SNRU JST
SNRU Journal of Science and Technology
Journal home page: snrujst.snru.ac.th

การวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาที่มีผลการเรียนปกติโดยใช้ต้นไม้ตัดสินใจ

Analysis on factors affecting normal-grade student dismissal using decision tree

จีระนันต์ เจริญรัตน์*

Jeeranan Chareonrat*

สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏสกลนคร สกลนคร 47000 ประเทศไทย

* **Corresponding Author**: jeechar@hotmail.com

Received: 28 November 2015; **Revised:** 20 December 2015; **Accepted:** 25 December 2015; **Available online:** 1 August 2016

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ วิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาระดับปริญญาตรี ที่มีผลการเรียนของเกรดเฉลี่ยสะสมตั้งแต่ 2.00 ขึ้นไป ในการวิเคราะห์ข้อมูลจะใช้ข้อมูลนักศึกษาระดับปริญญาตรี คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏสกลนคร ที่เข้าศึกษาระหว่างปี พ.ศ. 2553-2557 มีจำนวน 3,385 ชุดข้อมูล เลือกใช้เทคนิคการทำเหมืองข้อมูล แบบ Classification เลือกการทำนายข้อมูลด้วยวิธี Decision Tree และใช้อัลกอริธึมชนิด J48 การทดสอบโมเดลที่ได้จะทำการทดสอบแบบ 10-fold cross validation โดยใช้โปรแกรม WEKA ผลการวิเคราะห์พบว่าปัจจัยที่มีความสำคัญที่จะส่งผลต่อการพ้นสภาพของนักศึกษาที่มีผลการเรียนปกติ แบ่งเป็น 4 กลุ่ม ตามกลุ่มเกรดเฉลี่ยสะสม ดังนี้ กลุ่ม Best (GPA>3.50 ขึ้นไป) ปัจจัยคือ วุฒิการศึกษาเดิม กลุ่ม Excellent (GPA = 3.00-3.50) ปัจจัยคือ อาชีพมารดา และสาขาวิชาที่เรียน กลุ่ม Good (GPA = 2.50-2.99) ปัจจัยคือ ทุนกู้ยืมเพื่อการศึกษา สถานภาพของครอบครัว รายได้ของบิดา รายได้ของมารดา และจังหวัด กลุ่ม Medium (GPA = 2.00-2.49) ปัจจัยคือ ทุนกู้ยืมเพื่อการศึกษา สถานภาพครอบครัว และรายได้ของมารดา จากผลการวิจัยนี้สามารถนำกฎการจำแนกข้อมูลที่ได้มาใช้ในการพัฒนาระบบพยากรณ์การพ้นสภาพของนักศึกษาต่อไป และผู้บริหารสามารถใช้เป็นแนวทางในการบริหารจัดการหรืออาจารย์ที่ปรึกษาสามารถวางแผนการเรียน ดูแลการลงทะเบียนเรียนของนักศึกษา ให้ความช่วยเหลือ และส่งเสริมนักศึกษาได้อย่างเหมาะสม

คำสำคัญ: เหมืองข้อมูล; ต้นไม้ตัดสินใจ; การพ้นสภาพของนักศึกษา

Abstract

This research aimed to seek factors influencing student dismissal targeted 3,385 Management Science individual students obtained GPA of more than 2.00 whose studied during A.D. 2010 to A.D. 2014 for a Bachelor Degree. Data buildup was done by using The Classification Data Mining method then forecasted through The Decision Tree and J 48 Algorithm Techniques. Model testing was based on The 10 - fold Cross Validation with Weka Program. The findings divided students into 4 groups of which responded to different factors namely earlier educational attainment for group Best with GPA more than 3.50, student' s mother's occupation and field of study for group Excellent with GPA between 3.00 to 3.49, studying loans, family status, parents' incomes and province of birth for

group Good GPA between 2.50 to 2.99, and studying loans, family status, and the mother incomes for group Medium with GPA between 2.00 to 2.49. The results ensured the validity of the Data Classification Rules in predicting such dismissals and gave the school management ways to control and support both its own work and students learning plans.

Keywords: Data Mining; Decision Tree; Student Dismissal

1. บทนำ

ในปัจจุบันสถาบันการศึกษาทั้งภาครัฐและเอกชน ได้เห็นถึงความสำคัญของความเจริญก้าวหน้าอย่างรวดเร็วของเทคโนโลยีสารสนเทศและการสื่อสาร โดยได้นำเอาระบบคอมพิวเตอร์มาใช้ในการดำเนินงานทั้งทางด้านบริหารจัดการและการจัดการเรียนการสอนให้เกิดประสิทธิภาพ เช่น ระบบบริหารงบประมาณ ระบบบริหารงานพัสดุ ระบบฐานข้อมูลบุคลากร ระบบทะเบียนนักศึกษา และระบบลงทะเบียนออนไลน์ เป็นต้น ซึ่งระบบเหล่านี้ได้มีการเก็บข้อมูลไว้ที่ฐานข้อมูลมาอย่างต่อเนื่องและมีปริมาณมาก แต่ข้อมูลเหล่านั้นยังไม่ได้ถูกนำมาใช้ประโยชน์เท่าที่ควร ทั้งที่ข้อมูลนั้นน่าสนใจและสามารถนำมาใช้ในการสืบค้นความรู้ที่เป็นประโยชน์ได้ โดยการนำข้อมูลนักศึกษาในอดีตมาวิเคราะห์เพื่อสร้างต้นแบบในการพยากรณ์หรือทำนายแนวโน้มในอนาคต เช่น พยากรณ์โอกาสการสำเร็จการศึกษาของนักศึกษา หรือ การทำนายสถานภาพของนักศึกษา หรือพยากรณ์แนวโน้มการพ้นสภาพของนักศึกษา โดยใช้เทคนิคการทำเหมืองข้อมูล ซึ่งกระบวนการในการค้นหารูปแบบและความรู้ที่น่าสนใจจากข้อมูลที่มีปริมาณมากนั้นเรียกว่าเทคนิคการทำเหมืองข้อมูล (data mining) [1] เป็นเทคนิคที่ใช้ในการจัดการกับข้อมูลขนาดใหญ่ [2] ที่มีหลากหลายเทคนิคด้วยกัน ได้แก่ เทคนิคการหากฎความสัมพันธ์ของข้อมูล (association) เทคนิคการจัดกลุ่ม (clustering) เทคนิคการจำแนกประเภทของข้อมูล (classification) และการสร้างโมเดลเพื่อใช้ในการทำนายหรือพยากรณ์ ดังตัวอย่างในงานวิจัยของ Gulati H [3] ได้ทำนายการพ้นสภาพของนักศึกษาโดยใช้เทคนิคเหมืองข้อมูล Omkar and Parag [4] ได้ทำนายการพ้นสภาพของนักศึกษา โดยใช้เทคนิคเหมืองข้อมูล J48, RandomForest, RepTree and BFTree และ Songsee และคณะ [5] ได้ประยุกต์ใช้เหมืองข้อมูลเพื่อทำนายสถานภาพของนักศึกษา วิทยาลัยเทคนิคภาคใต้ ด้วยวิธี Decision Tree โดยใช้อัลกอริธึม J48

