การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่

นายปรัชญารัก เวียงสงค์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
แขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช
พ.ศ. 2564

Predictive Analytics for New Student Application Using Data Mining Techniques In Faculty of Education of Chiang Mai Rajabhat University

Mr.Prachyarak Weangsong

A Thesis Submitted in Partial Fulfillment of the Requirements for

the Degree of Master of Science in Information and Communication Technology

School of Science and Technology

Sukhothai Thammathirat Open University

2021

หัวข้อวิทยานิพนธ์ การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่

ชื่อและนามสกุล นายปรัชญารัก เวียงสงค์

แขนงวิชา เทคโนโลยีสารสนเทศและการสื่อสาร

สาขาวิชา วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช

อาจารย์ที่ปรึกษา 1. รองศาสตราจารย์ ณัฐพร เห็นเจริญเลิศ

2. รองศาสตราจารย์ ดร.วรัญญา ปุณณวัฒน์

วิทยานิพนธ์นี้ ได้รับความเห็นชอบให้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรระดับปริญญาโท เมื่อวันที่ 27 กันยายน 2565

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(อาจารย์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

..... กรรมการ
(รองศาสตราจารย์ ณัฐพร เห็นเจริญเลิศ)

..... กรรมการ
(รองศาสตราจารย์ ดร.วรัญญา ปุณณวัฒน์)

..... ประธานกรรมการบัณฑิตศึกษา
(รองศาสตราจารย์ ดร.นราธิป ศรีราม)

ชื่อวิทยานิพนธ์ การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล
คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่
ผู้วิจัย นายปรัชญารัก เวียงสงค์ **รหัสนักศึกษา** 2649600505 **ปริญญา** วิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร) อาจารย์ที่ปรึกษา (1) รองศาสตราจารย์ ณัฐพร เห็นเจริญเลิศ (2) รองศาสตราจารย์ ดร.วรัญญา ปุณณวัฒน์ **ปีการศึกษา** 2564

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อวิเคราะห์และคัดเลือกคุณลักษณะสำคัญที่สัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ 2) เพื่อสร้างแบบจำลองและประเมินประสิทธิภาพในการวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่

ข้อมูลการศึกษาประกอบด้วยข้อมูลพื้นฐานของผู้สมัครได้แก่ ปีที่สมัคร เพศ เกรดเฉลี่ย อาชีพบิดามารดา สาขาที่สมัครและข้อมูลอื่นที่เกี่ยวข้อง จำนวนทั้งหมด 17 คุณลักษณะมาหาค่าความสำคัญและคัดเลือกคุณลักษณะสำคัญด้วยวิธี Information Gain จากนั้นนำคุณลักษณะมาสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์และเทคนิคป่าสุ่ม และลดจำนวนคุณลักษณะที่มีค่าความสำคัญน้อยลงให้คงเหลือจำนวนคุณลักษณะที่สำคัญและสัมพันธ์กับการสมัครมากที่สุด นำวัดประสิทธิภาพแบบจำลองทั้งหมดด้วยวิธี 5-fold Cross Validation วิธี 10-fold Cross Validation วิธี Split Validation (70:30) และวิธี Split Validation (80:20) เพื่อประเมินประสิทธิภาพแบบจำลองด้วยเทคนิคต่าง ๆ โดยใช้โปรแกรม RapidMiner

ผลการวิจัยพบว่า 1) คุณลักษณะที่มีค่ามากที่สุด 5 คุณลักษณะคือ แผนการเรียนที่จบจากระดับมัธยม มีค่าน้ำหนักสูงสุดที่ 0.522 รองลงมากคือ เกรดเฉลี่ยมีค่าน้ำหนักที่ 0.290 เพศ มีค่าน้ำหนักที่ 0.207 อาชีพผู้ปกครองมีค่าน้ำหนักที่ 0.056 อาชีพมารดามีค่าน้ำหนักที่ 0.055 2) แบบจำลองด้วยเทคนิคป่าสุ่มมีความเหมาะสมมากที่สุด มีค่าความถูกต้องคิดเป็นร้อยละ 75.84 มีค่าความแม่นยำในการทำนายคิดเป็นร้อยละ 76.01 มีค่าความครบถ้วนคิดเป็นร้อยละ 75.71 และมีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมคิดเป็นร้อยละ 75.85

คำสำคัญ เหมืองข้อมูล การคัดเลือกคุณลักษณะ เทคนิคป่าสุ่ม เทคนิคต้นไม้ตัดสินใจ

Thesis title: Predictive Analytics for New Student Application Using Data Mining Techniques In Faculty of Education of Chiang Mai University
Researcher: Mr. Prachyarak Weangsong; **ID:** 2649600505; **Degree:** Master of Science (Information and Communication Technology) **Thesis advisors:** (1) Nuttaporn Hencharoenlert, Associate Professor; (2) Dr. Waranya Poonnawat, Associate Professor;
Academic year: 2021

Abstract

The study's goals were 1) to analyze and select the key features related to the application of new students at Chiang Mai Rajabhat University's Faculty of Education; 2) to create a model and assess the predictive behavior of new students using data mining at Chiang Mai Rajabhat University's Faculty of Education.

The study's information included the application's basic information such as the years of application, gender, average grade, parent's occupation, application documents, and other related information. 17 Features were selected for the key features and best feature values by the Information Gain method to create a model of Decision Tree, Naive Bayes, and Random Forest then reduced the number of less important features, leaving only the most important and relevant to the application. After that, the model's efficiency was assessed using 5-fold Cross Validation, 10-fold Cross Validation, Split Validation (70:30), and Split Validation (80:20). The model performance was validated using a variety of techniques by the RapidMiner program.

According to the study, 1) the five most valuable features were a high school graduation plan with a maximum weight of 0.522, the second most common grade with a weight of 0.290, gender with a weight of 0.207, parent's occupation with a weight of 0.056, and mother's occupation with a weight of 0.055. 2) The best fit model was the Random Forest technique which had an accuracy of 75.84%. The prediction accuracy (precision) was 76.01%. There was a recall value of 75.71% and an overall efficiency (F-measure) was 75.85%.

Keywords: Data Mining Feature Selection Random Forest Technique Decision Tree Technique

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปด้วยดี ผู้วิจัยจึงใคร่ขอขอบพระคุณรองศาสตราจารย์ ณัฐพร เห็นเจริญเลิศ อาจารย์ที่ปรึกษาซึ่งให้ความกรุณาสละเวลาให้คำแนะนำตลอดจนแนวทางในการแก้ไขข้อบกพร่องและช่วยเหลือจนวิทยานิพนธ์เล่มนี้สำเร็จลุล่วง รวมถึงรองศาสตราจารย์ ดร. วรัญญา ปุณณวัฒน์ อาจารย์ที่ปรึกษาที่ให้คำแนะนำแก้ไขข้อบกพร่องของวิทยานิพนธ์และช่วยแนะแนวทางในการหาข้อมูลเพื่อศึกษาเพิ่มเติมจนนำมาสู่การเปรียบเทียบแก้ไขจุดบกพร่องของการวิเคราะห์ข้อมูลให้สำเร็จ ผู้วิจัยขอขอบพระคุณคณะผู้บริหาร เพื่อนร่วมงานในคณะครุศาสตร์ ตลอดจนคณะร่วมผลิตและสำนักงานต่างๆ ของมหาวิทยาลัยราชภัฏเชียงใหม่ ที่ได้อนุเคราะห์ข้อมูลและช่วยเหลือเป็นกำลังใจในการศึกษาข้อมูลที่นำมาจัดทำในวิทยานิพนธ์

สุดท้ายนี้ผู้วิจัยขอขอบพระคุณครอบครัวอันได้แก่คุณพ่อ คุณแม่ น้องชาย ญาติและครอบครัวของผู้วิจัยเองที่ได้ช่วยเหลือและสนับสนุนตลอดจนให้กำลังใจในการศึกษาและการทำงานวิจัยในครั้งนี้จนสำเร็จ

นายปรัชญารัก เวียงสงค์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

การศึกษาเป็นสิทธิขั้นพื้นฐานที่สำคัญที่สุดของคนในทุกประเทศ ที่รัฐต้องจัดให้เพื่อพัฒนาคนในประเทศของตนในทุกช่วงวัย ให้มีความเจริญงอกงามในทุกด้าน เพื่อเป็นต้นทุนทางปัญญาที่สำคัญในการพัฒนาทักษะ คุณลักษณะ และสมรรถนะในการประกอบสัมมาชีพ และการดำรงชีวิตร่วมกับผู้อื่นในสังคมได้อย่างเป็นสุข อันจะนำไปสู่เสถียรภาพ และความมั่นคงของสังคมและประเทศชาติที่ต้องพัฒนาให้เจริญก้าวหน้า ทัดเทียมนานาประเทศ ในเวทีโลกท่ามกลางกระแสการเปลี่ยนแปลงอย่างรวดเร็วของโลกศตวรรษที่ ๒๑ โดยประเทศไทยได้ให้ความสำคัญด้านการศึกษาในฐานะกลไกหลัก ในการพัฒนาสังคมให้คนมีคุณภาพ คุณธรรม กล่าวคือการศึกษาช่วยสร้างจิตสำนึกในการเป็นมนุษย์ มีจิตวิญญาณของผู้มี อารยะธรรมทางปัญญาและความงดงามทางจิตใจ การศึกษาสร้างคนให้มีความรู้ในการดำรงชีวิต การประกอบอาชีพ มีความอดทนในการต่อสู้กับอุปสรรคของชีวิตมาโดยตลอด โดยมุ่งจัดการศึกษา ให้คนไทยทุกคนสามารถเข้าถึงโอกาสและความเสมอภาคในการศึกษาที่มีคุณภาพ การพัฒนาระบบ การบริหารจัดการศึกษาที่มีประสิทธิภาพ พัฒนากำลังคนให้มีสมรรถนะในการทำงานที่สอดคล้อง กับความต้องการของตลาดงานและการพัฒนาประเทศให้เกิดความยั่งยืนในการพัฒนาประเทศชาติ นั้น

มหาวิทยาลัยราชภัฏเชียงใหม่ เป็นสถาบันอุดมศึกษาในสังกัดกระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม ได้ตระหนักถึงความสำคัญของการบริหารงานโดยการใช้แผนยุทธศาสตร์เพื่อพัฒนาท้องถิ่น โดยมีการเรียนการสอนการรับนักศึกษาเข้าศึกษาต่อในสาขาวิชาต่างๆ กว่า 80 สาขา ใน 10 หลักสูตร ได้แก่ หลักสูตรครุศาสตรบัณฑิต หลักสูตรวิทยาศาสตรบัณฑิต หลักสูตรศิลปศาสตรบัณฑิต หลักสูตรบริหารธุรกิจบัณฑิต หลักสูตรบัญชีบัณฑิต หลักสูตรเศรษฐศาสตรบัณฑิต หลักสูตรนิเทศศาสตรบัณฑิต หลักสูตรนิติศาสตรบัณฑิต หลักสูตรรัฐประศาสนศาสตรบัณฑิต และหลักสูตรสาธารณสุขศาสตร(กองนโยบายและแผน มหาวิทยาลัยราชภัฏเชียงใหม่, 2563) การแนะแนวสัญจรหรือการแนะนำผ่านทางกิจกรรมต่างๆ และการประชาสัมพันธ์ผ่านสื่อต่าง ๆ เช่น สื่อวิทยุโทรทัศน์ สื่อสิ่งพิมพ์ เว็บไซต์ และอีกหลายช่องทางที่ได้จัดขึ้นมา เพื่อช่วยในการตัดสินใจเข้าสมัครเพื่อศึกษาต่อในระดับปริญญาตรีของนักศึกษามหาวิทยาลัยราชภัฏ ซึ่งโอกาสได้รับข้อมูลที่เป็นประโยชน์มากขึ้นและมีช่องทางเลือกในการตัดสินใจเลือกหลักสูตร สาขา คณะ สถาบันการศึกษาได้มากยิ่งขึ้น

คณะครุศาสตร์ถือเป็นหนึ่งในคณะที่มีความสำคัญ ด้านการผลิตบัณฑิตที่มีคุณภาพตลอดจนการพัฒนาวิชาชีพครูที่มีคุณภาพเพื่อออกสู่สังคม โดยนักศึกษาสามารถเลือกหลักสูตรที่มีความหลากหลายสอดคล้องกับความสามารถและความสนใจของตนเอง เพื่อการวางแผนนำไปประกอบอาชีพในอนาคตได้ การเลือกวิชาสาขาเรียนมีความสำคัญอย่างมากเพราะมีผลต่อการ

ตัดสินใจประกอบสัมมาอาชีพของตน และอาจส่งผลกระทบต่อนักศึกษาหากพบภายหลังการเข้าเรียนว่าสาขาที่ตัดสินใจเลือกเข้าศึกษานั้นไม่เหมาะสมและไม่ตรงกับความต้องการของตนเอง เนื่องมาจากการตัดสินใจที่ขาดประสบการณ์ ไม่ทราบถึงความต้องการและทักษะที่แน่นอนของตนเอง ไม่มีรายละเอียดของหลักสูตรที่สนใจเพียงพอหรือวิชาที่ต้องเรียนในหลักสูตรนั้นๆ ซึ่งนักศึกษาส่วนใหญ่มักจะใช้ความรู้สึกชอบสภาพแวดล้อม เลือกเรียนตามเพื่อน สังคมหรือความชอบของผู้ปกครอง เป็นต้น ก่อให้เกิดปัญหาแก่นักศึกษาในภายหลังเข้ามาศึกษาต่อในระดับปริญญาตรี ทำให้นักศึกษาขอย้ายสาขาวิชา ขอพักการศึกษาต่อ จนถึงการลาออกจากการเป็นนักศึกษา ดังนั้นเพื่อให้เป็นข้อมูลเบื้องต้นสำหรับผู้เรียนในการตัดสินใจเข้าศึกษาต่อในคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่

จากการศึกษาข้อมูลข้างต้นผู้วิจัยได้เล็งเห็นการใช้ประโยชน์จากการนำข้อมูลที่เก็บเป็นข้อมูลการสมัครเรียนของนักศึกษาในแต่ละสาขาวิชาของคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ หากนำมาวิเคราะห์ตามกระบวนผลลัพธ์ที่ได้จะช่วยในการตัดสินใจของนักศึกษาในการเลือกหลักสูตรในการเข้ารับการศึกษาต่อในระดับปริญญาตรีเบื้องต้น และยังช่วยในด้านการวางแผนการเปิดรับนักศึกษาใหม่ประจำปีการศึกษาต่อไปให้เข้าถึงกลุ่มผู้สมัครที่มีประสิทธิภาพ ประสิทธิผล เข้าถึงกลุ่มนักศึกษาที่สนใจและตรงกับความสามารถ ความถนัดในวิชาชีพที่แท้จริง เพื่อเพิ่มโอกาสในการคัดเลือกและคัดกรองนักศึกษาที่มีความพร้อมความสามารถตรงกับความต้องการที่จะศึกษาต่อให้จบและประกอบอาชีพตามหลักสูตรที่ได้สำเร็จการศึกษา นอกจากนี้ข้อมูลที่ได้จากการศึกษาครั้งนี้จะได้ทราบถึงปัจจัยที่มีอิทธิพลต่อการตัดสินใจเพื่อจะนำไปเป็นแนวทางในการพัฒนาปรับปรุงการประชาสัมพันธ์ข้อมูลข่าวสารของทางมหาวิทยาลัยราชภัฏเชียงใหม่ และการวางแผนด้านงบประมาณในการพัฒนาหลักสูตรการเรียนการสอนให้สอดคล้องกับความต้องการในปัจจุบัน เพื่อให้นโยบายการรับนักศึกษาเข้าศึกษาในมหาวิทยาลัยราชภัฏเชียงใหม่ บรรลุเป้าหมายต่อไป

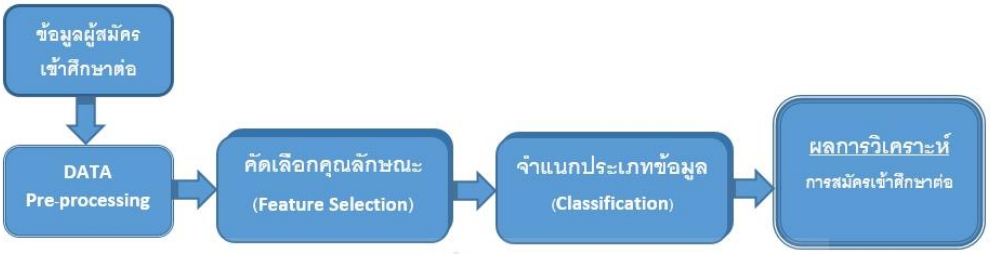
2. วัตถุประสงค์การวิจัย

2.1 เพื่อวิเคราะห์และคัดเลือกคุณลักษณะสำคัญที่สัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่

2.2 เพื่อสร้างแบบจำลองและประเมินประสิทธิภาพในการวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่

3. กรอบแนวคิดการวิจัย

ผู้วิจัยได้ดำเนินการวิจัยนี้ตามกรอบแนวความคิดตามภาพที่ 1.1



ภาพที่ 1.1 กรอบแนวความคิดการวิจัย

4. ขอบเขตของการวิจัย

4.1 ขอบเขตข้อมูล

การวิจัยครั้งนี้ทำการศึกษาในหลักสูตร 4 ปี ระดับปริญญาตรี คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ ข้อมูลในการวิจัยนี้ คือ ข้อมูลผู้เข้าสมัครทุกรอบ ผู้สมัครที่ผ่านการคัดเลือกและผู้สมัครที่ยืนยันการลงทะเบียนเรียนแล้ว เฉพาะการเรียนภาคปกติ จำนวน 3,195 รายการ ในปีการศึกษา 2562- 2564 (สำนักทะเบียนและประมวลผล, 2564)

4.2 ขอบเขตด้านระยะเวลา

การวิจัยนี้ใช้ระยะเวลา 1 ปี (เดือนกันยายน 2564 – เดือนสิงหาคม 2565)

4.3 ขอบเขตด้านเนื้อหา

งานวิจัยนี้ได้นำแนวคิดการทำเหมืองข้อมูล (Data Mining) มาประยุกต์ใช้ในการวิเคราะห์ข้อมูลผู้เข้าสมัครเรียนเพื่อศึกษาต่อในคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ โดยการจัดเก็บข้อมูลของผู้เข้าสมัครเรียนที่ผ่านการคัดเลือกและยืนยันการลงทะเบียนเรียน มาวิเคราะห์โดยใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) ตามกระบวนการทำเหมืองข้อมูลของ Cross – Industry Standard Process for Data Mining หรือ CRISP-DM (Jyoti, Nidhi and Sanjeev, 2013) โดยการคัดเลือกคุณลักษณะที่มีความเหมาะสมจากข้อมูล สร้างแบบจำลองและเปรียบเทียบแบบจำลองที่มีความเหมาะสม มีประสิทธิภาพ ด้วยโปรแกรม Rapid Miner Studio

5. นิยามศัพท์เฉพาะ

5.1 การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการที่ค้นหาหรือสกัดข้อมูลจากชุดข้อมูลขนาดใหญ่ เพื่อให้ได้รูปแบบความสัมพันธ์ของข้อมูลหรือพฤติกรรมที่ซ่อนอยู่ในชุดข้อมูล โดยอาศัยวิธีการจากปัญญาประดิษฐ์ (Artificial Intelligence) โดยแบ่งออกเป็นการหากฎความสัมพันธ์ (Association Rule) การจำแนกข้อมูล (Data Classification) การแบ่งกลุ่มข้อมูล (Data Clustering) และจินตทัศน์ (Visualization) ซึ่งข้อมูลที่ได้จากการทำเหมืองข้อมูลสามารถนำไปใช้ประโยชน์ในด้านการตัดสินใจและนำไปใช้ในด้านการพัฒนาด้านต่าง ๆ ได้

5.2 การคัดเลือกคุณลักษณะ (Feature Selection) เป็นกระบวนการจัดกลุ่มของข้อมูลคุณลักษณะหรือตัวแปรของข้อมูล ซึ่งการคัดเลือกคุณลักษณะช่วยลดตัวแปรหรือมิติของข้อมูล อาจทำให้เหลือคุณลักษณะที่มีความสำคัญและดีที่สุดเพียงหนึ่งตัวหรือกลุ่มของตัวแปรเพียงหนึ่งกลุ่มที่มีความสำคัญต่อการพยากรณ์ ซึ่งหากได้คุณลักษณะที่ดีจะช่วยให้ตัวแบบมีประสิทธิและทำงานได้เร็วขึ้น

5.3 การจำแนกประเภทข้อมูล (Classification) กระบวนการสร้างแบบจำลองสำหรับแยกแอตทริบิวต์ (Attribute) หรือฟีเจอร์ (Feature) จำนวนมากในข้อมูล เพื่อจัดข้อมูลให้อยู่ในกลุ่มที่กำหนด ตัวอย่างเช่น การแยกกลุ่มผลการเรียนออกเป็นกลุ่มดีมาก ดี ปานกลาง และต่ำ ซึ่งพิจารณาจากผลการเรียน หรือการแยกประเภทกลุ่มการประกอบอาชีพออกเป็นกลุ่มที่ประกอบอาชีพค้าขาย อาชีพการเกษตร เป็นต้น

5.4 เทคนิคการจำแนกประเภทข้อมูล (Classification Techniques) เทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่าง ๆ ที่ได้มีการกำหนดไว้ นำมาสร้างแบบจำลองเพื่อการพยากรณ์ค่าข้อมูล (Predictive Model) ในอนาคตเรียกว่า Supervised Learning ตัวอย่างเช่น เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าสุ่ม (Random Forest) ที่จะใช้ในการดำเนินการวิจัยนี้ เป็นต้น

5.5 การพยากรณ์ (Prediction) การทำนายเหตุการณ์ในอนาคต โดยนำข้อมูลการสมัครเรียนของนักศึกษาใหม่ คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ มาพยากรณ์กลุ่มผู้สมัครเรียนในแต่ละหลักสูตรและจำนวนผู้ที่จะสมัครด้วยเทคนิคการทำเหมืองข้อมูล

5.6 คุณลักษณะ หมายถึง องค์ประกอบหรือกลุ่มของข้อมูลผู้เข้าสมัครที่จะนำมาคัดเลือก ตัวอย่างเช่น ปีที่สมัคร เพศ เกรดเฉลี่ย หลักสูตรที่สมัคร ปีที่สมัคร แผนการเรียนที่จบจากระดับมัธยม สัญชาติบิดามารดา อาชีพบิดามารดา

5.7 ความถูกต้อง หมายถึงค่าความถูกต้องของตัวแบบจำลองจากการวัดประสิทธิภาพด้วยวิธี Cross Validation หรือ Split Validation ที่มีความถูกต้องแม่นยำในการทำนายผลข้อมูล

5.8 การวัดประสิทธิภาพ หมายถึงการวัดประเมินประสิทธิภาพของแบบจำลองด้วยวิธี Cross Validation หรือ Split Validation และเปรียบเทียบค่าเปรียบเทียบจากค่าความถูกต้องโดยรวมของแบบจำลอง (Accuracy) ความแม่นยำของการทำนาย (Precision) ค่าความครบถ้วน (Recall) และค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure)

5.9 คณะครุศาสตร์ หมายถึง คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ จัดการศึกษาในการผลิตบัณฑิตสาขาการศึกษาออกเป็น 6 หลักสูตร ได้แก่หลักสูตรครุศาสตรบัณฑิต (ระดับปริญญาตรี) หลักสูตรศิลปศาสตรบัณฑิต (ระดับปริญญาตรี) หลักสูตรประกาศนียบัตรบัณฑิตวิชาชีพครู หลักสูตรครุศาสตรมหาบัณฑิต (ระดับปริญญาโท) หลักสูตรวิทยาศาสตรมหาบัณฑิต (ระดับปริญญาโท) หลักสูตรปรัชญาดุษฎีบัณฑิต (ระดับปริญญาเอก) ซึ่งในหลักสูตรครุศาสตรบัณฑิตเปิดแยกตามสาขาวิชาออกเป็นจำนวน 22 สาขาวิชา โดยการเปิดรับสมัครจัดออกเป็นรอบด้วยระบบการคัดเลือกกลางบุคคลเข้าศึกษาในระดับอุดมศึกษา (TCAS) ภายหลังจากผ่านการคัดเลือกแล้วจะผู้สมัครที่ผ่านการคัดเลือกจะต้องลงทะเบียนเพื่อยืนยันสิทธิ์ในการเข้าศึกษาต่อ ซึ่งการจัดการด้านการรับสมัครดำเนินการโดยสำนักทะเบียนและประมวลผล ของมหาวิทยาลัยราชภัฏเชียงใหม่

6. ประโยชน์ที่คาดว่าจะได้รับ

6.1 ได้แบบจำลองในการวิเคราะห์จากข้อมูลผู้สมัครและทำนายจำนวนผู้ที่จะเลือกสมัครเรียนในแต่ละหลักสูตรโดยใช้กระบวนการเทคนิคการทำเหมืองข้อมูล

6.2 แบบจำลองที่ได้สามารถนำไปประยุกต์ใช้เพื่อสามารถนำมาวิเคราะห์ผู้สนใจสมัครเรียนในสาขาวิชาเรียนในระดับบัณฑิตศึกษา

6.3 เป็นเครื่องมือในการวิเคราะห์และทำนายกลุ่มเป้าหมาย เพื่อวางแผนการจัดสรรงบประมาณ สวัสดิการสำหรับนักศึกษาในปีต่อไป

6.4 เป็นเครื่องมือในการวางแผนการประชาสัมพันธ์ การลงพื้นที่สำหรับรับสมัครและคัดเลือกนักศึกษาที่จะเข้าสมัครเรียน

6.5 เป็นเครื่องมือในการวางแผนการจัดการบริหารพัฒนาหลักสูตรการเรียนการสอนให้สอดคล้องแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ

บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

การศึกษาวิจัย เรื่อง การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ ผู้วิจัยได้ทำการค้นคว้าแนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อใช้เป็นแนวทางในการศึกษาดังนี้

1. การทำเหมืองข้อมูล
2. การคัดเลือกคุณสมบัติ
3. การจำแนกประเภทข้อมูล
4. เกณฑ์ที่ใช้ในการเลือกคุณลักษณะ
5. การประเมินผลและการวัดประสิทธิภาพแบบจำลอง
6. โปรแกรม Rapid Miner Studio
7. งานวิจัยที่เกี่ยวข้อง

1. การทำเหมืองข้อมูล

1.1 แนวคิดและความหมายของการทำเหมืองข้อมูล

เทคโนโลยีสารสนเทศในยุคปัจจุบันนับได้ว่ามีความเกี่ยวข้องกับการดำรงชีวิต การติดต่อสื่อสาร การเผยแพร่และรับข้อมูลข่าวสารที่มีการเปลี่ยนแปลงให้อยู่ในรูปแบบของเว็บไซต์หรือโปรแกรมที่รองรับในเครื่องคอมพิวเตอร์และสมาร์ตโฟน การวางแผนเพื่อการพัฒนาระบบการเผยแพร่ข้อมูลให้ไปถึงผู้ใช้งานจึงเป็นแนวทางสำคัญและถือเป็นเครื่องมือที่สำคัญที่จะใช้วางแผนหรือกลยุทธ์ในการดำเนินธุรกิจ เมื่อมีการเก็บข้อมูลการใช้งานหรือพฤติกรรมการใช้งานของผู้เข้าถึงเว็บไซต์ที่นำเสนอข้อมูล หากนำมาวิเคราะห์ข้อมูลโดยกระบวนการเหมืองข้อมูลอาจทำให้ได้ข้อมูลที่ซ่อนอยู่ภายในข้อมูลเหล่านั้น

สายชล สินสมบูรณ์ทอง (2558) ได้ให้ความหมายการทำเหมืองข้อมูลว่า เป็นกระบวนการทำงานที่สกัดข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่เพื่อให้ได้สารสนเทศที่มีประโยชน์ที่เรายังไม่ทราบโดยเป็นสารสนเทศที่มีเหตุผลและสามารถนำไปใช้ได้ซึ่งเป็นสิ่งสำคัญที่จะช่วยการตัดสินใจในการดำเนินงานต่าง ๆ โดยการทำเหมืองข้อมูลเป็นกระบวนการที่สำคัญในรการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in Databases Process: KDD) ซึ่งการทำเหมืองข้อมูลจะสามารถนำมาคาดการณ์การพัฒนาการวิวัฒนาของอนาคตได้ซึ่งการทำเหมืองข้อมูลนับเป็น 1 ใน 10 เทคโนโลยีที่เกิดขึ้นใหม่ที่จะทำให้เกิดการเปลี่ยนแปลงเนื่องจากองค์กรต่าง ๆ ได้มีการเก็บข้อมูลไว้ในคลังข้อมูลจำนวนเพิ่มมากขึ้นสารสนเทศที่จะนำมาวิเคราะห์เพื่อให้ได้ประสบผลสำเร็จตามกลยุทธ์และเป้าหมายนั้นจะต้องพิจารณาจากข้อมูลที่มีอยู่ว่าสามารถนำมาทำอะไรได้บ้าง

เอกสิทธิ์ พัชรวงศ์ศักดา (2557) ได้ให้ความหมายของเหมืองข้อมูล หมายถึงคำศัพท์ที่ใช้เปรียบเทียบกับการขุดเหมืองแร่ โดยการขุดเหมืองแร่ผลลัพธ์ที่ต้องการคือ แร่ที่มีค่า เช่น

เพชร พลอย เป็นต้น ขั้นตอนในการทำเหมืองจะต้องระเบิดภูเขาใหญ่จำนวนหลายภูเขา เพื่อค้นหาแร่ที่ต้องการ ซึ่งแร่ที่พบจะมีจำนวนน้อยเมื่อเทียบกับการระเบิดภูเขา เช่นเดียวกับองค์กรหรือบริษัทที่มีข้อมูลภายในองค์กรจำนวนมาก จึงต้องพยายามค้นหาสิ่งที่มีค่าในข้อมูลที่ซ่อนอยู่ในข้อมูลเหล่านั้น

คำนาย อภิปรัชญาสกุล (2557) ได้ให้ความหมายของเหมืองข้อมูล หมายถึง โปรแกรมหรือซอฟต์แวร์ที่ใช้ในการสร้างคำถามเพื่อให้ได้คำตอบจากข้อมูลที่ซ่อนอยู่ ตามรูปแบบที่กำหนดความสัมพันธ์ระหว่างข้อมูลและกฎเกณฑ์สำหรับอ้างอิงในฐานข้อมูลขนาดใหญ่ ผลลัพธ์ที่ได้รับคือแนวโน้มและพฤติกรรมต่าง ๆ สำหรับคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต เมื่อนำข้อมูลและผลลัพธ์ที่ได้มาใช้ร่วมกันกับเครื่องมือสนับสนุนการตัดสินใจทำให้สามารถสร้างความได้เปรียบและสามารถตอบคำถามในทางธุรกิจที่จะเกิดขึ้นในอนาคตได้อย่างแม่นยำ

สุรพงศ์ เอื้อวัฒนามงคล (2557) ได้ให้ความหมายการทำเหมืองข้อมูลว่า เป็นขบวนการวิเคราะห์ข้อมูลที่มีขั้นตอนเพื่อให้ได้มาซึ่งตัวแบบ (Pattern) ซึ่งแสดงความสัมพันธ์ระหว่างข้อมูลโดยผลลัพธ์ความรู้เกี่ยวกับข้อมูลที่ถูกต้องสามารถนำไปใช้ในการตัดสินใจและดำเนินงานได้โดยไม่ผิดพลาดหรือสร้างความเสียหายจากการนำไปใช้งาน

วิจิตรสวัสดิ์ สุขสวัสดิ์ ณ อยุธยา (2555) ได้ให้ความหมายของการทำเหมืองข้อมูลว่า การทำเหมืองข้อมูลเป็นกระบวนการหาความสัมพันธ์ของรูปแบบข้อมูลและนำความรู้ที่ได้จากข้อมูลมาใช้ประโยชน์ การทำเหมืองข้อมูลอาศัยวิธีการจากปัญญาประดิษฐ์ (Artificial Intelligence) ในการค้นหารูปแบบหรือพฤติกรรมจากกลุ่มของข้อมูล โดยแบ่งออกเป็นการหากฎความสัมพันธ์ (Association Rule) การจำแนกข้อมูล (Data Classification) การแบ่งกลุ่มข้อมูล (Data Clustering) และจินตทัศน์ (Visualization)

การทำเหมืองข้อมูล (Data Mining) กล่าวได้ว่าเป็นกระบวนการกระทำกับข้อมูลที่มีจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ ที่ซ่อนอยู่ในชุดข้อมูลที่นำมาเข้าสู่กระบวนการทำเหมืองข้อมูลให้ ซึ่งในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำมาประยุกต์ใช้กับงานหลายประเภททั้งในด้านการดำเนินงานธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ด้านวิทยาศาสตร์และการแพทย์ในด้านการวิเคราะห์และทำนายการเกิดของโรคตลอดจนแนวทางในการป้องกันและการรักษาเพื่อไม่ให้เกิด รวมถึงในด้านเศรษฐกิจ ด้านการศึกษาและสังคม

การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บข้อมูลและการตีความหมายของข้อมูลที่มีการเก็บเพิ่มเติมจากเดิมด้วยวิธีการอย่างง่าย มาสู่การปรับปรุงการจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบข้อมูลสำคัญที่ซ่อนอยู่ในข้อมูล

ดังนั้นการทำเหมืองข้อมูลหมายถึง การจัดเก็บข้อมูล การรวบรวมข้อมูลขององค์กรหรือบริษัทที่มีกระบวนการทางด้านเทคโนโลยีสารสนเทศ ด้านฐานข้อมูลซึ่งมีการจัดเก็บจำนวนมากในระบบโดยแยกออกเป็นแอตทริบิวต์ (Attribute) ซึ่งเมื่อนำมาผ่านกระบวนการทำเหมืองข้อมูลจะทำให้สามารถค้นหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ซึ่งอาจเป็นสารสนเทศที่สามารถนำไปใช้พยากรณ์ประกอบการตัดสินใจต่อไปได้

1.2 รูปแบบข้อมูลของการทำเหมืองข้อมูล

ข้อมูลที่จัดเก็บในรูปแบบของข้อมูลที่มีขนาดใหญ่นั้น มีการจัดเก็บในหลายรูปแบบ การนำข้อมูลที่ได้มาเพื่อวิเคราะห์ด้วยวิธีการทำเหมืองข้อมูลนั้น อาจแบ่งออกเป็น 2 รูปแบบ (สุรพงศ์ เอื้อวัฒนามงคล, 2557) ดังนี้

รูปแบบที่หนึ่ง ข้อมูลแบบมีโครงสร้าง (Structured Data) ตัวอย่างเช่น ข้อมูลที่จัดเก็บในรูประเบียน (Record) ตาราง (Table) หรือรูปแบบรายการข้อมูล (Transactional Data) เป็นต้น นอกจากนี้ เอกสิทธิ์ พัชรวงศ์ศักดา (2557) ได้กล่าวว่า ข้อมูลแบบมีโครงสร้างโดยทั่วไปจะอยู่ในรูปแบบตารางซึ่งประกอบด้วยแถวและคอลัมน์ ในการวิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูล ส่วนใหญ่จะเรียกข้อมูลแต่ละแถว ว่า “ตัวอย่าง (Example)” หรือ “อินสแตนซ์ (Instance)” และเรียกข้อมูลแต่ละคอลัมน์ ว่า “แอตทริบิวต์ (Attribute)” หรือ “ฟีเจอร์ (Feature)”

รูปแบบที่สอง ข้อมูลแบบไม่มีโครงสร้างแน่นอน (Unstructured Data) ตัวอย่างเช่น ข้อมูลที่อยู่ในรูปแบบข้อความ (Text) ข้อมูลในเว็บไซต์ซึ่งประกอบด้วยข้อความและข้อมูลเชื่อมโยงที่ชี้ไปยังเว็บไซต์อื่น ๆ ข้อมูลที่อยู่ในรูปแบบกราฟ เป็นต้น นอกจากนี้ เอกสิทธิ์ พัชรวงศ์ศักดา (2557) ได้กล่าวว่า ข้อมูลส่วนใหญ่ที่จะอยู่ในรูปแบบของข้อมูลแบบไม่มีโครงสร้าง ตัวอย่างเช่น ข้อความหรือรูปภาพต่าง ๆ ข้อมูลเหล่านี้มีความสำคัญด้วยเช่นกัน

ข้อมูลที่นำมาทำเหมืองข้อมูลโดยส่วนใหญ่มักจะอยู่ในรูปแบบที่มีโครงสร้าง ตัวอย่างเช่น ระเบียนของข้อมูลหรือตารางข้อมูล เป็นต้น โดยในปัจจุบันการทำเหมืองข้อมูลกับข้อมูลที่ไม่มีโครงสร้างได้นำมาวิเคราะห์ด้วยวิธีการทำเหมืองข้อมูลมากขึ้น ตัวอย่างเช่น การทำเหมืองข้อมูลบนข้อมูลข้อความ (Text Mining) การทำเหมืองข้อมูลกับข้อมูลที่อยู่ในรูปแบบของเว็บไซต์ (Web Mining) เป็นต้น ข้อมูลส่วนใหญ่ที่นิยมนำมาวิเคราะห์ด้วยการทำเหมืองข้อมูลมักจะเป็นแบบมีโครงสร้าง ซึ่งประกอบด้วยแอตทริบิวต์ (Attribute) หรือตัวแปรของข้อมูล ตัวอย่างเช่น ระเบียนข้อมูลของสมาชิกร้านค้าแต่ละรายมีตัวแปรประกอบด้วย หมายเลขบัตรประจำตัวประชาชน อายุ เพศ เบอร์โทรศัพท์ เป็นต้น ซึ่งตัวแปรของข้อมูลอาจมีหลายชนิด (สุรพงศ์ เอื้อวัฒนามงคล, 2557) ดังนี้

ข้อมูลที่บ่งบอกกลุ่ม (Categorical Data) มีลักษณะข้อมูลที่มีค่าที่ไม่ต่อเนื่อง (Discrete) สามารถแทนค่าด้วยตัวอักษร ตัวอย่างเช่น เพศ สี เกรด เป็นต้น ข้อมูลประเภทนี้ยังแบ่งย่อยออกเป็น 2 ชนิด ได้แก่ ชนิดที่ 1 Nominal Data ข้อมูลระบุประเภท ใช้สำหรับบ่งบอกลักษณะข้อมูลเพื่อจัดประเภทหรือกลุ่ม โดยที่ในแต่ละประเภทหรือกลุ่มที่ระบุนั้นเป็นอิสระต่อกัน อาจไม่มีความเกี่ยวข้องกัน และไม่ได้เป็นลำดับต่อเนื่อง ตัวอย่างเช่น เพศ สี เป็นต้น และชนิดที่ 2 Ordinal Data ข้อมูลวัดระดับ ข้อมูลที่ใช้แสดงลำดับของข้อมูล โดยแต่ละลำดับสามารถนำมาเปรียบเทียบว่าเท่ากันหรือไม่เท่ากันก็ได้ ตัวอย่างเช่น ลำดับที่ 1 2 และ 3 ของผลการสอบ หรือเกรดของนักเรียน เป็นต้น

ข้อมูลที่บ่งบอกปริมาณ (Numerical Data) ค่าของข้อมูลจะมีความต่อเนื่อง (Continuous) นอกจากจะสามารถนำมาเปรียบเทียบได้เช่นเดียวกับ Categorical Data แล้วยังสามารถนำมาคำนวณการบวก ลบ คูณ หรือหารได้อีกด้วย ตัวอย่างเช่น น้ำหนัก ความสูง อายุ เป็นต้น สำหรับ Numerical Data ที่สามารถนำมาบวกลบกันได้เท่านั้น เรียกข้อมูลชนิดนี้ว่า Interval

Data ตัวอย่างเช่น วัน เวลา อุณหภูมิ เป็นต้น Numerical Data ที่สามารถนำมาบวก ลบ คูณ และหาร (หาค่าสัดส่วนระหว่างกันได้) เรียกข้อมูลชนิดนี้ว่า Ratio Data ตัวอย่างเช่น จำนวนนับ อายุและความสูง เป็นต้น

1.3 กระบวนการทำเหมืองข้อมูล

การนำข้อมูลขนาดใหญ่ที่มีมาเพื่อวิเคราะห์ข้อมูลที่ซ่อนอยู่จะนำข้อมูลมาเข้าสู่กระบวนการทำเหมืองข้อมูล (สุวิมล สิทธิชาติ, 2560) ประกอบด้วยขั้นตอนต่าง ๆ คล้ายขั้นตอนการพัฒนาซอฟต์แวร์ สามารถนำไปปฏิบัติทำให้การทำเหมืองข้อมูลมีประสิทธิภาพและได้ผลลัพธ์ที่มีความถูกต้อง กระบวนการมาตรฐานนี้เรียกว่า Cross-Industry Standard Process for Data Mining หรือ CRISP-DM (Jyoti, Nidhi and Sanjeev, 2013) มีขั้นตอนในการดำเนินการประกอบด้วย 6 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 การทำความเข้าใจปัญหา (Business Understanding) เป็นขั้นตอนแรกในกระบวนการซึ่งต้องทำความเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ กำหนดขอบเขตของข้อมูลที่จะนำวิเคราะห์เพื่อหาความได้เปรียบด้านต่าง ๆ เพื่อนำมาแก้ไขปัญหาองค์กรและต้องสามารถระบุผลลัพธ์ที่มีได้

ขั้นตอนที่ 2 การทำความเข้าใจและรวบรวมข้อมูลที่เกี่ยวข้อง (Data Understanding) เริ่มจากการเก็บรวบรวมข้อมูลที่เกี่ยวข้องให้อยู่รวมกัน ตรวจสอบข้อมูลที่ได้จากการรวบรวมเพื่อให้ได้ข้อมูลที่มีความถูกต้อง คัดเลือกและพิจารณาข้อมูลที่จะนำมาใช้สำหรับวิเคราะห์

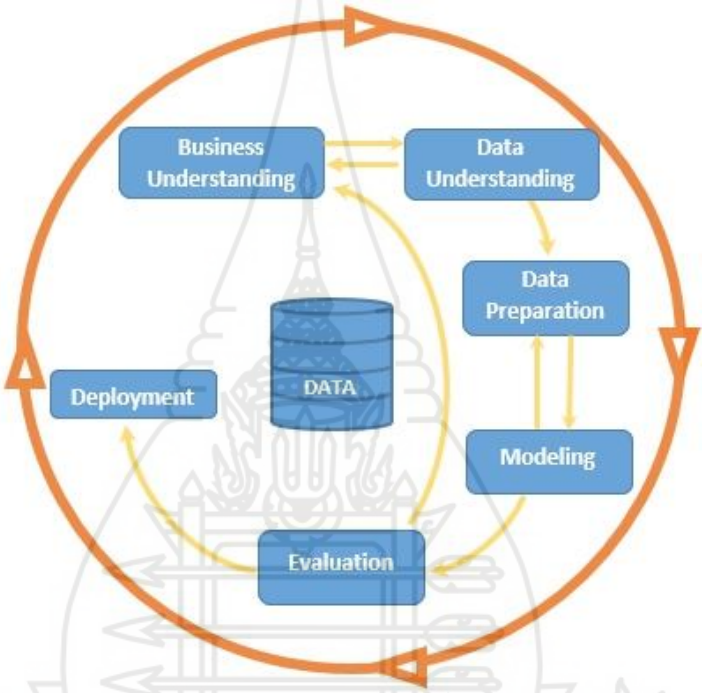
ขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation) ภายหลังจากการคัดเลือกข้อมูลที่ได้จากการรวบรวม ขั้นตอนต่อไปคือการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (Raw Data) ให้อยู่ในรูปแบบข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (Data Cleaning) ตัวอย่างเช่น ข้อมูลที่อยู่ในรูปแบบที่ผิดจากข้อมูลในกลุ่มเดียวกัน ตัวอักษรหรือตัวเลข การเติมข้อมูลที่ขาดหายไป เป็นต้น

ขั้นตอนที่ 4 การสร้างแบบจำลอง (Modeling Phase) เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทาง Data Mining ได้แก่ การจำแนกประเภทข้อมูล (Classification) การแบ่งกลุ่มข้อมูล (Clustering) หรือการหากฎความสัมพันธ์ (Association Rules) ในขั้นตอนนี้จะนำเทคนิคหลายเทคนิคมาใช้เพื่อให้ได้ผลลัพธ์ที่ถูกต้องและดีที่สุด ซึ่งอาจจะต้องย้อนกลับไปทำซ้ำในขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation) เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคที่นำมาใช้ ตัวอย่างเทคนิคที่ใช้ในการวิเคราะห์ข้อมูลต่าง ๆ เช่น การแบ่งกลุ่ม (Clustering) การหากฎความสัมพันธ์ (Association Rules) การจำแนกประเภทข้อมูล (Classification) ซึ่งมีหลายหลายวิธี อันได้แก่ เทคนิค Decision Tree เทคนิค Naïve Bayes เทคนิค Neural Network และเทคนิค Support Vector Machine (SVM) เป็นต้น

ขั้นตอนที่ 5 การประเมินประสิทธิภาพแบบจำลอง (Evaluation Phase) ในขั้นตอนนี้จะได้ผลการวิเคราะห์ข้อมูลด้วยเทคนิคทาง Data Mining แล้ว แต่ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานต่อจะต้องมีการวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรก หรือมีความน่าเชื่อถือมากน้อยเพียงใด ซึ่งอาจจะย้อนกลับไปยังขั้นตอนก่อนหน้าเพื่อเปลี่ยนแปลงแก้ไขเพื่อให้ได้ผลลัพธ์ตามที่ต้องการได้ สำหรับการสร้างโมเดลด้วยเทคนิค Classification มีการทดสอบ

ประสิทธิภาพของโมเดลอยู่ 3 แบบใหญ่ คือ 1) Self-Consistency Test 2) Split Test 3) Cross-validation Test

ขั้นตอนที่ 6 การหาผลลัพธ์และองค์ความรู้ที่ได้มาประยุกต์ใช้ (Deployment Phase) เป็นการนำโมเดลที่เหมาะสมที่สุดไปใช้งานจริง เพื่อวิเคราะห์และแก้ปัญหาที่ต้องการ ในกระบวนการทำงานของ CRISP-DM นั้นไม่ได้หยุดเพียงแค่ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคทาง Data Mining เท่านั้น แม้ผลลัพธ์ที่ได้จะแสดงถึงองค์ความรู้ที่มีประโยชน์แต่จะต้องนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริงในองค์กรหรือบริษัท



ภาพที่ 2.1 ภาพแสดงกระบวนการมาตรฐาน CRISP-DM

จากภาพที่ 2.1 แสดงกระบวนการมาตรฐาน 6 ขั้นตอน เรียกว่า Cross-Industry Standard Process for Data Mining หรือ CRISP-DM (Jyoti, Nidhi and Sanjeev, 2013) ดังอธิบายไว้ในหัวข้อ 1.3

