



การจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี
สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้วิธีการเรียนรู้แบบรวมกลุ่ม
Classifying Dropout for Undergraduate Students in The Department of Statistics:
Using Ensemble Method

นายศิริพัฒน์ จานเชื้อ

รหัสประจำตัว 613020198-8

นางสาวยุวลักษณ์ ดวงมะลา

รหัสประจำตัว 613020196-2

รายงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต
หลักสูตรสถิติ สาขาวิชาสถิติ คณะวิทยาศาสตร์
มหาวิทยาลัยขอนแก่น
ปีการศึกษา 2564

การจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี
สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้วิธีการเรียนรู้แบบรวมกลุ่ม
**Classifying Dropout for Undergraduate Students in The Department of Statistics:
Using Ensemble Method**

นายศิริพัฒน์	งานเนื่อง	รหัสประจำตัว 613020198-8
นางสาวยุวลักษณ์	ดวงมะลา	รหัสประจำตัว 613020196-2

รายงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต
หลักสูตรสถิติ สาขาวิชาสถิติ คณะวิทยาศาสตร์
มหาวิทยาลัยขอนแก่น
ปีการศึกษา 2564

หลักสูตรสถิติ สาขาวิชาสถิติ คณะวิทยาศาสตร์

มหาวิทยาลัยขอนแก่น

ปีการศึกษา 2564

หัวข้อโครงงานวิจัย

การจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาในระดับปริญญาตรี
สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้วิธีการเรียนรู้แบบรวมกลุ่ม

นักศึกษา

นายศิริพัฒน์ จานแข็ง รหัสประจำตัว 613020198-8

นางสาวชวลักษณ์ ดวงมะลา รหัสประจำตัว 613020196-2

อาจารย์ที่ปรึกษา

อ.ดร.จิตรจิรา ไชยฤทธิ์ ที่ปรึกษา

อ.ดร.พิชญ์ วิรัชโชติเสถียร ที่ปรึกษาร่วม

สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น อนุมัติให้รายงานฉบับนี้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรวิทยาศาสตรบัณฑิต (สถิติ)

.....*จิตรจิรา ไชยฤทธิ์*.....อาจารย์ที่ปรึกษา
(อาจารย์ ดร.จิตรจิรา ไชยฤทธิ์)

วันที่ 10 เดือน พฤษภาคม พ.ศ. 2565

.....*พิชญ์ วิรัชโชติเสถียร*.....อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร.พิชญ์ วิรัชโชติเสถียร)

วันที่ 10 เดือน พฤษภาคม พ.ศ. 2565

.....หัวหน้าสาขาวิชาสถิติ
(รองศาสตราจารย์ ดร.วิชุดา ไชยสีวามงคล)
วันที่ 10 เดือน พฤษภาคม พ.ศ. 2565

หัวข้อโครงการวิจัย	การจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี	
	สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้วิธีการเรียนรู้แบบรวมกลุ่ม	
นักศึกษา	นายศิริพัฒน์ จานแข็ง	รหัสประจำตัว 613020198-8
	นางสาวยุวลักษณ์ ควงมะลา	รหัสประจำตัว 613020196-2
อาจารย์ที่ปรึกษา	อ.ดร.จิตรจิรา ไชยฤทธิ์	ที่ปรึกษา
	อ.ดร.พิชญ์ วิรัชโชติเสถียร	ที่ปรึกษาร่วม

บทคัดย่อ

ในปัจจุบันการฟื้นฟูสภาพของนักศึกษาระดับอุดมศึกษาเป็นปัญหาที่สำคัญลำดับต้น ๆ ที่สถาบันการศึกษา สาขาวิชา และหน่วยงานที่เกี่ยวข้องต้องให้ความสำคัญในการแก้ปัญหา อีกทั้งยังส่งผลด้านลบต่อประวัติของนักศึกษา การใช้เทคนิคการทำเหมืองข้อมูล (Data Mining Techniques) ในการวิเคราะห์ปัจจัยที่ส่งผลต่อการฟื้นฟูสภาพของนักศึกษานำไปสู่การวางแผนและจัดการเพื่อลดการฟื้นฟูสภาพของนักศึกษาที่จะเกิดขึ้นในอนาคต งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาโดยใช้วิธีการเรียนรู้แบบเดี่ยว (Single Model) และแบบรวมกลุ่ม (Multiple Model) ของเทคนิคการทำเหมืองข้อมูล และศึกษาปัจจัยที่ส่งผลต่อการฟื้นฟูสภาพของนักศึกษา โดยอาศัยข้อมูลพื้นฐานของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ ปีการศึกษา 2558 – 2564 นำมาผ่านกระบวนการทำความสะอาดข้อมูล (Clean Data) สร้างตัวแปรหุ่น (Dummy Variable) ทำการคัดเลือกข้อมูล (Selection Data) ปรับสมดุลข้อมูล (Balancing Data) และปรับปรุงขอบเขตข้อมูล (Data Scaling) จากนั้นใช้เทคนิคเหมืองข้อมูล ประกอบด้วยวิธีการสร้างต้นไม้ช่วยตัดสินใจ (Decision Tree) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) สำหรับการจำแนกประเภทแบบเดี่ยว และใช้วิธีการ Bagging, Boosting, Random Forest สำหรับการจำแนกประเภทแบบรวมกลุ่ม ผลจากการเปรียบเทียบประสิทธิภาพการจำแนกประเภทแสดงให้เห็นว่า วิธีการการจำแนกประเภทแบบรวมกลุ่มให้ประสิทธิภาพในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาดีกว่าวิธีการการจำแนกประเภทแบบเดี่ยว การจำแนกประเภทแบบรวมกลุ่มโดยใช้ Random Forest ให้ประสิทธิภาพที่ดีที่สุดโดยให้ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) ค่าความถ่วงดุล (F1-Score) มากกว่า 98% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC มากกว่า 99%

นอกจากนี้ยังพบว่าปัจจัยสำคัญที่มีผลต่อการพัฒนาของนักศึกษาได้แก่ เกรดเฉลี่ย, เกรดเฉลี่ยระดับมัธยม, รายได้บิดา, รายได้มารดา, อายุมารดา, อายุบิดา การคิด F รายวิชา เกรด F รายวิชา STATISTICAL ANALYSIS I, ELEMENTARY PHYSICS, CALCULUS FOR PHYSICAL SCIENCE II, GENERAL CHEMISTRY LABORATORY ตามลำดับ

คำสำคัญ : เหมืองข้อมูล, การเรียนรู้แบบเดี่ยว, การเรียนรู้แบบรวมกลุ่ม, การพัฒนาของนักศึกษา

สาขาวิชาสถิติ

ปีการศึกษา 2564

ลายมือชื่อนักศึกษา.....

ศิริพัฒน์

(นายศิริพัฒน์ จานแข็ง)

ลายมือชื่อนักศึกษา.....

ยวลักษณ์ ดวงมะลา

(นางสาวยวลักษณ์ ดวงมะลา)

ลายมือชื่ออาจารย์ที่ปรึกษา

จตุรจิรา ไชยฤทธิ์

(อาจารย์ ดร.จตุรจิรา ไชยฤทธิ์)

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

พินญา วิรัชโชติเสถียร

(อาจารย์ ดร.พินญา วิรัชโชติเสถียร)

Classifying Dropout for Undergraduate Students in The Department of Statistics: Using Ensemble Method

Student	Mr. Siripat Jankhuang	Student ID 613020198-8
	Miss Yuwaluck Duangmala	Student ID 613020196-2
Project Advisor	Dr. Jitjira Chaifarit	
	Dr. Pitchaya Wiratchotisatian	

Abstract

At present, the dropout of undergraduate students is a major problem for educational institutions, disciplines, and related agencies must focus on solving problems. It also negatively affects the student's profile. Using data mining techniques to analyze factors affecting student dropout leads to planning and management to reduce dropout in the future. This research compared the classification efficiency for student dropout by using a single-model and multiple-model techniques, and study the factors affecting the dropout of a student. The information used in this research is obtained from undergraduate student in Department of Statistics, Faculty of Science, in the academic year 2015-2021. The process includes cleaning data, creating dummy variables, selecting data, balancing data, and enhancing data scaling. The data mining techniques for single classification used in this study include decision tree and support vector machine. Whereas, the data mining techniques for ensemble classification include the Bagging, Boosting, Random Forest method for ensemble classification. The results of the classification efficiency comparison showed that the ensemble classification methods were more effective in classifying students' dropout than the single classification methods. Ensemble classification using Random Forest revealed the optimum performance with over 98% of accuracy and F1-score, and over 99% of the area under the ROC curve. The essential factors affecting the dropout of a student found by this study were GPA, high school grade point average, father's income, maternal income, mother's age, father's age, and received F grade in STATISTICAL ANALYSIS I, ELEMENTARY PHYSICS, CALCULUS FOR PHYSICAL SCIENCE II, GENERAL CHEMISTRY LABORATORY, respectively.

Keywords: Data Mining, Single Model, Multiple Model, Dropout for Students

Department of Statistics

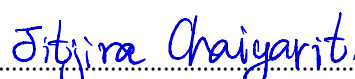
Academic year 2021

Signature of student..... 


(Mr. Siripat Jankhuang)

Signature of student..... 

(Miss Yuwaluck Duangmala)

Signature of project advisor..... 

(Dr. Jitjira Chaiyarit)

Signature of project co-advisor..... 

(Dr. Pitchaya Wiratchotisat)

กิตติกรรมประกาศ

โครงการฉบับนี้สำเร็จลุล่วงไปได้ด้วยความกรุณาและการอนุเคราะห์จาก อ.ดร.จิตรจิรา ไชยฤทธิ์ อาจารย์ที่ปรึกษางานวิจัยทางสถิติ และ อ.ดร.พิชญา วิรัชโชติเสถียร อาจารย์ที่ปรึกษาร่วม ซึ่งได้เสียสละเวลาให้คำปรึกษา คำแนะนำ ความรู้ ตลอดจนการตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ในโครงการ ขอขอพระคุณอาจารย์เป็นอย่างสูงไว้ ณ ที่นี้

การจัดทำโครงการฉบับนี้ได้สำเร็จลุล่วงตามวัตถุประสงค์ คณะผู้จัดทำต้องขอขอบคุณ นายอดิศักดิ์ ศรีรัตนประพันธ์ นักวิชาการคอมพิวเตอร์ ชำนาญการสำนักบริหารและพัฒนาวิชาการ มหาวิทยาลัยขอนแก่น ที่ให้ความอนุเคราะห์ในการให้ข้อมูลแก่ผู้ศึกษานำมาวิเคราะห์ในโครงการครั้งนี้ รวมทั้ง ผศ.ดร.ธิปไตย พงษ์ศาสตร์ และ ผศ.ดร.กฤษณา พัฒนากุล กรรมการสอบ ที่ให้ความรู้ และคำแนะนำในการแก้ไขโครงการให้ถูกต้องและสมบูรณ์ยิ่งขึ้น

คณะผู้จัดทำหวังเป็นอย่างยิ่งว่าโครงการฉบับนี้จะเป็นประโยชน์ในการพัฒนาหลักสูตรการเรียนการสอน นักศึกษาในสถาบัน หรือผู้ที่มีความสนใจในโครงการฉบับนี้ หากโครงการฉบับนี้มีข้อผิดพลาดประการใด คณะผู้จัดทำก็ขออภัยมา ณ ที่นี้ด้วย

ศิริพัฒน์ จานแข็ง
ยุวลักษณ์ ดวงมะลา

สารบัญเนื้อหา

บทที่ 1.....	1
บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของการวิจัย	2
1.3 สมมติฐานของการวิจัย.....	2
1.4 ขอบเขตของการวิจัย	3
1.5 ความหมายหรือนิยามคำศัพท์เฉพาะ	3
1.6 ประโยชน์ที่คาดว่าจะได้รับการวิจัย.....	3
บทที่ 2.....	4
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	4
2.1 ทฤษฎีการฟื้นฟูสภาพ	5
2.1.1 การฟื้นฟูสภาพ.....	5
2.1.1 ระเบียบมหาวิทยาลัยขอนแก่น ว่าด้วยการศึกษาชั้นปริญญาตรี พ.ศ. 2562	5
2.2 การทำเหมืองข้อมูล (Data Mining)	7
1) ทฤษฎีการทำเหมืองข้อมูล.....	7
2) ประเภทการทำเหมืองข้อมูล.....	7
3) การทำเหมืองข้อมูล.....	8
2.3 เทคนิคการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกประเภท (Imbalanced Data Problem Solving in Classification Technique).....	8
1) วิธีสุ่มเกิน.....	9
2) วิธีสุ่มลด	9

สารบัญเนื้อหา (ต่อ)

3) วิธีผสมผสาน	9
4) วิธีสังเคราะห์ข้อมูลเพิ่ม	9
2.4 เทคนิคการจำแนกประเภทข้อมูล (Data Classification Techniques)	10
1) ต้นไม้ตัดสินใจ (Decision Tree: DT)	10
2) ซัพพอร์ตเวกเตอร์แมชชีน	11
2.5 การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning)	12
1) Bagging Method	13
2) Boosting Method	15
3) แรนดอมฟอเรส	16
2.6 มาตรวัดและการทดสอบประสิทธิภาพแบบจำลอง	17
1) การวัดประสิทธิภาพแบบจำลอง	17
2) Receiver Operating Characteristic (ROC) curve	18
2.7 งานวิจัยที่เกี่ยวข้อง	19
2.7 กรอบแนวคิดงานวิจัย	22
บทที่ 3	23
วิธีการดำเนินการวิจัย	23
3.1 เครื่องมือที่ใช้ในการวิจัย	23
3.1.1 เทคนิคที่ใช้ในการเหมืองข้อมูล	23
3.1.2 โปรแกรมที่ใช้ในการประมวลผล	24
3.2 กรอบวิธีการดำเนินการวิจัย	24

สารบัญเนื้อหา (ต่อ)

3.3 การเตรียมข้อมูลสำหรับการใช้ในการสร้างแบบจำลองการจำแนกประเภท.....	26
บทที่ 4.....	44
ผลการวิจัย.....	44
4.1 กระบวนการจัดเตรียมข้อมูลและการสำรวจข้อมูลเบื้องต้น	45
4.2 ผลลัพธ์อัลกอริทึมสำหรับการจำแนกประเภทแบบเดี่ยว.....	53
4.3 ผลลัพธ์อัลกอริทึมสำหรับการจำแนกประเภทแบบรวมกลุ่ม	58
4.4 การเปรียบเทียบประสิทธิภาพความถูกต้องในการจำแนกประเภท.....	70
บทที่ 5.....	78
สรุปผลการวิจัย	78
5.1 สรุปผลการวิจัย	78
5.2 อภิปรายผลการวิจัย	79
5.3 ประโยชน์ของสถิติที่ใช้ในการวิเคราะห์.....	80
5.4 ข้อเสนอแนะ	80
เอกสารอ้างอิง	82
ภาคผนวก ก	84

สารบัญภาพ

ภาพที่ 1 Decision Tree	11
ภาพที่ 2 Support Vector Machine	12
ภาพที่ 3 แสดงโครงสร้างพื้นฐานของการเรียนรู้แบบรวมกลุ่ม	13
ภาพที่ 4 โครงสร้างวิธีการทำงานแบบ Bagging.....	14
ภาพที่ 5 แสดงโครงสร้างวิธีการทำงานแบบ Boosting (AdaBoost)	15
ภาพที่ 6 Random Forest	17
ภาพที่ 7 ROC curve	19
ภาพที่ 8 กรอบแนวคิดการวิจัย.....	22
ภาพที่ 9 กรอบวิธีการดำเนินงานวิจัยสำหรับการทำแนกประเภท	25
ภาพที่ 10 ขั้นตอนการแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้ DT	32
ภาพที่ 11 ขั้นตอนการแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้ SVM.....	33
ภาพที่ 12 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Bagging โดยใช้ DT เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	35
ภาพที่ 13 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Bagging โดยใช้ SVM เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	36
ภาพที่ 14 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Boosting โดยใช้ DT เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	38
ภาพที่ 15 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Boosting โดยใช้ SVM เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	39
ภาพที่ 16 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Random Forest.....	41
ภาพที่ 17 ขั้นตอนการทำการพยากรณ์สภาพของนักศึกษาปีการศึกษา 2564	43
ภาพที่ 18 แผนภูมิวงกลมสถานะนักศึกษา	45
ภาพที่ 19 ปัจจัยที่ส่งผลต่อการฟื้นสภาพของนักศึกษาของการฟื้นสภาพกรณีที่ 1	71
ภาพที่ 20 ปัจจัยที่ส่งผลต่อการฟื้นสภาพของนักศึกษาของการฟื้นสภาพกรณีที่ 2	73
ภาพที่ 21 ปัจจัยที่ส่งผลต่อการฟื้นสภาพของนักศึกษาของการฟื้นสภาพกรณีที่ 3	74
ภาพที่ 22 แผนภูมิวงกลมของการทำการพยากรณ์สภาพนักศึกษกรณีที่ 1.....	75
ภาพที่ 23 แผนภูมิวงกลมของการทำการพยากรณ์สภาพนักศึกษกรณีที่ 2.....	76

สารบัญภาพ (ต่อ)

ภาพที่ 24 แผนภูมิวงกลมของการทำนายการฟื้นสภาพนักศึกษากรณี 3.....	77
---	----

สารบัญตาราง

ตารางที่ 1 Confusion Matrix	17
ตารางที่ 2 เปรียบเทียบวิธีการสร้างแบบจำลอง ความแม่นยำ และปัจจัยที่มีความสำคัญกับการฟื้นฟูสภาพของ นักศึกษาระหว่างการศึกษาต่างๆ.....	21
ตารางที่ 3 ชื่อตัวแปรและความหมาย.....	26
ตารางที่ 4 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Decision Tree (DT).....	31
ตารางที่ 5 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Support Vector Machine (SVM).....	31
ตารางที่ 6 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Bagging โดยใช้ Decision Tree (DT) และ Support Vector Machine (SVM) เป็นอัลกอริทึมพื้นฐาน	34
ตารางที่ 7 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Boosting โดยใช้ Decision Tree (DT) และ Support Vector Machine (SVM) เป็นอัลกอริทึมพื้นฐาน	37
ตารางที่ 8 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Random Forest.....	40
ตารางที่ 9 ชื่อตัวแปรและความหมายของข้อมูลการฟื้นฟูสภาพกรณีที่ 1	46
ตารางที่ 10 ชื่อตัวแปรและความหมายของข้อมูลการฟื้นฟูสภาพกรณีที่ 2	48
ตารางที่ 11 ชื่อตัวแปรและความหมายของข้อมูลการฟื้นฟูสภาพกรณีที่ 3	50
ตารางที่ 12 ค่าพารามิเตอร์ที่เหมาะสม สำหรับอัลกอริทึม DT	53
ตารางที่ 13 Confusion Matrix โดยใช้อัลกอริทึม DT ของการฟื้นฟูสภาพกรณีที่ 1	54
ตารางที่ 14 Confusion Matrix โดยใช้อัลกอริทึม DT ของการฟื้นฟูสภาพกรณีที่ 2	54
ตารางที่ 15 Confusion Matrix โดยใช้อัลกอริทึม DT ของการฟื้นฟูสภาพกรณีที่ 3	54
ตารางที่ 16 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม DT	55
ตารางที่ 17 ค่าพารามิเตอร์ที่เหมาะสม สำหรับอัลกอริทึม SVM	56
ตารางที่ 18 Confusion Matrix โดยใช้อัลกอริทึม SVM ของการฟื้นฟูสภาพกรณีที่ 1	56
ตารางที่ 19 Confusion Matrix โดยใช้อัลกอริทึม SVM ของการฟื้นฟูสภาพกรณีที่ 2	56
ตารางที่ 20 Confusion Matrix โดยใช้อัลกอริทึม SVM ของการฟื้นฟูสภาพกรณีที่ 3	56
ตารางที่ 21 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม SVM.....	57
ตารางที่ 22 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้	58