การพ้นสภาพหรือการออกกลางคันของนักศึกษาถือว่าเป็นปัญหาที่สำคัญของสถาบันการศึกษาซึ่งจะส่งผลกระทบต่อประสิทธิภาพในการจัดการศึกษา การบริหารจัดการ รวมไปถึงการบริหารจัดการงบประมาณขององค์กร คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏสกลนคร เป็นหน่วยงานหนึ่งที่ได้รับผลกระทบจากการพ้นสภาพของนักศึกษาเช่นเดียวกัน จากข้อมูลระบบทะเบียนนักศึกษา สำนักส่งเสริมวิชาการและงานทะเบียน ตั้งแต่ปี 2553-2557 พบว่ามีนักศึกษาพ้นสภาพคิดเป็น ร้อยละ 20.68 ถือว่ามาก สำหรับนักศึกษาที่มีผลการเรียนปกติเกรดเฉลี่ยสะสมตั้งแต่ 2.00 ขึ้นไปแล้วตัดสินใจออกกลางคันนั้น นั่นคือนักศึกษาไม่สำเร็จการศึกษาตามหลักสูตรที่กำหนดไว้ถือว่าเป็นความสูญเสียโอกาสในการผลิตบัณฑิต และการสูญเสียเศรษฐกิจครอบครัว มหาวิทยาลัยและประเทศชาติ หรือที่เรียกว่า “เกิดความสูญเปล่าในการลงทุนเพื่อการศึกษา” กล่าวคือ คณะและมหาวิทยาลัยย่อมเสียเวลาในการบริหารจัดการ เสียทรัพยากรในการลงทุนและเสียโอกาสในการสร้างคน ส่วนนักศึกษาเสียเวลา เสียค่าใช้จ่าย และประการสำคัญ คือ เสียขวัญและกำลังใจในการถอยหลังเพื่อไปเริ่มต้นใหม่ รวมทั้งภาครัฐก็จำเป็นต้องจัดสรรเงินงบประมาณเพื่อสนับสนุนการอุดมศึกษาเป็นจำนวนมากเช่นกัน ดังนั้น การออกกลางคันจึงเป็นประเด็นปัญหาที่คณะวิทยาการจัดการประสบ โดยมีปัจจัยหลายอย่างที่ส่งผลต่อการออกกลางคัน ควรแก่การปรับแก้โดยเร่งด่วน ดังนั้นผู้วิจัยจึงได้นำเอาเทคนิคการทำเหมืองข้อมูล มาใช้ในการวิจัยเพื่อวิเคราะห์ข้อมูลนักศึกษาในอดีตที่จัดเก็บไว้ในฐานข้อมูลถึงปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษา ทั้งนี้เพื่อให้ผู้บริหารหรืออาจารย์ที่ปรึกษาสามารถวางแผนการเรียน ดูแลการลงทะเบียนเรียนของนักศึกษา ให้ความช่วยเหลือ และส่งเสริมนักศึกษาได้อย่างเหมาะสม รวมถึงสามารถนำกฎการจำแนกข้อมูลที่ได้มาใช้ในการพัฒนาระบบการพยากรณ์การพ้นสภาพของนักศึกษาได้

2. อุปกรณ์และวิธีดำเนินการวิจัย

งานวิจัยนี้ได้ดำเนินการตามแนวคิด คริสป์-ดีเอ็ม (CRIPS-DM : Cross-Industry Standard Process for Data Mining) [6] ซึ่งเป็นกระบวนการมาตรฐานอุตสาหกรรมสำหรับการทำเหมืองข้อมูลที่ได้รับความนิยมในปัจจุบัน แบ่งวิธีการดำเนินงาน 6 ขั้นตอนดังนี้

1) การทำความเข้าใจสภาพของปัญหา

เป็นการวิเคราะห์และศึกษาสภาพปัญหาเกี่ยวกับการออกกลางคันระหว่างกำลังศึกษาอยู่ของนักศึกษาระดับปริญญาตรี คณะวิทยาการจัดการ โดยสอบถามข้อมูลจากสำนักส่งเสริมวิชาการ มหาวิทยาลัยราชภัฏสกลนคร จากการศึกษาได้ตัวแปรที่สำคัญ จำนวน 14 แอททริบิวต์ ซึ่งได้ทำการคัดเลือกข้อมูล โดยวิเคราะห์ความสัมพันธ์ระหว่าง แอททริบิวต์ กับ คลาส ซึ่งคลาสที่ใช้แบ่งเป็น 2 คลาส คือ Yes กับ No และคัดเลือกเฉพาะ แอททริบิวต์ ที่สอดคล้องกับการวิเคราะห์ปัจจัยการพ้นสภาพของนักศึกษา ดังแสดงในตารางที่ 1

ตารางที่ 1 รายละเอียดแอททริบิวต์ที่ใช้ในงานวิจัย

ลำดับที่	แอททริบิวต์	คำอธิบาย
1	Sex	เพศ
2	Province	จังหวัด
3	Occup_farther	อาชีพบิดา
4	Revenue_far	รายได้บิดาต่อเดือน
5	occup_mother	อาชีพมารดา
6	Revenue_mom	รายได้มารดาต่อเดือน
7	Parent_status	สถานภาพครอบครัว
8	GPA_school	เกรดเฉลี่ยโรงเรียนเดิม
9	Old_Edu	วุฒิการศึกษาเดิม
10	curriculum	หลักสูตร
11	Major	สาขาวิชาที่เรียน
12	GPA	เกรดเฉลี่ยสะสม
13	Loan	ทุนกู้ยืมเพื่อการศึกษา
14	DropOut	สถานะพ้นสภาพ*

***หมายเหตุ** คลาสที่ใช้ในการทดสอบ คือ Yes, No

2) การทำความเข้าใจข้อมูลและแหล่งที่มา

แหล่งที่มาของข้อมูลที่ใช้ในงานวิจัยได้มาจาก ฝ่ายทะเบียนและวัดผล สำนักส่งเสริมวิชาการและงานทะเบียน มหาวิทยาลัยราชภัฏสกลนคร โดยใช้ฐานข้อมูลที่บันทึกในช่วงปีการศึกษา 2553-2557 คณะวิทยาการจัดการ ซึ่งเป็นนักศึกษาชั้นปีที่ 1- 4 จำนวน 4,163 ชุดข้อมูล เลือกเฉพาะข้อมูลที่มีเกรดเฉลี่ยสะสมตั้งแต่ 2.00 ขึ้นไป ได้ข้อมูล 3,385 ชุดข้อมูล เพื่อวิเคราะห์หาปัจจัยที่ส่งผลให้นักศึกษาที่มีผลการเรียนปกติตัดสินใจออกกลางคันระหว่างที่กำลังศึกษาอยู่

