The Sci J of Phetchaburi Rajabhat University
2019.16(2): 61-71

การวิเคราะห์การลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล

Analysis for Student Dropout in UndergraduateUsing Data Mining Technique

ภาภรณ์ เหล่าพิลัย[1*] และ จรัญ แสนราช[2]

PapornLaopilai[1*] and Charun Sanrach[2]

[1]*คณะศิลปะศาสตร์และวิทยาศาสตร์ มหาวิทยาลัยราชภัฏศรีสะเกษ จังหวัดศรีสะเกษ 33000*

[2]*คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร 10800*

[1] *Faculty of Liberal Arts and Sciences Sisaket Rajabhat University, Sisaket, 33000*

[2] *Faculty of Technical Education King Mongkuts University of Technology North Bangkok, 10800*

***Corresponding author; E-mail:** paporn.l@sskru.ac.th

Received: 5 December 2019 /Revised: 6 February 2020 / Accepted: 26 February 2020

บทคัดย่อ

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อ 1) วิเคราะห์หาปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี 2) สังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี และ 3) เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของโมเดลด้วยเทคนิควิธี Rule Induction, K-Nearest Neighbor, Decision Tree และ Naive Bayes โดยใช้ข้อมูลจากฐานข้อมูลงานทะเบียนของมหาวิทยาลัยเทคโนโลยีราชมงคลอีสานของนักศึกษาระดับปริญญาตรี ระหว่างปีการศึกษา 2557-2561 มีจำนวน 14 แอททริบิวต์และ 10,151 ชุดข้อมูล เมื่อนำมาวิเคราะห์ค่าน้ำหนักของแอททริบิวต์ พบว่ามีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาจำนวน 12 ปัจจัย เมื่อนำปัจจัยที่ได้มาทำการสร้างเป็นโมเดล ทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold Cross Validation และวัดประสิทธิภาพด้วยค่า Accuracy เพื่อหาวิธีการที่มีความถูกต้องมากที่สุด ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลพบว่าโมเดลที่สร้างด้วยเทคนิควิธี Rule Induction มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้อง 94.70 % และมีปัจจัยที่เกี่ยวข้องสูงสุด 5 อันดับ ได้แก่ เกรดเฉลี่ย ปีการศึกษา โรงเรียนเดิมสาขาวิชา และอาชีพของบิดา

คำสำคัญ:การลาออกกลางคันการทำเหมืองข้อมูลกฎการอุปนัยเคเนียร์เรสเนเบอร์ต้นไม้ตัดสินใจ

Abstract

The purposes of this research were 1) to analyze the factors related to dropout of undergraduate students 2) to propose a model for predicting dropout of undergraduate students and 3) to compare the data classification performance of the models using Rule Induction, K-Nearest Neighbor, Decision Tree and Naive Bayes algorithms. The data was collected from the undergraduate student's registration database of Rajamangala University of Technology Isanduring the academic years from 2014 to 2018. The dataset has 14 attributes and 10,151 records. The data were analyzed the weight of the attributes and showed that 12 factors influencing student's dropout. Those 12 factors were used to build models with different techniques. Moreover, the cross-validation with 10 folds method was also used to evaluate the best prediction accuracy of each technique. The result suggested that the rule induction model has the best performance among all techniques.It has an average accuracy of 94.70%. The findings also indicated that students' decision to dropout was significantly influenced by the GPA, academic year, former school, major and the occupation of the father.

Keywords:Student Dropout, Data Mining, Rule Induction, K-Nearest Neighbor, Decision Tree