สารบัญตาราง (ต่อ)

ตารางที่ 23 Confusion Matrix ด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ พ่นสภาพกรณ์ที่ 1.....	58
ตารางที่ 24 Confusion Matrix ด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ พ่นสภาพกรณ์ที่ 2.....	59
ตารางที่ 25 Confusion Matrix ด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ พ่นสภาพกรณ์ที่ 3.....	59
ตารางที่ 26 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธี Bagging โดยใช้ DT เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	60
ตารางที่ 27 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการ เรียนรู้.....	61
ตารางที่ 28 Confusion Matrix ด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของ การพ่นสภาพกรณ์ที่ 1.....	61
ตารางที่ 29 Confusion Matrix ด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของ การพ่นสภาพกรณ์ที่ 2.....	61
ตารางที่ 30 Confusion Matrix ด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของ การพ่นสภาพกรณ์ที่ 3.....	62
ตารางที่ 31 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธีการ Bagging โดยใช้ SVM เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	62
ตารางที่ 32 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการ เรียนรู้.....	63
ตารางที่ 33 Confusion Matrix วิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ พ่นสภาพกรณ์ที่ 1.....	64
ตารางที่ 34 Confusion Matrix วิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ พ่นสภาพกรณ์ที่ 2.....	64
ตารางที่ 35 Confusion Matrix วิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ พ่นสภาพกรณ์ที่ 3.....	64

สารบัญตาราง (ต่อ)

ตารางที่ 36 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธีการ Boosting โดยใช้ DT เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	65
ตารางที่ 37 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการ เรียนรู้	66
ตารางที่ 38 Confusion Matrix ด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการผันสภาพกรณีที่ 1	66
ตารางที่ 39 Confusion Matrix ด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการผันสภาพกรณีที่ 2	66
ตารางที่ 40 Confusion Matrix ด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการผันสภาพกรณีที่ 3	67
ตารางที่ 41 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธีการ Boosting โดยใช้ SVM เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้	67
ตารางที่ 42 ค่าพารามิเตอร์ที่เหมาะสมโดยใช้อัลกอริทึม Random Forest.....	68
ตารางที่ 43 Confusion Matrix ด้วยวิธี Random Forest ของการผันสภาพกรณีที่ 1	68
ตารางที่ 44 Confusion Matrix ด้วยวิธี Random Forest ของการผันสภาพกรณีที่ 2	69
ตารางที่ 45 Confusion Matrix ด้วยวิธี Random Forest ของการผันสภาพกรณีที่ 3	69
ตารางที่ 46 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม Random Forest	69
ตารางที่ 47 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทการผันสภาพของข้อมูลการผันสภาพกรณี ที่ 1	71
ตารางที่ 48 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทการผันสภาพของข้อมูลการผันสภาพกรณี ที่ 2	72
ตารางที่ 49 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทการผันสภาพของข้อมูลการผันสภาพกรณี ที่ 3	74
ตารางที่ 50 รายละเอียดค่าใช้จ่ายในการดำเนินงาน	85
ตารางที่ 51 การดำเนินงานโครงการวิจัย	86

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

สถาบันการศึกษามีบทบาทที่สำคัญในการพัฒนาประเทศ การศึกษาเป็นรากฐานที่สำคัญในการสร้างบุคคลให้มีความรู้ความสามารถในการปฏิบัติหน้าที่ และสามารถดำรงชีวิตอยู่ในสังคมอย่างสันติสุข (ชนิดาภา บุญประสม และ จรรย์ แสนราช, 2561) การที่นักศึกษาสามารถเล่าเรียนจนจบหลักสูตรได้นั้น จำเป็นจะต้องมีความรู้ความเข้าใจในวิชาชีพของตน นอกเหนือจากนี้ผู้ที่เกี่ยวข้องกับวงการศึกษาจำเป็นจะต้อง วางแผน ติดตาม และควบคุม ในแต่ละกระบวนการเพื่อส่งเสริมการพัฒนากระบวนการเรียนการสอนให้มีประสิทธิภาพ ตลอดจนช่วยกันหาแนวทางในการป้องกัน และการแก้ปัญหาการฟื้นฟูสภาพของนักศึกษาในระหว่างเรียน

ในสถาบันการศึกษามีปัญหาการฟื้นฟูสภาพของนักศึกษาเป็นสิ่งที่ต้องหาแนวทางในการแก้ไขปัญหานี้ เนื่องจากปัจจุบันมีอัตราการฟื้นฟูสภาพของนักศึกษาในระดับที่เป็นปัญหา จากข้อมูลของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ระหว่างปีการศึกษา 2558 - 2563 มีการฟื้นฟูสภาพของนักศึกษามากถึง 205 คน คิดเป็นร้อยละ 30.19 จากจำนวนทั้งหมด 679 คน และพบว่าอัตราการฟื้นฟูสภาพของนักศึกษาเฉลี่ยแต่ละปีอยู่ที่ ร้อยละ 29.24 (สำนักงานทะเบียนมหาวิทยาลัยขอนแก่น, 2564) ซึ่งมาจากหลายปัจจัย เช่น ปัญหาครอบครัว ปัญหาเศรษฐกิจ ปัญหาการเรียน และปัจจัยที่มาจากตัวนักศึกษาเองเป็นต้น (นนทวัฒน์ ทวีชาติ, อรยา เฟื่องประจัญ, วิไลรัตน์ ยาทองไชย, และชูศักดิ์ ยาทองไชย, 2563) ปัจจัยเหล่านี้จะส่งผลกระทบต่อคณะและมหาวิทยาลัย ทำให้เสียเวลาในการบริหารจัดการ และเสียทรัพยากรในการลงทุน ส่วนนักศึกษาเสียเวลา และเสียค่าใช้จ่าย ดังนั้นทางคณะและมหาวิทยาลัยควรส่งเสริม และพัฒนากระบวนการเรียนการสอนให้มีประสิทธิภาพ ตลอดจนช่วยกันหาแนวทางในการป้องกันและแก้ไขปัญหาการฟื้นฟูสภาพของนักศึกษา หากนักศึกษาฟื้นฟูสภาพก่อนจะจบการศึกษาถือว่าเป็นความสูญเสียทางการศึกษา จะส่งผลกระทบต่อด้านเศรษฐกิจของประเทศ และเศรษฐกิจของครอบครัวซึ่งต้องสิ้นเปลืองค่าใช้จ่ายไปโดยไม่ได้รับประโยชน์ที่คุ้มค่า

การทำเหมืองข้อมูล (Data Mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหา รูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ดังนั้นการทำเหมืองข้อมูลเป็นการนำเอาข้อมูลมาวิเคราะห์เพื่อให้ได้ความรู้ใหม่ออกมาเพื่อนำไปใช้ประโยชน์ในการตัดสินใจ (นิสานันท์ พลอาสา, 2558) ในช่วงหลายปีที่ผ่านมา มีการนำเทคนิคการทำเหมืองข้อมูลมาใช้ในการศึกษาปัจจัย และทำนายการฟื้นฟูสภาพของนักศึกษาอย่าง

แพร่หลาย เช่น การทำนายการลาออกของนักศึกษา (Tenpipat & Akkarajitsakul, 2020) และ การใช้แบบจำลองต้นไม้ตัดสินใจแบบรวมกลุ่มเพื่อทำนายการลาออกของนักศึกษา (Naseem, Chaudhary, Sharma, & Lal, 2020) เป็นต้น

งานวิจัยนี้มีแนวคิดที่จะพัฒนาประสิทธิภาพในการจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้การเรียนรู้แบบรวมกลุ่ม เพื่อแก้ไขปัญหาในด้านความถูกต้องและแม่นยำสำหรับการจำแนกการฟื้นฟูสภาพของนักศึกษา โดยได้นำเอาหลักทฤษฎีการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) มาใช้เพื่อช่วยเพิ่มประสิทธิภาพในความถูกต้องในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาให้สูงขึ้น และใช้อัลกอริทึมในการตรวจจับการฟื้นฟูสภาพของนักศึกษา โดยมีนำข้อมูลก่อนที่เหตุการณ์การฟื้นฟูสภาพของนักศึกษาจะเกิดขึ้นมาทำนายการฟื้นฟูสภาพของนักศึกษา เพื่อที่จะแก้ไขปัญหาและวางกลยุทธ์ในการจัดการการฟื้นฟูสภาพของนักศึกษาได้อย่างทันท่วงที

1.2 วัตถุประสงค์ของการวิจัย

- 1) สร้างแบบจำลองการจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้วิธีการเรียนรู้แบบเดี่ยวและแบบรวมกลุ่ม
- 2) เปรียบเทียบประสิทธิภาพการจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้วิธีการเรียนรู้แบบเดี่ยวและแบบรวมกลุ่ม

1.3 สมมติฐานของการวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ทำการตั้งสมมติฐานไว้ คือ การนำเทคนิควิธีการเรียนรู้แบบเดี่ยวและแบบรวมกลุ่มมาใช้สำหรับการจำแนกประเภทในการฟื้นฟูสภาพของนักศึกษา เพื่อที่จะได้แบบจำลองที่มีประสิทธิภาพความถูกต้องและความแม่นยำที่สุด

1.4 ขอบเขตของการวิจัย

- 1) การวิจัยนี้ใช้ข้อมูลของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้ข้อมูลเฉพาะนักศึกษาปี 1 ตั้งแต่ปีการศึกษา 2558 – 2564 จำนวน 719 คน
- 2) ใช้อัลกอริทึม Decision Tree และ Support Vector Machine สำหรับขั้นตอนการจำแนกประเภทแบบเดี่ยว รวมถึงใช้เป็นอัลกอริทึมพื้นฐานการเรียนรู้ร่วมกับวิธีการเรียนรู้แบบรวมกลุ่มแบบ Bagging และ Boosting
- 3) ใช้วิธีการเรียนรู้แบบรวมกลุ่ม 3 วิธี ได้แก่ Bagging, Boosting, และ Random Forest

1.5 ความหมายหรือนิยามคำศัพท์เฉพาะ

- 1) การฟื้นสภาพ คือการที่นักศึกษาที่มีสถานะฟื้นสภาพ โดยแบ่งออกเป็น ฟื้นสภาพเนื่องจากลาออก ฟื้นสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา ฟื้นสภาพเนื่องจากตกออก และฟื้นสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด

1.6 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

- 1) ทราบปัจจัยที่มีผลต่อการฟื้นสภาพนักศึกษาของนักศึกษาระดับปริญญาตรีสาขาวิชาสถิติ คณะวิทยาศาสตร์
- 2) ได้แบบจำลองการจำแนกประเภทที่มีประสิทธิภาพความถูกต้องและแม่นยำในการจำแนกประเภทสำหรับการฟื้นสภาพของนักศึกษา
- 3) นำข้อมูลสารสนเทศที่ได้จากการจำแนกประเภทสำหรับการฟื้นสภาพของนักศึกษาไปประยุกต์ใช้ในการปรับปรุงหลักสูตร วางแผน และพัฒนานักศึกษา
- 4) ใช้แบบจำลองในการลดเวลาปฏิบัติงานของเจ้าหน้าที่ที่เกี่ยวข้องในการประเมิน และวางแผนในการฟื้นสภาพของนักศึกษา

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบการจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษา โดยใช้วิธีการเรียนรู้แบบรวมกลุ่ม ทางผู้วิจัยได้ทำการศึกษาทฤษฎีที่เกี่ยวข้อง ศึกษางานวิจัย บทความ และเอกสารทางวิชาการต่างๆ เพื่อเป็นแนวทางและกรอบในการศึกษา โดยสรุปเป็นหัวข้อต่างๆดังต่อไปนี้

- 2.1 ทฤษฎีการฟื้นฟูสภาพ
- 2.2 การทำเหมืองข้อมูล (Data Mining)
- 2.3 เทคนิคการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกประเภท (Imbalanced Data Problem Solving in Classification Technique)
 - 2.3.1 วิธีสุ่มเกิน (Over Sampling)
 - 2.3.2 วิธีสุ่มลด (Under Sampling)
 - 2.3.3 วิธีผสมผสาน (Hybrid Methods)
 - 2.3.4 วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE)
- 2.4 เทคนิคการจำแนกประเภทข้อมูล (Data Classification Techniques)
 - 2.4.1 ต้นไม้ตัดสินใจ (Decision Tree)
 - 2.4.2 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
- 2.5 การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning)
 - 2.5.1 Bagging Method
 - 2.5.2 Boosting Method
 - 2.5.3 แรนดอมฟอเรส (Random Forest)
- 2.6 มาตรฐานและการทดสอบประสิทธิภาพแบบจำลอง
- 2.7 งานวิจัยที่เกี่ยวข้อง
- 2.8 กรอบแนวคิดงานวิจัย

2.1 ทฤษฎีการพ้นสภาพ

- 2.1.1 การพ้นสภาพการพ้นสภาพหมายถึง การที่นักศึกษาพ้นจากสภาพนักศึกษา โดยมหาวิทยาลัยให้ออกจากสถาบันก่อนเรียนครบตามหลักสูตรที่กำหนดไว้ 2 กรณี ได้แก่ กรณีที่ 1 พ้นจากสภาพนักศึกษาเนื่องจากคะแนนเฉลี่ยสะสมไม่ถึงเกณฑ์ที่มหาวิทยาลัยกำหนด กรณีที่ 2 พ้นจากสภาพนักศึกษาเนื่องจากขาดการติดต่อกับมหาวิทยาลัย (สภามหาวิทยาลัย, 2562)
- 2.1.2 ระเบียบมหาวิทยาลัยขอนแก่น ว่าด้วยการศึกษาชั้นปริญญาตรี พ.ศ. 2562จากระเบียบมหาวิทยาลัยขอนแก่น ว่าด้วยการศึกษาชั้นปริญญาตรี พ.ศ. 2562การวัดและประเมินผล และการพ้นสภาพนักศึกษาในมหาวิทยาลัยขอนแก่น (สำนักทะเบียนมหาวิทยาลัยขอนแก่น, 2562) มีดังนี้
- 2.1.2.1 การวัดและประเมินผล

มหาวิทยาลัยขอนแก่นจัดให้มีการวัดผลแต่ละรายวิชาที่นักศึกษาลงทะเบียน ซึ่งอาจารย์ผู้สอนต้องแจ้งเกณฑ์และเงื่อนไขในการประเมินผลในแต่ละวิชาให้นักศึกษาทราบล่วงหน้า และการประเมินผลในแต่ละรายวิชาจะใช้ระดับคะแนนตัวอักษร ดังนี้

ระดับคะแนนตัวอักษร	ความหมาย	ค่าคะแนนต่อหน่วยกิต
A	ผลการประเมินขั้นดีเยี่ยม (Excellent)	4.0
B+	ผลการประเมินขั้นดีมาก (Very Good)	3.5
B	ผลการประเมินขั้นดี (Good)	3.0
C+	ผลการประเมินขั้นค่อนข้างดี (Fairly Good)	2.5
C	ผลการประเมินพอใช้ (Fair)	2.0
D+	ผลการประเมินขั้นอ่อน (Poor)	1.5
D	ผลการประเมินขั้นอ่อนมาก (Very Poor)	1.0
F	ผลการประเมินขั้นตก (Fail)	0

และตัวอักษรที่มีความหมายเฉพาะซึ่งแสดงถึงสถานภาพนักศึกษา คือ I P R S T U และ W ตัวอักษรเหล่านี้ไม่มีค่าคะแนนยกเว้น ตัวอักษร T

ตัวอักษร	ความหมาย
I	ยังไม่สมบูรณ์ (Incomplete)
P	กำลังดำเนินการอยู่ (In Progress)
R	ซ้ำชั้น (Repeat)
S	พอใจ (Satisfactory)
T	รับโอน (Transfer)
U	ไม่พอใจ (Unsatisfactory)
W	การถอนรายวิชา (Withdrawal)

2.1.2.2 การพ้นสภาพนักศึกษา

จากระเบียบมหาวิทยาลัยขอนแก่น ว่าด้วยการศึกษาชั้นปริญญาตรี พ.ศ. 2562 นักศึกษาจะพ้นสภาพนักศึกษาก็ต่อเมื่อ

- 1) ตาย
- 2) ลาออก
- 3) ตกออก

นักศึกษาจะถูกพิจารณาให้ตกออกในกรณีดังต่อไปนี้

- ก. ระดับคะแนนเฉลี่ยสะสมไม่ถึง 1.50 เมื่อได้ลงทะเบียนเรียนมาแล้ว และมีหน่วยกิตสะสมตั้งแต่ 30-59 หน่วยกิต
 - ข. ระดับคะแนนเฉลี่ยสะสมไม่ถึง 1.75 เมื่อได้ลงทะเบียนเรียนมาแล้ว และมีหน่วยกิตสะสมตั้งแต่ 60 หน่วยกิตขึ้นไป
 - ค. สำหรับนักศึกษาหลักสูตรแพทยศาสตรบัณฑิต ให้เป็นไปตามหลักเกณฑ์ที่มหาวิทยาลัยกำหนด
- 4) ถูกสั่งให้พ้นสภาพตามระเบียบข้อบังคับของมหาวิทยาลัย
 - 5) ขาดคุณสมบัติการเข้าเป็นนักศึกษาของมหาวิทยาลัย ตามระเบียบฯ
 - 6) เรียนสำเร็จตามหลักสูตร
 - 7) ไม่ลงทะเบียนให้เสร็จสิ้นภายในเวลาที่มหาวิทยาลัยกำหนดในแต่ละภาคการศึกษา
 - 8) ไม่ชำระค่าธรรมเนียมเพื่อขึ้นโดยต่อทะเบียนภายในระยะเวลาที่มหาวิทยาลัยกำหนด

- 9) ศึกษาในมหาวิทยาลัยเกินจำนวนสองเท่าของระยะเวลาการศึกษาที่กำหนดไว้ในหลักสูตร
- 10) ต้องโทษโดยคำพิพากษาถึงที่สุดให้จำคุกเว้นแต่ความผิดโทษ หรือความผิดที่ได้กระทำโดยประมาท
- 11) โอนไปเป็นนิสิตนักศึกษาของสถาบันอุดมศึกษาอื่น
- 12) เหตุอื่นตามที่มหาวิทยาลัยกำหนด

2.2 การทำเหมืองข้อมูล (Data Mining)

