การใช้เทคนิคเหมืองข้อมูล เพื่อพยากรณ์การสำเร็จการศึกษา
Forecasting Graduation By Data Mining Techniques

กันต์ ศิระพรธนารัชต์ (Kan Siraphonthanarat)* และ ชุติพนธ์ ศรีสวัสดิ์ (Chutiphon Srisawat)*

Received: December 10, 2023
Revised: March 24, 2024
Accepted: May 29, 2024

*ผู้นิพนธ์ประสานงาน: ชุติพนธ์ ศรีสวัสดิ์ (Chutiphon Srisawat) อีเมล: chutiphon@psru.ac.th

DOI:10.14416/j.it.2025.v1.004

บทคัดย่อ

งานวิจัยนี้มีวัตุประสงค์เพื่อ 1) เพื่อศึกษาการคัดเลือกคุณลักษณะที่สำคัญใช้ในการวิเคราะห์ข้อมูล 2) เพื่อพัฒนาแบบจำลองพยากรณ์การสำเร็จการศึกษาของนักศึกษาคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม และ 3) เพื่อหาประสิทธิภาพจากแบบจำลองโดยใช้เทคนิคเหมืองข้อมูล (Data Mining) ข้อมูลที่นำมาวิเคราะห์รวบรวมจากกองบริการการศึกษา มหาวิทยาลัยราชภัฏพิบูลสงคราม ปีการศึกษา 2560-2562 จำนวน 1,082 ระเบียน 30 แอตทริบิวต์ ในการวิเคราะห์หาปัจจัยที่มีผลต่อการสำเร็จการศึกษาใช้วิธีการคัดเลือกคุณลักษณะ (Feature Selection) ใช้ค่า Information Gain และ Chi Squared โดยแต่ละเทคนิคจะมีการลดลิมิตปัจจัยที่มีค่าน้ำหนักน้อยออกตามเงื่อนไข มีข้อมูลทั้งหมด 15 ชุดข้อมูล ในการสร้างโมเดลจะแบ่งข้อมูลออกเป็น 2 ส่วน Training Data 80% และ Test Data 20% โดยใช้เทคนิคเหมืองข้อมูล 3 เทคนิค คือ ต้นไม้ตัดสินใจ (Decision Tree) ป่าไม้สุ่ม (Random Forest) และการเรียนรู้แบบเบย์ (Naïve Bayes) ในการทดสอบประสิทธิภาพของโมเดลใช้ 10-Fold Cross Validation ซึ่งวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) และค่า F1-Score ผลการทดลองพบว่าการลดลิมิตเมื่อพิจารณาจากค่า Information Gain จากชุดข้อมูล IG5 คู่กับตัวแบบจำลองป่าไม้สุ่ม (Random Forest) ให้ประสิทธิภาพความถูกต้องโดยรวมเหมาะสมที่สุด ซึ่งค่าความถูกต้อง (Accuracy) เท่ากับ 96.03% ค่าเฉลี่ย F1-score เท่ากับ 88.65%

คำสำคัญ: เหมืองข้อมูล พยากรณ์ การสำเร็จการศึกษา

Abstract

This study intends to: 1) investigate the selection of significant features for data analysis; 2) create a predictive model for students' academic success in Pibulsongkram Rajabhat University's Faculty of Science and Technology; and 3) use data mining techniques to evaluate the model's efficacy. The 1, 082 records with 30 features that made up the data evaluated in this study were obtained from Rajabhat Pibulsongkram University's educational services department during the academic years 2560-2562. The study used information gain, and Chi Squared as feature selection methods in the analysis to determine the elements influencing academic achievement. According to certain parameters, each strategy entailed lowering variables with low weights. There were fifteen sets of data in the dataset. The data were split into two categories for model creation: training data (80%) and test data (20%). The model's performance was assessed by 10-Fold Cross Validation using data mining techniques, specifically Decision Tree, Random Forest, and Naïve Bayes. Accuracy and F1-Score were the evaluation criteria. According to experimental results, the Random Forest model performed with the best overall accuracy when Information Gain values from dataset IG5 were matched with it. The accuracy was 96.03%, and the average F1-Score was 88.65%.

Keywords: Data Mining, Forecasting, Graduation.

*ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม
*Department of Computer Science, Faculty of Science and Technology, University Pibulsongkram Rajabhat University.

1. บทนำ

กระแสโลกยุคโลกาภิวัตน์ที่มีการเปลี่ยนแปลงอยู่ตลอดเวลาของการเมือง เศรษฐกิจ สังคมและเทคโนโลยี การศึกษาในปัจจุบันจึงเป็นส่วนสำคัญในการขับเคลื่อนสังคม และสร้างบุคลากรที่มีศักยภาพ เพื่อสามารถแข่งขันได้ในระดับสากล

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม [1] ที่เป็นองค์กรแห่งการเรียนรู้ภายใต้พันธกิจที่มุ่งเน้นสร้างคนเพื่อพัฒนาท้องถิ่น เป็นแหล่งผลิตบัณฑิตนักปฏิบัติที่มีคุณภาพ ซื่อสัตย์ อดทน เชี่ยวชาญด้านวิทยาศาสตร์และเทคโนโลยี สร้างสรรค์นวัตกรรมสีเขียวสู่สังคม มีการจัดการเรียนการสอน 2 ภาควิชา 18 สาขา ระดับปริญญาตรี 13 สาขา คือ คหกรรมศาสตร์ เคมี จุลชีววิทยา ชีววิทยา ฟิสิกส์ เทคโนโลยีสารสนเทศ วิทยาการคอมพิวเตอร์ สาธารณสุขศาสตร์ คณิตศาสตร์ วิทยาศาสตร์และเทคโนโลยีสิ่งแวดล้อม วิทยาการข้อมูลและการจัดการสารสนเทศ บูรณาการสุขภาพ ความงาม สปา และวิทยาศาสตร์ศึกษา ระดับปริญญาโท 2 สาขา คือ คหกรรมศาสตร์ และวิทยาศาสตร์และเทคโนโลยีศึกษา และระดับปริญญาเอก 3 สาขา คือ การจัดการสารสนเทศวิทยาศาสตร์และเทคโนโลยีศึกษา และคหกรรมศาสตร์ ปัญหาที่สำคัญของคณะวิทยาศาสตร์และเทคโนโลยี คือ นักศึกษาที่ไม่สำเร็จการศึกษาตามหลักสูตรที่กำหนดไว้ ถือเป็นความสูญเสียทั้งโอกาส และเวลา เพื่อให้สอดคล้องตามพระราชบัญญัติการศึกษาแห่งชาติ [2] ฉบับที่ 4 (พ.ศ. 2562) มาตรา 47 ว่าด้วยสถานศึกษาต้องมีระบบการประกันคุณภาพการศึกษาเพื่อพัฒนาคุณภาพและมาตรฐานการศึกษารองรับการประเมินคุณภาพทั้งภายในและภายนอก การได้รับการสนับสนุนจากสถานศึกษาก็มีส่วนสำคัญในการพัฒนาคุณภาพของนักศึกษา ช่วยสร้างโอกาสทางการศึกษารวมถึงการสำเร็จการศึกษา เพื่อนำไปสู่คุณภาพชีวิตของประชาชนในท้องถิ่นและประเทศชาติที่ดีขึ้น