บทนำ

การศึกษาระดับอุดมศึกษาเป็นองค์ประกอบที่สำคัญสำหรับการพัฒนาทรัพยากรมนุษย์ทั่วโลก ทำให้คนมีโอกาสพัฒนาทักษะระดับสูงที่จำเป็นในตลาดแรงงานทุกๆด้านและส่งประโยชน์ไปถึงการพัฒนาอาชีพ สังคมและประเทศชาติโดยรวม ซึ่งตามพระราชบัญญัติการศึกษาแห่งชาติกำหนดว่าการจัดหลักสูตรการศึกษาต้องมีลักษณะหลากหลาย นอกจากมุ่งพัฒนาคุณภาพชีวิตแล้ว ยังพัฒนาวิชาการ วิชาชีพชั้นสูงและการค้นคว้า วิจัย เพื่อพัฒนาองค์ความรู้และพัฒนาสังคมด้วยการยกระดับคุณภาพการศึกษาในระดับอุดมศึกษาเป็นสิ่งที่สำคัญยิ่ง ตัวบ่งชี้ในการประเมินคุณภาพการศึกษาของมหาวิทยาลัยอย่างหนึ่งคือจำนวนนักศึกษาที่สำเร็จการศึกษาในระยะเวลาที่กำหนดไว้ในหลักสูตรในการจัดทำแผนพัฒนาการศึกษาระดับอุดมศึกษาฉบับที่ 11 (พ.ศ.2555-2559) ได้กำหนดเป้าหมายให้มีผู้สำเร็จการศึกษาภายใน 4 ปี คิดเป็นร้อยละ 70 อย่างไรก็ตามผู้ที่มาศึกษาในระดับอุดมศึกษามีส่วนหนึ่งพ้นสภาพนักศึกษาทำให้ต้องออกจากการศึกษากลางคัน ส่งผลเสียหายต่อสถานศึกษาทำให้เสียเวลาในการบริหารจัดการ และเสียทรัพยากรในการลงทุน ส่วนผู้เรียนเสียเวลาและทรัพย์สิน เป็นปัญหาที่สถาบันอุดมศึกษาหลายแห่งประสบ[1] แม้ว่าใน

ปัจจุบันจะมีสถาบันอุดมศึกษาที่เปิดสอนเพิ่มมากขึ้นทั้งภาครัฐและเอกชน ทำให้ผู้เรียนมีทางเลือกที่หลากหลายแต่ยังมีนักศึกษาจำนวนหนึ่งที่มีการออกจากระบบการศึกษากลางคัน โดยมีปัจจัยสาเหตุแตกต่างกันไป ซึ่งการที่ผู้เรียนสามารถเรียนได้จนจบหลักสูตรหรือจบการศึกษาได้นั้นจำเป็นต้องอาศัยผู้ที่เกี่ยวข้องกับการศึกษา โดยการส่งเสริมและพัฒนากระบวนการเรียนการสอนให้มีประสิทธิภาพ ตลอดจนช่วยกันหาแนวทางในการป้องกันและแก้ไขปัญหาการลาออกกลางคันของนักศึกษา หากนักศึกษาลาออกกลางคันก่อนที่จะจบการศึกษาถือว่าเป็นความสูญเสียทางการศึกษา ส่งผลกระทบด้านเศรษฐกิจของประเทศและเศรษฐกิจของครอบครัว ซึ่งต้องสิ้นเปลืองค่าใช้จ่ายไปไม่ได้รับประโยชน์ที่คุ้มค่า

ในสถานศึกษาหลายแห่งปัญหาการลาออกกลางคันของนักศึกษาเป็นสิ่งที่ต้องหาแนวทางในการแก้ไขจากงานวิจัยของพิชัย ระเวงวันและพุธษดี ศิริแสงตระกูล[2]ศึกษาปัจจัยที่มีผลต่อสถานภาพของนักศึกษาของมหาวิทยาลัยเอกชนแห่งหนึ่งซึ่งเปิดสอนในภาคตะวันออกเฉียงเหนือ จากข้อมูลของนักศึกษาที่ศึกษา ระหว่างปี 2551-2557 พบว่ามีอัตราการออกกลางคันอยู่ที่19.16 % และจากฐานข้อมูลของมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ระหว่างปี 2557-2561 มีการออกกลางคันของนักศึกษามากถึง 3,577 คน คิดเป็นร้อยละ 35.23 จากจำนวนนักศึกษาทั้งหมด 10,151 คน

การทำเหมืองข้อมูล เป็นวิธีการในการค้นหารูปแบบและแนวโน้มที่มีประโยชน์จากแหล่งข้อมูลขนาดใหญ่ [3] เช่น ฐานข้อมูล คลังข้อมูล เว็บ หรือแหล่งจัดเก็บข้อมูลอื่น ๆ โดยเฉพาะในด้านการศึกษาถือว่าเป็นแหล่งที่มีข้อมูลขนาดใหญ่ถูกเก็บรวบรวมอยู่จำนวนมาก เช่นประวัติผู้เรียน ประวัติผลการเรียนและข้อมูลรายวิชาต่างๆเป็นต้น

จากเหตุผลดังกล่าว งานวิจัยฉบับนี้จึงเสนอการวิเคราะห์การลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูลด้วยเทคนิค Rule Induction, K-Nearest Neighbor, Decision Tree และ Naive Bayes เพื่อวิเคราะห์หาปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรีสังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี และเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของโมเดลด้วยเทคนิคทั้ง 4 ข้อมูลที่ได้สามารถนำไปรับใช้เพื่อการดูแล ติดตาม ช่วยเหลือให้นักศึกษาสามารถสำเร็จการศึกษาและนำไปใช้เป็นแนวทางในการป้องกันและแก้ไขปัญหาการลาออกกลางคันของนักศึกษาในแต่ละปีการศึกษาต่อไป

