ตัวแบบการจำแนกการเลือกหลักสูตรการศึกษา คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม โดยใช้เทคนิคเหมืองข้อมูล

Classification Model for Selection of Program Studies in Faculty of Information Technology in Rajabhat MahaSarakham University Using Data Mining Techniques

ธาดา จันตะคุณ
สาขาวิชาระบบสารสนเทศเพื่อการจัดการ
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม
thada.phd@gmail.com

บทคัดย่อ

ปัจจุบันนี้การรับสมัครนักศึกษาใหม่โดยการสอบคัดเลือกของแต่ละสถาบันอุดมศึกษามีลักษณะที่เป็นเชิงรุกมากขึ้น ซึ่งแต่ละสถาบันมีการประชาสัมพันธ์หลักสูตรและการแนะแนวตามโรงเรียนเนื่องจากสภาวะการแข่งขันของแต่ละสถาบัน ในปัจจุบันงานวิจัยนี้ได้นำเอาเทคนิคเหมืองข้อมูลมาช่วยวิเคราะห์ข้อมูลผู้สมัครเข้าศึกษาต่อจากข้อมูลที่ถูกเก็บไว้ในฐานข้อมูลเพื่อช่วยในการวางแผนการรับนักศึกษาในอนาคต งานวิจัยนี้ประกอบด้วย การเปรียบเทียบตัวแบบการจำแนก 4 เทคนิค คือ Decision Tree, Naïve Bayes, k-NN และ Rule Induction ผลปรากฏว่าเทคนิค Decision Tree ค่าความถูกต้องสูงสุดได้ 83.97%.

คำสำคัญ: ตัวแบบการจำแนก เหมืองข้อมูล

Abstract

Nowadays, a recruitment examination of a new student for each institution of higher education has a more aggressive policy such as course public relations or Guidance to the school curriculum. Due to the higher competition of each institution. This research has gotten data mining concept for helping the analysis of the candidate collected in the database and planning in the future. This research has two main components. Performance comparison of Decision Tree, Naïve Bayes, k-NN, Rule Induction for Classification Model. The results show that the highest accuracy is Decision Tree 83.97%.

Keyword: classification model, data mining

1. บทนำ

มหาวิทยาลัยราชภัฏมหาสารคามในปีการศึกษา 2558 ที่ผ่านมา ได้ทำการรับสมัครนักศึกษาเพื่อทำการคัดเลือกให้เข้าศึกษาต่อใน 3 รูปแบบ คือ รับนักศึกษาผ่านระบบการสอบคัดเลือกของ สกอ. การสอบคัดเลือกด้วยวิธีสอบตรง และการสอบคัดเลือกแบบโควตาทั้งเรียนดีและกีฬา เพื่อให้

มหาวิทยาลัยได้ผู้เรียนที่มีความรู้ ความสามารถ ความถนัดตรงตามหลักสูตรที่เรียน และเป็นการส่งเสริมให้การเรียนการสอนในระดับปริญญาตรีเป็นไปตามปรัชญาและวัตถุประสงค์ของหลักสูตร โดยมีองค์ประกอบในการพิจารณาคัดเลือกผู้สมัครสอบ ดังนี้ พิจารณาจากผลรวมคะแนนสอบวิชาศึกษาทั่วไปและวิชาชีพเฉพาะโดยมหาวิทยาลัยจะพิจารณาคัดเลือก จะพิจารณาจากหลักสูตรที่ผู้สมัครคัดเลือกเป็นอันดับ 1 วิธีและขั้นตอนการเลือกหลักสูตร ผู้สมัครสามารถเลือกหลักสูตรที่สมัครสอบได้สูงสุด 3 อันดับ

จากสภาวะการแข่งขันที่สูงขึ้นสำหรับความต้องการนักศึกษาเข้าใหม่ในแต่ละสถาบันอุมดมศึกษาทั้งในด้านคุณภาพการเรียนและปริมาณของนักศึกษาใหม่ ส่งผลให้รูปแบบการรับสมัครเป็นนโยบายเชิงรุกมากขึ้น โดยกระบวนการประชาสัมพันธ์ถือได้ว่าเป็นกระบวนการที่มีความสำคัญยิ่ง สำหรับการสมัครเข้าเป็นนักศึกษาใหม่ของแต่ละสถาบันอุดมศึกษาโดยเฉพาะวิธีการประชาสัมพันธ์ตามโรงเรียนจำเป็นต้องใช้งบประมาณสูง และอาจไม่ตรงกับกลุ่มเป้าหมายที่ต้องการของแต่ละสาขาวิชา การวิเคราะห์ถึงพฤติกรรมการเลือกสมัครเรียนเพื่อให้ทราบถึงกลุ่มพฤติกรรมในการเลือกสมัครเรียนจะทำให้ได้สารสนเทศประกอบการตัดสินใจเพื่อการวางแผนสำหรับการรับสมัครนักศึกษาใหม่