จากความก้าวหน้าของเทคโนโลยีทางด้านปัญญาประดิษฐ์ การทำเหมืองข้อมูล ทำให้สถานศึกษาได้นำเครื่องมือและเทคนิคต่าง ๆ มาใช้ในด้านการศึกษาช่วยในการวิเคราะห์สกัดข้อมูลจากชุดข้อมูลที่มีปริมาณมาก ๆ และซับซ้อน เพื่อให้ได้องค์ความรู้ที่เป็นประโยชน์ สามารถนำมาบริหารจัดการ วางแผนและแก้ไขปัญหาต่าง ๆ ได้ เช่น การใช้เทคนิคเหมืองข้อมูลศึกษาปัจจัยที่มีผลต่อการเรียน อาทิ เช่น ในงานวิจัยของทิพย์หทัย ทองธรรมชาติ [3] ได้นำเสนอบทความการคัดเลือกคุณลักษณะเพื่อสร้างโมเดลสำหรับการพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูล หรือการใช้เทคนิคเหมืองข้อมูลเพื่อทำนายผลการเรียน ในงานวิจัยของ จิราภรณ์ เจริญยิ่ง [4] ได้นำเสนอบทความการพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner

จากที่กล่าวมา ผู้วิจัยได้เห็นถึงความสำคัญของปัญหา จึงมีแนวคิดใช้เทคนิคเหมืองข้อมูล เพื่อพยากรณ์การสำเร็จการศึกษาของนักศึกษาคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม ใช้ข้อมูลย้อนหลังระหว่างปีการศึกษา 2560-2562 และเลือกเฉพาะข้อมูลนักศึกษาภาคปกติ เพื่อมาวิเคราะห์พยากรณ์ผลการสำเร็จการศึกษา ด้วยโปรแกรม RapidMiner Studio Education 10.0 ใช้เทคนิคในกลุ่ม Classification คือ Decision Tree, Random Forest และ Naïve Bayes และหาค่าความถูกต้อง (Accuracy) และค่า F1-Score ค่าเฉลี่ยระหว่าง Precision กับ Recall เพื่อให้ได้โมเดลที่เหมาะสมที่สุด และนำผลวิจัยมาเป็นแนวทางในการปรับปรุงแก้ไขปัญหานักศึกษาที่อาจสำเร็จการศึกษาเกินระยะเวลาตามหลักสูตร

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data Mining)

เป็นกระบวนการที่ใช้อัลกอริทึมและคอมพิวเตอร์ช่วยในการวิเคราะห์เพื่อประมวลผล สำรวจ และค้นหารูปแบบของข้อมูลความสัมพันธ์ที่ซ่อนอยู่ เทคนิคการทำเหมืองข้อมูลแบ่งออกได้ 2 ประเภท คือ 1) เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เช่น เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule) และการแบ่งกลุ่มข้อมูล (Clustering) 2) เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) เช่น การจำแนกประเภทข้อมูล (Classification) และการประมาณค่า (Regression) [5]

2.2 การคัดเลือกคุณลักษณะ (Feature Selection)

เป็นเทคนิคในการช่วยลดจำนวนตัวแปร คัดเลือกปัจจัยที่สำคัญ ที่ส่งผลต่อการพยากรณ์ เพื่อลดมิติของข้อมูล และช่วยในการเรียนรู้ของเครื่องจักร (Machine Learning) เพิ่มประสิทธิภาพของแบบจำลองให้มากยิ่งขึ้น สามารถแบ่งได้เป็น 2 กลุ่มใหญ่ ดังนี้ [6]

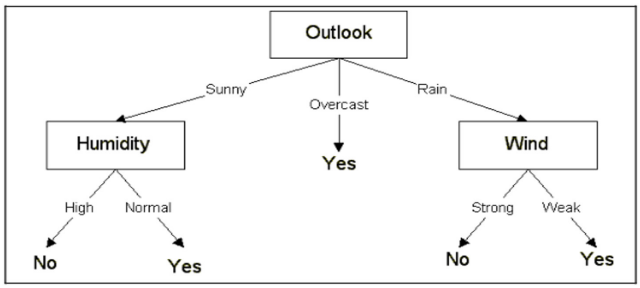
2.2.1 Wrapper Approach เป็นการคำนวณค่าน้ำหนักโดยใช้ Model แบ่งได้ 2 แบบ คือ Forward Selection และ Backward Elimination

2.2.2 Filter Approach เป็นการคำนวณหาค่าน้ำหนักความสัมพันธ์ระหว่างแต่ละฟีเจอร์และคลาสต่าง ๆ โดยเลือกฟีเจอร์ตามค่าน้ำหนักที่คำนวณได้ มีหลายวิธี เช่น Information Gain Chi-Square และ Correlation

2.3 เทคนิคการจำแนกประเภทข้อมูล (Classification Techniques)

เทคนิคการสร้างโมเดลจำแนกประเภทข้อมูล (Classification) มีหลายเทคนิค และผลลัพธ์ที่ได้ออกมาก็จะอยู่ในรูปแบบที่แตกต่างกัน ซึ่งมีรายละเอียดของแต่ละเทคนิค ดังนี้ [7]

2.3.1 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ โดยจุดเริ่มต้นคือโหนดที่อยู่ตำแหน่งบนสุดของแผนภาพต้นไม้ (Root Node) ไปจนถึงจุดสิ้นสุดที่โหนดใบ (Leaf Node) ในแต่ละโหนดการตัดสินใจจะแสดงการทดสอบคุณลักษณะ ของข้อมูลในแต่ละกิ่งก้านแสดงผลลัพธ์ที่เป็นไปได้จากการทดสอบ และนำไปสู่โหนดการตัดสินใจอีกโหนดหนึ่ง หรือสิ้นสุดอยู่ที่โหนดของใบ ซึ่งแสดงตัวอย่าง ดังภาพที่ 1



ภาพที่ 1 *การสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ Decision Tree*

ที่มาภาพ: N. Hongboonmee and P. Trepanichkul. 2562 [8]

2.3.2 เทคนิคนาอีฟเบย์ (Naïve Bayes) เป็นวิธีการที่อาศัยหลักการความน่าจะเป็นตามทฤษฎีของเบย์ (Bayes Theorem) ซึ่งเป็นเทคนิคที่ไม่ค่อยซับซ้อน อธิบายได้ด้วยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขสำหรับแต่ละความสัมพันธ์ โดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ ดังสมการที่ 1

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \quad (1)$$

โดย $P(A \mid B)$ คือ ความน่าจะเป็นในการเกิดเหตุการณ์ A โดยมี B เป็น Condition

$P(B \mid A)$ คือ ความน่าจะเป็นในการเกิดเหตุการณ์ B โดยมี A เป็น Condition

$P(A)$ คือ โอกาสในการเกิดเหตุการณ์ A จากเหตุการณ์ทั้งหมด

$P(B)$ คือ โอกาสในการเกิดเหตุการณ์ B จากเหตุการณ์ทั้งหมด

2.3.3 เทคนิคป่าไม้สุ่ม (Random Forest) เป็นอัลกอริทึมประเภทหนึ่งของต้นไม้ตัดสินใจที่มีลักษณะแบบไม่ตัดแต่งกิ่งหรือต้นไม้ถดถอยซึ่ง Random Forest มีหลักการทำงาน คือ จะแบ่งข้อมูลออกเป็นต้นไม้ตัดสินใจหลาย ๆ ต้น โดยแต่ละต้นจะได้รับคุณลักษณะและข้อมูลที่ไม่เหมือนกัน เพื่อทำให้ได้ต้นไม้ที่มีความหลากหลายและมีความอิสระต่อกันมากขึ้น ทำให้ประสิทธิภาพการทำงานสูงขึ้น และลดปัญหาการเกิด Overfitting ซึ่งอาจเกิดขึ้นใน Decision Tree