2.2.1 ทฤษฎีการทำเหมืองข้อมูล

ทฤษฎีการทำเหมืองข้อมูล (Data Mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบ และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ดังนั้นการทำเหมืองข้อมูล เป็นการนำเอาข้อมูลมาวิเคราะห์เพื่อให้ได้ความรู้ใหม่ออกมาเพื่อนำไปใช้ประโยชน์ในการตัดสินใจ (นิสานันท์ พลอาสา, 2558) ในปัจจุบันมีความก้าวหน้าทางเทคโนโลยีทำให้มีการจัดเก็บข้อมูลเป็นจำนวนมาก เพื่อช่วยวิเคราะห์ปัญหา ตัดสินใจ และดำเนินงานในหน่วยงานในปัจจุบัน การวิเคราะห์ข้อมูลเพื่อให้ความรู้ที่เกี่ยวข้องกับข้อมูลได้แก่ ตัวแบบที่แสดงความสัมพันธ์ระหว่างข้อมูล ความรู้เหล่านี้สามารถนำมาใช้ประโยชน์ในการดำเนินงาน และการตัดสินใจภายในองค์กร การทำเหมืองข้อมูลที่สำคัญได้แก่ การคัดเลือกข้อมูล (Selection) การเตรียมข้อมูล (Preprocessing) การแปลงข้อมูล (Transformation) การวิเคราะห์และค้นหารูปแบบข้อมูล (Data Mining) และการแปล/ประเมินผลการวิเคราะห์ข้อมูล (Interpretation/Evaluation)

2.2.2 ประเภทการทำเหมืองข้อมูล

การทำเหมืองข้อมูลแบบทำนาย (Predictive Mining) คือการนำข้อมูลที่มีอยู่มาใช้ในการทำนายผลข้อมูลในอนาคตที่ไม่ทราบมาก่อน ซึ่งการสร้างแบบจำลองรูปแบบนี้จะเน้นการแบ่งข้อมูลออกเป็นกลุ่มตามคุณสมบัติของข้อมูล ในกรณีที่ข้อมูลไม่ต่อเนื่องจะใช้เทคนิคการจำแนกประเภทข้อมูล (Classification) และในกรณีที่ข้อมูลมีความต่อเนื่องจะใช้เทคนิคการถดถอย (Regression)

การทำเหมืองข้อมูลแบบพรรณนา (Descriptive Mining) คือการนำข้อมูลที่มีอยู่มาศึกษาหา คำอธิบายคุณลักษณะทั่วไปของข้อมูล เพื่อใช้เป็นแนวทางในการตัดสินใจ เช่น เทคนิคการหา ความสัมพันธ์ (Association) หรือเทคนิคการจัดกลุ่ม (Clustering)

2.2.3 การทำเหมืองข้อมูล

Cluster Analytic คือ การจัดกลุ่มข้อมูลซึ่งมีลักษณะคล้ายกับการแบ่งประเภท (Classification) แต่จะไม่เหมือนกันโดยการแบ่งประเภทจะวิเคราะห์ข้อมูลที่กำหนดผลลัพธ์ แต่สำหรับการแบ่งกลุ่มเป็น การวิเคราะห์โดยไม่พิจารณาข้อมูลที่กำหนดผลลัพธ์ แต่จะใช้ขั้นตอนวิธีการจัดกลุ่มเพื่อค้นหากลุ่มที่ สามารถยอมรับได้เพื่อจัดเข้ากลุ่ม กล่าวคือ กลุ่มของวัตถุมีการสร้างขึ้นโดยเปรียบเทียบวัตถุที่มีความ เหมือนกันจัดเข้ากลุ่มเดียวกัน

Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูล โดยค้นหาความสัมพันธ์ของข้อมูล ทั้งสองชุดหรือมากกว่าสองชุดขึ้นไปไว้ด้วยกัน ความสำคัญของกฎทำการวัดโดยใช้ข้อมูลสองตัวด้วยกัน คือค่าสนับสนุน (Support) ซึ่งเป็นเปอร์เซ็นต์ของการดำเนินการที่กฎสามารถนำไปใช้ หรือเป็น เปอร์เซ็นต์ของการดำเนินการที่กฎที่ใช้มีความถูกต้อง และข้อมูลตัวที่สองที่นำมาใช้วัดคือค่าความมั่นใจ (Confidence) ซึ่งเป็นจำนวนของกรณีที่ถูกถูกต้องโดยสัมพันธ์กับจำนวนของกรณีที่สามารถนำไปใช้ ได้ ในการหาความสัมพันธ์นั้นจะมีขั้นตอนวิธีการหาหลายวิธีด้วยกัน แต่ขั้นตอนวิธีที่เป็นที่รู้จักและ ใช้อย่างแพร่หลายคือ อัลกอริทึม Apriori

Classification Analytic เป็นการจัดประเภทของ วัตถุประสงค์เพื่อให้สามารถใช้เป็นตัวแทน ทำนายประเภท ซึ่งตัวแทนสร้างจากการวิเคราะห์ชุดของข้อมูลฝึกสอน (Training Data) โดยใช้ข้อมูลที่ ระบุผลลัพธ์เรียบร้อยแล้ว รูปแบบของตัวแทนแสดงได้หลายแบบเช่น Classification Rules, Decision Trees หรือ Neural Networks เป็นต้น

2.3 เทคนิคการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกประเภท (Imbalanced Data Problem

Solving in Classification Technique)

ข้อมูลไม่สมดุลของกลุ่มตัวแปรผลลัพธ์ที่นำมาศึกษามีผลต่อความถูกต้องของสมการการทำนาย ซึ่งเป็น ปัญหาหลักที่นักวิจัยให้ความสนใจในปัญหาความไม่สมดุลของข้อมูลที่พบได้บ่อยครั้งในข้อมูลจริง เมื่อนำ ข้อมูลเหล่านี้มาใช้งานทางด้านการเรียนรู้ของเครื่อง (Machine Learning) และการทำเหมืองข้อมูล (Data Mining) จะส่งผลกระทบต่อการเรียนรู้ของอัลกอริทึมในการจำแนกข้อมูลด้วยวิธีการจำแนกข้อมูลแบบปกติที่ ให้ความสำคัญกับข้อมูลกลุ่มผลลัพธ์เท่ากันจะทำให้ประสิทธิภาพในการจำแนกประเภทข้อมูลส่วนน้อยมี ความถูกต้องน้อยลง (วิชญ์วิสิฐ เกษรสิทธิ์, วิจิต หล่อจิระชุนห์กุล, และจิราวัลย์ จิตรถเวช, 2561)

ซึ่งการแก้ปัญหาข้อมูลไม่สมดุลมีใช้เทคนิควิธีดังนี้

2.3.1 วิธีสุ่มเกิน (Over Sampling)

วิธีการสุ่มเกินเป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก ซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มเลือกจากข้อมูลเดิมในกลุ่มส่วนน้อยโดยใช้วิธีการสุ่มแบบเป็นระบบ

2.3.2 วิธีสุ่มลด (Under Sampling)

วิธีสุ่มลดเป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยโดยใช้วิธีการสุ่มแบบเป็นระบบ

2.3.3 วิธีผสมผสาน (Hybrid Methods)

วิธีผสมผสานเป็นวิธีการที่นำเทคนิควิธีสุ่มเกินและวิธีสุ่มลดมาทำงานร่วมกัน โดยพยายามหาค่ากลางในการชักตัวอย่างให้ได้ตามจำนวนที่อยู่ตรงกลางระหว่างข้อมูลในกลุ่มส่วนมากกับข้อมูลในกลุ่มส่วนน้อย

2.3.4 วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE)

วิธีสังเคราะห์ข้อมูลเพิ่มเป็นเทคนิคการสุ่มตัวอย่างแบบพิเศษของการสุ่มเพิ่ม แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่ จากข้อมูลเดิมที่มีอยู่โดยใช้ อัลกอริทึมเพื่อนบ้านที่อยู่ใกล้ที่สุด (K-Nearest Neighbor) ในการขยายขอบเขตการตัดสินใจของตัวแบบ ซึ่งในขั้นตอนการสังเคราะห์ข้อมูลมีขั้นตอนดังนี้คือระบุเพื่อนบ้านที่ใกล้เคียงที่สุด k ค่าของข้อมูลเดิม

สำหรับข้อมูลเดิม M หา $k=1$ ที่มีระยะใกล้เคียงกับข้อมูลเดิมโดย l คือจำนวนเพื่อนบ้านใกล้เคียงกับจุด M และสุ่มเลือกจุดระหว่างสองจุดและสร้างกรณีใหม่

ตัวอย่างเช่นสร้างจุด $m_1(c_1, c_2, \dots, c_n)$ ระหว่าง $M(a_1, a_2, \dots, a_n)$ และ $M_1(b_1, b_2, \dots, b_n)$

เมื่อ

$$c_1 = a_1 + (b_1 - a_1) \times \text{rand}('UNIFORM')$$

$$c_2 = a_2 + (b_2 - a_2) \times \text{rand}('UNIFORM')$$

.

$$c_n = a_n + (b_n - a_n) \times \text{rand}('UNIFORM')$$

โดยที่ m_1 คือจุดที่สังเคราะห์ขึ้นมาใหม่ระหว่าง $M(a_1, a_2, \dots, a_n)$ และ $M_1(b_1, b_2, \dots, b_n)$

a_1, a_2, \dots, a_n คือข้อมูลในค่าสังเกตที่จุด M และ b_1, b_2, \dots, b_n คือข้อมูลในค่าสังเกตที่จุด M_1

2.4 เทคนิคการจำแนกประเภทข้อมูล (Data Classification Techniques)

2.4.1 ต้นไม้ตัดสินใจ (Decision Tree: DT)

Decision Tree (DT) เป็นหนึ่งในอัลกอริทึมการเรียนรู้ของเครื่องที่เร็วและโดดเด่น ต้นไม้การตัดสินใจจำลองตรรกะการตัดสินใจ เช่น ทดสอบความสอดคล้องผลลัพธ์สำหรับการจัดประเภทรายการข้อมูลให้เป็นโครงสร้างแบบต้นไม้ โหนดของต้นไม้โดยปกติมีหลายระดับโดยที่โหนดแรกหรือบนสุดเรียกว่ารูตโหนด (Root Node) โหนดภายใน (Internal Node) ทั้งหมด (เช่น โหนดที่มีโหนดย่อยอย่างน้อยหนึ่งรายการ) เป็นโหนดที่แสดงถึงคุณลักษณะ (Feature) ที่ใช้ในการแบ่งกลุ่มข้อมูล โดยมีรูตโหนด (Root Node) อยู่บนสุดเป็นโครงสร้าง อัลกอริทึมการจำแนกประเภทต้นไม้ตัดสินใจจะแยกตัวไปยังโหนดย่อยที่เหมาะสมขึ้นอยู่กับผลการทดสอบ โดยที่กระบวนการทดสอบและการแตกแขนงจะทำซ้ำๆ จนกว่าจะถึงใบโหนด (Leaf Node) การตัดสินใจต้นไม้ถูกพบว่ายากต่อการตีความ และเรียนรู้ได้รวดเร็ว เมื่อสำรวจกฎที่สร้างมาจากต้นไม้ตัดสินใจจะพบว่าเส้นทางจะให้ข้อมูลที่เพียงพอต่อการคาดเดาเกี่ยวกับผลลัพธ์ต่างๆ (Uddin, Khan, & Moni, 2019)

จากภาพที่ 1 แสดงตัวอย่างต้นไม้ตัดสินใจ (Decision Tree) ที่มีตัวแปรแต่ละตัว (C1, C2 และ C3) จะถูกแทนด้วยวงกลม และผลการตัดสินใจ (Class A และ Class B) จะแสดงด้วยสี่เหลี่ยม เพื่อที่จะจำแนกตัวอย่างไปยังผลลัพธ์ได้สำเร็จ แต่ละเส้นทางจะมีป้ายกำกับว่า ‘จริง’ หรือ ‘เท็จ’ ตามค่าผลลัพธ์จากการทดสอบ สูตรที่ใช้ในการคำนวณค่า Information Gain ต้องเริ่มจากการหาค่า Entropy ดังนี้

Entropy:

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

โดย S คือ Attribute ที่นำมาวัดค่า Entropy

p_i คือ สัดส่วนของจำนวนสมาชิกของกลุ่ม i กับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

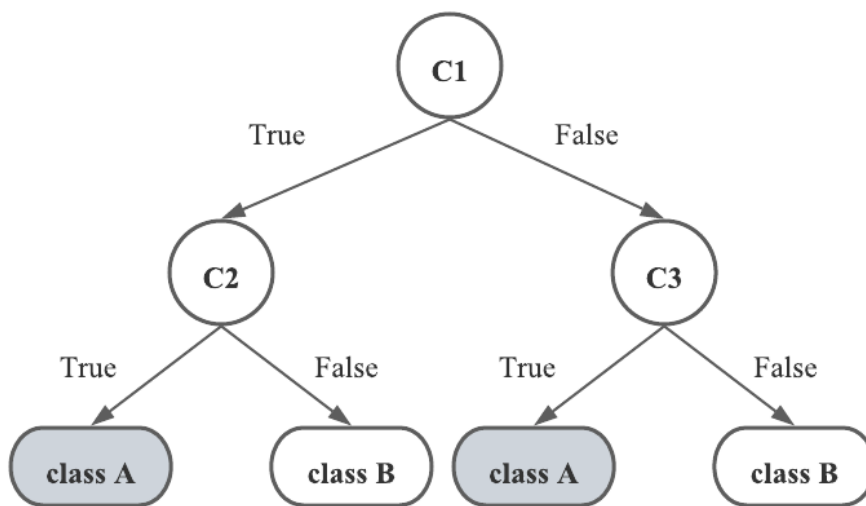
Information Gain:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

โดย A คือ Attribute A

$|S_v|$ คือ สมาชิกของ Attribute A ที่มีค่า v

$|S|$ คือ จำนวนสมาชิกของกลุ่มตัวอย่าง



ภาพที่ 1 Decision Tree

2.4.2 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

Support Vector Machine (SVM) เป็นเทคนิคการจำแนกประเภทของข้อมูลที่มีพื้นฐานมาจากทฤษฎีการเรียนรู้ทางสถิติ และสามารถจำแนกได้ทั้งข้อมูลเชิงเส้นและไม่เชิงเส้น ซึ่งคล้ายเทคนิคโครงข่ายประสาทเทียม โดย SVM ใช้หลักการลดค่าความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด (Structural Risk Minimization) เพื่อลดค่าความผิดพลาดของการทำนาย (Minimized Error) พร้อมกับเพิ่มระยะการแบ่งแยกให้มากที่สุด (Maximized Margin) ระยะขอบ (Margin) คือระยะห่างระหว่างการตัดสินใจไฮเปอร์เพลน (Decision Hyperplane) และตัวอย่างที่ใกล้ที่สุดจะเป็นสมาชิกของผลลัพธ์นั้น

หลักการของ SVM คือการหาสมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกประเภทข้อมูลทำโดยการเลือกเส้นหรือระนาบเพื่อแบ่งแยกประเภทข้อมูลที่เหมาะสมที่สุด (ปัทมญา บุญรักษา และ จาริ ทองคำ, 2560)

จากภาพที่ 2 Support Vector Machine คือวิธีการทำงาน SVM ได้ระบุไฮเปอร์เพลน (Hyperplane) เป็นเส้นซึ่งเป็นการแยกระหว่างผลลัพธ์ ‘ดาว’ และ ‘วงกลม’ ให้เหมาะสมที่สุด

กำหนดให้ $(x_i, y_i), \dots, (x_n, y_n)$ เมื่อ $x \in R^m, y \in \{-1, 1\}$ เป็นตัวอย่างที่ใช้สำหรับการสอน

โดย n คือ จำนวนข้อมูลตัวอย่าง

m คือ จำนวนมิติข้อมูล

x คือ ข้อมูลนำเข้า

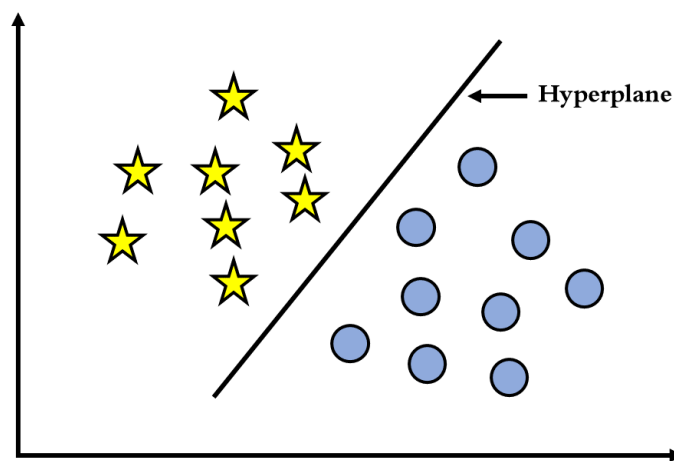
y คือ ประเภทข้อมูล ซึ่งประกอบไปด้วย 2 กลุ่ม มีค่า +1 หรือ -1

(+1 = “ข้อมูลบวก” และ -1 = “ข้อมูลลบ”)

การสร้างเส้นระนาบตัดสินใจเพื่อแบ่งแยกกลุ่มผลลัพธ์ของข้อมูลสามารถคำนวณได้ดังนี้

$$(W * x_1) + b > 0 \text{ ถ้า } y_i = +1 \text{ และ } (W * x_2) + b < 0 \text{ ถ้า } y_i = -1$$

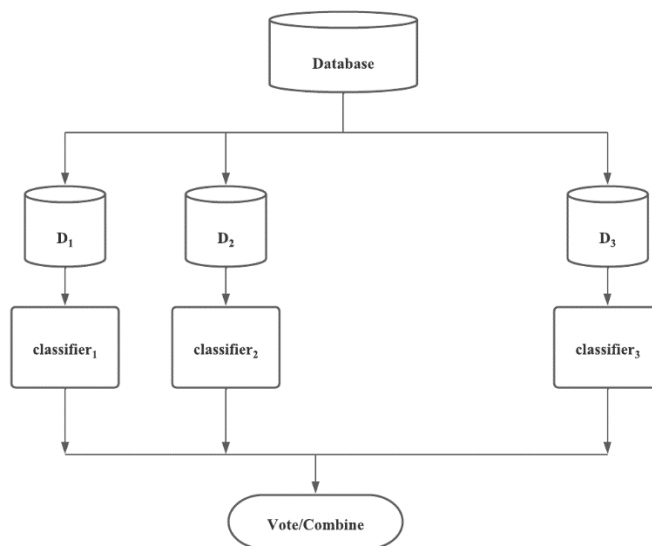
โดย w คือ เวกเตอร์น้ำหนัก
 x_1 คือ เวกเตอร์ข้อมูลที่มีค่าเป็นบวก
 x_2 คือ เวกเตอร์ข้อมูลที่มีค่าเป็นลบ
 b คือ ค่าความคลาดเคลื่อน (Bias)



ภาพที่ 2 Support Vector Machine

2.5 การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning)

การเรียนรู้แบบรวมกลุ่มเป็นวิธีการรวมเอากลุ่มของตัวจำแนกประเภทข้อมูลที่สร้างขึ้นหลายๆ ตัวจำแนก และมีความเป็นอิสระต่อกันมาพิจารณาร่วมกัน เพื่อช่วยในการตัดสินใจสำหรับการหาคำตอบโดยใช้วิธีการรวม (Combine) หรือ วิธีการโหวต (Voting) เพื่อให้ได้ผลลัพธ์ในการจำแนกข้อมูลที่มีประสิทธิภาพสูงซึ่งเทคนิคการเรียนรู้แบบรวมกลุ่มมีอยู่หลากหลายวิธีแต่สำหรับวิธีการที่มีประสิทธิภาพได้รับความนิยมได้แก่ วิธี Bagging และ Boosting (ปัทม์ อุปการ, 2560)