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1)กฎการอุปนัย (Rule Induction)

กฎการอุปนัย (Rule Induction) คือกฎการอุปนัยเป็นวิธีสำหรับการดึงเอาชุดกฎเกณฑ์ต่างๆ มาเพื่อจัดแบ่งเงื่อนไขหรือกรณีโครงสร้างต้นไม้สามารถสร้างชุดของกฎต่างๆ และขณะที่บางครั้งเรียกวิธีการแบบนี้ว่าการสร้างกฎใหม่จากตัวอย่าง [4]แต่วิธีการนี้ยังมีความหมายที่แตกต่างกันเนื่องจากวิธีการใช้การอุปนัยจะสร้างชุดของกฎที่เป็นอิสระ ซึ่งไม่จำเป็นต้องอยู่ในรูปโครงสร้างของต้นไม้ เพราะตัวสร้างการอนุมานกฎ (Rule Induction) ไม่ได้บังคับการแตก

ข้อมูลเป็นแต่ละระดับ แต่อาจจะสามารถค้นหารูปแบบ (Pattern) ที่แตกต่างกันได้

2)ต้นไม้ตัดสินใจ (Decision tree)

ต้นไม้ตัดสินใจ เป็นวิธีหนึ่งที่สำคัญในการจำแนกกฎโดยจะมีลักษณะเป็นการทำงานเหมือนโครงสร้างต้นไม้ที่แต่ละโหนด (Node) แสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูลแต่ละกิ่งแสดงผลในการทดสอบและลีฟโหนด (Leafnode) แสดงกลุ่มหรือคลาส (Class) ที่กำหนดไว้ซึ่งต้นไม้ตัดสินใจนี้ง่ายต่อการเข้าใจและการปรับเปลี่ยนเป็นกฎการจำแนก (Classification rules) [3]โดย Decisiontree จะทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของ tree เรียกว่าโหนดราก (Rood node) จากนั้นจะเลือกคุณสมบัติที่มีความสัมพันธ์ถัดไปเรื่อยๆ จากการคำนวณ InformationGain (IG) โดยเลือกคุณสมบัติที่มีค่า IG สูงที่สุด คำนวณได้จากสมการดังนี้

$$IG(parent, child) = Entropy(parent) - [p(c1)xEntropy(c1) + p(c2)x\ Entropy(c2) + \ldots] \quad (1)$$

เมื่อ

Entropy(c1) คือ –p(c1) log p(c1)

P(c1) คือ ค่าความน่าจะเป็นของค่า c1

cคือ ปัจจัย (Attribute) แต่ละตัวที่เกี่ยวข้อง

ซึ่งค่า Entropy นี้จะใช้ในการวัดความแตกต่างกันของข้อมูล ถ้าข้อมูลมีความแตกต่างกันน้อย ค่า Entropy จะมีค่าต่ำ แต่ถ้าข้อมูลมีความแตกต่างกันมากค่า Entropy จะมีค่าสูง ดังนั้นถ้าข้อมูล Entropy ของโหนดลูก (Child) สามารถสร้างโมเดลของ Decision Tree จะคำนวณค่า IG ของแต่ละแอททริบิวต์เทียบกับคลาสเพื่อหาแอททริบิวต์ที่มีค่า IG มากที่สุดมาเป็น Rood ของโมเดล Decision Tree

3) อัลกอริทึมนาอีฟเบย์ (Naïve Bayes)

นาอีฟเบย์เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น(Probability) ตามทฤษฎีของเบย์ (Bayes'stheorem)ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อนเป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ หลักการของนาอีฟเบย์ใช้หลักการของความน่าจะเป็น โดยมีสมมติฐานว่าปริมาณของความสนใจขึ้นอยู่กับการกระจายความน่าจะเป็น (Probability distribution) [5] เป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณลักษณะ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad (2)$$

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

$$P(C|A) = \frac{P(A|C)X\ P(C)}{P(A)} \quad (3)$$

จากสมการ Bayes อธิบายว่าถ้าต้องการทำนายคลาส C เมื่อทราบแอททริบิวต์ สามารถคำนวณจากความน่าจะเป็นของแอททริบิวต์A ที่มีคลาส C ใน Training dataและค่าความน่าจะเป็นของแอททริบิวต์A และ C มีสมการ [6] ดังนี้

