศ. 1993 (Quinlan, 1993) J48 ขยายมาจาก ID3 ซึ่งอาศัยหลักการของ Information Gain หรือ Entropy Reduction สำหรับจำแนกกิ่ง (node) ของต้นไม้ (tree) เกณฑ์ที่ใช้ช่วยประกอบการเลือกแอตทริบิวต์ (Attribute) คือ ทดลองเลือกแต่ละแอตทริบิวต์ (Attribute) มาทำหน้าที่เป็น Root Node และวัดค่า Gain ซึ่งเป็นค่าที่ชี้ว่าแอตทริบิวต์ (Attribute) นั้นจะช่วยจำแนกคลาส (Class) ของข้อมูลได้ดีเพียงใด โดยการจำแนกที่ดีที่สุดคือให้ Leaf Node ที่เป็นข้อมูลเดียวกันทั้งหมด และค่า Gain ที่สูงที่สุดหมายถึงการจำแนกคลาส

จากข้อมูลสถิติของสำนักทะเบียนและประมวลผลของภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง พบว่าในปีการศึกษา 2561 จำนวนนักศึกษาชั้นปีที่ 1 ที่สอบได้เกรดเฉลี่ยสะสมต่ำกว่า 2 มีจำนวน 19 คน จากจำนวนทั้งหมด 115 คน คิดเป็นร้อยละ 16.52 ทำให้นักศึกษาเหล่านั้นมีโอกาสที่จะพ้นสภาพหรือสำเร็จการศึกษาช้ากว่ากำหนด นับการสูญเสียเวลาและค่าใช้จ่ายของนักศึกษาและสถาบันการศึกษา

ดังนั้นวัตถุประสงค์ในการศึกษานี้เพื่อหารูปแบบกฎความสัมพันธ์ของรายวิชาที่มีผลต่อประสิทธิภาพการเรียนของนักศึกษาชั้นปีที่ 1 โดยใช้เทคนิคต้นไม้ตัดสินใจและอัลกอริทึมอพริโอริ (Apriori Algorithm) โดยข้อมูลนำมาศึกษาในครั้งนี้เป็นข้อมูลของนักศึกษาภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง โดยประกอบไปด้วยข้อมูลผลการเรียนในวิชาต่างๆ ของชั้นปีที่ 1 ส่วนที่ 2 จะกล่าวถึงวิธีการดำเนินการวิจัย กรอบการดำเนินการวิจัย ข้อมูลและเครื่องมือที่ใช้ในการวิจัย รายละเอียดของแนวคิดและเทคนิคการทำเหมืองข้อมูล ได้แก่ การหากฎความสัมพันธ์ (Association Rule Discovery) และการจำแนกข้อมูล(Data Classification) ในส่วนที่ 3 แสดงกฎความสัมพันธ์ และกฎการทำนายผลการเรียนของนักศึกษาชั้นปีที่ 1 ส่วนที่ 5 เป็นการสรุปงานวิจัยและแนวทางการปรับปรุงงานวิจัยในอนาคต

## วิธีดำเนินการวิจัย

### 1. กรอบการดำเนินการวิจัย

การดำเนินการวิจัยนี้อาศัยกระบวนการมาตรฐานในการวิเคราะห์ข้อมูลการทำเหมืองข้อมูลซึ่งพัฒนาขึ้นในปี ค.ศ.1996 โดยความร่วมมือของ 3 บริษัท คือ DaimlerChrysler SPSS และ NCR กระบวนการทำงานนี้เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือเรียกย่อว่า

“CRISP-DM” (Shearer, 2000) โดยประกอบด้วย 6 ขั้นตอนทำต่อเนื่องกัน ดังภาพที่ 1 คือ

1) Business Understanding เป็นขั้นตอนการเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการทำเหมืองข้อมูล

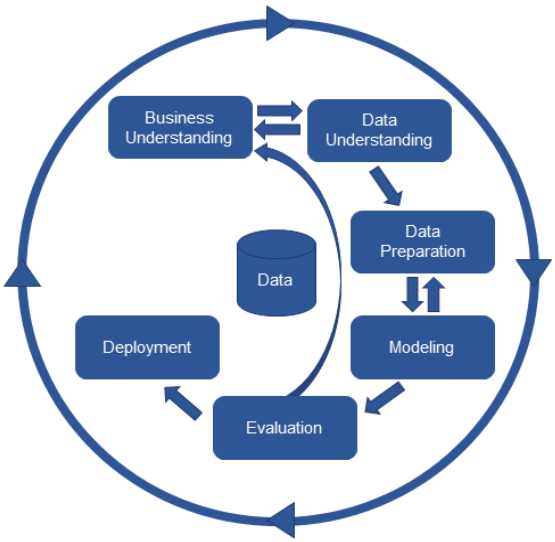
2) Data Understanding เป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์

3) Data Preparation เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (Raw Data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยต้องมีการทำข้อมูลให้ถูกต้อง (Data Cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (Scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป (Missing Value) เป็นต้น

4) Modeling เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทำเหมืองข้อมูล ในที่นี้อาศัยเทคนิคการหากฎความสัมพันธ์ และ การจำแนกข้อมูล ในบางครั้งอาจจะต้องมีการย้อนกลับไปที่ขั้นตอน Data Preparation เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย

5) Evaluation เป็นขั้นตอนที่วัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้กำหนดไว้ในขั้นตอนแรก เพื่อจะได้เปลี่ยนแปลงแก้ไขเพื่อให้ได้ผลลัพธ์ตามที่ต้องการได้

6) Deployment เป็นการนำเอาองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริงในองค์กรหรือบริษัท




**ภาพที่ 1** กระบวนการมาตรฐานในการทำเหมืองข้อมูล

## 2. ข้อมูลที่ใช้ในการวิจัย

ในงานวิจัยนี้จะพิจารณาข้อมูลผลการเรียนของนักศึกษา ภาควิชาสถิติ ในแต่ละรายวิชาของชั้นปีที่ 1 ตามแผนการศึกษาหลักสูตรวิทยาศาสตรบัณฑิต(สถิติประยุกต์) ระหว่างปีการศึกษา 2559 – 2561 จำนวน 383 ระเบียน ซึ่งสามารถแบ่งรายวิชาออกเป็นรายวิชาในกลุ่มวิชาต่างๆ ดังนี้

1) กลุ่มวิชาแกน ประกอบด้วยรายวิชาดังนี้
   - รหัสวิชา 05016201 แคลคูลัส 1
   - รหัสวิชา 05016202 แคลคูลัส 2
   - รหัสวิชา 05016213 พีชคณิตเชิงเส้น 1
   - รหัสวิชา 05106030 เคมีทั่วไป
   - รหัสวิชา 05206500 ชีววิทยาทั่วไป
   - รหัสวิชา 05306003 ฟิสิกส์ทั่วไป
   - รหัสวิชา 05406002 หลักสถิติ
2) กลุ่มวิชาบังคับ ประกอบด้วยรายวิชาดังนี้
   - รหัสวิชา 05406004 สถิติวิเคราะห์
   - รหัสวิชา 05506003 การเขียนโปรแกรมขั้นพื้นฐาน
   - รหัสวิชา 05506004 การเขียนโปรแกรมเชิงออบเจกต์

3) กลุ่มวิชาภาษาศาสตร์ ประกอบด้วยรายวิชาดังนี้

- รหัสวิชา 90201001 ภาษาอังกฤษพื้นฐาน 1
- รหัสวิชา 90201002 ภาษาอังกฤษพื้นฐาน 2

**3. เครื่องมือที่ใช้ในการวิจัย**

**3.1 โปรแกรม Microsoft Excel 2010**

เป็นโปรแกรมทางด้านตารางคำนวณ หรือที่เรียกว่า เสปรตชีต (Spreadsheet) เป็นโปรแกรมในชุด Microsoft Office มีความสามารถในด้านการสร้างตาราง การคำนวณ การวิเคราะห์ การออกรายงานในรูปแบบตารางและกราฟ โดยนำมาใช้ในการเตรียมข้อมูล

**3.2 โปรแกรม Weka 3.7**

โปรแกรมประยุกต์ RapidMiner เกิดขึ้นในปี 2006 จาก Ingo Mierswa and Ralf Kilnkenberg ผลสำรวจจาก Kdnuggets ปี 2014 พบว่า เป็นอันดับ 1 ในกลุ่มโปรแกรมด้านการวิเคราะห์ข้อมูล และ Gartner จัดให้ RapidMiner อยู่ในกลุ่ม Leaders ในปี 2015 ร่วมกับ ซอฟต์แวร์กลุ่มนี้ คือ SAS, IBM, KNIME และ RapidMiner ตัวอย่างบริษัทที่ใช้ RapidMiner บริษัท Paypal จำกัด การพัฒนาด้าน Business Intelligence เป็นการวิเคราะห์ข้อมูลในคลังข้อมูลเพื่อช่วยสนับสนุนการตัดสินใจในการบริหารจัดการองค์กร ประกอบด้วยระบบจัดการข้อมูลและโปรแกรมแอพพลิเคชั่นด้านการวิเคราะห์ข้อมูล (สายชล, 2558)

**4. เทคนิคการทำเหมืองข้อมูล**

**4.1 การหากฎความสัมพันธ์ (Association Rule)**

เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูล (Data Mining) โดยการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปหารูปแบบที่เกิดขึ้นบ่อยๆ (Frequent Pattern) ผลลัพธ์ที่ได้เป็นกฎความสัมพันธ์ (Association Rule) สามารถเขียนได้ในรูปเซตของรายการที่เป็นเหตุ ไปสู่เซตของรายการที่เป็นผล ซึ่งมีรากฐานมาจากการวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) เช่น ลูกค้าที่ซื้อผ้าอ้อมส่วนใหญ่จะซื้อเบียร์ด้วย ข้อมูลที่นำมาใช้จะอยู่ในรูปแบบ Nominal หรือ Ordinal เท่านั้น

การหากฎความสัมพันธ์ด้วยอัลกอริทึมอพริโอริ (Apriori Algorithm) ซึ่งจะมีอยู่ 2 ขั้นตอนใหญ่ๆ คือ