2.4 การวัดประสิทธิภาพของอัลกอริทึม

วิธีการทดสอบการตรวจสอบข้าม (Cross-Validation) เป็นเทคนิคประเมินประสิทธิภาพของแบบจำลอง โดยแบ่งข้อมูลออกเป็นส่วนย่อย ๆ เพื่อทำการฝึกสอนและทดสอบโมเดลจำนวนหลายรอบ (Folds) โดยแต่ละรอบจะใช้ข้อมูลในส่วนหนึ่งเป็นข้อมูลทดสอบ (Test Set) และใช้ข้อมูลในส่วนที่เหลือเป็นข้อมูลฝึกสอน (Training Set) เพื่อประเมินประสิทธิภาพในแต่ละรอบ [9]

2.5 งานวิจัยที่เกี่ยวข้อง

2.5.1 พิชัย ระเวงวัน และพุธษดี ศิริแสงตระกูล [10] ได้นำเสนอบทความ โมเดลเพื่อการพยากรณ์สถานภาพทางการศึกษาของนักศึกษา ซึ่งใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคการเรียนรู้แบบเบย์เชี่ยนเน็ตเวิร์ค (Bayesian Networks) และโลจิสติก รีเกรสชัน (Logistic Regression) โดยใช้ข้อมูลระหว่างปีการศึกษา 2551-2553 จำนวนทั้งสิ้น 2,272 คน 9 คุณลักษณะ โดยทดสอบประสิทธิภาพของโมเดลด้วยวิธีการ Cross Validation ผลการวิจัยพบว่า เทคนิคต้นไม้ตัดสินใจเป็นเทคนิคที่มีความเหมาะสมที่สุด โดยให้ค่าความถูกต้องเท่ากับ 82.85%

2.5.2 ธนพร คล้ายทอง และชุติพนธ์ ศรีสวัสดิ์ [11] ได้นำเสนอบทความการพยากรณ์การตกออกของนักศึกษาระดับปริญญาตรี มหาวิทยาลัยราชภัฏพิบูลสงคราม ด้วยเทคนิคเหมืองข้อมูลใช้เทคนิคต้นไม้ตัดสินใจ การเรียนรู้แบบเบย์เชี่ยนเน็ตเวิร์ค และกฎการอุปนัย (Rule Induction)

ข้อมูลทั้งหมด 20,093 รายการ 16 แอตทริบิวต์ แต่เลือกใช้แค่ 10 แอตทริบิวต์ โดยทดสอบประสิทธิภาพของโมเดล ด้วยวิธีการ 5-Fold Cross-Validation และ 10-Fold Cross-Validation ซึ่งวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) และค่าความผิดพลาด (Mean Absolute Error: MAE) ผลการวิจัยแบบจำลองที่เหมาะสม คือ เทคนิคต้นไม้ตัดสินใจได้ค่าสูงที่สุด ในการแบ่งข้อมูลทดสอบ 10-Fold Cross-Validation ให้ค่าความถูกต้อง (Accuracy) 97.81% และค่าความผิดพลาด (MAE) เท่ากับ 0.026

2.5.3 สุวิมล สิทธิชาติ [12] ได้นำเสนอบทความ การวิเคราะห์คุณลักษณะพื้นฐานทางการศึกษาด้วยเทคนิคเหมืองข้อมูล ซึ่งใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree-J48) และโครงข่ายประสาทเทียม (Artificial Neural Networks) ตัวแปรที่ใช้ คือ ข้อมูลส่วนบุคคล ผลการเรียน และปัจจัยอื่น ๆ รวม 50 คุณลักษณะ ผลกการวิจัยพบว่า ความแม่นยำในการจำแนกข้อมูลจาก 50 คุณลักษณะ วิธีโครงข่ายประสาทเทียม มีค่าความแม่นยำ 71.52% และ Decision Tree-J48 มีค่าความแม่นยำ 66.23% และหลังการคัดเลือกคุณลักษณะจาก 5 คุณลักษณะแรก พบว่า วิธีโครงข่ายประสาทเทียม ให้ค่าความถูกต้องสูงสุด คือ 80.13% และ Decision Tree-J48 ก็ให้ผลไปในทางเดียวกัน มีค่าความถูกต้องที่ 75. 83%

2.5.4 นนทศักดิ์ จันทร์ชุม และ ชลิตา ชีววิริยะนนท์ [13] ได้นำเสนอบทความ การใช้เทคนิคทางเหมืองข้อมูล เพื่อพัฒนาโมเดลการประเมินผลการเขียนโปรแกรมภาษาสแครช ซึ่งใช้เทคนิคการเรียนรู้แบบเบย์เซี่ยนเน็ตเวิร์ค เทคนิคต้นไม้ตัดสินใจและเทคนิคเพื่อนบ้านใกล้สุด จำนวน 113 ตัวอย่าง 9 คุณลักษณะ เป็นตัวแปรต้นและผลการเรียนเป็นตัวแปรทำนาย ผลวิจัยพบว่า เทคนิคต้นไม้ตัดสินใจทำนายผลการเรียนรู้การเขียนโปรแกรมภาษาสแครชได้สูงที่สุด จาก 3 โมเดล ซึ่งมีค่าความถูกต้อง 93.67%

2.5.5 สำราญ วานนท์ ธรัช อารีราษฎร์ และ จรัญ แสนราช [14] ได้นำเสนอบทความ การศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีสาขาคอมพิวเตอร์ โดยใช้เทคนิคเหมืองข้อมูล เทคนิคที่ใช้ต้นไม้ตัดสินใจ เทคนิคแรนดอมฟอร์เรส และเทคนิคแบ็กกิง ตัวแปรที่ใช้ คือ ข้อมูลส่วนตัว และผลการเรียน ประกอบด้วย 11 คุณลักษณะ ข้อมูลย้อนหลัง 5 ปี คือปี 2555-2559 จำนวน 65,335 ระเบียน ผลการวิจัยพบว่า เทคนิคการจำแนกข้อมูลด้วยแรนดอมฟอร์เรส ให้ความถูกต้องสูงที่สุด 84.29%

3. วิธีดำเนินการวิจัย

จากกรอบแนวคิดภาพที่ 2 สำหรับขั้นตอนในการดำเนินงานวิจัยประกอบด้วย 5 ระยะ ได้แก่ ระยะที่ 1 การเก็บรวบรวมข้อมูล (Data Collection) ระยะที่ 2 การเตรียมข้อมูล (Data Preprocessing) การทำความสะอาดข้อมูล (Data Cleaning) และการแปลงรูปแบบของข้อมูล (Data Transformation) ระยะที่ 3 การเลือกคุณลักษณะ (Feature Selection) ระยะที่ 4 การสร้างแบบจำลอง (Modeling) และระยะที่ 5 การประเมินประสิทธิภาพโมเดลที่เหมาะสม (Evaluation Optimal)

3.1 ระยะที่ 1 การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูล (Data Collection) ข้อมูลที่นำมาใช้ในการพยากรณ์การสำเร็จการศึกษา คือ ข้อมูล ของนักศึกษา คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม ใช้ข้อมูลย้อนหลังระหว่างปีการศึกษา 2560-2562 โดยรวบรวมมาจากกองบริการการศึกษา มหาวิทยาลัยราชภัฏพิบูลสงคราม ข้อมูลมีจำนวน 1,209 ระเบียน มีข้อมูลนักศึกษาอยู่ 2 ภาคการศึกษา คือ การศึกษาภาคพิเศษ และการศึกษาภาคปกติ ในการนำข้อมูลไปใช้ผู้วิจัยเลือกใช้ข้อมูลเฉพาะของนักศึกษาภาคปกติ จำนวนทั้งสิ้น 1,082 ระเบียน และประกอบด้วย 30 แอตทริบิวต์ ดังตารางที่ 1

