Technical Education Journal King Mongkut's University of Technology North Bangkok
Vol.9 No.1 January - April 2018

การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูล

ชณิดาภา บุญประสม[1*] และ จรัญ แสนราช[2]

บทคัดย่อ

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อ 1) วิเคราะห์หาปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี 2) สังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี และ 3) เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของโมเดลด้วยเทคนิควิธี Decision Tree, K-Nearest Neighbors, Naive Bayes โดยใช้ข้อมูลจากฐานข้อมูลงานทะเบียนของมหาลัยวิทยาลัยราชภัฏอุบลราชธานี ของนักศึกษาระดับปริญญาตรี ระหว่างปีการศึกษา 2558-2560 มีจำนวน 11 แอททริบิวต์และ 13,729 ชุดข้อมูล เมื่อนำมาวิเคราะห์ค่าน้ำหนักของแอททริบิวต์ด้วยวิธีการ Information Theory พบว่า 1) มีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาจำนวน 8 ปัจจัย 2) นำปัจจัยที่ได้มาทำการสร้างเป็นโมเดลทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold Cross Validation และวัดประสิทธิภาพด้วยค่า Accuracy เพื่อหาวิธีการที่มีความถูกต้องมากที่สุด 3) ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลพบว่าโมเดลที่สร้างด้วยเทคนิควิธี Naive Bayes มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้อง 93.58 % มากกว่าเทคนิควิธี Decision Tree มีค่าเฉลี่ยความถูกต้อง 93.52 % และเทคนิควิธี K-Nearest Neighbors มีค่าเฉลี่ยความถูกต้อง 87.95 % และมีปัจจัยที่เกี่ยวข้องสูงสุด 5 อันดับ ได้แก่ การกู้ยืมกองทุนเพื่อการศึกษา, สาขาวิชา, เกรดเฉลี่ย, อาชีพของมารดา และอาชีพของบิดา

คำสำคัญ: การลาออกกลางคัน, การทำเหมืองข้อมูล, ต้นไม้ตัดสินใจ, เคเนียร์เรสเนเบอร์, การเรียนรู้นาอีฟเบย์

[1] อาจารย์ประจำสาขาวิศวกรรมซอฟต์แวร์ คณะวิทยาการคอมพิวเตอร์ มหาวิทยาลัยราชภัฏอุบลราชธานี
[2] ผู้ช่วยศาสตราจารย์ ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
* ผู้นิพนธ์ประสานงาน โทร. 08-3932-0845 อีเมล: chanidapa.b@ubru.ac.th

Predictive Analytic for Student Dropout in Undergraduate Using Data Mining Technique

Chanidapa Boonprasom[1*] and Charun Sanrach[2]

Abstract

The purposes of this research were 1) to analyze the factors that involved with the dropout of undergraduate students 2) to propose a model for predicting the dropout of undergraduate synthesize 3) to compare the performance of 3 different classification techniques, including Decision Tree, K-Nearest Neighbors, and Naive algorithms. The data was collected from the undergraduate student's registration database of Ubon Ratchathani Rajabhat University during the academic years from 2015 to 2017. The dataset has 11 attributes and 13,729 records. The data were analyzed using the Information theory selection method. The results showed that 1) there are 8 factors that influencing student's dropout 2) Those factors were used to build models with the different techniques, Moreover, the cross-validation with 10 folds method was used to evaluate the best prediction accuracy of each technique. 3) the result suggested that the Naive Bayes model has the best performance among all techniques. It has the average accuracy of 93.58 %, which are higher than Decision tree and K-Nearest Neighbors which have the average accuracy of 93.52 % and 87.95 %, accordingly. The findings also indicated that students' decision to dropout was significantly influenced by the student loan, major of study, grade point average, and the occupation of their parents.

Keywords: Data Mining, Decision Tree, K-Nearest Neighbors, Naive Bayes, student dropout

[1] Lecturer, Department of Software Engineering, Faculty of Computer Science, Ubon Ratchathani Rajabhat University

[2] Assistant Professor, Department of Computer Studies, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok

[*] Corresponding Author Tel. 08-3932-0845 e-mail: chanidapa.b@ubru.ac.th

1. บทนำ

สถาบันการศึกษามีบทบาทสำคัญในการพัฒนาประเทศ การศึกษาเป็นรากฐานที่สำคัญในการสร้างบุคคลให้มีความรู้ความสามารถในการปฏิบัติหน้าที่และสามารถดำรงชีวิตอยู่ในสังคมได้อย่างสันติสุข การที่เยาวชนของชาติสามารถเล่าเรียนได้จนจบหลักสูตรหรือจบการศึกษาได้นั้นจำเป็นต้องอาศัยผู้ที่เกี่ยวข้องกับวงการศึกษา โดยการส่งเสริมและพัฒนากระบวนการเรียนการสอนให้มีประสิทธิภาพ ตลอดจนช่วยกันหาแนวทางในการป้องกันและแก้ไขปัญหาการลาออกกลางคันของนักศึกษาในระหว่างเรียน หากนักศึกษาเกิดความล้มเหลวทางการศึกษาโดยการลาออกกลางคันก่อนที่จะจบการศึกษา ศึกษาไม่สำเร็จตามหลักสูตรที่กำหนดได้นั้นถือว่าเป็นความสูญเสียทางการศึกษาที่ทำให้การลงทุนของรัฐบาลสูญเปล่าไม่สามารถจะผลิตนักศึกษาได้ตามความต้องการ จึงเป็นการสูญเสียทรัพยากรที่มีคุณค่ายิ่งทำให้ส่งผลกระทบด้านเศรษฐกิจของประเทศและเศรษฐกิจของครอบครัว ซึ่งต้องสิ้นเปลืองค่าใช้จ่ายไปไม่ได้รับประโยชน์ที่คุ้มค่า