3) การจัดเตรียมข้อมูล

เตรียมข้อมูลก่อนการวิเคราะห์ซึ่งขั้นตอนนี้เป็นขั้นตอนที่สำคัญมาก เพื่อให้ข้อมูลเหมาะสมกับเทคนิคที่นำมาวิเคราะห์ และให้อยู่ในรูปแบบที่สามารถนำไปใช้กับโปรแกรม WEKA ได้ [7] ประกอบด้วยขั้นตอนย่อยดังนี้

3.1) ทำความสะอาดข้อมูล (data cleaning) หลังจากเก็บรวบรวมข้อมูลแล้วตรวจสอบพบว่าข้อมูลยังไม่สมบูรณ์ เช่น ข้อมูลสูญหาย (missing value) มีสิ่งรบกวน (noisy data) หรือมีค่าที่นักศึกษากรอกข้อมูลผิดพลาด (error) ซึ่งค่าเหล่านี้จะเป็นค่าที่นำมาใช้ประกอบการวิเคราะห์ข้อมูลแต่ถ้าค่าข้อมูลไม่สมบูรณ์ ในงานวิจัยนี้จึงออกแบบการแก้ปัญหาด้วยการตัดชุดข้อมูลนั้นออกไปจากการวิเคราะห์ จึงได้ชุดข้อมูลที่สมบูรณ์ทั้งหมด 4,163 ชุดข้อมูล จากทั้งหมด 4,366 ชุดข้อมูล และเลือกเฉพาะข้อมูลที่มีเกรดเฉลี่ยสะสมตั้งแต่ 2.00 ขึ้นไป ได้ข้อมูล 3,385 ชุดข้อมูล

3.2) ปรับเปลี่ยนรูปแบบข้อมูล (data transformation) เนื่องจากข้อมูลมีทั้งเป็นตัวเลขและข้อมูลที่เป็นตัวอักษรไม่อยู่ในรูปแบบที่วิเคราะห์ได้ จึงต้องทำการแทนค่าข้อมูลให้อยู่ในรูปแบบที่สามารถวิเคราะห์ได้ เช่น เพศหญิง แปลงเป็น F เพศชาย แปลงเป็น M และรายได้ของบิดา <12,500 บาทต่อเดือน แปลงเป็น Rev_far1 รายได้ของบิดา = 12,500-25,000 บาทต่อเดือน แปลงเป็น Rev_far2 รายได้ของบิดา >25,000 บาทต่อเดือน แปลงเป็น Rev_far3 เป็นต้น

4) การสร้างโมเดล (modeling) แล้วเลือกเทคนิคที่เหมาะสม

ในการสร้างโมเดล ผู้วิจัยได้นำชุดข้อมูลจำนวน 3,385 ชุดข้อมูล มาสร้างโมเดลด้วยเทคนิคเหมืองข้อมูล แบบ Classification เลือกการทำนายข้อมูลด้วยวิธี Decision Tree ใช้อัลกอริธึมชนิด J48 กำหนดรูปแบบการทดสอบผลลัพธ์ด้วยวิธี 10-fold cross validation ในโปรแกรม WEKA ได้ผลลัพธ์ดังนี้

Classifier output

```
Number of Leaves  :     334

Size of the tree :      398


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3211               94.8597 %
Incorrectly Classified Instances       174                5.1403 %
Kappa statistic                          0.0545
Mean absolute error                      0.0735
Root mean squared error                  0.2244
Relative absolute error                 92.4086 %
Root relative squared error            112.7213 %
Total Number of Instances             3385
```

ภาพที่ 1 แสดงจำนวนโหนดของโมเดลต้นไม้ตัดสินใจที่มีขนาดใหญ่และซับซ้อน

จากภาพที่ 1 จะเห็นได้ว่าโมเดลต้นไม้ตัดสินใจที่ได้มีขนาดใหญ่และจำนวนโหนดมากทำให้โมเดลมีความซับซ้อน วิเคราะห์และแปลผลได้ยาก ดังนั้นเพื่อให้ได้โมเดลต้นไม้ตัดสินใจที่มีขนาดเล็กลง และจำนวนโหนดน้อยลง ซึ่งจะทำให้แปลผลได้ง่ายแต่ในขณะเดียวกันค่าความถูกต้องเท่าเดิมหรือใกล้เคียงกัน ผู้วิจัยจึงได้ย้อนกลับไปดำเนินการขั้นตอนที่ 2 และ 3 อีกครั้ง ดังนี้

4.1) ทำความเข้าใจข้อมูลด้วยการเลือกเฉพาะข้อมูลที่มีเกรดเฉลี่ยสะสม เท่ากับ 2.00 ขึ้นไป ได้ข้อมูล 3,385 ชุดข้อมูล แบ่งข้อมูลเฉพาะที่มีเกรดเฉลี่ยสะสม (GPA) ตั้งแต่ 2.00 ขึ้นไป ออกเป็น 4 กลุ่ม ดังตารางที่ 2 เพื่อแบ่งโมเดลต้นไม้ตัดสินใจออกเป็น 4 โมเดล

4.2) ปรับการเตรียมข้อมูล ด้วยวิธีการคัดเลือกแอททริบิวต์ (Attribute Selection) ของแต่ละกลุ่มย่อย เพื่อเลือกเฉพาะแอททริบิวต์ที่สำคัญ โดยใช้โปรแกรม WEKA ด้วยอัลกอริธึม CfsSubsetEval, InfoGainAttributeEval และผู้วิจัยพิจารณาคัดเลือก ซึ่งจะพิจารณาตัดแอททริบิวต์ที่สอดคล้องน้อยที่สุดกับการพ้นสภาพของนักศึกษา คือ Sex, Old_Edu และ GPA โดยทั้ง 4 กลุ่มย่อยผู้วิจัยได้ตัดแอททริบิวต์ออกเหมือนกัน คงเหลือ 11 แอททริบิวต์ ดังตารางที่ 3

ตารางที่ 2 จำนวนชุดข้อมูลแยกตามกลุ่มของเกรดเฉลี่ยสะสม

กลุ่มเกรดเฉลี่ยสะสม	รายละเอียด	จำนวนชุดข้อมูล
Best	GPA มีค่าตั้งแต่ 3.50 ขึ้นไป	1,185
Excellent	GPA มีค่าเท่ากับ 3.00-3.50	360
Good	GPA มีค่าเท่ากับ 2.50-2.99	1,305
Medium	GPA มีค่าเท่ากับ 2.00-2.49	535
	รวม	3,385