P(C|A) คือ ค่าความน่าจะเป็นที่ข้อมูลที่มีแอททริบิวต์ A จะมีคลาส C

P(A|C) คือ ค่าความน่าจะเป็นที่ข้อมูล Training data ที่มีคลาส C และมีแอททริบิวต์A โดยที่ A = a1⋂ a2 ... ⋂aMโดยที่ M คือจำนวนแอททริบิวต์ใน Training data

P(C) คือ ความน่าจะเป็นของคลาส C

P(A) คือ ความน่าจะเป็นของคลาส A

แต่การที่แอททริบิวต์A = a1 ⋂ a2 ... ⋂aM ที่เกิดขึ้นในTraining data อาจจะมีจำนวนน้อยมากหรือไม่มีรูปแบบของแอททริบิวต์แบบนี้เกิดขึ้นเลย ดังนั้นจึงได้ใช้หลักการที่ว่าแต่ละแอททริบิวต์เป็น Independent ต่อกันทำให้สามารถเปลี่ยนสมการ P(A|C) ได้เป็น

$$P(A|C) = P(a_1|C)\ X\ P(a_2|C) X \ldots X P(a_M|C) \quad (4)$$

โดยสามารถตัดส่วนของ P(A) ออกได้เนื่องจากเป็นส่วนของการปรับค่าให้อยู่ในช่วงนั้น (Normalization)

4)K-Nearest Neighbors

วิธีการค้นหาเพื่อนบ้านใกล้ที่สุด [5] เป็นการแบ่งกลุ่มข้อมูล และทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ตัว และคำตอบที่ทำนายได้ดีคือคลาสที่พบมากที่สุดของข้อมูลที่เป็นเพื่อนบ้านทั้ง K ตัวมักจะใช้วิธีการวัดระยะห่างแบบ Euclidean เกิดจากรากที่สองของผลต่างระหว่างแอททริบิวต์ต่าง ๆ ยกกำลังสองดังสมการ

$$Distance = \sqrt{\sum_{k=1}^{n}(p_k - (p_k - q_k)^2} \quad (5)$$

5) วิธีการ 10-Fold Cross Validation [6] การวัดประสิทธิภาพของโมเดลการพยากรณ์ที่สร้างด้วยวิธีการ 10-Fold cross validation จะแบ่งข้อมูลออกหลายส่วน(แสดงด้วยค่า k) ในการดำเนินการวิจัยครั้งนี้จะแบ่งข้อมูลออกเป็น 10 ส่วน โดยแต่ละส่วนมีจำนวนข้อมูลเท่ากัน จากนั้นข้อมูลส่วนหนึ่งจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดลทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ ประสิทธิภาพของโมเดล 10 ครั้ง

6) ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล[7] โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัย4 ค่า คือ

- ค่าความถูกต้อง (Accuracy) คือจำนวนข้อมูลที่ทำนายถูกทุกคลาสเป็นการวัดความถูกต้องของโมเดลโดยพิจารณารวมทุกกรณี
- ค่าความระลึก (Recall) คือจำนวนที่ทำนายถูกกี่ตัว เป็นการวัดความถูกต้องของโมเดล
- ค่าความแม่นยำ (Precision) คือค่าที่ดูสิ่งที่ทำนายออกมาแล้วทายถูกได้กี่เปอร์เซ็นต์
- ค่าความถ่วงดุล (F-measure) คือค่าเฉลี่ยของค่าความแม่นยำและค่าความระลึก

7) วิเคราะห์ปัจจัยที่ส่งผลต่อการลาออกกลางคันของนักศึกษาระดับปริญญาตรี[7]เพื่อกำหนดเป็นดัชนีและเรียงลำดับความสำคัญของดัชนีโดยการลดมิติของข้อมูล (Attribute selection)

วิธีดำเนินการวิจัย

ในงานวิจัยนี้ได้เสนอวิธีการทำเหมืองข้อมูลโดยใช้โปรแกรม Rapid Miner Studio 9 โดยขั้นตอนการทำเหมืองข้อมูลสำหรับการทำเหมืองข้อมูลแบบ CRISP-DM เป็น Workflow มาตรฐานสำหรับการทำ Data mining ประกอบด้วย 6 ขั้นตอน คือ 1) ความเข้าใจในธุรกิจ (Business understanding) 2) ความเข้าใจข้อมูล (Data understanding)3) การเตรียมข้อมูล(Data preparation)4) การจัดทำตัวแบบ Modeling)5) การประเมินผล(Evaluation)6) การนำเอาตัวแบบไปใช้งาน (Deployment) โดยแต่ละขั้นตอนจะเป็นขั้นตอนที่ต่อเนื่องกัน นั่นคือขั้นตอนต่อไปต้องมีผลลัพธ์จากขั้นตอนก่อนหน้าด้วยลูกศรที่เชื่อมระหว่างแต่ละขั้นตอน ดังแสดงในFigure 1