1.4 รูปแบบการทำเหมืองข้อมูล

การทำเหมืองข้อมูลจำแนกเป็น 2 รูปแบบ (สายชล สินสมบูรณ์ทอง, 2558) ดังนี้

รูปแบบที่หนึ่ง แบบจำลองการทำนาย (Predictive Modeling) หรือ การเรียนรู้แบบมีผู้สอน (Supervised learning) ผลลัพธ์ที่ได้สร้างจากการอนุมาน (Interface) โดยการนำชุดข้อมูลในอดีตมาจำลองและสร้างเป็นตัวแบบ ร่วมกับการใช้ข้อมูลแบบฝึกหัด (Training Data) สำหรับใช้ในการทำนายประเภทตัวอย่างในอนาคต ข้อมูลทุกตัวจะมีคุณสมบัติที่จะใช้ในการทำนาย อัลกอริธึมประเภทนี้จึงมุ่งเน้นการแบ่งแยกข้อมูลออกเป็นกลุ่มตามค่าคุณสมบัติของข้อมูล ถ้า

คุณสมบัติของข้อมูลมีค่าที่ไม่ต่อเนื่องจะเรียกกระบวนการที่ใช้แบ่งแยกนี้ว่า การจำแนกประเภท (Classification) แต่หากค่าคุณสมบัติของข้อมูลมีค่าต่อเนื่องจะเรียกกระบวนการที่ใช้แบ่งแยกนี้ว่า การถดถอย (Regression) หรือการพยากรณ์ (Forecasting)

รูปแบบที่สอง แบบจำลองในการบรรยาย (Descriptive Modeling) หรือการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) เป็นการนำข้อมูลที่มีอยู่มาวิเคราะห์และสร้างแบบแผนจากข้อมูลที่ได้รับเข้าไปเพื่อศึกษาหาความสัมพันธ์โดยใช้อัลกอริธึมการค้นหาความสัมพันธ์ (Association Algorithm) ของข้อมูลจากข้อมูลขนาดใหญ่หรือการใช้อัลกอริธึมจัดกลุ่มข้อมูล (Clustering Algorithm) ที่เป็นเทคนิคในการจำแนกกลุ่มของข้อมูลที่มีลักษณะคล้ายกันไว้ในกลุ่มเดียวกัน เป็นต้น

1.5 เทคนิคการทำเหมืองข้อมูล

นิสานันท์ พลอาสา (2558, น. 24-26) การแก้ปัญหาของงานชนิดต่าง ๆ โดยวิธีการ Data Mining มีเทคนิคหลากหลาย ซึ่งสามารถเลือกนำมาใช้ในแต่ละงานอย่างเหมาะสม โดยส่วนใหญ่จะใช้หลักการปัญญาประดิษฐ์ Artificial Intelligence (AI) หรือศาสตร์อื่น ๆ โดยแบ่งออกเป็น 3 เทคนิคดังนี้

การวิเคราะห์กลุ่ม (Cluster Analytic) คือการจัดการข้อมูลซึ่งมีลักษณะคล้ายกับการแบ่งประเภทแต่จะไม่เหมือนกันโดยการแบ่งประเภทจะวิเคราะห์ข้อมูลตามต้นแบบ แต่สำหรับการแบ่งกลุ่มเป็นการวิเคราะห์โดยไม่พิจารณาจัดกลุ่มตามประเภทที่มีหรือที่รู้จัก แต่จะใช้ขั้นตอนวิธีการจัดกลุ่มเพื่อค้นหากลุ่มที่สามารถยอมรับได้เพื่อจัดกลุ่ม กล่าวคือกลุ่มของวัตถุมีการสร้างขึ้นโดยเปรียบเทียบวัตถุที่มีความเหมือนกันจัดเข้ากลุ่มเดียวกัน

กฎการหาความสัมพันธ์ (Association Rule) เป็นการค้นหาความสัมพันธ์ของข้อมูลทั้งสองชุดเหรือมากกว่าสองชุดขึ้นไปไว้ด้วยกัน ความสำคัญของกฎทำการวัดโดยใช้ข้อมูลสองตัวด้วยกันคือค่าสนับสนุน (Support) ซึ่งเป็นเปอร์เซ็นต์ของการดำเนินการที่กฎสามารถนำมาใช้หรือเป็นเปอร์เซ็นต์ของการดำเนินการที่กฎที่ใช้มีความถูกต้องและข้อมูลตัวที่สองที่นำมาใช้วัดคือค่าความมั่นใจ (Confidence) ซึ่งเป็นจำนวนของกรณีที่กฎถูกต้องโดยสัมพันธ์กับจำนวนของกรณีที่กฎความสามารถนำไปใช้ได้ ในการหากฎความสัมพันธ์นั้นจะมีขึ้นตอนวิธีการหาหลายวิธีด้วยกัน แต่ขึ้นตอนที่เป็นที่นิยมและใช้กันอย่างแพร่หลายคือ อัลกอริธึม Apriori

การจำแนกข้อมูลวิเคราะห์ (Classification Analytic) เป็นการจัดแบ่งประเภทของข้อมูล โดยหาชุดต้นแบบหรือชุดของการทำงานที่อธิบายและแบ่งประเภทข้อมูล วัตถุประสงค์เพื่อให้สามารถใช้เป็นต้นแบบทำนายประเภทของวัตถุหรือข้อมูลที่ไม่มีการระบุประเภทหรือชนิดของข้อมูลซึ่งต้อนแบบจากการวิเคราะห์ชุดของข้อมูลฝึกสอน (Training Data) โดยอาจจะเป็นกลุ่มข้อมูลที่มีการระบุประเภทหรือกลุ่มเรียบร้อยแล้ว รูปแบบของต้นแบบแสดงได้หลายแบบเช่น Classification Rules, Decision Trees หรือ Neural Network เป็นต้น

เอกสิทธิ์ พัชรวงศ์ศักดา (2557, หน้า 15) กล่าวถึงการทำเหมืองข้อมูลไว้ 2 เทคนิค ดังนี้ 1) เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และ2) เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning)

เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการนำข้อมูลมาวิเคราะห์โดยพิจารณาหาความสัมพันธ์ของชุดข้อมูลที่มีอยู่ เพื่อนำรูปแบบผลลัพธ์จากการวิเคราะห์ข้อมูลที่ได้ไปใช้ประโยชน์โดยตรงกับชุดข้อมูลเดียวกัน รูปแบบการเรียนรู้แบบไม่มีผู้สอน คือ การวิเคราะห์ ความสัมพันธ์ และการวิเคราะห์การจัดกลุ่ม

เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการรวบรวมชุดข้อมูลในอดีต มาวิเคราะห์โดยพิจารณาความสัมพันธ์หรือรูปแบบเฉพาะของชุดข้อมูล เพื่อนำรูปแบบผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลไปใช้ในการคาดการณ์หรือทำนายสิ่งที่จะเกิดขึ้นในอนาคต เทคนิควิธีการวิเคราะห์ข้อมูลนี้คือการวิเคราะห์จำแนกประเภทข้อมูล

ดังนั้นสรุปได้ว่า การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการที่ค้นหาหรือสกัดข้อมูลจากชุดข้อมูลขนาดใหญ่ เพื่อให้ได้รูปแบบความสัมพันธ์ของข้อมูลหรือพฤติกรรมที่ซ่อนอยู่ในชุดข้อมูล โดยแบ่งออกเป็นการหากฎความสัมพันธ์ (Association Rule) การจำแนกข้อมูล (Data Classification) การแบ่งกลุ่มข้อมูล (Data Clustering) และจินตทัศน์ (Visualization) ซึ่งข้อมูลที่ได้จากการทำเหมืองข้อมูลสามารถนำไปใช้ประโยชน์ในด้านการตัดสินใจและนำไปใช้ในด้านการพัฒนาด้านต่าง ๆ ได้

2. การคัดเลือกคุณสมบัติ

นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) ให้ความหมายการคัดเลือกคุณลักษณะว่า การคัดเลือกคุณลักษณะเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่ใช้เป็นตัวแบบพยากรณ์ โดยอาจเลือกตัวแปรที่ดีที่สุดเพียงหนึ่งตัวแปรหรือเป็นกลุ่มของตัวแปรที่มีความสำคัญ ซึ่งกระบวนการคัดเลือกคุณสมบัติมีความสำคัญในการเตรียมข้อมูลเพื่อสร้างตัวแบบในการทำเหมืองข้อมูล เนื่องจากเป็นการลดมิติข้อมูลให้น้อยลงและช่วยให้ข้อมูลการเรียนรู้ที่นำมาเข้าสู่กระบวนการมีความรวดเร็วและมีประสิทธิภาพ โดยวิธี Correlation Based Feature Selection ถือเป็นการคัดเลือกคุณสมบัติอย่างง่ายโดยใช้หลักการคำนวณค่าความสัมพันธ์ระหว่างคุณสมบัติย่อยต่อค่าพยากรณ์ อาจใช้คำนวณค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson’s Correlation) มีการจัดลำดับความสัมพันธ์เพื่อประเมินค่าความสามารถในการพยากรณ์ของแต่ละคุณสมบัติ และยังสามารถพิจารณาคัดเลือกกลุ่มของคุณสมบัติที่มีความสัมพันธ์ภายในระหว่างคุณสมบัติย่อยกันเองต่ำเพื่อลดความซ้ำซ้อนของอิทธิพลการพยากรณ์

วิรัตน์ ชูนุ้ย (2555) ให้ความหมายการคัดเลือกคุณลักษณะว่า การคัดเลือกคุณลักษณะเป็นการลดขนาดและคัดเลือกด้วยวิธีการเลือกคุณลักษณะเฉพาะโดยเลือกกลุ่มของลักษณะเฉพาะใหม่จากเดิม ซึ่งกลุ่มใหม่ที่จะได้เป็นกลุ่มย่อยของกลุ่มลักษณะเฉพาะเดิม การคัดเลือกคุณลักษณะเฉพาะนี้เป็นการเลือกกลุ่มของลักษณะเฉพาะใหม่แยกจากกลุ่มเฉพาะเดิม โดยรูปแบบฟิลเตอร์เป็นวิธีการ

เลือกลักษณะเฉพาะที่มีค่าความสำคัญมากที่สุดและเป็นค่าที่คำนวณได้ง่าย ซึ่งวิธีการ Information Gain (IG) จะเลือกคุณลักษณะเฉพาะที่มีค่า Information Gain สูงที่สุดหรือมีค่า Entropy น้อยที่สุด

การคัดเลือกคุณสมบัติ (Feature Selection) เป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่จะใช้ในตัวแบบพยากรณ์ การเลือกตัวแปรอาจคัดเลือกตัวแปรที่ดีที่สุดเพียงตัวเดียวหรือเลือกกลุ่มของตัวแปรที่มีความสำคัญต่อการพยากรณ์ กระบวนการคัดเลือกคุณเป็นกระบวนการที่สำคัญในการเตรียมข้อมูลของการทำเหมืองข้อมูล ซึ่งจะส่งผลให้การสร้างตัวแบบพยากรณ์มีประสิทธิภาพเพราะจะช่วยลดมิติของข้อมูลและอาจช่วยให้การเรียนรู้ในวิธีการพยากรณ์ดำเนินการได้เร็วและมีประสิทธิภาพมากขึ้น

นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) ได้ทดลองใช้การคัดเลือกคุณสมบัติในการวิจัย เรื่องการวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์ โดยใช้วิธีการคัดเลือก 3 วิธีดังนี้

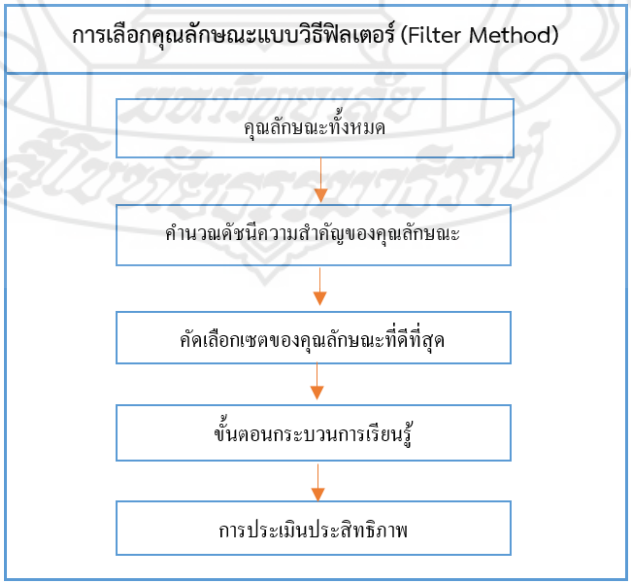
แบบที่ 1 การคัดเลือกคุณสมบัติแบบ Correlation Base Feature Selection เป็นการคัดเลือกกลุ่มคุณสมบัติอย่างง่าย ใช้หลักการคำนวณค่าความสัมพันธ์ระหว่างคุณสมบัติย่อยต่อค่าพยากรณ์ ซึ่งอาจใช้คำนวณด้วยค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's Correlation) มีการจัดลำดับตามค่าความสัมพันธ์เพื่อประเมินค่าความสามารถในการพยากรณ์ของแต่ละคุณสมบัติ นอกจากนั้นยังพิจารณาคัดเลือกกลุ่มของคุณสมบัติที่มีความสัมพันธ์ภายในระหว่างคุณสมบัติย่อยเพื่อลดความซ้ำซ้อนของอิทธิพลการพยากรณ์ (Hall and Smith. 1998)

แบบที่ 2 การคัดเลือกคุณสมบัติแบบ Consistency Based Feature Selection เป็นการคัดเลือกคุณสมบัติที่ต้องการกำหนดตัวชี้วัดความสอดคล้องมั่นคงไว้ก่อนเป็นอันดับแรก จากนั้นใช้ตัวชี้วัดนี้เพื่อประเมินความมั่นคงของกลุ่มคุณสมบัติที่เกี่ยวข้องกับค่าพยากรณ์เดียวกัน (Liu and Setiono. 1996; Liu et al. 1998) ตัวชี้วัดความมั่นคงที่ถูกกำหนดให้เป็นตัวชี้วัดนี้ อาจกำหนดมาตรวัดลักษณะเดียวกับการวัดระยะทางของคุณสมบัติย่อยและสามารถแปลผลคุณสมบัติที่สอดคล้องกันมากด้วยค่าเข้าใกล้ศูนย์ กระบวนการคัดเลือกคุณสมบัติจะกระทำซ้ำและเลือกกลุ่มคุณสมบัติที่มีค่าตัวชี้วัดน้อยและจำนวนคุณสมบัติเท่าเดิมหรือลดลงเท่านั้น วิธีการค้นหากลุ่มคุณสมบัติที่สอดคล้องมั่นคงนี้เป็นวิธีที่รวดเร็วและสามารถทราบความสัมพันธ์ระหว่างคุณสมบัติได้ (Hall and Holmes. 2003)

แบบที่ 3 การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection เป็นวิธีคัดเลือกตัวแปรโดยมีหลักการแบบเดียวกับการเลือกตัวแปรของการสร้างต้นไม้ตัดสินใจ เพื่อให้ได้ตัวแปรที่เป็นตัวแบ่งข้อมูลออกเป็นกลุ่มย่อยที่มีสมาชิกภายในกลุ่มเป็นชนิดเดียวกันมากที่สุด (Homogeneous) ด้วยมาตรวัดการได้ประโยชน์จากการแบ่งกลุ่มย่อยเรียกว่า อัตราส่วนเกน (Gain Ratio) ซึ่งเป็นอัตราส่วนของค่าเกน (Gain หรือ Information Gain) กับค่าสารสนเทศการแบ่งกลุ่ม (Split Info) อันเป็นการลดอิทธิพลของตัวแปรที่มีค่าหลายค่า ผลที่ได้รับจากการใช้เทคนิคนี้จะได้ลำดับของตัวแปรซึ่งตัวแปรอยู่ที่ลำดับแรก ๆ จะถือว่ามีอิทธิพลในการพยากรณ์ตัวแปรเป้าหมายมากกว่าตัวแปรในลำดับถัดไป ทำให้เราสามารถพิจารณาเลือกจำนวนตัวแปรที่เหมาะสมได้อย่างมีประสิทธิภาพ (Tan, Steinbach and Kumar. 2006; Asha, Manjunath and Jayaram. 2010)

การคัดเลือกคุณลักษณะ (Feature Selection) (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557) เป็นกระบวนการจัดกลุ่มของข้อมูล ตัวแปรของข้อมูล ซึ่งการคัดเลือกคุณลักษณะช่วยลดตัวแปรหรือมิติของข้อมูลอาจทำให้เหลือคุณลักษณะที่มีความสำคัญและดีที่สุดเพียงหนึ่งตัวหรือกลุ่มของตัวแปรเพียงหนึ่งกลุ่มที่มีความสำคัญต่อการพยากรณ์ หากได้คุณลักษณะที่ดีจะช่วยให้ตัวแบบมีประสิทธิและทำงานได้เร็วขึ้น โดยวิธีการคัดเลือกคุณลักษณะสามารถแบ่งออกได้เป็น 3 วิธี ได้แก่ วิธีฟิลเตอร์ (Filter Method) วิธีแรปเปอร์ (Wrapper Method) และวิธีฝังตัว (Embed Method) ซึ่งกระบวนการเลือกตัวแปรหรือคุณลักษณะที่สำคัญ มีดังต่อไปนี้

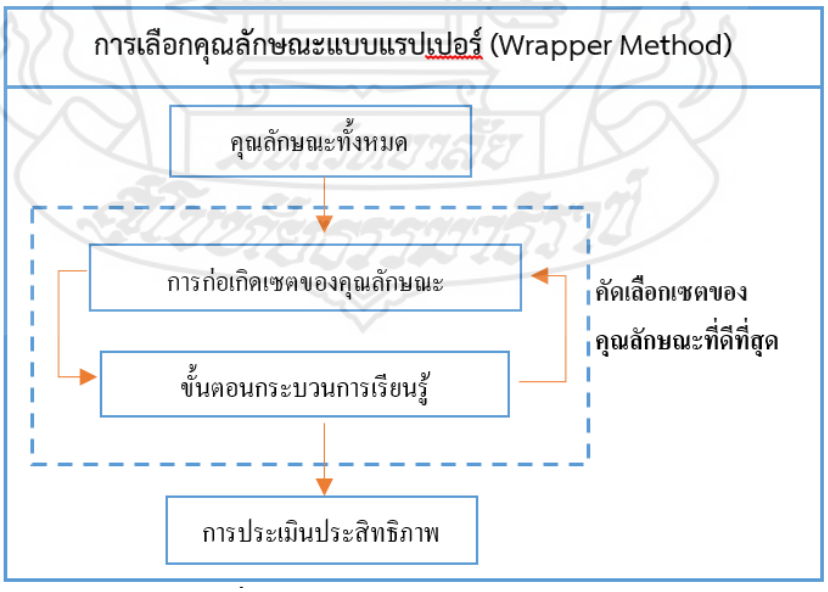
1) กระบวนการเลือกคุณลักษณะแบบวิธีฟิลเตอร์ (Filter Method) เป็นวิธีที่ใช้เทคนิคการจัดเรียงลำดับของคุณลักษณะหรือเรียกอีกชื่อว่า Feature Ranking technique สำหรับคัดเลือกคุณลักษณะที่เหมาะสม เริ่มจากคุณลักษณะทั้งหมดจะถูกวัดค่าความสำคัญจากนั้นกำหนดระดับ Threshold ขึ้นมาเพื่อตัดคุณลักษณะที่มีค่าดัชนีของความสำคัญน้อยกว่าระดับที่อ้างอิงออกไป ตัวอย่างของวิธีการเลือกคุณลักษณะ ได้แก่ วิธี Minimal Redundancy and Maximum-Relevance(mRMR), วิธี Relief Feature, Fisher Score Chi-square Test และ Information Gain เป็นต้น ข้อดีของวิธีเลือกคุณลักษณะแบบฟิลเตอร์นี้คือเป็นเทคนิคที่คำนวณได้ง่าย รวดเร็ว และหลีกเลี่ยงการเกิดโอเวอร์ฟิตติ้ง (Overfitting) เพราะวิธีนี้ไม่นำผลทดสอบประสิทธิภาพการเรียนรู้ของเครื่องมาพิจารณาร่วมด้วย ซึ่งคุณลักษณะที่ถูกคัดเลือกจะไม่ถูกเลือกด้วยความเอนเอียงหรือไบแอส (Bias) สำหรับข้อเสียของวิธีนี้ คือคุณลักษณะที่ถูกคัดเลือกเป็นคุณลักษณะที่เป็นอิสระต่อกันเพราะขั้นตอนการคำนวณค่าความสำคัญจะพิจารณาความสัมพันธ์เพียงด้านเดียว คือความสัมพันธ์ระหว่างคุณลักษณะนั้นกับข้อมูลเอาท์พุตเท่านั้น ไม่ได้คำนึงถึงความสัมพันธ์ระหว่างคุณลักษณะกันเองเซตของคุณลักษณะที่ถูกเลือกอาจมีความสัมพันธ์ระหว่างกันหรือไม่มีความสัมพันธ์ระหว่างกันก็ได้ เมื่อนำเซตคุณลักษณะที่ถูกเลือกเหล่านี้มาใช้ในการจำแนกกลุ่มข้อมูลจึงส่งผลทำให้ค่าความถูกต้องของการเรียนรู้ลดลง ดังนั้นการเลือกคุณลักษณะแบบฟิลเตอร์จึงเหมาะสมเฉพาะการวิเคราะห์ข้อมูลที่มีจำนวนมิติไม่สูงมากนัก



ภาพที่ 2.2 การเลือกคุณลักษณะวิธีฟิลเตอร์

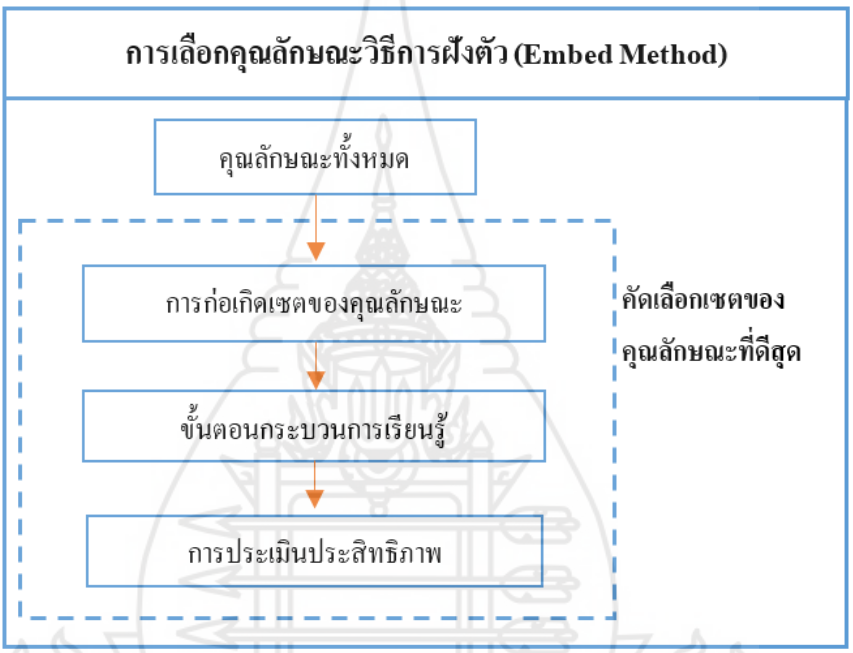
จากภาพที่ 2.2 แสดงกระบวนการการเลือกคุณลักษณะแบบวิธีฟิลเตอร์ โดยมีขั้นตอนตั้งแต่การนำคุณลักษณะทั้งหมด มาคำนวณดัชนีความสำคัญของคุณลักษณะ ลำดับต่อมาคัดเลือกเซตของคุณลักษณะที่ดีที่สุด ลำดับต่อมาเข้าสู่ขั้นตอนกระบวนการเรียนรู้ และการประเมินประสิทธิภาพในขั้นตอนสุดท้าย

2) วิธีเลือกคุณลักษณะแบบแรปเปอร์ (Wrapper Method) เป็นวิธีที่ถูกพัฒนามาเพื่อแก้ไขวิธีฟิลเตอร์ ซึ่งวิธีแรปเปอร์มีกระบวนการทำงานที่เพิ่มขึ้นคือคุณลักษณะทั้งหมดจะถูกจัดให้อยู่ในรูปของเซตคุณลักษณะ จากนั้นดำเนินการค้นหาเซตของคุณลักษณะที่เหมาะสมด้วยการประเมินด้วยฟังก์ชันความเหมาะสม (Fitness Function) เพื่อค้นหาเซตคุณลักษณะที่เหมาะสมที่สุดก่อนเข้าสู่กระบวนการจำแนกข้อมูล วิธีค้นหาเซตคุณลักษณะที่เหมาะสมประกอบด้วย 2 วิธี ได้แก่ 1) วิธีค้นหาเลือกคุณลักษณะโดยลำดับ (Sequential selection algorithm) เป็นวิธีการที่เลือกคุณลักษณะด้วยการพิจารณาคุณลักษณะที่เพิ่มทีละตัวตามลำดับหรือลดคุณลักษณะลงทีละตัวตามลำดับจนกว่าจะได้เซตคุณลักษณะที่เหมาะสม วิธีการนี้เป็นวิธีการที่มีกระบวนการดำเนินงานง่าย แต่มีข้อเสียคือ ใช้เวลาในการประมวลผลในขั้นตอนของการเรียนรู้นานมาก 2) วิธีการค้นหาเซตของคุณลักษณะด้วยวิธีการสุ่มเลือกหรือ วิธีค้นหาคำตอบที่เหมาะสมด้วยอัลกอริธึมของการแก้ไขปัญหาฮิวริสติก (Heuristic Search Algorithm) อันได้แก่ Genetic algorithm (GA), Particle Swarm Optimization (PSO) เป็นต้น กระบวนการค้นหาเซตของคุณลักษณะจะเริ่มต้นด้วยการกำหนดจำนวนของเซตของคุณลักษณะและจำนวนของคุณลักษณะที่ถูกเลือก แต่ละเซตของคุณลักษณะจะถูกประเมินด้วยฟังก์ชันความเหมาะสมและค่าฟังก์ชันความเหมาะสมของแต่ละเซตคุณลักษณะจะถูกนำมาเปรียบเทียบเพื่อค้นหาเซตของคุณลักษณะที่เหมาะสมที่สุด จากนั้นเซตของคุณลักษณะที่เหมาะสมที่สุดจะถูกนำไปใช้กระบวนการจำแนกข้อมูลต่อไป ข้อดีของวิธีการค้นหาแบบฮิวริสติกเมื่อเทียบกับวิธี SSA เป็นวิธีที่ใช้เวลาน้อยกว่า ซึ่งกระบวนการเลือกคุณลักษณะแบบแรปเปอร์ (Wrapper Method) ดังภาพที่ 2.3



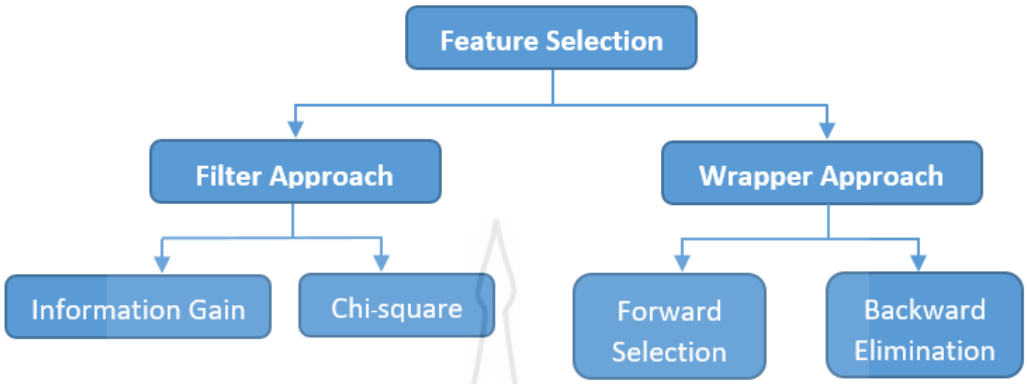
ภาพที่ 2.3 การเลือกคุณลักษณะวิธีแรปเปอร์

3) วิธีเลือกคุณลักษณะวิธีการฝังตัว (Embed Method) เป็นวิธีที่ออกแบบมาเพื่อแก้ไขข้อเสียของวิธีฟิลเตอร์ (Filter Method) และวิธีแรปเปอร์ (Wrapper) ซึ่งจะใช้เวลาในการประมวลผลน้อยกว่าวิธีแรปเปอร์ กระบวนการเลือกคุณลักษณะของวิธีนี้ได้รวมการเลือกคุณลักษณะไว้เป็นส่วนหนึ่งของกระบวนการเรียนรู้ซึ่งมีข้อดีคือ มีการค้นหาเซตของคุณลักษณะทั้ง Global Space และ Local Space จึงทำให้มีประสิทธิภาพในการค้นหามีประสิทธิภาพที่ดีขึ้น แต่อย่างไรก็ตามวิธีฝังตัวมีข้อเสีย คือการเลือกเซตคุณลักษณะไม่มีความยืดหยุ่นเนื่องจากขึ้นอยู่กับอัลกอริธึมการจำแนกกลุ่มข้อมูล ซึ่งกระบวนการเลือกคุณลักษณะวิธีการฝังตัว (Embed Method) ดังภาพที่ 2.4



ภาพที่ 2.4 การเลือกคุณลักษณะวิธีวิธีการฝังตัว

เอกสิทธิ์ พัชรวงศ์ศักดา (2557) ยังได้อธิบายการคัดเลือกคุณลักษณะ (Feature Selection) เพื่อนำมาใช้จำแนกประเภทข้อมูล โดยในการจำแนกประเภทของข้อมูล (Classification) ในหลาย ๆ ครั้ง พบว่าการแยกแอตทริบิวต์ (Attribute) หรือฟีเจอร์ (Feature) จำนวนมากในข้อมูลตัวอย่างเช่น เช่น การจำแนกประเภทข้อความทัศนคติ (Sentiment) ออกเป็นเชิงบวก (Positive) หรือเชิงลบ (Negative) นั้นจะมีจำนวนคำในข้อความต่างๆ ที่ใช้เป็นฟีเจอร์จำนวนมาก ฟีเจอร์เหล่านี้บางอันก็ไม่ได้มีความสำคัญในการแบ่งแยกคลาส (Class) ออกเป็นเชิงบวกหรือเชิงลบได้ ดังนั้นจึงจำเป็นต้องมีการคัดเลือกฟีเจอร์ที่สำคัญมาใช้งาน โดยมีขั้นตอนที่สามารถแบ่งออกเป็น 2 กลุ่มใหญ่ ดังภาพที่ 2.5



ภาพที่ 2.5 แผนผังแสดงการคัดเลือกคุณลักษณะ

กลุ่มที่ 1 Filter approach เป็นการคัดเลือกแอตทริบิวต์หรือฟีเจอร์โดยใช้การคำนวณหาค่าน้ำหนักซึ่งอาจจะเป็นค่าความสัมพันธ์ระหว่างแต่ละฟีเจอร์และคลาสต่าง ๆ และจะเลือกฟีเจอร์โดยเรียงลำดับตามค่าน้ำหนักที่คำนวณได้แล้วเลือกฟีเจอร์ที่มีค่าน้ำหนักมากกว่าที่ต้องการมาใช้งานต่อไป วิธีการนี้ต่างจากวิธีการ Wrapper ตรงที่วิธีการนี้จะไม่มีการสร้างโมเดลเพื่อคัดเลือกฟีเจอร์ เทคนิคในการคำนวณค่าน้ำหนักของฟีเจอร์ต่าง ๆ มีหลายวิธี ตัวอย่างเช่น Information Gain, Chi-Square หรือ Correlation

กลุ่มที่ 2 Wrapper approach เป็นการคัดเลือกแอตทริบิวต์หรือฟีเจอร์ด้วยการสร้างโมเดล (Classification model) ขึ้นมาจากเซตของฟีเจอร์ที่กำหนดไว้และวัดประสิทธิภาพการทำงานของโมเดลและเลือกเซตของฟีเจอร์ที่ทำให้โมเดลมีประสิทธิภาพมากที่สุดมาใช้งาน ตัวอย่างเช่น โมเดลที่ให้ค่าความถูกต้อง (Accuracy) มากที่สุด การคัดเลือกฟีเจอร์ด้วยวิธีการนี้ยังแบ่งย่อยได้เป็น 2 แบบใหญ่ๆ คือ 1) Forward Selection เป็นการสร้างโมเดลโดยการเพิ่มฟีเจอร์ทีละ 1 ฟีเจอร์ ถ้าฟีเจอร์ที่ใส่เพิ่มให้ประสิทธิภาพที่ดีก็จะเก็บไว้และเลือกฟีเจอร์อื่น ๆ มาเพิ่มต่อไปจนประสิทธิภาพของโมเดลไม่ได้ดีขึ้นก็จะหยุดทำงาน และ2) Backward Elimination เป็นการสร้างโมเดลที่เริ่มจากการใช้ฟีเจอร์ทั้งหมดก่อนและตัด (Eliminate) ฟีเจอร์ที่ไม่สำคัญทิ้งไปทีละฟีเจอร์ถ้าประสิทธิภาพดีขึ้นก็ตัดฟีเจอร์อื่น ๆ ต่อไป

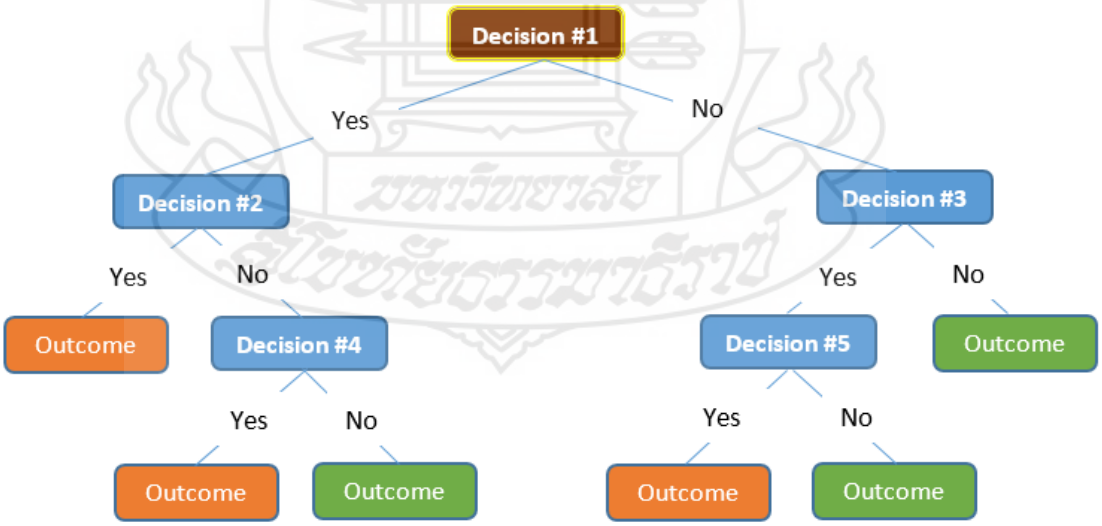
ดังนั้นสรุปได้ว่า การคัดเลือกคุณลักษณะ (Feature Selection) เป็นกระบวนการจัดกลุ่มของข้อมูล ตัวแปรของข้อมูล โดยวิธีการหาค่าน้ำหนักด้วยวิธี Filter Approach เป็นการนำค่าฟีเจอร์มาพิจารณาแต่จะไม่มีการสร้างโมเดลในการคัดเลือกฟีเจอร์และการคัดเลือกด้วยการวัดประสิทธิภาพของฟีเจอร์โดยการเพิ่มตัวที่มีประสิทธิภาพมากหรือตัดลดตัวที่มีประสิทธิภาพน้อยออกโดยใช้วิธี Wrapper approach ซึ่งวิธี Wrapper approach นี้จะมีการสร้างโมเดลขึ้นจากฟีเจอร์และวัดประสิทธิภาพจากการเพิ่มหรือลดฟีเจอร์ ซึ่งการคัดเลือกคุณลักษณะเป็นการช่วยลดตัวแปรหรือมิติของข้อมูลอาจทำให้เหลือคุณลักษณะที่มีความสำคัญและดีที่สุดเพียงหนึ่งตัวหรือกลุ่มของตัวแปรเพียง

หนึ่งกลุ่มที่มีความสำคัญต่อการพยากรณ์ ถือได้ว่าเป็นขั้นตอนที่สำคัญซึ่งหากได้คุณลักษณะที่ดีจะช่วยให้ผลลัพธ์ของตัวแบบมีประสิทธิภาพยิ่งขึ้นและทำงานได้เร็วขึ้น

3. การจำแนกประเภทข้อมูล

เทคนิคการจำแนกประเภทข้อมูล (Classification Techniques) เทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่าง ๆ ที่ได้มีการกำหนดไว้ นำมาสร้างแบบจำลองเพื่อการพยากรณ์ค่าข้อมูล (Predictive Model) ในอนาคตเรียกว่า Supervised Learning ตัวอย่างเช่น เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคโครงข่ายประสาทเทียม (Neural Network) เป็นต้น

3.1 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) ต้นไม้ตัดสินใจถูกพัฒนาโดย Quinlan (1986, p. 1) เป็นวิธีการที่ได้รับความนิยมอย่างแพร่หลายเนื่องจากโมเดลที่ได้จากการใช้เทคนิคนี้สามารถแปลความหมายและเข้าใจง่ายลักษณะของรูปแบบข้อมูล (Pattern) ซึ่งแต่ละโหนด (Node) จะแสดงคุณลักษณะหรือแอตทริบิวต์ (Attribute) ที่ใช้ทดสอบข้อมูล แต่ละกิ่งจะแสดงผลในการทดสอบและลีฟโหนด (Leaf Node) จะแสดงกลุ่มหรือคลาส (Class) ที่กำหนดไว้ อัลกอริธึมของเทคนิคต้นไม้ตัดสินใจส่วนใหญ่ไม่รองรับข้อมูลแบบต่อเนื่องจึงต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่องก่อน อัลกอริธึมสำหรับการแบ่งให้ข้อมูลเป็นแบบไม่ต่อเนื่อง ได้แก่ ID3, C4.5 และ C5.0 อัลกอริธึมของเทคนิคต้นไม้ตัดสินใจ C4.5 (J48) เป็นเทคนิคการสร้างตัวแบบต้นไม้ตัดสินใจ โดยการพิจารณาค่าบ่งชี้ความเหมาะสมของแอตทริบิวต์ (Attribute) เรียกว่า Gini (Gini Index) ดังแสดงตัวอย่างในภาพที่ 2.6



ภาพที่ 2.6 เทคนิคต้นไม้ตัดสินใจ

จากภาพที่ 2.6 แสดงรูปแบบการตัดสินใจด้วยเทคนิคต้นไม้ตัดสินใจ ในแต่ละกิ่งจะมีการติดสินใจและแยกออกเป็นกิ่งถัดไป

การสร้างต้นไม้ตัดสินใจ (โกเมศ อัมพวัน, 2548) ในช่วงปลายยุค 1970 มีนักวิจัยทางด้านการเรียนรู้ของเครื่อง (Machine Learning) คือ J. Ross Quinlan ได้คิดค้นอัลกอริธึมสำหรับสร้างต้นไม้ตัดสินใจที่มีชื่อว่า ID3 (Interactive Dichotomiser) ต่อมาได้มีการพัฒนาต่อยอด ID3 เป็น C4.5 ซึ่งถือได้ว่าเป็นอัลกอริธึมพื้นฐานที่ใช้สำหรับเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริธึมต่าง ๆ ทางด้านการเรียนรู้แบบมีผู้สอน (Supervised Learning)

ID3 และ C4.5 ได้ประยุกต์ใช้วิธีการเชิงละโมภ (Greedy Approach) ได้ประยุกต์ใช้วิธีการเชิงละโมภ (Greedy Approach) การสร้างต้นไม้ภายใต้วิธีการแบบ “Top-dawn recursive divide-and-conquer” โดยพิจารณาชุดข้อมูลสำหรับเรียนรู้ (Training data, เซตของเรคคอร์ดของข้อมูลที่แต่ละเรคคอร์ดจะประกอบไปด้วยเซตของแอตทริบิวต์ต่างๆ และแอตทริบิวต์ที่บ่งบอกถึงหมวดหมู่ของข้อมูลของข้อมูลเรคคอร์ดนั้น ๆ) ด้วยการแบ่งข้อมูลออกเป็นส่วนย่อยในระหว่างกระบวนการสร้างต้นไม้

กระบวนการต้นไม้จะเลือกแอตทริบิวต์ (Attribute) ที่มีความสัมพันธ์กับคลาส (Class) มากที่สุดมาเป็นโหนดแม่ (Root) เพื่อทำการแบ่งข้อมูลออกเป็นกลุ่มโดยจำนวนกลุ่มจะเท่ากับจำนวนค่าที่เป็นไปได้ทั้งหมดของแอตทริบิวต์ที่เลือกมาเป็นตัวแบ่ง จากนั้นก็จะทำการหาแอตทริบิวต์ถัดไปเรื่อย ๆ มาแบ่งกลุ่มข้อมูลที่อยู่ในโหนดลูกต่อไปจนกระทั่งในแต่ละกลุ่มที่แบ่งออกมานั้นมีคำตอบฉลากเป็นชนิดเดียวกันทั้งหมด หรือแอตทริบิวต์ทุกตัวถูกใช้ในการแบ่งกลุ่มหมดแล้ว จึงหยุดแบ่งกลุ่มข้อมูล ซึ่งในการหาความสัมพันธ์ระหว่างแอตทริบิวต์ใด ๆ กับคลาสจะใช้ตัววัด คือ ค่าเกนความรู้ (Information Gain หรือเรียกอีกชื่อว่า IG) คำนวณได้จากสมการ 2.1

$$IG(parent, child) = entropy(parent) - [p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + \ldots] \quad (2.1)$$

$$entropy(c_1) = -p(c_1) \log p(c_1)$$

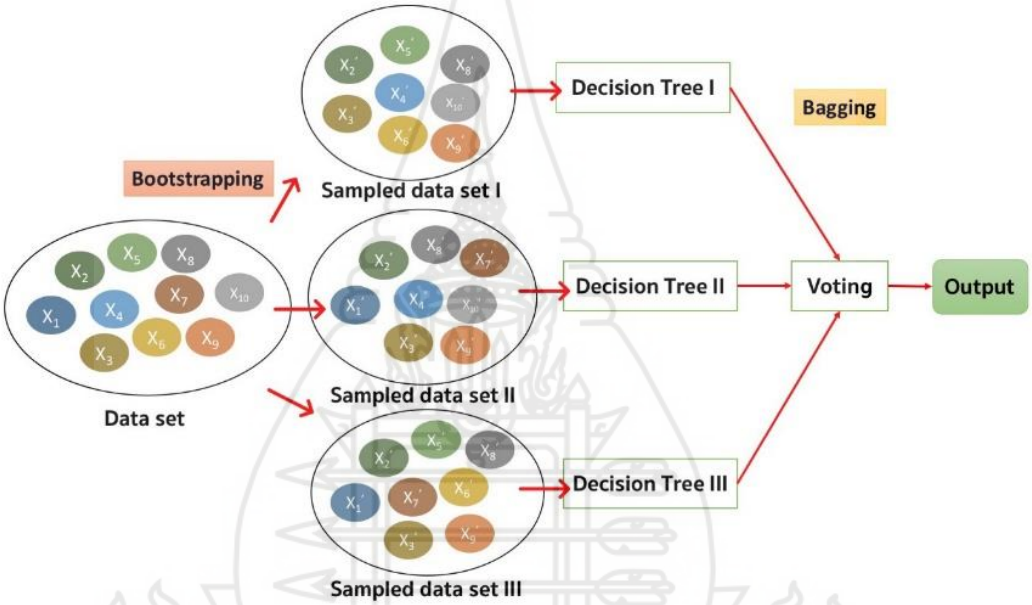
เมื่อ $entropy(c_1)$ คือ $-p(c_1) \log p(c_1)$

$p(c_1)$ คือ ค่าความน่าจะเป็นของค่า c_1

c คือ ปัจจัย (Attribute) ของข้อมูลแต่ละตัวที่เกี่ยวข้อง

ค่า Entropy ใช้วัดค่าความแตกต่างกันของข้อมูล ซึ่งถ้าค่า Entropy มีค่าต่ำจะหมายถึงข้อมูลมีความแตกต่างกันน้อย และถ้าค่า Entropy มีค่าสูงจะหมายถึงข้อมูลมีความแตกต่างกันมาก (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

3.2 เทคนิคป่าสุ่ม (Random Forest) แนวคิดของเทคนิคป่าสุ่ม (ปริญญา สงวนสัตย์. 2558) เป็นการสร้างแบบจำลอง (Model) ด้วยวิธีการต้นไม้ตัดสินใจ (Decision Tree) ขึ้นมาหลาย ๆ แบบจำลอง โดยวิธีการสุ่มตัวแปรภายหลังจากการสร้างนำผลลัพธ์ที่ได้แต่ละแบบจำลองมารวมกัน พร้อมนับจำวนผลที่มีจำนวนซ้ำกันมากที่สุด สกัดออกมาเป็นผลลัพธ์สุดท้าย วิธีการของ Decision Tree คือเทคนิคที่ให้ผลลัพธ์ในลักษณะเป็นโครงสร้างของต้นไม้ ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดจะมีเงื่อนไขของคุณลักษณะเป็นตัวทดสอบ กิ่งของต้นไม้ (Branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบและใบ (Leaf) เป็นสิ่งที่อยู่ล่างสุดของต้นไม้แสดงถึงกลุ่มของข้อมูล (Class) ก็คือผลลัพธ์ที่ได้จากการพยากรณ์ ซึ่งข้อดีของวิธีการนี้คือให้ผลการพยากรณ์ที่แม่นยำและเกิดปัญหา Overfitting น้อย ดังแสดงตัวอย่างในภาพที่ 2.7



ภาพที่ 2.7 หลักการทำ Random Forest

ที่มา : วิชญ์พงศ์ ดรุณธรรม (2561)

3.3 เทคนิคนาอีฟเบย์ (Naïve Bayes) การจำแนกประเภทข้อมูลด้วยวิธีนาอีฟเบย์ (Naïve Bayes) (T.Bayes and R.Price, 1763) เป็นวิธีที่ได้รับความนิยมเนื่องจากการสร้างโมเดลง่ายและไม่มีความซับซ้อนถูกพัฒนาขึ้นโดย Thomas Bayes อาศัยหลักการของความน่าจะเป็นเข้ามาช่วยในการหาคำตอบของประเภทตัวอย่างใหม่ หลักการการคำนวณความน่าจะเป็นแบบมีเงื่อนไขที่เรียกว่า Conditional Probability สามารถคำนวณได้จากสมการ 2.2

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (2.2)$$

โดย 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

ในลักษณะเดียวกันเราจะเขียน P(B|A) หรือค่าความน่าจะเป็นที่เหตุการณ์ A เกิดขึ้นก่อนและเหตุการณ์ B เกิดขึ้นตามมาทีหลังได้เป็นสมการ 2.3

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{2.3}$$

จากทั้งสองสมการที่กล่าวมาจะเห็นว่ามีค่า P(A ⋂ B) ที่เหมือนกันอยู่ ดังนั้นเราสามารถเขียนสมการของ P(A ⋂ B)ได้เป็นสมการ 2.4