ตารางที่ 3 การคัดเลือกแอททริบิวต์

อัลกอริธึม	กลุ่ม	แอททริบิวต์
CfsSubsetEval	Best	Province, Revenue_father, Old_Edu, Loan, DropOut
	Excellent	Province, Occup_father, Occup_mother, Parent_status, Major, Loan, DropOut
	Good	Parent_status, Loan, DropOut
	Medium	Parent_status, Loan, DropOut
InfoGainAttributeEval	Best	Revenue_father, Old_Edu, Occup_father, Loan, DropOut
	Excellent	Major, Occup_father, Occup_mother, Revenue_father, Parent_status, Province, DropOut
	Good	Loan, Occup_father, Occup_mother, Revenue_father, Revenue_mother, Parent_status, Province, Old_Edu, Major, GPA_School, curicurum, DropOut
	Medium	Loan, Occup_father, Occup_mother, Revenue_father, Revenue_mother, Major, Old_Edu, GPA_School, DropOut
ผู้วิจัยพิจารณาคัดเลือก	ทุกกลุ่ม	Province, Occup_father, Occup_mother, Major,Revenue_father, Loan,Revenue_mother, ,GPA_School, curicurum, Parent_status, DropOut

4.3) ปรับค่าพารามิเตอร์ในโปรแกรม WEKA ดังตารางที่ 4

ตารางที่ 4 รายละเอียดการปรับค่าพารามิเตอร์ ของอัลกอริธึม J48 ในโปรแกรม WEKA

พารามิเตอร์ที่ปรับใน WEKA	คำอธิบาย	ค่าที่ปรับ
ConfidenceFactor (CF)	ค่าระดับความเชื่อมั่นที่ใช้ประกอบการพิจารณาในการตัดกิ่งต้นไม้	0.25
minNumObj (mNO)	จำนวนตัวอย่างขั้นต่ำของโหนดใบแต่ละโหนด	2-10
Numfold (Nf)	จำนวนโฟลด์ของข้อมูลที่ใช้ในการตัดแต่งกิ่งของต้นไม้ โดยที่หนึ่งโฟลด์จะถูกใช้ในการตัดกิ่ง โฟลด์ที่เหลือจะใช้ในการสร้างต้นไม้	3
unpruned	กำหนดพารามิเตอร์ให้เป็น True เพื่อการสร้างต้นไม้โดยไม่มีการตัดกิ่งใด	T

5) การวัดประสิทธิภาพและความแม่นยำของโมเดล (evaluation)

5.1) ความเที่ยงตรงของโมเดล (K- fold Cross Validation) [8]

แบ่งชุดข้อมูลออกเป็น k ส่วน เท่าๆ กัน

ใช้ข้อมูล k-1 ส่วน เพื่อทำการสร้างตัวแบบ (Train Set)

ใช้ข้อมูลที่เหลือ 1 ส่วน เพื่อทำการทดสอบ (Test Set)

ทำซ้ำจนข้อมูลทุกส่วนถูกนำมาทดสอบ

5.2) ค่าความถูกต้อง (accuracy) [9]

เป็นการทดสอบหาค่าที่ทำนายค่าข้อมูลว่ามีความถูกต้องมากน้อยเพียงใด โดยคิดเป็นค่าร้อยละสูตรการคำนวณดังนี้

$$\text{accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

โดย TP คือ ค่าที่ทำนายถูกต้องเชิงบวก TN คือ ค่าที่ทำนายถูกต้องเชิงลบ FP คือ ค่าที่ทำนายผิดพลาดเชิงบวก และ FN คือ ค่าที่ทำนายผิดพลาดเชิงลบ

6) การนำไปใช้งาน

เมื่อได้โมเดลและตรวจสอบความถูกต้องแล้ววางแผนเพื่อนำไปพัฒนาระบบการพยากรณ์การพ้นสภาพของนักศึกษาต่อไป และผู้บริหารสามารถเป็นแนวทางในการบริหารจัดการการเรียนการสอน รวมถึงการบริการจัดการงบประมาณได้ หรืออาจารย์ที่ปรึกษาสามารถวางแผนการเรียน ดูแลการลงทะเบียนเรียนของนักศึกษา ให้ความช่วยเหลือ และส่งเสริมนักศึกษาได้อย่างเหมาะสม

3. ผลการวิจัย

ผลการสร้างโมเดลที่ผ่านการคัดเลือกแอททริบิวต์ ปรับค่าพารามิเตอร์ และเปรียบเทียบค่าความถูกต้อง ในแต่ละกลุ่มข้อมูลย่อย ทั้ง 4 กลุ่ม ผลที่ได้ดังตารางที่ 5–8 ได้ค่าความถูกต้องดังตารางที่ 9 และได้ผลลัพธ์และกฎการจำแนก ดังภาพที่ 2-5

จากตารางที่ 5 ค่าที่ดีที่สุด จะมีค่าความถูกต้องสูงที่สุดคือ 97.4684 โดยการคัดเลือกแอททริบิวต์ ด้วยอัลกอริธึม CfsSubsetEval ปรับค่าพารามิเตอร์ CF = 0.25 , mNO = 10 , Nf = 3 , unpruned = T

จากภาพที่ 2 ได้กฎการตัดสินใจทั้งหมด 4 กฎ และกฎการตัดสินใจที่สำคัญ คือ IF Old_Edu = Old_Edu4 THEN Student Drop Out.

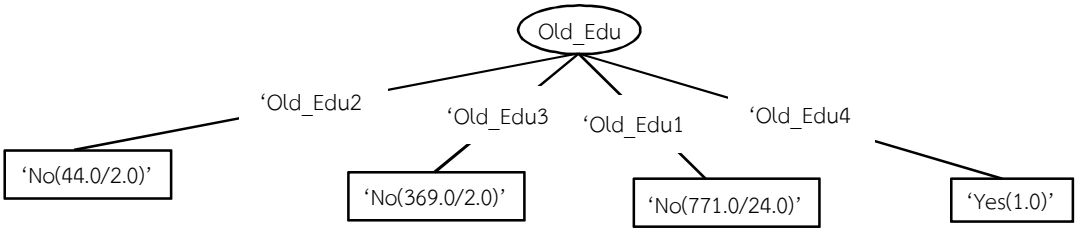
ตารางที่ 5 การปรับค่าพารามิเตอร์ การคัดเลือกข้อมูลและการเปรียบเทียบค่า Accuracy ในกลุ่ม Best

No	ปรับค่าพารามิเตอร์ใน WEKA				Number of leave / size of tree			Accuracy		
	CF	mNO	Nf	unpruned	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก
1	0.25	2	3	T	4 / 5	76/91	100/118	97.384	96.8776	96.7932
2	0.25	3	3	T	4 / 5	41/49	43/51	97.384	96.962	97.0464
3	0.25	4	3	T	4 / 5	41/49	31/37	97.384	96.8776	92.2152
4	0.25	5	3	T	4 / 5	41/49	24/29	97.384	97.0464	97.2996
5	0.25	6	3	T	4 / 5	41/49	24/29	97.384	97.1308	97.2996
6	0.25	7	3	T	4 / 5	41/49	24/29	97.384	97.1308	97.2996
7	0.25	8	3	T	4 / 5	24/41	13/16	97.4684	97.2996	97.2996
8	0.25	9	3	T	4 / 5	23/28	13/16	97.4684	97.2996	97.2996
9	0.25	10	3	T	4 / 5	23/28	13/16	97.4684	97.2996	97.3840