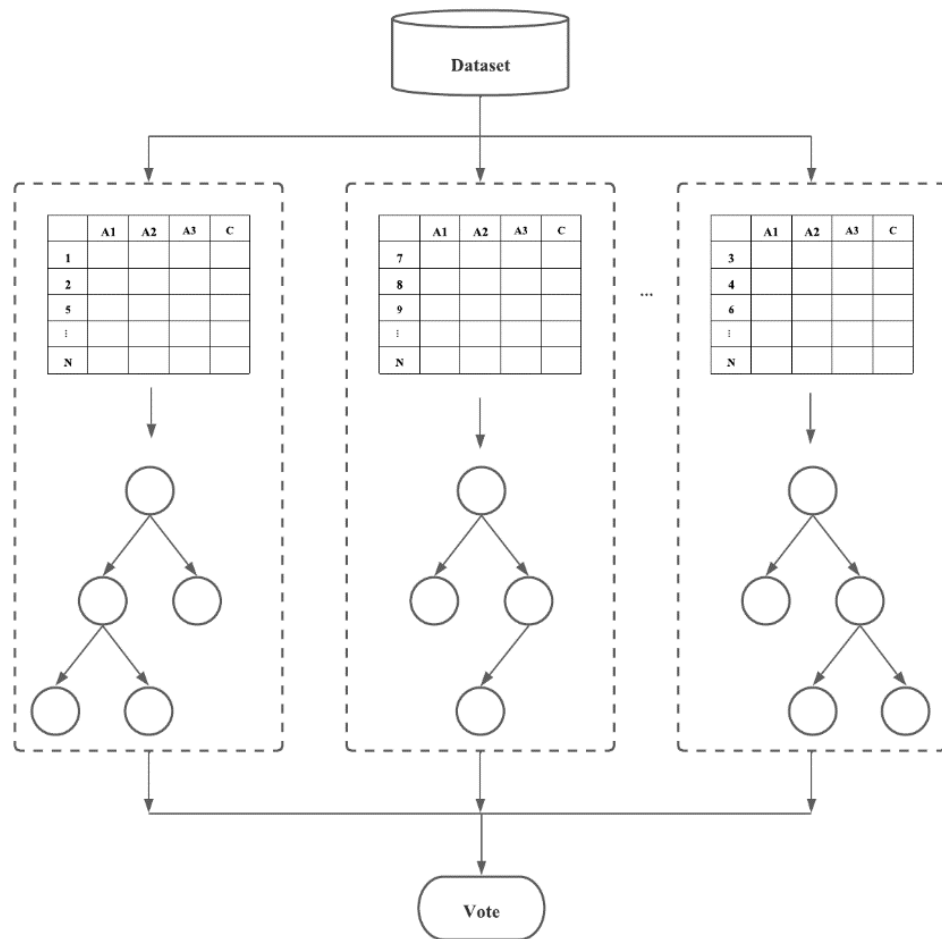


ภาพที่ 3 แสดงโครงสร้างพื้นฐานของการเรียนรู้แบบรวมกลุ่ม

จากภาพที่ 3 แสดงโครงสร้างพื้นฐานของแบบจำลองการเรียนรู้แบบรวมกลุ่ม ซึ่งจะประกอบด้วย 3 ขั้นตอนหลัก ได้แก่ ขั้นตอนที่ 1 การสร้างชุดข้อมูลตัวอย่างขึ้นมาหลายๆ ชุดที่มีลักษณะแตกต่างกันในแต่ละชุด ขั้นตอนที่ 2 สร้างแบบการจำแนกประเภทข้อมูลหลายๆ ตัวจำแนก เพื่อเรียนรู้ชุดข้อมูลที่สร้างขึ้นในแต่ละชุด และขั้นตอนสุดท้ายเป็นการรวบรวมตัวจำแนกประเภทหลายๆ ตัวจำแนกที่สร้างขึ้นจากขั้นตอนที่ 2 เพื่อร่วมกันตัดสินใจในการพิจารณาหาคำตอบ โดยใช้วิธีการรวมแบบจำลองหรือการโหวตจากเสียงข้างมาก (Majority Vote) เพื่อให้ได้คำตอบที่ดีที่สุด

2.5.1 Bagging Method

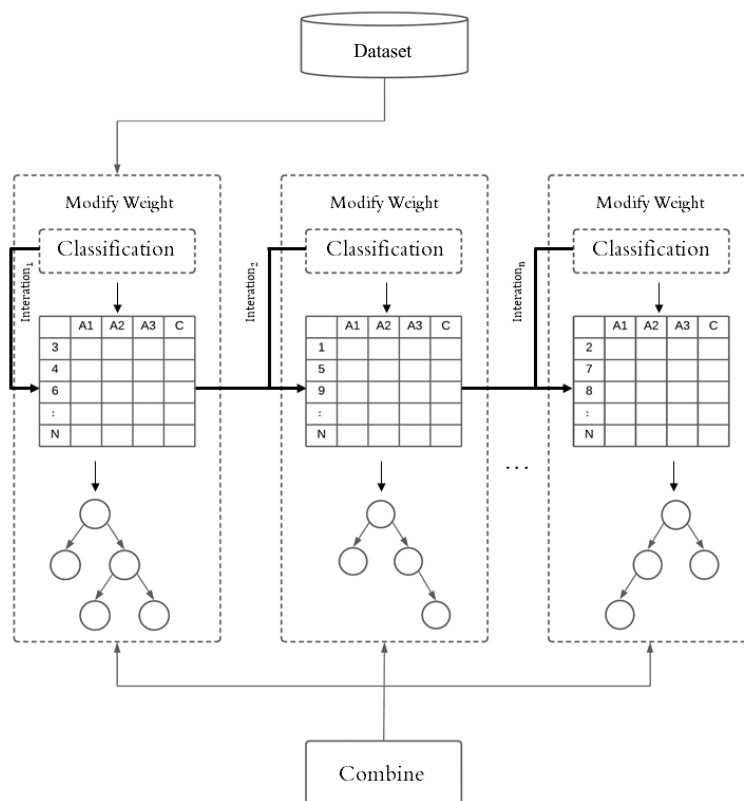
วิธีการเรียนรู้แบบรวมกลุ่มโดยใช้วิธี Bagging หรือเรียกอีกอย่างว่า Bootstrap Aggregating เป็นวิธีหนึ่งของการเรียนรู้แบบรวมกลุ่มที่มีประสิทธิภาพ โดยส่วนใหญ่จะใช้งานร่วมกับการจำแนกประเภทในรูปแบบต้นไม้ช่วยตัดสินใจแต่ก็สามารถใช้งานร่วมกับอัลกอริทึมการจำแนกประเภทในรูปแบบอื่นได้ ซึ่งกระบวนการทำงานจะสุ่มชุดข้อมูลด้วยตัวอย่างขึ้นมาใหม่โดยใช้วิธีที่เรียกว่า “Bootstrap Sampling” ซึ่งข้อมูลในแต่ละชุดที่ถูกสุ่มขึ้นมาจะถูกเรียนรู้ที่มีลักษณะต่างกันในการสร้างแบบจำลอง เนื่องจากข้อมูลที่ถูกเรียนรู้มีความแตกต่างกันทำให้เกิดความหลากหลายของแบบจำลอง ผลลัพธ์สุดท้ายจะทำการโหวตเสียงข้างมากจากแบบจำลองที่สร้างขึ้นเหล่านั้นเพื่อให้ได้ผลลัพธ์ที่ดีที่สุดเพียงหนึ่งคำตอบ วิธีการทำงานแบบ Bagging จะมีผลกระทบต่อการทำงานกับข้อมูลที่ไม่เป็นเชิงเส้น (Non-Linear) เมื่อข้อมูลที่ใช้ในการเรียนรู้เปลี่ยนแปลงเพียงเล็กน้อย สามารถแสดงวิธีการทำงานแบบ Bagging ได้ดังภาพที่ 4



ภาพที่ 4 โครงสร้างวิธีการทำงานแบบ Bagging

2.5.2 Boosting Method

Boosting เป็นวิธีการเรียนรู้แบบรวมกลุ่มอีกหนึ่งวิธีที่นิยมนำมาใช้ในการสร้างแบบจำลองการเรียนรู้แบบรวมกลุ่ม ลักษณะจะแตกต่างจากวิธีการของ Bagging ในส่วนของการถ่วงน้ำหนักให้กับข้อมูล ตัวอย่างที่ได้ทำการเรียนรู้โดยเน้นไปที่การหาค่าความผิดพลาดที่เกิดขึ้นจากกระบวนการเรียนรู้ข้อมูลเรียกลักษณะนี้ว่า “Weak Learning” และในขั้นตอนสุดท้ายจะใช้วิธีการรวมตัวแบบจำแนกประเภทที่สร้างขึ้นไปหลายๆ ตัว จำแนกโดยพิจารณาจากค่าเฉลี่ยในการถ่วงน้ำหนัก (Mean Weight) และทำการโหวตเพื่อให้ได้ผลลัพธ์เพียงคำตอบเดียว สำหรับวิธีการของ Boosting ที่นิยมนำมาใช้คือ Adaptive Boosting (AdaBoost) วิธีการของ AdaBoost จะทำการถ่วงน้ำหนักให้กับข้อมูลตัวอย่างที่ถูกเรียนรู้ของแต่ละรอบในการสร้างแบบจำลอง โดยข้อมูลตัวอย่างที่ทำการจำแนกประเภทให้ถูกต้องจะถูกลดค่าน้ำหนักลง ส่วนข้อมูลตัวอย่างที่จำแนกประเภทผิดพลาดจะถูกเพิ่มค่าน้ำหนักให้มีความสำคัญมากขึ้นเพื่อให้ข้อมูลนั้นมีโอกาสถูกเลือกในการเรียนรู้รอบต่อไป



ภาพที่ 5 แสดงโครงสร้างวิธีการทำงานแบบ Boosting (AdaBoost)

สามารถแสดงว่า Boosting แบบ AdaBoost ได้ดังภาพที่ 5 เนื่องจาก AdaBoost เป็น อัลกอริทึมที่มีการปรับปรุงค่าน้ำหนักให้กับข้อมูล ดังนั้นการหาค่าน้ำหนักจะอาศัยค่าความผิดพลาดที่เกิดขึ้นจากกระบวนการเรียนรู้ ของข้อมูลเป็นหลัก

$$\varepsilon_i = \sum_{k: c_i(x_k) \neq y_k} D_i(k) \quad (1)$$

เริ่มจากการคำนวณหาค่าความผิดพลาดที่เกิดขึ้นในการเรียนรู้ข้อมูลตามสมการ (1) โดยที่ เป็นการกำหนด D_i น้ำหนักให้กับชุดข้อมูล $D_i = 1/m$ เมื่อ m คือ จำนวนข้อมูลตัวอย่างทั้งหมด และ C_i เป็นตัวจำแนกประเภทที่จำแนกข้อมูลตัวที่ k ผิดพลาด ซึ่ง ε_i คือ ผลรวมของค่าความผิดพลาดที่ได้จากตัวจำแนกประเภทที่จำแนกข้อมูลตัวที่ k ผิดพลาดหลังจากนั้นทำการคำนวณค่าน้ำหนักของข้อมูลในสมการ (2)

$$\alpha_i = \frac{1}{2} \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right) \quad (2)$$

ทำการปรับปรุงค่าน้ำหนักข้อมูลตัวอย่าง แต่ละตัวจากสมการ (3) และที่ Z_i คือปัจจัยความเป็นปกติ (Normalization Factor) หาได้จากสมการ (4)

$$D_{i+1}(k) = \frac{D_i(k)}{Z_i} \times \begin{cases} e^{-\alpha} & \text{if } c_i(x_k) = y_k \\ e^{\alpha} & \text{if } c_i(x_k) \neq y_k \end{cases} \quad (3)$$

$$Z_i = 2\sqrt{\varepsilon_i(1-\varepsilon_i)} \quad (4)$$

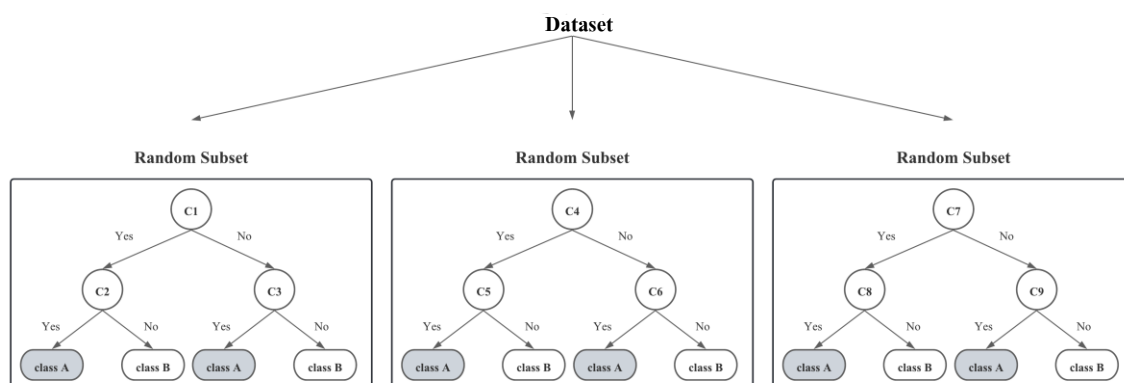
ผลลัพธ์สุดท้ายที่ได้จากกระบวนการจำแนกประเภทข้อมูลด้วยวิธี Boosting แบบ AdaBoost สามารถแสดงได้ดังสมการที่ (5)

$$C(x) = \text{Sign}\left(\sum_{i=1}^j \alpha_i c_i(x)\right) \quad (5)$$

โดยที่ α_i คือค่าน้ำหนักของข้อมูลทั้งหมดที่คำนวณได้ในแต่ละรอบจากการคำนวณหาค่าความผิดพลาดในการจำแนกข้อมูล และ C_i คือตัวจำแนกประเภทข้อมูล

2.5.3 แรนดอมฟอเรส (Random Forest: RT)

เป็นแบบจำลองที่ถูกจัดให้เป็นอัลกอริทึมการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) และใช้พื้นฐานจากต้นไม้ตัดสินใจเป็นการทำนายแบบชุดของ Decision Tree หลายต้น (Ensemble of Decision Trees) โดยสร้างจากการสุ่มข้อมูลตัวอย่างแบบเลือกแล้วใส่กลับคืน (Random Sampling with Replacement) เพื่อนำมาสร้างเป็นแบบจำลองต้นไม้โดยแต่ละต้นมีลักษณะที่สำคัญ โดยแต่ละแบบจำลองจะมีการทำนายผล ซึ่งผลจากการทำนายของต้นไม้แต่ละต้นจะทำการโหวต เลือกผลการทำนายที่ได้รับการโหวตมากที่สุด ภาพที่ 6 Random Forest คือภาพประกอบของ Random Forest ซึ่งประกอบด้วยต้นไม้ตัดสินใจที่แตกต่างกันสามต้น ต้นไม้การตัดสินใจทั้งสามต้นนั้นจะได้รับการฝึกอบรมโดยใช้ชุดย่อยแบบสุ่มของข้อมูลการฝึกอบรม (Train Data) (ธนัท จริยะสมบูรณ์ และ วราภรณ์ วิทยานนท์ 2561)



ภาพที่ 6 Random Forest

2.6 มาตรวัดและการทดสอบประสิทธิภาพแบบจำลอง

2.6.1 การวัดประสิทธิภาพแบบจำลอง

ความสามารถในการวินิจฉัยของตัวจำแนกประเภทถูกกำหนดโดย Confusion Matrix ตารางที่ 1 Confusion Matrix ประกอบไปด้วยตาราง 2X2 มีแนวตั้งคือผลการทำนาย (Prediction Class) และแนวนอนคือค่าจริง (Actual Class) ภายในตาราง 2X2 (Gatchalee, 2019) จะประกอบไปด้วย

True Positive (TP) คือความถี่สิ่งที่ทำนายว่า 'จริง' และมีค่าเป็น 'จริง'

True Negative (TN) คือความถี่สิ่งที่ทำนายว่า 'ไม่จริง' และมีค่าเป็น 'ไม่จริง'

False Positive (FP) คือความถี่สิ่งที่ทำนายว่า 'จริง' และมีค่าเป็น 'ไม่จริง' และ

False Negative (FN) คือความถี่สิ่งที่ทำนายว่า 'ไม่จริง' และมีค่าเป็น 'จริง'

ตารางที่ 1 Confusion Matrix

		Predicted Class	
		P	N
Actual Class	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

โดยทั่วไปมีการวัดประสิทธิภาพของตัวแยกประเภทโดยอิงจาก ตารางที่ 1 Confusion Matrix สามารถนำมาคำนวณเพื่อหาค่าวัดประสิทธิภาพสำหรับตัวจำแนกประเภท ซึ่งประกอบไปด้วย

- 1) ค่าความถูกต้อง (Accuracy) แสดงถึงความถูกต้องในการทำนายในภาพรวมทั้งกลุ่ม Positive และ Negative คำนวณได้ดังนี้

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

- 2) ค่าความครบถ้วน (Recall) คือ อัตราส่วนระหว่างจำนวน Positive ที่ถูกจำแนกได้อย่างถูกต้อง คำนวณได้ดังนี้

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

- 3) ค่าความแม่นยำ (Precision) คือ อัตราส่วนความถูกต้องของการทำนายกลุ่ม Positive เมื่อเทียบกับผลการทำนาย Positive ทั้งหมด คำนวณได้ดังนี้

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

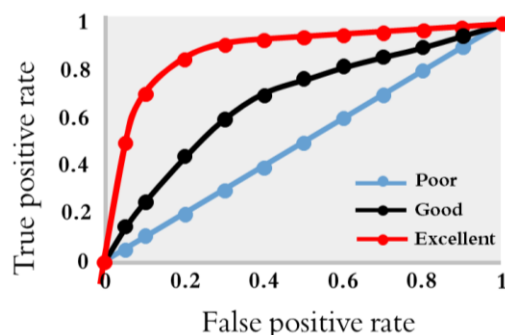
- 4) ค่าความถ่วงดุล (F1-Score) เป็นการวัดความถูกต้องโดยใช้ค่าเฉลี่ยฮาร์โมนิก ระหว่าง True Positive Rate และ Precision คำนวณได้ดังนี้

$$F_1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

2.6.2 Receiver Operating Characteristic (ROC) curve

Receiver Operating Characteristic (ROC) curve เป็นเครื่องมือพื้นฐานสำหรับการประเมินการทดสอบวินิจฉัย และสร้างขึ้นโดย True Positive Rate (1-Specificity) เทียบ False Positive Rate (Sensitivity) โดยทั่วไปแล้วเส้นโค้ง ROC (AUC) ยังใช้กำหนดความสามารถในการคาดการณ์ของตัวแยกประเภท ภาพที่ 7 ROC curve แสดงการนำเสนอเส้นโค้ง ROC สามเส้นตามชุดข้อมูล ค่า AUC สำหรับเส้นโค้ง ROC สีแดงมีความแม่นยำในการคาดการณ์ที่ดีมาก เส้นโค้ง ROC สีดำมีความแม่นยำในการคาดการณ์ที่ดี และเส้นโค้ง ROC สีฟ้ามีความแม่นยำในการคาดการณ์ที่ต่ำ โดยที่พื้นที่ใต้เส้นโค้ง ROC สีฟ้าคือครึ่งหนึ่งของสี่เหลี่ยมผืนผ้าที่แรงเงา ค่า AUC สำหรับเส้นโค้ง ROC สีฟ้าคือ 0.5