งานวิจัยนี้ผู้วิจัยได้นำเสนอเทคนิคเหมืองข้อมูล (Data Mining) โดยใช้ข้อมูลของนักศึกษาระดับปริญญาตรี คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม ที่นักศึกษาได้ทำการสมัครเข้าศึกษาและได้รับการคัดเลือกให้เป็นนักศึกษาตามหลักสูตรของมหาวิทยาลัยราชภัฏมหาสารคามแล้ว นำมาผ่านกระบวนการตามมาตรฐานสำหรับการทำเหมืองข้อมูล CRISP-DM (Cross Industry Standard Process for Data Mining) [1] ด้วยโปรแกรม Rapid Miner Studio [2] ซึ่งได้พัฒนาตัวแบบการจำแนกพฤติกรรมและคุณลักษณะของผู้เรียนจากการเลือกหลักสูตรการศึกษาโดยใช้เทคนิคเหมืองข้อมูล (Data Mining) [3] จาก 4 เทคนิค ได้แก่ Decision Tree, Naïve Bayes, k-NN แล้วทำการประเมินประสิทธิภาพเปรียบเทียบตัวแบบที่ได้พัฒนาขึ้น

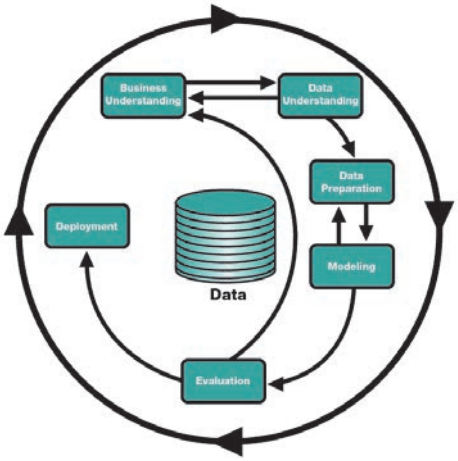
2. เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 เอกสารที่เกี่ยวข้อง

2.1.1 เหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือ กระบวนการวิเคราะห์เพื่อหาความสัมพันธ์และการสรุปผลข้อมูล ซึ่งสามารถเข้าใจได้และเป็นประโยชน์ต่อผู้ทำการรวบรวมข้อมูล [3]

2.1.2 กระบวนการมาตรฐานการทำเหมืองข้อมูลคริสป์-ดีเอม (Cross-Industry Standard Process Data Mining: CRISP-DM) [1] เริ่มต้นจากบริษัท DaimlerChysler บริษัท SPSS และบริษัท NCR มีความต้องการกระบวนการมาตรฐานเพื่อนำมาใช้ทำเหมืองข้อมูลที่เป็นระบบซึ่งในเวลานั้นยังไม่มีการกำหนดมาตรฐานเหล่านั้นออกมาอย่างเป็นรูปธรรม จึงริเริ่มจัดทำแนวทางให้มีกระบวนการที่เป็นมาตรฐานกลางขึ้นมาโดยไม่อิงกับระบบของบริษัทหรือซอฟต์แวร์ใดๆ และไม่เป็นวิชาการมากจนเกินไป แต่เน้นการใช้งานเป็นหลักเพื่อให้เกิดการยอมรับในวงกว้างทั้งในกลุ่มผู้เกี่ยวข้องกับการทำคลังข้อมูลและเหมืองข้อมูล จุดประสงค์หลักของการสร้างกระบวนการมาตรฐานเพื่อให้เกิดโครงการ หรืองานการทำเหมืองข้อมูลรวดเร็วและเป็นไปตามกำหนด โดยมีกระบวนการที่มีประสิทธิภาพและน่าเชื่อถือในการทำงาน เพื่อให้จัดการได้โดยใช้เงินไม่มากเกินไปหรือประหยัดงบประมาณมากที่สุด โดย มาตรฐานคริสป์-ดีเอม ถูกพัฒนาขึ้นในปี ค.ศ.1996 ในกระบวนการทำเหมืองข้อมูลแบบคริสป์-ดีเอ็ม ได้กำหนดไว้ 6 ขั้นตอน คือ 1) ความเข้าใจในธุรกิจ (business understanding) 2) ความเข้าใจข้อมูล (data understanding) 3) การเตรียมข้อมูล (data preparation) 4) การจัดทำตัวแบบ (modelling) 5) การประเมินผล (evaluation) 6) การนำเอาตัวแบบไปใช้งาน (deployment) ดังภาพที่ 1



ภาพที่ 1 : ภาพแสดงกระบวนการของคริสป์-ดีเอ็ม ทั้ง 6 ขั้นตอน [1]