1. การหาเซตไอเทม (itemset) เป็นการหารูปแบบของข้อมูลที่เกิดขึ้นร่วมกันบ่อยๆ ในฐานข้อมูลหรือมากกว่าค่า minimum support ที่กำหนดในขั้นตอนนี้จะแบ่งได้อีกเป็น 2 ขั้นตอนย่อย คือ

1.1. การสร้างรูปแบบของเซตไอเทม (itemset) โดยการ join จะใช้รูปแบบของเซตไอเทม (itemset) ที่มีค่ามากกว่า minimum support มาทำการสร้างรูปแบบของเซตไอเทม (itemset) ที่มีขนาดยาวมากขึ้นทีละหนึ่งขั้นไปเรื่อยๆ

1.2. การนับค่า support หลังจากที่สร้างรูปแบบของเซตไอเทม (itemset) ได้แล้ว ขั้นถัดมาจะทำการคำนวณค่า support ที่เกิดขึ้น โดยที่ support คือ จำนวนเปอร์เซ็นต์ที่พบเซตไอเทม (itemset) ในฐานข้อมูล (Database)

2. การสร้างกฎความสัมพันธ์ (Association Rule) หลังจากที่หาเซตไอเทม (itemset) ได้แล้วจะนำรูปแบบที่หาได้มาสร้างเป็นกฎความสัมพันธ์

การค้นหากฎความสัมพันธ์นี้มีเกณฑ์ในการวัดความน่าสนใจ 3 แบบ ได้แก่

- ค่าสนับสนุน (Support) แสดงถึงเปอร์เซ็นต์ของข้อมูลที่เป็นไปตามกฎจากข้อมูลทั้งหมดดังสมการ (1)

$$\text{Support (LHS,RHS)} = \frac{\text{จำนวนรายการโดยประกอบไอเทมLHS และ RHS}}{\text{จำนวนรายการใน Database}} \quad (1)$$

- ค่าความมั่นใจ (Confidence) แสดงถึงความน่าเชื่อถือของกฎ เมื่อรูปแบบไอเทมเซตด้านซ้ายของกฎความสัมพันธ์ (Left Hand Side: LHS) เกิดขึ้นแล้วมีโอกาสเกิดรูปแบบไอเทมเซตด้านขวาของกฎความสัมพันธ์ (Right Hand Side: RHS) มากน้อยเท่าใด ซึ่งจะมีค่าอยู่ระหว่าง 0 - 1 ถ้าใกล้เคียง 1 หมายถึงมีความเชื่อมั่นในการหาความสัมพันธ์มาก ดังสมการ (2)

$$\text{Confidence (LHS} \rightarrow \text{RHS)} = \frac{\text{Support(LHS,RHS)}}{\text{Support(LHS)}} \quad (2)$$

- ค่าสหสัมพันธ์ หรือเรียกว่าค่าลิฟต์ (Lift) คือค่าที่บ่งบอกว่าการเกิดรูปแบบไอเทมเซตด้านซ้ายของกฎความสัมพันธ์ (Left Hand Side: LHS) และรูปแบบไอเทมเซตด้านขวาของกฎความสัมพันธ์ (Right Hand Side: RHS) มีความสัมพันธ์กันมากหรือไม่ โดยถ้าค่าลิฟต์ มีค่าเท่ากับ 1 แสดงว่ารูปแบบ LHS และ RHS ไม่ขึ้นต่อกัน (Independent) แต่ถ้ามีค่ามากกว่า 1 มาก ๆ แสดงว่ากฎทั้งสองมีความสัมพันธ์กันมากด้วย ดังสมการ (3)

$$\text{Lift (LHS} \rightarrow \text{RHS)} = \frac{\text{Support(LHS,RHS)}}{\text{Support(LHS)} \times \text{Support(RHS)}} \quad (3)$$

เราจะพิจารณาเฉพาะความสัมพันธ์ที่มีค่าสนับสนุนและค่าความมั่นใจสูงกว่าค่าสนับสนุนต่ำสุด (Minimum Support) และค่าความมั่นใจต่ำสุด (Minimum Confidence)

**4.2 การจำแนกข้อมูล(DataClassification)**

เทคนิคการจำแนกประเภทข้อมูลเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลสอนระบบ (Training data) ที่แต่ละแถวของข้อมูลประกอบด้วยฟิลด์หรือแอตทริบิวต์อาจเป็นค่าต่อเนื่อง (Continuous) หรือค่ากลุ่ม (Categorical) โดยจะมีแอตทริบิวต์แบ่ง (Classifying Attribute) ซึ่งเป็นตัวบ่งชี้คลาสของข้อมูล จุดประสงค์ของการจำแนกประเภทข้อมูลคือการสร้างโมเดลการแยกแอตทริบิวต์หนึ่งโดยขึ้นกับแอตทริบิวต์อื่นๆ โดยตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกข้อมูล ได้แก่

1) ค่าความถูกต้อง (Accuracy) เป็นค่าที่ใช้วัดประสิทธิภาพในการจำแนกหรือเป็นตัววัดขนาดของความผิดพลาด หากค่าความถูกต้องมีค่ามากจะมีความผิดพลาดน้อย ดังสมการ (4)

$$\text{ความถูกต้อง} = \frac{\text{จำนวนข้อมูลที่ทำนายถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมด}} \quad (4)$$