CfsSubsetEval คือ การคัดเลือกแอททริบิวต์ โดยใช้โปรแกรม WEKA ด้วยอัลกอริธึม CfsSubsetEval

InfoGain คือ การคัดเลือกแอททริบิวต์ โดยใช้โปรแกรม WEKA ด้วยอัลกอริธึม InfoGainAttributeEval

ผู้วิจัยเลือก คือ ผู้วิจัยพิจารณาคัดเลือกแอททริบิวต์ ได้ 11 แอททริบิวต์

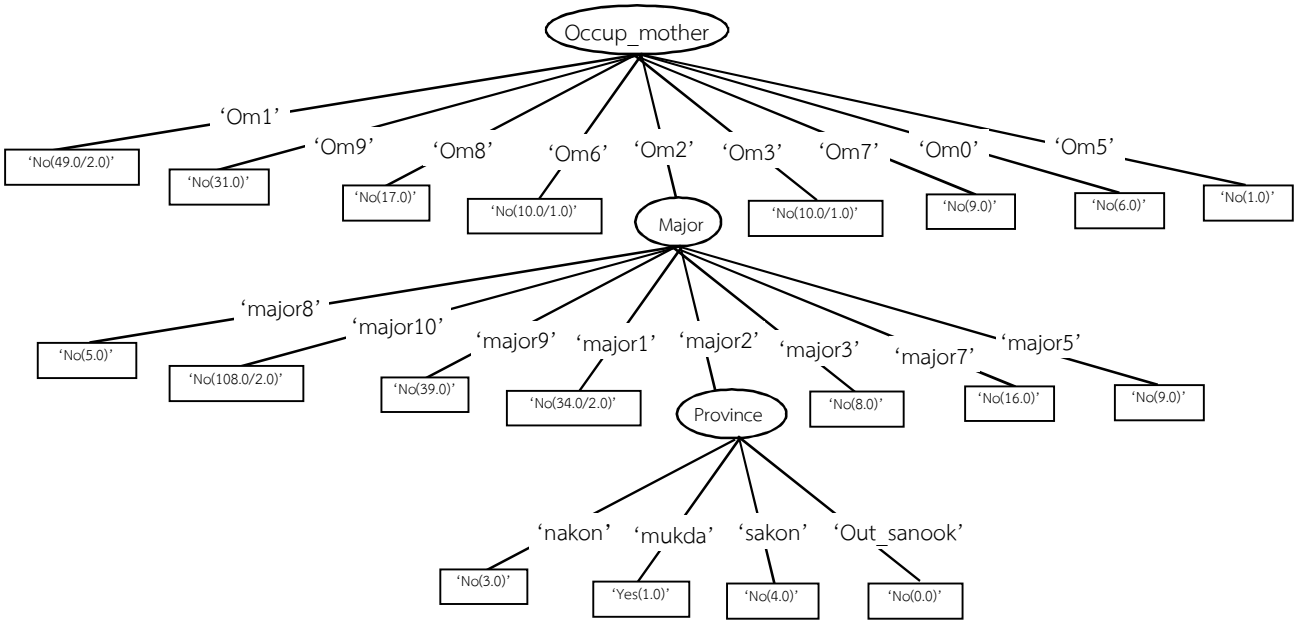


ภาพที่ 2 โมเดลต้นไม้ตัดสินใจในกลุ่ม Best

ตารางที่ 6 การปรับค่าพารามิเตอร์ การคัดเลือกข้อมูลและการเปรียบเทียบค่า Accuracy ในกลุ่ม Excellent

ปรับค่าพารามิเตอร์ใน WEKA					Number of leave / size of tree			Accuracy		
No	CF	mNO	Nf	unpruned	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก
1	0.25	2	3	T	35/ 40	35/40	42/49	96.9444	96.9444	96.3889
2	0.25	3	3	T	19 / 22	19/22	26/31	97.5	97.5	96.3889
3	0.25	4	3	T	1 / 1	1 / 1	1 / 1	97.5	97.5	97.5
4	0.25	5	3	T	1 / 1	1 / 1	1 / 1	97.5	97.5	97.5
5	0.25	6	3	T	1 / 1	1 / 1	1 / 1	97.5	97.5	97.5
6	0.25	7	3	T	1 / 1	1 / 1	1 / 1	97.5	97.5	97.5
7	0.25	8	3	T	1 / 1	1 / 1	1 / 1	97.5	97.5	97.5
8	0.25	9	3	T	1 / 1	1 / 1	1 / 1	97.5	97.5	97.5
9	0.25	10	3	T	1 / 1	1 / 1	1 / 1	97.5	97.5	97.5

จากตารางที่ 6 ค่าที่ดีที่สุด จะมีค่าความถูกต้องสูงที่สุดคือ 97.50 โดยการคัดเลือกแอททริบิวต์ ด้วยอัลกอริธึม CfsSubsetEval และ InfoGainAttributeEval ซึ่งมีค่าความถูกต้องเท่ากัน ปรับค่าพารามิเตอร์ CF = 0.25, mNO = 3, Nf = 3, unpruned = T