2.1.3 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ เป็นวิธีหนึ่งที่สำคัญในการจำแนกกฎ โดยจะมีลักษณะเป็นการทำงานเหมือนโครงสร้างต้นไม้ ที่แต่ละโหนด (Node) แสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูลแต่ละกิ่งแสดงผลในการทดสอบและลีฟโหนด (Leaf Node) แสดงกลุ่มหรือคลาส (Class) ที่กำหนดไว้ ซึ่งต้นไม้ตัดสินใจนี้ง่ายต่อการเข้าใจและการปรับเปลี่ยนเป็นกฎการจำแนก(Classification Rules) [3]

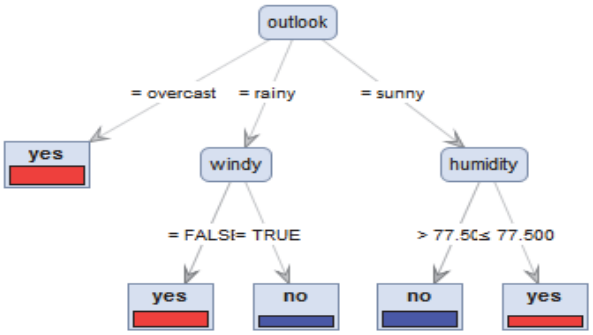
โดยจะสร้างต้นไม้จากบนลงล่างแบบวนซ้ำ (Recursive) ด้วยวิธีการแบ่งปัญหาใหญ่เป็นปัญหาย่อย (Divide-and-Conquer) ซึ่งรูปแบบของต้นไม้จะประกอบด้วย โหนดแรกสุดที่เรียกว่า Root Node จาก Root Node ก็จะแตกออกเป็นโหนดลูก และที่โหนดลูก ก็จะมีลูกของตัวเองซึ่งโหนดในระดับสุดท้ายจะเรียกว่า Leaf Node เช่น ตัวอย่างนี้เป็นข้อมูล weather data ซึ่งเป็นข้อมูลขนาดเล็กมีเพียง 14 กรณี จากตัวอย่างนี้ในโปรแกรม RapidMiner Studio [2] ใช้ชื่อว่า Golf Dataset ส่วนใน Weka [4] มีตัวอย่างข้อมูล 2 ชุด ใช้ชื่อว่า weather.nominal.arff และ weather.numeric.arff ซึ่งชุดข้อมูลนี้เป็นชุดข้อมูลเกี่ยวกับเงื่อนไขสภาพอากาศที่เหมาะสมในการเล่นกอล์ฟ ดังภาพที่ 2 และ 3

ต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคที่ค่อนข้างแพร่หลาย เนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่าย เทคนิคต้นไม้ตัดสินใจจะจำกัดข้อมูลที่เป็นตัวแปรตาม (Dependent Variable) 1 ตัวต่อ 1 แบบจำลอง ถ้าต้องการทำนายตัวแปรตามหลาย ๆ ตัว จะต้องสร้างแบบจำลอง สำหรับตัวแปรตามแต่ละตัวอัลกอริทึมของเทคนิคแบบต้นไม้ตัดสินใจ

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	play	outlook	temperature	humidity	windy
1	no	sunny	85	85	FALSE
2	no	sunny	80	90	TRUE
3	yes	overcast	83	86	FALSE
4	yes	rainy	70	96	FALSE
5	yes	rainy	68	80	FALSE
6	no	rainy	65	70	TRUE
7	yes	overcast	64	65	TRUE
8	no	sunny	72	95	FALSE
9	yes	sunny	69	70	FALSE
10	yes	rainy	75	80	FALSE
11	yes	sunny	75	70	TRUE
12	yes	overcast	72	90	TRUE
13	yes	overcast	81	75	FALSE
14	no	rainy	71	91	TRUE

ภาพที่ 2 : ตัวอย่างชุดข้อมูล Golf Dataset ที่รันบน Rapidminer Studio



ภาพที่ **3** : ตัวอย่างต้นไม้ตัดสินใจ (Decision Tree)จากชุดข้อมูลตัวอย่าง

2.1.4 อัลกอริทึม Naïve Bayes (Naïve Bayes Algorithm) [5] การจำแนกประเภทข้อมูลด้วย Naïve Bayes นี้เป็นวิธีที่ได้รับความนิยม เนื่องจากอาศัยความน่าจะเป็น (probability) เป็นหลัก โดยอาศัยสมการนี้

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

P(A|B) คือ ค่า conditional probability หรือค่าความน่าจะเป็นที่เกิดเหตุการณ์ B ขึ้นก่อนและจะมีเหตุการณ์ A ตามมา

P(A ⋂ B) คือ ค่า joint probability หรือค่าความน่าจะเป็นที่เหตุการณ์ A และเหตุการณ์ B เกิดขึ้นร่วมกัน