ดังนั้นตัวแยกประเภทที่สร้างเส้นโค้ง ROC สีแดงจะมีความแม่นยำในการคาดการณ์ที่สูงกว่าเมื่อเทียบกับตัวแยกประเภทอื่นที่สร้างเส้นโค้ง ROC สีดำ และสีฟ้า (Uddin et al, 2019)



ภาพที่ 7 ROC curve

2.6.3 การแบ่งข้อมูลเพื่อวัดประสิทธิภาพแบบจำลอง

Holdout คือ การแบ่งชุดข้อมูลออกเป็น 2 ชุดข้อมูลย่อยด้วยวิธีการสุ่ม โดยชุดข้อมูลที่ได้จะเป็นชุดข้อมูลฝึกสอนและชุดทดสอบ ซึ่งโดยปกติชุดข้อมูลฝึกสอนจะมีปริมาณข้อมูลเท่ากับ 2 ใน 3 ของชุดข้อมูลทั้งหมด และชุดข้อมูลทดสอบจะมีปริมาณ 1 ใน 3 ของชุดข้อมูล หลังจากแบ่งชุดข้อมูลแล้วจะนำชุดข้อมูลฝึกสอนจะถูกใช้ในการสร้างตัวจำแนกข้อมูล และชุดข้อมูลทดสอบจะถูกใช้ในการทดสอบตัวจำแนกที่สร้างขึ้น

Cross Validation คือ การแบ่งข้อมูลเพื่อใช้สำหรับวัดประสิทธิภาพแบบจำลอง โดยใช้วิธีการ Cross Validation ซึ่งเป็นวิธี ที่ได้รับความนิยมสำหรับการแบ่งข้อมูล เพื่อประสิทธิภาพแบบจำลองเนื่องจากผลลัพธ์ที่ได้มีความน่าเชื่อถือ โดยหลักในการแบ่งข้อมูลด้วยวิธีนี้จะเริ่มจากการกำหนดค่า k หรือการแบ่งข้อมูลออกเป็น k ส่วนเท่า ๆ กัน

2.7 งานวิจัยที่เกี่ยวข้อง

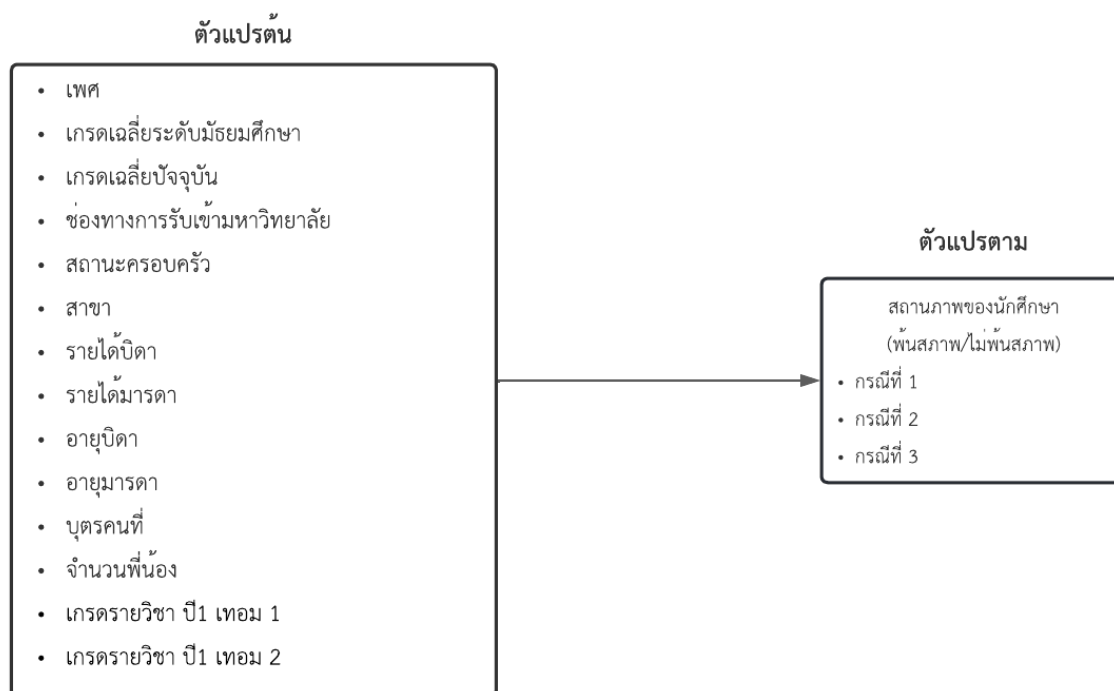
2.7.1 การทำนายผลการลาออกของนักศึกษา โดยทำการแบ่งเป็น ‘ลาออก’ และ ‘ไม่ลาออก’ และทำความสะอาดข้อมูล แปลงข้อมูล สกัดคุณลักษณะ และปรับสมดุลของข้อมูล ซึ่งเป็นการเตรียมข้อมูลสำหรับขั้นตอนการจำแนกประเภทสำหรับการลาออกของนักศึกษา โดยได้ทำการเปรียบเทียบอัลกอริทึม Decision Tree, Random Forest และ Gradient Boosting ผลการทดลองพบว่า การใช้อัลกอริทึม Gradient Boosting ให้ประสิทธิภาพความแม่นยำดีที่สุดในการจำแนกประเภทการลาออกของนักศึกษา (ร้อยละ 93) และคุณลักษณะที่สำคัญได้แก่ ปีการศึกษาของนักเรียน เกรดเฉลี่ยของโรงเรียนมัธยม ช่องทางการรับเข้ามหาวิทยาลัย คณาจารย์ของนักเรียน และเพศ (Tenpipat, & Akkarajitsakul, 2020)

- 2.7.2 การใช้แบบจำลองต้นไม้ตัดสินใจแบบรวมกลุ่มเพื่อทำนายการลาออกของนักศึกษา โดยทำการปรับขอบเขตข้อมูลโดยใช้วิธี Min-Max Normalization ให้อยู่ในช่วง $[0,1]$ สำหรับขั้นตอนการจำแนกประเภทของการลาออกของนักศึกษา โดยใช้อัลกอริทึม Random Forest และการใช้ K-Fold cross validation ที่ $K = 5,10$ ในการวัดประสิทธิภาพของตัวแบบ จากผลการทดลองพบว่า การใช้ Random Forest ที่ $K=5,10$ ให้ประสิทธิภาพความแม่นยำในการจำแนกการลาออกของนักศึกษาเท่ากับ 81.77% และ 80.11% ตามลำดับ (Naseem et al., 2020)
- 2.7.3 ระบบการทำนายในการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์ ด้วยเทคนิคการทำเหมืองข้อมูล ได้นำเสนอเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มาช่วยในการสร้างกฎเพื่อจะพัฒนาระบบการทำนายการฟื้นฟูสภาพของนักศึกษา โดยใช้ข้อมูล คุณลักษณะ สาขาวิชาที่ศึกษาในคณะวิทยาศาสตร์ เกรดเฉลี่ยในภาคเรียนที่ 1-6 เกรดเฉลี่ยจากโรงเรียนมัธยม แผนการเรียนที่ศึกษาในโรงเรียนมัธยม ขนาดโรงเรียน สถานะกู้ยืมเพื่อการศึกษา สถานะการฟื้นฟูสภาพ โดยรวมมีตัวแปรที่เกี่ยวข้องทั้งหมด 7 ตัวแปร ผลลัพธ์จากงานวิจัยพบว่า รูปแบบการทำนายการฟื้นฟูสภาพด้วยวิธีต้นไม้ตัดสินใจมีจำนวน 32 กฎ ประเมินโดยใช้ 10-Folds Cross Validation มีค่าความถูกต้องเฉลี่ย 95.57% (นนทวัฒน์ ทวีชาติ และคณะ, 2563)
- 2.7.4 การวิเคราะห์ปัจจัยที่มีผลต่อการฟื้นฟูสภาพของนักศึกษาโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษา หลักสูตรวิทยาการคอมพิวเตอร์หลักสูตรเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏยะลา ได้มีการนำเสนอเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคโครงข่ายประสาทเทียมแบบย้อนกลับ (Back Propagation Neural Network : BP-NN) และเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) มาทำการเปรียบเทียบกับ 10-Fold Cross Validation จากผลลัพธ์การทดลองพบว่า SVM ให้ประสิทธิภาพในการจำแนกประเภทโดยเฉลี่ยสูงที่สุด 97.75 % โดยแต่ละโมเดลให้ความแม่นยำเฉลี่ยมากกว่า 97% และปัจจัยที่ส่งผลต่อการฟื้นฟูสภาพของนักศึกษามี ผลการเรียนรู้รายวิชาพื้นฐานทางฟิสิกส์ วิชาแพลตฟอร์มเทคโนโลยี ผลการเรียนรู้เฉลี่ยปีที่สอง และผลการเรียนรู้รายวิชาโครงสร้างข้อมูล (ซอและ เกป็น, พิมลพรรณ ลีลาภทรพันธุ์, และอัจฉราพร ยกขุน, 2561)

ตารางที่ 2 เปรียบเทียบวิธีการสร้างแบบจำลอง ความแม่นยำ และปัจจัยที่มีความสำคัญกับการฟื้นฟูสภาพของนักศึกษาระหว่างการศึกษาดังกล่าว

ผู้วิจัย	Machine Learning Model	Train-test Data	ความแม่นยำ (%)	ปัจจัยการฟื้นฟูสภาพของนักศึกษา
Tenpipat, & Akkarajitsakul, 2020	1) Decision Tree 2) Random Forest 3) Gradient Boosting	1) 10-Fold CV	Gradient Boosting ให้ประสิทธิภาพความแม่นยำโดยเฉลี่ยดีที่สุด 93%	ปีการศึกษาของนักเรียน, เกรดเฉลี่ย, เกรดเฉลี่ยระดับมัธยมศึกษา, ช่องทางการรับเข้ามหาวิทยาลัย, คณาจารย์ของนักเรียน, และเพศ
Naseem et al., 2020	1) Random Forest	1) 5-Fold CV 2) 10-Fold CV	Random Forest ที่ K=5,10 ให้ประสิทธิภาพความแม่นยำเท่ากับ 81.77% และ 80.11% ตามลำดับ	การได้รับทุน, เพศ, อายุ, เกรด, คะแนนงานต่างๆ
นนทวัฒน์ ทวีชาติ และคณะ, 2563	1) Decision Tree	1) 10-Folds CV	วิธีต้นไม้ตัดสินใจมีจำนวน 32 กฎ และมีค่าความถูกต้องเฉลี่ย 95.57%	สาขาวิชา, เกรดเฉลี่ยในภาคเรียนที่ 1-6, เกรดเฉลี่ยจากโรงเรียนมัธยม, แผนการเรียนที่ศึกษาในโรงเรียนมัธยม, ขนาดโรงเรียน, สถานะกู้ยืมเพื่อการศึกษา และสถานะการฟื้นฟูสภาพ
ชอและ เกป็น, พิมพ์พรหม และคณะ, 2561	1) Decision Tree 2) Back Propagation Neural Network 3) Support Vector Machine	1) 10-Fold Cross Validation	SVM ให้ประสิทธิภาพในการจำแนกประเภทโดยเฉลี่ยสูงที่สุด 97.75 %	ผลการเรียนรายวิชาพื้นฐานทางฟิสิกส์, วิชาแพลตฟอร์มเทคโนโลยี, ผลการเรียนเฉลี่ยปีที่สอง และผลการเรียนรายวิชาโครงสร้างข้อมูล

2.7 กรอบแนวคิดงานวิจัย



ภาพที่ 8 กรอบแนวคิดการวิจัย

บทที่ 3

วิธีการดำเนินการวิจัย

การศึกษาวิจัยเรื่อง “การจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้วิธีการเรียนรู้แบบรวมกลุ่ม” ได้ทำการเปรียบเทียบระหว่างวิธีการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยว และแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม ผู้วิจัยได้ทำการศึกษาแนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง เพื่อพัฒนาประสิทธิภาพการจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้เทคนิควิธีการเรียนรู้แบบรวมกลุ่ม โดยได้กำหนดวิธีการดำเนินงาน วิจัยซึ่งประกอบด้วยเนื้อหาดังต่อไปนี้

3.1 เครื่องมือที่ใช้ในงานวิจัย

3.2 กรอบวิธีการดำเนินการวิจัย

3.3 การเตรียมข้อมูลสำหรับการใช้ในการทดลอง

- 3.3.1 การสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยว
- 3.3.2 การสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม
- 3.3.3 การเปรียบเทียบผลลัพธ์ของประสิทธิภาพความถูกต้องในการจำแนกประเภท
- 3.3.4 การทำนายการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2564

3.1 เครื่องมือที่ใช้ในการวิจัย

3.1.1 เทคนิคที่ใช้ในการเหมืองข้อมูล (Data Mining Tool)

การจัดประเภท (Classification) เป็นการจัดประเภทของข้อมูล โดยหาตัวแบบการจำแนกประเภทข้อมูล ซึ่งตัวแบบสร้างจากการวิเคราะห์ชุดของข้อมูลฝึกสอน (Training Data) โดยเป็นกลุ่มข้อมูลที่มีการระบุกลุ่มผลลัพธ์เรียบร้อยแล้ว วัตถุประสงค์เพื่อใช้เป็นตัวแบบในการทำนายข้อมูลที่ไม่มีเคยเห็นมาก่อน

ในการวิจัยครั้งนี้ได้ใช้อัลกอริทึมดังนี้

- 1) Decision Tree (DT)
- 2) Support Vector Machine (SVM)
- 3) Bagging (DT base model)
- 4) Bagging (SVM base model)
- 5) Boosting (DT base model)

6) Boosting (SVM base model)

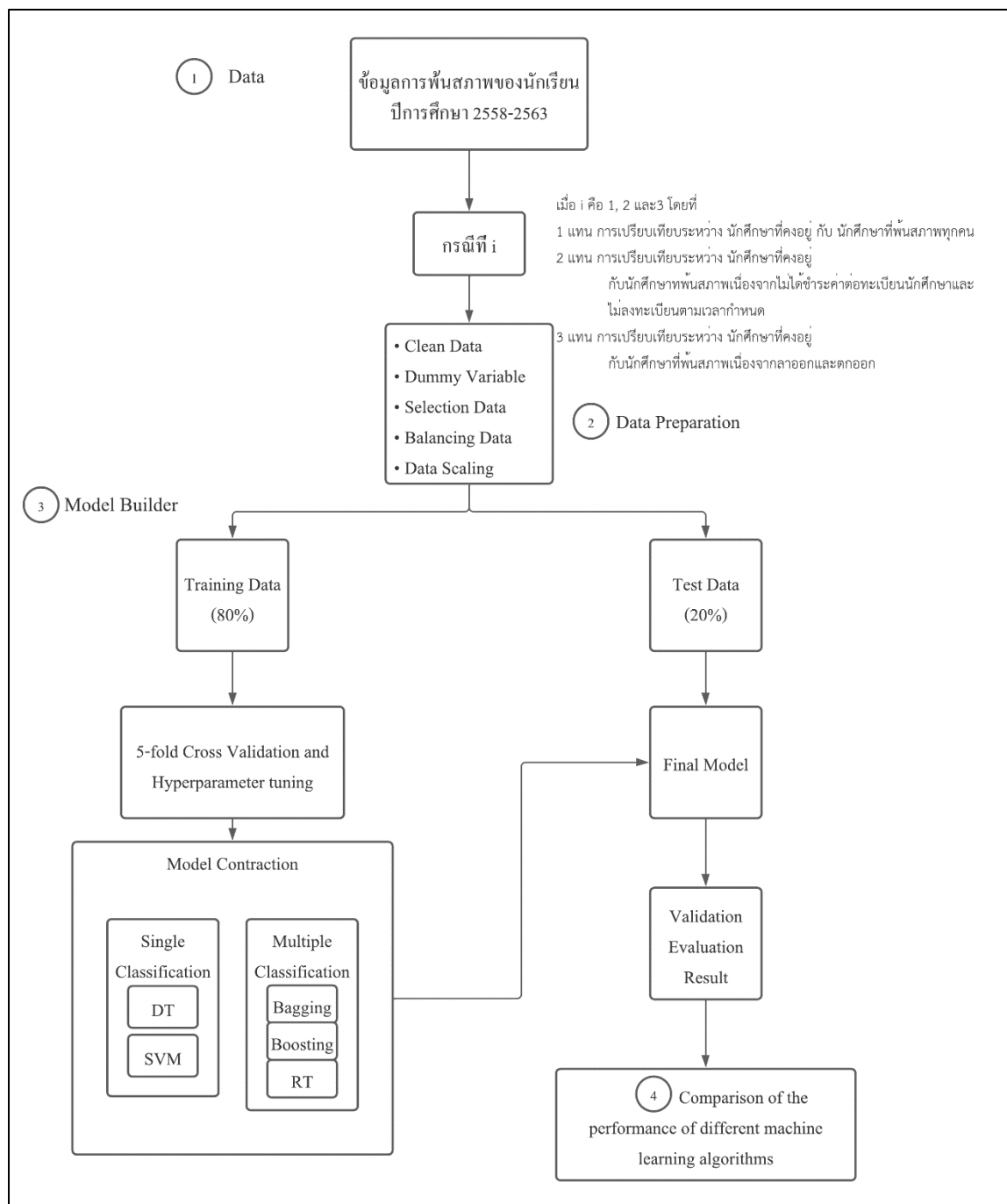
7) Random Forest (RT)

3.1.2 โปรแกรมที่ใช้ในการประมวลผล (Software Tools)

ผู้วิจัยได้ใช้โปรแกรมทางคอมพิวเตอร์ Google Colaboratory ในการประมวลผลทางข้อมูล ซึ่งเป็นโปรแกรมที่นักวิทยาศาสตร์ทางข้อมูลนิยมใช้กันเป็นจำนวนมากในการทำเหมืองข้อมูล และจัดการกับข้อมูลที่มีขนาดใหญ่ ในการวิจัยครั้งนี้ผู้วิจัยได้ใช้ Library Pandas ในการจัดการกับข้อมูลที่มีขนาดใหญ่ และ Library Scikit-Learn ในการเลือกใช้อัลกอริทึมต่างๆ ในการสร้างตัวจำแนกประเภท

3.2 กรอบวิธีการดำเนินการวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้เสนอกรอบวิธีการดำเนินงานวิจัย รวมถึงการออกแบบการทดลองโดยแบ่งออกเป็น 3 ส่วนหลักๆ แสดงดังภาพที่ 9 ในส่วนที่ 1 การจัดเตรียมข้อมูลการพื้นฐานของนักศึกษาปีการศึกษา 2558 – 2563 จำนวน 614 คน โดยได้ทำการแบ่งชุดข้อมูลที่ใช้สำหรับการทดลองออกเป็น 3 กรณี เพื่อใช้สำหรับการทดลองเป็นกระบวนการในการจัดเตรียมข้อมูลก่อนการสร้างโมเดลโดยจะนำข้อมูลมาทำความสะอาดข้อมูล ตัวแปรหุ่น ทำการคัดเลือกข้อมูล ปรับสมดุลของข้อมูล และการปรับปรุงขอบเขตข้อมูล หลังจากผ่านขั้นตอนกระบวนการเตรียมข้อมูลในส่วนที่ 1 ได้นำข้อมูลมาแบ่งเป็นข้อมูลฝึกสอน (Training Data) 80% และข้อมูลทดสอบ (Test Data) 20% จากนั้นนำข้อมูลฝึกสอน (Training Data) แบ่งข้อมูลออกเป็น 5 ชุดโดยใช้วิธี Cross Validation สำหรับการปรับพารามิเตอร์ของโมเดลให้เหมาะสมที่สุดโดยใช้ Grid Search หลังจากนั้นนำข้อมูลที่ได้เตรียมไว้เข้าสู่กระบวนการเรียนรู้การจำแนกประเภท ในขั้นตอนนี้ได้ใช้อัลกอริทึม Decision Tree และ Support Vector Machine เรียนรู้ข้อมูลตัวอย่างในแต่ละชุดเพื่อใช้ในการจำแนก ในการสร้างแบบจำลองการจำแนกประเภทแบบเดียว และทำการเปรียบเทียบการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม 3 วิธี ได้แก่ Bagging Boosting และ Random Forest โดยใช้อัลกอริทึม Decision Tree และ Support Vector Machine เป็นอัลกอริทึมพื้นฐานใช้งานร่วมกับวิธีการเรียนรู้แบบรวมกลุ่ม 2 วิธี คือ Bagging และ Boosting ส่วนวิธีการของ Random Forest จะใช้ต้นไม้ช่วยตัดสินใจซึ่งเป็นอัลกอริทึมพื้นฐานในการเรียนรู้ของตัวอัลกอริทึมเอง จากนั้นใช้ข้อมูลทดสอบ (Test Data) ที่อัลกอริทึมไม่เคยได้เห็นหรือยังไม่ได้เรียนรู้สำหรับวัดประสิทธิภาพแบบจำลองแต่ละแบบจำลอง เพื่อให้ได้แบบจำลองที่มีความน่าเชื่อถือ และในลำดับสุดท้ายจะดำเนินการเปรียบเทียบผลลัพธ์ของประสิทธิภาพความถูกต้องในการจำแนกประเภทที่ได้จากการทดลองทั้งหมด ซึ่งจะอธิบายแต่ละส่วนดังนี้



ภาพที่ 9 กรอบวิธีการดำเนินงานวิจัยสำหรับการทำแนกประเภท

3.3 การเตรียมข้อมูลสำหรับการสร้างแบบจำลองการจำแนกประเภท

1) ชุดข้อมูล

ในการวิจัยครั้งนี้ได้รับข้อมูลตัวอย่างการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี จากสำนักทะเบียนและประมวลผล มหาวิทยาลัยขอนแก่น โดยใช้ข้อมูลของนักศึกษาปีการศึกษา 2558-2563 จำนวน 614 คน แสดงในตารางที่ 3 โดยได้ทำการแบ่งชุดข้อมูลที่ใช้สำหรับการทดลองออกเป็น 3 กรณีดังนี้

1.1) ข้อมูลกรณีที่ 1

การฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากลาออก ฟื้นฟูสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา ฟื้นฟูสภาพเนื่องจากตกออก และฟื้นฟูสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด

1.2) ข้อมูลกรณีที่ 2

การฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา และฟื้นฟูสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด

1.3) ข้อมูลกรณีที่ 3

การฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากลาออก และฟื้นฟูสภาพเนื่องจากตกออก

ตารางที่ 3 ชื่อตัวแปรและความหมาย

ตัวแปร	ความหมาย
เพศ	0 = หญิง 1 = ชาย
เกรดเฉลี่ยระดับมัธยมศึกษา	คะแนนเฉลี่ยสะสมของโรงเรียนมัธยมศึกษา
เกรดเฉลี่ยปัจจุบัน	คะแนนเฉลี่ยสะสมปัจจุบัน
ช่องทางการรับเข้ามหาวิทยาลัย	0 = สอบคัดเลือกจากระบบกลาง (Admissions) 1 = สอบคัดเลือกประเภทโควตาภาค ตะวันออกเฉียงเหนือ 2 = โควตาทั่วประเทศ

ตัวแปร	ความหมาย
	<p>3 = โครงการรับนักเรียนที่เป็นผู้มีคุณธรรม จริยธรรม และบริการสังคม</p> <p>4 = โครงการร่วมรับนักศึกษาภาคใต้กับ มหาวิทยาลัยสงขลานครินทร์ โดยวิธีรับตรง</p> <p>5 = การคัดเลือกโดยวิธีพิเศษ</p> <p>6 = โครงการร่วมรับนักศึกษาภาคเหนือกับ มหาวิทยาลัยเชียงใหม่ โดยวิธีรับตรง</p> <p>7 = โครงการมูลนิธิส่งเสริมโอลิมปิกวิชาการ ฯ (สอวน.)</p>
สถานะครอบครัว	<p>0 = อยู่ด้วยกัน</p> <p>1 = หย่าขาดจากกัน</p> <p>2 = บิดาถึงแก่กรรม</p> <p>3 = แยกกันอยู่เพราะเหตุผลอื่นๆ</p> <p>4 = แยกกันอยู่เพราะความจำเป็นเกี่ยวกับ อาชีพ</p> <p>5 = มารดาถึงแก่กรรม</p> <p>6 = บิดามารดาถึงแก่กรรม</p>
สาขา	<p>0 = สารสนเทศสถิติ</p> <p>1 = สถิติ</p>
รายได้ปัจจุบันของมารดา	รายได้ปัจจุบันของมารดาต่อเดือน
รายได้ปัจจุบันของบิดา	รายได้ปัจจุบันของบิดาต่อเดือน
อายุบิดา	อายุบิดา ณ เวลาที่เก็บข้อมูล
อายุมารดา	อายุมารดา ณ เวลาที่เก็บข้อมูล
บุตรคนที่	บุตรคนที่
จำนวนพี่น้อง	จำนวนพี่น้องที่กำลังศึกษาอยู่
เกรดรายวิชา ปี 1 เทอม 1	<p>0 = A</p> <p>1 = B+</p> <p>2 = B</p>

ตัวแปร	ความหมาย
	3 = C+ 4 = C 5 = D+ 6 = D 7 = F 8 = S 9 = W 10 = S AU 11 = U
เกรดรายวิชา ปี 1 เทอม 2	0 = A 1 = B+ 2 = B 3 = C+ 4 = C 5 = D+ 6 = D 7 = F 8 = S 9 = W 10 = S AU 11 = U
สถานภาพ	0 = นักศึกษาปัจจุบัน สถานะปกติ 1 = สำเร็จการศึกษา 2 = พ้นสภาพเนื่องจากลาออก 3 = พ้นสภาพเนื่องจากไม่ชำระค่าต่อทะเบียน นักศึกษา 4 = พ้นสภาพเนื่องจากตกออก 5 = พ้นสภาพเนื่องจากไม่ลงทะเบียนตาม เวลากำหนด

ตัวแปร	ความหมาย
	6 = ลาพักการเรียน 7 = มารายงานตัวแล้ว

2) การทำความสะอาดข้อมูล (Clean Data)

การทำความสะอาดข้อมูล คือการลดข้อมูลที่ผิดปกติ และข้อมูลที่สูญหาย เพื่อให้ข้อมูลมีความสมบูรณ์และพร้อมสำหรับการทดลอง ซึ่งใช้วิธี KNNImputer ในการประมาณค่าข้อมูลสูญหาย สามารถประมาณค่าได้โดยใช้จุดข้อมูลที่ใกล้ที่สุดในตัวแปรอื่นๆ และพิจารณาความสัมพันธ์ของข้อมูล ซึ่ง KNNImputer สามารถการประมาณค่าข้อมูลสูญหายได้ทั้งตัวแปรที่เป็น Numerical และ Categorical

2.1) ตัวแปรรายได้บิดา รายได้มารดา อายุบิดาอายุมารดา บุตรคนที่ และจำนวนพี่น้องมีข้อมูลที่ผิดปกติคือ {"ไม่แน่นอน, ไม่ทราบ, -, 15,000บาท, 40,000 - 50,000"} ในข้อมูลที่มีการกรอกรายแบบช่วงรายได้จะทำการหาค่าเฉลี่ย ส่วนข้อมูลที่กรอกรายมา 15,000บาททำการลบข้อมูลที่เป็นหนังสือ และ นอกเหนือจากนี้ข้อมูลที่กรอกราย {"ไม่แน่นอน, ไม่ทราบ, -"} แทนเป็นค่าสูญหาย (Missing Value)

2.2) ตัวแปรรายได้บิดามารดา อายุบิดามารดา บุตรคนที่และจำนวนพี่น้องมีข้อมูลที่มีค่าสูญหาย (Missing Value) ซึ่งจะทำการประมาณค่าโดยใช้ KNNImputer

3) ตัวแปรหุ่น (Dummy Variable)

ตัวแปรหุ่น (Dummy Variable) คือ ตัวแปรที่ถูกกำหนดให้มีสองค่า (Binary) คือ 0 และ 1 โดยจะทำการแปลงข้อมูลเกรดรายวิชาปี 1 เทอม 1 และเกรดรายวิชาปี 1 เทอม 2 ให้เป็นตัวแปรหุ่นเพื่อต้องการศึกษาเกรดในรายวิชาต่างๆเป็นปัจจัยต่อการผันสภาพหรือไม่

4) การคัดเลือกข้อมูล (Selection Data)

ในขั้นตอนการการคัดเลือกข้อมูล คือการนำตัวแปรเพศ ช่องทางการรับเข้ามหาวิทยาลัย สถานะครอบครัว สาขา เกรดรายวิชาปี 1 เทอม 1 และเกรดรายวิชาปี 1 เทอม 2 ที่เป็นตัวแปรหุ่น (Dummy Variable) มาทำการคัดเลือกตัวแปรที่มีความสำคัญกับสถานภาพโดยใช้ค่าสถิติ Chi-Square ในการทดสอบ Test of Independence เพื่อที่จะตัดและลดขนาดของข้อมูลหรือคุณลักษณะที่ไม่จำเป็นในการวิเคราะห์

5) ปรับสมดุลของข้อมูล (Balancing Data)

ในการวิจัยครั้งนี้ได้ใช้วิธีสุ่มเกิน (Over Sampling) ในการปรับสมดุลของข้อมูล คือการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก ซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มเลือกจากข้อมูลเดิมในกลุ่มส่วนน้อยโดยใช้วิธีการสุ่มแบบเป็นระบบ

3.3.6 การปรับปรุงขอบเขตข้อมูล (Data Scaling)

นำข้อมูลที่ผ่านมากระบวนการคัดเลือกข้อมูล ทำความสะอาดข้อมูล และปรับสมดุลของข้อมูล มาทำการปรับปรุงขอบเขตข้อมูลที่เป็นข้อมูลเชิงปริมาณ โดยใช้วิธี Min-Max Normalization ทำให้ข้อมูลอยู่ในช่วง $[0,1]$

3.3.1 การสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยว

ในขั้นตอนการสร้างแบบจำลองการจำแนกประเภทประเภทเดี่ยว โดยใช้อัลกอริทึม Decision Tree และ Support Vector Machine เพื่อหาประสิทธิภาพความถูกต้องหรือแม่นยำสำหรับการจำแนกประเภท ซึ่งจะแสดงรายละเอียดดังต่อไปนี้

3.3.1.1 การสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม Decision Tree (DT) และ Support Vector Machine (SVM)

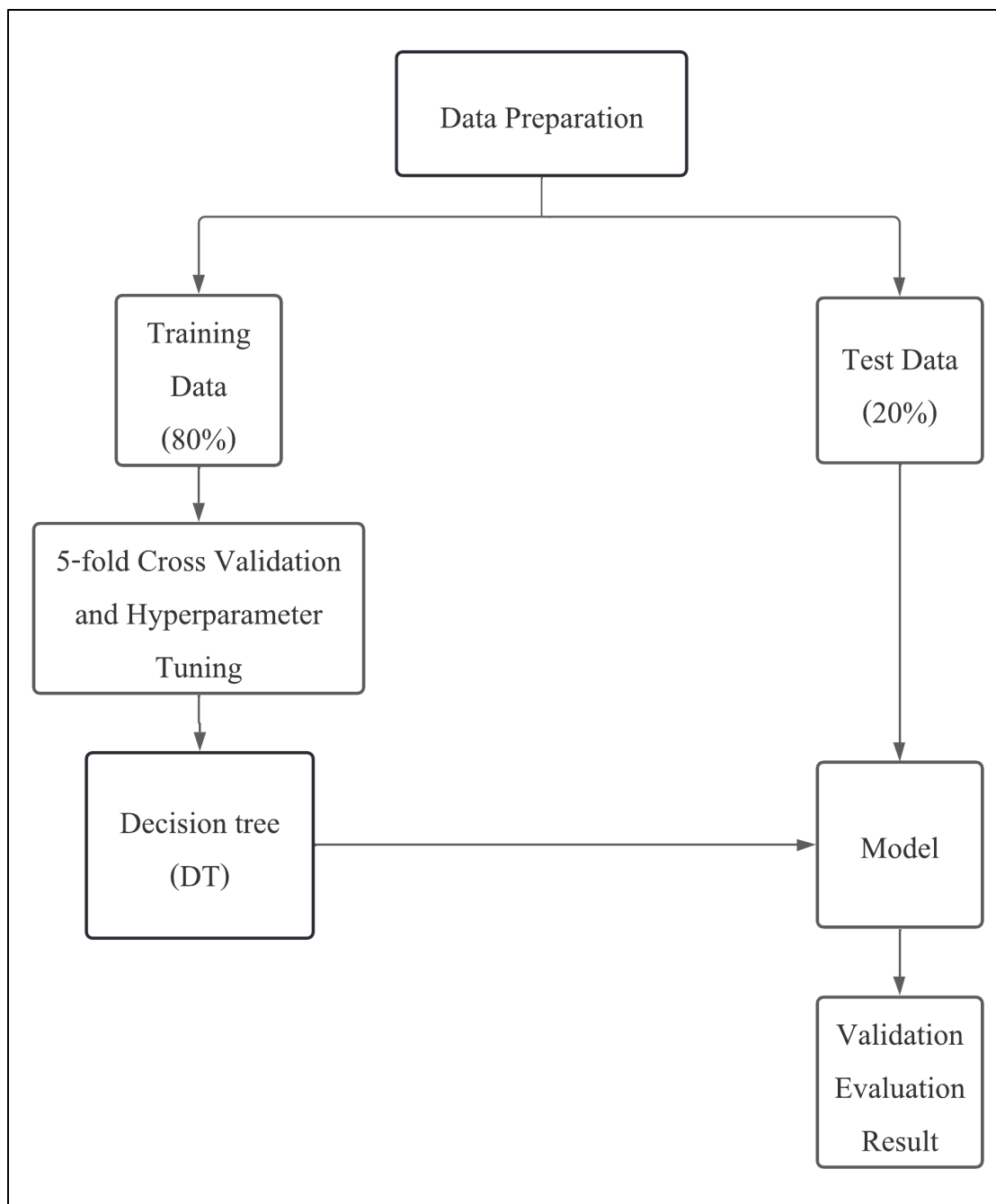
- 1) นำข้อมูลการพันสภาพกรณี 1 2 และ 3 ที่ผ่านกระบวนการเตรียมข้อมูลมาทำการแบ่งข้อมูลออกเป็น 2 ชุด คือข้อมูลฝึกสอน (Training Data) 80% และข้อมูลทดสอบ (Test Data) 20%
- 2) นำข้อมูลฝึกสอน (Training Data) มาทำการแบ่งข้อมูลให้เป็น 5 ส่วน ข้อมูลทุกส่วนจะถูกเรียนรู้และตรวจสอบในการวัดประสิทธิภาพ ด้วยวิธี 5-fold Cross Validation และทำการหาค่าพารามิเตอร์ที่เหมาะสม โดยใช้ Grid Search สำหรับอัลกอริทึม Decision Tree (DT) และ Support Vector Machine (SVM) ดังตารางที่ 4 - 5
- 3) กำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับอัลกอริทึม Decision Tree (DT) และ Support Vector Machine (SVM) หลังจากนั้นทำการเรียนรู้โดยใช้ข้อมูลทดสอบ (Test Data)
- 4) ผลลัพธ์สุดท้ายของการประเมินการตรวจสอบความถูกต้องของกระบวนการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม Decision Tree และ Support Vector Machine จะได้ค่าวัดประสิทธิภาพของโมเดล คือ Accuracy, Precision, F1-Measure, Recall, และ AUC ดังภาพที่ 10 – 11

ตารางที่ 4 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Decision Tree (DT)

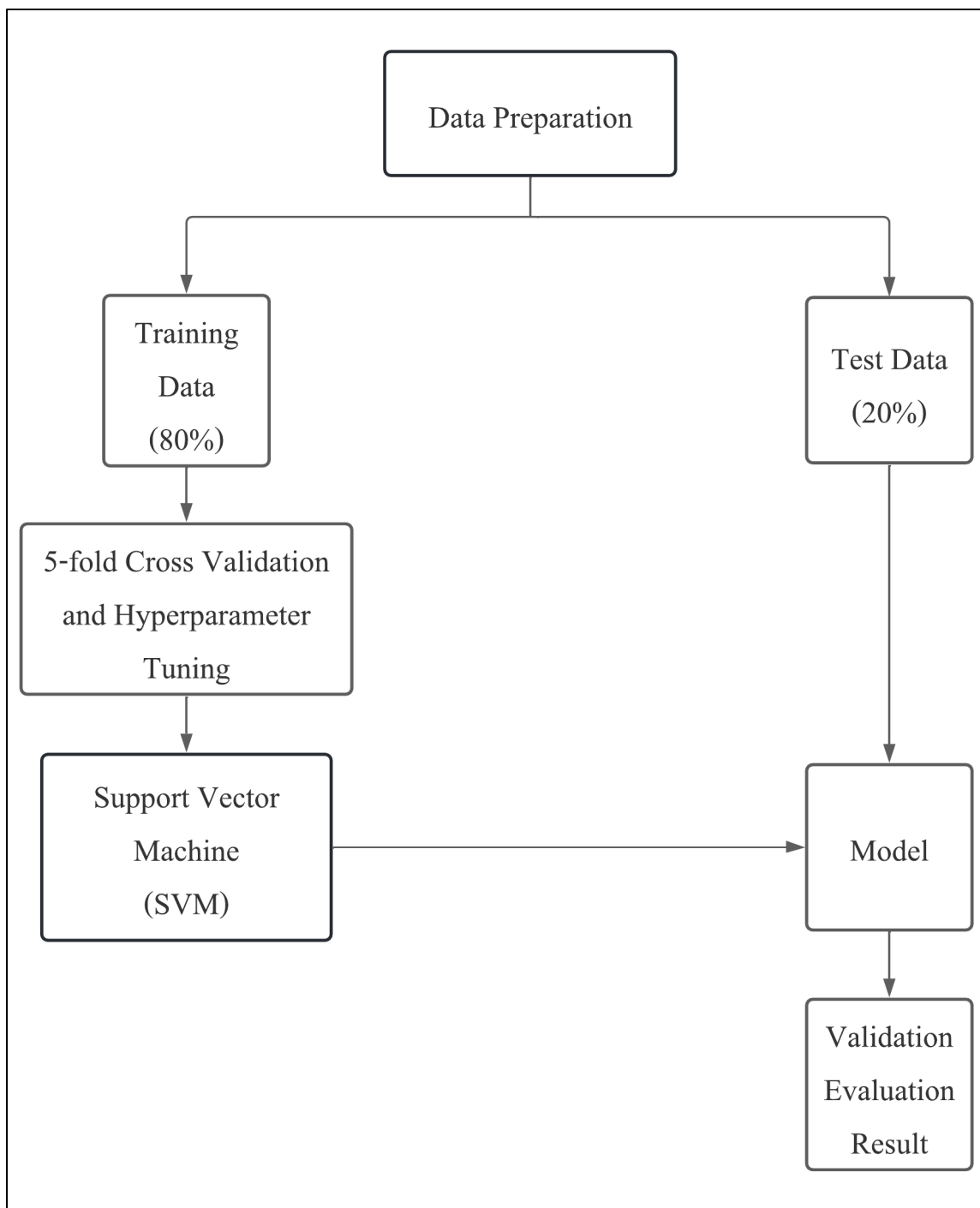
พารามิเตอร์	ค่าพารามิเตอร์
Criterion	Gini, Entropy
Max Depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
Min Samples_Leaf	2, 5, 10
Max Features	1, 2, 4, 6, 7, 8, 13, 15, 16, 17, 19