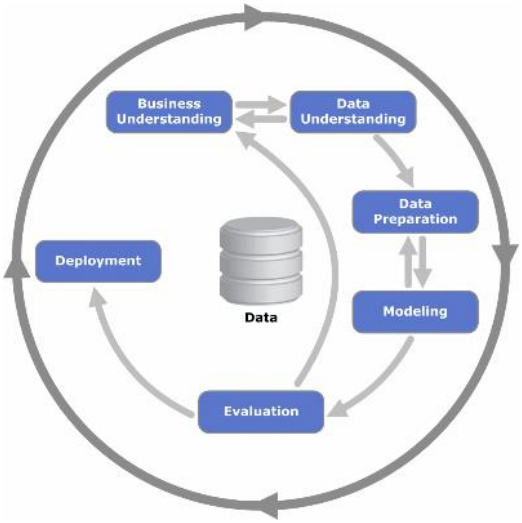

Figure 1.Process of CRISP-DM

1) ความเข้าใจเกี่ยวกับธุรกิจ (Business understanding)การเข้าใจปัญหาและทำการวิเคราะห์เหมืองข้อมูลเพื่อศึกษาสาเหตุและแนวทางแก้ไขการลาออกกลางคันของนักศึกษา วิเคราะห์เหมืองข้อมูลด้วยเทคนิค 1) Rule Induction 2) K-Nearest Neighbor 3) Decision Tree และ4) Naive Bayes เพื่อวิเคราะห์หาปัจจัยที่เกี่ยวข้องกับการสังเคราะห์โมเดลสำหรับทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรีและเปรียบเทียบประสิทธิภาพของโมเดล ผู้วิจัยได้ทำการวิเคราะห์โดยใช้ชุดข้อมูลของนักศึกษาระดับปริญญาตรีมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ตั้งแต่ปีการศึกษา 2557-2561 เมื่อได้ข้อมูลแล้วทำการเตรียมข้อมูลให้พร้อมที่จะนำไปทำการคัดกรองภายใต้หลักการทำงานของเหมืองข้อมูล

2) ความเข้าใจข้อมูล (Data understanding) ทำการรวบรวมข้อมูลพื้นฐานของนักศึกษาจากฐานข้อมูลนักศึกษาจากฐานข้อมูลงานทะเบียนของมหาวิทยาลัย ซึ่งมี 14 แอททรบิวต์และคัดกรองเฉพาะข้อมูลที่มีความสมบูรณ์ทั้งหมด 10,151 ชุดข้อมูลและทำการคัดกรองแอทริบิวต์เพื่อให้ได้แอทริบิวต์ที่จำเป็นที่สุดสำหรับการจำแนกประเภทเพื่อหาประสิทธิภาพในการจำแนกข้อมูลและสร้างความสัมพันธ์ของแอทริบิวต์

3) การเตรียมข้อมูล(Data preparation) ในขั้นตอนนี้เป็นการเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุด เนื่องจากโมเดลที่ได้จากการทำ Data mining จะให้ผลลัพธ์ที่ถูกต้องหรือไม่นั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ ผู้วิจัยทำการเตรียมข้อมูลที่ได้อยู่ในรูปแบบของตาราง Excelโดยเริ่มจาก

3.1) ทำการคัดเลือกข้อมูล (Data selection) เป็นการเลือกใช้เฉพาะข้อมูลที่เกี่ยวข้อง งานวิจัยในครั้งนี้ประกอบไปด้วย14 แอทริบิวต์ ดังแสดงใน Table1

Table 1.List of attributes used to predict student dropout

No	Attribute	Direction
1.	เกรดเฉลี่ย (GPA)	Input
2.	ปีเข้าศึกษา(Year)	Input
3.	โรงเรียนเดิม (Former school)	Input
4.	สาขาวิชา(Major)	Input
5.	อาชีพบิดา (Father's occupation)	Input
6.	อาชีพมารดา (Mother's occupation)	Input
7.	แผนการศึกษา(Study program)	Input
8.	รายได้บิดา (Father's income)	Input
9.	รายได้ครอบครัว (Family's income)	Input
10.	รายได้มารดา (Mother's income)	Input
11.	เพศ(Gender)	Input
12.	สถานะครอบครัว(Parent status)	Input
13.	ปี (Year)	Input
14.	สถานะนักศึกษา (Student status)	Input