ในสถานศึกษาหลายแห่งพบว่า ปัญหาการลาออกกลางคันของนักศึกษาเป็นสิ่งที่ต้องหาแนวทางในการแก้ไข ศักรินทร์และคณะ [1] จากฐานข้อมูลของวิทยาลัยเทคโนโลยีภาคตะวันออกเฉียงเหนือในปีการศึกษา 2556-2558 พบว่ามีจำนวนนักศึกษา ปวส. ทั้งหมด 3,146 คน จำนวนนักศึกษาที่ลาออกกลางคัน 326 คน คิดเป็นร้อยละ 10.36 ของภาพรวม และจากฐานข้อมูลของมหาวิทยาลัยราชภัฏอุบลราชธานี ระหว่างปี 2558-2560 มีการออกกลางคันของนักศึกษาทั้งภาคปกติและภาคพิเศษมีจำนวนมากถึง 2,364 คน คิดเป็นร้อยละ 17.21 จากจำนวนนักศึกษาทั้งหมด 13,729 คน ซึ่งส่งผลกระทบไปยังจำนวนนักศึกษาคงอยู่และจำนวนผู้สำเร็จการศึกษา มีผลการดำเนินงานที่ไม่สามารถบรรลุได้ตามเป้าหมายที่กำหนดไว้

ปัจจุบันการทำเหมืองข้อมูลทางการศึกษา (Education Data Mining) เป็นที่นิยมอย่างมากโดยเฉพาะการจำแนกประเภทข้อมูล (Classification) เพื่อหาความสัมพันธ์ของข้อมูลและเปรียบเทียบประสิทธิภาพของอัลกอริทึม ดังงานวิจัยของโชติกาและขวัญทัย [2] เป็นการทดสอบประสิทธิภาพของอัลกอริทึม Decision Tree, Naive Bays, Sequential minimal optimization เพื่อใช้ในสร้างแบบจำลองการพยากรณ์ประสิทธิภาพของผู้เรียน โดยผู้เรียนสามารถพยากรณ์การทำนายประสิทธิภาพของตนเองได้ว่าตนเองอยู่ในประสิทธิภาพระดับใดของผลการพยากรณ์ เพื่อนำไปสู่การตัดสินใจให้นิสิตวางแผนการเรียนตามเป้าหมายของตนเองได้อย่างมีประสิทธิภาพยิ่งขึ้น พบว่า อัลกอริทึม C4.5 มีประสิทธิภาพในการจำแนกข้อมูลโดยรวมได้ดีที่สุด แตกต่างกับงานวิจัยของจุฑาทิพย์และนิเวศ [3] ศึกษาการเปรียบเทียบประสิทธิภาพการจำแนกอีเมลที่เป็นสแปมโดยใช้เทคนิคการทำเหมืองข้อมูล Decision Tree, Naive Bays, K-Nearest Neighbor การวิเคราะห์และทำนายหาที่มาของจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมได้ถูกต้องและสามารถนำไปประยุกต์ใช้เป็นการลดภาระของเครื่องแม่ข่าย (Server) ในองค์กรธุรกิจในการกรองจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมได้อย่างมีประสิทธิภาพ พบว่า วิธี Naive Bays ให้ค่าความถูกต้องสูงที่สุด และแตกต่างกับงานวิจัยของประพัฒน์และคณะ [4] ที่เสนอการแบ่งกลุ่มข้อความจากข้อความรีวิว โดยใช้เทคนิคเหมืองข้อมูล ด้วยเทคนิค SVM, Decision Tree, Naive Bays, K-Nearest Neighbor พบว่าเทคนิค SVM ได้ค่าความถูกต้องสูงที่สุด เป็นเทคนิคที่น่าสนใจ ที่จะนำไปประยุกต์ใช้ในการจำแนกกลุ่มข้อความวิจารณ์ภาพยนตร์ ยังสามารถนำไปประยุกต์ใช้ในงานข้อมูลรีวิวสินค้าและข้อมูลการรีวิวโรงแรมเป็นต้น จากการทบทวนงานวิจัยที่เกี่ยวข้องแม้จะมีการเปรียบเทียบอัลกอริทึมจากงานวิจัยต่าง ๆ ก็ไม่สามารถระบุได้ว่า อัลกอริทึมใดที่จำแนกประเภทมูลได้ดีที่สุด

จากเหตุผลดังกล่าว งานวิจัยฉบับนี้จึงเสนอการวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูลด้วยเทคนิควิธี Decision Tree, Naive Bayes และ K-Nearest Neighbor เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องในการสังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรีและเปรียบเทียบประสิทธิภาพของโมเดลเพื่อหาค่าความแม่นยำที่เหมาะสมที่สุดในการทำนายการลาออกกลางคันของนักศึกษา เพื่อนำไปใช้เป็นแนวทางในการป้องกันและแก้ไขปัญหาการลาออกกลางคันของนักศึกษาในแต่ละปีการศึกษาต่อไป

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อวิเคราะห์หาปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี

2.2 เพื่อสังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี

2.3 เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของโมเดลด้วยเทคนิควิธี Decision Tree, K-Nearest Neighbors (K-NN), Naive Bayes