หมายเหตุ : Om หมายถึง Occ_mom

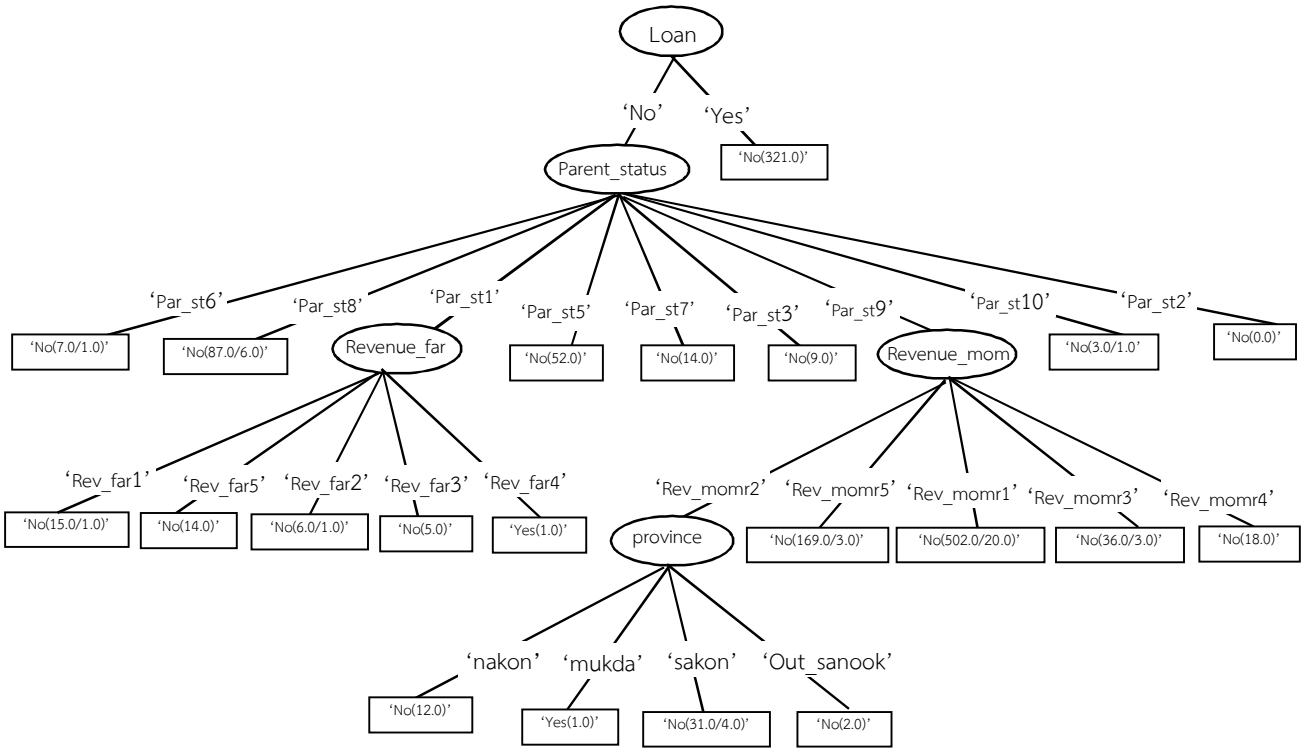
ภาพที่ 3 โมเดลต้นไม้ตัดสินใจในกลุ่ม Excellent

จากภาพที่ 3 ได้กฎการตัดสินใจทั้งหมด 19 กฎ กฎการตัดสินใจที่สำคัญ คือ IF Occup_mother = Occ_mom2 AND Major = Major2 AND Province = mukda THEN Student Drop Out.

ตารางที่ 7 การปรับค่าพารามิเตอร์ การคัดเลือกข้อมูลและการเปรียบเทียบค่า Accuracy ในกลุ่ม Good

No	ปรับค่าพารามิเตอร์ใน WEKA				Number of leave / size of tree			Accuracy		
	CF	mNO	Nf	unpruned	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก
1	0.25	2	3	T	1 / 1	116/137	80/95	96.7816	95.7854	95.8621
2	0.25	3	3	T	1 / 1	75/89	33/40	96.7816	95.9387	95.9387
3	0.25	4	3	T	1 / 1	75/89	33/40	96.7816	96.0153	96.0153
4	0.25	5	3	T	1 / 1	59/71	33/40	96.7816	96.1686	96.2452
5	0.25	6	3	T	1 / 1	48/58	33/40	96.7816	96.3218	96.3985
6	0.25	7	3	T	1 / 1	33/40	33/40	96.7816	96.3985	96.4751
7	0.25	8	3	T	1 / 1	29/35	29/35	96.7816	96.3985	96.5517
8	0.25	9	3	T	1 / 1	29/35	29/35	96.7816	96.3985	96.6284
9	0.25	10	3	T	1 / 1	21/26	21/26	96.7816	96.3985	96.6284

จากตารางที่ 7 ค่าที่ดีที่สุด จะมีค่าความถูกต้องสูงที่สุดคือ 96.6284 โดยการคัดเลือกแอททริบิวต์ ด้วยวิธีการที่ผู้วิจัยพิจารณาคัดเลือก ปรับค่าพารามิเตอร์ CF = 0.25, mNO = 10, Nf = 3, unpruned = T



ภาพที่ 4 โมเดลต้นไม้ตัดสินใจในกลุ่ม Good

จากภาพที่ 4 ได้กฎการตัดสินใจทั้งหมด 21 กฎ กฎการตัดสินใจที่สำคัญ 2 กฎ คือ

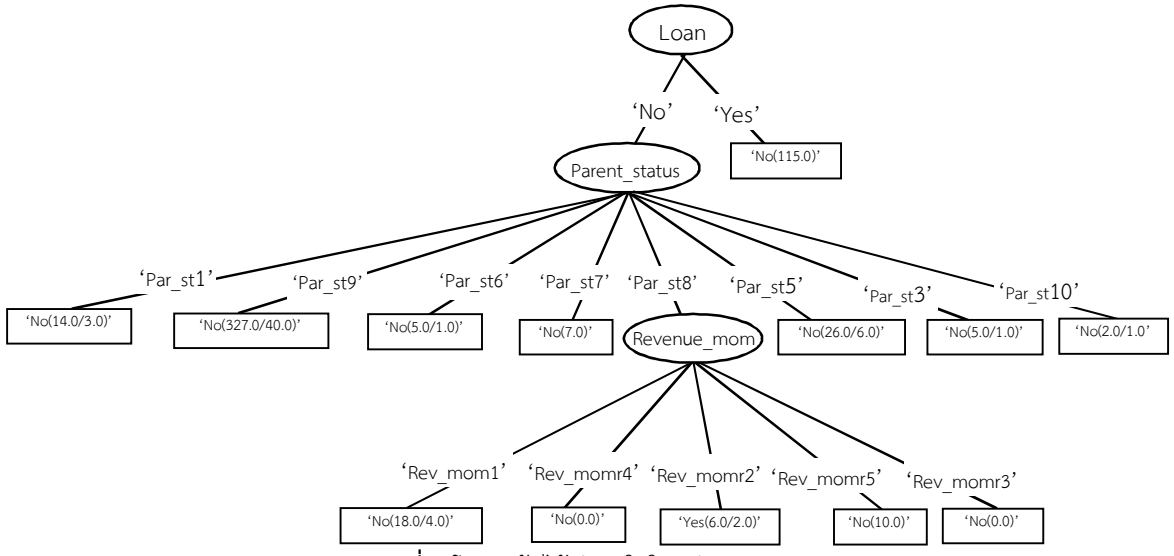
IF Loan = No AND Parent_Status = Par_st1 AND Revenue_far = Rev_far4 THEN Student Drop Out.

IF Loan = No AND Parent_Status = Par_st9 AND Revenue_mom = Rev_mom2 AND Province = mukda THEN Student Drop Out.