P(B) คือ ค่าความน่าจะเป็นที่เหตุการณ์ B เกิดขึ้น

2.1.5 อัลกอริทึมความใกล้เคียงกันมากที่สุด (k-Nearest Neighbor Algorithm) [5] หลักการทำงานของ k-NN คล้ายๆ กับการแบ่งกลุ่มข้อมูล คือ ทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนาย กับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน k ตัว และคำตอบที่ทำนายได้คือ คลาสที่พบมากที่สุดของข้อมูลที่เป็นเพื่อนบ้านทั้ง k ตัว ในเทคนิคนี้จะใช้วิธีวัดระยะห่างแบบ Euclidean ซึ่งเกิดจากรากที่สองของผลต่างระหว่างแอตทริบิวต์ต่างๆ ยกกำลังสอง ดังสมการ

$$\text{distance} = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

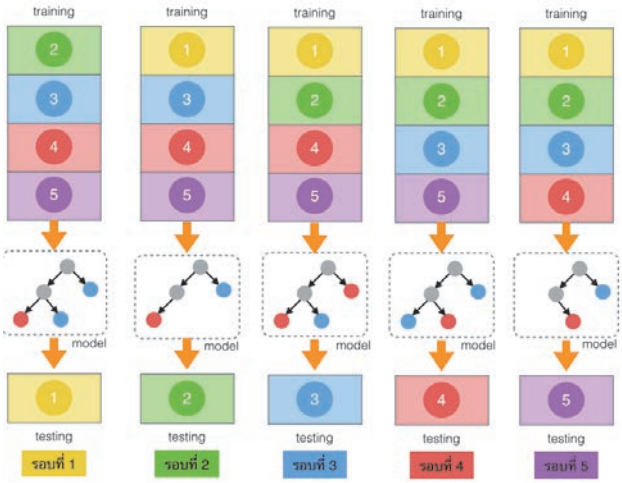
2.1.6 กฎการอุปนัย (Rule Induction)

กฎการอุปนัย (Rule Induction) [6] คือ กฎการอุปนัยเป็นวิธีสำหรับการดึงเอาชุดกฎเกณฑ์ต่างๆ มาเพื่อจัดแบ่งเงื่อนไขหรือกรณี โครงสร้างต้นไม้สามารถสร้างชุดของกฎต่างๆ และขณะที่บางครั้งเรียก วิธีการแบบนี้ว่า การสร้างกฎใหม่จากตัวอย่าง แต่วิธีการนี้ก็ยังมีความหมายที่แตกต่างกันเนื่องจากวิธีการใช้การอุปนัยจะสร้างชุดของกฎที่เป็นอิสระซึ่งไม่จำเป็นต้องอยู่ในรูปโครงสร้างต้นไม้ เพราะตัวสร้างกฎ (Rule Inducer) ไม่ได้บังคับการแตกข้อมูลแต่ละระดับ แต่อาจจะสามารถค้นหารูปแบบ (Pattern) ที่แตกต่างกันได้ และบางครั้งอาจดีกว่าสำหรับการจัดแบ่ง Class ของผลลัพธ์ ลักษณะการนำกฎอุปนัย ดังภาพที่ 4

2.1.7 การประเมินตัวแบบด้วยวิธีการ Cross-validation Test วิธีการนี้เป็นที่นิยมใช้ในการทดสอบประสิทธิภาพของตัวแบบ เนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี Cross-validation Test นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน เช่น 5-fold cross-validation คือการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยแต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของตัวแบบ ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ [5] ดังภาพที่ 5

```
For each class C
  Initialize to the set of all example E
  While E contains example in class C
  Create a rule R with an empty left-hand side that predicts
  class C
  Until R is 100% accurate (or there are no more attributes
  to use) do:
    For each attribute A not in R and each value v
    Consider adding the condition (attribute-value pair)
    A=v to the left hand side of R
    Sellect A and v to maximize the accuracy and oov
    erring of the attribute –value
  Add A=v to R
Remove the examples covered by R from E
```

ภาพที่ 4 : ตัวอย่างอัลกอริทึมสำหรับการนำกฎอุปนัยมาประยุกต์ใช้



ภาพที่ 5 : ตัวอย่างการแบ่งข้อมูลแบบ 5-fold cross-validation [5]

2.2 งานวิจัยที่เกี่ยวข้อง

แก้วสวรรค์ ศรีหริ่ง ได้ทำการศึกษาวิจัยเรื่อง การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาปริญญาตรีของมหาวิทยาลัย:รามคำแหงตามระยะเวลาที่กำหนดโดยเปรียบเทียบวิธีการทางโครงข่ายประสาทเทียม กับวิธีต้นไม้ตัดสินใจ [7] งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของโมเดลโครงข่ายประสาท วิธีต้นไม้ตัดสินใจและตรรกศาสตร์คลุมเครือ เพื่อคาดคะเนความสามารถในการศึกษาของนักศึกษาแต่ละคนที่ผ่านระบบการรับสมัคร