2) ค่าความแม่นยำ (Precision) เป็นอัตราส่วนของการทำนายข้อมูลในคลาสได้ถูกต้องจากจำนวนข้อมูลทั้งหมดในคลาสนั้น ดังสมการ (5)

$$\text{ความแม่นยำ} = \frac{\text{จำนวนข้อมูลที่ทำนายถูกต้องในคลาส}}{\text{จำนวนข้อมูลทั้งหมดที่ทำนายได้ในคลาสนั้น}} \quad (5)$$

3) ค่าความระลึก (Recall) เป็นอัตราส่วนของการค้นพบคลาสที่ใกล้เคียงกับคำขอและมีการค้นคืนให้กับผู้ใช้กับเอกสารที่ตรงตามคำขอทั้งหมดแม่นยำ ดังสมการ (6)

$$\text{ความระลึก} = \frac{\text{จำนวนข้อมูลที่ทำนายถูกต้องในคลาสนั้น}}{\text{จำนวนข้อมูลที่ถูกต้องทั้งหมด}} \quad (6)$$

4) ค่าความถ่วงดุล (F-Measure) เป็นค่าที่แสดงความสัมพันธ์ระหว่างค่าความแม่นยำและค่าความระลึกเพื่อหาค่าความถ่วงดุล โดยค่าที่ได้จากการคำนวณจะมีค่าอยู่ระหว่าง 0 ถึง 1 ถ้าค่าที่คำนวณได้เข้าใกล้ 1 หมายความว่า การให้ผลในการจำแนกมีประสิทธิภาพสูง และถ้าค่าคำนวณได้เข้าใกล้ 0 หมายความว่า การให้ผลในการจำแนกมีประสิทธิภาพต่ำ ดังสมการ (7)

$$\text{ความถ่วงดุล} = \frac{2(\text{ความแม่นยำ})(\text{ความระลึก})}{\text{ความแม่นยำ}+\text{ความระลึก}} \quad (7)$$

การสร้างต้นไม้ตัดสินใจ C4.5 ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ถ้าให้ชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้คือ $\{m_1, m_2, \ldots, m_n\}$ และให้ความน่าจะเป็นที่จะเกิดค่า $m_i$ เท่ากับ $P(m_i)$ จะได้ว่าค่าเกนสารสนเทศ (Information Gain) ของ M เขียนแทนด้วย I(M) ดังสมการ (8)

$$I(M) = \sum_{i=1}^{n} -P(M_i) \log_2 P(M_i) \quad (8)$$

กำหนดให้ T แทนชุดตัวอย่างข้อมูลสำหรับใช้สร้างตัวแบบและคุณลักษณะที่เป็นโหนด คือ x และมีค่าทั้งหมดที่เป็นไปได้ n ค่าโหนดปัจจุบันจะแบ่งตัวอย่าง T ออกตามกิ่งเป็น $\{t_1, t_2, \ldots, t_n\}$ ตามค่าที่เป็นไปได้ของ x ดังนั้นจึงสามารถคำนวณค่าเกนสารสนเทศหลังจากแบ่งตามคุณลักษณะ x ได้ดังสมการ (9)

$$\text{Gain}(x) = I(T) - I_n(T) \quad (9)$$

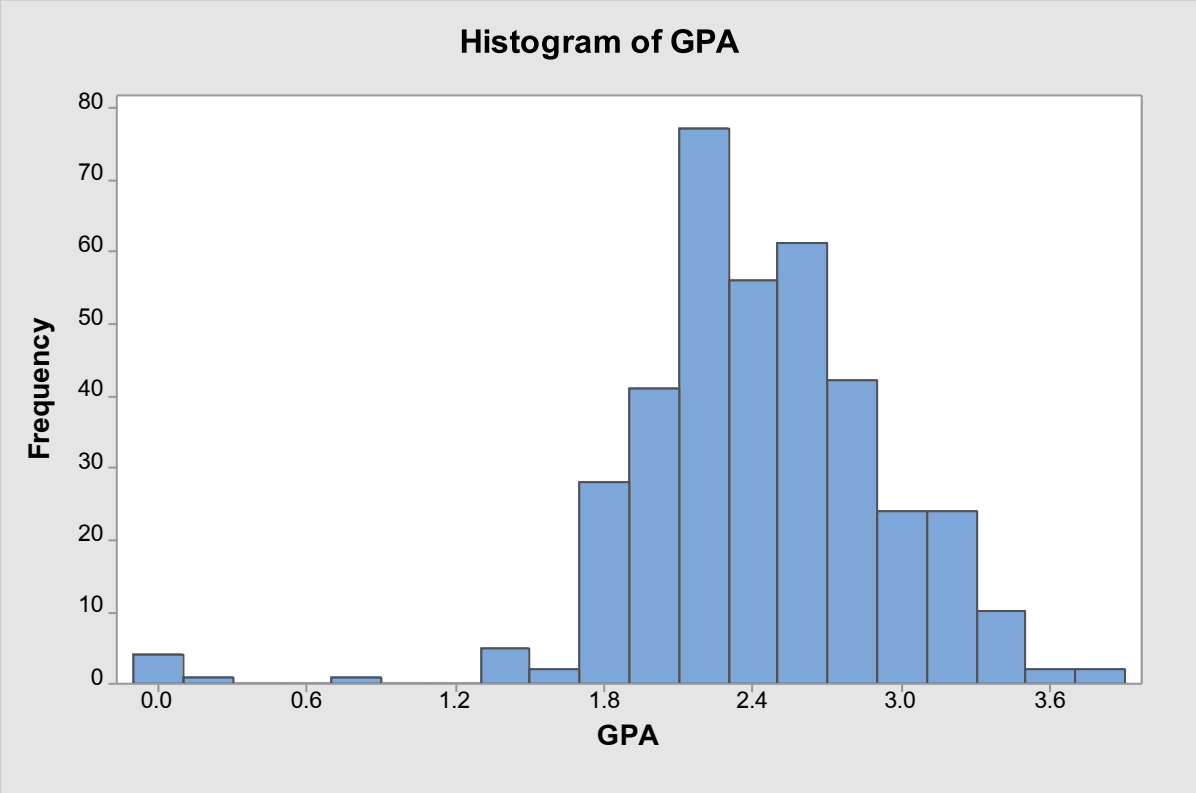
โดยที่ $I_n(T) = \sum_{i=1}^{n} \frac{|t_i|}{|T|} I(t_1)$ (2.9)