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 ต้นไม้ตัดสินใจ (Decision Tree) [5] เป็นการเรียนรู้โดยการจำแนกประเภท (Classification) ข้อมูลออกเป็นกลุ่ม (Class) ต่างๆ โดยใช้คุณลักษณะ หรือคุณสมบัติ (Attribute) ข้อมูลในการจำแนกประเภท ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่า คุณลักษณะใดเป็นตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวมีความสำคัญมากน้อยต่างกันอย่างไรโมเดลด้วยวิธี Decision Tree จะทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของ tree เรียกว่า โหนดราก (Rood Node) จากนั้นจะเลือกคุณสมบัติที่มีความสัมพันธ์ถัดไปเรื่อย ๆ จากการคำนวณ Information Gain (IG) โดยเลือกคุณสมบัติที่มีค่า IG สูงที่สุด คำนวณได้จากสมการดังนี้

$$\text{IG(parent, child)} = \text{Entropy(parent)} - [\text{p(c1)} \times \text{Entropy(c1)} + \text{p(c2)} \times \text{Entropy(c2)} + \dots] \quad (1)$$

เมื่อ

Entropy(c1) คือ $-\text{p(c1)} \log \text{p(c1)}$

p(c1) คือ ค่าความน่าจะเป็นของค่า $_{c1}$

$_{c}$ คือ 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

3.2 การเรียนรู้เบย์ (Naive Bayes) [6] จะใช้วิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น โดยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อนนิยมใช้ เนื่องจากเป็นรูปแบบการหาความสัมพันธ์ที่ไม่ซับซ้อนได้ผลลัพธ์ดีดังสมการ

$$\text{P(C|A)} = \frac{\text{P(A|C)} \times \text{P(C)}}{\text{P(A)}} \quad (2)$$

จากสมการ Bayes อธิบายว่าถ้าต้องการทำนายคลาส C เมื่อทราบแอททริบิวต์ สามารถคำนวณจากความน่าจะเป็นของแอททริบิวต์ A ที่มีคลาส C ใน Training data และค่าความน่าจะเป็นของแอททริบิวต์ A และ C มีสมการ 3 ส่วนดังนี้

P(C|A) คือ ค่าความน่าจะเป็นที่ข้อมูลที่มีแอททริบิวต์ A จะมีคลาส C

P(A|C) คือ ค่าความน่าจะเป็นที่ข้อมูล Training data ที่มีคลาส C และมีแอททริบิวต์ A โดยที่ A = $a_1 \cap a_2 \dots \cap a_M$ โดยที่ M คือจำนวนแอททริบิวต์ใน Training data

P(C) คือ ความน่าจะเป็นของคลาส C

P(A) คือ ความน่าจะเป็นของคลาส A

แต่การที่แอททริบิวต์ A = $a_1 \cap a_2 \dots \cap a_M$ ที่เกิดขึ้นใน Training data อาจจะมีจำนวนน้อยมากหรือไม่มีรูปแบบของแอททริบิวต์แบบนี้เกิดขึ้นเลย ดังนั้นจึงได้ใช้หลักการที่ว่าแต่ละแอททริบิวต์เป็น Independent ต่อกันทำให้สามารถเปลี่ยนสมการ P(A|C) ได้เป็น

$$\text{P(A|C)} = \text{P}(a_1|\text{C}) \times \text{P}(a_2|\text{C}) \times \dots \times \text{P}(a_M|\text{C}) \quad (3)$$

โดยสามารถตัดส่วนของ P(A) ออกได้เนื่องจากเป็นส่วนของการปรับค่าให้อยู่ในช่วงนั้น (Normalization)

3.3 วิธีการค้นหาเพื่อนบ้านใกล้ที่สุด K-Nearest Neighbors [7] เป็นการแบ่งกลุ่มข้อมูล และทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ตัว และคำตอบที่ทำนายได้ดีคือคลาสที่พบมากที่สุดของข้อมูลที่เป็นเพื่อนบ้านทั้ง K ตัว มักจะใช้วิธีการวัดระยะห่างแบบ Euclidean เกิดจากรากที่สองของผลต่างระหว่างแอททริบิวต์ต่าง ๆ ยกกำลังสองดังสมการ

$$\text{Euclidean Distance} = \sqrt{((x_1 - y_1)^2) + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2)} \quad (4)$$

โดยที่ x_1 คือ แอททริบิวต์ที่ 1 ของข้อมูลจุดที่ 1 และ y_1 คือ แอททริบิวต์ที่ 1 ของข้อมูลชุดที่ 2 โดยข้อมูลทั้ง 2 ตัว (x และ y) มีจำนวนแอททริบิวต์เท่ากับ L

3.4 วิธีการ 10-Fold Cross Validation [8] การวัดประสิทธิภาพของโมเดลการพยากรณ์ที่สร้างด้วย วิธีการ 10-Fold Cross Validation จะแบ่งข้อมูลออกหลายส่วน (แสดงด้วยค่า k) ในการดำเนินวิจัยครั้งนี้จะแบ่งข้อมูลเป็น 10 ส่วน โดยแต่ละส่วนมีจำนวนข้อมูลเท่ากัน จากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดลทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ ประสิทธิภาพของโมเดล 10 ครั้ง

3.5 ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล [5] โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัยและการทำงานต่าง ๆ อยู่ 4 ค่า คือ

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

3.6 วิเคราะห์ปัจจัยที่ส่งผลต่อการลาออกกลางคันของนักศึกษาระดับปริญญาตรี [5] เพื่อกำหนดเป็นดัชนีและเรียงลำดับความสำคัญของดัชนี โดยการลดมิติข้อมูล (Attribute Selection) แบบ Filter approach เป็นการคำนวณค่าน้ำหนัก (หรือค่าความสัมพันธ์) ของแต่ละแอททริบิวต์และเลือกเฉพาะแอททริบิวต์ที่สำคัญเก็บไว้ เช่น Information Theory, Chi-Square ในบทความนี้เป็นการคำนวณค่าน้ำหนักของแต่ละแอททริบิวต์ด้วยวิธีการ Information Theory