นักศึกษาใหม่ได้ว่า สามารถที่จะสำเร็จการศึกษาในหลักสูตรตามระยะเวลาที่มหาวิทยาลัยรามคำแหงกำหนดได้หรือไม่ โดยใช้ข้อมูลของมหาวิทยาลัยรามคำแหง สาขาวิทยบริการเฉลิมพระเกียรติจังหวัดลพบุรี ตั้งแต่ปีการศึกษา 2550 – 2554 จำนวน 437 คน พิจารณาจากตัวแปรที่จะนำมาศึกษาทั้งสิ้น 20 ตัวแปร และได้นำข้อมูลจำนวนปีที่สำเร็จการศึกษาจัดกลุ่มเป็น 6 กลุ่ม คือ ต่ำกว่า 4 ปี, 4 ปี, 5 ปี, 6 ปี, 7 ปี และ 8 ปี จากนั้นจึงนำข้อมูลไปวิเคราะห์ต่อโดยโครงข่ายประสาทเทียมขนาด 2 ชั้น, 3 ชั้น ตามลำดับ เพื่อทำการเรียนรู้ข้อมูลแล้วจึงเปรียบเทียบผลลัพธ์ วิธีต้นไม้ตัดสินใจในการสร้างแบบจำลองและตรรกศาสตร์คลุมเครือ พบว่าวิธีโครงข่ายประสาทเทียมขนาด 2 ชั้น ซึ่งมีขนาดแต่ละชั้น 16-6 ได้ผลลัพธ์ที่ 99.64% โดยใช้ฟังก์ชัน กระตุ้นแบบ Tansig-Purelin วิธีโครงข่ายประสาทเทียมขนาด 3 ชั้น ได้ผลลัพธ์ที่ 99.80% ซึ่งมีขนาดแต่ละชั้น 24-6-6 โดยใช้ฟังก์ชันกระตุ้นแบบ Tansig-Logsig-Purelin วิธีต้นไม้ตัดสินใจ ได้ผลลัพธ์ที่ 99.54% และตรรกศาสตร์คลุมเครือได้ผลลัพธ์ที่ 89.24% ผลการทดลองพบว่า แบบจำลองโครงข่ายประสาทเทียมขนาด 3 ชั้น เป็นโมเดลที่เหมาะแก่การนำไปพัฒนาต่อ เพราะให้ประสิทธิภาพมากที่สุด

ไพฑูรย์ จันทร์เรือง ได้ทำการศึกษาวิจัยเรื่อง ระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษา ระดับปริญญาตรี โดยใช้เทคนิคต้นไม้ตัดสินใจ [8] มีใจความสำคัญดังนี้ งานวิจัยนี้ได้ทำการพัฒนาระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคต้นไม้ตัดสินใจ ซึ่งจากการทดลองพบว่าในการสร้างตัวแบบสำหรับพัฒนาระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคต้นไม้ตัดสินใจนั้น ควรแยกสร้างตัวแบบสำหรับแต่ละสาขาการเรียน เนื่องจากคุณสมบัติของผู้เรียนแต่ละสาขามีความแตกต่างกัน เพื่อให้ได้ตัวแบบที่สามารถทำนายแนวโน้มของผลการเรียนที่เหมาะสมสำหรับแต่ละสาขา แต่เนื่องจากคะแนนเฉลี่ยของนักศึกษาที่นำมาพัฒนาตัวแบบนั้น ส่วนใหญ่จะมีเกณฑ์คะแนนเกาะกลุ่มกันอยู่ในช่วงกลางของข้อมูล (2.00 – 3.00) ทำให้ผลการตัดสินส่วนใหญ่จะโน้มเอียงไปในเกณฑ์พอใช้ (ช่วงคะแนน 2.00 – 2.49) และปานกลาง (ช่วงคะแนน 2.50 – 2.99)

ณัฐ พลอยอ่อง ได้ทำการศึกษาวิจัยเรื่อง ระบบคัดกรองนักศึกษาที่มีความสามารถในการเรียนวิชาเทคโนโลยีสารสนเทศผ่านสื่ออิเล็กทรอนิกส์โดยการเปรียบเทียบเทคนิคต้นไม้การตัดสินใจและซัพพอร์ตเวกเตอร์แมชชีน [9] งานวิจัยนี้มีประสงค์เพื่อคัดเลือกโมเดลระหว่างเทคนิคต้นไม้ตัดสินใจกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน เพื่อใช้พัฒนาระบบคัดกรองนักศึกษาที่มีความสามารถในการเรียนวิชาเทคโนโลยีสารสนเทศโดยการเรียนรู้ผ่านสื่ออิเล็กทรอนิกส์ของมหาวิทยาลัยราชภัฏสวนสุนันทาเพื่อเป็นการส่งเสริมผู้เรียนโดยเน้นผู้เรียนเป็นหลักให้สามารถเลือกรูปแบบการเรียนที่เหมาะสมกับตนเองได้ ผลการเปรียบเทียบโมเดลที่ได้คือเทคนิคต้นไม้การตัดสินใจให้ผลความถูกต้องสูงที่สุดจากการสร้างโมเดลแบบ 100 Folds ได้ค่าความถูกต้องที่74.46% ซึ่งมากกว่าเทคนิคซัพพอร์ตเวกเตอร์แมชชีนที่ให้ผลความถูกต้องสูงที่สุด จากการสร้างโมเดลแบบ 100 Folds ได้ค่าความถูกต้องที่ 72.48% จากผลดังกล่าวเทคนิคต้นไม้ตัดสินใจถูกนำมาใช้พัฒนาโปรแกรมระบบคัดกรองนักศึกษา