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A) \tag{2.4}$$

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

จากสมการที่กล่าวมานี้เรียกว่า Bayes Theorem หรือทฤษฎีของเบย์เมื่อนำมาใช้ในการจำแนกประเภทข้อมูลในการทำเหมืองข้อมูล มักจะเปลี่ยนสัญลักษณ์ B เป็น *C*โดยให้ A คือ แอตทริบิวต์ (Attribute*)* และ C คือ คลาส *(*Class*)* ดังสมการ 2.5

$$P(C|A) = \frac{P(A|C) \times P(C)}{P(A)} \tag{2.5}$$

จากสมการ Bayes อธิบายได้ว่าถ้าต้องการทำนายคลาส C เมื่อทราบแอตทริบิวต์ A แล้วสามารถคำนวณได้จากความน่าจะเป็นของแอตทริบิวต์ A ที่มีคลาส C ในการเทรนนิ่งข้อมูลและค่าความน่าจะเป็นของแอตทริบิวต์ A และคลาส C เพื่อให้สามารถเข้าใจได้ง่ายจึงแบ่งสมการของ Bayes ได้ดังนี้

โดย Posterior probability หรือ P(C|A) คือ ค่าความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น A จะมีคลาส C

Likelihood หรือ P(A|C) คือ ค่าความน่าจะเป็นที่เทรนนิ่งข้อมูล (Training data) มีคลาส C และมีแอตทริบิวต์ A โดยที่ A = a_1 ⋂ a_2 ... ⋂ am โดยที่ m คือจำนวนแอตทริบิวต์ในเทรนนิ่งข้อมูล (Training data)

P(C) คือ ค่าความน่าจะเป็นของคลาส C

P(A) คือ ความน่าจะเป็นของคลาส A

แต่การที่แอตทริบิวต์ A = $a_1 \cap a_2 \ldots \cap$ an ที่เกิดขึ้นใน training data อาจจะมีจำนวนน้อยมากหรือไม่มีรูปแบบของแอตทริบิวต์แบบนี้เกิดขึ้นเลย ดังนั้นจึงได้ใช้หลักการที่ว่าแต่ละแอตทริบิวต์เป็นอิสระ (Independent) ต่อกัน ทำให้สามารถเปลี่ยนสมการ P(A|C) ได้เป็นสมการ 2.6

$$P(A|C) = P(a_1|C) \times P(a_2|C) \times \ldots \times P(a_M|C) \times P(C) \quad (2.6)$$

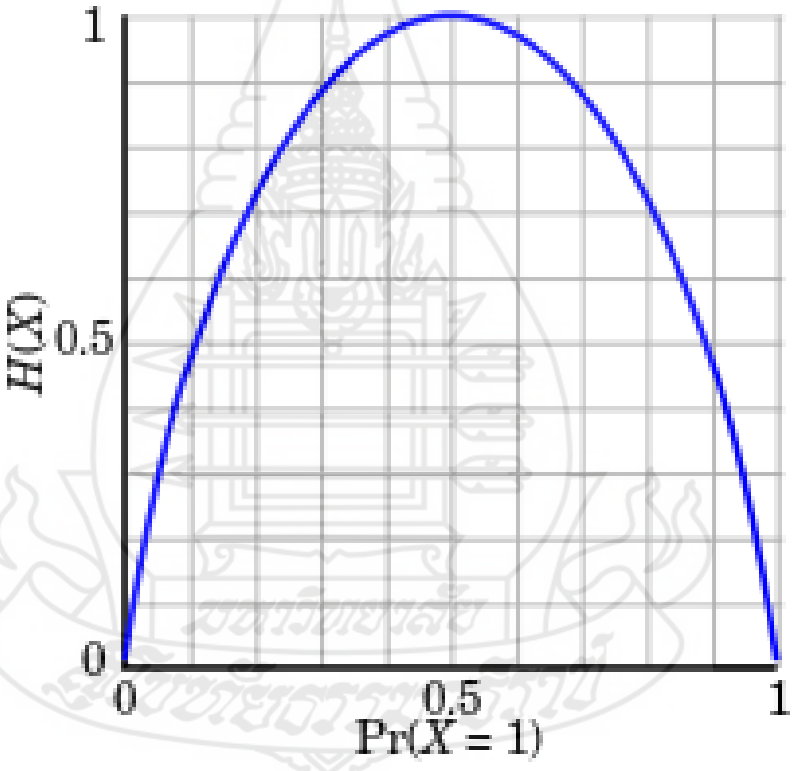
ผู้วิจัยสรุปได้ว่า การจำแนกประเภทข้อมูล (Classification) เป็นการจำแนกคุณลักษณะของข้อมูลโดยภายหลังจากผ่านกระบวนการคัดเลือกคุณลักษณะ (Feature Selection) แล้วนำมาเข้าสู่กระบวนการโดยใช้เทคนิคด้านการทำเหมืองข้อมูล (Data Mining) ตัวอย่างเช่น เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าสุ่ม (Random Forest) หรือวิธีนาอีฟเบย์ (Naïve Bayes) เพื่อหาผลลัพธ์ในการแยกกลุ่มข้อมูลที่มีความสนใจและนำผลลัพธ์ที่ได้จากการใช้เทคนิคที่ดีที่สุดไปใช้ประโยชน์ต่อไป

4. เกณฑ์ที่ใช้ในการเลือกคุณลักษณะ

การพิจารณาคุณลักษณะที่เหมาะสมเพื่อนำมาใช้ในการแบ่งข้อมูลออกเป็นชุดย่อยเพื่อให้ข้อมูลในแต่ละชุดนั้นมีความถูกต้องและบริสุทธ์ของข้อมูลมากที่สุด หมายถึง มีการปะปนกันของข้อมูลในคลาสใด ๆ น้อยที่สุด ซึ่งหมายถึงแอตทริบิวต์นั้นจะทำหน้าที่เป็นโหนดสำหรับการตัดสินใจที่จะส่งผลข้อมูลที่ถูกแบ่งออกมาเป็นข้อมูลที่อยู่ในคลาสเดียวกัน ซึ่งเกณฑ์ที่ใช้ในการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับการแบ่งข้อมูลนี้จะเรียกว่ากฎการแบ่ง (Splitting Rule) และแอตทริบิวต์ที่ถูกเลือกเพื่อทำหน้าที่เป็นโหนดสำหรับการตัดสินใจจะเรียกแอตทริบิวต์นี้ว่าแอตทริบิวต์แบ่ง (Splitting Attribute) ในปัจจุบันเกณฑ์ที่ใช้ในการเลือกแอตทริบิวต์ที่นิยมใช้กันอย่างแพร่หลายประกอบด้วย 1) เกนความรู้ (Information Gain) 2) อัตราส่วนเกน (Gain Ratio) และ 3) ดัชนีจีนี (Gini Index)

ค่าเกนความรู้ (Information Gain) (โกเมศ อัมพวัน, 2548) ค่าเกนความรู้เป็นตัวชี้วัดการแบ่งข้อมูลออกเป็นชุดข้อมูลย่อยที่ได้รับความนิยมอย่างแพร่หลายโดยจะถูกประยุกต์ใช้ในอัลกอริธึม ID3 ซึ่งจะทำการเลือกแอตทริบิวต์สำหรับแบ่งข้อมูลที่มีค่าเกนความรู้สูงที่สุด การเลือกแอตทริบิวต์ที่ต้องการข้อมูลที่น้อยที่สุด เกณฑ์ในการพิจารณาแอตทริบิวต์ตัวแปรที่ใช้ในแบ่งข้อมูลออกเป็นชุดย่อย ๆ เพื่อทำการสร้างต้นไม้ตัดสินใจ โดยแอตทริบิวต์ที่มีค่าเกนความรู้สูงสุดจะถูกเลือกเป็นแอตทริบิวต์สำหรับการแบ่งข้อมูล เนื่องจากแอตทริบิวต์นี้เป็นแอตทริบิวต์ที่ต้องการข้อมูลน้อยที่สุดในการแบ่งข้อมูลออกเป็นชุดย่อย ๆ นั่นคือเป็นแอตทริบิวต์ที่เมื่อใช้ในการแบ่งข้อมูลแล้ว ข้อมูลในแต่ละชุดนั้นจะมีการปะปนกันของข้อมูล (impurity) น้อยที่สุด

การคำนวณหาค่าเกนความรู้ (Information Gain) คือการวัดค่าหาค่าเอ็นโทรปี (Entropy) ก่อนที่จะมีการแบ่งข้อมูลออกตามคลาส (Class) และหลังจากการแบ่งมีประสิทธิภาพดีขึ้นหรือไม่ หากมีประสิทธิภาพดีขึ้นค่า Information Gain จะมีค่าสูง เริ่มจากการหาค่าเอ็นโทรปี (Entropy) ซึ่งเป็นค่าที่ใช้ในการพิจารณาความไม่มีแบบแผน (Randomness) ในข้อมูล โดยหากเอ็นโทรปีมีค่าต่ำจะหมายถึง ข้อมูลนั้นมีค่าปะปนกันหลายค่าทำให้ยากต่อการหาข้อสรุป และหากมีความแตกต่างกันมากค่าเอ็นโทรปี(Entropy)จะมีค่าสูง ดังแสดงตัวอย่างในภาพที่ 2.6 โดยแสดงข้อมูลการสุ่มด้านของเหรียญ ค่าเอ็นโทรปีจะเป็น 0 หากมีการสุ่มได้ด้านของเหรียญเป็นด้านเดียวกันทุกครั้ง ในขณะที่เอ็นโทรปีจะมีค่าเท่ากับ 1 เมื่อการสุ่มได้ด้านของเหรียญได้จำนวนครั้งเท่ากันทั้งสองด้าน ซึ่งหมายถึงข้อมูลมีการปะปนกันจนไม่สามารถหาข้อสรุปที่เหมาะสมจากข้อมูลได้ แต่เนื่องจากไม่สามารถคาดการณ์ได้ว่าผลลัพธ์ความแน่นอนของการสุ่มเหรียญจะมีผลลัพธ์ออกมาเป็นหัวและก้อยในจำนวนที่เท่ากัน ซึ่งค่าที่ได้หาความแน่นอนในผลที่ได้น้อยจะทำให้ค่าเอ็นโทรปีน้อยลง (Services,2015)



ภาพที่ 2.8 เอ็นโทรปีของการสุ่มด้านของเหรียญ

ที่มา : Matthew N. Bernstein (2020)

จากภาพที่ 2.8 แสดงค่าเอนโทรปี H(X) แสดงค่าความน่าจะเป็นหรือการคาดไว้ว่าจะได้ผลลัพธ์ออกมาเป็นหัว Pr(X=1) คือผลลัพธ์ที่แสดงออกมาโดย (X= 1) หมายถึงเหรียญด้านหัว ซึ่งในการพลิกเหรียญความเป็นไปได้ของผลลัพธ์จะมี 2 ค่า (ความน่าจะเป็นเท่ากันคือ ½) ผลลัพธ์ในการ

โยนเหรียญจะมีค่าข้อมูลเต็มคือ 1 ทุกครั้งที่มีการโยนเหรียญด้านหนึ่งมีแนวโน้มที่จะได้ผลลัพธ์อีกด้านหนึ่ง ความไม่แน่นอนที่จะเกิดขึ้นถูกวัดด้วยปริมาณในเอนโทรปี ผลลัพธ์ที่แสดงออกมา ซึ่งสามารถอธิบายถึงโดยสรุปคือ เมื่อเรามีการทำนายเหรียญว่าจะออกมาในอัตราที่เท่ากันแต่เมื่อโยนเหรียญแล้วผลลัพธ์ถ้าออกมาตามที่เราคาดหมายหรือกำหนด แสดงว่าค่าเอนโทรปีสูง คือความเป็นไปได้สูง แต่หากผลลัพธ์จากการสุ่มออกมาแล้วผลลัพธ์ที่ได้ไม่เป็นไปตามที่คาดหวังไว้ หมายถึงค่าเอนโทรปีมีค่าต่ำหรือลดน้อยลง ทำให้ความแม่นยำที่ได้ออกมาก็ไม่แม่นยำด้วยเช่นกัน

ค่า Information Gain สามารถแปลความหมายได้คือ ผลต่างระหว่างค่าเอ็นโทรปี (Entropy) ในสถานะปัจจุบันและค่าเอ็นโทรปี (Entropy) ก่อนหน้า โดยมาหาได้จากสมการ 2.7

$$Information\ Gain(IG) = Entropy(before) - Entropy(after) \qquad (2.7)$$

โดย Information Gain(IG) คือค่าเกนความรู้
Entropy(before) คือค่าเอ็นโทรปีก่อนหน้า
Entropy(after) คือค่าเอ็นโทรปีตัวหลัง

ซึ่งจะอธิบายความหมายของค่าเกนความรู้ (Information Gain) ได้ดังนี้

Information Gain>0 คือ ค่าเกนความรู้สูง การปะปนน้อยทำให้จำแนกประเภทได้ง่าย

Information Gain<0 คือ ค่าเกนความรู้ต่ำ การปะปนมากทำให้จำแนกประเภทได้ยาก

Information Gain=0 คือ ค่าเกนความรู้ ไม่มีการเปลี่ยนแปง (ปัจจัยที่นำมาไม่มีผลต่อการจำแนกประเภท)

ผู้วิจัยสรุปได้ว่า การพิจารณาคุณลักษณะเป็นการแยกแอตทริบิวต์ที่ปะปนกันอยู่ออกเพื่อทำหน้าที่เป็นโหนดแม่ในการตัดสินใจกลุ่มของชุดข้อมูล โดยคัดเลือกค่าน้ำหนักที่ดีที่สุดและแอตทริบิวต์นั้นจะทำหน้าที่เป็นโหนดสำหรับการตัดสินใจที่จะส่งผลข้อมูลที่ถูกแบ่งออกมาเป็นข้อมูลที่อยู่ในคลาสเดียวกัน เมื่อนำมาใช้กับเทคนิคต้นไม้ตัดสินใจโดยใช้ค่าเกนความรู้ (Information Gain) ค่าที่มีน้ำหนักมากจะถูกนำมาใช้เป็นโหนดแม่ในโมเดลการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ โดยไม่จำเป็นต้องใช้ทุกแอตทริบิวต์ในการแบ่งกลุ่มข้อมูลเพราะถ้าเลือกแอตทริบิวต์ที่มีค่าเกนความรู้ (Information Gain) สูง ๆ หลายค่าก็อาจจะเพียงพอต่อการแบ่งกลุ่มข้อมูลทั้งหมด ทำให้สามารถจำแนกข้อมูลได้เร็วขึ้นจึงเป็นเหตุผลที่ได้รับความนิยมในการใช้งาน

5. การประเมินผลและการวัดประสิทธิภาพแบบจำลอง

การประเมินตัวแบบการจำแนกข้อมูลถือเป็นขั้นตอนสำคัญในกระบวนการทำเหมืองข้อมูลเพื่อให้ทราบประสิทธิภาพของตัวแบบที่สร้างขึ้นรวมถึงสามารถเปรียบเทียบตัวแบบหลายตัวเพื่อเลือกตัวแบบที่ดีที่สุดในสำหรับนำไปใช้งานด้วยตาราง Confusion Matrix ซึ่งใช้ตัวชี้วัดต่อไปนี้ (สุรพงค์ เอื้อวัฒนามงคล, 2559, หน้า 70)

1) ค่าความถูกต้อง (Accuracy) คือ จำนวนที่ทำนายความถูกต้องแม่นยำในการจำแนกข้อมูล โดยพิจารณาค่ารวมการทำนายทั้งหมดของข้อมูล ได้แก่ สัดส่วนระหว่างจำนวนข้อมูลทั้งหมดที่จำแนกประเภทถูกต้องทั้งประเภท Positive และ Negative กับจำนวนข้อมูลทั้งหมดที่ถูกจำแนกประเภท โดยสามารถคำนวณได้จากสมการ 2.8

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

เมื่อ True Positive คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสที่กำลังสนใจอยู่
False Positive คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสที่กำลังสนใจอยู่

ตารางที่ 2.1 ตารางแสดงตัวอย่าง Confusion Matrix ขนาด 2 X 2 ในการทำนายผลลัพธ์

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	**TP**	**FP**
Predicted Negative (0)	**FN**	**TN**

จากตารางที่ 2.1 True Positive คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสที่กำลังสนใจอยู่ และ False Positive คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสที่กำลังสนใจอยู่ โดยแสดงรายละเอียดและอธิบายความหมายในตารางได้ดังต่อไปนี้

True Positive (TP) คือ ความเป็นจริง True แล้วทำนาย True ให้ผลเป็นจริง
True Negative (TN) คือ ความเป็นจริง No แล้วทำนายว่า No ให้ผลเป็นจริง
False Positive (FP) คือ ความเป็นจริง Yes แล้วทำนายว่า No ให้ผลเป็นเท็จ
False Negative (FN) คือ ความเป็นจริง No แล้วทำนายว่า Yes ให้ผลเป็นเท็จ

2) ค่าความแม่นยำ (Precision) คือ ค่าความถูกต้องของการจำแนกประเภทข้อมูลว่าเป็น Positive หรือดูค่าความถูกต้องจากการทำนายออกมาแล้วมีผลการทำนายถูกต้อง โดยพิจารณาจากผลการทำนายเป็นหลัก ได้แก่ สัดส่วนที่ตัวแบบจำแนกประเภทข้อมูลแบบ Positive ได้ถูกต้องเทียบกับจำนวนข้อมูลทั้งหมดที่ถูกจำแนกประเภทนี้ว่าเป็น Positive โดยสามารถคำนวณได้จากสมการ 2.9

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

3) ค่าความครบถ้วน(Recall) คือ การวัดค่าความถูกต้องของโมเดลโดยพิจารณาแยกทีละคลาส และมีความครบถ้วนโดยอัตราส่วนจากความถูกต้องของการค้นพบข้อมูลเทียบจากผลการเฉลยเป็นหลัก โดยสามารถคำนวณได้จากสมการ 2.10

$$Recall = \frac{TP}{TP + FN} \tag{2.10}$$

4) ค่าความถ่วงดุล (F-measure) หรือค่าประสิทธิภาพโดยรวม เป็นตัวชี้วัดที่คำนวณจากค่า Precision และ Recall โดยนำทั้งสองค่ามาคำนวณร่วมกัน ดังสมการ 2.11

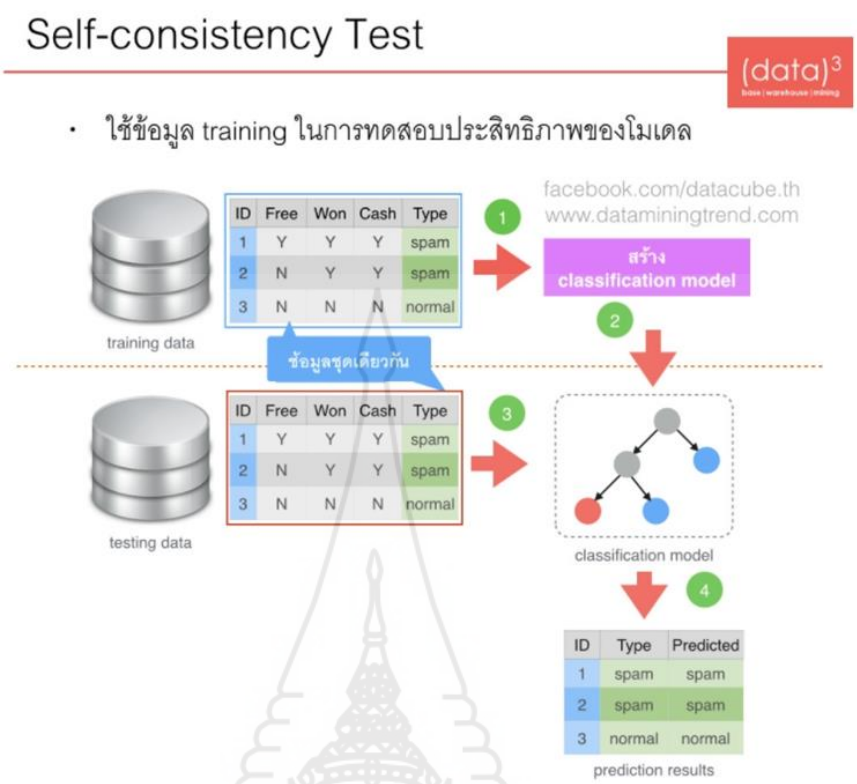
$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.11}$$

เมื่อ Precision คือ ค่าความแม่นยำของโมเดล (จากสมการ 2.9)

Recall คือ ค่าความแม่นยำหรือจำนวนที่ทำนายถูกกี่ตัว เป็นการวัดค่าความครบถ้วนของโมเดล (จากสมการ 2.10)

การแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพของโมเดล (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557) มีการโอยแบ่งเป็น 3 วิธีการใหญ่ ๆ ดังนี้

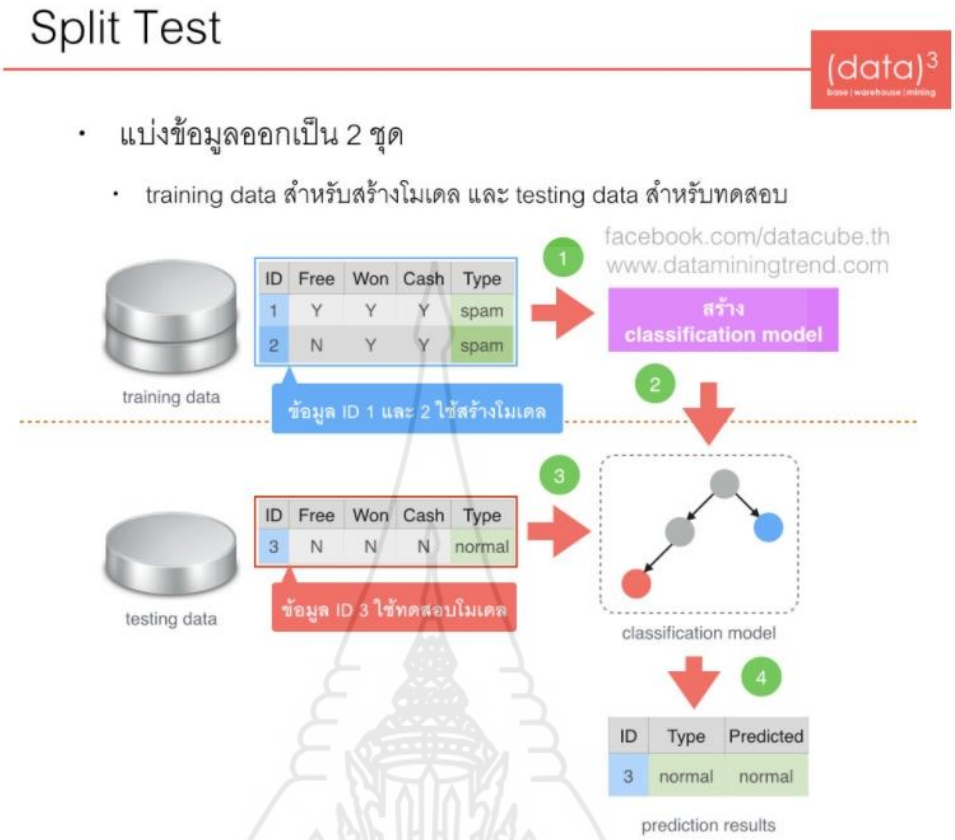
1) *วิธี Self Consistency Test หรือ Use Training Set* นี้เป็นวิธีการที่ง่ายที่สุด คือ ข้อมูลที่ใช้ในการสร้างแบบจำลอง (Model) และข้อมูลที่ใช้ในการทดสอบแบบจำลอง (Model) เป็นข้อมูลชุดเดียวกัน กระบวนการนี้เริ่มจากสร้างโมเดลด้วยข้อมูลชุดสอน (Training Data) จากนั้นนำแบบจำลองที่สร้างได้มาทำนายข้อมูลชุดสอน (Training Data) ชุดข้อมูลเดิม ตัวอย่างเช่น การนำข้อมูลชุดสอน (Training Data) ในตาราง มาสร้างแบบจำลองและทดสอบแบบจำลองเป็นต้น การวัดประสิทธิภาพด้วยวิธีนี้จะให้ผลการวัดประสิทธิภาพที่มีค่าสูงมาก (อาจจะเข้าใกล้ 100%) เนื่องจากเป็นข้อมูลชุดเดิมที่ระบบได้ทำการเรียนรู้มาแล้ว แต่ผลการวัดที่ได้ไม่เหมาะที่จะนำไปรายงานในงานวิจัยต่าง ๆ ซึ่งวิธีการนี้เหมาะสำหรับใช้ในการทดสอบประสิทธิภาพเพื่อดูแนวโน้มของแบบจำลองที่สร้างขึ้น ถ้าได้ผลการวัดที่น้อย แสดงว่าแบบจำลองไม่เหมาะสมกับข้อมูล จึงไม่ควรจะนำไปทดสอบด้วยวิธีการแบ่งข้อมูลแบบต่าง ๆ ดังแสดงในภาพที่ 2.9



ภาพที่ 2.9 วิธี Self Consistency Test

ที่มา : เอกสิทธิ์ พัชรวงศ์ศักดา (2557)

2) *วิธี Split Test* จะแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน ตัวอย่างเช่น 70% ต่อ 30% หรือ 80% ต่อ 20% โดยข้อมูลส่วนที่หนึ่ง (70% หรือ 80%) ใช้ในการสร้างแบบจำลอง (Model) และข้อมูลส่วนที่สอง (30% หรือ 20%) ใช้ใน การทดสอบประสิทธิภาพของแบบจำลอง (Model) ดังแสดงในภาพที่ 2.10



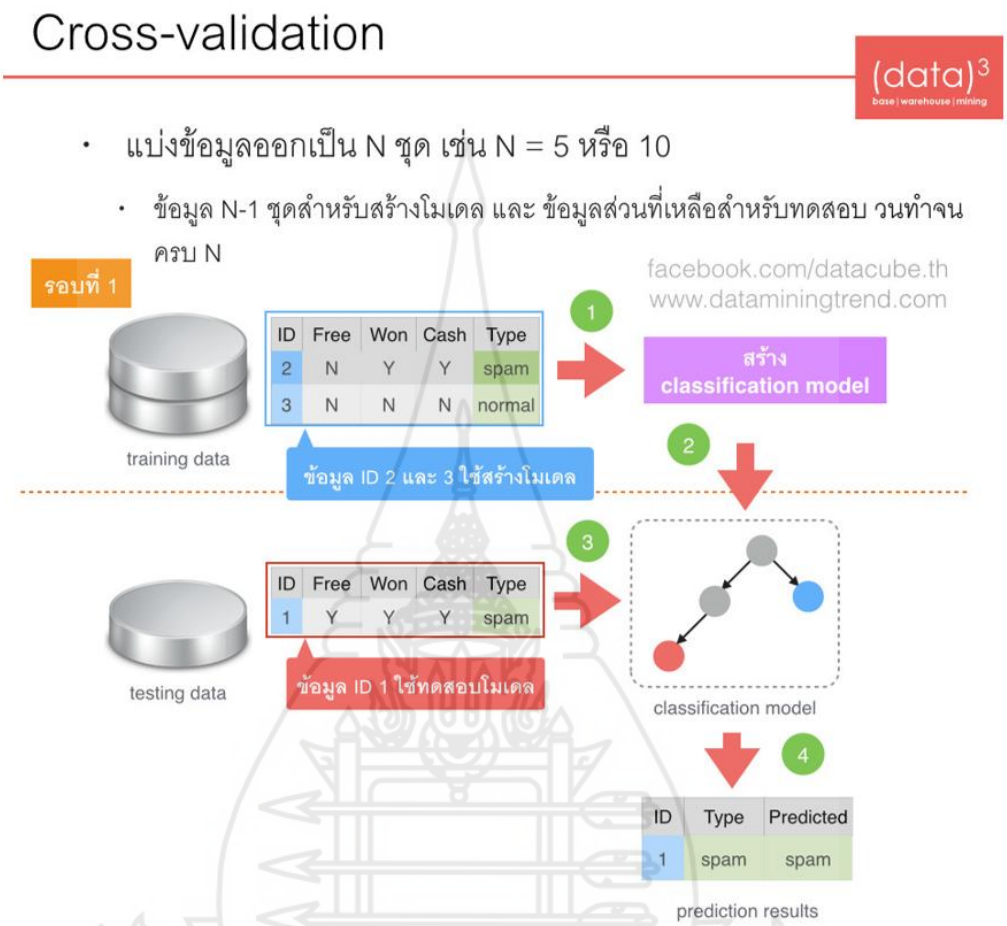
ภาพที่ 2.10 วิธี Split Test

ที่มา : เอกสิทธิ์ พัชรวงศ์ศักดา (2557)

จากภาพที่ 2.10 แสดงวิธี Split Test โดยแบ่งข้อมูลสำหรับการวัดประสิทธิภาพออกเป็น 2 ส่วน สำหรับการสร้างแบบจำลองและข้อมูลสำหรับใช้ในการทดสอบประสิทธิภาพของแบบจำลอง แต่การทดสอบแบบ Split Test นี้จะสุ่มข้อมูลเพียงครั้งเดียวซึ่งในบางครั้งถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะคล้ายกับข้อมูลที่ใช้สร้างแบบจำลองอาจทำให้ผลการวัดประสิทธิภาพที่ได้ออกมาดี และในทางตรงข้ามถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะแตกต่างกับข้อมูลที่ใช้สร้างแบบจำลองมากอาจจะทำให้ผลการวัดประสิทธิภาพได้ออกมาแย่ ดังนั้นจึงควรใช้วิธี Split Test นี้หรือทำการสุ่ม หลาย ๆ ครั้ง แต่ข้อดีของวิธีการนี้คือใช้เวลาในการสร้างโมเดลน้อยซึ่งเหมาะกับชุดข้อมูลที่มีขนาดใหญ่มาก

3) วิธี Cross – Validation Test วิธีการวัดนี้ถือเป็นวิธีที่นิยมในการทำงานวิจัย เพื่อใช้ในการทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัด ประสิทธิภาพด้วยวิธี Cross-validation นี้ทำได้โดยจะแบ่งข้อมูลออกเป็นหลายส่วน (มักจะแสดงด้วยค่า k) เช่น 5-fold cross-validation คือนำข้อมูลแบ่งออกเป็น 5 ส่วน โดยแต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน

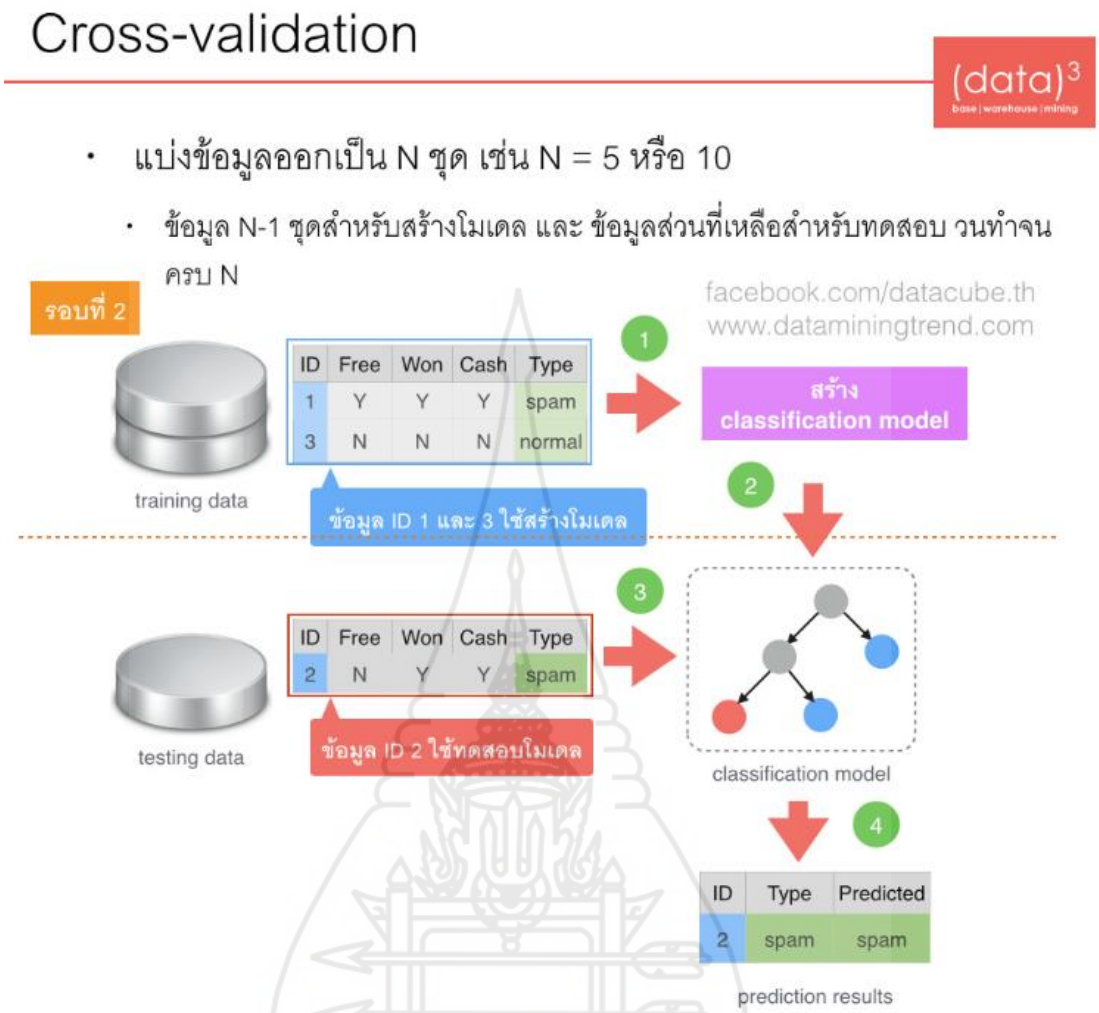
เมื่อแบ่งข้อมูลแล้วจะนำข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบตามจำนวนที่แบ่งไว้เช่นเดียวกันกับการทดสอบด้วยวิธี 5-fold cross-validation



ภาพที่ 2.11 การแบ่งข้อมูลแบบ 5-fold Cross-Validation (รอบที่ 1)

ที่มา : เอกสิทธิ์ พัชรวงศ์ศักดา (2557)

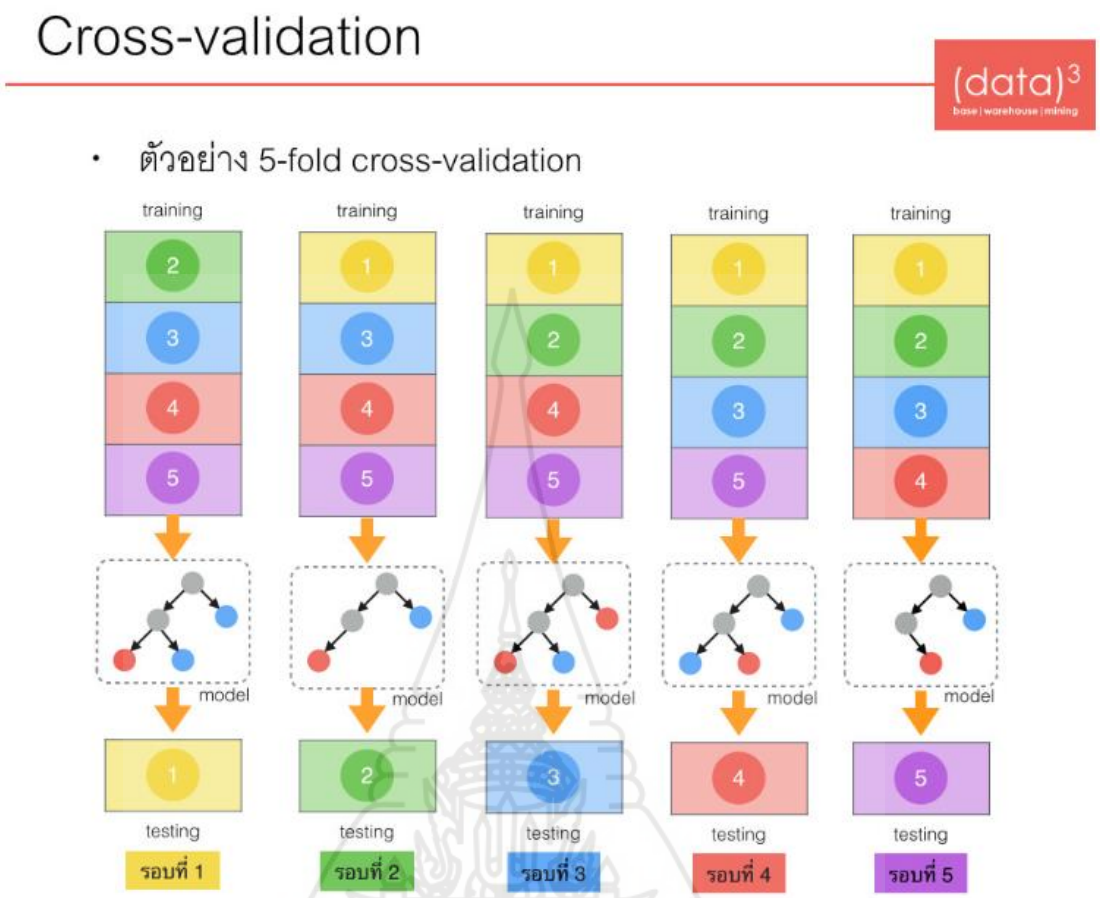
จากภาพที่ 2.11 แสดงการแบ่งข้อมูลแบบ 5-fold Cross-Validation ซึ่งแบ่งข้อมูลออกเป็น 5 ส่วนมีจำนวนเท่ากัน โดยรอบที่ 1 ใช้ข้อมูลส่วนที่ 2, 3, 4 และ 5 สร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 1 เพื่อทำการทดสอบประสิทธิภาพของโมเดล



ภาพที่ 2.12 การแบ่งข้อมูลแบบ 5-fold Cross-Validation (รอบที่ 2)

ที่มา : เอกสิทธิ์ พัชรวงศ์ศักดา (2557)

จากภาพที่ 2.12 แสดงการทำงานรอบที่ 2 ใช้ข้อมูลส่วนที่ 1, 3, 4 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 2 เพื่อทำการทดสอบประสิทธิภาพของโมเดล



ภาพที่ 2.13 วิธี 5-fold Cross-Validation

ที่มา : เอกสิทธิ์ พัชรวงศ์ศักดา (2557)

จากภาพที่ 2.13 แสดงให้เห็นการแบ่งข้อมูลทั้ง 5 ส่วนแต่ละส่วนมีจำนวนข้อมูลที่เท่ากัน โดยกระบวนการทดสอบประสิทธิภาพของโมเดลด้วยวิธี 5-fold Cross-validation สามารถอธิบายตามภาพได้ดังนี้

รอบที่ 1 ใช้ข้อมูลที่แบ่งไว้ในส่วนที่ 2, 3, 4 และ 5 จากนั้นสร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 1 เพื่อทำการทดสอบประสิทธิภาพ

รอบที่ 2 ใช้ข้อมูลที่แบ่งไว้ในส่วนที่ 1, 3, 4 และ 5 จากนั้นสร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 2 เพื่อทำการทดสอบประสิทธิภาพ

รอบที่ 3 ใช้ข้อมูลที่แบ่งไว้ในส่วนที่ 1, 2, 4 และ 5 จากนั้นสร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 3 เพื่อทำการทดสอบประสิทธิภาพ

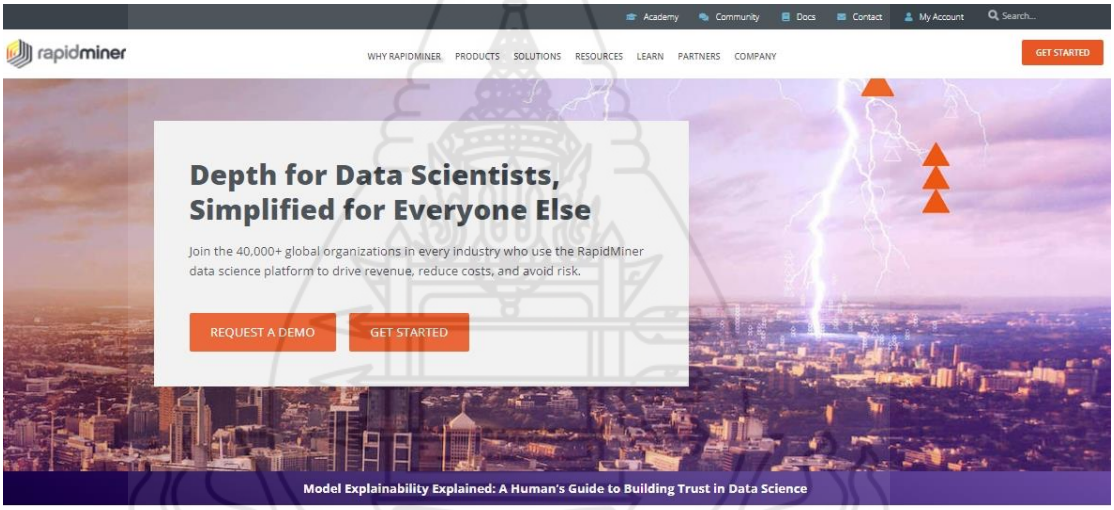
รอบที่ 4 ใช้ข้อมูลที่แบ่งไว้ในส่วนที่ 1, 2, 3 และ 5 จากนั้นสร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 4 เพื่อทำการทดสอบประสิทธิภาพ

รอบที่ 5 ใช้ข้อมูลที่แบ่งไว้ในส่วนที่ 1, 2, 3 และ 4 จากนั้นสร้างโมเดลและใช้โมเดลทำนายข้อมูลส่วนที่ 5 เพื่อทำการทดสอบประสิทธิภาพ

ผู้วิจัยสรุปได้ว่า การประเมินตัวแบบการจำแนกข้อมูลถือเป็นขั้นตอนสำคัญในกระบวนการทำเหมืองข้อมูลเพื่อให้ทราบประสิทธิภาพของตัวแบบที่สร้างขึ้น การเปรียบเทียบสามารถเปรียบเทียบความถูกต้องของตัวแบบหลายตัวเพื่อให้ได้ตัวแบบที่ดีที่สุดสำหรับและเหมาะสมกับชุดข้อมูลที่จะนำไปใช้งาน

6. โปรแกรม RapidMiner Studio

RapidMiner Studio เริ่มพัฒนาขึ้นจากบริษัท Rapid-I ในประเทศเยอรมนีภายหลัง เมื่อช่วงปลายปี ค.ศ. 2013 (พ.ศ.2556) ได้รับทุนจากนักลงทุนในประเทศสหรัฐอเมริกาจึงเปลี่ยนชื่อบริษัท จาก Rapid-I เป็น RapidMiner แทนชื่อเดิมและย้ายสำนักงานใหญ่มาอยู่ประเทศสหรัฐอเมริกา โดยสามารถเข้าไปดาวน์โหลดโปรแกรมและศึกษาข้อมูลเพิ่มเติมผ่านทางเว็บไซต์ https://rapidminer.com/ ดังภาพที่ 2.14



ภาพที่ 2.14 หน้าเว็บไซต์ https://rapidminer.com/

โปรแกรม RapidMiner Studio ถือเป็นเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล (Data Analytics) สร้างโมเดลคณิตศาสตร์ในงานเชิงวิศวกรรมและงานด้านธุรกิจ มีความสะดวกในการใช้งานผู้ใช้งานสามารถออกแบบผ่าน Graphical User Interface (GUI) ทำให้การใช้งานมีความสะดวกและง่ายต่อการใช้งาน

เอกสิทธิ์ พัชรวงศ์ศักดา (2557) กล่าวว่า ในยุคปัจจุบันได้ก้าวเข้าไปสู่ยุคที่เรียกว่า Big Data หรือข้อมูลมหาศาล เนื่องจากในแต่ละวันมีข้อมูลเกิดขึ้นมากมาย เช่น ข้อมูลสมาชิกของ Facebook ข้อมูลการซื้อสินค้าในซุปเปอร์มาร์เกตต่าง ๆ และเพื่อให้เกิดประโยชน์มากที่สุดจำเป็นต้องนำ ข้อมูลมหาศาลเหล่านี้มาทำการวิเคราะห์ (Analyze) ซึ่งเทคนิคหนึ่งที่ได้รับการนิยมอย่างสูงในปัจจุบัน คือ เทคนิค Data Mining ซึ่งเป็นเทคนิคที่ค้นหาความสัมพันธ์ในข้อมูล เช่น ถ้าลูกค้าซื้อเบียร์แล้วลูกค้าจะซื้อผ้าอ้อมร่วมไปด้วยหรือถ้าเรากด Like หน้า Facebook Page เราจะ

เห็นว่า Facebook มีระบบแนะนำ Page อื่น ๆ ที่เกี่ยวข้องมาให้ด้วย หรือ การสร้างโมเดลเพื่อทำนายสิ่งที่จะเกิดขึ้นในอนาคต เช่น ทำนายยอดขายในไตรมาสถัดไป หรือ การทำนายว่าพนักงานคนไหนที่จะลาออกจากบริษัทในช่วง 3 เดือนข้างหน้า ตัวอย่างเหล่านี้ล้วนเป็นผลมาจากการวิเคราะห์ข้อมูลทางด้าน Data Mining

การวิเคราะห์ข้อมูลด้วย Data Mining โดยใช้ซอฟต์แวร์ที่ช่วยให้ทำการวิเคราะห์ได้ง่ายขึ้นแต่ซอฟต์แวร์ส่วนใหญ่จะเป็นซอฟต์แวร์เชิงพาณิชย์ (Commercial Software) เช่น SAS Enterprise Miner หรือ IBM Intelligent Miner การลงทุนซื้อซอฟต์แวร์เชิงธุรกิจเหล่านี้มาใช้งานอาจจะไม่คุ้มค่าในการลงทุนสำหรับผู้ประกอบการวิสาหกิจขนาดกลางและขนาดย่อม (SMEs) หรืออาจารย์ นักวิจัยและนักศึกษาในมหาวิทยาลัยต่าง ๆ ดังนั้นวิธีการหนึ่งที่จะทำให้สามารถวิเคราะห์ข้อมูลเหล่านี้ได้คือ การใช้ซอฟต์แวร์เวอร์ชันฟรี (Free Version) ที่สามารถดาวน์โหลด (Download) มาใช้งานได้โดยไม่เสียค่าใช้จ่าย เช่น RapidMiner Studio Educational

ในการวิจัยนี้ผู้วิจัยเลือกใช้โปรแกรม RapidMiner Studio โดยเลือกใช้ Educational License Program สำหรับภาคการศึกษา ทำให้สามารถใช้งานระบบการทำงานหรือฟังก์ชัน (Function) ของโปรแกรมได้ครบถ้วน ทั้งในส่วนของการใช้บริการข้อมูลบนเซิร์ฟเวอร์ของโปรแกรมเพื่อเก็บข้อมูลและประมวลผลผ่านระบบออนไลน์ได้ รองรับการนำข้อมูลเข้าประมวลผลในระบบมากกว่า 10,000 รายการ ภายหลังจากการใช้งานทำให้การวิจัยมีความสะดวกและมีประสิทธิภาพในการดำเนินงานยิ่งขึ้น

7. งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องผู้วิจัยได้ศึกษาและค้นคว้างานวิจัยต่าง ๆ ดังนี้