3.2) ทำการกลั่นกรองข้อมูล (Data cleaning) ในขั้นตอนนี้จะทำการปรับปรุงข้อมูลให้ถูกต้อง ซึ่งอยู่ในรูปแบบของตาราง Excelและทำการลบข้อมูลซ้ำซ้อน แก้ไขข้อผิดพลาด ข้อมูลที่ผิดรูปแบบ ข้อมูลที่มีค่าว่างข้อมูลที่แปลกแยกจากส่วนอื่น

3.3) ทำการแปลงรูปแบบของข้อมูล (Data transformation) เมื่อทำการปรับข้อมูลแล้ว ในขั้นตอนนี้จะทำการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ข้อมูลตามอัลกอริทึ่มของ Data mining ที่เลือกใช้ เนื่องจากข้อมูลที่นำมาวิเคราะห์ครั้งนี้มีรูปแบบในการจัดเก็บทั้งประเภท Nominal และ Numeric จึงต้องแปลงให้อยู่ในรูปแบบเดียวกันและทำการ Clustering data แบ่งข้อมูลหลายๆกลุ่มตามความคล้ายคลึงกัน จากชุดข้อมูลทดสอบคำตอบ (Class) จะแบ่งเป็น 2 ประเภท คือ Yes ลาออก มีจำนวน 3,577 ชุดข้อมูล และNo ไม่ลาออก มีจำนวน 6,574 ชุดข้อมูล

3.4) การสร้างโมเดล (Modeling) การสร้างและทดสอบความถูกต้องของโมเดลเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิค Data mining ทำการสังเคราะห์โมเดลจากข้อมูลที่มีอยู่ เพื่อการทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรี หาค่าความถูกต้อง (Accuracy) ที่ออกมาเป็นตัวเลข เครื่องมือที่ใช้ทำเหมืองข้อมูลด้วยโปรแกรม Rapid Miner Studio 9 เป็นเครื่องมือในการวิเคราะห์ข้อมูล (Figure 2) ในการวิจัยครั้งนี้ผู้วิจัยวิเคราะห์ปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาด้วย Filter approach วิเคราะห์ค่าน้ำหนักของแอทริบิวต์ จากนั้นจึงนำปัจจัยที่ได้จากการวิเคราะห์ปัจจัยที่ส่งผลต่อการลาอออกกลางคันของนักศึกษาระดับปริญญาตรี มาทำการสร้างโมเดลการทำนายด้วยเทคนิคเหมืองข้อมูล 4 โมเดล ด้วยเทคนิควิธี1) Rule Induction 2) K-Nearest Neighbor 3) Decision Tree

และ4) Naive Bayes ทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold cross validation ในการวัดประสิทธิภาพของการจำแนกประเภทข้อมูล ได้แก่การหาค่าค่าความระลึก (Recall)ค่าความถ่วงดุล (F-measure) ความแม่นยำ (Precision)และค่าความถูกต้อง (Accuracy)