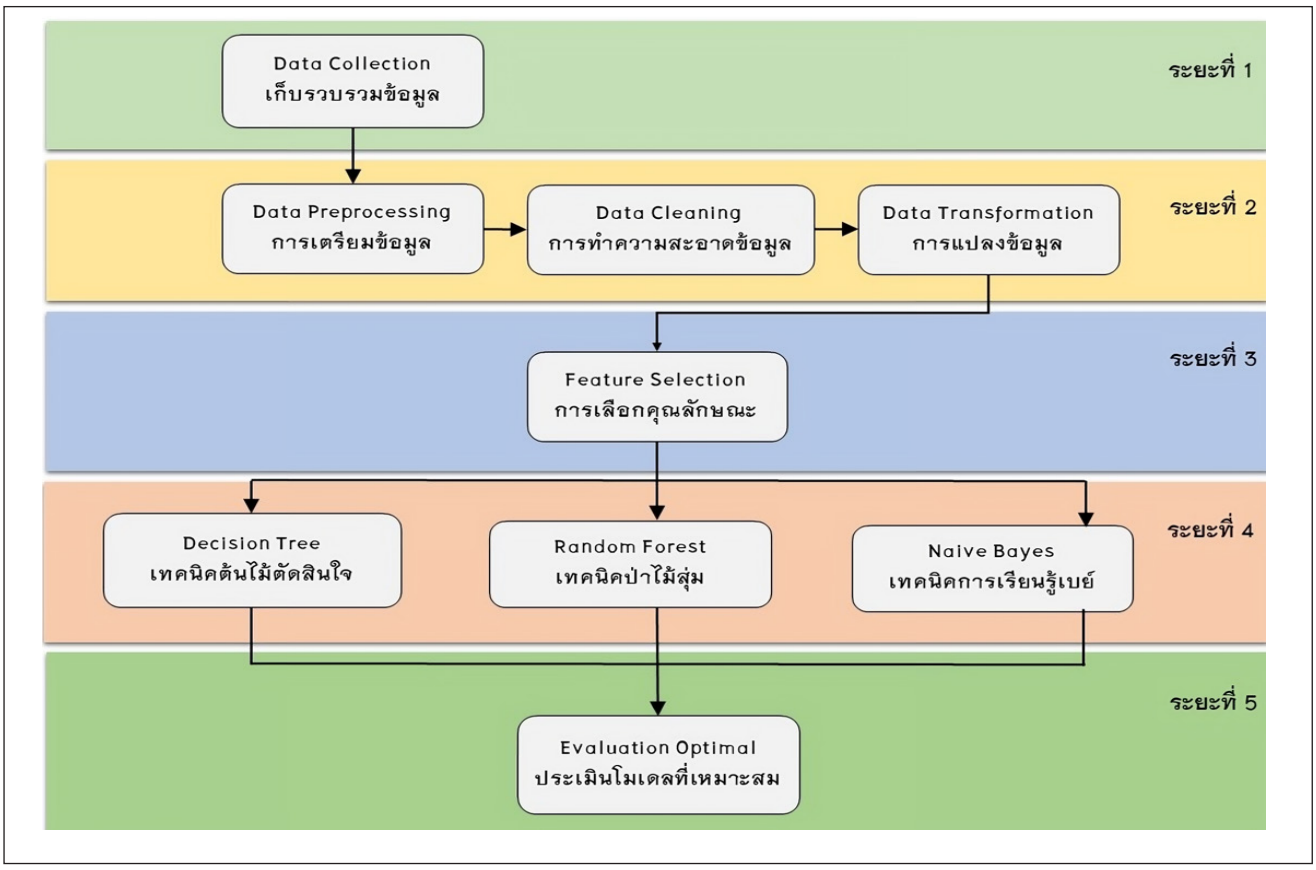
3.2 ระยะที่ 2 การเตรียมข้อมูล การทำความสะอาดข้อมูล และการแปลงรูปแบบของข้อมูล

3.2.1 การเตรียมข้อมูล (Data Preprocessing) เป็นการจัดระเบียบข้อมูลใหม่ของนักศึกษาคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงครามที่ได้รวบรวมมา โดยเก็บบันทึกข้อมูลอยู่ในรูปแบบไฟล์ .xlsx ให้เหมาะสมสำหรับการวิเคราะห์

3.2.2 การทำความสะอาดข้อมูล (Data Cleaning) เป็นกระบวนในการจัดการข้อมูลที่ผิดปกติเพื่อแก้ไขความไม่สมบูรณ์ของข้อมูล จัดการข้อมูลที่สูญหาย (Missing Value) และลดข้อมูลรบกวน (Noise) ก่อนการนำข้อมูลเข้าโปรแกรม RapidMiner Studio Education 10.0 เพื่อเลือกคุณลักษณะ (Feature Selection) จากข้อมูลที่ได้ทำการรวบรวมมา พบว่ามีข้อมูลที่เป็นค่าว่างซึ่งทำความสะอาดข้อมูล โดยกำจัดค่าว่างและแทนที่ค่าที่ผิดปกติที่ไม่สามารถนำข้อมูลเข้าโปรแกรม RapidMiner Studio Education 10.0 ได้

3.2.3 การแปลงรูปแบบของข้อมูล (Data Transformation) เป็นกระบวนการที่ปรับเปลี่ยนข้อมูลจากรูปแบบหนึ่งเป็นรูปแบบอื่น เพื่อทำให้ข้อมูลเหมาะสมสำหรับการวิเคราะห์หรือการใช้งาน จากชุดข้อมูลข้างต้น ผู้วิจัยได้ทำการแปลงรูปแบบของข้อมูลเกรด

จากประเภท Real ที่มีค่า 3.33, 2.65, 1.15 เป็นต้น ให้เป็นประเภท Nominal เพื่อให้ข้อมูลไม่ต่างกันเกินไปและเป็นประเภทเดียวกัน โดยผู้วิจัยจะเปลี่ยนตามระบบลำดับขั้น สัญลักษณ์ A, B+, B, C+, C, D+, D, และ F [15] แสดงตัวอย่างของข้อมูลที่จะนำไปใช้ในขั้นตอนถัดไป คือ การคัดเลือกคุณลักษณะ โดยมีรายละเอียดของปัจจัยดังตารางที่ 1



ภาพที่ 2 *กรอบแนวคิด*

ตารางที่ 1 *แสดงลักษณะข้อมูลที่จะนำมาวิเคราะห์ด้วยเทคนิคเหมืองข้อมูล*

ลำดับ	แอตทริบิวต์	ประเภท	คำอธิบาย	ค่าที่เป็นไปได้
1	ลำดับ	Integer	ไอดี	-
2	ความถนัด/ความ สามารถพิเศษ	Nominal	สิ่งที่ชอบและถนัด	{ด้านดนตรี, กีฬา เป็นต้น}
3	จังหวัด	Nominal	ภูมิลำเนาของนักศึกษา	{พิษณุโลก, พิจิตร เป็นต้น}
4	อาชีพบิดา	Nominal	อาชีพหรือการทำงานของบิดา	{เกษตร, ประมง, ค้าขาย เป็นต้น}
5	อาชีพมารดา	Nominal	อาชีพการทำ งานของมารดา	{เกษตร, ประมง, ค้าขาย เป็นต้น}
6	ปีที่เข้าศึกษา	Integer	ปีที่นักศึกษาเข้าศึกษา	{2560, 2561, 2562}
7	หลักสูตร	Nominal	หลักสูตรที่นักศึกษาเข้า	{ศิลปศาสตรบัณฑิต, วิทยาศาสตรบัณฑิต เป็นต้น}
8	สาขาวิชา	Nominal	สาขาวิชาที่นักศึกษาเข้า	{คหกรรมศาสตร์, เคมี เป็นต้น}
9	1T1	Nominal	เกรดเฉลี่ยปี 1 เทอม 1	{A, B+,B, C+, C, D+, D, F}
10	1T2	Nominal	เกรดเฉลี่ยปี 1 เทอม 2	{A, B+,B, C+, C, D+, D, F}
11	G1	Nominal	เกรดเฉลี่ยรวมปีการศึกษา 1	{A, B+,B, C+, C, D+, D, F}
12	G2	Nominal	เกรดเฉลี่ยรวมปีการศึกษา 1 และภาคฤดูร้อน ปี 1	{A, B+,B, C+, C, D+, D, F}
13	2T1	Nominal	เกรดเฉลี่ยปี 2 เทอม 1	{ A, B+,B, C+, C, D+, D, F}