3.7 งานวิจัยที่เกี่ยวข้อง

สุภาวดีและสมบรูณ์ [9] พยากรณ์ผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐานของนักเรียนชั้นประถมศึกษาปีที่ 6 โรงเรียนอนุบาลสิงห์บุรี โดยใช้เทคนิคเหมืองข้อมูล (Data Mining) หาความสัมพันธ์ระหว่างผลการสอบแต่ละรายวิชาของนักเรียน โดยรวบรวมผลการสอบ O-Net และผลการเรียนแต่ละรายวิชาที่เกี่ยวข้องมาใช้สร้างโมเดลการพยากรณ์ ผู้วิจัยได้นำเทคนิคมาเปรียบเทียบด้วยกัน 3 โมเดล คือ Decision Tree, K-Nearest Neighbors, Naive Bayes เพื่อหาวิธีการที่มีความถูกต้องมากที่สุด นำมาพยากรณ์ผลการสอบ O-Net ล่วงหน้า และนำผลการทำนายที่ได้ไปใช้ในการพัฒนาการเรียนการสอนวิธีการสอนและการสอนเสริม พบว่าโมเดล ส่วนใหญ่ใช้เทคนิควิธี Decision Tree สร้างตัวแบบการพยากรณ์ในวิชาภาษาไทยมีค่าเฉลี่ยความถูกต้องสูงสุดร้อยละ 74.14 วิชาวิทยาศาสตร์มีค่าเฉลี่ยความถูกต้องสูงสุดร้อยละ 81.46 ส่วนวิชาอังกฤษใช้เทคนิควิธี K-Nearest Neighbors มีค่าเฉลี่ยความถูกต้องสูงสุดร้อยละ 80.44 จากการทดสอบ 10 ครั้ง จะเห็นได้ว่าวิชาภาษาไทย วิชาภาษาอังกฤษ และวิชาวิทยาศาสตร์มีประสิทธิภาพให้ความถูกต้องมากที่สุด

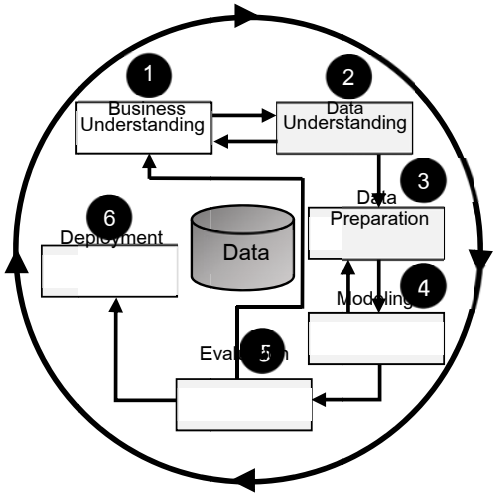
โชติกาและขวัญทัย [2] การสร้างแบบจำลองเพื่อในการพยากรณ์ประสิทธิภาพของผู้เรียนโดยใช้เทคนิควิธีการทำเหมืองข้อมูลสำหรับการจำแนกประเภทข้อมูลและการทำนาย (Classification and Prediction) ผู้วิจัยเลือกใช้อัลกอริทึมในการทำเหมือนข้อมูลจำนวน 3 อัลกอริทึมได้แก่ Decision Tree, Naive Bays, Sequential minimal optimization กับชุดข้อมูลผู้เรียนจำนวนทั้งสิ้น 338 คน ในขั้นตอนการหาประสิทธิภาพในแต่ละอัลกอริทึม ผู้วิจัย ใช้ 10-fold Cross-validation เพื่อใช้สำหรับวัดค่าประสิทธิภาพของแบบจำลองสำหรับค่าความถูกต้องของการจำแนกกลุ่ม (Accuracy) เป็นดัชนีชี้วัดประสิทธิภาพโดยรวมของอัลกอริทึม และใช้ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure) เพื่อเป็นดัชนีชี้วัดประสิทธิภาพของแต่ละคลาสคำตอบ (Target Class) ที่ได้จากแบบจำลองการพยากรณ์ของผู้เรียน ผลการวิจัยพบว่า Decision Tree มีประสิทธิภาพการจำแนกข้อมูลโดยรวมสูงที่สุดโดยมีค่าความถูกต้องในการจำแนกกลุ่มผู้เรียน (Accuracy) สูงสุดที่ระดับ 85.80%

จุฑาทิพย์และนิเวศ [3] การเปรียบเทียบประสิทธิภาพการจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยใช้เทคนิคการทำเหมืองข้อมูล ซึ่งประกอบด้วย Decision Tree, Naive Bayes, K-Nearest Neighbor ผลการทดลองเมื่อเปรียบเทียบจากค่าความถูกต้อง (Accuracy) พบว่าวิธี Naive Bayes ให้ค่าความถูกต้องสูงสุดเท่ากับ 92.48 %

4. วิธีการดำเนินการวิจัย

ในงานวิจัยนี้ได้เสนอวิธีการทำเหมืองข้อมูลโดยใช้โปรแกรม Rapid Miner Studio 7 โดยขั้นตอนการทำ