7.1 งานวิจัยในประเทศ

7.1.1 ภุมริน หรั่งน้อย คุณัญญา ส้มเกลี้ยง ปิยนันท์ เทียบศรไชย และประภาส ทองรัก (2564) ศึกษาการประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อแนะนำอาชีพด้านไอทีสำหรับนักศึกษาระดับปริญญาตรี กรณีศึกษามหาวิทยาลัยเทคโนโลยีราชมงคล จากข้อมูลภาวะการมีงานทำของผู้สำเร็จการศึกษาปริญญาตรี หลักสูตรด้านคอมพิวเตอร์ โดยเลือกข้อมูลที่เกี่ยวข้องเพื่อวิเคราะห์ความสัมพันธ์ต่อกันจำนวน 16 ลักษณะ วิเคราะห์ด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคการเรียนรู้เชิงลึก (Deep Learning) และเทคนิคนาอีฟเบย์ (Naïve Bayes) ผลการศึกษาพบว่าเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีความถูกต้อง 84.39% ค่าความคลาดเคลื่อน 15.61% และค่าความเที่ยงตรง 100%

7.1.2 รัชฎา เทพประสิทธิ์ และจรัญ แสนราช (2563) ศึกษาปัจจัยที่มีผลต่อการเลือกเข้าศึกษาต่อในสาขาวิชา คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงราย จากข้อมูลพื้นฐานนักศึกษาของงานส่งเสริมวิชาการและงานทะเบียน ในช่วงปีการศึกษา 2556 – 2560 จำนวนนักศึกษา 3,867 คน ด้วยกฎการจำแนกเทคนิคต้นไม้ตัดสินใจ ผลการศึกษาพบว่าปัจจัยที่มีผลต่อการเลือกสาขาวิชา คณะครุศาสตร์ 15 สาขาวิชา ได้แก่ แผนการเรียนก่อนเข้าศึกษาและเพศ จากตัวแปรทั้งหมด 9 ตัวแปรและการประเมินวัดประสิทธิภาพของโมเดลวัดความแม่นยำได้ค่า 72.5%

7.1.3 สำราญ วานนท์ และรจนา เมืองแสน (2563) ศึกษาและพัฒนาตัวแบบพยากรณ์คุณลักษณะความเหมาะสมสำหรับการเลือกสมัครสาขาเรียนโดยใช้เทคนิคเหมืองข้อมูล โดยใช้ข้อมูลการรับสมัครของนักศึกษามหาวิทยาลัยราชภัฏชัยภูมิ ด้วยโปรแกรม WEKA จำแนกกลุ่มข้อมูลด้วยเทคนิค Simple K-mean และหากฎความสัมพันธ์ของข้อมูลด้วยอะพริออลิอัลกอริธึม จากนั้นสร้างตัวแบบพยากรณ์ด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิค Random Forest เปรียบเทียบกฎความสัมพันธ์จากข้อมูลทั้งหมด 6 กลุ่ม ผลการศึกษาการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์ 2 เทคนิคคือเทคนิคต้นไม้ตัดสินใจ (Decision Tree) และ เทคนิค Random Forest โดยเทคนิค Random Forest ค่าความถูกต้องสูงสุดได้ 74.67% ถือเป็นเทคนิคที่สร้างตัวแบบที่ได้รับการยอมรับและค่าความถูกต้องมากที่สุด

7.1.4 อัจจิมา มณฑาพันธ์ (2562) ศึกษาการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม จากข้อมูลคนที่มีชีวิตจำนวน 569 คน โดยมีคุณลักษณะทางกายภาพ 30 ลักษณะ ทำนายค่าความถูกต้องด้วยเทคนิค Support Vector Machine ผลการศึกษาพบว่า เทคนิค Correlation Based Feature Selection คัดเลือกคุณลักษณะที่สำคัญ 3 คุณลักษณะและนำมาพยากรณ์วัดค่าความถูกต้องได้ 91.22% เทคนิค Information Gain คัดเลือกคุณลักษณะที่สำคัญ 6 คุณลักษณะและนำมาพยากรณ์วัดค่าความถูกต้องได้ 92.27% และเทคนิค Gain Ration คัดเลือกคุณลักษณะที่สำคัญ 6 คุณลักษณะและนำมาพยากรณ์วัดค่าความถูกต้องได้ 92.27%

7.1.5 สำราญ วานนท์ ธรัช อารีราษฎร์ และจรัญ แสนราช (2561) ศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาปริญญาตรีสาขาคอมพิวเตอร์โดยใช้เทคนิคเหมืองข้อมูล นำข้อมูลภาวการณ์มีงานทำของบัณฑิตและข้อมูลประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษาจำนวน 65,335 ระเบียน ในสาขาทางด้านคอมพิวเตอร์โดยมีคุณลักษณะประกอบด้วยผลการเรียน ความสามารถพิเศษ อาชีพของบิดามารดา รายได้ของบิดามารดาและข้อมูลพื้นฐานทดลองวัดความแม่นยำด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจ วิธีแรนดอมฟอร์เรสและเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิง (Bagging) ผลการศึกษาพบว่าความแม่นยำในการจำแนกประเภทข้อมูลด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) เท่ากับ 81.91% วิธีแรนดอมฟอร์เรส (Random Forest) เท่ากับ 84.29% และเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิง (Bagging) เท่ากับ 81.71% โดยเทคนิคที่ให้ความแม่นยำมากที่สุดคือวิธีแรนดอมฟอร์เรส (Random Forest)

7.1.6 ธนวรรณ วิสุทธิรัตน์ (2561) ศึกษาการใช้เหมืองข้อมูล (Data Mining) ในการวิเคราะห์และสร้างตัวแบบความสัมพันธ์ระหว่างคุณภาพสังคมและความสุขของประชาชนในจังหวัดจันทบุรี โดยการกำหนดกลุ่มตัวอย่างเพื่อดำเนินการวิจัยด้วยวิธีสุ่มกลุ่มตัวอย่างแบบหลายขั้นตอน การสุ่มตัวอย่างแบบเจาะจง (Purposive Sampling) การสุ่มตัวอย่างอย่างง่ายดาย (Sample Random Sampling) การเลือกตัวอย่างแบบชั้นภูมิ (Stratified Random Sampling) วิเคราะห์ตามกระบวนการ Cross Industry Standard Process for Data Mining (CRISP-DM) ได้ผลความเชื่อมั่นแบบสัมประสิทธิ์แอลฟาอยู่ระหว่าง 0.607-0.972 โดยใช้เทคนิคโครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ การวิเคราะห์การถดถอยแบบขั้นตอนและการจัดกลุ่มแบบ K-means

7.1.7 ศุภามณ จันทร์สกุล (2561) ศึกษาเทคนิคเหมืองข้อมูลในการวิเคราะห์ข้อมูลทางการพยาบาล โดยการแบ่งกลุ่มข้อมูลสำหรับการวิเคราะห์ออกเป็น 2 กลุ่มใหญ่คือ การวิเคราะห์เหมืองข้อมูลเพื่อทำนายมีการเรียนรู้แบบมีการสอนและการวิเคราะห์เหมืองข้อมูลเพื่อการอธิบายมีการเรียนรู้แบบไม่มีการสอนด้วยเทคนิคที่ได้รับความนิยมได้แก่ การจำแนกประเภทต้นไม้ตัดสินใจและการจำแนกประเภทเครือข่ายใยประสาท การจัดกลุ่มข้อมูลและการค้นหากฎความสัมพันธ์ของข้อมูล พบว่าการวิเคราะห์เครือข่ายใยประสาทโดยใช้โครงสร้างแบบ Multilayer Perception (MLP) ซึ่งเป็นแบบหลายชั้นโหนด (Node)และการวิเคราะห์แบบต้นไม้ตัดสินใจด้วยวิธี Classification and Regression Trees (CRT) แบ่งข้อมูลออกเป็นส่วนและพิจารณาความคล้ายคลึง การจัดประเภทกฎผู้เรียนมีสไตล์การเรียนแบบ VAR ในรูปแบบ Trimodal หรือ VARK/Multimod ผลการศึกษาพบว่ามีผลการเรียนดีคิดเป็นร้อยละ 57 และถ้าผู้เรียนมีสไตล์การเรียนแบบ VARK ในรูปแบบ Unimodal หรือ Bimodal มีผลสัมฤทธิ์การเรียนไม่ดีคิดเป็นร้อยละ 62.4

7.1.8 สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง (2560) ศึกษาการเปรียบเทียบประสิทธิ์ภาพวิธีจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย จากการนำข้อมูลผู้ป่วยโรคมะเร็งเรื้อรังจำนวน 400 ระเบียน ประกอบด้วยตัวแปรอิสระ 24 คุณลักษณะและตัวแปรตาม 1 คุณลักษณะ) แบ่งข้อมูลเป็นชุดสร้างตัวแบบและชุดทดสอบในอัตราส่วน 70 และ 30 โดยเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มด้วยวิธีทางด้านสถิติ ได้แก่ วิธีความใกล้เคียงกันมากที่สุด (K-nearest Neighbor) วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) วิธี Support Vector Machine วิธีฐานกฎ (Rule-based) วิธีการถดถอยลอจิสติก (Logistic Regression) และวิธีนาอีฟเบย์ (Naïve Bayes) ผลการศึกษาพบว่าการจำแนกประเภทข้อมูลที่ดีที่สุดคือ วิธีต้นไม้ตัดสินใจ (Decision Tree) ซึ่งให้ความถูกต้องคือ 100% และความคาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0059

7.1.9 ธาดา จันตะคุณ (2560) ศึกษาการพยากรณ์ความเป็นไปได้ในการเลือกสมัครสาขาวิชาโดยใช้เทคนิคเหมืองข้อมูล ของนักศึกษาระดับปริญญาตรี คณะเทคโนโลยีสารสนเทศมหาวิทยาลัยราชภัฏมหาสารคาม โดยคัดเลือกข้อมูลจากข้อมูลนักศึกษาที่ได้รับการคัดเลือกแล้วจำนวน 162 คน ใช้คุณลักษณะในการพิจารณาจำนวน 9 คุณลักษณะ เปรียบเทียบตัวแบบการจำแนก 4 เทคนิค คือ Decision Tree, Naïve Bayes, k-NN และ Rule Induction และทดสอบประสิทธิภาพของแบบจำลองด้วยวิธี Cross Validation ผลการศึกษาพบว่า เทคนิค Decision Tree มีค่าความถูกต้องสูงสุดได้ 83.97%

7.1.10 วันวิสาข์ ชนะประเสริฐ (2559) ศึกษาวิเคราะห์เทคนิคการใช้เหมืองข้อมูลเพื่อแนะนำอาชีพสำหรับนักศึกษา ปริญญาตรี คณะโบราณคดี จากข้อมูลบัณฑิตผู้สำเร็จการศึกษาจำนวน 400 คน โดยการศึกษาคัดเลือกคุณลักษณะที่เหมาะสมระหว่างวิธี InfoGainAttributeEval, CfsSubsetEval และการไม่คัดเลือกคุณลักษณะ พบว่าการคัดเลือกคุณลักษณะด้วยวิธี InfoGainAttributeEval มีความเหมาะสมที่สุด เมื่อได้คุณลักษณะที่เหมาะสมแล้วนำมาวัดประสิทธิภาพด้วยวิธี 10-fold Cross Validation และวัดค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) และเปรียบเทียบประสิทธิภาพแบบจำลองสมมติฐานด้วย T-test (Dependent Sample) ผลการศึกษาพบว่าเมื่อนำการคัดเลือกคุณลักษณะวิธี

InforGainAttributeEval ด้วยแบบจำลองเทคนิค Neural Network มีประสิทธิภาพมากที่สุด มีค่าความถูกต้องสูงสุดคิดเป็นร้อยละ 63.45 จึงเหมาะสมที่จะนำไปเป็นแนวทางการพัฒนาระบบแนะนำอาชีพสำหรับนักศึกษาปริญญาตรี คณะโบราณคดี มหาวิทยาศิลปากรต่อไป

7.1.11 ขันทอง ปทุมชาติ และพิมริมทร์ คีรีนทร์ (2558) ศึกษาเรื่องการวิเคราะห์พฤติกรรมการเลือกสมัครสาขาวิชาเรียนของนักศึกษาใหม่ โดยใช้เทคนิคการเหมืองข้อมูล นำข้อมูลการสมัครเข้าศึกษาประเภทการสมัคร เพศ ระดับเกรดเฉลี่ยประเภทโรงเรียนภายในและนอกจังหวัด กลุ่มการเรียน สาขาวิชาที่เลือกสมัครตามลำดับ 1 -3 ตามกรอบการทำเหมืองข้อมูลแบบ CRISP-DM ด้วยโปรแกรม WEKA และสร้างแบบจำลองการแบ่งกลุ่มด้วยวิธีเคมีนเพื่อหากลุ่มที่เหมาะสมด้วยการกำหนดกลุ่มออกเป็น 4-8 กลุ่มด้วยวิธี Devies-Bouldin’s index และ Dunn’s index นำผลที่ได้ไปปรึกษาผู้เชี่ยวชาญโดยสรุปการแบ่งข้อมูลออกเป็น 2 ประเภทและแบ่งกลุ่มย่อย 6 กลุ่ม หาความสัมพันธ์ของการเลือกสาขาวิชาผลที่ได้ภายหลังจากโปรแกรมกฎการแบ่งกลุ่มประเมินผลที่ได้ผลการศึกษาพบว่ามีความคลาดเคลื่อนไม่เกิน 0.003 และกฎความสัมพันธ์ที่ได้จากความเชื่อมั่น (Confidence) มีค่าความเชื่อมั่นร้อยละ 70 ขึ้นไปและ Minimum Support เท่ากับ 0.03

7.1.12 เสกสรรค์ วิลัยลักษณ์ วิภา เจริญภัณฑรักษ์ และดวงดาว วิชาดากุล (2558) ศึกษาการใช้เทคนิคการทำเหมืองข้อมูลเพื่อพัฒนาคลังข้อมูลและสร้างตัวแบบพยากรณ์ผลการเรียนของนักเรียนโรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษาโดยใช้ข้อมูลนักเรียนระดับมัธยมศึกษาปีที่ 4 ระหว่างปีการศึกษา 2553-2556 จำนวน 525 ระเบียนประกอบด้วย 16 คุณลักษณะสร้างตัวแบบพยากรณ์ผลการเรียนโดยใช้ชุดข้อมูล 2 แบบ คือ ข้อมูลแบบไม่จัดกลุ่มและข้อมูลแบบจัดกลุ่ม นำเข้าสู่กระบวนการคัดเลือกคุณลักษณะวิธี Correlation-based Feature Selection (CFS) และวิธี Information Gain (IG) ร่วมกับเทคนิคเหมืองข้อมูลโครงข่ายประสาทเทียมแบบ Multilayer Perception (MLP) เทคนิค Support Vector Machine และเทคนิค Decision Tree และวัดประสิทธิภาพด้วยวิธี 10-fold Cross Validation พบว่าแบบจำลองทำนายโดยชุดข้อมูลแบบไม่จัดกลุ่ม (Original Data) ที่คัดเลือกด้วยวิธี Correlation-based Feature Selection ร่วมกับเทคนิค Neural Network แบบ Multi-Layer Perceptron จำนวน 5 คุณลักษณะได้แก่ แผนการเรียน ผลการเรียนเฉลี่ยวิทยาศาสตร์ ผลการเรียนเฉลี่ยสังคมศึกษา ผลการเรียนเฉลี่ยภาษาอังกฤษและผลการเรียนเฉลี่ยระดับชั้นมัธยมศึกษาปีที่ 3 มีความเหมาะสมโดยให้ค่าความถูกต้องสูงที่สุดร้อยละ 94.48 และมีรากที่สองของความคลาดเคลื่อน (RMSE) น้อยที่สุดที่ 0.1880 จากนั้นนำมาพัฒนาระบบพยากรณ์ผลการเรียนโดยใช้ภาษา PHP

7.1.13 นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) ศึกษาการวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์ เพื่อประยุกต์ใช้เทคนิคเหมืองข้อมูลทำนายผลสัมฤทธิ์ทางการเรียนของนิสิตระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยนเรศวร จำนวน 180 ระเบียน ประกอบด้วย 23 คุณลักษณะ โดยแบ่งออกเป็นตัวแปรอิสระ 22 คุณลักษณะ ใช้ข้อมูลพื้นฐานและข้อมูลผลการเรียนรายวิชาในชั้นปีที่ 1 และ 2 ตัวแปรตามหรือที่ใช้สำหรับทำนายคือเกรดเฉลี่ยเมื่อสำเร็จการศึกษา ใช้วิธีคัดเลือกคุณลักษณะที่สำคัญ 3 วิธีได้แก่ วิธี Correlation-base วิธี Consistency-base และวิธี Gain Ratio จากนั้นนำมาพัฒนาแบบจำลองการทำนายด้วยเทคนิค Neural Network แบบ Back-propagation และเทคนิค Support Vector

Machines และวัดประสิทธิภาพด้วยวิธี 10-flod Cross Validation และวัดค่ารากที่สองของความคลาดเคลื่อน (RMSE) พบว่าข้อมูลพื้นฐานไม่ใช่ข้อมูลสำคัญในการทำนายผลสัมฤทธิ์ทางการเรียน ตัวแปรสำคัญคือผลการเรียนรายวิชาจำนวน 10 คุณลักษณะ การใช้เทคนิค Neural Network แบบ Back-propagation และเทคนิค Support Vector Machines จากคุณลักษณะสำคัญมีค่าความข้างต้นมีค่าความผิดพลาดอยู่ที่ระดับต่ำกว่าแบบจำลองทำนายที่ใช้ตัวแปรตั้งต้นจำนวน 22 คุณลักษณะ ในขณะที่เทคนิคการรวมกลุ่มจำแนกประเภทด้วยวิธี Bagging ร่วมกับเทคนิค Neural Network แบบ Back-propagation และเทคนิค Support Vector Machine ผลการศึกษาพบว่าผลการพยากรณ์ของ Bagging ร่วมกับเทคนิค Neural Network แบบ Back-propagation มีค่ารากสองของความคลาดเคลื่อน (RMSE) อยู่ระดับต่ำสุดที่ 0.1051 มีประสิทธิภาพที่สุดสำหรับการนำมาพยากรณ์

7.1.14 พรเทพ คงไชย และรัชฎา คงคะจันทร์ (2554) ได้วิจัยเรื่องการศึกษาเชิงเปรียบเทียบในการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับการทำเหมืองข้อมูลเพื่อพยากรณ์โอกาสการศึกษาของนักศึกษา เพื่อนำเสนอรูปแบบและเปรียบเทียบการคัดเลือกคุณลักษณะแบบ Filter โดยใช้ Raker Method ซึ่งมีการใช้ Evaluator สำหรับคัดเลือกคุณลักษณะที่เหมาะสมด้วย วิธี Correlation-based Feature Selection วิธี Consistency-based Subset Evaluation และ วิธี Wrapper Subset Evaluation และการคัดเลือกคุณลักษณะแบบ Wrapper โดยใช้ Genetic Algorithm ร่วมกับการจำแนกประเภทด้วยเทคนิค Support Vector Machine เทคนิค Neural Network แบบ Multi-Layer Perceptron และเทคนิค Bayesian Belief Networks ใช้ชุดข้อมูลนักศึกษาปริญญาตรีที่เข้าศึกษาระหว่างปีการศึกษา 2544 – 2550 จำนวน 42,665 ชุด 20 คุณลักษณะ เปรียบเทียบการวัดค่าความถูกต้องด้วยค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าถ่วงดุล (F-Measure) ผลการศึกษาพบว่าการจำแนกประเภทด้วยการคัดเลือกคุณลักษณะแบบFilter โดย Ranker Method ร่วมกับ ReliefFAttributeEval เป็นวิธีที่ดีที่สุดโดยมีค่าคุณลักษณะที่เหมาะสมจำนวน 5 คุณลักษณะลดลงจากเดิม 20 คุณลักษณะ สามารถลดการใช้คุณลักษณะที่ไม่เหมาะสมและลดระยะเวลาประมวลผลในการทำเหมืองข้อมูลถึง 75% และให้ความถูกต้องของค่าPrecision อยู่ที่ 95.10% ค่า Recall อยู่ที่ 94.90% และค่า F-Measure ที่ 94.90% โดยใช้การจำแนกประเภทด้วยเทคนิค Neural Network แบบ Multi-Layer Perceptron

7.2 งานวิจัยต่างประเทศ

7.2.1 Osiris Villacampa (2015) ได้ศึกษาข้อมูลและทำงานวิจัยเรื่อง Feature Selection and Classification Methods for Decision Making : A Comparative Analysis โดยมีวัตถุประสงค์เพื่อศึกษาการประมวลผลการทำเหมืองข้อมูลเพื่อลดมิติข้อมูลด้วยการคัดเลือกคุณลักษณะหรือการเลือกคุณลักษณะซึ่งจะช่วยให้การทำเหมืองข้อมูลมีความถูกต้องและมีประสิทธิภาพ โดยใช้ข้อมูลประวัติการบริการและการขายรถจากตัวแทนจำหน่ายรถจำนวน 15,417 ระเบียน มีจำนวนคุณลักษณะ 40 คุณลักษณะ และประวัติการเงินของลูกค้าจากธนาคารจำนวน 10,578 ระเบียน มีจำนวนคุณลักษณะ 17 คุณลักษณะ เพื่อหารูปแบบจำแนกประเภทลูกค้าที่มีฐานะอยู่ในกลุ่มผู้ซื้อรถใหม่ โดยนำเทคนิคการคัดเลือกคุณลักษณะในแบบ Filters, Wrappers และ

Hybrid ได้แก่วิธี Information Gain วิธี Correlation Based Feature Selection วิธี Relief-F และวิธี Wrappers นำมาใช้เพื่อลดจำนวนคุณลักษณะในชุดข้อมูล จากนั้นทำการจำแนกประเภทและพัฒนาแบบจำลองด้วยเทคนิค Decision Tree เทคนิค K-Nearest Neighbor และเทคนิค Support Vector Machines ทดสอบประสิทธิภาพแบบจำลองด้วยวิธี 5-Fold Cross Validation และวิเคราะห์เปรียบเทียบการคัดเลือกคุณลักษณะและความแตกต่างด้วยค่า Accuracy, ค่า Area Under Receiver Operating Characteristic Curve (AUC), ค่า F-Measure, ค่า TF Rate และค่า FP Rate พบว่าค่าความถูกต้องของเทคนิค Decision Tree ร่วมกับการคัดเลือกคุณลักษณะแบบ Wrapper ที่ค่าความเชื่อมั่น 0.1 มีความถูกต้องแม่นยำถึง 86.4% ในการจำแนกประเภทการคัดเลือกคุณลักษณะแบบเดียวกันนี้ยังให้ค่า AUC สูงสุดที่ 90.4% ส่วนเทคนิค K-Nearest Neighbor และเทคนิค Support Vector Machines มีค่าความถูกต้องเพียง 84.9% และ 79.8% ตามลำดับ

7.3 ตารางเปรียบเทียบงานวิจัยที่เกี่ยวข้อง

ตารางที่ 2.2 ตารางงานวิจัยด้านการคัดเลือกคุณลักษณะเฉพาะปัจจัย/ข้อมูลสำคัญ

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
การคัดเลือกคุณลักษณะเฉพาะปัจจัย/ข้อมูลสำคัญ																
Correlation base Feature Selection			X	X	X								X		X	
Consistency based Feature Selection				X									X			
Wrapper Method													X			
Filter Method													X		X	✓
Gain Ratio				X					X	X	X	X	X			
Information Gain (IG)		X	X		X			X	X			X	X		X	✓
Apriori	X						X									

ตารางที่ 2.3 ตารางงานวิจัยแบบจำลองด้วยเทคนิคเหมืองข้อมูล

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
แบบจำลองด้วยเทคนิคเหมืองข้อมูล																
Decision Tree		X	X		X		X	X	X	X	X			X	X	✓
Neural Network			X	X	X				X	X	X		X			
Naïve Bayes					X						X		X	X		✓
Random Forest					X		X	X	X							✓
K-means	X						X									
K-nearest Neighbor														X	X	
Support Vector Machine (SVM)			X	X								X	X		X	

ตารางที่ 2.4 ตารางงานวิจัยการประเมินผลและการวัดประสิทธิภาพแบบจำลอง

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
การประเมินผลและการวัดประสิทธิภาพแบบจำลอง																
K-fold Cross Validation		X	X	X	X		X			X			X	X	X	✓
Confusion Matrix								X			X					✓
Precision	X												X			✓
Root Mean Square Error (RMSE)			X	X			X	X								
Accuracy		X	X		X		X	X		X				X	X	✓
Mean Squared Error (MAE)					X		X	X								
Recall													X			✓
F-measure													X		X	✓

ลำดับและความหมายในตารางเปรียบเทียบงานวิจัยที่เกี่ยวข้อง

ลำดับที่ 1 หมายถึง งานวิจัยของขันทอง ปทุมชาติ และพิมริมทร์ คีรีนทร์ (2558)

ลำดับที่ 2 หมายถึง งานวิจัยของรัชฎา เทพประสิทธิ์ และจรัญ แสนราช (2563)

ลำดับที่ 3 หมายถึง งานวิจัยของเสกสรรค์ วิลัยลักษณ์ วิภา เจริญภัณฑรักษ์ และดวงดาว วิชาดากุล (2558)

ลำดับที่ 4 หมายถึง งานวิจัยของนิภาพร ชนะมาร และพรรณี สิทธิเดช (2557)

ลำดับที่ 5 หมายถึง งานวิจัยของวันวิสาข์ ชนะประเสริฐ (2559)

ลำดับที่ 6 หมายถึง งานวิจัยของสุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง (2560)

ลำดับที่ 7 หมายถึง งานวิจัยของสำราญ วานนท์ และรจนา เมืองแสน (2563)

ลำดับที่ 8 หมายถึง งานวิจัยของสำราญ วานนท์ ธรัช อารีราษฎร์ และจรัญ แสนราช (2561)

ลำดับที่ 9 หมายถึง งานวิจัยของธนวรรณ วิสุทธิรัตน์ (2561)

ลำดับที่ 10 หมายถึง งานวิจัยของศุภามณ จันทร์สกุล (2561)

ลำดับที่ 11 หมายถึง งานวิจัยของภุมริน หรั่งน้อย คุณัญญา ส้มเกลี้ยง ปิยนันท์ เทียบศรไชย และประภาส ทองรัก (2564)

ลำดับที่ 12 หมายถึง งานวิจัยของอัจจิมา มณฑาพันธ์ (2562)

ลำดับที่ 13 หมายถึง งานวิจัยของพรเทพ คงไชย และรัชฎา คงคะจันทร์ (2554)

ลำดับที่ 14 หมายถึง งานวิจัยของ ธาดา จันตะคุณ (2560)

ลำดับที่ 15 หมายถึง งานวิจัยของ Osiris Villacampa (2015)

ลำดับที่ 16 หมายถึง งานวิจัยเรื่อง “การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่” (งานวิจัยของวิทยานิพนธ์นี้)

จากการศึกษางานวิจัยที่กล่าวมาแล้วนั้นผู้วิจัยมีความสนใจในการทำเหมืองข้อมูลโดยการรวบรวมข้อมูลคุณลักษณะจากข้อมูลพื้นฐานการสมัครเข้าศึกษาต่อในหลักสูตรครุศาสตร์บัณฑิต คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ คัดเลือกคุณลักษณะที่มีความสัมพันธ์และเหมาะสมตามกระบวนการคัดเลือกคุณลักษณะ (Feature Selection) ด้วยเทคนิค Information Gain ที่มีความเหมาะสมกับโครงสร้างข้อมูลจากการเก็บรวบรวม โดยลักษณะข้อมูลการจัดเก็บอยู่ในรูปแบบของตัวอักษรและตัวเลข นำมาสร้างแบบจำลองและประเมินประสิทธิภาพแบบจำลองด้วยการจำแนกข้อมูล (Data Classification) ซึ่งจากการศึกษางานวิจัยข้างต้นและแบบจำลองที่มีความใกล้เคียงกับโครงสร้างข้อมูล ได้แก่ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest) สำหรับการวิเคราะห์เชิงทำนายการสมัครเรียนในคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ เพื่อให้มีความถูกต้องมากที่สุด โดยใช้โปรแกรม Rapid Miner Studio ในการดำเนินการวิจัย

การเลือกใช้เทคนิค Information Gain สำหรับงานวิจัยนี้ เป็นการหาค่าน้ำหนักที่มีความสัมพันธ์กับคุณลักษณะเป้าหมาย คือคุณลักษณะที่มีความเกี่ยวข้องกับการเลือกสมัครหรือไม่เลือกสมัครในสาขาวิชาเรียน โดยใช้คุณลักษณะจากข้อมูลพื้นฐานที่รวบรวมและนำมาวิเคราะห์เพื่อหาค่าคุณลักษณะที่ดีที่สุดและนำค่าคุณลักษณะที่มีค่าน้ำหนักมาก หมายถึงค่าคุณลักษณะที่มีความสัมพันธ์กับคุณลักษณะเป้าหมาย มาสร้างแบบจำลองในแต่ละรูปแบบ และลดจำนวนคุณลักษณะลงทีละตัวเพื่อวิเคราะห์ประสิทธิภาพของคุณลักษณะที่นำมาใช้ในการจำแนกประเภทของข้อมูล ซึ่งจะทำให้ทราบถึงคุณลักษณะที่สามารถนำไปใช้งานต่อไปและคัดเลือกคุณลักษณะที่ดีและมีประสิทธิภาพเมื่อนำไปประยุกต์ใช้งานต่อไป

การเลือกเทคนิคการสร้างแบบจำลองพิจารณาถึงประเภทของข้อมูลที่นำมาวิเคราะห์ โดยข้อมูลที่มีการจัดเก็บและรวบรวมเป็นลักษณะของกลุ่ม ตัวอย่างเช่น กลุ่มของเพศ ซึ่งแบ่งออกเป็นเพศชายและเพศหญิง หรือกลุ่มของผู้ที่สมัครในแต่ละสาขาวิชา เมื่อพิจารณาถึงผลลัพธ์ที่ต้องการจากการวิเคราะห์ คือผลลัพธ์ที่แสดงให้เห็นว่าเลือกหรือไม่เลือกสมัครในแต่ละสาขา โดยอาศัยทฤษฎีและค่าจากสถิติที่เกิดขึ้น นำมาคำนวณหาความน่าจะเป็นในการจำแนกประเภทของข้อมูล การใช้เทคนิค Decision Tree เทคนิค Naïve Bays และเทคนิค Random Forest มีความสอดคล้องกับแนวคิดและข้อมูลที่จะนำมาวิเคราะห์ ซึ่งจะทำให้ผลการวิเคราะห์มีประสิทธิภาพยิ่งขึ้น

เครื่องมือที่ใช้ในงานวิจัยโปรแกรม RapidMiner Studio ถือเป็นเครื่องมือที่พัฒนาขึ้นสำหรับงานด้านการทำเหมืองข้อมูล โดยการใช้โอเปอร์เรเตอร์ (Operator) ที่เกี่ยวข้องในการรับค่าเพื่อประมวลผล ซึ่งสามารถปรับค่าของตัวแปรหรือพารามิเตอร์ (Parameter) ที่เกี่ยวข้องภายในตัวโอเปอร์เรเตอร์ของโปรแกรม และมีรูปแบบการใช้งานที่ง่าย เมื่อนำมาใช้เป็นเครื่องมือในการวิจัยนี้จึงถือเป็นการเพิ่มประสิทธิภาพในการทำงานวิจัย

การวัดประสิทธิแบบแบบจำลอง ในงานวิจัยนี้ได้เลือกการวัดประสิทธิภาพแบบจำลองโดยการใช้ตาราง Confusion Matrix ซึ่งเป็นการเก็บสถิติและจำนวนของผลที่เกิดขึ้นจากการทดสอบโดยนำคุณลักษณะของข้อมูลที่คัดเลือกมาใช้ร่วมกับแบบจำลอง ค่าในตาราง Confusion Matrix จะเก็บเป็นจำนวนของผลข้อมูลจริงและจำนวนของผลข้อมูลที่ทำนายในรูปแบบตาราง จากนั้นนำมาพิจารณาค่าความถูกต้อง ค่าความแม่นยำ ค่าความครบถ้วนและค่าประสิทธิภาพโดยรวม เมื่อพิจารณาแล้วนำข้อมูลหรือผลลัพธ์ที่ได้มาจัดเก็บในรูปแบบของตาราง หาผลลัพธ์แต่ละแบบจำลองโดยการเปรียบเทียบค่าทางสถิติ

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยเรื่อง “การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่” เป็นการวิจัยเชิงพัฒนา ผู้วิจัยได้กำหนดวิธีการดำเนินการวิจัยโดยมีรายละเอียดไว้ดังนี้

1. ข้อมูลที่ใช้ในการวิจัย
2. เครื่องมือที่ใช้ในการวิจัย
3. การวิเคราะห์ข้อมูล

1. ข้อมูลที่ใช้ในการวิจัย

การวิจัยครั้งนี้ทำการศึกษาการสมัครเรียนในหลักสูตร 4 ปี ระดับปริญญาตรี ข้อมูลที่ใช้ในการวิจัย คือ ข้อมูลผู้เข้าสมัครทุกรอบ ที่ผ่านการคัดเลือกและผู้สมัครที่ยืนยันการลงทะเบียนเรียนแล้ว เฉพาะการเรียนภาคปกติ ในปีการศึกษา 2562- 2564 จำนวน 4 หลักสูตรของคณะครุศาสตร์ คือ 1.หลักสูตรการศึกษาปฐมวัย 2.หลักสูตรการประถมศึกษา 3.หลักสูตรพลศึกษา 4.หลักสูตรการศึกษาพิเศษ (สำนักทะเบียนและประมวลผล, 2564) จำนวนทั้งสิ้น 684 รายการ

2. เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการดำเนินการวิจัยโดยประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการวิเคราะนี้แบ่งออกเป็น 2 ส่วนด้วยกันคือ ด้านฮาร์ดแวร์และด้านซอฟต์แวร์ ดังนี้

2.1 เครื่องมือด้านฮาร์ดแวร์

เครื่องคอมพิวเตอร์ : PC Dell OptiPlex 3060 MT

2.2 เครื่องมือด้านซอฟต์แวร์

ระบบปฏิบัติการ : Microsoft Windows 10

โปรแกรมจัดการข้อมูล : Microsoft Excel

โปรแกรมวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล : Rapid Miner Studio 9.10 (Education License)

3. การวิเคราะห์ข้อมูล

การวิจัยครั้งนี้ได้ดำเนินการวิเคราะห์ข้อมูลเพื่อทำนายการเลือกสมัครเรียนในสาขาวิชาเป้าหมายตามกระบวนการเหมืองข้อมูลของ Cross – Industry Standard Process for Data

Mining หรือ CRISP-DM (Jyoti, Nidhi and Sanjeev, 2013) มีกระบวนการ 6 ขั้นตอนประกอบด้วยขั้นตอนการดำเนินการต่าง ๆ ดังนี้

3.1 การทำความเข้าใจกับปัญหา (Business Understanding)

การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่พบปัญหาการประชาสัมพันธ์ข้อมูลหลักสูตรไม่ตรงกับความสนใจของนักเรียนและการให้ข้อมูลกลุ่มนักเรียนเป้าหมายในพื้นที่ไม่ตรงกับกลุ่มเป้าหมายที่จะสมัครเรียน ส่งผลให้นักศึกษาที่เข้ามาสมัครและตัดสินใจเลือกสาขาวิชาเรียนลดลง (กองแผนและนโยบายคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่, 2564)

ตารางที่ 3.1 ตารางแสดงรายการจำนวนนักศึกษาทั้งหมดของคณะครุศาสตร์

ปีการศึกษา	แผนการรับจำนวน (คน)	จำนวนรับ (คน)	จำนวนปัจจุบัน (คน)
2564	1230	1154	1065
2563	1340	1241	1163
2562	1280	1187	1155

ที่มา : กองแผนและนโยบายคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ (2564)

จากตารางที่ 3.1 แสดงจำนวนการรับสมัครในแต่ละปีการศึกษา โดยอธิบายได้ดังนี้

ปีการศึกษา 2564 แผนการรับจำนวน 1,230 คน จำนวนที่รับเข้าศึกษา 1,154 คน จำนวนปัจจุบัน 1,065 คน

ปีการศึกษา 2563 แผนการรับจำนวน 1,340 คน จำนวนที่รับเข้าศึกษา 1,241 คน จำนวนปัจจุบัน 1,163 คน

ปีการศึกษา 2562 แผนการรับจำนวน 1,280 คน จำนวนที่รับเข้าศึกษา 1,187 คน จำนวนปัจจุบัน 1,155 คน

จากรายละเอียดข้างต้นแสดงให้เห็นว่าในแต่ละปีการศึกษาจำนวนที่รับนักศึกษาเข้าศึกษามีจำนวนน้อยกว่าแผนการรับและภายหลังจากการรับเข้าแล้ว จำนวนผู้เข้าศึกษาคงเหลือในแต่ละปีมีจำนวนลดลงจากจำนวนที่รับเข้าศึกษา

3.2 การทำความเข้าใจและรวบรวมข้อมูลที่เกี่ยวข้อง (Data Understanding)

ผู้วิจัยได้ศึกษาและรวบรวมข้อมูลที่ถูกเก็บในระบบฐานข้อมูลของสำนักทะเบียนและประมวลของมหาวิทยาลัยราชภัฏเชียงใหม่ ซึ่งเป็นเจ้าของและดูแลข้อมูลการรับสมัครนักศึกษา จากนั้นนำข้อมูลที่ได้มาพิจารณาความเป็นไปได้ในการนำมาวิเคราะห์ โดยข้อมูลที่ได้เป็นข้อมูลพื้นฐานและความสัมพันธ์ด้านครอบครัวของผู้สมัคร จึงพิจารณานำข้อมูลผู้สมัครเข้าศึกษาต่อในหลักสูตรของคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ ที่อยู่ในความรับผิดชอบของอาจารย์ประจำคณะมีจำนวน 4 หลักสูตร ได้แก่ 1.หลักสูตรการศึกษาปฐมวัย 2.หลักสูตรการประถมศึกษา 3.หลักสูตรพล

ศึกษา 4.หลักสูตรการศึกษาพิเศษ ในปีการศึกษา 2562-2564 จำนวนทั้งสิ้น 684 รายการ มีปัจจัยที่ต้องพิจารณาคุณลักษณะที่ต้องพิจารณาทั้งหมด 26 คุณลักษณะ (Attributes) โดยปัจจัยที่นำมาพิจารณาคุณลักษณะนั้นมีบางคุณลักษณะที่เก็บข้อมูลแบบเดียวกันซึ่งอยู่ในรูปของตัวแปรอื่น จึงนำมาพิจารณาและนำคุณลักษณะที่มีการเก็บข้อมูลซ้ำซ้อนกัน ตัวอย่างเช่น การเก็บข้อมูลคำนำหน้าซึ่งค่าที่พบในฐานข้อมูลมีเพียงนายและนางสาว ซึ่งสามารถบ่งบอกได้ว่านายหรือนางสาวเป็นเพศใด และการเก็บข้อมูลคุณลักษณะของเพศ เมื่อพิจารณาแล้วคุณลักษณะที่เก็บนายและนางสาวมีความใกล้เคียงกับการเก็บเพศชายและเพศหญิงจึงเลือกมาเพียงหนึ่งคุณลักษณะสำหรับนำมาวิเคราะห์ข้อมูล เป็นต้น ซึ่งข้อมูลคงเหลือปัจจัยในการพิจารณาทั้งหมดจำนวน 17 คุณลักษณะ ได้แก่ 1)ปีที่สมัคร 2)เพศ 3)สาขาที่สมัคร 4)แผนการเรียนที่จบจากระดับมัธยม 5)เกรดเฉลี่ย 6)สัญชาติ 7)ศาสนา 8)สัญชาติบิดา 9)สัญชาติมารดา 10)สถานะการมีชีวิตบิดา 11)สถานะการมีชีวิตมารดา 12)สถานะครอบครัว 13)จังหวัด 14)ผู้ปกครอง 15)อาชีพผู้ปกครอง 16)อาชีพบิดา 17)อาชีพมารดา ดังแสดงรายละเอียดในตารางที่ 3.2

3.3 การเตรียมข้อมูล (Data Preparation)

ข้อมูลที่ได้จากสำนักทะเบียนและประมลผลอยู่ในรูปแบบของไฟล์ Excel โดยในขั้นตอนการเตรียมข้อมูลได้นำข้อมูลมาตรวจสอบความถูกต้องของข้อมูล ด้วยโปรแกรม Microsoft Excel โดยพิจารณาข้อมูลเป็นลำดับดังนี้

3.3.1 คัดกรองและเลือกข้อมูล (Data Selection) ที่สนใจโดยคัดเลือกหลักสูตรของคณะครุศาสตร์ เพื่อเป็นการทดลองและเป็นแนวทางในการดำเนินการก่อนซึ่งเมื่อพิจารณาแล้วได้คัดเลือกกลุ่มข้อมูลจำนวน 4 หลักสูตรดังได้อธิบายไว้ในหัวข้อ 3.1 นำมาวิเคราะห์

3.3.2 ทำความสะอาดข้อมูล (Data Cleaning) ภายหลังจากการเลือกข้อมูลที่มีความสนใจแล้ว นำข้อมูลที่ได้มาทำความสะอาดโดยการตัดรายการข้อมูลที่ไม่ครบถ้วน (Missing Data) ออก ตัวอย่างเช่น เกรดเฉลี่ยของผู้เข้าสมัครที่ไม่ได้ใส่ข้อมูลหรือใส่เป็นตัวหนังสือ แผนการเรียนที่จบจากระดับชั้นมัธยมไม่ได้กรอกข้อมูล เพื่อให้การวิเคราะห์และทำนายผลมีความถูกต้อง

3.3.3 ตรวจสอบรูปแบบของข้อมูลให้ถูกต้องและแปลงข้อมูล (Data Transformation) ภายหลังจากการทำความสะอาดข้อมูลแล้ว นำข้อมูลมาตรวจสอบคัดเลือกโดยการจัดกลุ่มของข้อมูลให้เหมาะสมกับกระบวนการทำเหมืองข้อมูล ตัวอย่างเช่น ข้อมูลที่เป็นค่าตัวเลขแต่จัดเก็บอยู่ในรูปแบบของตัวหนังสือ เป็นต้น จากนั้นกำหนดค่าคุณลักษณะเพื่อนำข้อมูลที่ได้ทั้งหมดไปเข้าสู่กระบวนการต่อไป ดังรายละเอียดอธิบายคุณลักษณะตามตารางที่ 3.2

ตารางที่ 3.2 ตารางแสดงรายละเอียดคุณลักษณะของข้อมูลและการแปลงค่าข้อมูล

คุณลักษณะ	คำอธิบาย	การแปลงข้อมูล	ประเภท
Std_regyear	ปีที่สมัคร	2562	Polynominal
		2563	
		2564	

ตารางที่ 3.2 (ต่อ)

คุณลักษณะ	คำอธิบาย	การแปลงข้อมูล	ประเภท
Std_Gender	เพศ	MALE = ชาย FEMALE = หญิง	Biominal
Major	ชื่อหลักสูตรหรือสาขาวิชา	EDM01 = หลักสูตรการศึกษาปฐมวัย EDM02 = หลักสูตรการประถมศึกษา EDM03 = หลักสูตรพลศึกษา EDM04 = หลักสูตรการศึกษาพิเศษ	Polynominal
Std_program	แผนการเรียนที่จบจากระดับมัธยม	PLAN01 = วิทย์-คณิต PLAN02 = ศิลป์ (คำนวณ สังคม ทั่วไป ฯลฯ) PLAN03 = อื่น ๆ (สายอาชีพ,ไม่ลงข้อมูล)	Polynominal
Std_GPA	เกรดเฉลี่ยที่จบจากระดับมัธยม	Low = ผลการเรียนเฉลี่ย 2- 2.5 Medium = ผลการเรียนเฉลี่ย 2.51 – 3.00 Good = ผลการเรียนเฉลี่ย 3.01 – 3.50 Very Good = ผลการเรียนเฉลี่ย 3.51 - 4	Polynominal
Std_Local	มีภูมิลำเนาหรือพักอาศัย ในจังหวัดเชียงใหม่และแม่ฮ่องสอน	YES = จังหวัดเชียงใหม่และแม่ฮ่องสอน NO = จังหวัดอื่น	Biominal
Std_Nationality	การมีสัญชาติไทยของผู้สมัคร	YES = สัญชาติไทย NO = สัญชาติอื่น	Biominal
Std_F_Nationality	การมีสัญชาติไทยของบิดาผู้สมัคร	YES = สัญชาติไทย NO = สัญชาติอื่น	Biominal
Std_M_Nationality	การมีสัญชาติไทยของมารดาผู้สมัคร	YES = สัญชาติไทย NO = สัญชาติอื่น	Biominal
Region	ศาสนาที่นับถือ	Buddhism = ศาสนาพุทธ Other = ศาสนาอื่น	Biominal
Std_F_Life	สถานะการมีชีวิตของบิดา	Alive = มีชีวิต Dead = เสียชีวิตแล้ว	Biominal
Std_M_Life	สถานการณ์มีชีวิตของมารดา	Alive = มีชีวิต Dead = เสียชีวิตแล้ว	Biominal
Std_P_Reletional	ผู้ปกครอง	Father = พ่อ Mother = แม่ Other = ญาติหรือบุคคลอื่น	Polynominal
Std_Family	สถานภาพทางครอบครัว	Still together = พ่อแม่อยู่ด้วยกัน Divorce = พ่อแม่หย่าร้าง	Biominal
Std_F_Occupation	กลุ่มการประกอบอาชีพของบิดา	GRP01 = รับราชการ / พนักงานราชการ / รัฐวิสาหกิจ /พนักงานรัฐวิสาหกิจ GRP02 = พนักงานเอกชน / หน่วยงานเอกชน GRP03 = ค้าขาย /ธุรกิจส่วนตัว /รับจ้างอิสระ GRP04 = เกษตรกร / ประมง GRP05 = อื่น ๆ	Polynominal

ตารางที่ 3.2 (ต่อ)

คุณลักษณะ	คำอธิบาย	การแปลงข้อมูล	ประเภท
Std_M_Occupation	กลุ่มการประกอบอาชีพของมารดา	GRP01 = รับราชการ / พนักงานราชการ / รัฐวิสาหกิจ /พนักงานรัฐวิสาหกิจ GRP02 = พนักงานเอกชน / หน่วยงานเอกชน GRP03 = ค้าขาย /ธุรกิจส่วนตัว /รับจ้างอิสระ GRP04 = เกษตรกร / ประมง GRP05 = อื่น ๆ	Polynominal
Std_P_Occupation	กลุ่มการประกอบอาชีพของผู้ปกครอง	GRP01 = รับราชการ / พนักงานราชการ / รัฐวิสาหกิจ /พนักงานรัฐวิสาหกิจ GRP02 = พนักงานเอกชน / หน่วยงานเอกชน GRP03 = ค้าขาย /ธุรกิจส่วนตัว /รับจ้างอิสระ GRP04 = เกษตรกร / ประมง GRP05 = อื่น ๆ	Polynominal