ธีรพงษ์ สังข์ศร ได้ศึกษาวิจัยเรื่อง การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษา [10] งานวิจัยชิ้นนี้ได้นำเอาแนวคิดเหมืองข้อมูลมาช่วยวิเคราะห์โดยสามารถแบ่งส่วนการทำงานออกได้เป็น 2 ส่วนหลักประกอบด้วย 1) การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนเป็นการประยุกต์ใช้ เทคนิคการจัดกลุ่มข้อมูล (Clustering) แบบ Simple K-means สามารถแบ่งข้อมูลพฤติกรรมของผู้ สมัครได้ เป็น 4 กลุ่มและใช้ การหากฎความสัมพันธ์ ของข้อมูล (Association Rules) ด้วยเทคนิคแอพริออริ (Apriori) เพื่อหากฎความสัมพันธ์ของข้อมูลในแต่ละกลุ่มพฤติกรรมผู้สมัคร โดยใช้ ค่าความเชื่อมั่นเท่ากับ 0.9และ 2) การเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยตัวแบบพยากรณ์ ที่ถูกสร้างขึ้นด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีค่าความถูกต้องเท่ากับ 93.76% และตัวแบบที่ถูกสร้างขึ้นด้วยเทคนิคโครงข่าย

ประสาทเทียม(Artificial Neural Network) มีค่าความถูกต้องเท่ากับ 93.60%

3. วิธีการดำเนินงานวิจัย

งานวิจัยนี้ผู้วิจัยใช้กระบวนการของ CRISP-DM ดังนี้

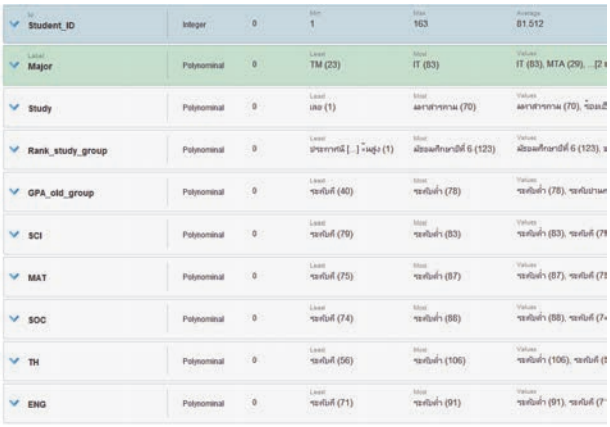
3.1 การทำความเข้าใจในด้านธุรกิจ (Business Understanding)

พบปัญหาของการประชาสัมพันธ์หลักสูตร ไม่ตรงกับความสนใจของนักเรียนทำให้จำนวนนักศึกษาลดลง จึงพัฒนาตัวแบบการจำแนกพฤติกรรมการเลือกสาขาวิชาให้กับนักศึกษา จำนวน 4 หลักสูตร ดังนี้ หลักสูตรเทคโนโลยีสารสนเทศ หลักสูตรเทคโนโลยีมัลติมีเดียและแอนิเมชั่น หลักสูตรเทคโนโลยีคอมพิวเตอร์และการสื่อสาร และหลักสูตรการจัดการเทคโนโลยี

3.2 การทำความเข้าใจข้อมูล (Data Understanding)

ปีการศึกษา 2558 คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคามในระดับปริญญาตรีมีการรับสมัครนักศึกษาจำนวน 4 หลักสูตร มีจำนวนนักศึกษาที่สมัครแล้วได้รับการคัดเลือกทั้ง 4 หลักสูตร รวมทั้งสิ้นจำนวน 162 คน มีปัจจัยที่ต้องพิจารณาทั้งหมด 8 ปัจจัย (แอตทริบิวต์) ได้แก่ Major หมายถึง หลักสูตรที่สมัครเข้าศึกษา, Study หมายถึง จังหวัดเดิมที่เคยเรียนมาก่อน, Rank_Study_group หมายถึง วุฒิการศึกษาที่ใช้สมัครสอบ, Gpa_old_group หมายถึง เกรดเฉลี่ยนของวุฒิการศึกษาเดิม, SCI หมายถึง ระดับคะแนนรายวิชาวิทยาศาสตร์, MAT หมายถึงระดับคะแนนวิชาคณิตศาสตร์, SOC หมายถึง ระดับคะแนนรายวิชาสังคม, TH หมายถึง ระดับคะแนนรายวิชาภาษาไทย, ENG หมายถึง ระดับคะแนนรายวิชาภาษาอังกฤษ แสดงชุดข้อมูล ดังภาพที่ 6