เหมืองข้อมูลแบบ CRISP-DM เป็น Workflow มาตรฐานสำหรับการทำ Data mining ประกอบด้วย 6 ขั้นตอนดังแสดงในรูป โดยแต่ละขั้นตอนจะเป็นขั้นตอนที่ต่อเนื่องกัน คือ ขั้นตอนต่อไปต้องมีผลลัพธ์จากขั้นตอนก่อนหน้าด้วยลูกศรที่เชื่อมระหว่างแต่ละขั้นตอน เช่น จากผลลัพธ์จากขั้นตอนการเตรียมข้อมูล (Data Preparation) แล้วสามารถสร้างโมเดลจำแนกประเภทข้อมูลในขั้นตอน Modeling และอาจทำซ้ำเพื่อข้อมูลที่ถูกต้องมากที่สุด



รูปที่ 1 ขั้นตอนในกระบวนการ CRISP-DM [10]

4.1 ความเข้าใจธุรกิจ (Business Understanding) การเข้าใจปัญหาและทำการวิเคราะห์เหมืองข้อมูล เพื่อศึกษาปัจจัยสาเหตุและแนวทางแก้ไขปัญหาการลาออกกลางคันของนักศึกษาโดยการวิเคราะห์การทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคการทำเหมืองข้อมูลด้วยเทคนิควิธี คือ 1) Decision Tree 2) K-Nearest Neighbors และ 3) Naive Bayes เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องในการสังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรีและเปรียบเทียบประสิทธิภาพของโมเดล เพื่อใช้เป็นแนวทางในการป้องกันและแก้ไขปัญหาการลาออกกลางคันของนักศึกษาในแต่ละปีการศึกษาต่อไป ผู้วิจัยได้ทำการวิเคราะห์โดยใช้ชุดข้อมูลของนักศึกษาระดับปริญญาตรี มหาวิทยาลัยราชภัฏอุบลราชธานี จังหวัดอุบลราชธานี ตั้งแต่ปีการศึกษา 2558-2560 เมื่อได้ข้อมูลแล้วให้เตรียมข้อมูลเพื่อให้พร้อมที่จะนำไปทำการคัดกรองภายใต้หลักการทำงานของเหมืองข้อมูล

4.2 ความเข้าใจข้อมูล (Data Understanding) ทำการรวบรวมข้อมูลพื้นฐานของนักศึกษาจากฐานข้อมูลงานทะเบียนของมหาลัยวิทยาลัยราชภัฏอุบลราชธานี จังหวัดอุบลราชธานี ซึ่งมี 11 แอททริบิวต์ และ คัดกรองเฉพาะข้อมูลที่มีความสมบูรณ์ทั้งหมด 13,729 ชุดข้อมูล และทำการคัดกรองแอททริบิวต์เพื่อให้ได้แอททริบิวต์ที่จำเป็นที่สุด สำหรับการจำแนกประเภทเพื่อหาประสิทธิภาพในการจำแนกข้อมูลและการสร้างความ สัมพันธ์ของแอททริบิวต์

4.3 การเตรียมข้อมูล (Data Preparation) หลังจากเสร็จขั้นตอน Data Understanding แล้วในขั้นตอนนี้เป็นขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุด เนื่องจากโมเดลที่ได้จากการทำ Data mining จะให้ผลลัพธ์ที่ถูกต้องหรือไม่นั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ ผู้วิจัยทำการเตรียมข้อมูลที่ได้เก็บรวบรวมมาซึ่งอยู่ในรูปแบบของตารางใน Excel โดยเริ่มจาก

4.3.1 ทำการคัดเลือกข้อมูล (Data Selection) เป็นการเลือกใช้เฉพาะข้อมูลที่เกี่ยวข้อง งานวิจัยในครั้งนี้จะประกอบไปด้วยคณะที่เรียน สาขาวิชา เพศของนักศึกษา โรงเรียนเดิมก่อนเข้ารับการศึกษา อาชีพของบิดาและมารดา การกู้ยืมเงินกองทุนเพื่อการศึกษา เกรดเฉลี่ยรวม จำนวนพี่น้องที่กำลังศึกษา อายุ และ สถานะการศึกษา ตั้งแต่ปีการศึกษา 2558-2560 ซึ่งมี 11 แอททริบิวต์

ตารางที่ 1 แสดงรายชื่อแอททริบิวต์ที่ใช้ในทำนายการออกกลางคันของนักศึกษา

No	Attribute	Direction
1.	คณะที่เรียน (Id_faculty)	Input
2.	สาขาวิชา (Major)	Input
3.	เพศ (Sex)	Input
4.	ประเภทโรงเรียนเดิม (Old_school)	Input
5.	อาชีพของบิดา (Father_occ)	Input
6.	อาชีพของมารดา (Mother_occ)	Input
7.	การกู้ยืมกองทุนเพื่อการศึกษา (Edu_Loan)	Input
8.	เกรดเฉลี่ย (GPA)	Input
9.	จำนวนพี่น้องที่ศึกษา (Num_brother)	Input
10.	อายุ (Age)	Input
11.	สถานะนักศึกษา (Status)	output

4.3.2 ทำการกลั่นกรองข้อมูล (Data Cleaning) ในขั้นตอนนี้จะทำการปรับข้อมูลให้ถูกต้อง ซึ่งอยู่ในรูปแบบของตาราง Excel และทำการลบข้อมูลซ้ำซ้อน แก้ไขข้อมูลที่ผิดพลาด ข้อมูลผิดรูปแบบ ข้อมูลที่มีค่าว่าง