## ผลการวิจัย

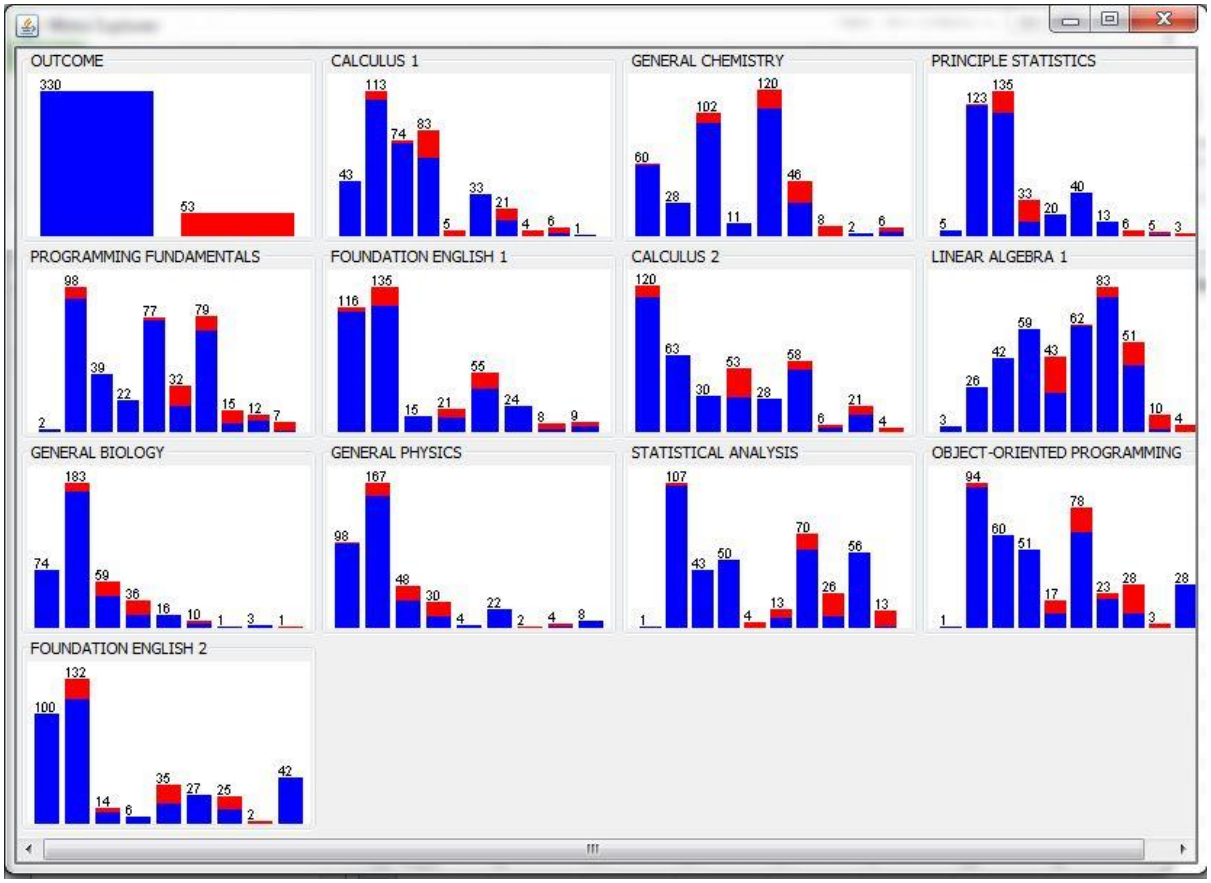
**1. ผลการวิเคราะห์เบื้องต้นเกี่ยวกับผลการเรียนของนักศึกษา**

จากภาพที่ 2 แสดงการแจกแจงความถี่ของเกรดเฉลี่ยสะสมของนักศึกษาชั้นปีที่ 1 ภาควิชาสถิติ