จากตารางที่ 3.2 อธิบายการแทนค่าคุณลักษณะและการแปลงข้อมูลที่เกี่ยวข้องทั้งหมดจำนวน 17 คุณลักษณะเพื่อนำไปใช้ในการวิเคราะห์ โดยประเภทของข้อมูลได้แบ่งออกเป็น 2 รูปแบบ คือ ข้อมูลรูปแบบ Polynominal คือข้อมูลกลุ่มที่ไม่ใช่ตัวเลขนำมาคำนวณไม่ได้แต่นับจำนวนได้ ซึ่งจะมีจำนวนกลุ่มมากกว่า 2 กลุ่ม ตัวอย่างเช่น กลุ่มของหลักสูตรการศึกษาปฐมวัย หลักสูตรการประถมศึกษา หลักสูตรพลศึกษา หลักสูตรการศึกษาพิเศษซึ่งมีจำนวนมากกว่า 2 กลุ่ม เป็นต้น และข้อมูลรูปแบบ Biominal คือข้อมูลกลุ่มที่ไม่ใช่ตัวเลขนำมาคำนวณไม่ได้แต่นับจำนวนได้ แต่ข้อมูลรูปแบบนี้จะมีค่าเพียง 2 ค่าเท่านั้น ตัวอย่างเช่น เพศชาย เพศหญิง เป็นต้น

ข้อมูลที่ได้จากแปลงข้อมูลตามกระบวนการข้างต้นแล้ว หากนำมาแสดงผลในรูปแบบข้อมูล CSV ด้วยโปรแกรม Microsoft Excel แล้วสามารถดูข้อมูลดังแสดงภาพที่ 3.1 ซึ่งถือเป็นข้อมูลที่พร้อมจะนำเข้าสู่โปรแกรม Rapid Miner ที่ผู้วิจัยนำมาเป็นเครื่องมือในวิเคราะห์ในการวิจัยนี้

std_id	std_regyear	std_gender	major	std_program	std_gpa	std_nation	std_region
15	2562	FEMALE	EDM03	PLAN01	GOOD	YES	Buddhism
17	2562	FEMALE	EDM02	PLAN01	MEDIUM	YES	Buddhism
25	2562	FEMALE	EDM04	PLAN02	GOOD	YES	Buddhism
26	25						Buddhism
31	25						Other
33	2562	FEMALE	EDM02	PLAN01	MEDIUM	YES	Buddhism

ข้อมูลที่ถูกจัดรูปแบบแล้วพร้อมจะนำเข้าโปรแกรมเพื่อวิเคราะห์

ภาพที่ 3.1 ข้อมูลที่จะนำไปเข้าสู่กระบวนการวิเคราะห์

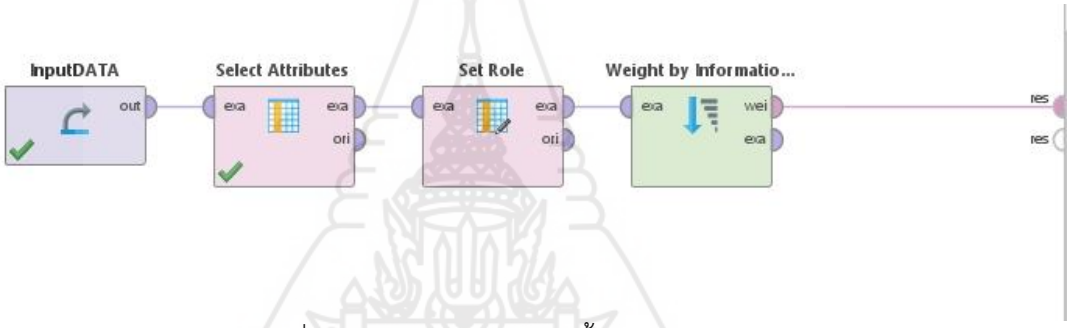
จากภาพที่ 3.1 แสดงข้อมูลคุณลักษณะที่ถูกแปลงค่าซึ่งอธิบายตัวอย่างเช่น คอลัมน์ Std_id หมายถึงค่ารหัสของรายการซึ่งจะนำมาเป็น Primary Key ในการเรียงลำดับ และในตัวอย่างคอลัมน์ Std_gpa การแปลงค่าเกรดเฉลี่ยให้อยู่ระดับ Low Medium Good และ Very Good ดังได้อธิบายและแสดงรายละเอียดในตารางแสดงรายละเอียดคุณลักษณะของข้อมูลและการแปลงค่าข้อมูลข้างต้น

3.4 การสร้างแบบจำลอง (Modeling)

ภายหลังจากการนำข้อมูลเข้าในโปรแกรม Rapid Miner ผู้วิจัยได้ดำเนินการตามขั้นตอน ดังนี้

ขั้นตอนที่ 1 การคัดเลือกคุณลักษณะโดยการหาค่าน้ำหนักจากข้อมูล คุณลักษณะทั้งหมดที่ได้จากการเตรียมข้อมูลจำนวน 17 คุณลักษณะนำมาเข้าสู่กระบวนการคัดเลือกเพื่อหาค่าคุณลักษณะที่มีความสำคัญและสัมพันธ์กับการเลือกสมัครหลักสูตร โดยการหาค่าน้ำหนักด้วยเทคนิค Information Gain

ผู้วิจัยเลือกใช้โอเปอร์เรเตอร์ Select Attributes ของโปรแกรม Rapid Miner สำหรับการเลือกตัวแปรและหาค่าน้ำหนักโดยใช้โอเปอร์เรเตอร์ Weight By Information Gain หาค่าน้ำหนักของตัวแปรที่นำมาวิเคราะห์ด้วยเทคนิค Information Gain จากนั้นกำหนดให้แอตทริบิวต์ตัวแปรชื่อ std_id เป็นประเภท Id และกำหนดให้แอตทริบิวต์ ตัวแปรชื่อ Major เป็นประเภท Label ดังแสดงภาพที่ 3.2



ภาพที่ 3.2 กระบวนการหาค่าน้ำหนัก Information Gain

จากภาพที่ 3.2 แสดงกระบวนการทำงานในการหาค่าน้ำหนักด้วยเทคนิค Information Gain ซึ่งสามารถอธิบายในแต่ละขั้นตอนดังนี้

InputData คือข้อมูลทั้งหมดที่นำมาวิเคราะห์

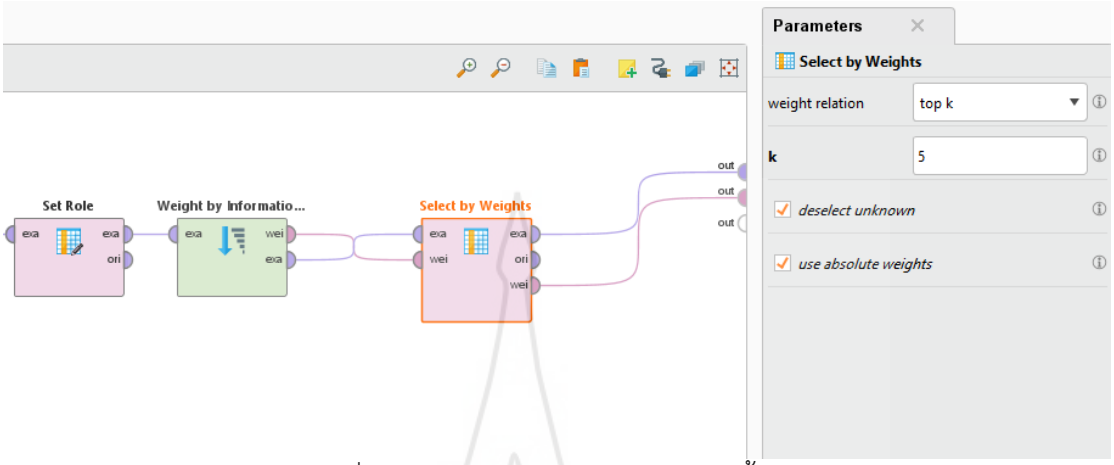
Select Attributes คือ การเลือกหรือกำหนดคุณลักษณะจากข้อมูลทั้งหมดมาเข้าสู่กระบวนการคัดเลือก

Set Role คือ การกำหนดกลุ่มหรือกฎในการคัดเลือกคุณลักษณะ ซึ่งในขั้นตอนนี้ได้กำหนดตัวแปร Std_id เป็นประเภท Id เพราะจะไม่นำตัวแปรนี้มาคำนวณในการคัดเลือก และกำหนดตัวแปร Major เป็นประเภท Label เพื่อเป็นป้ายกำกับในการค้นหาค่าน้ำหนักของคุณลักษณะที่มีความเกี่ยวข้องและสัมพันธ์กับตัวแปร Major

Weight by Information Gain คือการหาค่าคุณลักษณะด้วยเทคนิค Information Gain

ภายหลังจากการรันโปรแกรมแล้ว จะแสดงผลค่าน้ำหนักของคุณลักษณะทั้งหมดผ่านหน้าต่างแสดงผลของโปรแกรม Rapid Miner เมื่อทราบค่าน้ำหนักของแต่ละแอตทริบิวต์แล้ว เราสามารถเลือกแอตทริบิวต์ที่ให้ค่าสูงมาสร้างแบบจำลองโดยการเลือกโอเปอร์เรเตอร์ Select By Weight เลือกค่า Weight Relation เป็น Top K หมายถึงเลือกจำนวนที่มีค่ามากที่สุดและปรับลดค่า

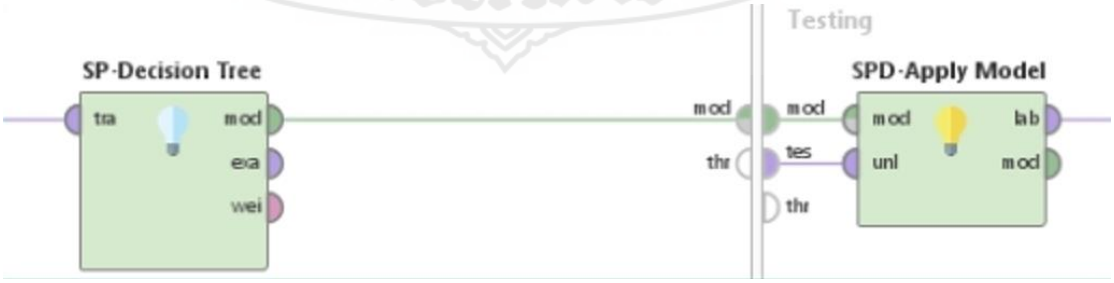
จำนวนแอตทริบิวต์ (ค่า k) จำนวนค่าที่ใช้และมีค่าความถูกต้องมากที่สุดอยู่ที่จำนวน 5 คุณลักษณะ (Attributes)



ภาพที่ 3.3 การเลือกคุณลักษณะด้วยค่าน้ำหนัก

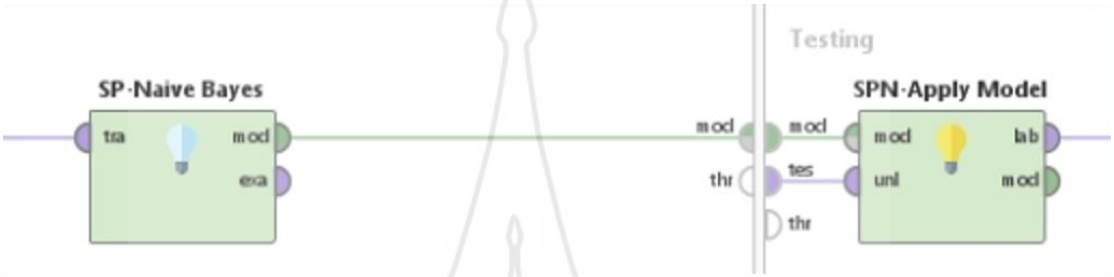
จากภาพที่ 3.3 แสดงการเลือกคุณลักษณะโดยใช้โอเปอร์เรเตอร์ Select by Weights ที่จำนำไปเข้าสู่กระบวนการวิเคราะห์ในขั้นตอนต่อไป

ขั้นตอนที่ 2 สร้างตัวแบบจำลองเพื่อวิเคราะห์ข้อมูล จากขั้นตอนที่ 1 เมื่อได้ผลการคัดเลือกคุณลักษณะที่มีความเหมาะสมที่จะมาสร้างแบบจำลองแล้ว ผู้วิจัยได้นำเทคนิคการจำแนกประเภทข้อมูลได้แก่เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest) ซึ่งจากการศึกษาข้อมูลแล้วมีความเหมาะสมกับรูปแบบและลักษณะของข้อมูลที่นำมาวิเคราะห์ เพื่อเปรียบเทียบแบบจำลองที่มีความเหมาะสมกับข้อมูลที่นำมาวิเคราะห์มากที่สุด โดยใช้งานร่วมกับการใช้คุณลักษณะที่ได้คัดเลือกจากเทคนิค Information Gain ซึ่งผู้วิจัยได้ศึกษาและเลือกใช้โอเปอร์เรเตอร์ในโปรแกรม Rapid Miner ด้วยโอเปอร์เรเตอร์ด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest) ตามลำดับ ดังแสดงในภาพที่ 3.4 – 3.6



ภาพที่ 3.4 แบบจำลองด้วยเทคนิค Decision Tree

จากภาพที่ 3.4 ผู้วิจัยเลือกใช้โอเปอร์เรเตอร์ Decision Tree สำหรับการสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ โดยรับค่าคุณลักษณะที่ได้จากการเลือกค่าน้ำหนัก ในขา wei หมายถึง คุณลักษณะที่เลือกโดยค่า wei ดังอธิบายไว้ในขั้นตอนที่ 1 มาเข้าขารับข้อมูล tra หมายถึงข้อมูลที่จะนำเข้าเรียนรู้และนำผลลัพธ์ออกในขา mod หมายถึง แบบจำลองที่ได้จากกระบวนการจำแนกประเภท (Classification) จากนั้นนำเข้าสู่กระบวนการวัดประสิทธิภาพซึ่งแสดงให้เห็นในส่วนของการ Testing ของโปรแกรม



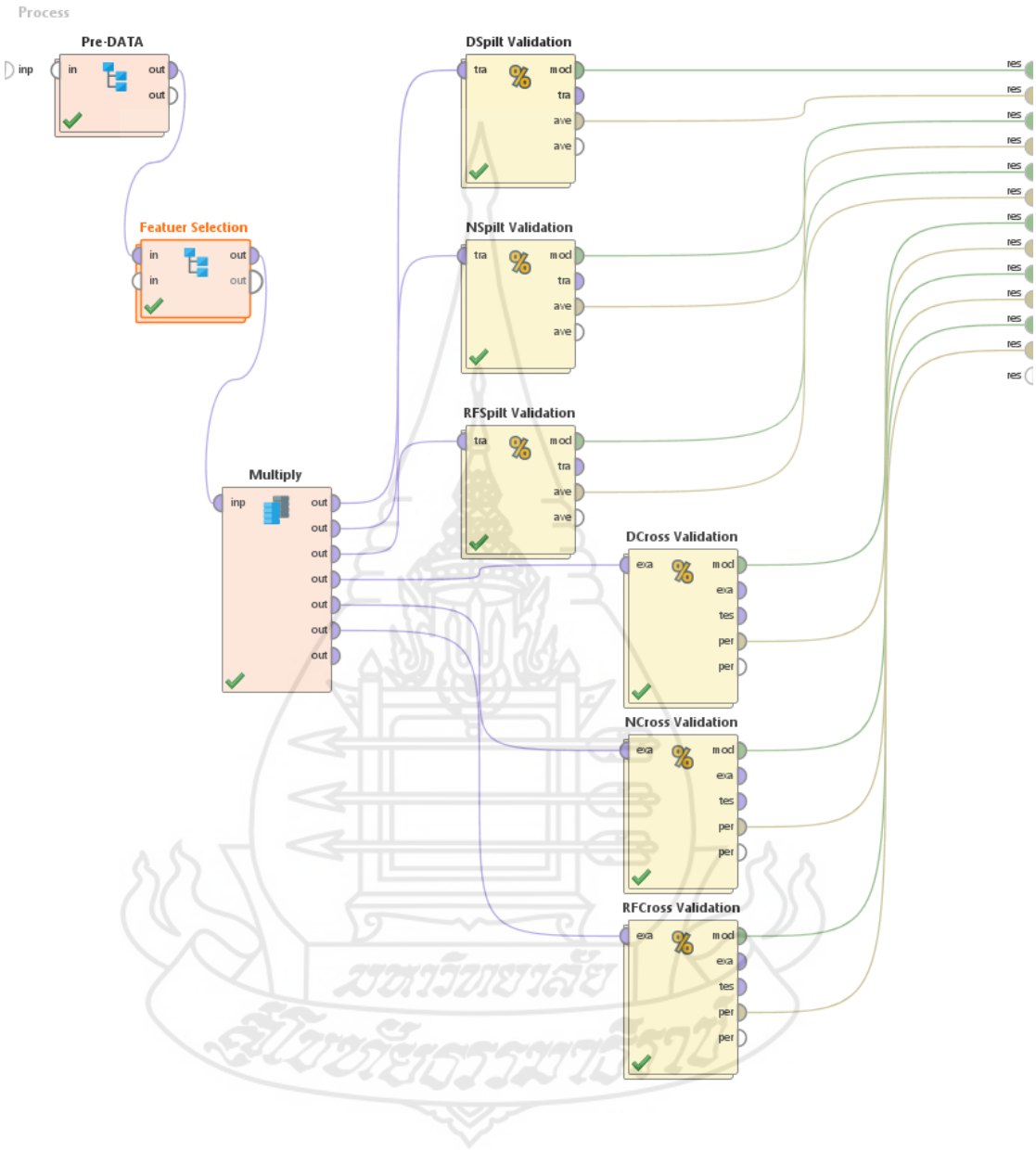
ภาพที่ 3.5 แบบจำลองด้วยเทคนิค Naïve Bays

จากภาพที่ 3.5 ผู้วิจัยเลือกใช้โอเปอร์เรเตอร์ Naïve Bays สำหรับการสร้างแบบจำลองด้วยเทคนิคนาอีฟเบย์ โดยรับค่าที่ได้จากการเลือกค่าน้ำหนัก ในขา wei หมายถึง คุณลักษณะที่เลือกโดยค่า wei ดังอธิบายไว้ในขั้นตอนที่ 1 มาเข้าขารับข้อมูล tra หมายถึงข้อมูลที่จะนำเข้าเรียนรู้และนำผลลัพธ์ออกในขา mod หมายถึง แบบจำลองที่ได้จากกระบวนการจำแนกประเภท (Classification) จากนั้นนำเข้าสู่กระบวนการวัดประสิทธิภาพซึ่งแสดงให้เห็นในส่วนของการ Testing ของโปรแกรม

ภาพที่ 3.6 แบบจำลองด้วยเทคนิค Random Forest

จากภาพที่ 3.6 ผู้วิจัยเลือกใช้โอเปอร์เรเตอร์ Random Forest สำหรับการสร้างแบบจำลองด้วยเทคนิคนาป่าสุ่ม ซึ่งการรับค่าและการนำผลลัพธ์ออกมีกระบวนการทำงานรูปแบบเดียวกันกับการสร้างแบบจำลองและการวัดประสิทธิภาพด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคนาอีฟเบย์ (Naïve Bays) ดังได้แสดงในภาพที่ 3.4 และ 3.5

สำหรับขั้นตอนที่ 2 การสร้างแบบจำลองนี้ผู้วิจัยได้เลือกใช้โอเปอร์เรเตอร์ Multiply เป็นตัวแยกข้อมูลเพื่อให้สามารถใช้ข้อมูลชุดเดียวกันร่วมกับหลายเทคนิค โดยการทำงานประมวลผลในครั้งเดียวซึ่งแสดงให้เป็นกระบวนการทั้งหมดดังภาพที่ 3.7

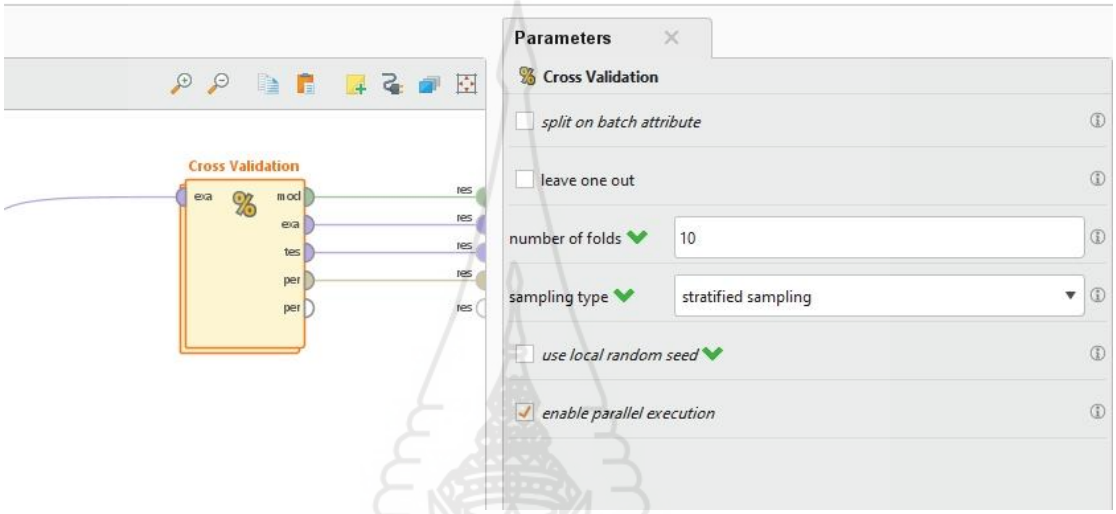


ภาพที่ 3.7 กระบวนการสร้างแบบจำลองและวัดประสิทธิภาพ

3.5 การประเมินวัดประสิทธิภาพของโมเดล (Evaluation)

การวัดประสิทธิภาพในการวิจัยนี้ใช้วิธีการ Cross-Validation Test ด้วย 5-fold Cross-Validation, 10-fold Cross-Validation และวิธี Split Test โดยการแบ่งข้อมูลออกเป็น 70% : 30% และ 80% : 20 % ของแบบจำลองที่สร้างด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟย์ และเทคนิคป่าสุ่มกับข้อมูลคุณลักษณะที่ได้จากการหาค่าน้ำหนักนำมาทดสอบเพื่อวัดประสิทธิภาพด้วย

การเปรียบเทียบค่าความถูกต้อง (Accuracy) ความแม่นยำของการทำนาย (Precision) ค่าความครบถ้วน(Recall) และค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure) ของการจำแนกประเภทข้อมูลด้วยเทคนิคที่ได้กล่าวไว้ข้างต้น ซึ่งการวัดประสิทธิภาพผู้วิจัยเลือกใช้โอเปอร์เรเตอร์ Apply Model เพื่อใช้ในการนำข้อมูลที่จัดเตรียมไว้ไปใช้งานร่วมกับแบบจำลอง และใช้โอเปอร์เรเตอร์ Performance สำหรับการวัดประสิทธิภาพแต่ละแบบจำลอง โดยปรับค่า parameter เพื่อให้สอดคล้องกับวิธีการวัดประสิทธิภาพ



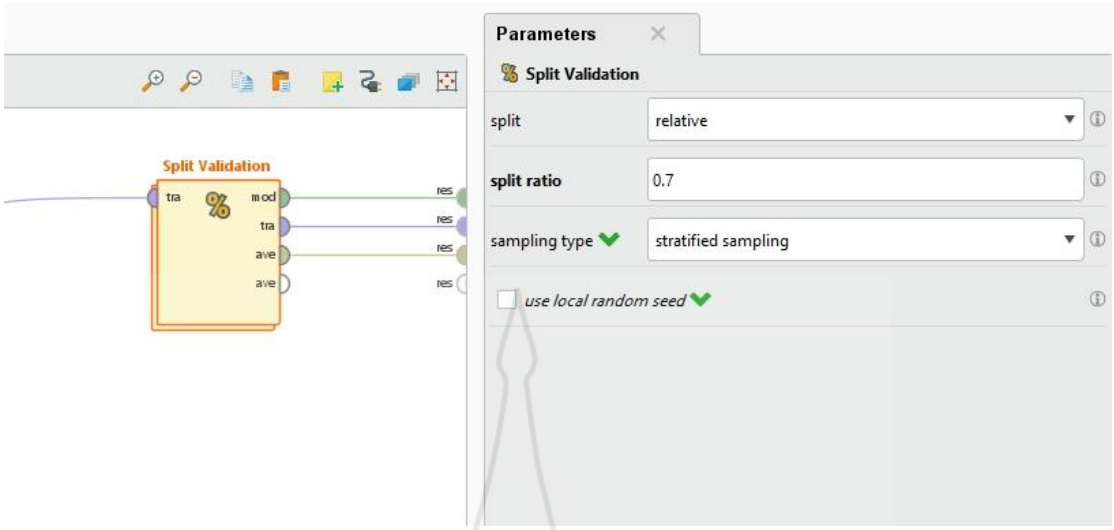
ภาพที่ 3.8 การวัดประสิทธิภาพด้วยวิธี Cross-Validation

จากภาพที่ 3.8 การวัดประสิทธิภาพด้วยวิธี Cross-Validation โดยการกำหนดค่าเพื่อทดสอบประสิทธิภาพดังนี้

การตั้งค่า folds สำหรับวิธี Cross-Validation โดยเลือกค่า Number of folds เลือก 10 สำหรับวิธี 10-fold Cross-Validation

การเลือกค่าการแบ่งประเภทกลุ่มตัวอย่าง ในช่อง Sampling type ของโปรแกรม ผู้วิจัยเลือกเป็น stratified sampling ซึ่งหมายถึง โปรแกรมจะสุ่มตัวอย่างจากข้อมูลที่นำวิเคราะห์โดยแยกประชากรออกเป็นกลุ่มประชากรย่อย ๆ หรือแบ่งเป็นชั้นภูมิก่อน

การวัดประสิทธิภาพด้วยวิธี Split Test ใช้โอเปอร์เรเตอร์ การตั้งค่า Split เป็น relative, ค่า Split ratio เป็น 0.7 (70%) และเลือก Sampling type เป็น stratified sampling ดังแสดงภาพที่ 3.9



ภาพที่ 3.9 การวัดประสิทธิภาพด้วยวิธี Split Test

จากภาพที่ 3.9 การวัดประสิทธิภาพด้วยวิธี Split Test โดยการกำหนดค่าเพื่อทดสอบประสิทธิภาพดังนี้

Split หมายถึง วิธีการแบ่งข้อมูลเพื่อทำการทดสอบ ผู้วิจัยเลือก relative ซึ่งหมายถึงไม่ได้ระบุจำนวนที่แน่นอนแต่เลือกจากค่าของข้อมูลทั้งหมดมาแบ่งจำนวนข้อมูล

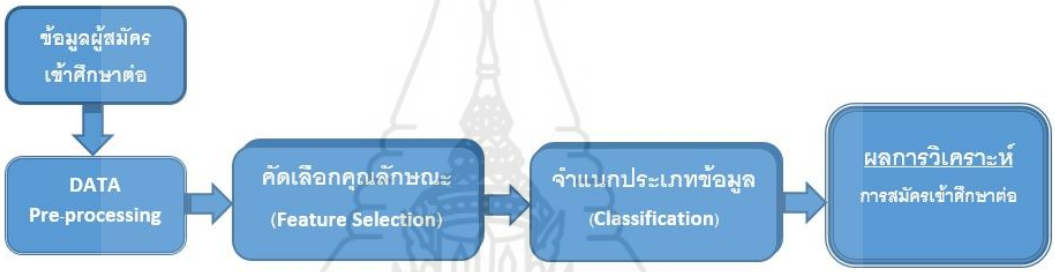
Split Ratio คืออัตราส่วนในการแบ่งข้อมูลสำหรับเรียนรู้ (Training Set) ดังตัวอย่าง 0.7 หมายถึงแบ่งข้อมูลทั้งหมดออกมา 70% สำหรับการเรียนรู้และคงเหลือ 30% จากข้อมูลทั้งหมดสำหรับทดสอบประสิทธิภาพ

3.6 การหาผลลัพธ์และองค์ความรู้ที่ได้มาประยุกต์ใช้ (Deployment)

ผลจากตัวแบบจำลองที่วัดประเมินประสิทธิภาพที่มีค่ามากที่สุด โดยวัดค่าจากความถูกต้องโดยรวมของโมเดล (Accuracy) จากตาราง Confusion Matrix แสดงค่าความแม่นยำ (Precision) โดยพิจารณาจากผลการทำนายเป็นหลัก และค่าความครบถ้วน(Recall) โดยพิจารณาจากความถูกต้องแม่นยำเทียบกับผลการเฉลยการทำนาย ค่าความครบถ้วน (Recall) และค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure) ที่ได้จากการทดสอบด้วยโปรแกรมดังได้อธิบายขั้นตอนไว้ในหัวข้อ 3.5 และ 3.6 ซึ่งค่าความถูกต้องจากการวัดประสิทธิภาพของแบบจำลองที่มากที่สุดคือ เทคนิคการจำแนกประเภทที่ดีที่สุด สามารถนำเทคนิคที่ได้มาเพื่อมาเป็นตัวแบบในการพัฒนาระบบสนับสนุนการวางแผนดำเนินงานและแนะนำข้อมูลให้กับผู้บริหารในคณะครุศาสตร์ใน และเพื่อข้อมูลในการตัดสินใจและวางแผนการลงพื้นที่ประชาสัมพันธ์ข้อมูลให้กับผู้เข้าสมัครเรียนหรือกลุ่มเป้าหมายในปีถัดไป

บทที่ 4
ผลการวิเคราะห์ข้อมูล

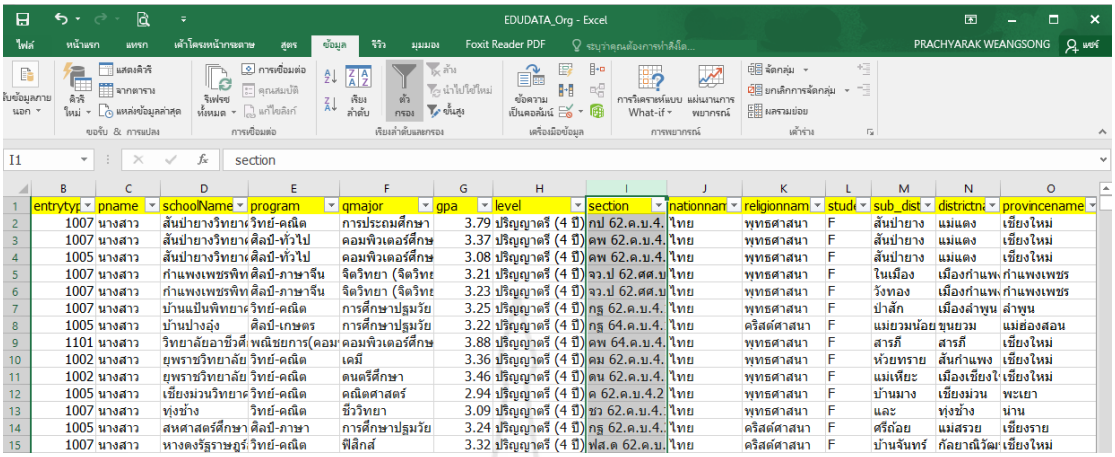
การวิจัยเรื่อง “การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่” ผู้วิจัยได้ดำเนินการวิจัยกระบวนการเหมืองข้อมูลของ Cross – Industry Standard Process for Data Mining หรือ CRISP-DM (Jyoti, Nidhi and Sanjeev, 2013) โดยมีกรอบแนวความคิดการวิจัยเพื่อให้ได้ผลการวิเคราะห์ตรงตามวัตถุประสงค์ดังอธิบายตามกรอบแนวความคิดการวิจัย ดังนี้



ภาพที่ 4.1 กรอบแนวความคิดการวิจัย

จากภาพที่ 4.1 แสดงกรอบแนวความคิดการวิจัย ผู้วิจัยได้ดำเนินการตามแนวความคิดในการวิจัยและมีกระบวนการแต่ละขั้นตอนดังนี้

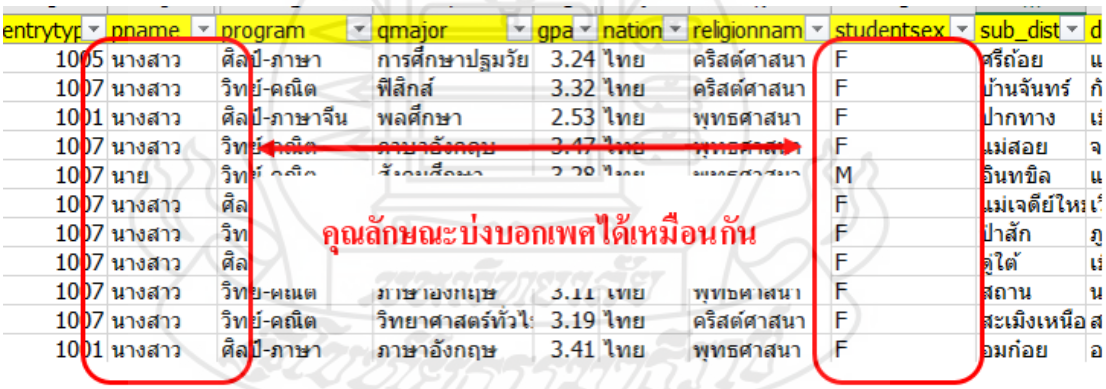
ข้อมูลผู้สมัครเข้าศึกษาต่อ ผู้วิจัยได้รวบรวมข้อมูลเพื่อนำมาใช้ในการวิจัยโดยได้รวบรวมข้อมูลพื้นฐานและความสัมพันธ์ด้านครอบครัวของผู้สมัครที่สามารถเปิดเผยได้จากสำนักทะเบียนและประมวลผลของมหาวิทยาลัยราชภัฏเชียงใหม่ โดยข้อมูลที่ได้รับอยู่ในรูปแบบของ CSV File ผู้วิจัยใช้โปรแกรม Microsoft Excel สำหรับการตรวจสอบและพิจารณาข้อมูล ดังภาพที่ 4.2



entrytyp	pname	schoolName	program	qmajor	gpa	level	section	nationnam	religionnam	stud	sub_dist	districtna	provincename
1007	นางสาว	สันป่ายางวิทยา	วิทย์-คณิต	การประถมศึกษา	3.79	ปริญญาตรี (4 ปี)	กป 62.ค.บ.4.	ไทย	พุทธศาสนา	F	สันป่ายาง	แม่แดง	เชียงใหม่
1007	นางสาว	สันป่ายางวิทยา	ศิลป์-ทั่วไป	คอมพิวเตอร์ศึกษ	3.37	ปริญญาตรี (4 ปี)	คพ 62.ค.บ.4.	ไทย	พุทธศาสนา	F	สันป่ายาง	แม่แดง	เชียงใหม่
1005	นางสาว	สันป่ายางวิทยา	ศิลป์-ทั่วไป	คอมพิวเตอร์ศึกษ	3.08	ปริญญาตรี (4 ปี)	คพ 62.ค.บ.4.	ไทย	พุทธศาสนา	F	สันป่ายาง	แม่แดง	เชียงใหม่
1007	นางสาว	กำแพงเพชรพิทา	ศิลป์-ภาษาจีน	จิตวิทยา (จิตวิทย	3.21	ปริญญาตรี (4 ปี)	จว.ป 62.ศศ.บ	ไทย	พุทธศาสนา	F	ในเมือง	เมืองกำแพ	กำแพงเพชร
1007	นางสาว	กำแพงเพชรพิทา	ศิลป์-ภาษาจีน	จิตวิทยา (จิตวิทย	3.23	ปริญญาตรี (4 ปี)	จว.ป 62.ศศ.บ	ไทย	พุทธศาสนา	F	วังทอง	เมืองกำแพ	กำแพงเพชร
1007	นางสาว	บ้านแป้นพิทยาค	วิทย์-คณิต	การศึกษาปฐมวัย	3.25	ปริญญาตรี (4 ปี)	กฐ 62.ค.บ.4.	ไทย	พุทธศาสนา	F	ป่าสัก	เมืองลำพูน	ลำพูน
1005	นางสาว	บ้านป่างอุ้ง	ศิลป์-เกษตร	การศึกษาปฐมวัย	3.22	ปริญญาตรี (4 ปี)	กฐ 64.ค.บ.4.	ไทย	คริสต์ศาสนา	F	แม่ยวมน้อย	ขุนยวม	แม่ฮ่องสอน
1101	นางสาว	วิทยาลัยอาชีวศึ	พณิชยการ(คอม	คอมพิวเตอร์ศึกษ	3.88	ปริญญาตรี (4 ปี)	คพ 64.ค.บ.4.	ไทย	พุทธศาสนา	F	สารภี	สารภี	เชียงใหม่
1002	นางสาว	ยุพราชวิทยาลัย	วิทย์-คณิต	เคมี	3.36	ปริญญาตรี (4 ปี)	คม 62.ค.บ.4.	ไทย	พุทธศาสนา	F	ห้วยทราย	สันกำแพง	เชียงใหม่
1002	นางสาว	ยุพราชวิทยาลัย	วิทย์-คณิต	คนตรีศึกษา	3.46	ปริญญาตรี (4 ปี)	ดน 62.ค.บ.4.	ไทย	พุทธศาสนา	F	แม่เหียะ	เมืองเชียงใ	เชียงใหม่
1005	นางสาว	เชียงม่วนวิทยาค	วิทย์-คณิต	คณิตศาสตร์	2.94	ปริญญาตรี (4 ปี)	ค 62.ค.บ.4.2	ไทย	พุทธศาสนา	F	บ้านมาง	เชียงม่วน	พะเยา
1007	นางสาว	ทุ่งช้าง	วิทย์-คณิต	ชีววิทยา	3.09	ปริญญาตรี (4 ปี)	ชว 62.ค.บ.4.	ไทย	พุทธศาสนา	F	และ	ทุ่งช้าง	น่าน
1005	นางสาว	สหศาสตร์ศึกษา	ศิลป์-ภาษา	การศึกษาปฐมวัย	3.24	ปริญญาตรี (4 ปี)	กฐ 62.ค.บ.4.	ไทย	คริสต์ศาสนา	F	ศรีถ้อย	แม่สรวย	เชียงราย
1007	นางสาว	หางดงรัฐราษฎรว	วิทย์-คณิต	ฟิสิกส์	3.32	ปริญญาตรี (4 ปี)	ฟส.ค 62.ค.บ.	ไทย	คริสต์ศาสนา	F	บ้านจันทร์	กัลยาณิวัฒ	เชียงใหม่

ภาพที่ 4.2 ข้อมูลผู้สมัครเข้าศึกษาต่อ

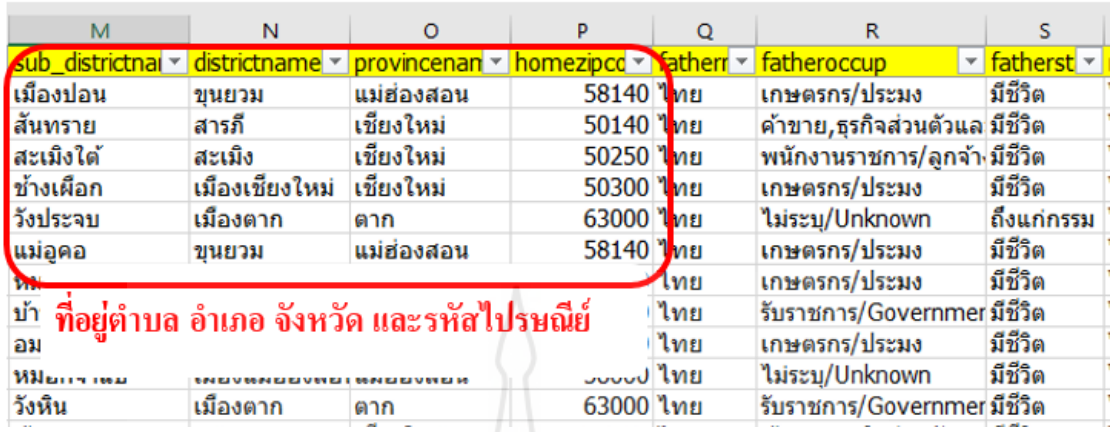
การเตรียมข้อมูล ภายหลังจากการรวบรวมข้อมูล ผู้วิจัยนำข้อมูลมาพิจารณาตัวแปรคุณลักษณะและค่าของข้อมูล ซึ่งจากการพิจารณาพบว่าการเก็บข้อมูลคุณลักษณะมีความซ้ำซ้อนกัน ตัวอย่างเช่นกลุ่มข้อมูลของเพศชายและเพศหญิง ที่เก็บอยู่ในตัวแปรคุณลักษณะ Studentsex และกลุ่มข้อมูลคำนำหน้านายและนางสาว ที่เก็บอยู่ในตัวแปรคุณลักษณะ pname เมื่อพิจารณาแล้วตัวแปรคุณลักษณะทั้งสองสามารถระบุเพศได้ จึงคัดเลือกคุณลักษณะเพียงหนึ่งคุณลักษณะเพื่อนำมาใช้ในการวิเคราะห์ ดังภาพที่ 4.3



entrytyp	pname	program	qmajor	gpa	nation	religionnam	studentsex	sub_dist	d
1005	นางสาว	ศิลป์-ภาษา	การศึกษาปฐมวัย	3.24	ไทย	คริสต์ศาสนา	F	ศรีถ้อย	แ
1007	นางสาว	วิทย์-คณิต	ฟิสิกส์	3.32	ไทย	คริสต์ศาสนา	F	บ้านจันทร์	กั
1001	นางสาว	ศิลป์-ภาษาจีน	พลศึกษา	2.53	ไทย	พุทธศาสนา	F	ปากทาง	เ
1007	นางสาว	วิทย์-คณิต	ภาษาอังกฤษ	3.47	ไทย	พุทธศาสนา	F	แม่สอย	จ
1007	นาย	วิทย์-คณิต	ภาษาอังกฤษ	2.28	ไทย	พุทธศาสนา	M	อินทขิล	แ
1007	นางสาว	ศิล					F	แม่เจดีย์ใหม่	เ
1007	นางสาว	วิท					F	ป่าสัก	ภู
1007	นางสาว	ศิล					F	ดู่ใต้	เ
1007	นางสาว	วิทย์-คณิต	ภาษาอังกฤษ	3.11	ไทย	พุทธศาสนา	F	สถาน	น
1007	นางสาว	วิทย์-คณิต	วิทยาศาสตร์ทั่วไ	3.19	ไทย	คริสต์ศาสนา	F	สะเมิงเหนือ	ส
1001	นางสาว	ศิลป์-ภาษา	ภาษาอังกฤษ	3.41	ไทย	พุทธศาสนา	F	อมก๋อย	อ

ภาพที่ 4.3 ตัวอย่างคุณลักษณะที่สามารถบ่งบอกความหมายเดียวกัน

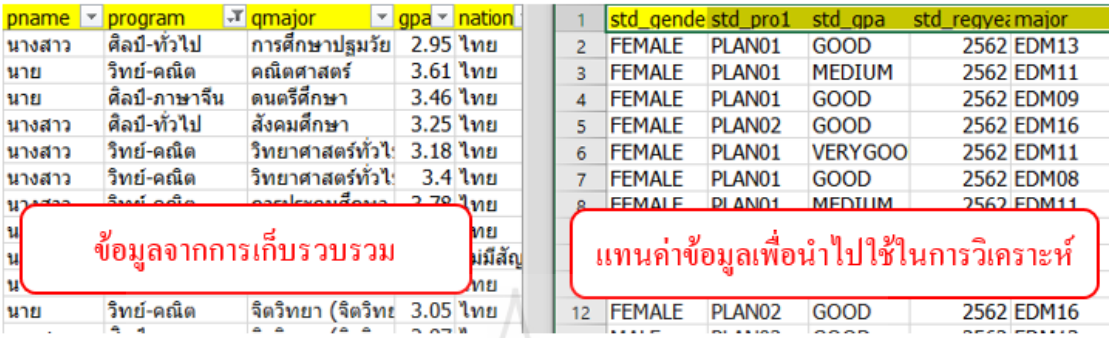
จากภาพที่ 4.3 แสดงข้อมูลคุณลักษณะที่มีความซ้ำซ้อนกัน ระหว่างตัวแปรคุณลักษณะ pname ที่เก็บค่านายและนางสาว กับตัวแปร Studentsex ตัวแปรที่เก็บค่าเพศชายและเพศหญิง เมื่อพิจารณาแล้วคัดเลือกคุณลักษณะทั้งสองสามารถบ่งบอกเพศได้จึงนำมาใช้วิเคราะห์เพียงหนึ่งคุณลักษณะ



ภาพที่ 4.4 ตัวอย่างคุณลักษณะที่อยู่

จากภาพที่ 4.4 แสดงข้อมูลคุณลักษณะของที่อยู่โดยเก็บข้อมูลตำบล อำเภอ จังหวัดและรหัสไปรษณีย์ ในการวิจัยนี้เป็นงานวิจัยเชิงพัฒนาเพื่อเป็นต้นแบบผู้วิจัยศึกษาข้อมูลของกลุ่มนักศึกษาที่อยู่ภายในจังหวัดเชียงใหม่และจังหวัดแม่ฮ่องสอน ซึ่งเป็นวิทยาเขตของมหาวิทยาลัยภัฏเชียงใหม่ จึงมุ่งเน้นที่จะนำข้อมูลของกลุ่มผู้สมัครที่อยู่ในจังหวัดและต่างจังหวัดมาเป็นตัวแบบในการวิเคราะห์เพื่อให้ได้ต้นแบบและแนวทางในการพัฒนางานวิจัยในลำดับต่อไป ผู้วิจัยจึงได้คัดเลือกคุณลักษณะที่ระบุจังหวัด และตัดคุณลักษณะตำบล อำเภอ และรหัสไปรษณีย์ออก จากนั้นนำคุณลักษณะมาจัดกลุ่มตามกลุ่มของผู้สมัครในพื้นที่ด้วยการแทนค่าให้ความหมายสำหรับผู้สมัครที่อาศัยอยู่ในจังหวัดเชียงใหม่จังหวัดแม่ฮ่องสอน และกลุ่มที่อยู่ต่างจังหวัดเพื่อที่จะนำไปวิเคราะห์ในขั้นตอนต่อไป