ลำดับ	แอตทริบิวต์	ประเภท	คำอธิบาย	ค่าที่เป็นไปได้
14	2T2	Nominal	เกรดเฉลี่ยปี 2 เทอม 2	{ A, B+,B, C+, C, D+, D, F}
15	G3	Nominal	เกรดเฉลี่ยรวม ปีการศึกษา 1 และ 2	{ A, B+,B, C+, C, D+, D, F}
16	2T3	Nominal	เกรดเฉลี่ยปี 2 ภาคฤดูร้อน	{ A, B+,B, C+, C, D+, D, F}
17	G4	Nominal	เกรดเฉลี่ยรวม ปีการศึกษา 1,2 และ ภาคฤดูร้อน ปี 2	{ A, B+,B, C+, C, D+, D, F}
18	3T1	Nominal	เกรดเฉลี่ยปี 3 เทอม 1	{ A, B+,B, C+, C, D+, D, F}
19	3T2	Nominal	เกรดเฉลี่ยปี 3 เทอม 2	{ A, B+,B, C+, C, D+, D, F}
20	G5	Nominal	เกรดเฉลี่ยรวม ปีการศึกษา 1, 2 และ 3	{ A, B+,B, C+, C, D+, D, F}
21	3T3	Nominal	เกรดเฉลี่ยปี 3 ภาคฤดูร้อน	{ A, B+,B, C+, C, D+, D, F}
22	G6	Nominal	เกรดเฉลี่ยรวม ปีการศึกษา 1,2,3 และภาคฤดูร้อน ปี 3	{ A, B+,B, C+, C, D+, D, F}
23	4T1	Nominal	เกรดเฉลี่ยปี 4 เทอม 1	{ A, B+,B, C+, C, D+, D, F}
24	4T2	Nominal	เกรดเฉลี่ยปี 4 เทอม 2	{ A, B+,B, C+, C, D+, D, F}
25	G7	Nominal	เกรดเฉลี่ยรวม ปีการศึกษา 1,2,3 และ 4	{ A, B+,B, C+, C, D+, D, F}
26	4T3	Nominal	เกรดเฉลี่ยปี 4 ภาคฤดูร้อน	{ A, B+,B, C+, C, D+, D, F}
27	G8	Nominal	เกรดเฉลี่ยรวม ปีการศึกษา 1,2,3,4 และภาคฤดูร้อน ปี 4	{ A, B+,B, C+, C, D+, D, F}
28	Total	Nominal	เกรดเฉลี่ยรวม ทั้งหมด	{ A, B+,B, C+, C, D+, D, F}
29	สถานะ นักศึกษา	Nominal	เป็นคำตอบ	{อนุมัติผล, ออกตามระเบียบ}

3.3 ระยะที่ 3 การเลือกคุณลักษณะ (Feature Selection)

งานวิจัยนี้ผู้วิจัยเลือกใช้เทคนิค Filter Approach ใช้ 2 ค่า คือ Information Gain และ Chi Squared ในการคำนวณค่าของแต่ละมิติข้อมูล โดยใช้โปรแกรม RapidMiner Studio Education 10.0

3.3.1 เทคนิค Information Gain ในการทำงานหลักจะใช้ Probability หรือความน่าจะเป็น เมื่อนำมารวมกัน Entropy เหมาะกับเงื่อนไขที่ไม่ได้ซับซ้อน ค่า Information Gain จะอยู่ในช่วงระหว่าง 0 ถึง 1 ค่าที่สูงแสดงถึงคุณสมบัติที่ให้ข้อมูลที่เป็นประโยชน์มากในการแบ่งข้อมูล ดังสมการที่ 2 แสดงการคำนวณค่า Information Gain หรือค่า Entropy ของชุดข้อมูลทั้งหมด สมการที่ 3 แสดงการคำนวณค่า Entropy ของชุดมิติข้อมูลในแต่ละลักษณะ สมการที่ 4 เป็นการคำนวณหาค่า Information Gain สำหรับการพิจารณามิติข้อมูลคุณลักษณะ [16]

$$E(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (2)$$

โดย $E(D)$ คือ Entropy ของชุดข้อมูล D

$\sum$ คือ การบวกผลรวมทุกค่าของ i ที่มีค่าเริ่มต้นจาก $i = 1$ และสิ้นสุดที่ $i = n$

p_i คือ ความน่าจะเป็นที่ข้อมูลจะอยู่ในกลุ่มหรือคลาสที่ i

$\log_2$ คือ ฟังก์ชัน log ฐาน 2

$$E_A(D) = -\sum_{j=1}^{m} \frac{|D_j|}{D} E(D_j) \quad (3)$$

โดย $E_A(D)$ คือ Conditional Entropy ของชุดข้อมูล D เมื่อแบ่งด้วยตัวแบ่ง A

$|D_j|$ คือ จำนวนตัวอย่างกลุ่มย่อย D_j ที่เกิดจากการแบ่งชุดข้อมูล D ด้วยตัวแบ่ง A

D คือ จำนวนตัวอย่างทั้งหมดในชุดข้อมูล D

$E(D_j)$ คือ Entropy ของกลุ่มย่อย D_j ซึ่งคำนวณได้จากสมการของ Entropy $E(D_j)$ โดยใช้ข้อมูลจากกลุ่มย่อย D_j

$$Gain(A) = E(D) - E_A(D) \quad (4)$$

โดย $Gain(A)$ คือ Information Gain ที่ได้จากการแบ่งชุดข้อมูล D เมื่อใช้ตัวแบ่ง A

$E(D)$ คือ Entropy ของชุดข้อมูล D

$E_A(D)$ คือ Conditional Entropy ของชุดข้อมูล D เมื่อแบ่งด้วยตัวแบ่ง A

3.3.2 เทคนิค Chi-Square มีหลักการคำนวณน้ำหนักของแอตทริบิวต์ ด้วยหลักการทางสถิติไคสแควร์ ค่าน้ำหนักของแอตทริบิวต์สูงถือว่ามีความเกี่ยวข้องสูง [17] สมการที่ 5

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (5)$$

โดย *O1, O2 ...On* คือ ความถี่ของตัวแปรที่ได้จากการศึกษา
E1, E2 ...En คือ ความถี่ที่คาดหวัง (หรือความถี่ที่ควรจะเป็น)