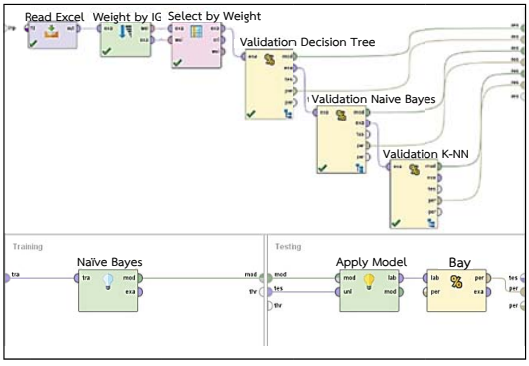
(Missing Values) ข้อมูล (Outlier) ที่แปลกแยกจากส่วนอื่น

4.3.3 ทำการแปลงรูปแบบของข้อมูล (Data Transformation) เมื่อปรับข้อมูลแล้ว ในขั้นตอนนี้จะเป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ข้อมูลตามอัลกอริทึมของ Data mining ที่เลือกใช้ เนื่องจากข้อมูลที่นำมาวิเคราะห์ในครั้งนี้ มีรูปแบบในการจัดเก็บทั้งประเภท Nominal และ Numeric จึงต้องแปลงให้อยู่ในรูปแบบเดียวกัน เช่น ประเภทโรงเรียนเดิม โรงเรียนเป็น School วิทยาลัยเป็น College กศน. เป็น NFE และทำการ Clustering Data แบ่งข้อมูลหลาย ๆ กลุ่มตามความคล้ายคลึง เช่น เกรดเฉลี่ย จะแบ่งออกเป็น 4 ช่วง 0.00-1.00 เป็น VeryLow 1.01-2.00 เป็น Low 2.01-3.00 เป็น Middle 3.01-4.00 เป็น High จากชุดข้อมูลทดสอบคำตอบ (Class) จะแบ่งเป็น 2 ประเภทคือ Yes เป็นลาออก มีจำนวน 2,364 ชุดข้อมูล และ No เป็นไม่ลาออก มีจำนวน 13,729 ชุดข้อมูลโดยมีรายละเอียดดังแสดงในตารางที่ 2

ตารางที่ 2 การแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ข้อมูล

แอททริบิวต์	รายละเอียด
Id_faculty	คณะสาธารณสุขศาสตร์ = Public, คณะวิทยาศาสตร์ = Science, คณะวิทยาการคอมพิวเตอร์= Comscience, คณะมนุษยศาสตร์และสังคมศาสตร์ = Human, คณะแพทย์แผนไทยและแพทย์ทางเลือก = Thai traditional, คณะพยาบาลศาสตร์ = Nurse, คณะบริหาร ธุรกิจและการจัดการ = Business, คณะนิติศาสตร์ = Law, คณะเทคโนโลยีอุตสาหกรรม = Techno, คณะครุศาสตร์ = Edu, คณะเกษตรศาสตร์ = Agriculture
Sex	ชาย = Male, หญิง = Female
Old_school	โรงเรียน = School, วิทยาลัย = College, กศน. = NFE
Father_occ	เกษตรกรรม = Agriculturalist, ลูกจ้างหรือรับจ้างทั่วไป = Employee, ค้าขายหรือธุรกิจส่วนตัว = Merchant, ข้าราชการ = Government officer, รัฐวิสาหกิจ = State Enterprise
Mother_occ	เกษตรกรรม = Agriculturalist, ลูกจ้างหรือรับจ้างทั่วไป = Employee, ค้าขายหรือธุรกิจส่วนตัว = Merchant, ข้าราชการ = Government officer, รัฐวิสาหกิจ = State Enterprise
Edu_Loan	กู้ยืม = Yes, ไม่กู้ยืม = No
GPA	0.00-1.00 = VeryLow, 1.01-2.00 = Low, 2.01-3.00 = Middle, 3.01-4.00 = High
Num_brother	1 คน = One, 2 คน = Two, 3 คนขึ้นไป = Threeup
Age	18 -19 = One, 20 -21= Two, 22 -23 = Three, 24 ปี ขึ้นไป = Fourup
Status	ลาออก = Yes, ไม่ลาออก = No

4.4 การสร้างโมเดล/แบบจำลอง (Modeling) การสร้างและการทดสอบความถูกต้องของโมเดลเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิค Data mining ทำการสังเคราะห์โมเดลจากข้อมูลที่มีอยู่เพื่อการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี หาค่าความถูกต้อง (Accuracy) ที่ออกมาเป็นตัวเลข เครื่องมือที่ใช้ในการทำเหมืองข้อมูลใช้โปรแกรม Rapid Miner Studio 7 เป็นเครื่องมือในการวิเคราะห์ข้อมูล ในการวิจัยครั้งนี้ผู้วิจัยวิเคราะห์ปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาด้วย Filter approachวิเคราะห์ค่าน้ำหนักของแอททริบิวต์ด้วยวิธีการ Information Theory จากนั้นจึงนำปัจจัยที่ได้ในข้อ 3.6 มาทำการสร้างโมเดลการทำนายด้วยเทคนิคเหมืองข้อมูล 3 โมเดล ด้วยเทคนิควิธีคือ 1) Decision Tree 2) K-Nearest Neighbors และ 3) Naive Bayes เพราะเหมาะสมกับการจำแนกประเภทข้อมูลประเภทตัวอักษร (Text Classification) ทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold Cross Validation ในการวัดประสิทธิภาพของการจำแนกประเภทข้อมูลได้แก่ การหาค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-measure) และค่าความถูกต้อง (Accuracy)



รูปที่ 2 สร้างโมเดลด้วย Rapid Miner Studio 7

4.5 การประเมินผล (Evaluation) จะได้ผลจากการสร้างโมเดล พบว่ามีจำนวน 8 ปัจจัยที่เกี่ยวข้องในการ