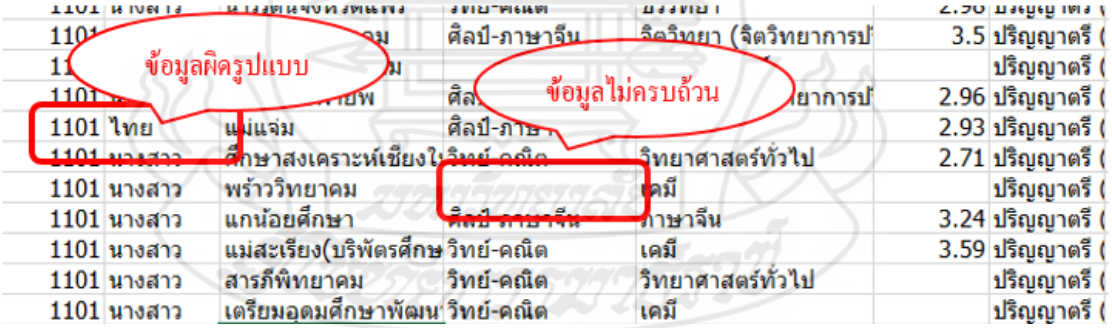
การแทนค่าข้อมูลถือเป็นกระบวนการที่อยู่ในขั้นตอนการเตรียมข้อมูล จากการศึกษาคุณลักษณะที่มีความคล้ายกันและคัดเลือกคุณลักษณะที่จะนำมาวิเคราะห์แล้ว ขั้นตอนต่อไปผู้วิจัยศึกษาข้อมูลที่เก็บภายในแต่ละคุณลักษณะเพื่อจัดกลุ่มข้อมูลและแปลงค่าข้อมูลให้อยู่ในรูปแบบที่จะนำไปเข้าสู่กระบวนการต่อไป โดยการแทนค่าข้อมูลที่เก็บในรูปแบบคำอธิบาย หมายถึง ข้อมูลที่เก็บในแต่ละรายการ ตัวอย่างเช่นการเก็บข้อมูลเพศจากข้อมูลเดิมเก็บในรูปแบบตัวอักษร “M” แทนค่าข้อมูลเพศชาย และตัวอักษร “F” แทนค่าข้อมูลเพศหญิง การเก็บข้อมูลสายวิชาที่จบจากระดับมัธยมมีค่าข้อมูล “วิทย์-คณิต” “ศิลป์-ภาษาจีน” “ศิลป์-ภาษา” เมื่อพิจารณาแล้วนำมาจัดกลุ่มและแปลงค่าข้อมูลเพื่อให้มีความเหมาะสมและพร้อมที่จะนำไปเข้าสู่โปรแกรมประมวลผล จึงได้แทนค่าให้กับข้อมูลเดิม ตัวอย่างเช่น จากเดิมข้อมูลคือ “วิทย์-คณิต” แทนค่าข้อมูลใหม่เป็น “PLAN01” ซึ่งหมายถึงกลุ่มผู้สมัครที่จบจากระดับมัธยมสายวิทย์-คณิต ดังได้แสดงในตารางที่ 3.2 ตารางแสดงรายละเอียดคุณลักษณะของข้อมูลและการแปลงค่าข้อมูล ในบทที่ 3 วิธีการดำเนินการวิจัย



ภาพที่ 4.5 ตัวอย่างการแทนค่าข้อมูล

จากภาพที่ 4.5 แสดงตัวอย่างการแทนค่าข้อมูลเพื่อนำไปใช้ในการวิเคราะห์ ด้านซ้ายคือข้อมูลที่ได้จากการเก็บรวบรวม ด้านขวาคือข้อมูลที่จัดกลุ่มและแทนค่าข้อมูลแล้ว ซึ่งเป็นข้อมูลที่พร้อมที่จะนำเข้าสู่โปรแกรม Rapid Miner ที่เป็นเครื่องมือในการวิเคราะห์ข้อมูล

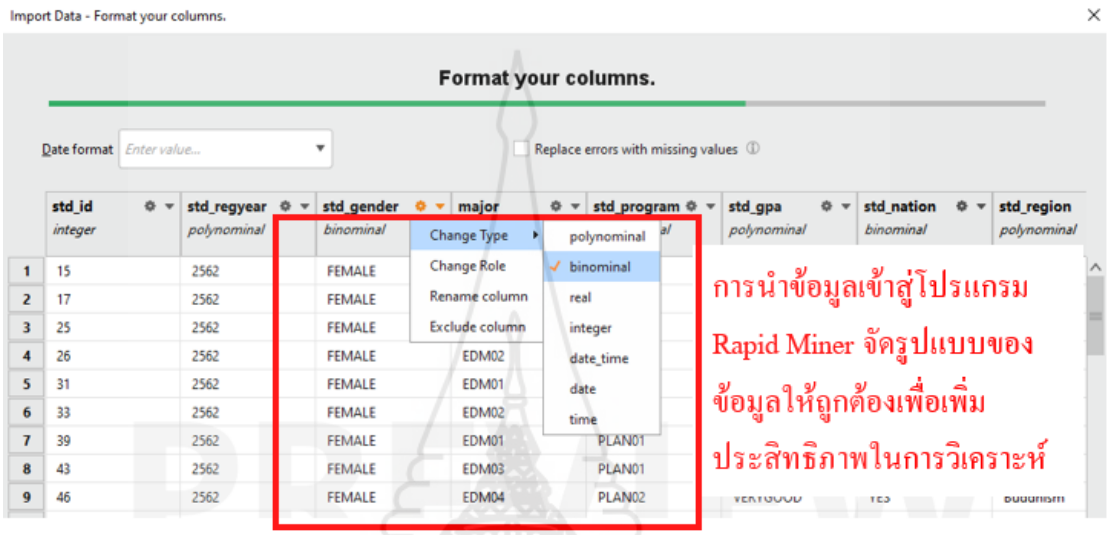
ขั้นตอนที่สำคัญในการนำข้อมูลไปใช้ในการวิเคราะห์คือการทำความสะอาดข้อมูล เพื่อให้การวิเคราะห์และการทำนายข้อมูลที่มีความแม่นยำ ข้อมูลที่อยู่ในรูปแบบที่ผิดหรือรูปแบบที่ไม่ครบถ้วน เมื่อนำไปวิเคราะห์จะทำให้ค่าผลลัพธ์ออกมาไม่มีความแม่นยำ ในขั้นตอนพิจารณาข้อมูลเพื่อนำข้อมูลมาจัดกลุ่มและแปลงข้อมูลสามารถตรวจสอบข้อมูลที่อยู่ในรูปแบบที่ไม่ถูกต้องหรือไม่ครบถ้วนจะต้องลบรายการข้อมูลนั้นออก ดังภาพที่ 4.6 แสดงรายการที่ข้อมูลที่ไม่ถูกต้องและข้อมูลไม่ครบถ้วน



ภาพที่ 4.6 ทำความสะอาดข้อมูล

ข้อมูลที่ได้จากการเตรียมข้อมูลถือเป็นข้อมูลเบื้องต้น เมื่อนำข้อมูลเข้าสู่โปรแกรมแล้วภายหลังจากการรันโปรแกรมหากพบว่าค่าความถูกต้องหรือการจำแนกข้อมูลออกมาแล้วมีประสิทธิภาพน้อยจะต้องกลับมาพิจารณาการจัดกลุ่มของข้อมูลและแก้ไขเพื่อให้ได้ประสิทธิภาพออกมาดีที่สุด ซึ่งอยู่ในกระบวนการเหมืองข้อมูลของ Cross – Industry Standard Process for Data Mining หรือ CRISP-DM

จากข้อมูลที่ได้รวบรวม ทำความสะอาดข้อมูลและคัดเลือกคุณลักษณะ ที่จะนำมาวิเคราะห์ จะได้ข้อมูลที่จะนำมาสู่กระบวนการต่อไปคือการนำข้อมูลเข้าสู่โปรแกรม Rapid Miner ซึ่งต้องเลือกประเภทข้อมูลที่จะนำมาวิเคราะห์ให้ถูกต้องและสอดคล้องกับประเภทของข้อมูลภายในโปรแกรมเพื่อให้ผลการวิเคราะห์มีความถูกต้องและมีประสิทธิภาพในการประมวลผล



ภาพที่ 4.7 การเลือกประเภทของข้อมูล

จากภาพที่ 4.7 แสดงการเลือกประเภทของข้อมูลให้ตรงกับคุณลักษณะของข้อมูล โดยประเภท polynominal หมายถึงประเภทข้อมูลที่เป็นตัวอักษรและมีค่ามากกว่า 2 ค่า และประเภท binominal หมายถึงประเภทข้อมูลที่เป็นตัวอักษร มีค่าเพียง 2 ค่าเท่านั้น ตัวอย่างเช่นคุณลักษณะเพศ มีค่า “MALE” และ “FEMALE” จะเลือกข้อมูลเป็นประเภท binominal เป็นต้น

std_regyear Category	std_gender Category	std_program Category	std_gpa Category	std_nation Category	std_region Category
2563	MALE	PLAN01	MEDIUM	YES	Buddhism
2563	MALE	PLAN01	MEDIUM	YES	Buddhism
2563	FEMALE	PLAN01	VERYGOOD	YES	Buddhism
2563	FEMALE	PLAN01	GOOD	YES	Buddhism
2564	FEMALE	PLAN02	GOOD	YES	Buddhism
2564	MALE	PLAN01	VERYGOOD	YES	Buddhism
2564	MALE	PLAN01	GOOD	YES	Buddhism

ภาพที่ 4.8 ข้อมูลในโปรแกรม Rapid Miner

จากภาพที่ 4.8 แสดงข้อมูลที่นำเข้าโปรแกรม Rapid Miner โดยข้อมูลคุณลักษณะแสดงในรูปของ Category หมายถึงข้อมูลที่เป็นกลุ่มตัวอักษรที่สามารถนับจำนวนได้แต่ไม่สามารถคำนวณจากข้อมูลได้ ซึ่งข้อมูลนี้เป็นข้อมูลที่พร้อมจะนำไปสู่กระบวนการทำเหมืองข้อมูลโดยใช้โอเปอร์เรเตอร์ของโปรแกรมด้วยเทคนิคที่ได้คัดเลือกมาใช้ในการวิเคราะห์

ขั้นตอนการคัดเลือกคุณลักษณะ ผู้วิจัยได้นำข้อมูลที่ได้จากการเตรียมข้อมูลเข้าสู่กระบวนการคัดเลือกคุณลักษณะที่สำคัญและมีความสัมพันธ์กับการเลือกสมัครเรียน โดยใช้เทคนิคที่มีความเหมาะสมกับลักษณะของข้อมูลที่จะนำมาใช้ในงานวิจัย พบว่าเทคนิค Information Gain มีความเหมาะสมมากที่สุด เพื่อใช้ในการหาค่าน้ำหนักของคุณลักษณะของข้อมูลเพื่อนำไปใช้เป็นคุณลักษณะที่จะจำแนกประเภทของข้อมูลให้มีความถูกต้องแม่นยำ

ขั้นตอนการจำแนกประเภทข้อมูล ผู้วิจัยได้ศึกษาและนำเทคนิคในการสร้างแบบจำลองที่มีความเหมาะสมกับลักษณะของข้อมูลที่นำมาวิจัยจำนวน 3 เทคนิคได้แก่การสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest) โดยนำค่าคุณลักษณะที่ได้จากการคัดเลือกด้วยเทคนิค Information Gain ในขั้นตอนการคัดเลือกคุณลักษณะมาใช้ร่วมกับแบบจำลอง และนำแบบจำลองมาเข้าสู่กระบวนการวัดประสิทธิภาพของแบบจำลอง ด้วยเทคนิคการวัดประสิทธิภาพแบบจำลอง ด้วยวิธี 5-fold Cross Validation วิธี 10-fold Cross Validation วิธี Split Validation (70:30) และ วิธี Split Validation (80:20) เพื่อประเมินประสิทธิภาพและให้ได้แบบจำลองที่มีประสิทธิภาพมากที่สุด โดยเปรียบเทียบจากค่าความถูกต้องโดยรวมของแบบจำลอง (Accuracy) ความแม่นยำของการทำนาย (Precision) ค่าความครบถ้วน (Recall) และค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure)

จากกระบวนการในกรอบแนวความคิดในการวิจัยข้างต้น ผู้วิจัยได้นำเสนอผลการวิเคราะห์ข้อมูลด้วยผลการวิเคราะห์ที่ได้จากการประมวลผลด้วยโปรแกรม Rapid Miner ดังนี้

1. ผลการวิเคราะห์ข้อมูลที่ใช้ในการวิจัย
2. ผลการวิเคราะห์ข้อมูลคุณลักษณะ
3. ผลการวิเคราะห์ประสิทธิภาพแบบจำลอง
4. ผลการทำนายข้อมูล

1. ผลการวิเคราะห์ข้อมูลที่ใช้ในการวิจัย

การวิจัยนี้ผู้วิจัยได้ศึกษาและรวบรวมข้อมูลจากสำนักทะเบียนและประมวลผลของมหาวิทยาลัยราชภัฏเชียงใหม่ ซึ่งเป็นเจ้าของและดูแลข้อมูลการรับสมัครนักศึกษา โดยได้ขอความอนุเคราะห์นำข้อมูลพื้นฐานและความสัมพันธ์ด้านครอบครัวของผู้สมัคร ตลอดจนข้อมูลที่สามารถเปิดเผยนำมาวิเคราะห์ โดยได้ดำเนินการวิเคราะห์ข้อมูลเป็น 2 ส่วนดังนี้

1.1 ผลการเตรียมข้อมูลจากการคัดเลือกคุณลักษณะ จากข้อมูลที่ได้รวบรวมมีจำนวนทั้งหมด 3,364 รายการ จำนวนคุณลักษณะทั้งหมด 26 คุณลักษณะ ประกอบด้วยลำดับข้อมูลดังนี้ 1) รหัสการรับนักศึกษา 2) รหัสประเภทการรับสมัคร 3) คำนำหน้าชื่อ 4) ชื่อโรงเรียนเดิม 6) สายวิชาที่เรียน 7) หลักสูตรที่เรียน 8) เกรดเฉลี่ย 9) กลุ่มชั้นเรียนที่ศึกษา 10) สถานการณ์เป็นนักศึกษาในปัจจุบัน 11) สัญชาติ 12) ศาสนา 13) เพศ 14) ตำบล 15) อำเภอ 16)จังหวัด 17) รหัสไปรษณีย์ที่พักอาศัย 18)สัญชาติบิดา 19) สัญชาติมารดา 20) สถานภาพครอบครัว 21) สถานะการมีชีวิตของบิดา 22) สถานะการมีชีวิตของมารดา 23) ผู้ปกครอง 24) กลุ่มการประกอบอาชีพของบิดา 25) กลุ่มการประกอบอาชีพของมารดา และ 26)กลุ่มการประกอบอาชีพของผู้ปกครอง

โดยได้ดำเนินการในขั้นตอนการเตรียมข้อมูลดังได้อธิบายไว้ในบทที่ 3 วิธีการดำเนินวิจัย เริ่มจากการคัดกรองข้อมูลที่มีลักษณะคล้ายกันและลักษณะข้อมูลที่มีความซ้ำซ้อนกันออก ตัวอย่างเช่น ข้อมูลคำนำหน้านามและข้อมูลเพศ ซึ่งคำนำหน้านามของเพศชายคือนาย และคำนำหน้านามของเพศหญิงคือนางสาว ซึ่งข้อมูลเพศสามารถที่จะระบุคำนำหน้านามได้อย่างชัดเจน จึงเลือกตัดข้อมูลคำนำหน้านามคงเหลือไว้เพียงข้อมูลเพศ ข้อมูลตำบล และอำเภอ ในการวิจัยนี้ศึกษาข้อมูลของกลุ่มนักศึกษาที่อยู่ภายในจังหวัดเชียงใหม่และจังหวัดแม่ฮ่องสอน ซึ่งจังหวัดแม่ฮ่องสอนมีวิทยาลัยที่เป็นวิทยาเขตของมหาวิทยาลัยภัฏเชียงใหม่ จึงมุ่งเน้นที่จะนำข้อมูลของกลุ่มผู้สมัครที่อยู่ในจังหวัดและต่างจังหวัด จากการคัดเลือกคุณลักษณะที่นำมาศึกษาทั้งหมด ได้ทำการตัดคุณลักษณะที่มีลักษณะข้อมูลที่มีความซ้ำซ้อนกันและข้อมูลที่ไม่ครบถ้วน เพื่อเป็นการทำความสะอาดข้อมูลและเพื่อให้ได้คุณลักษณะที่สำคัญไปใช้ในการวิเคราะห์ที่มีความถูกต้องสูงที่สุด ดังนั้น ข้อมูลที่ได้รวบรวมมาจากเดิม 3,364 รายการ ภายหลังจากการทำความสะอาดข้อมูลแล้วคงเหลือข้อมูลจำนวน 3,195 รายการ และคุณลักษณะของข้อมูลจากเดิมทั้งหมดจำนวน 26 คุณลักษณะ คงเหลือคุณลักษณะที่จะนำไปใช้ในงานวิจัยนี้จำนวน 17 คุณลักษณะ ประกอบด้วย 1) ปีที่สมัคร 2) เพศ 3) สาขาที่สมัคร 4) แผนการเรียนที่จบจากระดับมัธยม 5) เกรดเฉลี่ย 6) สัญชาติ 7) ศาสนา 8) สัญชาติบิดา 9) สัญชาติมารดา 10) สถานะการมีชีวิตบิดา 11) สถานะการมีชีวิตมารดา 12) สถานะครอบครัว 13) จังหวัด 14) ผู้ปกครอง 15) อาชีพผู้ปกครอง 16) อาชีพบิดา 17) อาชีพมารดา

1.2 ผลการเตรียมข้อมูลที่ใช้ในการวิจัย ภายหลังจากการคัดเลือกคุณลักษณะและการทำความสะอาดข้อมูลที่ได้จากการรวบรวมข้อมูลในขั้นตอนที่ 1.1 ข้อมูลทั้งหมดมีจำนวน 3,195 รายการ จำนวนคุณลักษณะ 17 คุณลักษณะ การวิจัยนี้ผู้วิจัยได้ศึกษาข้อมูลที่จะใช้ในการวิจัย คือ ข้อมูลผู้เข้าสมัครทุกรอบ ผู้สมัครที่ผ่านการคัดเลือกและผู้สมัครที่ยืนยันการลงทะเบียนเรียนแล้ว เฉพาะการเรียนภาคปกติ ในปีการศึกษา 2562- 2564 จำนวน 4 หลักสูตรของคณะครุศาสตร์ โดยนำข้อมูลหลักสูตรที่อยู่ในการกำกับดูแลของคณะครุศาสตร์โดยตรง ประกอบด้วย 1.หลักสูตรการศึกษาปฐมวัย 2.หลักสูตรการประถมศึกษา 3.หลักสูตรพลศึกษา 4.หลักสูตรการศึกษาพิเศษ โดยคัดเลือกข้อมูลที่จะนำมาใช้ในการวิเคราะห์คงเหลือจำนวนทั้งสิ้น 684 รายการ โดยใช้คุณลักษณะจำนวน 17 คุณลักษณะ ดังนี้ 1) ปีที่สมัคร 2) เพศ 3) สาขาที่สมัคร 4) แผนการเรียนที่จบจากระดับมัธยม 5) เกรดเฉลี่ย 6) สัญชาติ 7) ศาสนา 8) สัญชาติบิดา 9) สัญชาติมารดา 10) สถานะการมีชีวิตบิดา 11)

สถานะการมีชีวิตมารดา 12) สถานะครอบครัว 13) จังหวัด 14) ผู้ปกครอง 15) อาชีพผู้ปกครอง 16) อาชีพบิดา 17) อาชีพมารดา

2. ผลการวิเคราะห์ข้อมูลคุณลักษณะ

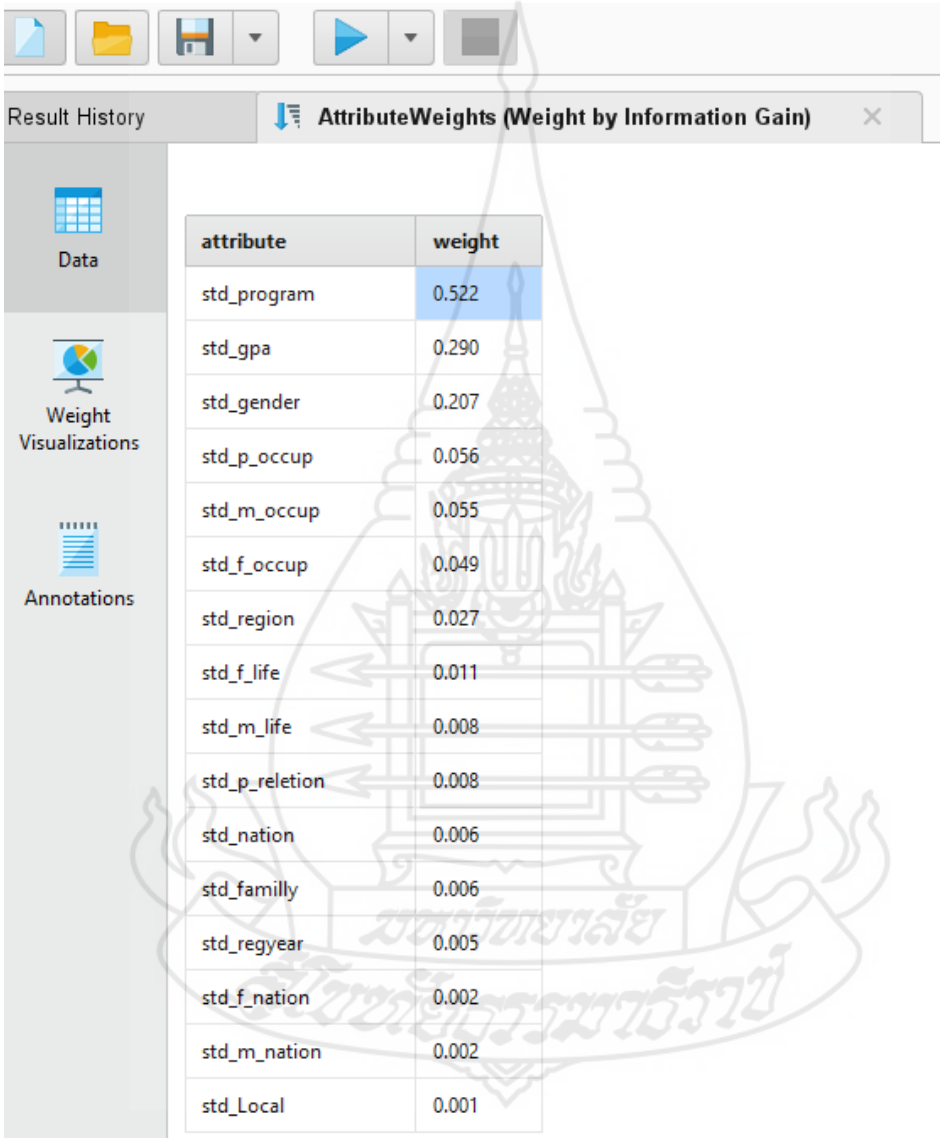
การวิจัยนี้นำข้อมูลคุณลักษณะดังได้อธิบายการคัดเลือกและเตรียมข้อมูลไว้ในตอนที่ 1 นำมาคัดเลือกคุณลักษณะ (Feature Selection) ด้วยเทคนิค Information Gain เพื่อหาค่าน้ำหนักของคุณลักษณะที่มีความสำคัญมีความสัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ โดยเทคนิค Information Gain จะนำคุณลักษณะทั้งหมดมาคำนวณความสัมพันธ์กับคุณลักษณะเป้าหมายคือสาขาวิชาที่สมัคร ซึ่งคุณลักษณะเป้าหมายจะไม่ถูกนำมาหาค่าน้ำหนัก เมื่อนำคุณลักษณะที่ได้จากการเตรียมข้อมูลจำนวน 17 คุณลักษณะนำมาคำนวณค่าจะแสดงผลในการหาค่าจำนวน 16 คุณลักษณะ ผลค่าน้ำหนัก คุณลักษณะที่มีความสำคัญจะมีค่าน้ำหนักมากและคุณลักษณะที่สำคัญน้อยจะมีค่าน้ำหนักน้อย ดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 ตารางแสดงข้อมูลคุณลักษณะ

ลำดับ	คุณลักษณะข้อมูล	ชื่อตัวแปร	ค่าน้ำหนัก
1	แผนการเรียนที่จบจากระดับมัธยม	std_program	0.522
2	เกรดเฉลี่ย	std_gpa	0.290
3	เพศ	std_gender	0.207
4	อาชีพผู้ปกครอง	std_p_occup	0.056
5	อาชีพมารดา	std_m_occup	0.055
6	อาชีพบิดา	std_f_occup	0.049
7	สัญชาติผู้สมัครเรียน	std_region	0.027
8	บิดามีชีวิต	std_f_life	0.011
9	มารดามีชีวิต	std_m_life	0.008
10	ผู้ปกครอง	std_p_reletion	0.008
11	สัญชาติผู้สมัครเรียน	std_nation	0.006
12	สถานะครอบครัว	std_familly	0.006
13	ปีที่สมัคร	std_regyear	0.005
14	สัญชาติบิดา	std_f_nation	0.002
15	สัญชาติมารดา	std_m_nation	0.002
16	จังหวัด	std_Local	0.001

จากตารางที่ 4.1 คุณลักษณะที่นำมาวิเคราะห์ จำนวน 17 คุณลักษณะ เมื่อนำมาคำนวณหาค่าน้ำหนัก โดยคุณลักษณะเป้าหมาย คือ สาขาวิชา จะไม่นำมาคำนวณ คงเหลือจำนวน 16 คุณลักษณะ แสดงตารางลำดับคุณลักษณะที่มีค่าน้ำหนักมากที่สุดลงมาตามลำดับจนถึงคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุด ดังนี้ ลำดับที่ 1 แผนการเรียนที่จบจากระดับมัธยม มีค่าน้ำหนักที่ 0.522 ลำดับที่ 2 เกรดเฉลี่ย มีค่าน้ำหนักที่ 0.290 ลำดับที่ 3 เพศ มีค่าน้ำหนักที่ 0.207 ลำดับที่ 4 อาชีพผู้ปกครอง มีค่าน้ำหนักที่ 0.056 ลำดับที่ 5 อาชีพมารดา มีค่าน้ำหนักที่ 0.055 ลำดับที่ 6 อาชีพบิดา มีค่าน้ำหนักที่ 0.049

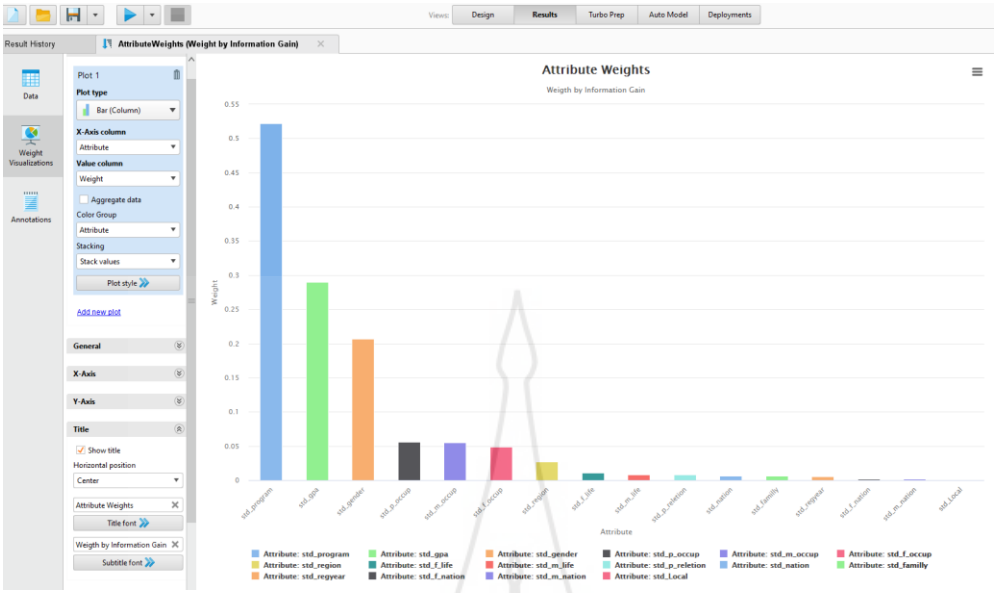
ลำดับที่ 7 สัญชาติผู้สมัครเรียน มีค่าน้ำหนักที่ 0.027 ลำดับที่ 8 บิดามีชีวิต มีค่าน้ำหนักที่ 0.011 ลำดับที่ 9 มารดามีชีวิต มีค่าน้ำหนักที่ 0.008 ลำดับที่ 10 ผู้ปกครอง มีค่าน้ำหนักที่ 0.008 ลำดับที่ 11 สัญชาติผู้สมัครเรียน มีค่าน้ำหนักที่ 0.006 ลำดับที่ 12 สถานะครอบครัว มีค่าน้ำหนักที่ 0.006 ลำดับที่ 13 ปีที่สมัคร มีค่าน้ำหนักที่ 0.005 ลำดับที่ 14 สัญชาติบิดา มีค่าน้ำหนักที่ 0.002 ลำดับที่ 15 สัญชาติมารดา มีค่าน้ำหนักที่ 0.002 และคุณลักษณะที่มีค่าน้ำหนักน้อยที่สุดคือ จังหวัด มีค่าน้ำหนักที่ 0.001



attribute	weight
std_program	0.522
std_gpa	0.290
std_gender	0.207
std_p_occup	0.056
std_m_occup	0.055
std_f_occup	0.049
std_region	0.027
std_f_life	0.011
std_m_life	0.008
std_p_reletion	0.008
std_nation	0.006
std_familly	0.006
std_regyear	0.005
std_f_nation	0.002
std_m_nation	0.002
std_Local	0.001

ภาพที่ 4.9 ผลค่าน้ำหนักของคุณลักษณะ

จากภาพที่ 4.9 ผลลัพธ์จากการหาค่าน้ำหนักของคุณลักษณะ เมื่อนำคุณลักษณะจำนวนทั้งหมด 17 คุณลักษณะและตั้งค่าคุณลักษณะเป้าหมาย(สาขาวิชา) เพื่อหาค่าน้ำหนักความสัมพันธ์กัน คงเหลือคุณลักษณะที่นำมาคำนวณด้วยเทคนิค Information Gain จำนวน 16 คุณลักษณะโดยใช้โปรแกรม Rapid Minor แสดงผลในรูปแบบของตาราง



ภาพที่ 4.10 กราฟแสดงค่าน้ำหนักของคุณลักษณะ

จากภาพที่ 4.10 ผลลัพธ์จากการหาค่าน้ำหนักของคุณลักษณะ เมื่อนำคุณลักษณะจำนวนทั้งหมด 17 คุณลักษณะและตั้งค่าคุณลักษณะเป้าหมาย(สาขาวิชา) เพื่อหาค่าน้ำหนักความสัมพันธ์กันคงเหลือคุณลักษณะที่นำมาคำนวณด้วยเทคนิค Information Gain จำนวน 16 คุณลักษณะโดยใช้โปรแกรม Rapid Minor แสดงผลในรูปแบบกราฟ สามารถอธิบายได้ดังนี้

ลำดับที่ 1 ตัวแปร std_program หมายถึงแผนการเรียนที่จบจากระดับมัธยม มีค่าน้ำหนักที่ 0.522

ลำดับที่ 2 ตัวแปร std_gdp หมายถึงเกรดเฉลี่ย มีค่าน้ำหนักที่ 0.290

ลำดับที่ 3 ตัวแปร std_gender หมายถึงเพศ มีค่าน้ำหนักที่ 0.207

ลำดับที่ 4 ตัวแปร std_p_occup หมายถึงอาชีพผู้ปกครอง มีค่าน้ำหนักที่ 0.056

ลำดับที่ 5 ตัวแปร std_m_occup หมายถึงอาชีพมารดา มีค่าน้ำหนักที่ 0.055

ลำดับที่ 6 ตัวแปร std_f_occup หมายถึงอาชีพบิดา มีค่าน้ำหนักที่ 0.049

ลำดับที่ 7 ตัวแปร std region หมายถึงสัญชาติผู้สมัครเรียน มีค่าน้ำหนักที่ 0.027

ลำดับที่ 8 ตัวแปร std_f_life หมายถึงบิดามีชีวิต มีค่าน้ำหนักที่ 0.011

ลำดับที่ 9 ตัวแปร std_m_life หมายถึงมารดามีชีวิต มีค่าน้ำหนักที่ 0.008

ลำดับที่ 10 ตัวแปร std_p_relation หมายถึงผู้ปกครอง มีค่าน้ำหนักที่ 0.008

ลำดับที่ 11 ตัวแปร std_nation หมายถึงสัญชาติผู้สมัครเรียน มีค่าน้ำหนักที่ 0.006

ลำดับที่ 12 ตัวแปร std_family หมายถึงสถานะครอบครัว มีค่าน้ำหนักที่ 0.006

ลำดับที่ 13 ตัวแปร std_regyear หมายถึงปีที่สมัคร มีค่าน้ำหนักที่ 0.005

ลำดับที่ 14 ตัวแปร std_f_nation หมายถึงสัญชาติบิดา มีค่าน้ำหนักที่ 0.002

ลำดับที่ 15 ตัวแปร std_m_nation หมายถึงสัญชาติมารดา มีค่าน้ำหนักที่ 0.002

ลำดับที่ 16 ตัวแปร std_local หมายถึงจังหวัด มีค่าน้ำหนักที่ 0.001

3. ผลการวิเคราะห์ประสิทธิภาพแบบจำลอง

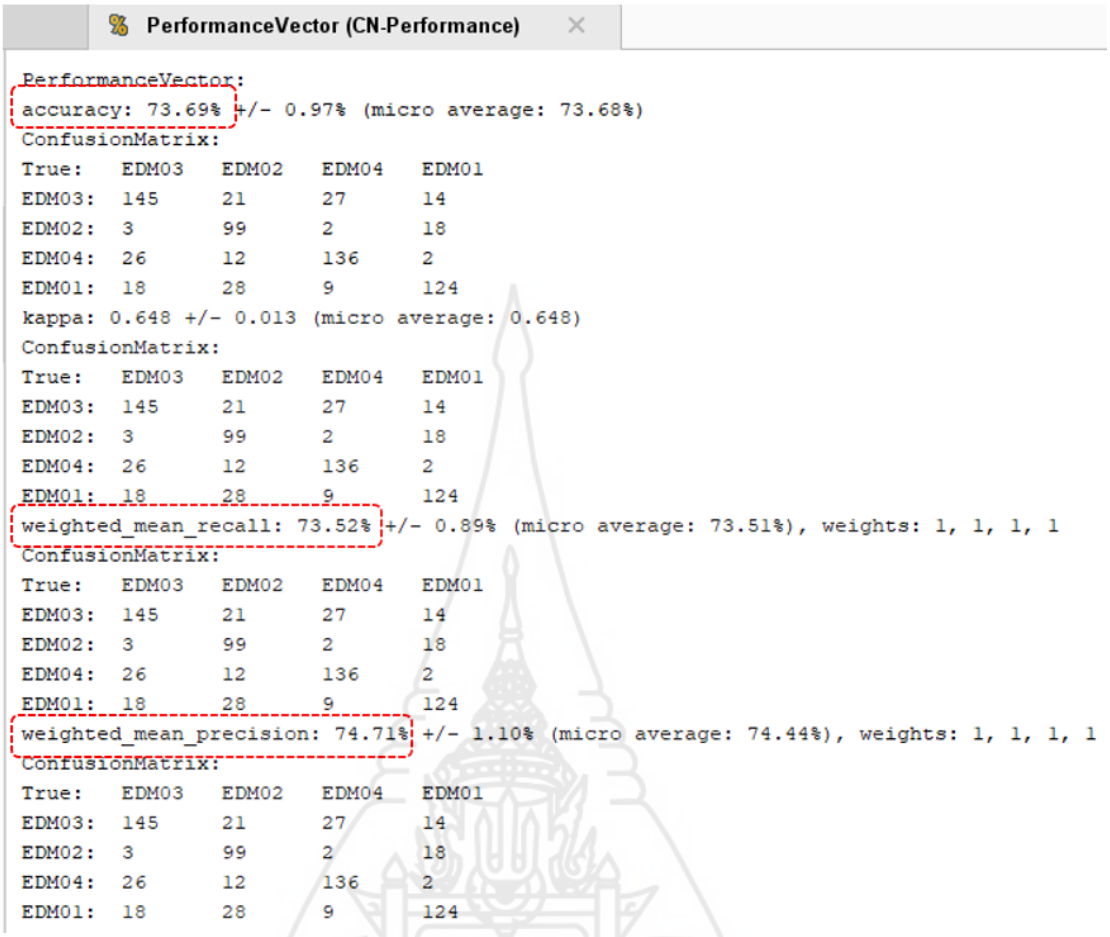
การวิจัยนี้นำผลจากการคัดเลือกคุณลักษณะสำคัญที่มีความสัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่มาสร้างแบบจำลองจำแนกประเภทข้อมูลและการทำนายด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest) และวัดประสิทธิภาพแบบจำลองทั้งหมดด้วยกระบวนการวัดประสิทธิภาพแบบจำลอง 4 วิธีได้แก่ 1) วิธี 5-fold Cross Validation 2)วิธี 10-fold Cross Validation 3)วิธี Split Validation (70:30) และ 4) วิธี Split Validation (80:20) เพื่อประเมินประสิทธิภาพแบบจำลองด้วยเทคนิคต่าง ๆ จากค่าความถูกต้องโดยรวมของแบบจำลอง (Accuracy) ความแม่นยำของการทำนาย (Precision) ค่าความครบถ้วน (Recall) และค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure)

PerformanceVector (CD-Performance)

```
PerformanceVector:
accuracy: 73.53% +/- 3.52% (micro average: 73.54%)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     11      19      8
EDM02:  11      109     6       30
EDM04:  31      13      140     2
EDM01:  14      27      9       118
kappa: 0.647 +/- 0.046 (micro average: 0.647)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     11      19      8
EDM02:  11      109     6       30
EDM04:  31      13      140     2
EDM01:  14      27      9       118
weighted_mean_recall: 73.54% +/- 3.06% (micro average: 73.53%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     11      19      8
EDM02:  11      109     6       30
EDM04:  31      13      140     2
EDM01:  14      27      9       118
weighted_mean_precision: 73.58% +/- 2.98% (micro average: 73.38%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     11      19      8
EDM02:  11      109     6       30
EDM04:  31      13      140     2
EDM01:  14      27      9       118
```

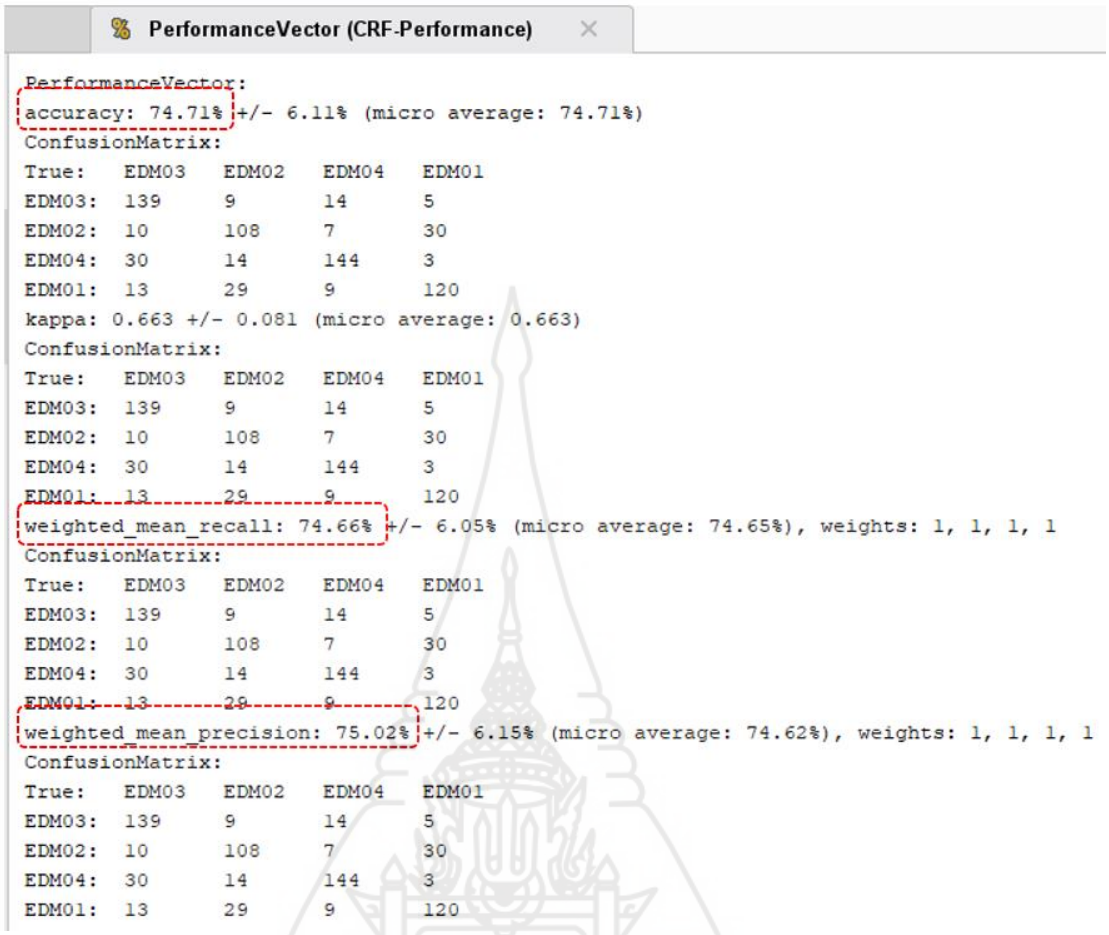
ภาพที่ 4.11 ผลทดสอบแบบจำลองเทคนิคต้นไม้ตัดสินใจด้วยวิธี 5-folds Cross - Validation

จากภาพที่ 4.11 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 73.53 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 73.54 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 73.58



ภาพที่ 4.12 ผลทดสอบแบบจำลองเทคนิคนาอีฟเบย์ด้วยวิธี 5-folds Cross-Validation

จากภาพที่ 4.12 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 73.69 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 73.52 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 74.71



PerformanceVector (CRF-Performance)

```
PerformanceVector:
accuracy: 74.71% +/- 6.11% (micro average: 74.71%)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     9       14      5
EDM02:  10      108     7       30
EDM04:  30      14      144     3
EDM01:  13      29      9       120
kappa: 0.663 +/- 0.081 (micro average: 0.663)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     9       14      5
EDM02:  10      108     7       30
EDM04:  30      14      144     3
EDM01:  13      29      9       120
weighted_mean_recall: 74.66% +/- 6.05% (micro average: 74.65%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     9       14      5
EDM02:  10      108     7       30
EDM04:  30      14      144     3
EDM01:  13      29      9       120
weighted_mean_precision: 75.02% +/- 6.15% (micro average: 74.62%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     9       14      5
EDM02:  10      108     7       30
EDM04:  30      14      144     3
EDM01:  13      29      9       120
```

ภาพที่ 4.13 ผลทดสอบแบบจำลองเทคนิคป่าสุ่มด้วยวิธี 5-folds Cross-Validation

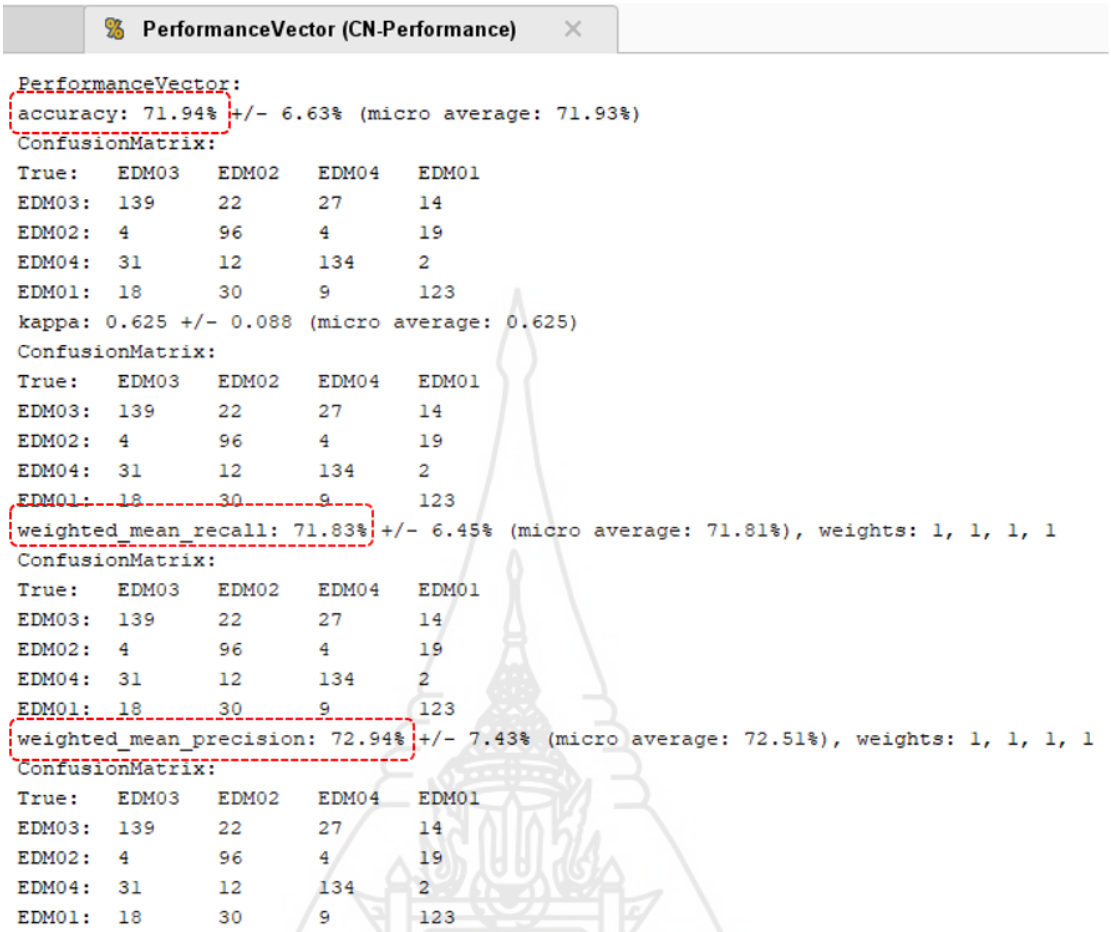
จากภาพที่ 4.13 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 74.71 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 74.66 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 75.02

PerformanceVector (CD-Performance)

```
PerformanceVector:
accuracy: 73.82% +/- 4.87% (micro average: 73.83%)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     9       18      6
EDM02:  12      113     7       34
EDM04:  32      14      140     2
EDM01:  12      24      9       116
kappa: 0.651 +/- 0.065 (micro average: 0.651)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     9       18      6
EDM02:  12      113     7       34
EDM04:  32      14      140     2
EDM01:  12      24      9       116
weighted_mean_recall: 73.82% +/- 4.67% (micro average: 73.83%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     9       18      6
EDM02:  12      113     7       34
EDM04:  32      14      140     2
EDM01:  12      24      9       116
weighted_mean_precision: 74.64% +/- 4.65% (micro average: 73.77%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  136     9       18      6
EDM02:  12      113     7       34
EDM04:  32      14      140     2
EDM01:  12      24      9       116
```

ภาพที่ 4.14 ผลทดสอบแบบจำลองเทคนิคต้นไม้ตัดสินใจด้วยวิธี 10-folds Cross - Validation

จากภาพที่ 4.14 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 73.82 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 74.64 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 75.02