ตารางที่ 5 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Support Vector Machine (SVM)

พารามิเตอร์	ค่าพารามิเตอร์
C	0.1, 1, 10, 100
Gamma	1, 0.1, 0.01, 0.001
Kernel	Rbf, Poly, Sigmoid



ภาพที่ 10 ขั้นตอนการแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้ DT



ภาพที่ 11 ขั้นตอนการแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้ SVM

3.3.2 การสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม

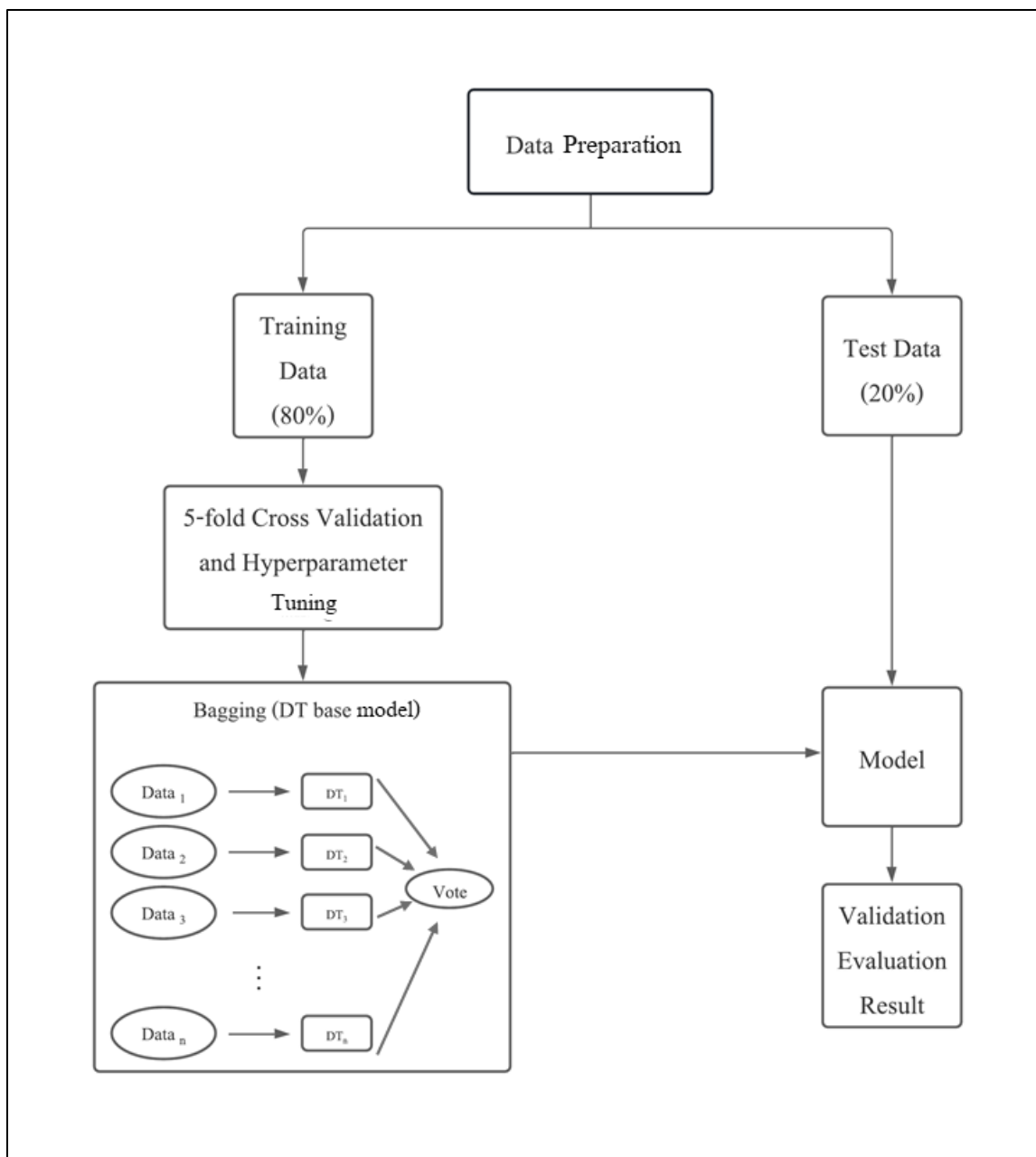
ในขั้นตอนวิธีการสร้าง Multiple Classification โดยใช้อัลกอริทึม Decision Tree (DT) และ Support Vector Machine (SVM) เป็นอัลกอริทึมพื้นฐานที่ได้กำหนดพารามิเตอร์ที่เหมาะสมไว้แล้ว ในขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยวใช้งานร่วมกับวิธีการเรียนรู้แบบรวมกลุ่ม 2 วิธี คือ Bagging และ Boosting ส่วนวิธีการของ Random Forest จะใช้ต้นไม้ช่วยตัดสินใจ ซึ่งเป็นอัลกอริทึมพื้นฐานในการเรียนรู้ของตัวอัลกอริทึมเอง

3.3.2.1 การสร้างแบบจำลอง Bagging โดยใช้ Decision Tree (DT) และ Support Vector Machine (SVM) เป็นอัลกอริทึมพื้นฐาน

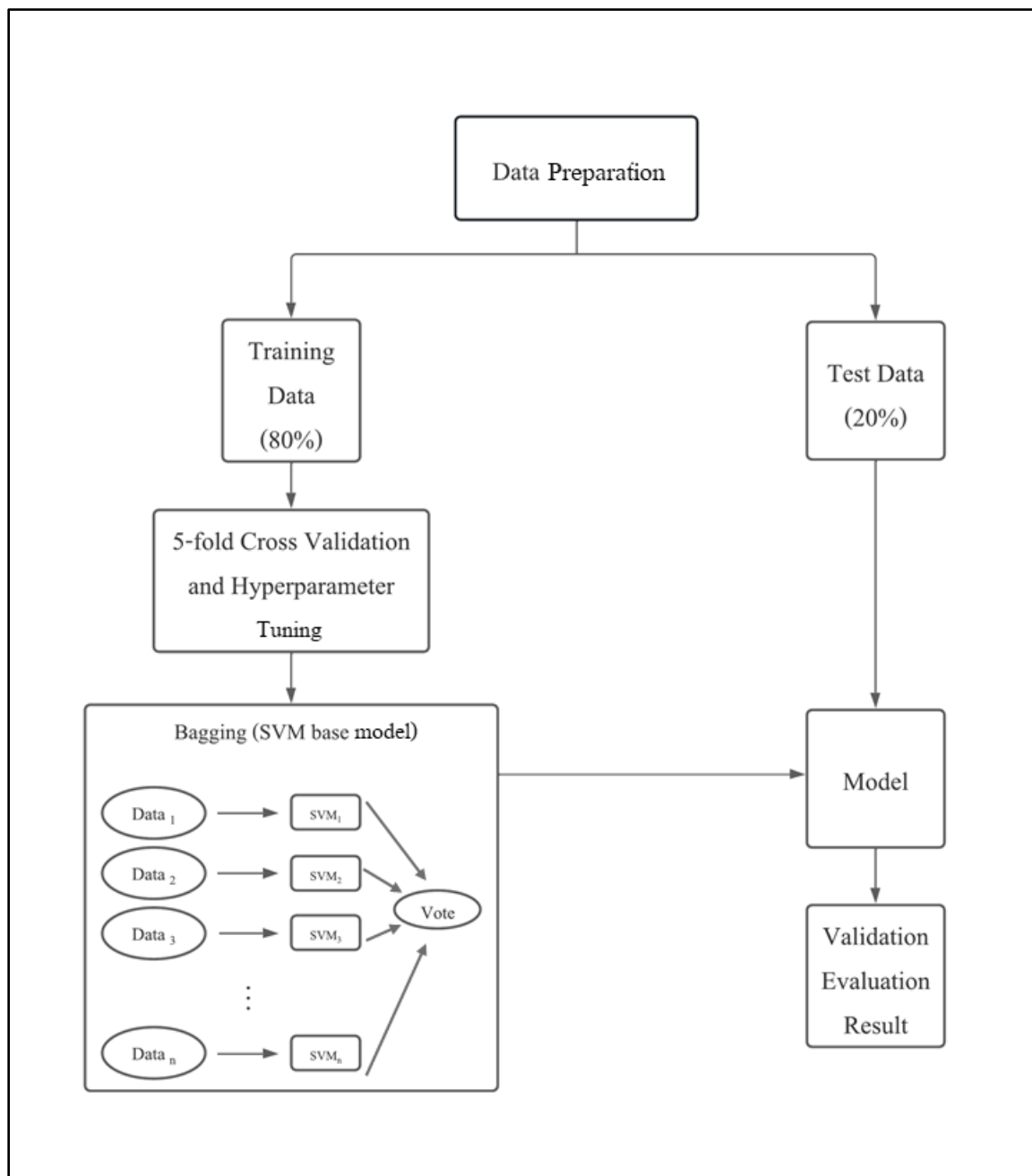
- 1) นำข้อมูลการฝึกสภาพกรณีที่ 1 2 และ 3 ที่ผ่านกระบวนการเตรียมข้อมูลมาทำการแบ่งข้อมูลออกเป็น 2 ชุด คือข้อมูลฝึกสอน (Training Data) 80% และข้อมูลทดสอบ (Test Data) 20%
- 2) นำข้อมูลฝึกสอน (Training Data) มาทำการแบ่งข้อมูลให้เป็น 5 ส่วน ข้อมูลทุกส่วนจะถูกเรียนรู้และตรวจสอบในการวัดประสิทธิภาพ ด้วยวิธี 5-fold Cross Validation และทำการหาค่าพารามิเตอร์ที่เหมาะสมโดยใช้ Grid Search สำหรับแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Bagging โดยใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ดังตารางที่ 6
- 3) กำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Bagging โดยใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ หลังจากนั้นทำการเรียนรู้โดยใช้ข้อมูลทดสอบ (Test Data)
- 4) ผลลัพธ์สุดท้ายของการประเมินการตรวจสอบความถูกต้องของกระบวนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Bagging โดยใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้และทำการเรียนรู้ จะได้ค่าวัดประสิทธิภาพของโมเดล คือ Accuracy, Precision, F1-Measure, Recall, และ AUC ดังภาพที่ 12 – 13

ตารางที่ 6 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Bagging โดยใช้ Decision Tree (DT) และ Support Vector Machine (SVM) เป็นอัลกอริทึมพื้นฐาน

พารามิเตอร์	ค่าพารามิเตอร์
n_estimators	10, 50, 100, 500
learning_rate	0.0001, 0.001, 0.01, 0.1, 1.0



ภาพที่ 12 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Bagging โดยใช้ DT เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้



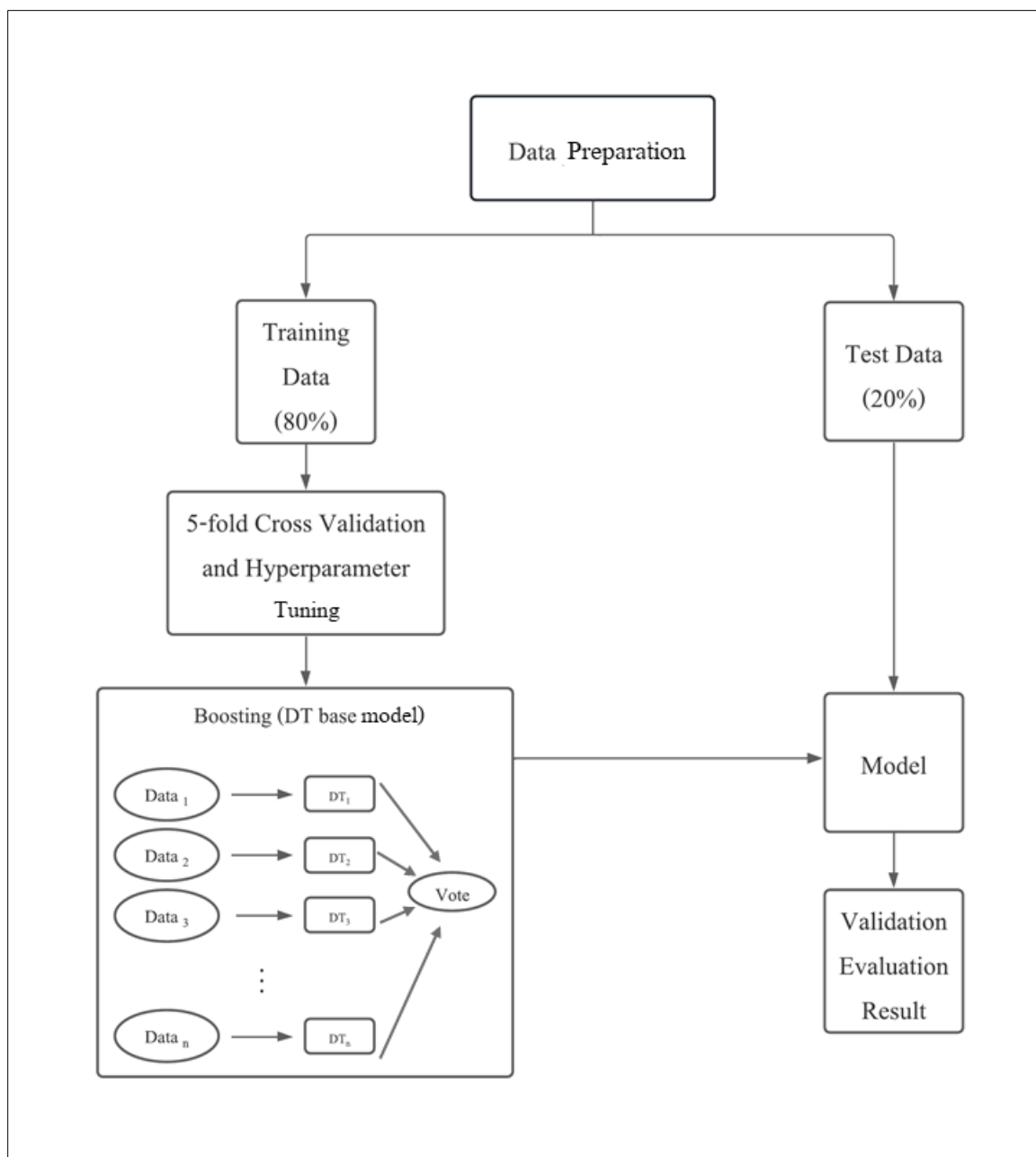
ภาพที่ 13 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Bagging โดยใช้ SVM เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้

3.3.2.2 การสร้างแบบจำลอง Boosting โดยใช้ Decision Tree (DT) และ Support Vector Machine (SVM) เป็นอัลกอริทึมพื้นฐาน

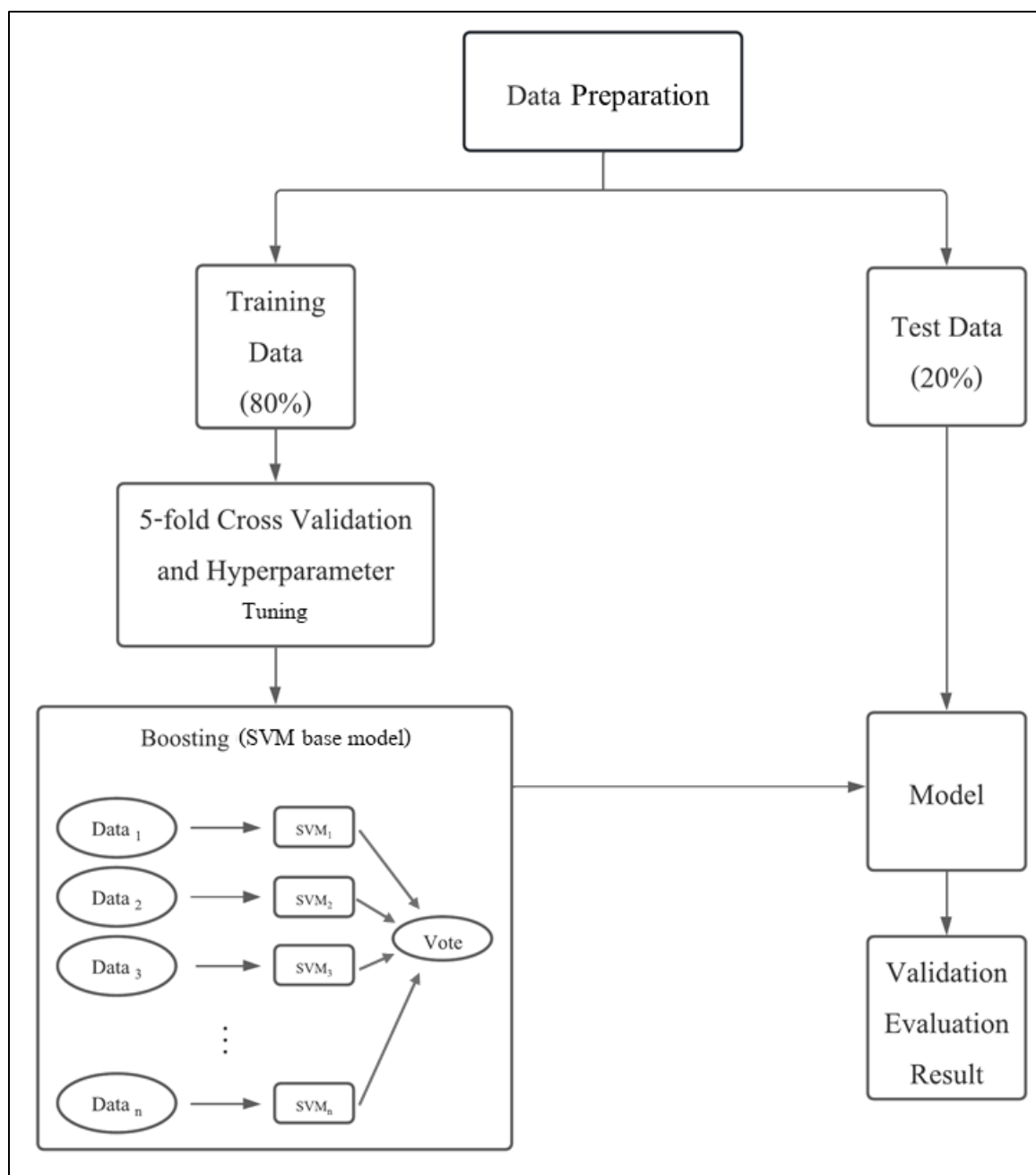
- 1) นำข้อมูลการฟื้นฟูสภาพกรณี 1 2 และ 3 ที่ผ่านกระบวนการเตรียมข้อมูลมาทำการแบ่งข้อมูลออกเป็น 2 ชุด คือข้อมูลฝึกสอน (Training Data) 80% และข้อมูลทดสอบ (Test Data) 20%
- 2) นำข้อมูลฝึกสอน (Training Data) มาทำการแบ่งข้อมูลให้เป็น 5 ส่วน ข้อมูลทุกส่วนจะถูกเรียนรู้และตรวจสอบในการวัดประสิทธิภาพ ด้วยวิธี 5-fold Cross Validation และทำการหาค่าพารามิเตอร์ที่เหมาะสม โดยใช้ Grid Search สำหรับแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Boosting โดยใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ดังตารางที่ 7
- 3) กำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Boosting โดยใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ หลังจากนั้นทำการเรียนรู้โดยใช้ข้อมูลทดสอบ (Test Data)
- 4) ผลลัพธ์สุดท้ายของการประเมินการตรวจสอบความถูกต้องของกระบวนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Boosting โดยใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้และทำการเรียนรู้ จะได้ค่าวัดประสิทธิภาพของโมเดล คือ Accuracy, Precision, F1-Measure, Recall, และ AUC ดังภาพที่ 14 - 15