ในงานวิจัยครั้งนี้ผู้วิจัยได้มีการทดลองลดลิมิตของปัจจัยของแต่ละเทคนิคออกเป็น 5 ชุดข้อมูล จากการหาค่าน้ำหนักของแอททริบิวต์แล้วจึงค่อยเลือกชุดข้อมูลตามเกณฑ์ โดยตัดปัจจัยที่มีน้ำหนักน้อยออกเพื่อนำไปสร้างโมเดลและทดสอบประสิทธิภาพ เริ่มจากการใช้ปัจจัยทั้งหมด 28 แอททริบิวต์ และลดปัจจัยที่มีค่าน้ำหนักน้อยออกให้เหลือ 20 แอททริบิวต์ เหลือ 15 แอททริบิวต์ เหลือ 10 แอททริบิวต์ และเหลือ 5 แอททริบิวต์ ก็จะมีข้อมูลทั้งหมด 10 ชุดข้อมูล เพื่อนำไปสร้างโมเดลและเปรียบเทียบประสิทธิภาพของโมเดลในการพยากรณ์การสำเร็จการศึกษาของนักศึกษา ดังตารางที่ 2

ตารางที่ 2 *ชุดข้อมูลที่นำไปสร้างโมเดล*

ลำดับ	ชื่อชุดข้อมูล	Attribute	หมายเหตุ
1	IG1	28	เรียงตามค่า Information Gain
2	IG2	20	
3	IG3	15	
4	IG4	10	
5	IG5	5	
6	Ch1	28	เรียงตามค่า Chi Squared
7	Ch2	20	
8	Ch3	15	
9	Ch4	10	
10	Ch5	5	

3.4 ระยะที่ 4 การสร้างแบบจำลอง (Modeling)

เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมน์นิง ซึ่งในขั้นตอนนี้หลายเทคนิคจะถูกนำมาใช้เพื่อให้ได้คำตอบที่ดีที่สุด [18] หลังจากการเลือกคุณลักษณะ จะแบ่งข้อมูลเป็น 2 กลุ่ม ผ่านโอเปอร์เรเตอร์ Split Data กลุ่มข้อมูลที่จะใช้สอน (Training Data) 80% เพื่อเทรนโมเดลและใช้กลุ่มข้อมูลทดสอบ (Test Data) 20% ทดสอบโมเดล โดยจะใช้ 3 เทคนิคในการสร้างโมเดล ได้แก่

3.4.1 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) โดย Operator ที่ใช้ได้แก่ 1) Decision Tree ใช้สำหรับสร้างโมเดล Decision Tree 2) Apply Model ใช้สำหรับ predict ข้อมูลใหม่และ 3) Performance ใช้สำหรับแสดงตัวชี้วัดของโมเดล Classification

3.4.2 เทคนิคป่าไม้สุ่ม (Random Forest) โดย Operator ที่ใช้ได้แก่ 1) Random Forest ใช้สำหรับสร้างโมเดล Random Forest 2) Apply Model ใช้สำหรับ predict ข้อมูลใหม่และ 3) Performance ใช้สำหรับแสดงตัวชี้วัดของโมเดล Classification

3.4.3 เทคนิคการเรียนรู้แบบเบย์ (Naive Bayes) โดย Operator ที่ใช้ได้แก่ 1) Naïve Bayes ใช้สำหรับสร้างโมเดล Naïve Bayes 2) Apply Model ใช้สำหรับ predict ข้อมูลใหม่และ 3) Performance ใช้สำหรับแสดงตัวชี้วัดของโมเดล Classification

3.5 ระยะที่ 5 การประเมินประสิทธิภาพโมเดลที่เหมาะสม (Evaluation Optimal)

ในการทดสอบประสิทธิภาพของโมเดลจะใช้ 10 Folds Cross Validation-Test โดยจะแบ่งข้อมูลออกเป็นกลุ่มเท่า ๆ กัน ในการทดสอบจะใช้ข้อมูล 9 ส่วน เป็นข้อมูลชุดสอน และใช้ข้อมูล 1 ส่วน เป็นชุดทดสอบ เพื่อทดสอบประสิทธิภาพของโมเดล โดยจะทำวนจนครบ 10 รอบ ซึ่ง Operator ที่ใช้ในการประเมินประสิทธิภาพ คือ Cross-Validation ในส่วนของการวัดประสิทธิภาพจะใช้ค่า Accuracy และ F1-Score [19]

3.5.1 Accuracy เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุก Class คือ ค่าที่โมเดลทายถูกเทียบกับค่าคำตอบทั้งหมด ดังสมการ 6

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

โดย *TP* คือ ค่าทำนายที่ทำนายว่าจริงซึ่งตรงกับค่าจริง
TN คือ ค่าทำนายที่ทำนายว่าไม่จริงซึ่งตรงกับค่าจริง
FP คือ ค่าทำนายที่ทำนายว่าไม่จริงซึ่งไม่ตรงกับค่าจริง
FN คือ ค่าทำนายที่ทำนายว่าจริงซึ่งไม่ตรงกับค่าจริง

3.5.2 F1-Score คือ ค่าเฉลี่ยระหว่าง Precision และ Recall เป็น Single Metric ที่วัดความสามารถของโมเดล ไม่ต้องเลือกระหว่าง Precision หรือ Recall ดังสมการ 7

$$F1 = 2\left(\frac{\mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call}\right) \quad (7)$$

โดย *F*1 คือ ผลลัพธ์ค่าเฉลี่ย
Pr*ecision* คือ ค่าความแม่นยำ
Re*call* คือ ค่าความครบถ้วน

ตารางที่ 3 *เปรียบเทียบผลการทดสอบประสิทธิภาพโมเดล Decision Tree, Random Forest และ Naive Bayes*

	Select Attributes	**Evaluation** of Model Classification Techniques					
		Information Gain					
		Decision Tree		**Random Forest**		**Naive Bayes**	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
	IG1	94.09	84.55	95.66	87.52	94.83	86.71
	IG2	94.09	84.57	95.72	87.53	94.92	86.71
	IG3	94.09	84.85	95.75	87.70	94.92	86.84
	IG4	95.01	86.17	95.85	87.97	94.92	86.94
Data set 217	IG5	95.11	86.40	96.03	88.65	94.92	87.10
		Chi Squared					
		Decision Tree		**Random Forest**		**Naive Bayes**	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
	Ch1	94.09	84.57	95.66	84.44	94.83	86.71
	Ch2	94.18	85.02	95.66	87.33	94.92	86.94
	Ch3	94.09	86.17	95.66	87.38	94.92	87.10
	Ch4	94.09	86.17	95.66	87.53	94.92	87.10
	Ch5	95.57	87.44	96.03	88.58	95.57	87.22

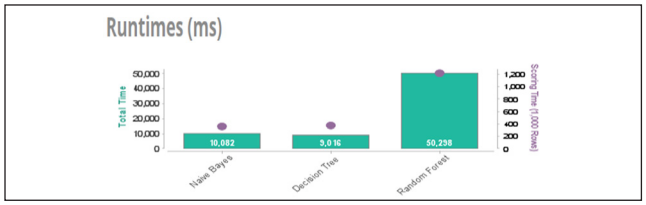
4. สรุป

การเปรียบเทียบผลการทดสอบประสิทธิภาพโมเดล Decision Tree, Random Forest และ Naive Bayes จากการเลือกคุณลักษณะ ทั้ง 2 เทคนิค Information Gain (IG) และ Chi Squared โดยมีชุดข้อมูล 10 แบบ ผลการทดสอบประสิทธิภาพของโมเดลจากกลุ่มข้อมูลทดสอบ (Test Data) 20% ใช้วิธีการ 10 Folds Cross Validation-Test ในการพยากรณ์การสำเร็จการศึกษาของนักศึกษา แสดงดังตารางที่ 3