ตารางที่ 8 การปรับค่าพารามิเตอร์ การคัดเลือกข้อมูลและการเปรียบเทียบค่า Accuracy ในกลุ่ม Medium

No	ปรับค่าพารามิเตอร์ใน WEKA				Number of leave / size of tree			Accuracy		
	CF	mNO	Nf	unpruned	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก	CfsSubsetEval	InfoGain	ผู้วิจัยเลือก
1	0.25	2	3	T	1 / 1	126/149	137/162	88.5981	84.8598	85.0467
2	0.25	3	3	T	1 / 1	75/88	108/127	88.5981	85.2336	85.9813
3	0.25	4	3	T	1 / 1	67/78	84/99	88.5981	85.7944	87.1028
4	0.25	5	3	T	1 / 1	51/60	65/77	88.5981	85.4206	86.9159
5	0.25	6	3	T	1 / 1	43/51	54/65	88.5981	85.9813	87.4766
6	0.25	7	3	T	1 / 1	39/46	42/50	88.5981	86.3551	87.6636
7	0.25	8	3	T	1 / 1	32/38	35/42	88.5981	86.5421	87.2897
8	0.25	9	3	T	1 / 1	32/38	13/16	88.5981	87.1028	87.2897
9	0.25	10	3	T	1 / 1	13/16	13/16	88.5981	87.4766	87.4766

จากตารางที่ 8 ค่าที่ดีที่สุด จะมีค่าความถูกต้องสูงที่สุดคือ 87.4766 โดยการคัดเลือกแอททริบิวต์ด้วยอัลกอริธึม InfoGainAttributeEval และ ผู้วิจัยพิจารณาคัดเลือก ซึ่งมีค่าความถูกต้องเท่ากัน ปรับค่าพารามิเตอร์ CF = 0.25, mNO = 10, Nf = 3 , unpruned = T



ภาพที่ 5 โมเดลต้นไม้ตัดสินใจในกลุ่ม Medium

จากภาพที่ 5 ได้กฎการตัดสินใจทั้งหมด 13 กฎ กฎการตัดสินใจที่สำคัญ คือ IF Loan = No AND Parent_Status = Par_st8 AND Revenue_mom = Rev_mom2 THEN Student Drop Out.

ตารางที่ 9 ค่าความถูกต้องของแต่ละกลุ่มเกรดเฉลี่ย

	Classifier J48			
กลุ่มเกรดเฉลี่ย	Best	Excellent	Good	Medium
Accuracy	97.5527	97.5000	96.6284	87.4766
TP Rate	1	1	0.998	0.981
FP Rate	1	1	1	0.967
TN Rate	0	0	0	0.033
FN Rate	0	0	0.002	0.019

4. สรุปและอภิปรายผลการวิจัย

การสร้างโมเดลต้นไม้ตัดสินใจเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาที่มีผลการเรียนปกตินั้นพบว่า โมเดลที่ได้มีขนาดต้นไม้ที่ใหญ่ และขนาดของโหนดที่มากเกินไป ทำให้โมเดลมีความซับซ้อนการทำความเข้าใจและแปลผลทำได้ยาก ซึ่งสอดคล้องกับแนวคิดของ นิตยา เกิดประสพ [10] ที่กล่าวว่า โมเดลหรือความสัมพันธ์ที่สร้างได้มาแล้วนั้น จะต้องถูกนำมาทดสอบอัตราความผิดพลาดและวิเคราะห์ความซับซ้อนของรูปแบบโมเดล ถ้าอัตราความผิดพลาดยังสูงเกินไป อาจจะต้องย้อนกลับไปที่ขั้นตอนในการค้นหาโมเดลอีกครั้ง หรือบางครั้งอาจจะต้องย้อนกลับไปที่ขั้นตอนคัดเลือกข้อมูลเพื่อปรับปรุงโมเดลให้ถูกต้องยิ่งขึ้น ในทำนองเดียวกัน ถ้าโมเดลที่หามาได้มีรูปแบบที่ซับซ้อนเกินไปจนยากต่อการทำความเข้าใจ อาจจะต้องย้อนกระบวนการไปขั้นตอนการสร้างและค้นหาโมเดล เพื่อให้หาโมเดลใหม่ที่มีความถูกต้องเท่าเดิมแต่มีรูปแบบที่ซับซ้อนน้อยลง ดังนั้นผู้วิจัยจึงได้ย้อนกลับไปดำเนินการตามขั้นตอนที่ 2 และขั้นตอนที่ 3 ใหม่ ทำให้พบว่าวิธีการคัดเลือกแอททริบิวต์และการปรับค่าพารามิเตอร์ที่เหมาะสม สามารถลดความซับซ้อนของโมเดลและเพิ่มประสิทธิภาพการจำแนกข้อมูลได้ ซึ่งได้ผลการวิเคราะห์ปัจจัยการพ้นสภาพของนักศึกษาที่มีผลการเรียนปกติ แบ่งตามกลุ่มเกรดเฉลี่ยสะสมได้ 4 กลุ่ม ดังนี้

กลุ่มที่ 1 (best) เกรดเฉลี่ยสะสม ตั้งแต่ 3.50 ขึ้นไป ปัจจัยที่ได้คือ วุฒิการศึกษาเดิมที่จบปริญญาตรี ค่าความถูกต้องเท่ากับ 97.5527 %

กลุ่มที่ 2 (excellent) เกรดเฉลี่ยสะสม 3.00-3.50 ปัจจัยที่ได้คือ มารดามีอาชีพทำนา/เกษตรกร/ประมง เรียนสาขาวิชาการตลาด และอาศัยอยู่ในจังหวัดมุกดาหาร ค่าความถูกต้องเท่ากับ 97.50 %

กลุ่มที่ 3 (good) เกรดเฉลี่ยสะสม 2.50-2.99 ปัจจัยที่ได้คือ

1) ไม่ได้รับทุนกู้ยืมเพื่อการศึกษา สถานภาพครอบครัวบิดา-มารดาแยกกันอยู่ และบิดาไม่มีรายได้

2) ไม่ได้รับทุนกู้ยืมเพื่อการศึกษา สถานภาพครอบครัวบิดา-มารดาหย่าร้าง มารดามีรายได้ 12,500-25,000 บาท และอาศัยอยู่ในจังหวัดมุกดาหาร ค่าความถูกต้องเท่ากับ 96.6284 %

กลุ่มที่ 4 (medium) เกรดเฉลี่ยสะสม 2.00-2.49 ปัจจัยที่ได้คือ ไม่ได้รับทุนกู้ยืมเพื่อการศึกษา สถานภาพครอบครัวมารดาแต่งงานใหม่ และมารดามีรายได้ 12,500-25,000 บาทต่อเดือน ค่าความถูกต้องเท่ากับ 87.4766 %