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ระหว่างปีการศึกษา 2559 – 2561 โดยแกน X แทนช่วงของเกรดเฉลี่ยสะสม และ แกน Y แทนจำนวนนักศึกษาที่ได้เกรดเฉลี่ยสะสมในแต่ละช่วงชั้น พบว่ามีลักษณะเบ้ซ้าย (หรือลาดทางซ้าย) โดยที่ค่าความเบ้ (Skewness) เท่ากับ -0.79 และค่าเฉลี่ยของเกรดเฉลี่ยสะสมเท่ากับ 2.42 และมีนักศึกษาจำนวน 53 คน ที่ได้เกรดเฉลี่ยสะสมน้อยกว่า 2.00 จาก จำนวนทั้งหมด 383 คน คิดเป็นร้อยละ 14 เมื่อพิจารณาแยกตามรายวิชาต่างๆ ดังภาพที่ 3 แสดงการแจกแจงความถี่ของผลการเรียนในรายวิชาต่างๆ ของชั้นปีที่ 1 พบว่านักศึกษาที่เรียนรายวิชาหลักสถิติ สถิติเบื้องต้น และพีชคณิตเชิงเส้น มีผลการเรียนน้อยกว่า C มีโอกาสสูงที่จะได้เกรดเฉลี่ยสะสมต่ำกว่า 2.00 เนื่องจากรายวิชาดังกล่าวเป็นวิชาแกนซึ่งเป็นพื้นฐานสำคัญและจำเป็นต่อการศึกษาในรายวิชาต่างๆ ต่อไปของภาควิชาสถิติ คณะวิทยาศาสตร์




**ภาพที่ 2** ฮิสโทแกรมเกรดเฉลี่ยสะสมของนักศึกษา ชั้นปีที่ 1 ภาควิชาสถิติ ระหว่างปีการศึกษา 2559 – 2561




**ภาพที่ 3** แผนภูมิแท่งแสดงการแจกแจงของเกรดในแต่ละวิชาจำแนกโดยตามผลการเรียน

**ตารางที่ 1** กฎความสัมพันธ์ของรายวิชาที่มีผลต่อการสอบได้ของนักศึกษาชั้นปีที่ 1

| ข้อ | กฎความสัมพันธ์ | ตัวชี้วัด | | |
|---|---|---|---|---|
| | | ค่าความเชื่อมั่น | ค่าสนับสนุน | ค่าสหสัมพันธ์ |
| 1 | 05016202 = “Withdraw” → Result = “Failed” | 1.00 | 0.45 | 2.22 |
| 2 | 05016201 = “C” → Result = “Failed” | 1.00 | 0.42 | 2.38 |
| 3 | 05016213 = “Withdraw” → Result = “Failed” | 1.00 | 0.40 | 2.50 |
| 4 | 05406002 = “D+” → Result = “Failed” | 1.00 | 0.38 | 2.63 |
| 5 | 05406002 = “C” → Result = “Failed” | 1.00 | 0.38 | 2.63 |

**2. กฎความสัมพันธ์**

จากตารางแสดงกฎความสัมพันธ์ที่เกิดขึ้น สามารถอธิบายกฎทั้ง 5 ข้อที่ได้ดังนี้

- กฎข้อที่ 1 ถ้านักศึกษาถอนรายวิชาแคลคูลัส 2 แล้วนักศึกษามีความน่าจะเป็นที่จะได้เกรดเฉลี่ยสะสมต่ำกว่า 2.00 ด้วยค่าความเชื่อมั่นที่ 1.00 ค่าสนับสนุนที่ 0.45 ค่าสหสัมพันธ์ที่ 2.22

- กฎข้อที่ 2 ถ้านักศึกษาเรียนรายวิชาแคลคูลัส 1 มีผลการเรียนเท่ากับ C แล้วนักศึกษามีความน่าจะเป็นที่จะได้เกรดเฉลี่ยสะสมต่ำกว่า 2.00 ด้วยค่าความเชื่อมั่นที่ 0.99 ค่าสนับสนุนที่ 0.42 ค่าสหสัมพันธ์ที่ 2.38

- กฎข้อที่ 3 ถ้านักศึกษาถอนรายวิชาพีชคณิตเชิงเส้น แล้วนักศึกษามีความน่าจะเป็นที่จะได้

เกรดเฉลี่ยสะสมต่ำกว่า 2.00 ด้วยค่าความเชื่อมั่นที่ 0.99 ค่าสนับสนุนที่ 0.40 ค่าสหสัมพันธ์ที่ 2.50

- กฎข้อที่ 4 ถ้านักศึกษาเรียนรายวิชาหลักสถิติมีผลการเรียนเท่ากับ D+ แล้ว นักศึกษามีความน่าจะเป็นที่จะได้เกรดเฉลี่ยสะสมต่ำกว่า 2.00 ด้วยค่าความเชื่อมั่นที่ 0.98 ค่าสนับสนุนที่ 0.38 ค่าสหสัมพันธ์ที่ 2.63

- กฎข้อที่ 5 ถ้านักศึกษาเรียนรายวิชาหลักสถิติมีผลการเรียนเท่ากับ C แล้ว นักศึกษามีความน่าจะเป็นที่จะได้เกรดเฉลี่ยสะสมต่ำกว่า 2.00 ด้วยค่าความเชื่อมั่นที่ 0.97 ค่าสนับสนุนที่ 0.38 ค่าสหสัมพันธ์ที่ 2.63