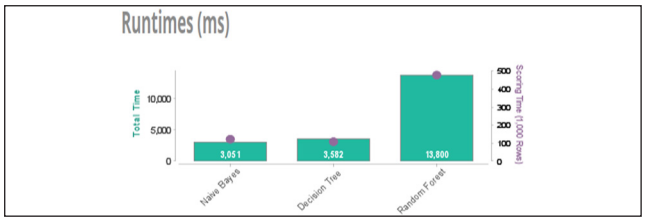
จากตารางที่ 3 สามารถสรุปได้ว่าผลการเปรียบเทียบการทดสอบประสิทธิภาพโมเดล Decision Tree, Random Forest และ Naive Bayes จากรูปแบบของชุด ข้อมูลทั้ง 10 แบบ ที่มีความแตกต่างกันตามเงื่อนไข พบว่าการลดลิมิตของปัจจัย โดยพิจารณาจากค่า Information Gain ตามชุดข้อมูล IG5 คือ 5 แอตทริบิวต์ คู่กับตัวแบบจำลองเทคนิคป่าไม้สุ่ม (Random Forest) ให้ประสิทธิภาพความถูกต้องโดยรวมออกมาเหมาะสมที่สุดในการพยากรณ์การสำเร็จการศึกษาของนักศึกษา ซึ่งค่าความถูกต้อง (Accuracy) เท่ากับ 96.03% และค่าเฉลี่ย (F1-score) ระหว่าง Precision และ Recall เท่ากับ 88.65% รองลงมาคือ การพิจารณาจากค่า Chi Squared ตามชุดข้อมูล Ch5 คือ 5 แอตทริบิวต์ คู่กับตัวแบบจำลองเทคนิคป่าไม้สุ่ม (Random Forest) ซึ่งค่าความถูกต้อง (Accuracy) เท่ากับ 96.03% และค่าเฉลี่ย (F1-score) ระหว่าง Precision และ Recall เท่ากับ 88.58% และจากชุดข้อมูล IG5 กับ Ch5 มีปัจจัยที่ใช้และน้ำหนักของแต่ละปัจจัยแสดงดังตารางที่ 4 และในการลดลิมิตของข้อมูลยังช่วยลดระยะเวลาในการวิเคราะห์ข้อมูลเพื่อสร้างตัวแบบจำลอง แสดงดังภาพที่ 3-4

ตารางที่ 4 ปัจจัยที่ใช้ชุดข้อมูล IG5 และ Ch5

ชุดข้อมูล IG5			ชุดข้อมูล Ch5		
แอตทริบิวต์	คำอธิบาย	น้ำหนัก	แอตทริบิวต์	คำอธิบาย	น้ำหนัก
2T1	เกรดเฉลี่ยปี 2 เทอม 1	1	G8	เกรดเฉลี่ยรวม ปีการศึกษา 1, 2, 3, 4 และภาคฤดูร้อน ปี 4	1
2T2	เกรดเฉลี่ยปี 2 เทอม 2	0.996	G7	เกเกรดเฉลี่ยรวม ปีการศึกษ 1, 2, 3 และ 4	1
G8	เกรดเฉลี่ยรวม ปีการศึกษา 1, 2, 3, 4 และภาคฤดูร้อน ปี 4	0.992	G5	เกรดเฉลี่ยรวม ปีการศึกษา 1, 2, และ 3	0.998
G7	เกรดเฉลี่ยรวม ปีการศึกษ 1, 2, 3 และ 4	0.991	G6	เกรดเฉลี่ยรวม ปีการศึกษา 1, 2, 3 และภาคฤดูร้อน ปี 3	0.998
Total	เกรดเฉลี่ยรวมทั้งหมด	0.987	Total	เกรดเฉลี่ยรวมทั้งหมด	0.994



ภาพที่ 3 แสดงระยะเวลาจากการใช้ข้อมูลทั้งหมด



ภาพที่ 4 แสดงระยะเวลาจากการใช้ข้อมูลที่มีการลดลิมิต

โดยภาพที่ 4 แสดงเวลาในการรันโมเดลเมื่อใช้ปัจจัยทั้งหมด ตัวอย่างเทคนิค Random Forest ใช้เวลารันประมาณ 50 วินาที และภาพที่ 5 แสดงเวลาในการรันโมเดลเมื่อลดลิมิตของข้อมูลลง ตัวอย่างเทคนิค Random Forest ชุดข้อมูล IG5 ใช้เวลารันประมาณ 13 วินาที ลดเวลารันลง 26%

5. อภิปรายผล

งานวิจัยนี้มีวัตุประสงค์เพื่อศึกษาการคัดเลือกคุณลักษณะที่สำคัญ เพื่อพัฒนาแบบจำลองพยากรณ์การสำเร็จการศึกษา และเพื่อหาประสิทธิภาพจากแบบจำลอง ผลจากการคัดเลือกคุณลักษณะที่สำคัญพบว่าค่าจาก Information Gain จากชุดข้อมูล IG5 คือ 5 แอตทริบิวต์ที่เกี่ยวข้อง 1) เกรดเฉลี่ยปี 2 เทอม 1 2) เกรดเฉลี่ยปี 2 เทอม 2 3) เกรดเฉลี่ยรวม ปีการศึกษา 1, 2, 3, 4 และภาคฤดูร้อน ปี 4 4) เกรดเฉลี่ยรวม ปีการศึกษา 1, 2, 3 และ 4 5) เกรดเฉลี่ยรวมทั้งหมด คู่กับตัวแบบจำลองป่าไม้สุ่ม (Random Forest) ให้ประสิทธิภาพโดยรวมออกมาเหมาะสมที่สุดในการพยากรณ์การสำเร็จการศึกษาของนักศึกษา ซึ่งค่าความถูกต้อง (Accuracy) เท่ากับ 96.03% และค่าเฉลี่ย (F1-score) ระหว่าง Precision และ Recall เท่ากับ 88.65% จากประสิทธิภาพที่ได้ของโมเด ล Random Forest เป็นไปในทิศทางเดียวกับบทความของ สำราญ วานนท์, ธรัช อารีราษฎร์ และ จรัญ แสนราช [14] และผลการวิจัยมีความแตกต่างจากบทความของ พิชัย ระเวงวัน และพุธษดี ศิริแสตระกูล [10] และนนทศักดิ์ จันทร์ชุม และชลิตา ชีววิริยะนนท์ [13] ที่พบว่าเทคนิคต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความถูกต้องสูงที่สุด

ผลจากการวิเคราะห์การคัดเลือกคุณลักษณะไม่ว่าจะพิจารณาจากค่า Information Gain หรือ Chi Squared ที่เหมาะสมสามารถเพิ่มประสิทธิภาพของโมเดลให้ดีขึ้นได้ และยังสรุปได้อีกว่าปัจจัยที่มีผลต่อการพยากรณ์การสำเร็จการศึกษาจากชุดข้อมูล IG5 พบว่าปัจจัยที่สำคัญเป็นเกรดเฉลี่ยผลการเรียนที่ให้ค่าน้ำหนักมาก แสดงให้เห็นถึงความสำคัญและความอยากง่ายของสาขาวิชาที่นักศึกษากำลังเรียนอยู่ สอดคล้องกับบทความของ ธนพร คล้ายทอง และชุติพนธ์ ศรีสวัสดิ์ [11] โดยในการช่วยเหลือสามารถนำผลที่ได้ให้อาจารย์ที่ปรึกษาหรือผู้ที่เกี่ยวข้องสามารถวางแผนแก้ไขรูปแบบการสอนเพื่อให้ความช่วยเหลือกับนักศึกษาที่อาจสำเร็จการศึกษาเกินระยะเวลาได้ ทั้งนี้ในส่วนของปัจจัยอื่น ๆ ที่ไม่ได้ถูกเลือกใช้เพื่อการพยากรณ์ หากดูรายละเอียดในเชิงลึกแล้ว มีความสำคัญต่อการสำเร็จการศึกษาของนักศึกษาเช่นกัน เช่น ปัจจัยอาชีพของบิดามารดาที่ไม่มั่งคง มีผลต่อการตกออกของนักศึกษาในระหว่างเรียนเนื่องจากความไม่พร้อมของครอบครัวและความยากจน ทำให้ขาดโอกาสในการศึกษาต่อให้จบตามหลักสูตร รวมถึงความสามารถหรือความถนัดในสายวิชาที่เลือกเรียนส่งผลให้นักศึกษาออกกลางคัน เนื่องจากพึ่งได้ค้นพบตัวเอง