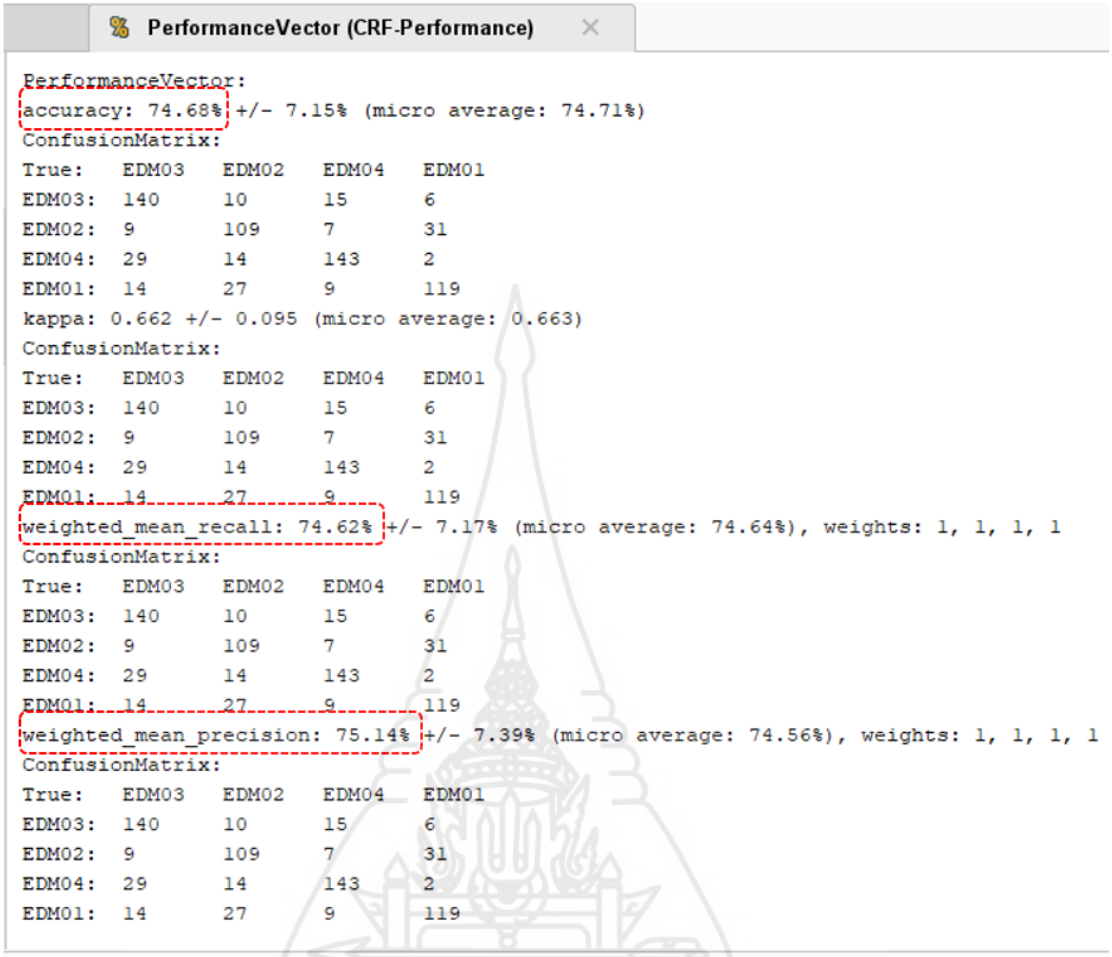


PerformanceVector (CN-Performance)

```
PerformanceVector:
accuracy: 71.94% +/- 6.63% (micro average: 71.93%)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     22      27      14
EDM02:  4       96      4       19
EDM04:  31      12      134     2
EDM01:  18      30      9       123
kappa: 0.625 +/- 0.088 (micro average: 0.625)
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     22      27      14
EDM02:  4       96      4       19
EDM04:  31      12      134     2
EDM01:  18      30      9       123
weighted_mean_recall: 71.83% +/- 6.45% (micro average: 71.81%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     22      27      14
EDM02:  4       96      4       19
EDM04:  31      12      134     2
EDM01:  18      30      9       123
weighted_mean_precision: 72.94% +/- 7.43% (micro average: 72.51%), weights: 1, 1, 1, 1
ConfusionMatrix:
True:   EDM03   EDM02   EDM04   EDM01
EDM03:  139     22      27      14
EDM02:  4       96      4       19
EDM04:  31      12      134     2
EDM01:  18      30      9       123
```

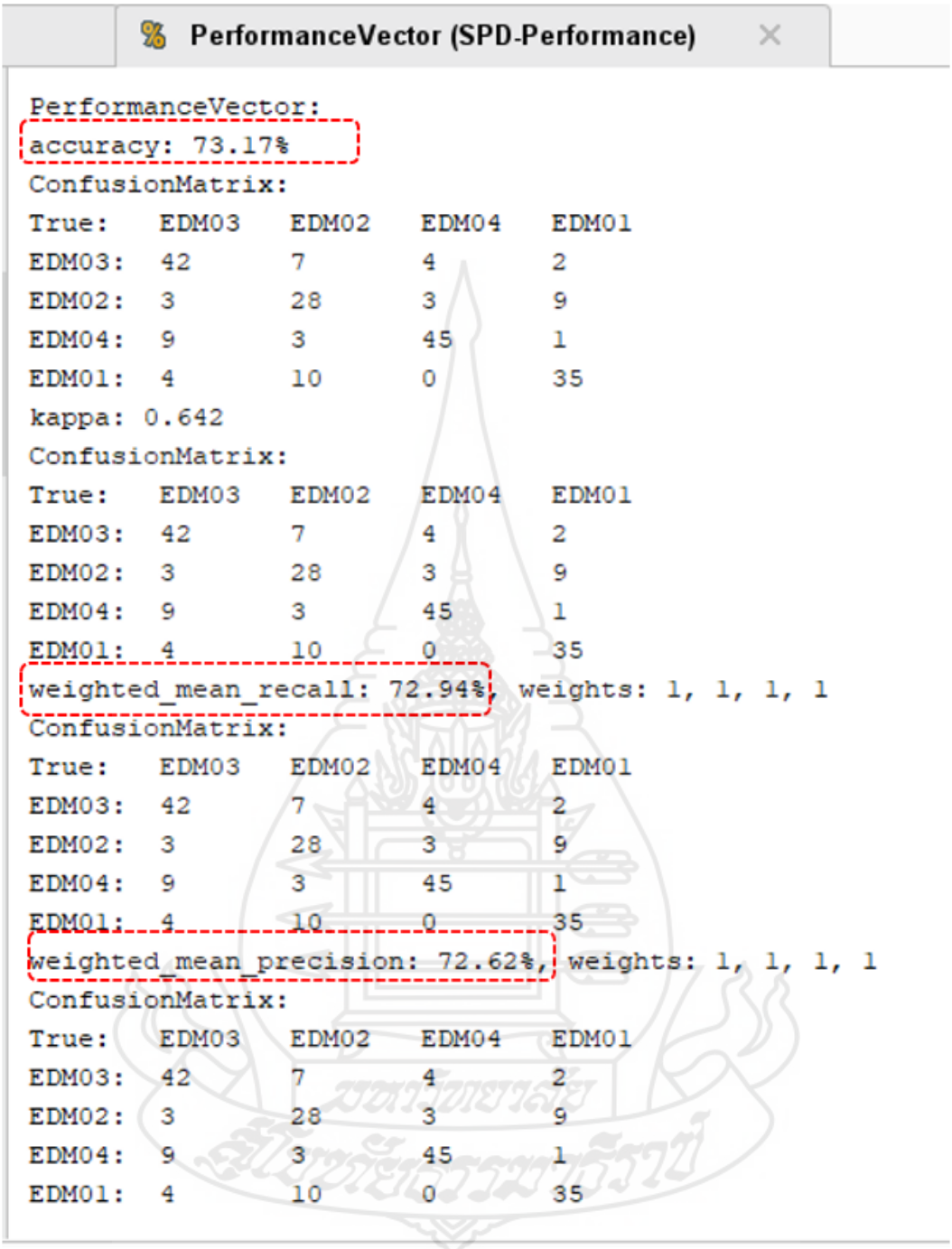
ภาพที่ 4.15 ผลทดสอบแบบจำลองเทคนิคนาอีฟเบย์ด้วยวิธี 10-folds Cross-Validation

จากภาพที่ 4.15 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 71.94 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 71.83 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 72.94



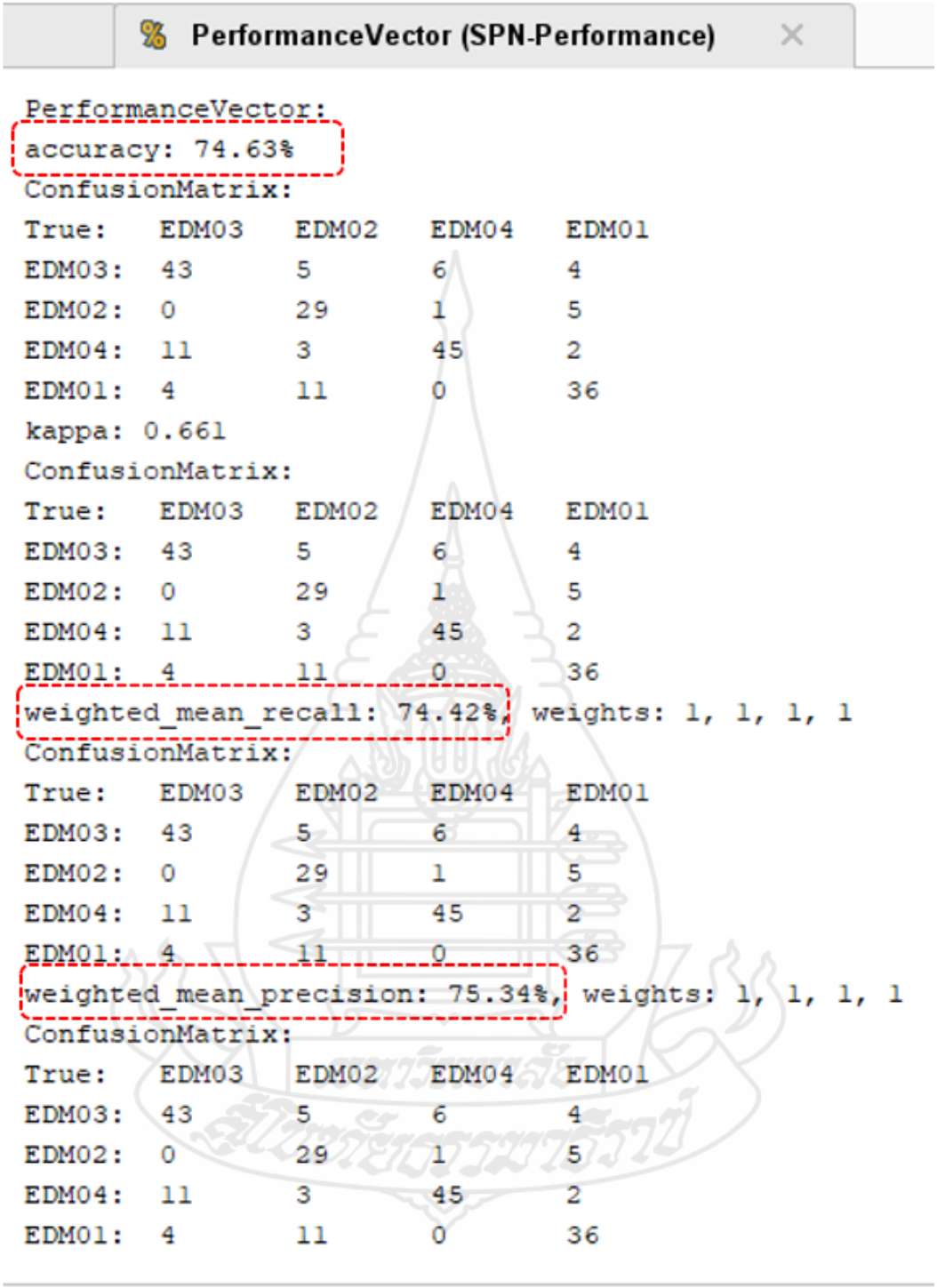
ภาพที่ 4.16 ผลทดสอบแบบจำลองเทคนิคป่าสุ่มด้วยวิธี 10-folds Cross-Validation

จากภาพที่ 4.16 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 74.68 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 74.62 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 75.14



ภาพที่ 4.17 ผลทดสอบแบบจำลองเทคนิคต้นไม้ตัดสินใจด้วยวิธี Split Test (70:30)

จากภาพที่ 4.17 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 73.17 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 72.94 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 72.62



ภาพที่ 4.18 ผลทดสอบแบบจำลองเทคนิคนาอีฟเบย์ด้วยวิธี Split Test (70:30)

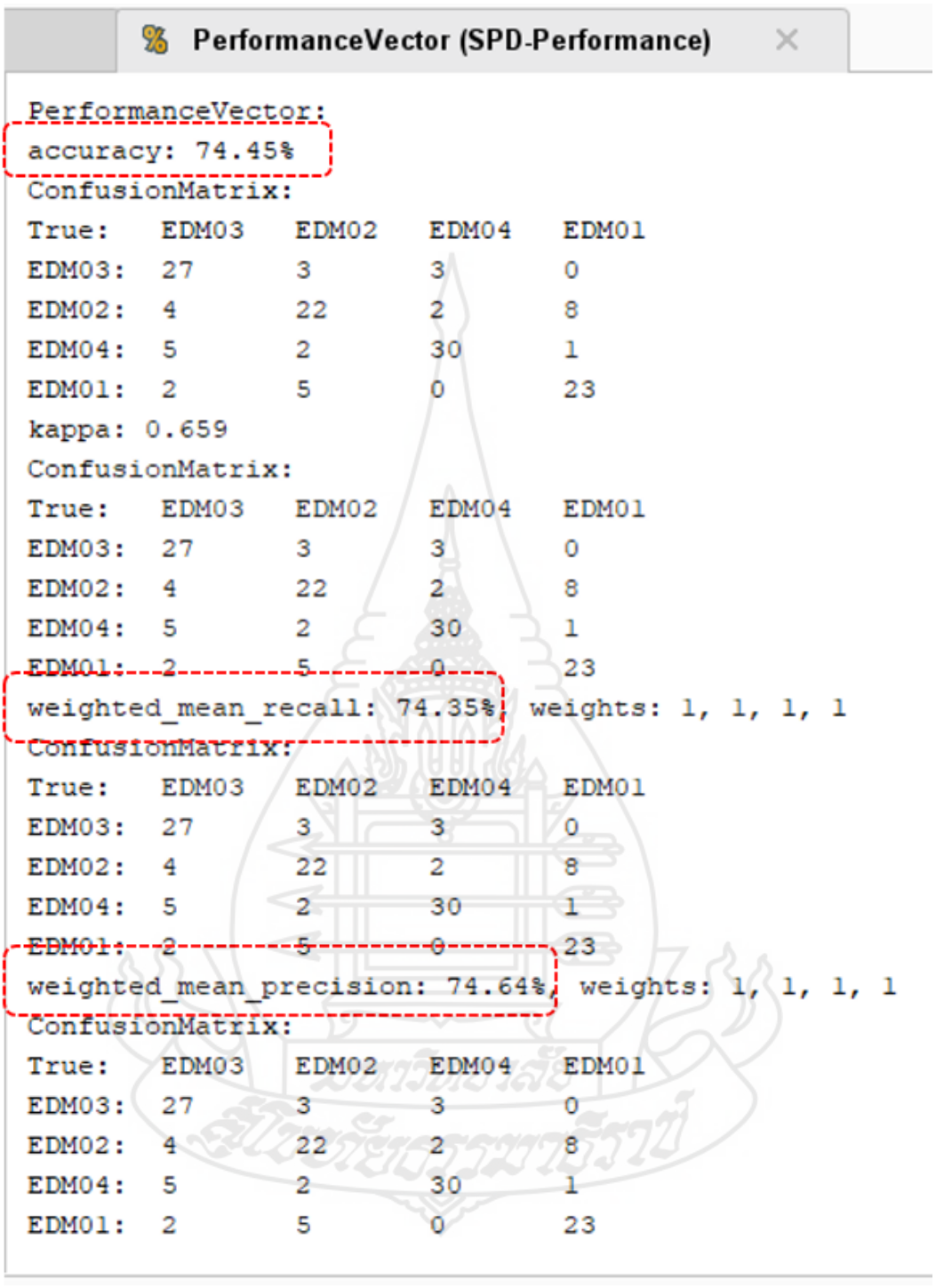
จากภาพที่ 4.18 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 74.63 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 74.42 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 75.34

PerformanceVector (SPRF-Performance)

```
PerformanceVector:
accuracy: 76.59%
ConfusionMatrix:
True:    EDM03    EDM02    EDM04    EDM01
EDM03:   43       3        5        1
EDM02:   2        33       0        8
EDM04:   9        3        45       2
EDM01:   4        9        2        36
kappa: 0.687
ConfusionMatrix:
True:    EDM03    EDM02    EDM04    EDM01
EDM03:   43       3        5        1
EDM02:   2        33       0        8
EDM04:   9        3        45       2
EDM01:   4        9        2        36
weighted_mean_recall: 76.51%, weights: 1, 1, 1, 1
ConfusionMatrix:
True:    EDM03    EDM02    EDM04    EDM01
EDM03:   43       3        5        1
EDM02:   2        33       0        8
EDM04:   9        3        45       2
EDM01:   4        9        2        36
weighted_mean_precision: 76.57%, weights: 1, 1, 1, 1
ConfusionMatrix:
True:    EDM03    EDM02    EDM04    EDM01
EDM03:   43       3        5        1
EDM02:   2        33       0        8
EDM04:   9        3        45       2
EDM01:   4        9        2        36
```

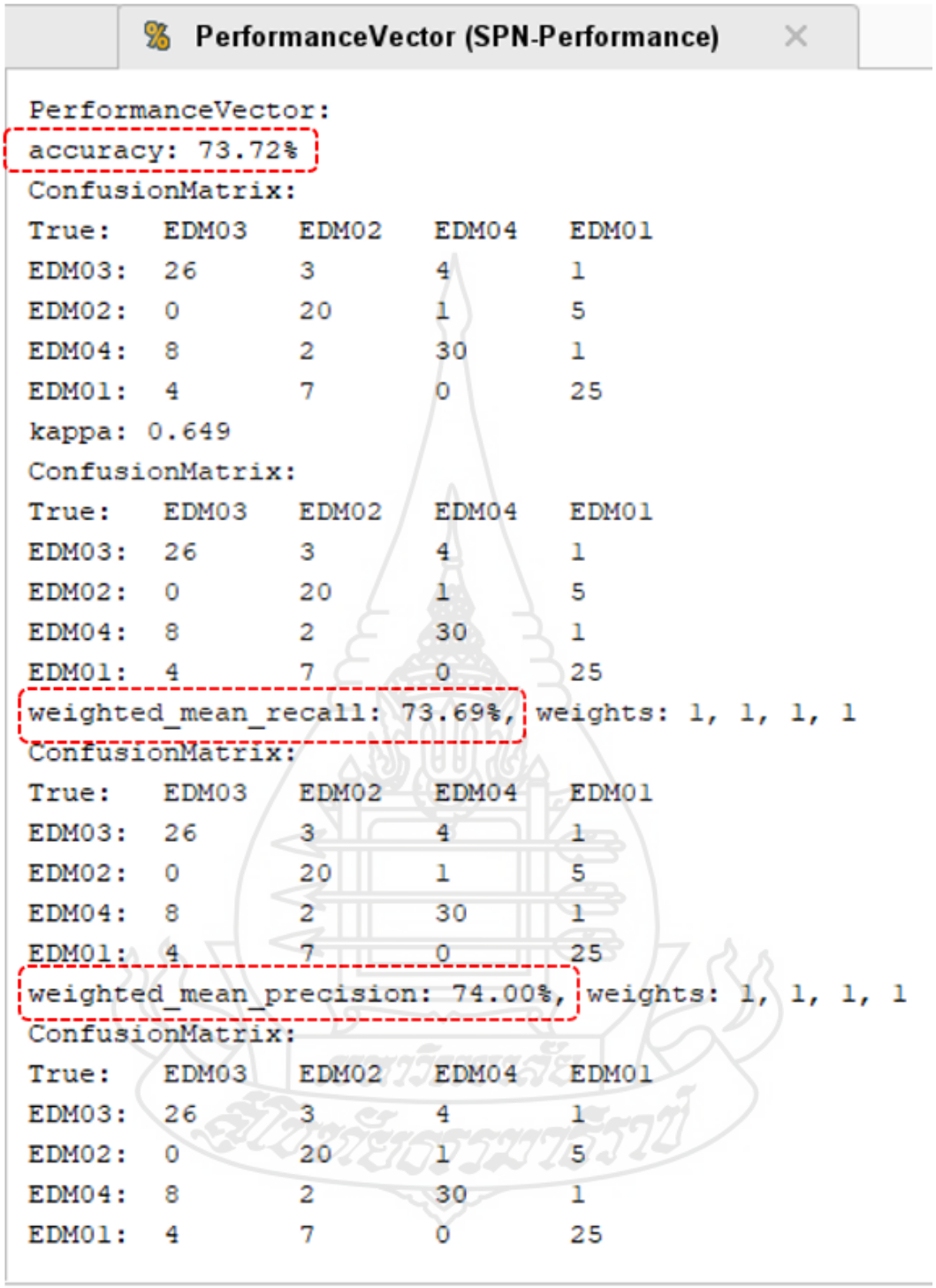
ภาพที่ 4.19 ผลทดสอบแบบจำลองเทคนิคป่าสุ่มด้วยวิธี Split Test (70:30)

จากภาพที่ 4.19 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 76.59 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 76.51 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 76.57



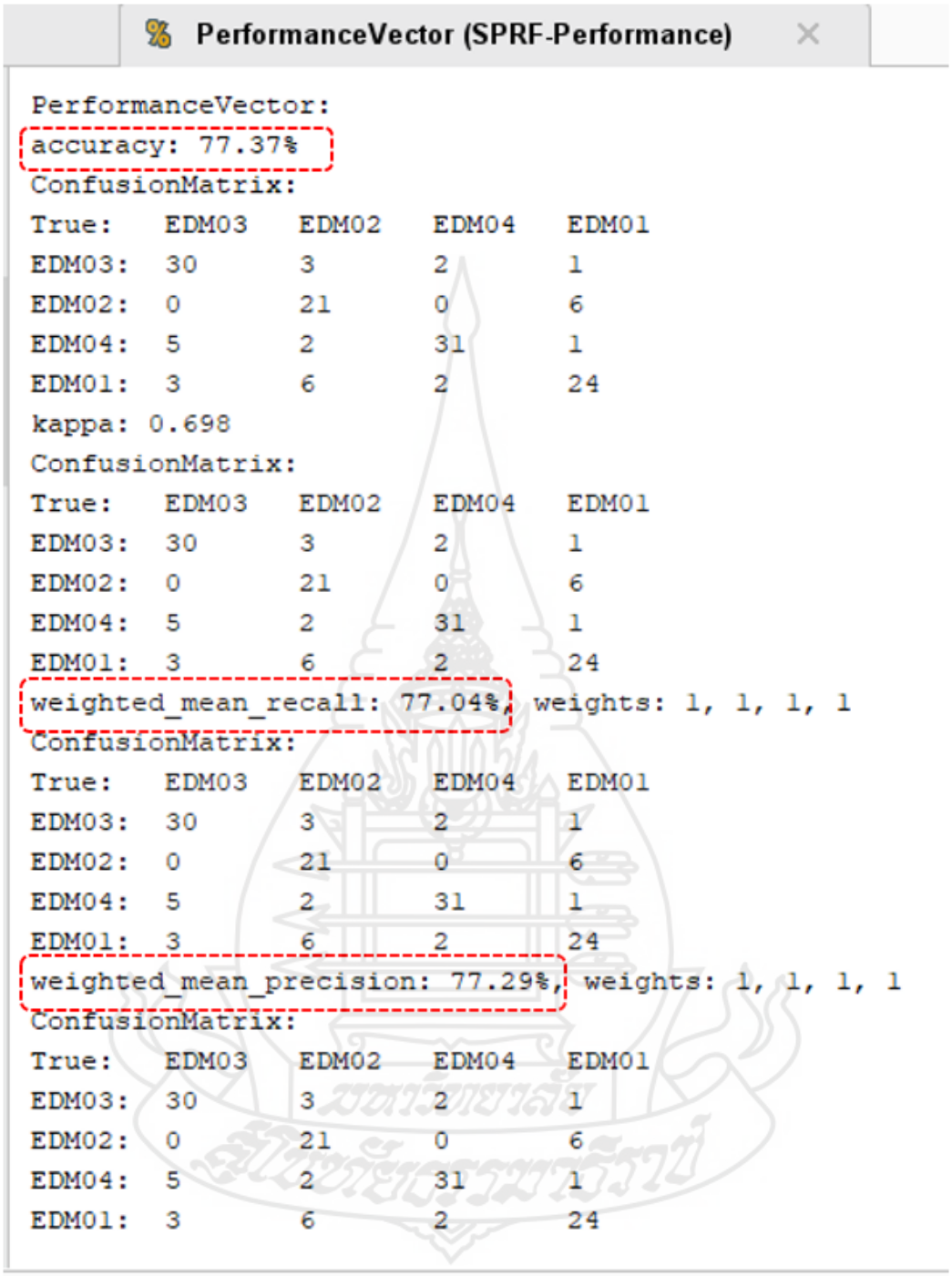
ภาพที่ 4.20 ผลทดสอบแบบจำลองเทคนิคต้นไม้ตัดสินใจด้วยวิธี Split Test (80:20)

จากภาพที่ 4.20 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 74.45 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 74.35 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 74.64



ภาพที่ 4.21 ผลทดสอบแบบจำลองเทคนิคนาอีฟเบย์ด้วยวิธี Split Test (80:20)

จากภาพที่ 4.21 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 73.72 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 73.69 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 74.00



ภาพที่ 4.22 ผลทดสอบแบบจำลองเทคนิคป่าสุ่มด้วยวิธี Split Test (80:20)

จากภาพที่ 4.22 แสดงผลลัพธ์จากการรันโปรแกรมมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 77.37 มีค่าความครบถ้วน (Recall)คิดเป็นร้อยละ 77.04 และค่าความแม่นยำ (Precision) คิดเป็นร้อยละ 77.29

ผู้วิจัยได้นำผลจากการรันโปรแกรมเพื่อวัดประสิทธิภาพแบบจำลองด้วยเทคนิคการสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest) โดยนำค่าคุณลักษณะที่ได้จากการคัดเลือกด้วยเทคนิค Information Gain ในขั้นตอนการคัดเลือกคุณลักษณะมาใช้ร่วมกับแบบจำลองและนำแบบจำลองมาเข้าสู่กระบวนการวัดประสิทธิภาพของแบบจำลอง ด้วยเทคนิคการวัดประสิทธิภาพแบบจำลอง ด้วยวิธี 5-fold Cross Validation วิธี 10-fold Cross Validation วิธี Split Validation (70:30) และ วิธี Split Validation (80:20) เพื่อประเมินประสิทธิภาพและให้ได้แบบจำลองที่มีประสิทธิภาพมากที่สุดมาสรุปโดยแยกการเปรียบเทียบจากค่าความถูกต้องโดยรวมของแบบจำลอง (Accuracy) ความแม่นยำของการทำนาย (Precision) ค่าความครบถ้วน (Recall) และค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure) โดยพิจารณาจากค่าที่ให้ผลลัพธ์สูงที่สุด และนำการเปรียบเทียบด้วยวิธีต่าง ๆ มารวมเป็นค่าเฉลี่ยเพื่อหาค่าที่ดีที่สุด ซึ่งจะเป็นแบบจำลองที่มีประสิทธิภาพและเหมาะสมกับการนำไปใช้งานมากที่สุด

ตารางที่ 4.2 ผลลัพธ์ความถูกต้องในการจำแนกประเภทข้อมูล

No	Classification Model	5-fold Cross Validation	10-fold Cross Validation	Split Validation (70/30)	Split Validation (80/20)	ค่าเฉลี่ย (%)
		Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	
1	Decision Tree	73.53	73.82	73.17	74.45	73.74
2	Naive Bays	73.69	71.94	74.63	73.72	73.50
3	Random Forest	74.71	74.68	76.59	77.37	75.84

ตารางที่ 4.2 แสดงค่าความถูกต้องในการจำแนกข้อมูลเฉลี่ยรวมพบว่า แบบจำลอง Random Forest มีค่าความถูกต้องสูงสุดคิดเป็นร้อยละ 75.84 ลำดับต่อมาคือแบบจำลอง Decision Tree มีค่าความถูกต้องคิดเป็นร้อยละ 73.74 และแบบจำลอง Naïve Bays มีค่าความถูกต้องคิดเป็นร้อยละ 73.50

ตารางที่ 4.3 ผลลัพธ์การวัดค่าความแม่นยำ

No	Classification Model	5-fold Cross Validation	10-fold Cross Validation	Split Validation (70/30)	Split Validation (80/20)	ค่าเฉลี่ย (%)
		Precision (%)	Precision (%)	Precision (%)	Precision (%)	
1	Decision Tree	73.58	74.64	72.62	74.64	73.87
2	Naive Bays	74.71	72.94	75.34	74.00	74.25
3	Random Forest	75.02	75.14	76.57	77.29	76.01

จากตารางที่ 4.3 แสดงค่าความแม่นยำในการจำแนกประเภทข้อมูลพบว่า แบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความแม่นยำสูงสุดคิดเป็นร้อยละ 76.01 ลำดับต่อมาคือแบบจำลองด้วยเทคนิคนาอีฟเบย์ (Naïve Bays) มีค่าความแม่นยำคิดเป็นร้อยละ 74.25 และแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีความแม่นยำคิดเป็นร้อยละ 73.87

ตารางที่ 4.4 ผลลัพธ์การวัดค่าความครบถ้วน

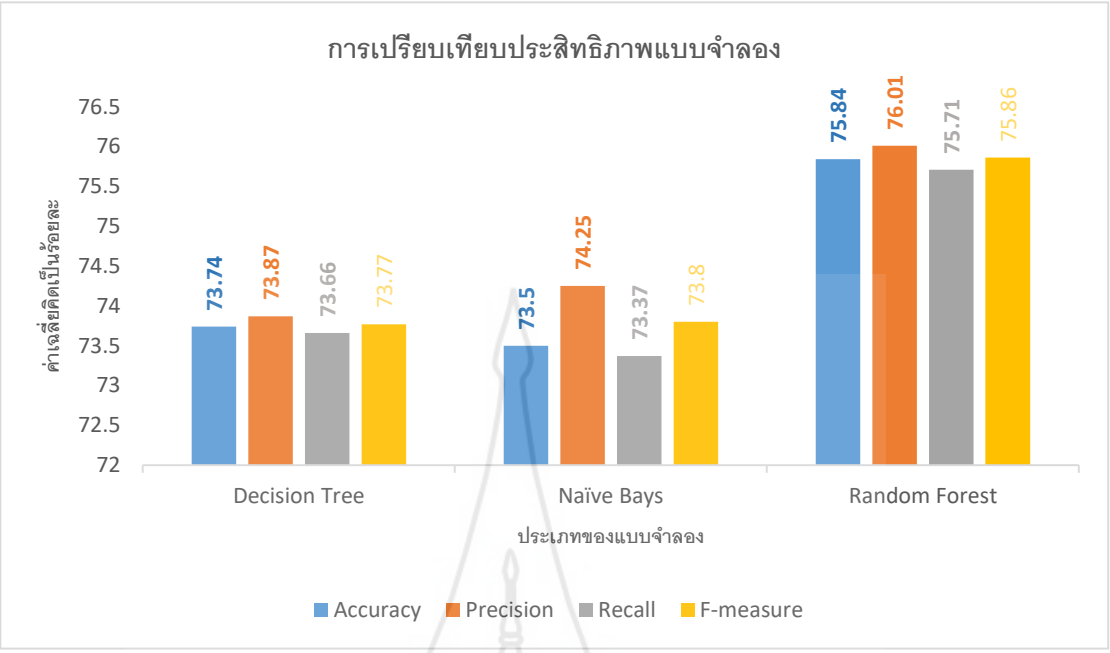
No	Classification Model	5-fold Cross Validation	10-fold Cross Validation	Split Validation (70/30)	Split Validation (80/20)	ค่าเฉลี่ย (%)
		Recall (%)	Recall (%)	Recall (%)	Recall (%)	
1	Decision Tree	73.54	73.82	72.94	74.35	73.66
2	Naive Bays	73.52	71.83	74.42	73.69	73.37
3	Random Forest	74.66	74.62	76.51	77.04	75.71

จากตารางที่ 4.4 แสดงค่าความครบถ้วนพบว่า แบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความครบถ้วนสูงสุดคิดเป็นร้อยละ 75.71 ลำดับต่อมาแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีค่าความครบถ้วนคิดเป็นร้อยละ73.66 และแบบจำลองด้วยเทคนิคนาอีฟเบย์ (Naïve Bays) มีความครบถ้วนคิดเป็นร้อยละ 73.37

ตารางที่ 4.5 ผลลัพธ์การวัดค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม

No	Classification Model	5-fold Cross Validation	10-fold Cross Validation	Split Validation (70/30)	Split Validation (80/20)	ค่าเฉลี่ย (%)
		F-Measure (%)	F-Measure (%)	F-Measure (%)	F-Measure (%)	
1	Decision Tree	73.56	74.23	72.78	74.49	73.77
2	Naive Bays	74.11	72.38	74.88	73.84	73.80
3	Random Forest	74.84	74.88	76.54	77.17	75.86

จากตารางที่ 4.5 แสดงค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมพบว่า แบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมสูงสุดคิดเป็นร้อยละ 75.86 ลำดับต่อมาแบบจำลองด้วยเทคนิคนาอีฟเบย์ (Naïve Bays) มีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมคิดเป็นร้อยละ 73.80 และแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมคิดเป็นร้อยละ 73.77



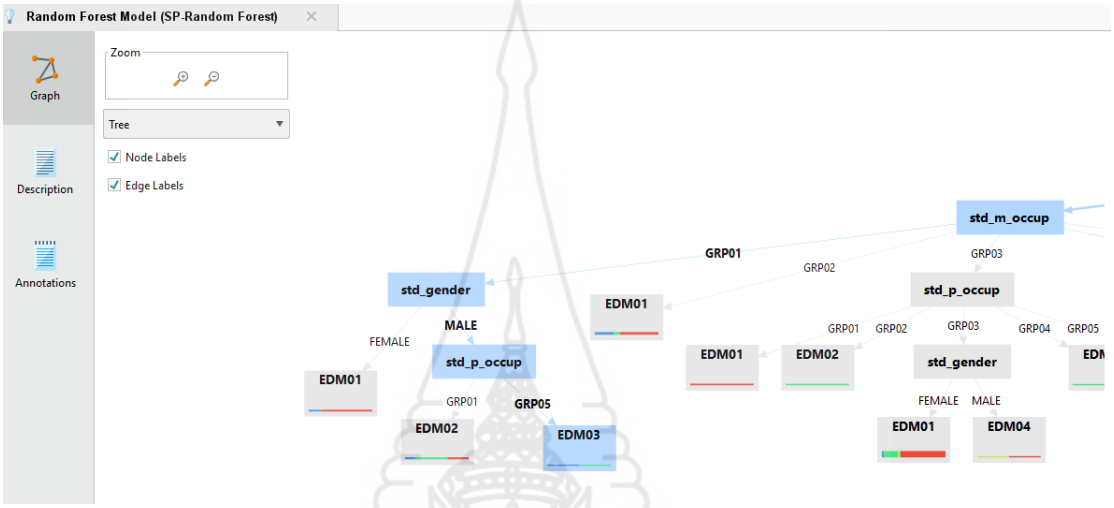
ภาพที่ 4.23 การเปรียบเทียบประสิทธิภาพแบบจำลอง

จากภาพที่ 4.23 แสดงการเปรียบเทียบประสิทธิภาพแบบจำลองแสดงในรูปแบบแผนภูมิแท่งโดยแสดงค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) และค่าความถ่วงดุล ของแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest)

จากผลการวัดประสิทธิภาพแบบจำลองด้วยเทคนิคการวัดประสิทธิภาพแบบจำลอง ด้วยวิธี 5-fold Cross Validation วิธี 10-fold Cross Validation วิธี Split Validation (70:30) และ วิธี Split Validation (80:20) เพื่อประเมินประสิทธิภาพและให้ได้แบบจำลองที่มีประสิทธิภาพมากที่สุดพบว่า แบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) ด้วยวิธีการแยกทดสอบด้วยวิธี Split Test (80:20) คือแบ่งข้อมูลสำหรับการเรียนรู้คิดเป็นร้อยละ 80 ของข้อมูลทั้งหมดและแบ่งข้อมูลสำหรับการทำนายคิดเป็นร้อยละ 20 ของข้อมูลทั้งหมด มีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 75.84 มีค่าความแม่นยำในการทำนาย (Precision) คิดเป็นร้อยละ 76.01 มีค่าความครบถ้วน (Recall) คิดเป็นร้อยละ 75.71 และมีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure) คิดเป็นร้อยละ 75.85 เป็นแบบจำลองและวิธีการที่จะนำไปใช้เป็นต้นแบบเพื่อเป็นแนวทางการพัฒนาระบบและประยุกต์ใช้เพื่อทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ มากที่สุด

4. ผลการทำนาย

การวิจัยนี้นำคุณลักษณะที่ได้จากการคัดเลือกคุณลักษณะมาวิเคราะห์เชิงทำนายด้วยแบบจำลอง Random Forest โดยใช้โปรแกรม Rapid Miner แสดงตัวแบบในรูปแบบต้นไม้ตัดสินใจ ดังภาพที่ 4.24



ภาพที่ 4.24 ตัวแบบจำลองในรูปแบบต้นไม้ตัดสินใจ

จากภาพที่ 4.24 แสดงผลลัพธ์ที่ได้จากการประมวลผลของโปรแกรม Rapid Miner เมื่อเลือกการแสดงแบบจำลองในลักษณะของกราฟ ผลลัพธ์จะแสดงในรูปแบบต้นไม้ตัดสินใจ โดยในโปรแกรม Rapid Miner สามารถแสดงรายละเอียดการอธิบายตัวแบบในรูปแบบต้นไม้ตัดสินใจ เมื่อเลือกผลลัพธ์จะมีสีหรือข้อมูลที่บ่งบอกถึงการตัดสินใจในแต่ละคุณลักษณะที่นำมาวิเคราะห์และย้อนกลับไปยังโหนดแม่เพื่อให้มีความเข้าใจได้ง่ายขึ้น การใช้งานหรือการเลือกดูผลลัพธ์สามารถย่อ ขยายและเลื่อนเพื่อไปดูในโหนดอื่นได้ ตัวอย่างดังภาพแสดงให้เห็นผลลัพธ์สุดท้ายคือEDM03 (หลักสูตรการศึกษาพิเศษ) เมื่อนำเมาส์เลือกจะแสดงแถบสีย้อนกลับไปยัง std_p_occup (กลุ่มอาชีพของผู้ปกครอง) โดยจะเป็น GRP05 (กลุ่มอาชีพอื่น ๆ) และย้อนกลับขึ้นไปยังคุณลักษณะ std_gender (เพศ) ซึ่งจะเป็น MALE (เพศชาย) เป็นต้น จากรูปแบบต้นไม้ตัดสินใจนี้สามารถที่จะดูย้อนจากผลลัพธ์การจำแนกข้อมูลไปยังโหนดแม่หรือจากโหนดแม่มายังผลลัพธ์ เพื่อให้เข้าใจผลลัพธ์ของข้อมูลได้

ภาพที่ 4.25 ผลลัพธ์การรันโปรแกรมด้วยเทคนิค Random Forest

จากภาพที่ 4.5 แสดงผลการจำแนกประเภทข้อมูลโดยนำคุณลักษณะทั้งหมดที่นำมาวิเคราะห์จำนวน 17 คุณลักษณะ โดยการหาค่าน้ำหนักเพื่อให้ได้ซึ่งคุณลักษณะที่มีความสัมพันธ์กับคุณลักษณะเป้าหมาย เมื่อนำคุณลักษณะทั้งหมดมาใช้งานร่วมกับแบบจำลองและลดจำนวนคุณลักษณะลงทีละตัว นำมาวัดประสิทธิภาพแบบจำลองด้วยวิธีเปรียบเทียบค่าความถูกต้อง ค่าความแม่นยำ ค่าความครบถ้วนและค่าประสิทธิภาพโดยรวม จากนั้นหาค่าเฉลี่ยของรูปแบบการจำแนกแบบจำลองแต่ละแบบเพื่อให้ได้รูปแบบเทคนิคการจำแนกประเภทที่มีความเหมาะสมกับคุณลักษณะที่ต้องการ

จากกระบวนการวิเคราะห์นี้ผลที่ได้จากการนำคุณลักษณะที่มีค่าน้ำหนักมากที่สุด จำนวน 5 คุณลักษณะ ได้แก่ 1) แผนการเรียนที่จบจากระดับมัธยม 2) เกรดเฉลี่ย 3) เพศ 4) อาชีพผู้ปกครอง 5) อาชีพมารดา นำมาใช้ร่วมกับเทคนิคป่าสุ่ม (Random Forest) โดยคุณลักษณะเป้าหมายที่จะนำมาหาผลลัพธ์คือ สาขาวิชา เมื่อนำมาวิเคราะห์เชิงทำนายในลักษณะของต้นไม้สามารถอธิบายผลจากการจำแนกประเภทข้อมูลและการทำนายผลที่ได้จากการประมวลผลด้วยโปรแกรม RapidMiner ในรูปแบบคำอธิบายตามลำดับตามผลลัพธ์ดังนี้

std_program = PLAN01

| std_gender = FEMALE

| | std_m_occup = GRP01

| | | std_gpa = GOOD: EDM01 {EDM03=3, EDM02=1, EDM04=0, EDM01=5}

กลุ่มที่ 1 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับดี เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_gpa = MEDIUM: EDM03 {EDM03=6, EDM02=1, EDM04=0, EDM01=0}

กลุ่มที่ 2 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับปานกลาง เลือกสมัครหลักสูตรพลศึกษา

| | | std_gpa = VERYGOOD: EDM02 {EDM03=0, EDM02=4, EDM04=0, EDM01=2}

กลุ่มที่ 3 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับดีมาก เลือกสมัครหลักสูตรการประถมศึกษา

| | std_m_occup = GRP02

| | | std_p_occup = GRP02: EDM02 {EDM03=0, EDM02=3, EDM04=0, EDM01=3}

กลุ่มที่ 4 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพพนักงานเอกชน/หน่วยงานเอกชน เลือกสมัครหลักสูตรการประถมศึกษา

| | | std_p_occup = GRP03: EDM02 {EDM03=1, EDM02=2, EDM04=0, EDM01=0}

กลุ่มที่ 5 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิงมารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรการประถมศึกษา

| | | std_p_occup = GRP04

| | | | std_gpa = GOOD: EDM02 {EDM03=0, EDM02=3, EDM04=0, EDM01=0}

กลุ่มที่ 6 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับดี เลือกสมัครหลักสูตรการประถมศึกษา

| | | | std_gpa = MEDIUM: EDM01 {EDM03=0, EDM02=0, EDM04=0, EDM01=3}

กลุ่มที่ 7 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับปานกลาง เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_p_occup = GRP05: EDM01 {EDM03=0, EDM02=0, EDM04=0, EDM01=2}

กลุ่มที่ 8 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพอื่น ๆ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | std_m_occup = GRP03

| | | std_p_occup = GRP01: EDM02 {EDM03=1, EDM02=2, EDM04=0, EDM01=0}

กลุ่มที่ 9 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ เลือกสมัครหลักสูตรการประถมศึกษา

| | | std_p_occup = GRP03: EDM01 {EDM03=4, EDM02=13, EDM04=2, EDM01=30}

กลุ่มที่ 10 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_p_occup = GRP04: EDM01 {EDM03=0, EDM02=0, EDM04=0, EDM01=2}

กลุ่มที่ 11 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพเกษตรกร/ประมง เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_p_occup = GRP05: EDM01 {EDM03=1, EDM02=0, EDM04=0, EDM01=5}

กลุ่มที่ 12 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพอื่น ๆ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | std_m_occup = GRP04

| | | std_gpa = GOOD: EDM01 {EDM03=1, EDM02=4, EDM04=1, EDM01=50}

กลุ่มที่ 13 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับดี เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_gpa = MEDIUM: EDM01 {EDM03=0, EDM02=3, EDM04=0, EDM01=7}

กลุ่มที่ 14 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับปานกลาง เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_gpa = VERYGOOD: EDM02 {EDM03=0, EDM02=28, EDM04=3, EDM01=8}

กลุ่มที่ 15 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับดีมาก เลือกสมัครหลักสูตรการประถมศึกษา

| | std_m_occup = GRP05

| | | std_gpa = GOOD: EDM01 {EDM03=0, EDM02=5, EDM04=6, EDM01=18}

กลุ่มที่ 16 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพอื่นๆ มีเกรดเฉลี่ยในระดับดี เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_gpa = MEDIUM: EDM01 {EDM03=1, EDM02=0, EDM04=2, EDM01=6}

กลุ่มที่17 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพอื่นๆ มีเกรดเฉลี่ยในระดับปานกลาง เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_gpa = VERYGOOD: EDM02 {EDM03=0, EDM02=12, EDM04=0, EDM01=2}

กลุ่มที่18 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพอื่นๆ มีเกรดเฉลี่ยในระดับดีมาก เลือกสมัครหลักสูตรการประถมศึกษา

| std_gender = MALE
| | std_gpa = GOOD
| | | std_m_occup = GRP01: EDM02 {EDM03=1, EDM02=5, EDM04=0, EDM01=1}

กลุ่มที่ 19 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_m_occup = GRP02: EDM03 {EDM03=2, EDM02=0, EDM04=0, EDM01=0}

กลุ่มที่ 20 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน เลือกสมัครหลักสูตรพลศึกษา

| | | std_m_occup = GRP03: EDM01 {EDM03=1, EDM02=3, EDM04=1, EDM01=4}

กลุ่มที่ 21 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_m_occup = GRP04: EDM02 {EDM03=0, EDM02=7, EDM04=0, EDM01=2}

กลุ่มที่ 22 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพเกษตรกร/ประมง เลือกสมัครหลักสูตรการประถมศึกษา

| | | std_m_occup = GRP05: EDM01 {EDM03=1, EDM02=1, EDM04=0, EDM01=5}

กลุ่มที่ 23 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพอื่น ๆ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | std_gpa = LOW: EDM03 {EDM03=7, EDM02=0, EDM04=0, EDM01=0}

กลุ่มที่ 24 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับต่ำ เลือกสมัครหลักสูตรพลศึกษา

| | std_gpa = MEDIUM
| | | std_m_occup = GRP01: EDM01 {EDM03=2, EDM02=1, EDM04=0, EDM01=6}

กลุ่มที่ 25 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_m_occup = GRP03: EDM03 {EDM03=2, EDM02=0, EDM04=0, EDM01=0}