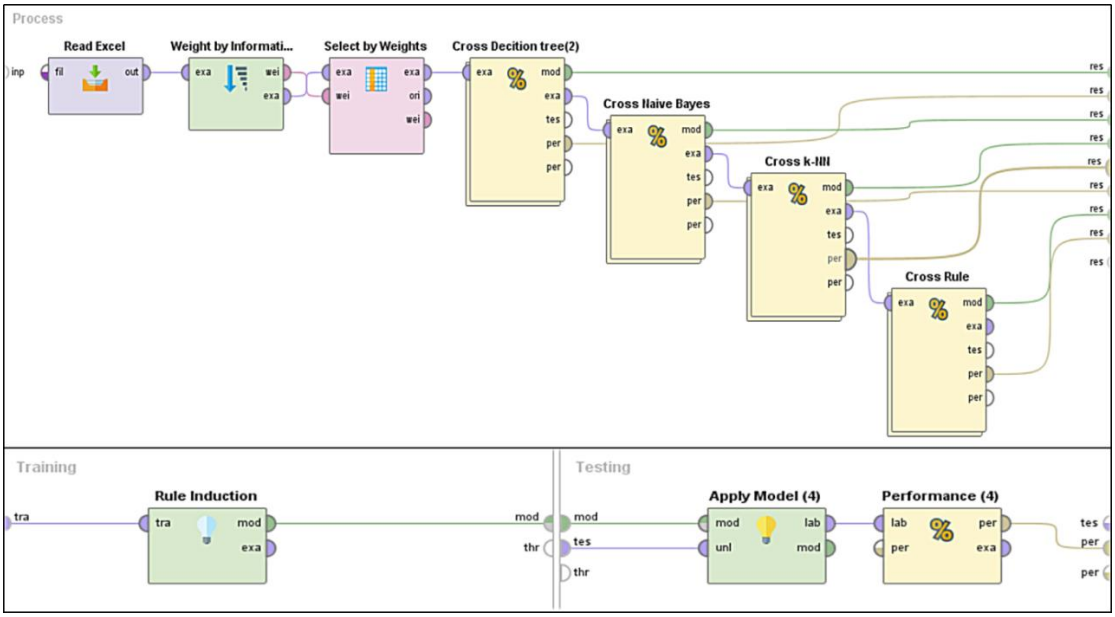

Figure2. Model ofRapid Miner Studio9

3.5) การประเมินผล (Evaluation) จะได้ผลจากการสร้างโมเดล พบว่ามีจำนวน 12 ปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษา โมเดลที่สร้างด้วยเทคนิควิธี Rule Induction ให้ผลลัพธ์ที่ดีที่สุดโดยมีค่าความถูกต้องร้อยละ 94.70 (Table 2)

Table 2. The results of the comparison of the efficiency

Algorithm	Accuracy
Rule Induction	94.70%
K-Nearest Neighbor	89.63%
Decision Tree	90.12%
Naive Bayes	90.62%

5) การนำแบบจำลองไปใช้งาน (Deployment) นำโมเดลที่ได้ผ่านกระบวนการเปรียบเทียบประสิทธิภาพของโมเดลที่ได้ค่าความถูกต้องสูงสุด ไปใช้ในการทำนายการลาออกกลางคันของนักศึกษาเพื่อใช้เป็นแนวทางในการป้องกันและแก้ไขปัญหาการลาออกกลางคันของนักศึกษาในแต่ละปีการศึกษาต่อไป

ผลการศึกษา

1.ผลการวิเคราะห์ปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยการลดมิติข้อมูล (Attribute selection) การคำนวณค่าน้ำหนักของแอททริบิวต์ พบว่ามีจำนวน 12 ปัจจัย

ที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษา ได้แก่ เกรดเฉลี่ย ชั้นปี โรงเรียนเดิม สาขาวิชา อาชีพของบิดา อาชีพของมารดา วุฒิการศึกษา รายได้บิดา รายได้มารดา รายได้ครอบครัว เพศและสถานะครอบครัว รายละเอียดดังแสดงใน Table3

Table 3. Factors related to student's dropout

No	Attribute	Weight
1.	GPA	0.375
2.	Year	0.053
3.	Former school	0.044
4.	Major	0.023
5.	Mother's occupation	0.005
6.	Mother's occupation	0.004
7.	Study program	0.002
8.	Father's income	0.001
9.	Family's income	0.001
10.	Mother's income	0.001
11.	Gender	0.001
12.	Parent status	0.001

2.ผลการสังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรีและวัดประสิทธิภาพของโมเดลด้วยวิธีการ 10-Fold crossvalidation

Table4. The result of comparing the efficiency with the accuracy of the model

Algorithm	Accuracy	Precision	Recall
Rule Induction	94.70%	96.46%	88.23%
K-NN	89.63%	94.44%	74.14%
Decision Tree	90.12%	96.67%	74.53%
Naive Bayes	90.62%	90.05%	82.36%

Table 4. (Cont.)

จาก Table 4 จะเห็นว่าสามารถใช้เหมืองข้อมูลในการทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรีด้วยเทคนิคเหมืองข้อมูล 4 โมเดลด้วยเทคนิคคือ 1) Rule Induction 2) K-Nearest Neighbors 3) Decision Tree 4)Naive Bayes เพื่อวัดประสิทธิภาพของโมเดลด้วยวิธีการ 10-Fold cross validation โมเดลที่สร้างด้วยเทคนิควิธี Rule Induction มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้องร้อยละ 94.70 เทคนิควิธี Naive Bayes มีค่าเฉลี่ยความถูกต้องร้อยละ90.62 Decision tree มีค่าเฉลี่ยความถูกต้องร้อยละ 90.12 และ เทคนิควิธี K-Nearest Neighbors มีค่าเฉลี่ยความถูกต้องร้อยละ 89.63

อภิปรายผล

จากการรวบรวมข้อมูลจากฐานข้อมูลงานทะเบียนของมหาวิทยาลัยเทคโนโลยีราชมงคลอีสานของนักศึกษาระดับปริญญาตรี มีจำนวนนักศึกษาที่ลาออกจำนวน3,577 คน จำนวนนักศึกษาที่ไม่ลาออกมีจำนวน 10,151คน และเมื่อทำการวิเคราะห์ค่าน้ำหนักของแอททริบิวต์พบว่ามีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันสูงสุด 5 อันดับ ได้แก่เกรดเฉลี่ย ปีการศึกษา โรงเรียนเดิมสาขาวิชา และอาชีพของบิดา