กล่าวโดยสรุป ปัจจัยที่ส่งผลให้นักศึกษาตัดสินใจออกกลางคันระหว่างกำลังศึกษาอยู่ แตกต่างกันออกไป แต่ปัจจัยที่เหมือนกันของกลุ่มที่ 3 และ กลุ่มที่ 4 คือ การไม่ได้รับทุนกู้ยืมเพื่อการศึกษา และสถานภาพครอบครัวไม่สมบูรณ์ ซึ่งสอดคล้องกับงานศึกษาของ Phannarat และคณะ [11] ที่พบว่า หากสถานภาพครอบครัวไม่สมบูรณ์ ย่อมส่งผลต่อการลาออกของนักศึกษา สำหรับการวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาระดับปริญญาตรี ที่มีผลการเรียนของเกรดเฉลี่ยสะสมตั้งแต่ 2.00 ขึ้นไป ในครั้งนี้ พบว่าค่าความถูกต้องของโมเดลทั้ง 4 กลุ่ม มีค่าความถูกต้องอยู่ระหว่าง 87.4766 - 97.5527 ถือว่าเป็นค่าที่สูงที่สามารถนำกฎการจำแนกข้อมูลที่ได้จากโมเดลไปใช้ในการพัฒนาระบบพยากรณ์การพ้นสภาพของนักศึกษาต่อไปได้ ฉะนั้นการลาออกกลางคันของนักศึกษาที่มีผลการเรียนปกติ เป็นประเด็นที่ผู้บริหาร คณาจารย์ และอาจารย์ที่ปรึกษา ต้องเข้าใจสภาพครอบครัวของนักศึกษา และช่วยเหลือทางการเงิน

ให้กับนักศึกษาเพื่อบรรเทาปัญหาการลาออกกลางคัน รวมทั้งช่วยวางแผนการเรียน โดยอาจารย์ที่ปรึกษาให้ความช่วยเหลือในการวางแผนการลงทะเบียนเรียนในแต่ละปีการศึกษา หรือวางแผนการจัดตารางเวลาเรียนเพื่อให้นักศึกษามีเวลาว่าง ซึ่งจะทำให้มีโอกาสทำงานนอกเวลาเรียนจะได้มีรายได้เพิ่มขึ้น สำหรับผู้บริหาร อาจจะมีโครงการช่วยเหลือนักศึกษาที่มีปัญหาด้านการเงิน ด้วยการจ้างนักศึกษาทำงานชั่วคราวรายชั่วโมง ในหน่วยงานต่างๆ ของมหาวิทยาลัย เช่น สถาบันวิจัยและพัฒนา อาจจะมีโครงการสนับสนุนให้นักศึกษาเป็นผู้ช่วยวิจัย ของโครงการวิจัยต่างๆ ของมหาวิทยาลัย เป็นต้น หรือผู้บริหารวางแผนด้วยการอาจจะพิจารณาจัดสรรงบประมาณ เพื่ออุดหนุนทุนการศึกษาเพิ่มเติม เช่น สนับสนุนทุนเรียนดีแต่ยากจน หรือ พิจารณาผ่อนผันชำระค่าลงทะเบียน ให้กับนักศึกษาที่อยู่ในกลุ่มเสี่ยงที่จะออกกลางคันระหว่างศึกษาอยู่ได้

5. ข้อเสนอแนะ

5.1 เนื่องจากข้อมูลที่นำมาวิเคราะห์ คือข้อมูลที่ได้บันทึกไว้ในระบบฐานข้อมูล จากฝ่ายทะเบียนและวัดผล สำนักส่งเสริมวิชาการและงานทะเบียน ซึ่งอาจจะมีปัจจัยอื่นที่เกี่ยวข้องกับการตัดสินใจออกกลางคันระหว่างกำลังศึกษาอยู่ เช่น ปัจจัยเกี่ยวกับหลักสูตรและการเรียนการสอน ปัจจัยด้านอาจารย์ผู้สอน หรือปัจจัยเกี่ยวกับสถานศึกษา เป็นต้น ดังนั้นอาจจะสอบถามหรือสัมภาษณ์เพิ่มเติมจากนักศึกษาอีกทางหนึ่ง

5.2 ในการนำเทคนิคเหมืองข้อมูลมาใช้ ควรใช้เทคนิคเหมืองข้อมูลหลายๆ เทคนิค เพื่อเปรียบเทียบค่าความถูกต้องที่ดีที่สุด ซึ่งจะส่งผลให้ผลการดำเนินงานมีประสิทธิภาพมากยิ่งขึ้น

5.3 การเตรียมข้อมูลก่อนการวิเคราะห์นั้นถือว่าเป็นขั้นตอนที่สำคัญมาก ถ้าทำ Data Cleaning โดยใช้วิธีการตัด Missing Data ซึ่งบางทีข้อมูลที่ตัดออกไปอาจมีผลต่อปัจจัยก็ได้ การทำ Data Cleaning จึงต้องพิจารณาให้ดีและรอบคอบ

6. กิตติกรรมประกาศ

ขอขอบคุณสำนักส่งเสริมวิชาการและงานทะเบียน มหาวิทยาลัยราชภัฏสกลนคร ที่อนุเคราะห์ข้อมูลเพื่อนำมาใช้ในการวิจัยครั้งนี้ และขอขอบคุณ รศ.ดร.กิตติศักดิ์ เกิดประสพ และ รศ.ดร.นิตยา เกิดประสพ ที่ได้ให้คำแนะนำที่เป็นประโยชน์ และให้คำปรึกษาในการทำวิจัยเกี่ยวกับเทคนิคการทำเหมืองข้อมูล ด้วยความเมตตาตลอดมา ผู้วิจัยรู้สึกซาบซึ้งและขอขอบคุณมา ณ โอกาสนี้

7. References

[1] J. Han, M. Kamber, J. Pei, Data mining concepts and techniques, 3rd ed., ElSevier, USA, 2011.

[2] S. Sinsomboon, Data Mining, JamjureeProduct, Bangkok, Thailand, 2015.

[3] H. Gulati, Predictive analytics using data mining technique, Computing for Sustainable Global Development (INDIACom), IEEE Conference Publications, 11-13 March, pp.713-716. New Delhi, India, 2015.

[4] S. Omkar, M. Parag, Predicting Dropout Students Using Data-Mining Techniques, IJR. 2(1) (2015) 365-375.

[5] M. Songsee, C. Palaman, W. Wuttisak, Application to Predict Student Status Using Data Mining for Southern College of Technology, SCT. 3(1) (2010) 73-89

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0 Step-by-step data mining guide, Technical report, SPSS inc. 2000.

[7] R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, D. Scuse, WEKA Manual for Version 3-7-8, University of Waikato, Hamilton, New Zealand, 2013.

[8] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer, 2008.

[9] J. Mohammed, J.R. Meira .Wagner, Data Mining and Analysis Fundamental Concepts and Algorithms, Cambridge University Press, New York, USA, 2014.

[10] N. Kerdprasop, Knowledge Discovery and Data Mining, School of computer Engineering, Suranaree University of Technology, Nakorn Ratchasima, Thailand, n.d.

[11] N. Phannarat, K. Wangrangsimakul, M. Pipatpen, The Problems of Withdrawal of the Youths Receiving Scholarship from World Vision Foundation of Thailand in Songkhla Province, PNUJ. 1(3) (2009) 130-144.