3.3 ขั้นตอนการเตรียมข้อมูล ข้อมูลที่ได้มาอยู่ในรูปแบบของแฟ้มข้อมูลไมโครซอฟท์แอคเซส โดยทำการคัดเลือกแอตทริบิวต์ที่เกี่ยวข้องแล้วแปลงให้อยู่ในรูปแบบของแฟ้มข้อมูลที่มีนามสกุลเป็น .CSV จากนั้นทำการกำหนดแอตทริบวิต์ Student_id เป็นประเภท id กำหนดให้แอตทริบิวต์ Major เป็นประเภท Label จากนั้นทำการแปลงข้อมูลแอตทริบิวต์ที่เกี่ยวข้องทั้ง 8 แอตทริบิวต์ให้อยู่ในรูปแบบ ดังตารางที่ 1



ภาพที่ 6 : ภาพแสดงชุดข้อมูล (Data Set)

ตารางที่ 1 อธิบายแอตทริบิวต์ที่เกี่ยวข้อง

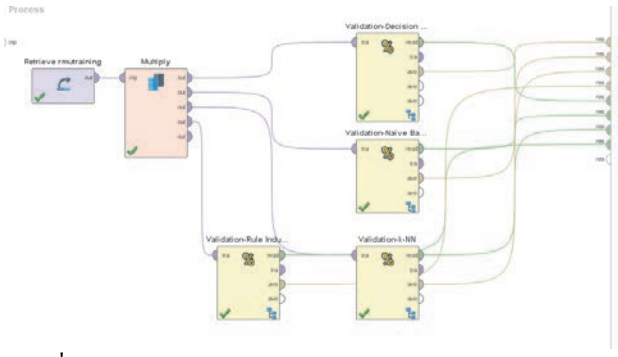
แอตทริบิวต์	ลักษณะของข้อมูลที่ถูกแปลงก่อนนำไปสร้างตัวแบบ
Major	เทคโนโลยีสารสนเทศ, เทคโนโลยีมัลติมีเดียและแอนิเมชั่น, เทคโนโลยีคอมพิวเตอร์และการสื่อสาร, การจัดการเทคโนโลยี
Study	ระบุชื่อจังหวัดที่จบการศึกษามา เช่น “มหาสารคาม”
Rank_Study_group	สายสามัญ, สายอาชีพ
Gpa_old_group	ระดับต่ำ, ระดับปานกลาง, ระดับดี
SCI	ระดับต่ำ, ระดับปานกลาง, ระดับดี
MAT	ระดับต่ำ, ระดับปานกลาง, ระดับดี
SOC	ระดับต่ำ, ระดับปานกลาง, ระดับดี
TH	ระดับต่ำ, ระดับปานกลาง, ระดับดี
ENG	ระดับต่ำ, ระดับปานกลาง, ระดับดี

3.4 การสร้างตัวแบบ (Modeling) สร้างตัวแบบด้วย RapidMiner Studio โดยใช้โอเปอร์เรเตอร์ Multiply เพื่อให้สามารถสร้างตัวแบบด้วยเทคนิค Decision Tree, Naïve Bayes, k-NN และ Rule Induction ในโปรเซสเดียว ดังภาพที่ 7

3.5 การประเมินประสิทธิภาพตัวแบบ (Evaluation)

ใช้วิธีการ Cross-validation Test ด้วย 5-fold cross-validation, 10-fold cross-validation เพื่อหาค่าความถูกต้อง (accuracy) โดยตัวแบบที่พัฒนาจากเทคนิค Decision Tree ที่ถูกทดสอบด้วยข้อมูลที่แบ่งออกเป็น 5 ส่วน นั้น ได้ค่าความถูกต้องเท่ากับ 83.31% และเมื่อทดสอบด้วยข้อมูลที่แบ่ง