กลุ่มที่ 26 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรพลศึกษา

| | | std_m_occup = GRP04: EDM03 {EDM03=5, EDM02=0, EDM04=0, EDM01=0}

กลุ่มที่ 27 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพเกษตรกร/ประมง เลือกสมัครหลักสูตรพลศึกษา

| | | std_m_occup = GRP05: EDM03 {EDM03=9, EDM02=0, EDM04=0, EDM01=0}

กลุ่มที่ 28 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพอื่น ๆ เลือกสมัครหลักสูตรพลศึกษา

| | std_gpa = VERYGOOD: EDM02 {EDM03=1, EDM02=33, EDM04=0, EDM01=2}

กลุ่มที่ 29 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก เลือกสมัครหลักสูตรการประถมศึกษา

std_program = PLAN02

| std_gender = FEMALE

| | std_gpa = GOOD

| | | std_p_occup = GRP01: EDM04 {EDM03=2, EDM02=0, EDM04=9, EDM01=0}

กลุ่มที่ 30 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับดี เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_p_occup = GRP02

| | | | std_m_occup = GRP02: EDM04 {EDM03=0, EDM02=0, EDM04=2, EDM01=0}

กลุ่มที่ 31 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพพนักงานเอกชน/หน่วยงานเอกชน มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | | | std_m_occup = GRP05: EDM01 {EDM03=0, EDM02=0, EDM04=0, EDM01=2}

กลุ่มที่ 32 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพพนักงานเอกชน/หน่วยงานเอกชน มารดาอาชีพอื่น ๆ เลือกสมัครหลักสูตรการศึกษาปฐมวัย

| | | std_p_occup = GRP03: EDM04 {EDM03=1, EDM02=1, EDM04=21, EDM01=2}

กลุ่มที่ 33 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | | std_p_occup = GRP04: EDM04 {EDM03=6, EDM02=1, EDM04=18, EDM01=0}

กลุ่มที่ 34 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพเกษตรกร/ประมง เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | | std_p_occup = GRP05: EDM04 {EDM03=11, EDM02=1, EDM04=23, EDM01=0}

กลุ่มที่ 35 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพอื่น ๆ เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | std_gpa = LOW: EDM03 {EDM03=5, EDM02=0, EDM04=0, EDM01=0}

กลุ่มที่ 36 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับต่ำ เลือกสมัครหลักสูตรพลศึกษา

| | std_gpa = MEDIUM

| | | std_p_occup = GRP03: EDM04 {EDM03=0, EDM02=0, EDM04=5, EDM01=0}

กลุ่มที่ 37 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับปานกลาง ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | | std_p_occup = GRP04: EDM04 {EDM03=0, EDM02=0, EDM04=3, EDM01=0}

กลุ่มที่ 38 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับปานกลาง ผู้ปกครองอาชีพเกษตรกร/ประมง เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | | std_p_occup = GRP05: EDM03 {EDM03=12, EDM02=0, EDM04=3, EDM01=0}

กลุ่มที่ 39 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับปานกลาง ผู้ปกครองอื่น ๆ เลือกสมัครหลักสูตรพลศึกษา

| | std_gpa = VERYGOOD: EDM04 {EDM03=6, EDM02=4, EDM04=56, EDM01=0}

กลุ่มที่ 40 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับดีมาก เลือกสมัครหลักสูตรการศึกษาพิเศษ

| std_gender = MALE

| | std_gpa = GOOD: EDM03 {EDM03=35, EDM02=4, EDM04=4, EDM01=0}

กลุ่มที่ 41 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี เลือกสมัครหลักสูตรพลศึกษา

| | std_gpa = LOW: EDM03 {EDM03=5, EDM02=0, EDM04=0, EDM01=0}

กลุ่มที่ 42 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับต่ำ เลือกสมัครหลักสูตรพลศึกษา

| | std_gpa = MEDIUM: EDM03 {EDM03=32, EDM02=0, EDM04=6, EDM01=0}

กลุ่มที่ 43 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง เลือกสมัครหลักสูตรพลศึกษา

| | std_gpa = VERYGOOD

| | | std_m_occup = GRP03: EDM04 {EDM03=0, EDM02=4, EDM04=7, EDM01=0}

กลุ่มที่ 44 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | | std_m_occup = GRP04: EDM04 {EDM03=1, EDM02=0, EDM04=2, EDM01=1}

กลุ่มที่ 45 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก มารดาอาชีพเกษตรกร/ประมง เลือกสมัครหลักสูตรการศึกษาพิเศษ

| | | std_m_occup = GRP05: EDM03 {EDM03=1, EDM02=1, EDM04=0, EDM01=0}

กลุ่มที่ 46 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ เลือกสมัครหลักสูตรพลศึกษา

std_program = PLAN03: EDM03 {EDM03=24, EDM02=0, EDM04=5, EDM01=0}

กลุ่มที่ 47 ผู้สมัครจบแผนการเรียนจากมัธยมสายอื่น ๆ เลือกสมัครหลักสูตรพลศึกษา

4.1 ผลการทำนายกลุ่มผู้สมัครแยกตามหลักสูตร

จากการอธิบายผลลัพธ์จากตัวแบบต้นไม้ในรูปแบบต้นไม้ตัดสินใจเมื่อพิจารณาและนำมาจัดกลุ่มและแยกตามหลักสูตรการสมัคร เพื่อให้เกิดความเข้าใจง่ายและสามารถพิจารณาคุณลักษณะที่มีความสัมพันธ์กับสาขาวิชาเป้าหมาย โดยนำผลลัพธ์มาจัดกลุ่มตามหลักสูตรได้ดังนี้

หลักสูตรการศึกษาปฐมวัย ผู้สมัครที่จะเลือกสมัครหลักสูตรการศึกษาปฐมวัยอยู่ในกลุ่มดังนี้

กลุ่มที่ 1 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับดี

กลุ่มที่ 7 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับปานกลาง

กลุ่มที่ 8 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพอื่น ๆ

กลุ่มที่ 10 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 11 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพเกษตรกร/ประมง

กลุ่มที่ 12 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพอื่น ๆ

กลุ่มที่ 13 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับดี

กลุ่มที่ 14 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับปานกลาง

กลุ่มที่ 16 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพอื่นๆ มีเกรดเฉลี่ยในระดับดี

กลุ่มที่17 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพอื่นๆ มีเกรดเฉลี่ยในระดับปานกลาง

กลุ่มที่ 19 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ

กลุ่มที่ 21 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 23 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพอื่น ๆ

กลุ่มที่ 25 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ

กลุ่มที่ 30 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับดี

กลุ่มที่ 32 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพพนักงานเอกชน/หน่วยงานเอกชน มารดาอาชีพอื่น ๆ

หลักสูตรการประถมศึกษา ผู้สมัครที่จะเลือกสมัครหลักสูตรการประถมศึกษาอยู่ในกลุ่มดังนี้

กลุ่มที่ 3 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับดีมาก

กลุ่มที่ 4 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพพนักงานเอกชน/หน่วยงานเอกชน

กลุ่มที่ 5 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิงมารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 6 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน ผู้ปกครองอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับดี

กลุ่มที่ 9 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ ผู้ปกครองอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ

กลุ่มที่ 15 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพเกษตรกร/ประมง มีเกรดเฉลี่ยในระดับดีมาก

กลุ่มที่18 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพอื่นๆ มีเกรดเฉลี่ยในระดับดีมาก

กลุ่มที่ 22 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพเกษตรกร/ประมง

กลุ่มที่ 29 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก

หลักสูตรพลศึกษา ผู้สมัครที่จะเลือกสมัครหลักสูตรพลศึกษาอยู่ในกลุ่มดังนี้

กลุ่มที่ 2 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศหญิง มารดาอาชีพรับราชการ/พนักงานราชการ/รัฐวิสาหกิจ/พนักงานรัฐวิสาหกิจ มีเกรดเฉลี่ยในระดับปานกลาง

กลุ่มที่ 20 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน

กลุ่มที่ 24 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับต่ำ

กลุ่มที่ 26 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 27 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพเกษตรกร/ประมง

กลุ่มที่ 28 ผู้สมัครจบแผนการเรียนจากมัธยมสายวิทย์-คณิต เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง มารดาอาชีพอื่น ๆ

กลุ่มที่ 36 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับต่ำ

กลุ่มที่ 39 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับปานกลาง ผู้ปกครองอื่น ๆ

กลุ่มที่ 41 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดี

กลุ่มที่ 42 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับต่ำ

กลุ่มที่ 43 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับปานกลาง

กลุ่มที่ 46 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 47 ผู้สมัครจบแผนการเรียนจากมัธยมสายอื่น ๆ

หลักสูตรการศึกษาพิเศษ ผู้สมัครที่จะเลือกสมัครหลักสูตรการศึกษาพิเศษอยู่ในกลุ่มดังนี้

กลุ่มที่ 31 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพพนักงานเอกชน/หน่วยงานเอกชน มารดาอาชีพพนักงานเอกชน/หน่วยงานเอกชน

กลุ่มที่ 33 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 34 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพเกษตรกร/ประมง

กลุ่มที่ 35 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง ผู้ปกครองอาชีพอื่น ๆ

กลุ่มที่ 37 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับปานกลาง ผู้ปกครองอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 38 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับปานกลาง ผู้ปกครองอาชีพเกษตรกร/ประมง

กลุ่มที่ 40 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศหญิง มีเกรดเฉลี่ยในระดับดีมาก

กลุ่มที่ 44 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก มารดาอาชีพค้าขาย/ธุรกิจส่วนตัว/รับจ้างอิสระ

กลุ่มที่ 45 ผู้สมัครจบแผนการเรียนจากมัธยมสายศิลป์ เป็นเพศชาย มีเกรดเฉลี่ยในระดับดีมาก มารดาอาชีพเกษตรกร/ประมง

บทที่ 5
สรุปการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยเรื่อง “การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่” มีวัตถุประสงค์ 1) เพื่อวิเคราะห์และคัดเลือกคุณลักษณะสำคัญที่สัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ 2) เพื่อสร้างแบบจำลองและประเมินประสิทธิภาพในการวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ มีรายละเอียดผลการวิจัยดังนี้

1. สรุปการวิจัย

1.1 ผลการคัดเลือกคุณลักษณะสำคัญ

ผลการคัดเลือกคุณลักษณะสำคัญที่สัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ ด้วยวิธี Information Gain เพื่อวิเคราะห์และหาค่าความสัมพันธ์กันระหว่างคุณลักษณะที่นำมาวิเคราะห์กับคุณลักษณะเป้าหมาย โดยคุณลักษณะที่มีความสัมพันธ์กันจะมีค่าน้ำหนักมากและคุณลักษณะที่ไม่มีความสัมพันธ์กันจะมีค่าน้ำหนักน้อย

จากคุณลักษณะที่ได้จากการคัดเลือกนำมาใช้ในงานวิจัยนี้ทั้งหมดจำนวน 17 คุณลักษณะ โดยคุณลักษณะเป้าหมาย คือ สาขาวิชา เมื่อนำมาหาค่าความสัมพันธ์กันกับคุณลักษณะคงเหลือจำนวน 16 คุณลักษณะ ประกอบด้วย 1) ปีที่สมัคร 2) เพศ 3) สาขาที่สมัคร 4) แผนการเรียนที่จบจากระดับมัธยม 5) เกรดเฉลี่ย 6) สัญชาติ 7) ศาสนา 8) สัญชาติบิดา 9) สัญชาติมารดา 10) สถานการณ์มีชีวิตบิดา 11) สถานการณ์มีชีวิตมารดา 12) สถานะครอบครัว 13) จังหวัด 14) ผู้ปกครอง 15) อาชีพผู้ปกครอง 16) อาชีพบิดาและอาชีพมารดา ซึ่งคุณลักษณะสำคัญที่สัมพันธ์กับการสมัครเรียนจะมีค่าน้ำหนักมากและลดลงตามลำดับความสำคัญที่สัมพันธ์กัน

เมื่อนำมาใช้งานร่วมกับแบบจำลองโดยนำจำนวนคุณลักษณะทั้งหมดมาทดสอบและลดจำนวนคุณลักษณะที่มีค่าน้ำหนักน้อยออกไปครั้งละหนึ่งคุณลักษณะ จากนั้นเปรียบเทียบค่าประสิทธิภาพของแบบจำลอง ผลพบว่า การใช้คุณลักษณะที่มีค่ามากที่สุดจำนวน 5 คุณลักษณะยังคงมีผลการวิเคราะห์ที่มีประสิทธิภาพดี ซึ่งคุณลักษณะจำนวน 5 คุณลักษณะแรกคือแผนการเรียนที่จบจากระดับมัธยม มีค่าน้ำหนักที่ 0.522 รองลงมาคือ เกรดเฉลี่ย มีค่าน้ำหนักที่ 0.290 รองลงมาคือ เพศ มีค่าน้ำหนักที่ 0.207 รองลงมาคืออาชีพผู้ปกครอง มีค่าน้ำหนักที่ 0.056 รองลงมาคืออาชีพมารดา มีค่าน้ำหนักที่ 0.055 และจังหวัดมีค่าน้ำหนักน้อยที่สุดคิดเป็นร้อยละ 0.001 เมื่อนำคุณลักษณะที่มีค่าน้ำหนักมากที่สุด 5 คุณลักษณะเมื่อนำไปใช้ร่วมกับแบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความถูกต้อง (Accuracy) สูงที่สุดคิดเป็นร้อยละ 77.37 มีความแม่นยำ (Precision) สูงสุดคิดเป็นร้อยละ 77.29 มีค่าความครบถ้วน (Recall) สูงสุดคิดเป็นร้อยละ 77.04 และมีค่าความความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมสูงสุดคิดเป็นร้อยละ 77.17

1.2 ผลการประเมินประสิทธิภาพ

ผลการประเมินประสิทธิภาพในการวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ โดยนำคุณลักษณะสำคัญที่มีความสัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ที่มีค่าน้ำหนักที่ดีที่สุดจำนวน 5 คุณลักษณะประกอบด้วย 1) แผนการเรียนที่จบจากระดับมัธยม 2) เกรดเฉลี่ย 3) เพศ 4) อาชีพผู้ปกครองและ 5) อาชีพมารดา เมื่อนำมาใช้ร่วมกับการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคป่าสุ่ม (Random Forest) นำมาวัดประสิทธิภาพแบบจำลองด้วยวิธี 5-fold Cross Validation วิธี 10-fold Cross Validation วิธี Split Validation (70:30) และวิธี Split Validation (80:20) โดยนำผลวัดประสิทธิภาพแบบจำลองแต่ละวิธีมาคำนวณหาค่าเฉลี่ยของเทคนิคการจำแนกประเภทข้อมูล ผลการศึกษามีรายละเอียดดังนี้

แบบจำลองที่ให้ค่าความถูกต้อง (Accuracy) สูงสุดคือแบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความถูกต้องสูงสุดคิดเป็นร้อยละ 75.84 ลำดับต่อมาคือแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีค่าความถูกต้องคิดเป็นร้อยละ 73.74 และแบบจำลองด้วยเทคนิค (Naïve Bays) มีค่าความถูกต้องคิดเป็นร้อยละ 73.50

แบบจำลองที่มีค่าความแม่นยำ (Precision) สูงสุดคือแบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความแม่นยำสูงสุดคิดเป็นร้อยละ 76.01 ลำดับต่อมาคือแบบจำลองด้วยเทคนิคนาอีฟเบย์ (Naïve Bays) มีค่าความแม่นยำคิดเป็นร้อยละ 74.25 และแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีความแม่นยำคิดเป็นร้อยละ 73.87

แบบจำลองที่มีค่าความครบถ้วน (Recall) สูงสุดคือแบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความครบถ้วนสูงสุดคิดเป็นร้อยละ 75.71 ลำดับต่อมาแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีค่าความครบถ้วนคิดเป็นร้อยละ73.66 และแบบจำลองด้วยเทคนิคนาอีฟเบย์ (Naïve Bays) มีความครบถ้วนคิดเป็นร้อยละ 73.37

แบบจำลองที่มีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure) สูงสุดคือแบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) มีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมสูงสุดคิดเป็นร้อยละ 75.86 ลำดับต่อมาแบบจำลองด้วยเทคนิคนาอีฟเบย์ (Naïve Bays) มีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมคิดเป็นร้อยละ 73.80 และแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวมคิดเป็นร้อยละ 73.77

จากผลการวัดประสิทธิภาพแบบจำลองด้วยเทคนิคการวัดประสิทธิภาพแบบจำลอง ด้วยวิธี 5-fold Cross Validation วิธี 10-fold Cross Validation วิธี Split Validation (70:30) และ วิธี Split Validation (80:20) ซึ่งวิธีดังกล่าวเป็นการจัดชุดข้อมูลเพื่อเรียนรู้ซึ่งชุดข้อมูลจะประกอบด้วยชุดของข้อมูลและชุดผลลัพธ์ของข้อมูล จากนั้นนำค่าประสิทธิภาพที่ได้จากการทดสอบมาเปรียบเทียบค่าทางสถิติของแต่ละแบบจำลองเพื่อประเมินประสิทธิภาพและให้ได้แบบจำลองที่มีประสิทธิภาพมากที่สุด ผลพบว่า แบบจำลองด้วยเทคนิคป่าสุ่ม (Random Forest) ด้วยวิธีการแยกทดสอบด้วยวิธี Split Test (80:20) คือแบ่งข้อมูลสำหรับการเรียนรู้คิดเป็นร้อยละ 80 ของข้อมูลที่นำมาวิเคราะห์ทั้งหมดและแบ่งข้อมูลสำหรับการทำนายคิดเป็นร้อยละ 20 ของข้อมูลข้อมูลที่นำมาวิเคราะห์ทั้งหมด

โดยมีค่าความถูกต้อง (Accuracy) คิดเป็นร้อยละ 75.84 มีค่าความแม่นยำในการทำนาย (Precision) คิดเป็นร้อยละ 76.01 มีค่าความครบถ้วน (Recall) คิดเป็นร้อยละ 75.71 และมีค่าความถ่วงดุลหรือค่าประสิทธิภาพโดยรวม (F-measure) คิดเป็นร้อยละ 75.85

2. อภิปรายผล

การวิจัยเรื่อง “การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่” สามารถอภิปรายผลการวิจัยได้ดังนี้

ในการวิจัยนี้ทำการวิเคราะห์และคัดเลือกคุณลักษณะสำคัญที่มีสัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ การเตรียมข้อมูลและการแปลงข้อมูลเพื่อนำวิเคราะห์ มีความสอดคล้องกับรัชฎา เทพประสิทธิ์ และจรัญ แสนราช (2563) ศึกษาปัจจัยที่มีผลต่อการเลือกเข้าศึกษาต่อในสาขาวิชา คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงราย โดยนำปัจจัยที่มีผลได้แก่แผนการเรียนก่อนเข้าศึกษา เพศและตัวแปรอื่นจำนวน 9 ตัวแปรมาวิเคราะห์ และยังสอดคล้องกับขันทอง ปทุมชาติ และพิมริมทร์ คีรีนทร์ (2558) จัดกลุ่มและแปลงข้อมูลการสมัครเข้าศึกษา เพศ ระดับเกรดเฉลี่ยประเภทโรงเรียนภายนอกและนอกจังหวัด เพื่อวิเคราะห์พฤติกรรมการเลือกสมัครสาขาวิชาเรียนของนักศึกษาใหม่

จากการเตรียมข้อมูลการวิจัยนี้ได้คุณลักษณะนำมาวิเคราะห์จำนวน 17 คุณลักษณะ โดยนำคุณลักษณะหาค่าความสัมพันธ์กับคุณลักษณะเป้าหมาย ด้วยวิธี Information Gain เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญมีความสัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่ โดยคุณลักษณะเป้าหมายคือสาขาวิชา จะไม่ถูกนำมาคำนวณหาค่า คงเหลือจำนวนคุณลักษะที่นำมาจำนวน 16 คุณลักษณะ จากนั้นนำมาใช้งานร่วมกับแบบจำลองและลดจำนวนคุณลักษณะที่มีค่าน้อยออกครั้งละหนึ่งคุณลักษณะ เพื่อวัดประสิทธิภาพของแบบจำลอง การใช้คุณลักษณะจำนวน 5 คุณลักษณะที่ดีที่สุดยังคงมีประสิทธิภาพดี ซึ่งสอดคล้องกับงานวิจัยของอัจจิมา มณฑพันธ์ (2562) เปรียบเทียบวิธีคัดเลือกคุณลักษณะในการปรับปรุงการพยากรณ์มะเร็งเต้านม ด้วยวิธี Information Gain จำนวน 6 คุณลักษณะเมื่อนำมาพยากรณ์วัดค่าความถูกต้องดีที่สุดคิดเป็นร้อยละ 92.27 และยังสอดคล้องกับงานวิจัยของรัชฎา เทพประสิทธิ์และจรัญ แสนราช (2563) นำคุณลักษณะแผนการเรียนก่อนเข้าศึกษาและคุณลักษณะอื่นรวม 9 คุณลักษณะมาประเมินประสิทธิภาพของแบบจำลองเพื่อวัดความแม่นยำซึ่งได้ค่าความแม่นยำร้อยละ 72.5 ในการศึกษาปัจจัยที่มีผลต่อเลือกเข้าศึกษาต่อ การคัดเลือกคุณลักษณะด้วยวิธี Information Gain ยังสอดคล้องกับวันวิสาข์ ชนะประเสริฐ (2559) เปรียบเทียบวิธีการคัดเลือกคุณลักษณะพบว่าวิธีการคัดเลือกคุณลักษณะด้วยวิธี Information Gain มีความเหมาะสมที่สุด วิธีการคัดเลือกคุณลักษณะในการวิจัยนี้สอดคล้องกับแนวคิดของเอกสิทธิ์ พัชรวงศ์ศักดา (2557) ที่กล่าวว่าการคัดเลือกคุณลักษณะที่สำคัญที่มีความสัมพันธ์กับคุณลักษณะเป้าหมายเป็นการลดตัวแปรหากได้คุณลักษณะที่ดีจะช่วยให้ตัวแบบมีประสิทธิภาพและสามารถลดจำนวนคุณลักษณะลงได้ เมื่อพิจารณาจากผลลัพธ์ค่าความถูกต้องแม่นยำที่ได้

การวัดประสิทธิภาพแบบจำลองเพื่อให้ได้แบบจำลองที่มีประสิทธิภาพที่สุด ด้วยวิธี Cross-Validation ถือเป็นวิธีที่นิยมใช้กันอย่างแพร่หลายซึ่ง ธาดา จันตะคุณ (2560) เลือกใช้ในการ

ทดสอบประสิทธิภาพเพื่อหาแบบจำลองการพยากรณ์ความเป็นไปได้ในการเลือกสมัครสาขาวิชา นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) นำวิธี 10-folds Cross-Validation เพื่อวัดประสิทธิภาพแบบจำลองในการวิเคราะห์ปัจจัยในการคัดเลือกคุณสมบัติและการพยากรณ์ การวิจัยนี้ใช้วิธี Split Test วัดประสิทธิภาพแบบจำลอง สอดคล้องกับสุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง (2560)เปรียบเทียบประสิทธิภาพวิธีจำแนกกลุ่มการเป็นโรคไตเรื้อรัง โดยแบ่งข้อมูลเป็นเรียนรู้สร้างตัวแบบและชุดทดสอบในอัตราส่วน 70 และ30 เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มด้วยวิธีทางด้านสถิติ

การวัดประสิทธิภาพของแบบจำลอง การวิจัยนี้นำคุณลักษณะที่ได้จากการคัดเลือกมาสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bays) และเทคนิคป่าสุ่ม (Random Forest) เพื่อเปรียบเทียบหารูปแบบการจำแนกประเภทของข้อมูลที่มีความเหมาะสมกับข้อมูลมากที่สุด โดยได้นำวิธีการวัดค่าประสิทธิภาพแบบจำลองด้วยวิธี 5-fold Cross Validation วิธี 10-fold Cross Validation วิธี Split Validation (70:30) และ วิธี Split Validation (80:20) ซึ่งเป็นการทดสอบประสิทธิโดยอาศัยการเรียนรู้จากข้อมูลซึ่งประกอบด้วยชุดข้อมูลที่นำเข้าไปเรียนรู้และชุดข้อมูลที่เป็นผลลัพธ์ที่ต้องการจากนั้นนำมาเปรียบเทียบค่าทางสถิติโดยใช้ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) และค่าความถ่วงดุลและค่าประสิทธิภาพโดยรวม (F-measure) มาเปรียบเทียบด้วยวิธีทางด้านสถิติ สอดคล้องกับพรเทพ คงไชย และรัชฎา คงคะจันทร์ (2554) นำมาใช้ในการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับการทำเหมืองข้อมูลเพื่อพยากรณ์โอกาสการศึกษาของนักศึกษา และ Osiris Villacampa (2015) วิเคราะห์เปรียบเทียบการคัดเลือกคุณลักษณะและความแตกต่างด้วยค่า Accuracy, ค่า Area Under Receiver Operating Characteristic Curve (AUC), ค่า F-Measure, ค่า TF Rate และค่า FP Rate เพื่อลดมิติข้อมูลการคัดเลือกคุณลักษณะ

จากการเปรียบเทียบประสิทธิภาพแบบจำลองโดยนำเทคนิค Random เทคนิค Naïve Bayes เทคนิค Decision Tress มาใช้ร่วมกับคุณลักษณะที่ดีที่สุด พบว่าแบบจำลองด้วยเทคนิค Random Forest มีความถูกต้องแม่นยำและมีประสิทธิภาพมากที่สุด ซึ่งสอดคล้องกับงานวิจัยของสำราญ วานนท์และรจนา เมืองแสน (2563) ที่พัฒนาตัวแบบคุณลักษณะความเหมาะสมกับการเลือกสมัครสาขาเรียน โดยเปรียบเทียบประสิทธิภาพของเทคนิค Random Forest และเทคนิค Decision Tree ผลการเปรียบเทียบเทคนิค Random Forest มีค่าความถูกต้องสูงที่สุด และยังสอดคล้องกับงานวิจัยของสำราญ วานนท์ ธรัช อารีราษฎ์และจรัญ แสนราช (2561) นำข้อมูลผลการเรียน อาชีพของบิดาและมารดา เพื่อหาอาชีพที่เหมาะสมสำหรับนักศึกษาปริญญาตรีสาขาคอมพิวเตอร์ โดยใช้เทคนิค Decision Tree และ Random Forest มาเปรียบเทียบเพื่อพยากรณ์พบว่าเทคนิค Random Forest ให้ค่าความแม่นยำมากที่สุด

จากการวิจัยการคัดเลือกคุณลักษณะสำคัญที่มีสัมพันธ์กับการสมัครเรียนของนักศึกษาใหม่และการวัดประสิทธิภาพแบบจำลองโดยใช้คุณลักษณะที่สำคัญ 5 คุณลักษณะและผลการเปรียบเทียบแบบจำลองเทคนิค Decision Tree กับการวัดประสิทธิภาพแบบ Cross Validation ให้ค่าความแม่นยำสูงที่สุดคิดเป็นร้อยละ 71.93 ซึ่งการดำเนินการวิจัยด้านการวัดผลประสิทธิภาพแบบจำลองด้วยวิธี Cross Validation มีความคล้ายคลึงกับงานของธาดา จันตะคุณ (2560) การ

พยากรณ์ความเป็นได้ในการเลือกสมัครสาขาวิชาเรียน ซึ่งใช้คุณลักษณะในการพิจารณาจำนวน 9 คุณลักษณะด้วยเทคนิค Decision Tree มีความถูกต้องแม่นยำถึงร้อยละ 83.97 นั้นแต่ด้วยคุณลักษณะของข้อมูลในการเก็บรวบรวมและการนำมาวิเคราะห์อยู่ในค่าของข้อมูลที่ต่างกันจึงทำให้ผลลัพธ์จากการวัดประสิทธิภาพของเทคนิคได้ผลออกมาต่างกันคืองานวิจัยนี้เทคนิคที่ดีที่สุดคือเทคนิคป่าสุ่ม (Random Forest) นอกจากนั้นผู้วิจัยได้พิจารณาคุณลักษณะที่มีความสำคัญที่ได้แก่ข้อมูลการเรียนในแต่ละวิชา ซึ่งคุณลักษณะข้อมูลของแต่ละวิชานั้นถือเป็นคุณลักษณะที่สำคัญในการใช้ Information Gain ดังที่โกเมศ อัมพวัน (2548) กล่าวว่าค่าเกณฑ์ความรู้หาได้จากค่าเอ็นโทรปี หากค่าเอ็นโทรปีมีค่าสูงจะหมายถึงข้อมูลมีการปะปนกันทำให้ยากต่อการสรุป โดยได้เปรียบเทียบกับการสุ่มของเหรียญ เมื่อนำมาเทียบกับงานวิจัยนี้หากมีคุณลักษณะข้อมูลของแต่ละสาขาวิชาตัวอย่างเช่นข้อมูลผู้สมัครที่เป็นเพศชายและเพศหญิง เมื่อมีข้อมูลการเลือกสาขาเพียงหนึ่งสาขาอาจจะทำให้สามารถทราบว่าเพศใดจะเลือกหรือไม่เลือก แต่เมื่อมีคุณลักษณะอื่นที่สามารถบ่งบอกลักษณะจำเพาะออกมาเพิ่มเช่น เพศชายที่จบสายวิทย์-คณิตจะเลือกสาขาวิชา เพศชายที่จบสายศิลป์จะไม่เลือกสาขาวิชา เป็นต้น โดยความน่าจะเป็นหากมีการสุ่มข้อมูลมาแล้วได้ผลลัพธ์ที่เท่ากันค่าเอนโทรปีจะมีค่ามาก แต่หากมีการสุ่มแล้วผลลัพธ์ออกมาไม่สม่ำเสมอและเป็นไปตามคาดค่าเอนโทรปีก็จะมีค่าน้อย ซึ่งหมายความว่าระดับความไม่แน่นอนในการที่จะได้ผลลัพธ์ที่ถูกต้องน้อยลง

3. ข้อเสนอแนะ

1. การศึกษาการวิจัยเรื่อง “การวิเคราะห์เชิงทำนายการสมัครเรียนของนักศึกษาใหม่ด้วยเทคนิคเหมืองข้อมูล คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่” เป็นต้นแบบในการจำลองควรนำผลการวิจัยที่ได้ไปศึกษาเพื่อเป็นแนวทางในการพัฒนาระบบสนับสนุน โดยใช้คุณลักษณะที่นำมาวิเคราะห์เป็นแนวทางการพัฒนาและสร้างระบบรายงาน เพื่อการวางแผนการรับสมัครนักศึกษาในปีถัดไป

2. การศึกษาการวิจัยนี้มีกลุ่มตัวอย่างและคุณลักษณะที่มีข้อจำกัดจากการเก็บข้อมูลและความเป็นส่วนตัวของข้อมูลทำให้อาจขาดคุณลักษณะที่สำคัญที่จะนำมาพิจารณาเพื่อให้ได้ความถูกต้องและความแม่นยำที่มากขึ้น ในกรณีที่จะใช้กับหลักสูตรอื่น ๆ อาจต้องพิจารณาเพิ่มคุณลักษณะที่มีความเฉพาะหรือบ่งบอกถึงความสัมพันธ์ที่เกี่ยวข้องกับหลักสูตร เพื่อให้การจำแนกทำได้อย่างมีประสิทธิภาพ

3. ศึกษาบริบทตลอดจนกลยุทธ์ของสถานศึกษา เพื่อให้มีความสอดคล้องกับกลยุทธ์และแนวทางการดำเนินทางยุทธศาสตร์และเก็บข้อมูลของกลุ่มเป้าหมายได้ถูกต้อง

4. การวิจัยในครั้งต่อไปควรพิจารณาคุณลักษณะที่มีความเฉพาะหรือบ่งบอกความสัมพันธ์ที่สำคัญ การจัดกลุ่มของข้อมูลในแต่ละคุณลักษณะและจำนวนข้อมูล บริบทของพื้นที่เป้าหมายเพื่อให้ผลที่ได้มีความถูกต้องและแม่นยำมากยิ่งขึ้น ตัวอย่างเช่นคุณลักษณะการประกอบอาชีพในแต่ละพื้นที่ คุณลักษณะความสนใจด้านกิจกรรม คุณลักษณะความสามารถพิเศษ คุณลักษณะกลุ่มโรงเรียนหรือพื้นที่การจัดการศึกษา เป็นต้น

บรรณานุกรม

บรรณานุกรม

กองแผนและนโยบายคณะครุศาสตร์. (2564). *การประเมินตนเอง.* เชียงใหม่: มหาวิทยาลัยราชภัฏเชียงใหม่

โกเมศ อัมพวัน. (2548*). วิธีการหากฎความสัมพันธ์แบบใหม่โดยต้นไม้แสดงรายการความถี่.* กรุงเทพมหานคร : จุฬาลงกรณ์มหาวิทยาลัย.

ขันทอง ปทุมชาติ และพิมริมทร์ คีรีนทร์. (2558). การวิเคราะห์พฤติกรรมการเลือกสมัครสาขาวิชาเรียนของนักศึกษาใหม่โดยใช้เทคนิคการเหมืองข้อมูล. ใน *การประชุมวิชาการระดับชาติ ครั้งที่ 2 สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏกำแพงเพชร* (หน้า 174-185). กำแพงเพชร: มหาวิทยาลัยราชภัฏกำแพงเพชร.

คำนาย อภิปรัชญาสกุล. (2557). *คู่มือซอฟต์แวร์การวางแผนทรัพยากรองค์กร (ERP).* กรุงเทพมหานคร: โฟกัสมีเดีย แอนด์ พับลิชชิ่ง.

ธนวรรณ วิสุทธิรัตน์. (2561). *การใช้เหมืองข้อมูล (Data Mining) ในการวิเคราะห์และสร้างตัวแบบความสัมพันธ์ระหว่างคุณภาพสังคมและความสุขของประชาชนในจังหวัดจันทบุรี.* สถาบันบัณฑิตพัฒนบริหารศาสตร์, จันทบุรี.

ธาดา จันตะคุณ. *การพยากรณ์ความเป็นไปได้ในการเลือกสมัครสาขาวิชาโดยใช้เทคนิคเหมืองข้อมูล.* มหาวิทยาลัยราชภัฏมหาสารคาม. 2559

นิภาพร ชนะมาร และพรรณี สิทธิเดช. (2557). การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์. *วารสารมหาวิทยาราชภัฏสกลนคร*, 6(12), 34-35.

นิสานันท์ พลอาสา. (2558). *การสร้างแบบจำลองการขายผลิตภัณฑ์และพยากรณ์ยอดขายประกันชีวิต โดยเทคนิคการทำเหมืองข้อมูล (กรณีศึกษา บริษัทประกันชีวิตแห่งหนึ่ง ไม่ได้ตีพิมพ์).* มหาวิทยาลัยธรรมศาสตร์, กรุงเทพมหานคร.

ปริญญา สงวนสัตย์. (2558). *Artificial Intelligence with Machine Learning.* นนทบุรี: ไอดีซี พรีเมียร์.

พรเทพ คงไชย และรัชฎา คงคะจันทร์. (2554). *การศึกษาเชิงเปรียบเทียบในการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับการทำเหมืองข้อมูลเพื่อพยากรณ์โอกาสการสำเร็จการศึกษาของนักศึกษา.* มหาวิทยาลัยธรรมศาสตร์, กรุงเทพมหานคร.

ภุมริน หรั่งน้อย คุณัญญา ส้มเกลี้ยง ปิยนันท์ เทียบศรไชย และประภาส ทองรัก. (2564). การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อแนะนำอาชีพด้านไอทีสำหรับนักศึกษาระดับปริญญาตรี กรณีศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี. ใน *การประชุมหาดใหญ่วิชาการระดับชาติและนานาชาติครั้งที่ 12 มหาวิทยาลัยหาดใหญ่* (หน้า 1572-1585). สงขลา: มหาวิทยาลัยหาดใหญ่.

รัชฎา เทพประสิทธิ์ และจรัญ แสนราช. การวิเคราะห์ปัจจัยที่มีผลต่อการเลือกสาขาวิชาของนักศึกษาระดับปริญญาตรี คณะครุศาสตร์ โดยใช้เทคนิคการทำเหมืองข้อมูล. *วารสารบัณฑิตศึกษา มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์.* 14(2563), 134-146.

วันวิสาข์ ชนะประเสริฐ. (2559). *การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อแนะนำอาชีพสำหรับนักศึกษาปริญญาตรีคณะโบราณคดีมหาวิทยาลัยศิลปากร*. มหาวิทยาลัยศิลปากร, กรุงเทพมหานคร.

วิจิตรสวัสดิ์ สุขสวัสดิ์ ณ อยุทธยา. (2555). *การประยุกต์ใช้การทำเหมืองข้อมูลในระบบโลจิสติกส์*. สืบค้นจาก http://www.freightmaxad.com/magazine/?p=2744.

วิชญ์พงศ์ ดรุณธรรม. (2561). *เจาะลึก Random Forest !!!— Part 2 of "รู้จัก Decision Tree, Random Forest, และ XGBoost* สืบค้นจาก https://medium.com/@witchapongdaroontham/เจาะลึก-random-forest-part-2-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c.

วิรัตน์ ชูนุ้ย. (2555). *การลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ*. มหาวิทยาลัยสงขลานครินทร์, สงขลา.

ศุภมณ จันทร์สกุล. (2561). เทคนิคเหมืองข้อมูลในการวิเคราะห์ข้อมูลทางการพยาบาล. *วารสารวิชาการมหาวิทยาลัยอีสเทิร์นเอเชีย.* 12(2), 83-96.

สายชล สินสมบูรณ์ทอง. (2558). *การทำเหมืองข้อมูล*. กรุงเทพมหานคร: จามจุรีโปรดักส์.

สุรพงค์ เอื้อวัฒนามงคล. (2559). *การทำเหมืองข้อมูล*. กรุงเทพมหานคร: สถาบันบัณฑิตพัฒนบริหารศาสตร์.

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง. (2560). การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์*. 25(5), 839-853.

สุวิมล สิทธิชาติ. (2560). การวิเคราะห์คุณลักษณะพื้นฐานการศึกษาด้วยเทคนิคเหมืองข้อมูล. *วารสารเทคโนโลยีสารสนเทศ*, 13(2), 20-27.

เสกสรรค์ วิลัยลักษณ์ วิภา เจริญภัณฑารักษ์ และดวงดาว วิชาดากุล. (2558). การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียน โรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์. *Veridian E-Journal Science and Technology Silpakorn University*, 2(2), 1-17.

สำราญ วานนท์ ธรัช อารีราษฎร์ และจรัลแสนราช. การศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรี สาขาคอมพิวเตอร์ โดยใช้เทคนิคเหมืองข้อมูล. *วารสารวิชาการการจัดการเทคโนโลยีสารสนเทศและนวัตกรรม คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม*. 5(2561), 164-171.

สำราญ วานนท์ และรจนา เมืองแสน. การศึกษาและพัฒนาตัวแบบพยากรณ์คุณลักษณะความเหมาะสมสำหรับการเลือกสมัครสาขาวิชาเรียนโดยใช้เทคนิคเหมืองข้อมูล. *วารสารวิทยาการจัดการ*. 7(2563), 135-152.

อัจจิมา มณฑาพันธ์. *การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม*. มหาวิทยาลัยศรีปทุม. 2560.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิ่ง เบื้องต้น*. กรุงเทพมหานคร: ดาต้าคิวบ์.

Hall, M. A. & Smith, L. A. (1998). Practical feature subset selection for machine learning. In C. McDonald(Ed.), Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98, Perth, 4-6 February, 1998(pp. 181-191). Berlin: Springer.

H. Liu and R. Setiono, "A probabilistic approach to feature selection—a filter solution," in Proceedings of International Conference on Machine Learning (ICML-96), edited by L. Saitta, Morgan Kaufmann Publishers, 1996, pp. 319- 327.

Jyoti, A., Nidhi, B. and Sanjeev, R. (2013). *A review on association rule mining algorithms. International Journal of Innovative Research in Computer and Communication Engineering*, 1(5), 1246-1251.

Mark A. Hall, Geoffrey Holmes. (2003). *Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering*, 15(3), 1437 – 1447.

Matthew N. Bernstein. (2020*). Information entropy (Foundations of information theory: Part 2)*. Retrieved February 14,2021, from https://mbernste.github .io /posts/entropy/

Osiris Villacampa. (2015). Feature Selection and Classification Methods for Decision Making: A Comparative Analysis. Ph.D. dissertation, Information Systems, College of Engineering and Computing, Nova Southeastern University.

Tan, Steinbach and Kumar. (2006). Introduction to Data Mining. United States of America: Pearson Education Limited.

Thomas Bayes and Richard Price. (1763). An Essay towards solving a Problem in the Doctrin of Chance. By the late Rev. Mr. Bayes, communicated by M.Price, in a letter in John Canton, A.M.F.R.S., Philosophical Transaction of the Royal Society of London, 53(0): 370-418.

ภาคผนวก

ภาคผนวก ก

บันทึกข้อความขอข้อมูลเพื่อใช้ในงานวิจัย

มหาวิทยาลัยราชภัฏเชียงใหม่
เลขรับ 10960
วันที่ 3 ต.ค. 2564
เวลา 10.19 น.

คณะครุศาสตร์
รับเลขที่ 3667
วันที่ 6 ต.ค. 2564
เวลา 14:13

คณะครุศาสตร์
รับเลขที่ 3451
วันที่ 21 ก.ย. 2564
เวลา 11:18

สำนักทะเบียนและประมวลผล
รับเลขที่ 2807
วันที่ 21 ก.ย. 2564
เวลา 14.10 น.

บันทึกข้อความ

ส่วนราชการ สำนักงานคณบดีคณะครุศาสตร์ โทร. ๕๕๐๐

ที่ อว ๐๖๑๒.๐๒.๐๑/พิเศษ **วันที่** ๒๑ กันยายน ๒๕๖๔

เรื่อง ขอความอนุเคราะห์ข้อมูลในระบบฐานข้อมูล

เรียน ผู้อำนวยการสำนักทะเบียนและประมวลผล

ด้วยข้าพเจ้า นายปรัชญารัก เวียงสงค์ ตำแหน่งนักวิชาการคอมพิวเตอร์ สังกัดสำนักงานคณบดีคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ ได้อนุญาตศึกษาต่อในระดับปริญญาโท สาขาวิชาเทคโนโลยีสารสนเทศและการสื่อสาร หลักสูตรวิทยาศาสตร์มหาบัณฑิต มหาวิทยาลัยสุโขทัยธรรมธิราช ทั้งนี้ในการศึกษาค้นคว้าและทำวิจัยในสาขาวิชาดังกล่าวมีความเกี่ยวข้องในการพัฒนาและประยุกต์ใช้ระบบสารสนเทศภายในคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ ที่อำนวยความสะดวกในการให้บริการนักศึกษา เช่น ระบบฐานข้อมูลนักศึกษาที่ออกฝึกประสบการณ์วิชาชีพครูในแต่ละภาคการศึกษา ระบบฐานข้อมูลการบริการงานวิชาการให้กับนักศึกษาโดยจัดเก็บในฐานข้อมูลเฉพาะของคณะครุศาสตร์และสามารถนำมาเป็นข้อมูลการบริการ ตลอดการตรวจสอบสถานะการให้บริการนักศึกษาผ่านระบบเทคโนโลยีสารสนเทศ

เพื่อให้การศึกษาค้นคว้าและการทำวิจัยเป็นไปด้วยความเรียบร้อย จึงขอความอนุเคราะห์ข้อมูลจากระบบฐานข้อมูลจากสำนักทะเบียนและประมวลผล ในการนำข้อมูลมาวิเคราะห์ให้มีความแม่นยำและถูกต้อง

จึงเรียนมาเพื่อโปรดพิจารณา

(นายปรัชญารัก เวียงสงค์)
ตำแหน่ง นักวิชาการคอมพิวเตอร์
คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่

① เรียน คณบดี
☐ เพื่อโปรดทราบ
☑ เพื่อโปรดพิจารณา
เห็นควร นำเสนอสำนักทะเบียน พิจารณา

(นายสุรชาติ ...)
รักษาการในตำแหน่งหัวหน้าสำนักงานคณบดี
21 ก.ย. 64

② เรียน ผอ. สำนักทะเบียนและประมวลผล
เพื่อโปรดพิจารณา

(ผู้ช่วยศาสตราจารย์เยี่ยมลักษณ์ อุดาการ)
คณบดีคณะครุศาสตร์

③ เรียน ผอ.สทป.
☐ เพื่อโปรดทราบ
☑ เพื่อโปรดพิจารณา
เห็นควร มอบราชการ ฝ่ายสารสนเทศ จัดเตรียมข้อมูล

มอบงานสารสนเทศ ฝ่ายสารสนเทศ ประสานนักศึกษา เพื่อทราบ ข้อมูลตามที่ร้องขอ เพื่อขอความอนุเคราะห์

22 ก.ย. 64

21 ก.ย. 2564 22-9-64

เรียน คณบดี
เพื่อโปรด ทราบ พิจารณา
เห็นควร

(นายสุรชาติ ถึงกาศ)
รักษาการในตำแหน่งหัวหน้าสำนักงานคณบดี

(ผู้ช่วยศาสตราจารย์เยี่ยมลักษณ์ อดาการ)
คณบดีคณะครุศาสตร์

(ผู้ช่วยศาสตราจารย์เกษรา ปัญญา)
ผู้อำนวยการสำนักทะเบียนและประมวลผล

(นายนนท์กฤต ผัดวัง)
เจ้าหน้าที่บริหารงานทั่วไปปฏิบัติการ
- 4 ต.ค. 2564

(นายภาคภูมิ ทะนุดี)
รักษาการในตำแหน่งผู้อำนวยการกองกลาง

รองอธิการบดี ปฏิบัติราชการแทน
อธิการบดีมหาวิทยาลัยราชภัฏเชียงใหม่

ข้อมูลรับสมัครปี 64 63 62 ในการสมัครคณะครุศาสตร์ทุกรอบ(นักเรียนอาจไม่มารายงานตัวก็ได้แต่สนใจสมัครคณะครุศาสตร์) + การรายงานตัว

ข้อมูลการรับสมัคร(ประกาศผลการคัดเลือก)	ข้อมูลการรายงานตัว
โรงเรียนที่จบ	ระดับการศึกษา ที่เรียนในปัจจุบัน
แผนการเรียนก่อนเข้าศึกษา (มัธยม)	หลักสูตรสาขาวิชาที่ศึกษา
เกรดเฉลี่ย ก่อนเข้าศึกษา (มัธยม)	ชั้นปีของนักศึกษา
หลักสูตรสาขาวิชาที่สมัคร	อาชีพของบิดา
ปีที่สมัคร	อาชีพของมารดา
	อาชีพของผู้ปกครอง
	ระดับชั้น
	หมู่เรียนเรียน
	เพศ
	สัญชาติ
	ศาสนา
	หมู่ที่
	ซอย
	ถนน
	ตำบล/แขวง
	อำเภอ/เขต
	จังหวัด
	รหัสไปรษณีย์
	สถานภาพของบิดา
	สัญชาติของบิดา
	สถานภาพของมารดา
	สัญชาติของมารดา
	สถานะครอบครัว

ประวัติผู้วิจัย

ชื่อ	นายปรัชญารัก เวียงสงค์
วัน เดือน ปีเกิด	8 มีนาคม 2524
สถานที่เกิด	อำเภอเมือง จังหวัดกาฬสินธุ์
ประวัติการศึกษา	วิทยาศาสตร์บัณฑิต มหาวิทยาลัยมหิดล พ.ศ. 2545
สถานที่ทำงาน	คณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่ จังหวัดเชียงใหม่
ตำแหน่ง	นักวิชาการคอมพิวเตอร์ สำนักงานคณบดีคณะครุศาสตร์ มหาวิทยาลัยราชภัฏเชียงใหม่