**3. กฎการตัดสินใจ**

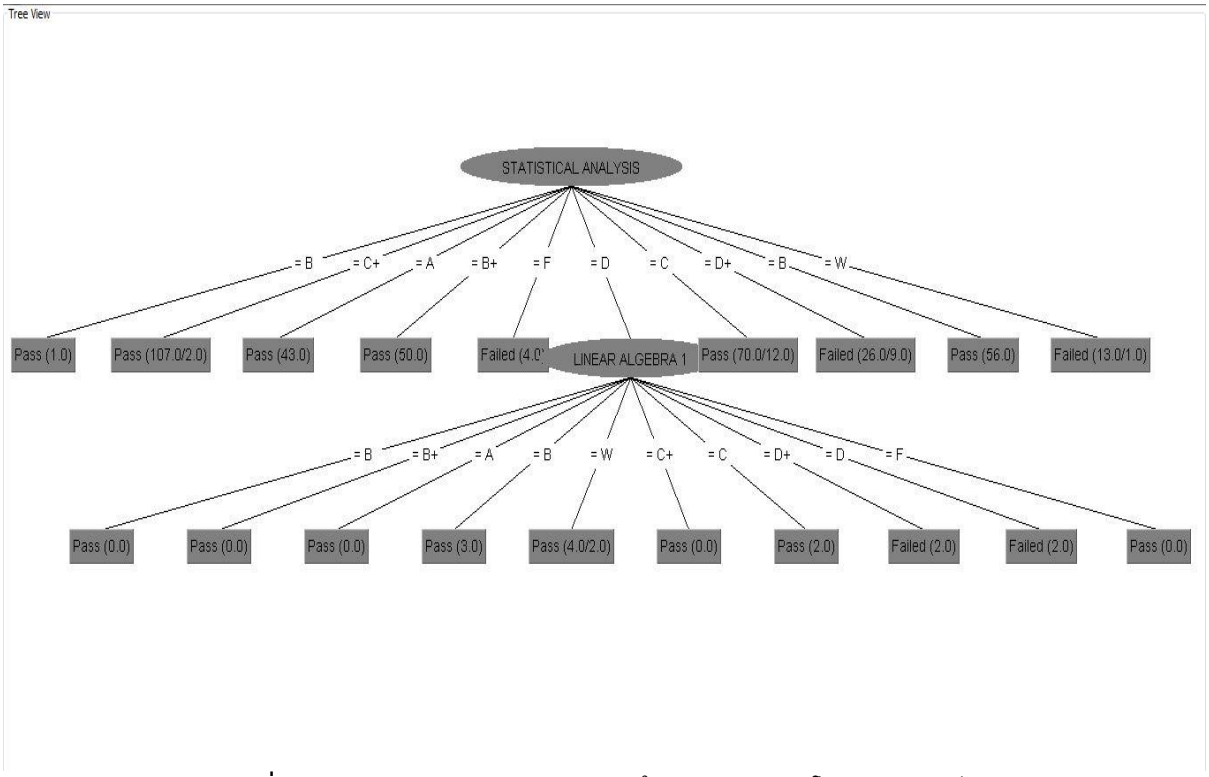
จากการวิเคราะห์ข้อมูลโดยต้นไม้ตัดสินใจด้วยเทคนิค J48 พบว่าความถูกต้อง (Accuracy) เท่ากับ 91.00% ความแม่นยำ (Precision) เท่ากับ 90.70% ความระลึก (Recall) เท่ากับ 91.10% และ F-Measure เท่ากับ 90.90% กฎการตัดสินใจสามารถเขียนอธิบายได้ดังนี้

STATISTICAL ANALYSIS = B : Pass (1.0)
STATISTICAL ANALYSIS = C+: Pass (107.0/2.0)
STATISTICAL ANALYSIS = A: Pass (43.0)
STATISTICAL ANALYSIS = B+: Pass (50.0)
STATISTICAL ANALYSIS = F: Failed (4.0)
STATISTICAL ANALYSIS = D
| LINEAR ALGEBRA 1 = B : Pass (0.0)
| LINEAR ALGEBRA 1 = B+: Pass (0.0)
| LINEAR ALGEBRA 1 = A: Pass (0.0)
| LINEAR ALGEBRA 1 = B: Pass (3.0)
| LINEAR ALGEBRA 1 = W: Pass (4.0/2.0)
| LINEAR ALGEBRA 1 = C+: Pass (0.0)
| LINEAR ALGEBRA 1 = C: Pass (2.0)
| LINEAR ALGEBRA 1 = D+: Failed (2.0)
| LINEAR ALGEBRA 1 = D: Failed (2.0)
| LINEAR ALGEBRA 1 = F: Pass (0.0)
STATISTICAL ANALYSIS = C: Pass (70.0/12.0)
STATISTICAL ANALYSIS = D+: Failed (26.0/9.0)
STATISTICAL ANALYSIS = B: Pass (56.0)
STATISTICAL ANALYSIS = W: Failed (13.0/1.0)