และต้องการศึกษาต่อในสาขาวิชาชีพที่ตนเองคาดหวัง เพราะไม่ต้องการเสียเวลาในสาขาวิชาที่เลือกเรียนในครั้งแรก และหากมีการดำเนินงานครั้งต่อไป เพิ่มปัจจัยที่เกี่ยวข้องกับนักศึกษามากขึ้น เช่น แรงจูงใจหรือเป้าหมาย สถานภาพครอบครัว ปัจจัยทางเศรษฐกิจ เป็นต้น เพื่อนำมาใช้วิเคราะห์และคัดเลือก พัฒนาตัวแบบจำลองในการพยากรณ์ให้มีประสิทธิภาพมากยิ่งขึ้น และจะช่วยในการวางแผนในการบริหารจัดการในการลดจำนวนการตกออกระหว่างทางของผู้เรียนได้

6. ข้อเสนอแนะ

จากการวิจัยครั้งนี้เป็นการศึกษาปัจจัย และเทคนิคในการพยากรณ์การสำเร็จการศึกษา เพื่อเป็นแนวทางในการทำวิจัยครั้งต่อไป ผู้วิจัยจึงขอเสนอแนวทางดังนี้

6.1 ในงานวิจัยนี้ผู้วิจัยได้เลือกใช้เทคนิค Decision Tree, Random Forest และ Naive Bayes ในการสร้างแบบจำลอง ซึ่งยังมีเทคนิคอื่น เช่น Neural Network, K-Nearest Neighbors (KNN) และ Linear Regression เป็นต้น หากมีการใช้เทคนิคร่วมกันอาจจะส่งผลให้ประสิทธิภาพแม่นยำมากขึ้น

6.2 ควรมีการศึกษาปัจจัยที่เกี่ยวข้องเพิ่มเติมเพื่อนำมาใช้วิเคราะห์และคัดเลือก ในพัฒนาตัวแบบจำลองในการพยากรณ์ให้มีประสิทธิภาพมากยิ่งขึ้น และช่วยให้มีผลนำไปใช้ประกอบการวางแผนการทำงานที่เกี่ยวข้องกับการตกออกของนักศึกษาในระหว่างเรียนได้

7. กิตติกรรมประกาศ

ขอขอบคุณกองบริการการศึกษา มหาวิทยาลัยราชภัฏพิบูลสงคราม ที่อนุเคราะห์ข้อมูลเพื่อนำมาใช้ในงานวิจัยในครั้งนี้

8. เอกสารอ้างอิง

[1] SC., *Vision Mission Philosophy.* Available Online at http://202.29.80.54/vision/, accessed on 05 April 2023.

[2] Office of Educational Quality Assurance, *Educational quality assurance system.* Available Online at https://qa.chandra.ac.th/index.php, accessed on 15 June 2023.

[3] T. Thongthammachart. "The Feature Selection to Creating Models for Predicting Learning Achievement using Data Mining Techniques." *Report following the 4th national academic conference Kamphaeng Phet Rajabhat University,* pp. 338-347, 2017.

[4] J. Jareanying. *The Prediction of Student Performance Using Data Mining Techniques with RapidMiner.* Master's Thesis, Information Technology Program Faculty of Science Srinakharinwirot University, 2020.

[5] S. Sinsomboon. *Data Mining.* 1st ed., Bangkok: Jamjuree Product, 2015.

[6] P. N. wichian, P. Manair, Y. Chuchuen, and S. Mak-on. "Optimization Feature Selection for Classification of Manuscript Grouping." *Journal of Science and Technology Songkla University,* Vol. 1, No. 1, January-June, 2020.

[7] S. Euawatthanamongkol. *Data Mining.* 2nd ed., Bangkok: National Institute of Development Administration, (n.d.), 2019.

[8] N. Hongboonmee and P. Trepanichkul. "Comparison of Data Classification Efficiency to Analyze Risk Factors that Affect the Occurrence of Hyperthyroidusing Data Mining Techniques." *Journal of Information Science and Technology,* Vol. 9, No. 1, pp. 41-51, January-June, 2019.

[9] K. Satangmongkol. *K-Fold Cross Validation.* Available Online at https://datarockie.com/blog/k-fold-cross-validation/, accessed on 10 June 2023.

[10] P. Rawengwan and P. Seresangtakul. "A model for forecasting educational status of students." *Proceedings of the Graduate Research Presentation Conference National and International Levels Khon Kaen University,* Vol. 10, pp. 273-283, March, 2017.

[11] T. Klaythong and C. Srisawat. "Forecasting Dropout of Undergraduates Pibulsongkram Rajabhat University with Data Mining Technique." *Journal of Applied Informatics and Technology,* Vol. 5, No. 1, pp. 1-17, January-June, 2023.

[12] S. Sittichat. "Study of Educational Attributes Using Data Mining Technique." *Information Technology Journal,* Vol. 13, No. 2, pp. 20-28, July-December, 2017.

[13] N. Janchum and C. Cheewaviriyanon. "Using Data Mining Techniques to Develop a Model for Scratch Programming Assessment." *Information Technology Journal,* Vol. 18, No. 1, pp. 96-105, January-June, 2022.

[14] S. Vanont, T. Areerat, and C. Saenrat. "A Study of Techniques in Predicting Career Counseling for Undergraduate Students of the Computer Program by Using Data Mining Technique." *Journal of Technology Management Rajabhat Maha Sarakham University,* Vol. 5, No. 1, January-June, 2018.

[15] REG PSRU, *Measurement of educational evaluation.* Available Online at https://reg.psru.ac.th/reg2018/student.php, accessed on 20 June 2023.

[16] W. Jaidee and N. Wannapee. "The Study of Factors Affecting for On-time Graduation of Ungraduated Student Using Feature Selection Technique on Imbalanced Datasets." *Journal of Information Science and Technology,* Vol. 10, No. 1, pp. 75-84, January-June, 2020.

[17] A. Montaphan. "Comparison of Feature Selection Methods to Improve Breast Cancer Prediction." *Royal Thai Air Force Medical Gazette,* Vol. 65, No. 2, pp. 49-56, May-August, 2019.

[18] A. Phutthala and S. Saensri. "The Searching Relationship of Results High School and Bachelor in Case Study: The Graduate Student in Year 2017 at KU.CSC." *The 8th Asia Undergraduate Conference on Computing (AUCC 2020),* pp. 277-285, 2023.

[19] D. Hunthong, T. Ngernwilai, and S. Sinsomboonthong. "Efficiency Comparison inReplace Missing Value Using Regression Imputation, Multiple Imputation and Expectation Maximizationfor Classification in Data Mining." *Thai Journal of Science and Technology,* Vol. 9, No. 5, pp. 575-588, September-October, 2020.