ลาออกกลางคันของนักศึกษา โมเดลที่สร้างด้วยเทคนิควิธี Naive Bayes ให้ผลลัพธ์ที่ดีที่สุดโดยมีค่าความถูกต้องร้อยละ 93.58

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพในการจำแนกการลาออกกลางคันของนักศึกษา

อัลกอริทึม	ค่าความถูกต้อง Accuracy
Naive Bayes	93.58%
Decision tree	93.52%
k-NN	87.95%

4.6 การนำแบบจำลองไปใช้งาน (Deployment) นำโมเดลที่ได้ผ่านกระบวนการเปรียบเทียบประสิทธิภาพของโมเดล ที่ได้ค่าความถูกต้องสูงสุด ไปใช้ในการทำนายการลาออกกลางคันของนักศึกษาเพื่อใช้เป็นแนวทางในการป้องกันและแก้ไขปัญหาการลาออกกลางคันของนักศึกษาในแต่ละปีการศึกษาต่อไป

5. ผลการวิจัย

5.1 ผลการวิเคราะห์ปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยการลดมิติข้อมูล (Attribute Selection) การคำนวณค่าน้ำหนักของแอททริบิวต์ด้วยวิธีการ Information Theory พบว่ามีจำนวน 8 ปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษา ได้แก่ การกู้ยืมกองทุนเพื่อการศึกษา สาขาวิชา เกรดเฉลี่ย อาชีพของมารดา อาชีพของบิดา คณะที่เรียน อายุ และโรงเรียนเดิมรายละเอียดดังตารางที่ 4

ตารางที่ 4 ปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษา

No	Attribute	Weight
1.	การกู้ยืมกองทุนเพื่อการศึกษา (Edu_Loan)	1.6
2.	สาขาวิชา (Major)	1.2
3.	เกรดเฉลี่ย (GPA)	0.8
4.	อาชีพของมารดา (Mother_occupation)	0.5
5.	อาชีพของบิดา (Father_occupation)	0.5
6.	คณะที่เรียน (Id_faculty)	0.4
7.	อายุ (Age)	0.1
8.	โรงเรียนเดิม (Old_school)	0.1

5.2 ผลการสังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรีและวัดประสิทธิภาพของโมเดลด้วยวิธีการ 10-Fold Cross Validation

ตารางที่ 5 ผลการเปรียบเทียบประสิทธิภาพค่าความแม่นยำของโมเดล

อัลกอริทึม	Accuracy	Precision	Recall
Naive Bayes	93.58%	93.80%	99.40%
Decision tree	93.52%	94.66%	98.19%
k-NN	87.95%	98.34%	73.18%

จากตารางที่ 5 จะเห็นว่าสามารถใช้เหมืองข้อมูลในการทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรี ด้วยเทคนิคเหมืองข้อมูล 3 โมเดล ด้วยเทคนิควิธีคือ 1) Decision Tree 2) K-Nearest Neighbors และ 3) Naive Bayes เพื่อวัดประสิทธิภาพของโมเดลด้วยวิธีการ 10-Fold Cross Validation โมเดลที่สร้างด้วยเทคนิควิธี Naive Bayes มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้องร้อยละ 93.58 เทคนิควิธี Decision tree มีค่าเฉลี่ยความถูกต้องร้อยละ 93.52 และ เทคนิควิธี K-Nearest Neighbors มีค่าเฉลี่ยความถูกต้องร้อยละ 87.95

6. อภิปรายผล

จากการรวบรวมข้อมูลจากฐานข้อมูลงานทะเบียนของมหาลัยวิทยาลัยราชภัฏอุบลราชธานี ของนักศึกษาระดับปริญญาตรี มีจำนวนนักศึกษาที่ลาออกจำนวน 2,364 คน จำนวนนักศึกษาที่ไม่ลาออกมีจำนวน 13,729 คน และเมื่อทำการวิเคราะห์ค่าน้ำหนักของแอททริบิวต์ด้วยวิธีการ Information Theory พบว่ามีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันสูงสุด 5 อันดับ ได้แก่ การกู้ยืมกองทุนเพื่อการศึกษา สาขาวิชา เกรดเฉลี่ อาชีพของมารดา และอาชีพของบิดา เพราะเหตุนี้ผู้บริหารจะต้องเข้าใจถึงสภาพการเป็นอยู่ในครอบครัวของนักศึกษาและให้ความช่วยเหลือทางด้านการเงินแก่นักศึกษาที่ขาดแคลนทุนทรัพย์ โดยอาจจะพิจารณาผ่อนผันการชำระค่าเทอม พิจารณาให้กู้ยืมกองทุนเพื่อการศึกษา พิจารณาให้ทุนการศึกษาสำหรับผู้เรียนดีแต่ยากจนหรือทุนทางด้านกิจกรรม รวมไปถึงการแก้ปัญหาทางด้านการเรียนของนักศึกษา ในแต่ละสาขาวิชา เพื่อให้ผลการเรียนดีขึ้น ดังนั้นอาจารย์ประจำสาขาวิชาจะต้องคอยดูแลนักศึกษาอย่างใกล้ชิดเพื่อลดความเสี่ยงในการลาออกกลางคันของนักศึกษา