โดยรายละเอียดของกฎของต้นไม้ตัดสินใจ แสดงดังภาพที่ 4 และ Confusion Matrix แสดงในตารางที่ 2 ซึ่งค่าอัตราความถูกต้องเชิงบวก (TP Rate) เท่ากับ 316 ค่าอัตราความถูกต้องเชิงลบ (TN Rate) เท่ากับ 33 ค่าอัตราความผิดพลาดเชิงบวก (FP Rate) เท่ากับ 14 และ ค่าอัตราความผิดพลาดเชิงลบ (FN Rate) เท่ากับ 20

**ตารางที่ 2** Confusion Matrix

| **Predicted/Actual** | **Pass** | **Failed** |
|---|---|---|
| **Pass** | 316 | 14 |
| **Failed** | 20 | 33 |




**ภาพที่ 4** แผนภูมิแท่งแสดงการแจกแจงของเกรดในแต่ละวิชาจำแนกโดยตามผลการเรียน

## สรุปและวิจารณ์ผล

งานวิจัยนี้เป็นการหากฎความสัมพันธ์ของรายวิชาที่มีผลต่อการสอบผ่านนักศึกษาโดยใช้เทคนิคอัลกอริทึมอพริโอริ ทำให้ทราบถึงรายวิชาในชั้นปีที่ 2 ที่มีผลต่อผลการเรียน

ข้อมูลที่ใช้ในการสร้างกฎอาศัยรายวิชาในกลุ่มวิชาบังคับในชั้นปีที่ 1 จำนวน 10 วิชา และผลการเรียน การสร้างกฎความสัมพันธ์ได้กำหนดค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำไว้ที่ 0.1 และ 0.8 ตามลำดับ จากการวิจัยพบกฎความสัมพันธ์ของรายวิชาที่มีผลต่อการสอบผ่านของนักศึกษาชั้นปีที่ 1 ได้แก่ รายวิชาหลักสถิติ วิชาฟิสิกส์ วิชาเคมี และวิชาชีววิทยาซึ่งเป็นรายวิชาในกลุ่มวิชาแกน ตามโครงสร้างหลักสูตร รวมทั้งรายวิชาหลักสถิติเป็นวิชาพื้นฐานที่สำคัญต่อการศึกษาต่อในภาควิชาสถิติต่อไป

สำหรับกฎการจำแนกกลุ่มสำหรับนำไปพยากรณ์ด้วยเทคนิค J48 ตัวอย่างกฎที่พบเช่น นักศึกษาที่ได้เกรดเฉลี่ยสะสมต่ำกว่า 2.00 มักจะมีผลการเรียนรายวิชาสถิติวิเคราะห์เท่ากับ D และรายวิชาพีชคณิตเชิงเส้นเท่ากับ D โดยมีความถูกต้อง (Accuracy) เท่ากับ 91.00% ความแม่นยำ (Precision) เท่ากับ 90.70% ความระลึก (Recall) เท่ากับ 91.10% และ F-Measure เท่ากับ 90.90%

ดังนั้นงานวิจัยนี้สามารถนำกฎความสัมพันธ์และกฎการตัดสินใจไปประยุกต์ใช้ในการพัฒนาระบบแนะนำการเรียนสำหรับนักศึกษาชั้นปีที่ 1 เพื่อให้สามารถสอบผ่านไปเรียนในระดับชั้นที่สูงขึ้นต่อไปได้ รวมไปถึงช่วยอาจารย์ที่ปรึกษาในการวางแผนการเรียนของนักศึกษา และงานอื่นๆ ที่เกี่ยวข้องเพื่อปรับปรุงคุณภาพของการศึกษาให้ดีขึ้น

## เอกสารอ้างอิง

จิราภา เลาหะวรนันท์, รชต ลิ้มสุทธิวันภูมิ, บัณฑิต ฐานะโสภณ และพรฤดี เนติโสภากุล. 2558. การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ. วารสารเทคโนโลยีสารสนเทศ. 4(2): 45-53.

ทิพย์ธิดา วงศ์พิพันธ์. 2556. การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ. งานค้นคว้าอิสระสาขาวิชาเทคโนโลยีคอมพิวเตอร์และการ

สื่อสาร คณะวิศวกรรมศาสตร์ กรุงเทพ. มหาวิทยาลัยธุรกิจบัณฑิตย์.

บุษราภรณ์ มหัทธนชัย, ครรชิต มาลัยวงศ์, เสมอแข สมหอมและณัฐิยา ตันตรานนท์. 2559. กฎความสัมพันธ์ของรายวิชาที่มีผลต่อการพันสภาพนักศึกษาโดยใช้อัลกอริทึมอพริโอริ [Online].1:456-469.http://www.cmruir.cmru.ac.th/bitstream/123456789/448/1/Dropout_Mining.pdf.

สายชล สินสมบรูณ์ทอง.2558. การทำเหมืองข้อมูล. กรุงเทพฯ: จามจุรีโปรดักท์.

ศุภามณ จันทร์สกุล. 2561. เทคนิคเหมืองข้อมูลในการวิเคราะห์ข้อมูลทางการพยาบาล. วารสารวิชาการมหาวิทยาลัยอีสเทิร์นเอเชีย ฉบับวิทยาศาสตร์และเทคโนโลยี. 12(2):83-96.

Agrawal, R., Imielinski, T. and Swami, A. 1993. Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington DC, May 1993, 207-216.

Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Shearer, C. 2000. The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.