ออกเป็น 10 ส่วนได้ค่าความถูกต้องเท่ากับ 83.97% ตัวแบบที่พัฒนาจากเทคนิค Naïve Bayes ที่ถูกทดสอบด้วยข้อมูลที่แบ่งออกเป็น 5 ส่วน นั้น ได้ค่าความถูกต้องเท่ากับ 74.07% และเมื่อทดสอบด้วยข้อมูลที่แบ่งออกเป็น 10 ส่วนได้ค่าความถูกต้องเท่ากับ 73.53% ตัวแบบที่พัฒนาจากเทคนิค k-NN ที่ถูกทดสอบด้วยข้อมูลที่แบ่งออกเป็น 5 ส่วน นั้น ได้ค่าความถูกต้องเท่ากับ 73.41% และเมื่อทดสอบด้วยข้อมูลที่แบ่งออกเป็น 10 ส่วนได้ค่าความถูกต้องเท่ากับ 74.71% ตัวแบบที่พัฒนาจากเทคนิค Rule Induction ที่ถูกทดสอบด้วยข้อมูลที่แบ่งออกเป็น 5 ส่วน นั้น ได้ค่าความถูกต้องเท่ากับ 82.75% และเมื่อทดสอบด้วยข้อมูลที่แบ่งออกเป็น 10 ส่วนได้ค่าความถูกต้องเท่ากับ 75.18% ดังตารางที่ 2



ภาพที่ 7 : ภาพแสดงสร้างตัวแบบและประมวลผลเพื่อเปรียบเทียบได้ในโปรเซสเดียวบน RapidMiner Studio

ตารางที่ 2 เปรียบเทียบประสิทธิภาพของตัวแบบที่พัฒนาขึ้น

Classification Model	5-fold cross-validation (Accuracy)	10-fold cross-validation (Accuracy)
Decision Tree	83.31%	83.97%
Naïve Bayes	74.07%	73.53%
k-NN	73.41%	74.71%
Rule Induction	82.75%	75.18%

3.6 การนำไปใช้งาน (Deployment)

นำตัวแบบที่พัฒนาด้วยเทคนิค Decision Tree ซึ่งได้ค่าความถูกต้องสูงที่สุดไปแนะนำหลักสูตรให้กับนักเรียน โดยให้เปรียบเทียบกับตัวแบบกับนักเรียนโดยให้เริ่มไล่จากโหนดบนสุดที่เป็นวิชาต่างๆ และไล่เรียงลงมาจนถึงโหนดสุดท้าย (leaf) ดังภาพที่ 8

ภาพที่ 8 : แสดงตัวแบบในรูปแบบต้นไม้ตัดสินใจ

4. สรุปผลการวิจัย

สรุปผลการดำเนินการวิจัยครั้งนี้โดยพัฒนาและเปรียบเทียบตัวแบบการจำแนกทั้ง 4 เทคนิค ได้แก่ Decision Tree, Naïve Bayes, k-NN, Rule Induction ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ Decision Tree ซึ่งได้ค่าที่สูงที่สุดจากการแบ่งข้อมูลทดสอบออกเป็น 10 ชุด ค่าความถูกต้อง (Accuracy) ได้ 83.97% จึงสรุปได้ว่าDecision Tree เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำไปใช้ประชาสัมพันธ์หลักสูตร

5. เอกสารอ้างอิง

[1] IBM Corporation, "http://www.ibm.com/us-en/,". [Online]. Available: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf.

[2] RapidMiner USA, "https://rapidminer.com/,". [Online]. Available: https://rapidminer.com/products/studio/.

[3] J. Han, Data Mining Concepts and Techniques Third Edition, United States of America: Morgan Kaufmann Publishers, 2012.

[4] Data Mining Software in Java, "http://www.cs.waikato.ac.nz/ml/weka/,". [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/book.html.

[5] เอกสิทธิ์ พัชรวงศีศักดา, การวิเคราะห์ข้อมูลด้วยดาต้า ไมน์นิง เบื้องต้น, กรุงเทพฯ: บริษัท เอเชีย ดิจิตอลการพิมพ์ จำกัด, 2557.

[6] ฝนทิพย์ คูนแก้ว, "การสังเคราะห์โมเดลเพื่อการจำแนกตามข้อกำหนดของผู้ใช้," มหาวิทยาลัยสุรนารี, นครราชสีมา, 2555.

[7] แก้วสวรรค์ ศรีหริ่ง, "การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาปริญญาตรีของมหาวิทยาลัยรามคำแหงตามระยะเวลาที่กำหนดโดยเปรียบเทียบวิธีการทางโครงข่ายประสาทเทียม วิธีต้นไม้ตัดสินใจและตรรกศาสตร์คลุมเครือ," มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2556.

[8] ไพฑูรย์ จันทร์เรือง, " ระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคต้นไม้ตัดสินใจ,"

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 2550.

[9] ณัฐ พลอยอ่อง, "ระบบคัดกรองนักศึกษาที่มีความสามารถในการเรียนวิชาเทคโนโลยีสารสนเทศผ่านสื่ออิเล็กทรอนิกส์ โดยการเปรียบเทียบเทคนิคต้นไม้การตัดสินใจและซัพพอร์ตเวก," มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 2555.

[10] ธีรพงษ์ สังข์ศรี, "การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิคการทำเหมืองข้อมูล," การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่10, กรุงเทพฯ, 2557.