การเปรียบเทียบประสิทธิภาพการจำแนกการลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิคเหมืองข้อมูล 3 โมเดล ด้วยเทคนิควิธี คือ 1) Decision Tree 2) K-Nearest Neighbors และ 3) Naive Bayes เพื่อวัดประสิทธิภาพของโมเดลด้วยวิธีการ 10-Fold Cross Validation ผลการทดลองเมื่อเปรียบเทียบจากค่าความถูกต้อง (Accuracy) พบว่า Naive Bayes มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้องร้อยละ 93.58 ดังนั้นจึงสามารถสรุปได้ว่า แบบโมเดลที่ได้จากการทดลองนี้สามารถนำไปวิเคราะห์และทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรีได้ถูกต้องในระดับที่ยอมรับได้ เมื่อเปรียบเทียบกับงานวิจัยที่อ้างอิงถึง จะใช้วิธีการเหมืองข้อมูลที่แตกต่างกัน เช่น สุภาวดีและสมบูรณ์ [9] การพยากรณ์ผลการทดสอบทางการศึกษาระดับชาติขั้นพื้นฐาน (O-Net) มหาวิทยาลัยรังสิต ได้นำเสนอวิธีการทำเหมืองข้อมูลโดยวิธี Decision Tree เป็นวิธีการที่ดีที่สุด และ จุฑาทิพย์และนิเวศ [3] การจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยใช้เทคนิคการทำเหมืองข้อมูล ซึ่งประกอบด้วย Decision Tree, Naive Bayes, K-Nearest Neighbor ผลการทดลองเมื่อเปรียบเทียบจากความค่าถูกต้อง (Accuracy) พบว่าวิธี Naive Bayes ให้ค่าความถูกต้องสูงสุดเท่ากับ 92.48 % จากงานวิจัยนี้พบว่า การวิเคราะห์และทำนายการลาออกกลางคันของนักศึกษาวิธีการที่ดีที่สุดคือ เทคนิควิธีการของนาอีฟเบย์ (Naive Bayes) มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้องร้อยละ 93.58 ซึ่งการใช้วิธีการทำเหมืองข้อมูลจะให้มีประสิทธิภาพมากที่สุด จะต้องทดสอบด้วยกันหลายวิธี เนื่องจากความถูกต้องที่ทำนายได้ขึ้นอยู่กับข้อมูลที่จะนำมาใช้วิเคราะห์ด้วย

7. ข้อเสนอแนะ

การวิเคราะห์ข้อมูลให้ละเอียดหรือให้มีประสิทธิภาพมากยิ่งขึ้น ควรมีการเก็บรวบรวมข้อมูลปัจจัยตัวอื่นๆ ร่วมด้วย เช่น ปัจจัยด้านส่วนตัวของนักศึกษา ปัจจัยด้านสถานศึกษา ปัจจัยด้านสภาพครอบครัว ปัจจัยด้านสังคม ข้อมูลพฤติกรรมการเรียนของนักศึกษา ข้อมูลสภาพการเรียนและการสอน ข้อมูลโครงสร้างของหลักสูตร ข้อมูลเหล่านี้จะต้องจัดเก็บเพิ่มเติมโดยอาจจะใช้เครื่องมืออื่น เช่น แบบสอบถาม การสัมภาษณ์ เป็นต้น

8. เอกสารอ้างอิง

[1] Sakkarin Phupanna. (2017). "Predictive Analytic for Student Dropout in High Vocational Certificate Using Data Mining Technique." The 13Th National Conference on Computing and Information Technology King Mongkut's University of Technology North Bangkok. 6-7 July 2017. Bangkok. (51-56). (in Thai)

[2] Chotika Lamprik and Kwanruetai Sudadet. (2017). "Predicting Student's Data for Performance by Data Mining." The 13Th National Conference on Computing and Information Technology King Mongkut's University of Technology North Bangkok. 6-7 July 2017. Bangkok. (32-37). (in Thai)

[3] Jutathip Thipphol and Nivet Jirawichitchai. (2016). "Email Spam Classification Using Data Mining Techniques." Journal of Science and Technology, Rajamangala University of Technology Thanyaburi. Vol.6 No.1 : 102-109. (in Thai)

[4] Prapat Promnamang, Vasuvat Pongkhajorn and Nivet Jirawichitchai (2016). "Text Review using data mining Classification Technique." Journal of Science and Technology, Rajamangala University of Technology Thanyaburi. Vol.6 No.1 : 94-101. (in Thai)

[5] Eakasit Pacharawongsakda.(2014). An Introduction to Data Mining Techniques. Bangkok : Asia Digital Printing Co., Ltd. (in Thai)

[6] Thanapon Panprom and Somboon Anekritmongkol. (2016). "A comparison of the effectiveness of the data mining techniques in the prediction of the results in Medical Technology license examination : A case study of Medical Technology

graduates, Rangsit University." Journal of the Thai Medical Informatics Association. Vol.3 No.1 : 32-40. (in Thai)

[7] Thada Jantakoon. (2016). "Classification Model for Selection of Program Studies in Faculty of Information Technology in Rajabhat MahaSarakham University Using Data Mining Technique." The 9th National Conference on Technical Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok. 24 November 2016. Bangkok. (336-343). (in Thai)

[8] Kitsana Waiyamai, Chidchanok Songsiri and Thanawin Raktummanon. (2001). "The use of data mining techniques to improve the quality of education." Journal of National Electronics and Computer Technology Center. Vol.3 No.11 : 134-142. (in Thai)

[9] Supawadee Juntasuwan and Somboon Anekritmongkol. (2017). "The Prediction Model for Ordinary National Educational Test Results through Using Data Mining Technique Case Study of Grade 6 Students at Anubansingburi School." the 4th International Conference on Management Science, Innovation, and Technology. 16 June 2017. Bangkok. (552-566). (in Thai)

[10] Foster Provost. (2013). Data Science for Business : What you need to know about data Mining and data-analytic thinking. America : O'Reilly Media, Inc. cited in Eakasit Pacharawongsakda. (2016). Practical Data Mining with RapidMiner Studio7. Bangkok : Asia Digital Printing Co., Ltd.