เพราะเหตุนี้จึงต้องเข้าใจถึงเกรดเฉลี่ยของนักศึกษาและความยากของสาขาวิชา ให้ความช่วยเหลือทางด้านการเงินแก่นักศึกษาที่ขาดแคลนทุนทรัพย์ โดยอาจจะพิจารณาผ่อนผันการชำระค่าเทอม พิจารณาให้กู้ยืมกองทุนเพื่อการศึกษา พิจารณาให้ทุนการศึกษาสำหรับผู้เรียนดีแต่ยากจนหรือทุนทางด้านกิจกรรม รวมไปถึงการแก้ปัญหาทางด้านการเรียนของนักศึกษา ในแต่ละสาขาวิชา เพื่อให้ผลการเรียนดีขึ้น ดังนั้นอาจารย์ประจำสาขาวิชาจะต้องกำกับดูแลนักศึกษาอย่างใกล้ชิดเพื่อลดความเสี่ยงในการลาออกกลางคันของนักศึกษาซึ่งสอดคล้องกับชณิดาภา บุญประสม,และจรัญ แสนราช[8]ที่ได้รวบรวมข้อมูลจากฐานข้อมูลงานทะเบียนของมหาลัยวิทยาลัยราชภัฏอุบลราชธานีของนักศึกษาระดับปริญญาตรี เมื่อทำการวิเคราะห์ค่าน้ำหนักของแอททริบิวต์ด้วยวิธีการ Information theory พบว่ามีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันสูงสุด 5 อันดับ ได้แก่ การกู้ยืมกองทุนเพื่อการศึกษา สาขาวิชา เกรดเฉลี่ย อาชีพของมารดาและอาชีพของบิดา

สรุป

การดำเนินการวิจัยครั้งนี้ได้ศึกษาและเปรียบเทียบตัวแบบการจำแนกทั้ง 4 เทคนิค ได้แก่ วิธี Rule Induction, K-Nearest Neighbor, Decision Tree และ Naive Bayes ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ Rule Induction ซึ่งได้ค่าที่สูงที่สุดจากการแบ่งข้อมูลทดสอบออกเป็น 10 ชุด ค่าความถูกต้อง 94.70 % จึงสรุปได้ว่า Rule Induction เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำไปวิเคราะห์หาปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี

เอกสารอ้างอิง

1.บุษราภรณ์มหัทธนชัย,ครรชิต มาลัยวงศ์, เสมอแข สมหอมและณัฐิยา ตันตรานนท์. 2559. กฎความสัมพันธ์ของรายวิชาที่มีผลต่อการพ้นสภาพนักศึกษาโดยใช้อัลกอริทึมอพริโอริ. *การประชุมวิชาการระดับชาติมหาวิทยาลัยราชภัฏกำแพงเพชร*.3(1):456-469.

2.พิชัย ระเวงวัน และพุธษดี ศิริแสงตระกูล.โมเดลเพื่อการพยากรณ์สถานภาพทางการศึกษาของนักศึกษา. [online] เข้าถึงได้จาก. https://gsbooks.gs.kku.ac.th/60/nigrc2017/pdf/PMP6.pdf. 2562.

3.Daniel T. L., Chantal D. L. 2015. *Datamining and predictive analytics*.United States of America: John Wiley and Sons incorporate.

4. กฤษณะไวยมัย, ชิดชนก ส่งศิริและธนาวินท์ รักธรรมานนท์. 2549. การใช้เทคนิคดาต้าไมน์นิ่งเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์. *วารสารวิชาการเนคเทค*. 3(11):134-142.

5.ธาดา จันตะคุณ.2559.ตัวแบบการจำแนกการเลือกหลักสูตรการศึกษา คณะเทคโนโลยีสารสนเทศมหาวิทยาลัยราชภัฏมหาสารคาม โดยใช้เทคนิคเหมืองข้อมูล.*การประชุมวิชาการครุศาสตร์อุตสาหกรรมระดับชาติ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*. 9: 336-343.

6. Han,J. 2012.*Data Mining Concepts and Techniques*,Third Edition. United States of America: Morgan Kaufmann Publishers.
7. เอกสิทธิ์ พัชรวงศ์ศักดา.2557. *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมน์นิงเบื้องต้น*. กรุงเทพมหานคร: เอเชียดิจิตอลการพิมพ์.
8. ชณิดาภา บุญประสม และจรัญ แสนราช. 2561. การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูล. *วารสารวิชาการครุศาสตร์อุตสาหกรรม มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ*. 9: 142-151.