ตารางที่ 7 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Boosting โดยใช้ Decision Tree (DT) และ Support Vector Machine (SVM) เป็นอัลกอริทึมพื้นฐาน

พารามิเตอร์	ค่าพารามิเตอร์
n_estimators	10, 50, 100, 500
learning_rate	0.0001, 0.001, 0.01, 0.1, 1.0



ภาพที่ 14 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Boosting โดยใช้ DT เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้



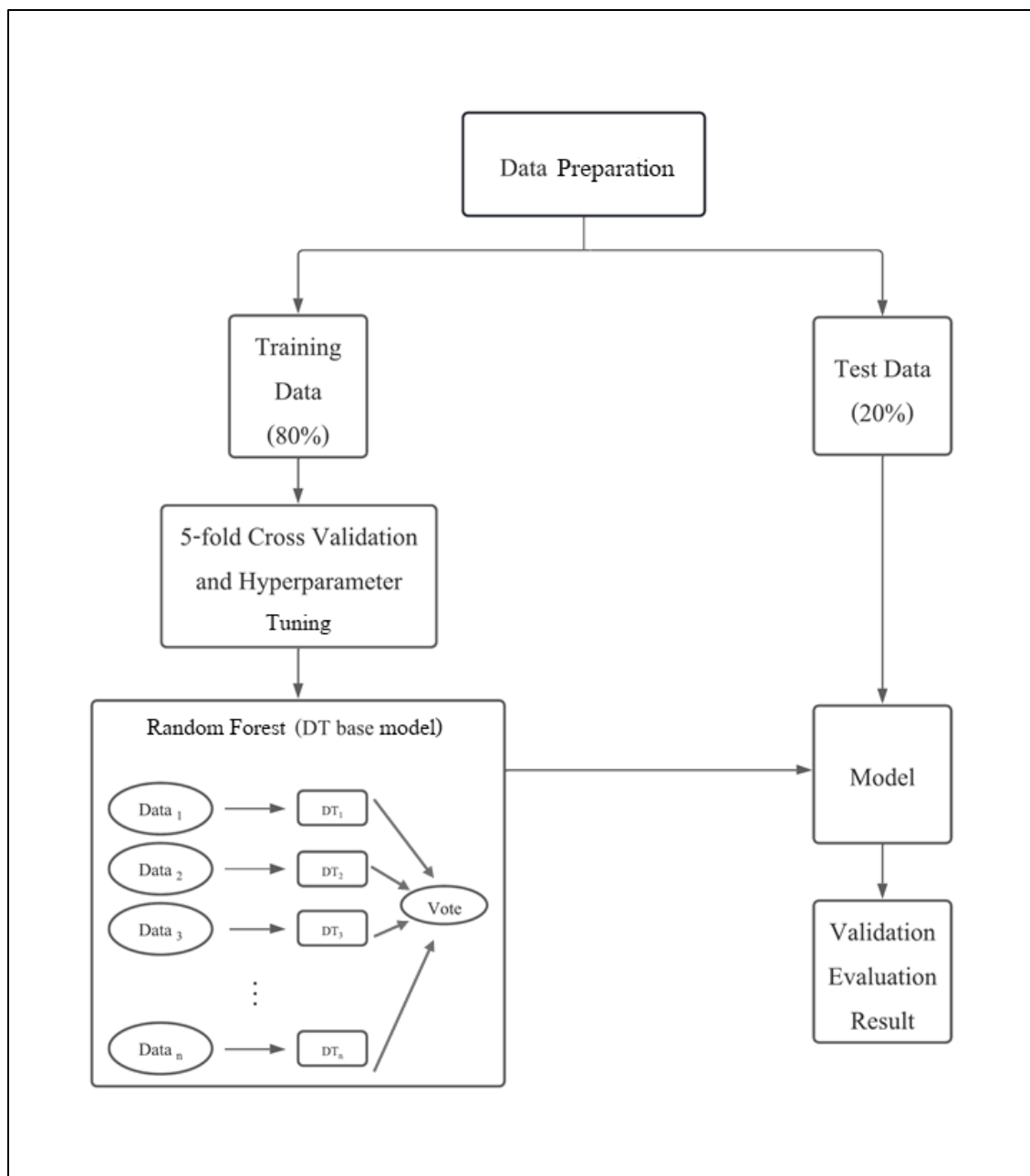
ภาพที่ 15 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

3.3.2.3 การสร้างแบบจำลอง Random Forest

- 1) นำข้อมูลการฟื้นฟูสภาพกรณี 1 2 และ 3 ที่ผ่านกระบวนการเตรียมข้อมูลมาทำการแบ่งข้อมูลออกเป็น 2 ชุด คือ ข้อมูลฝึกสอน (Training Data) 80% และข้อมูลทดสอบ (Test Data) 20%
- 2) นำข้อมูลฝึกสอน (Training Data) มาทำการแบ่งข้อมูลให้เป็น 5 ส่วน ข้อมูลทุกส่วนจะถูกเรียนรู้และตรวจสอบในการวัดประสิทธิภาพ ด้วยวิธี 5-fold Cross Validation และทำการหาค่าพารามิเตอร์ที่เหมาะสมโดยใช้ Grid Search สำหรับแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Random Forest ดังตารางที่ 8
- 3) กำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Random Forest หลังจากนั้นทำการเรียนรู้โดยใช้ข้อมูลทดสอบ (Test Data)
- 4) ผลลัพธ์สุดท้ายของการประเมินการตรวจสอบความถูกต้องของกระบวนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Random Forest จะได้ค่าวัดประสิทธิภาพของโมเดล คือ Accuracy, Precision, F1-Measure, Recall, และ AUC ดังภาพที่ 16

ตารางที่ 8 กำหนดพารามิเตอร์ให้กับอัลกอริทึม Random Forest

พารามิเตอร์	ค่าพารามิเตอร์
Max_Depth	3, 5, 10, 20, 30
Criterion	Gini, Entropy
Max_Features	Auto, Sqrt, Log2
Min_Samples_Split	2, 5, 10



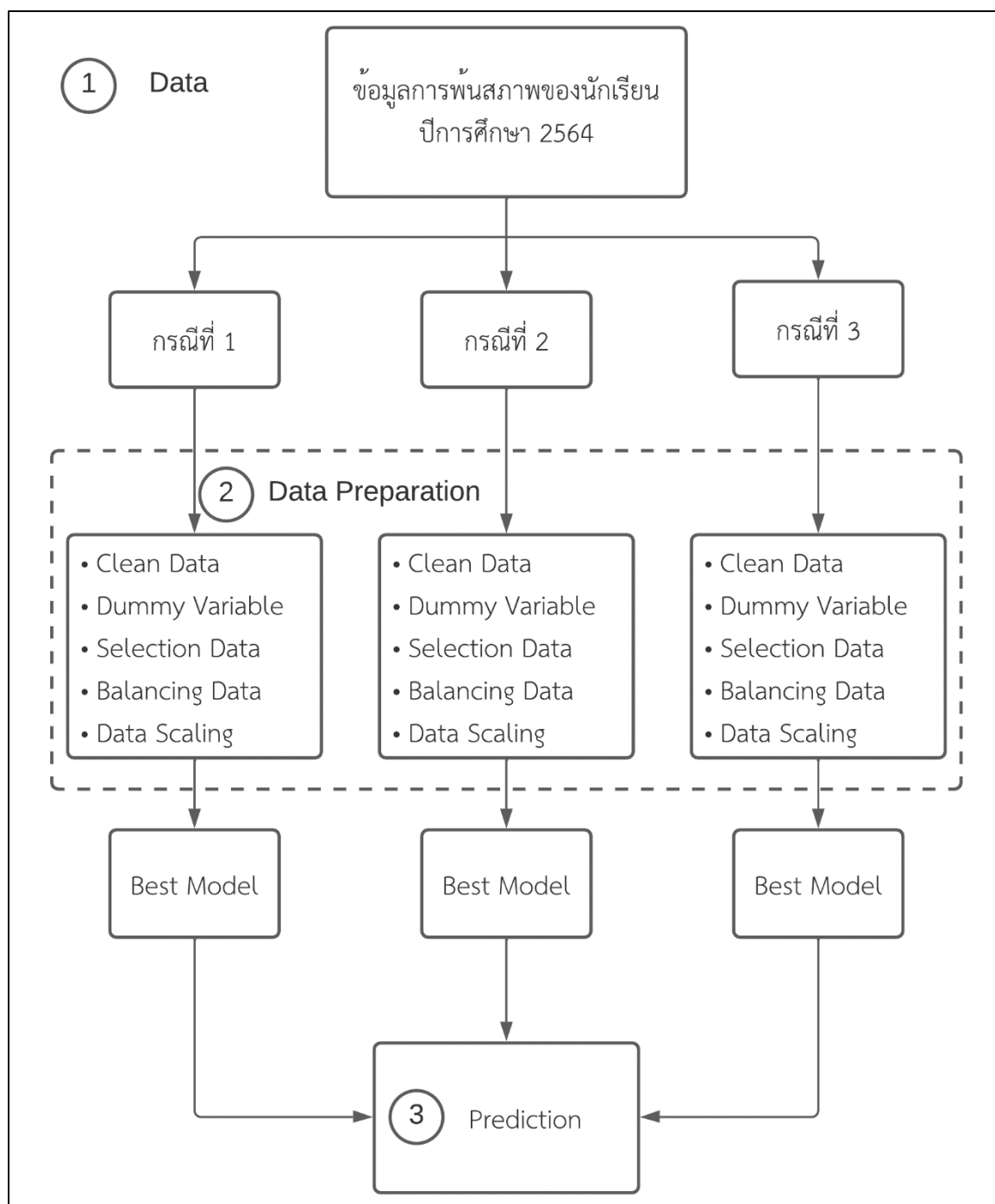
ภาพที่ 16 ขั้นตอนการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มด้วยวิธี Random Forest

3.3.3 การเปรียบเทียบผลลัพธ์ของประสิทธิภาพความถูกต้องในการจำแนกประเภท

ในส่วนนี้จะนำผลลัพธ์จากการวัดประสิทธิภาพของโมเดล คือ Accuracy, Precision, F1-Measure, Recall, และ AUC ที่ได้จากการทดลองทั้งหมดนำมาทำการเปรียบเทียบ โดยจะพิจารณาเปรียบเทียบค่า F1-Measure และ AUC แต่ละโมเดลเพื่อที่จะเลือกโมเดลที่ดีที่สุดในการจำแนกประเภทสำหรับการฟื้นฟูสภาพของนักศึกษา

3.3.4 การทำนายการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2564

หลังจากผ่านกระบวนการเปรียบเทียบผลลัพธ์ของประสิทธิภาพความถูกต้องในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาทั้ง 3 กรณี จากนั้นนำข้อมูลการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2564 ซึ่งทำการแบ่งข้อมูลการฟื้นฟูสภาพของนักศึกษาเป็น 3 กรณี โดยได้ผ่านกระบวนการจัดเตรียมข้อมูล หลังจากนั้นให้โมเดลที่ดีที่สุดสำหรับการการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2558 – 2563 ทั้ง 3 กรณีมาทำนายการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2564 ทั้ง 3 กรณี แสดงดังภาพที่ 17



ภาพที่ 17 ขั้นตอนการทำนายการพื้นฐานของนักศึกษาปีการศึกษา 2564

บทที่ 4

ผลการวิจัย

ในบทนี้ผู้วิจัยได้นำเสนอผลลัพธ์ที่ได้จากกระบวนการจัดเตรียมข้อมูล การสำรวจข้อมูลเบื้องต้น และการทดลองสำหรับการวัดประสิทธิภาพในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา รวมไปถึงแสดงการปรับค่าพารามิเตอร์ให้เหมาะสมในอัลกอริทึมแบบเดี่ยวและแบบรวมกลุ่มที่ได้กล่าวมาข้างต้น ซึ่งใช้ข้อมูลของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้ข้อมูลเฉพาะนักศึกษาปี 1 ตั้งแต่ปีการศึกษา 2558 – 2563 จำนวน 614 คน โดยในการทดลองได้ทำการเปรียบเทียบระหว่างการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยว และการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม เพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาที่ดีที่สุด และใช้ค่าจากตาราง Confusion Matrix ซึ่งเป็นค่าสำหรับใช้วัดประสิทธิภาพความถูกต้องสำหรับการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาทั้งหมดสามารถนำมาคำนวณค่า Accuracy, Precision, F1-Measure, Recall, และ AUC เพื่อวัดและเปรียบเทียบประสิทธิภาพการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา จากนั้นนำข้อมูลการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2564 มาทำนายการฟื้นฟูสภาพของนักศึกษา โดยในงานวิจัยได้ทำการแบ่งผลลัพธ์ของการทดลองออกเป็น 5 ส่วนดังต่อไปนี้

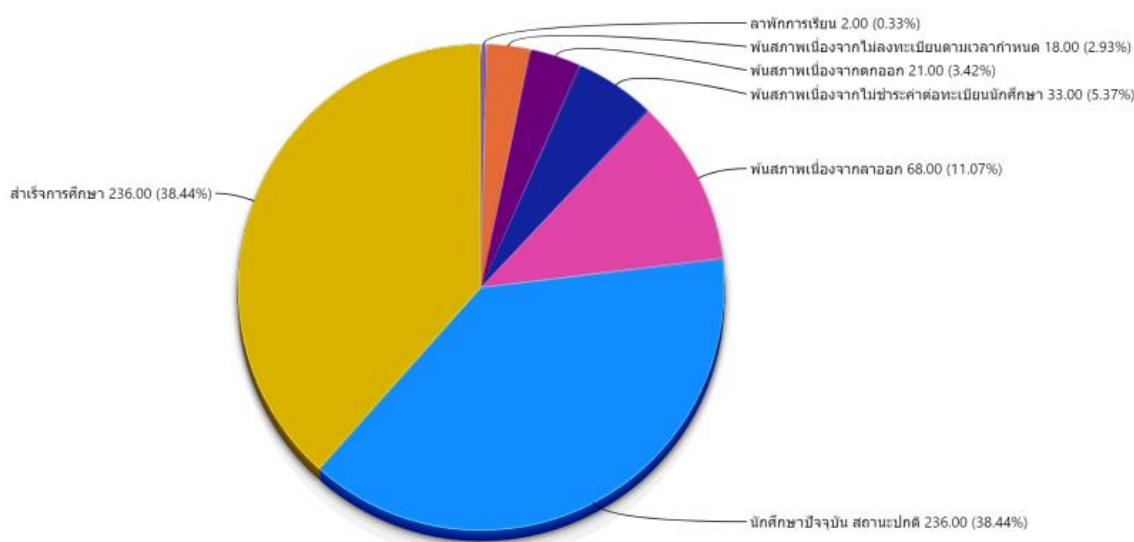
- 4.1 การสำรวจข้อมูลเบื้องต้นและกระบวนการจัดเตรียมข้อมูล
- 4.2 ผลลัพธ์อัลกอริทึมสำหรับการจำแนกประเภทแบบเดี่ยว
 - 4.2.1 อัลกอริทึม DT
 - 4.2.2 อัลกอริทึม SVM
- 4.3 ผลลัพธ์อัลกอริทึมสำหรับการจำแนกประเภทแบบรวมกลุ่ม
 - 4.3.1 วิธีการแบบ Bagging
 - 4.3.2 วิธีการแบบ Boosting
 - 4.3.3 วิธีการแบบ Random Forest
- 4.4 การเปรียบเทียบผลลัพธ์ของประสิทธิภาพความถูกต้องในการจำแนกประเภท
- 4.5 การทำนายการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2564

4.1 กระบวนการจัดเตรียมข้อมูลและการสำรวจข้อมูลเบื้องต้น

ผลลัพธ์จากการนำข้อมูลของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้ข้อมูลเฉพาะนักศึกษาปี 1 ตั้งแต่ปีการศึกษา 2558 – 2563 มีจำนวน 614 คน นำมาแสดงผลลัพธ์ ได้แก่กระบวนการจัดเตรียมข้อมูล และการสำรวจข้อมูลเบื้องต้น

4.1.1 ผลลัพธ์จากการสำรวจข้อมูลเบื้องต้น

ผลลัพธ์จากการสำรวจข้อมูลเบื้องต้นทำการ Data Visualization ข้อมูลของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ ตั้งแต่ปีการศึกษา 2558 – 2563 มีจำนวน 614 คน



ภาพที่ 18 แผนภูมิวงกลมสถานะนักศึกษา

จากภาพที่ 18 แสดงประเภทการพ้นสภาพของนักศึกษาแบ่งออกเป็น พ้นสภาพเนื่องจากลาออก พ้นสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา พ้นสภาพเนื่องจากตกลูก พ้นสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด ลาพักการเรียน พบว่า นักศึกษาที่สำเร็จการศึกษาอยู่ที่ร้อยละ 38.44 นักศึกษาในปัจจุบันที่มีสถานะปกติมีสัดส่วนอยู่ที่ร้อยละ 38.44 พ้นสภาพเนื่องจากลาออกอยู่ที่ร้อยละ 11.07 พ้นสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษาอยู่ที่ร้อยละ 5.37 พ้นสภาพเนื่องจากตกลูกอยู่ที่ร้อยละ 3.42 พ้นสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนดอยู่ที่ร้อยละ 2.93 ลาพักการเรียนมีสัดส่วนอยู่ที่ร้อยละ 0.33 จากจำนวนทั้งหมด 614 คน

4.1.2 ผลลัพธ์จากกระบวนการจัดเตรียมข้อมูล

ผลลัพธ์จากกระบวนการจัดเตรียมข้อมูลโดยใช้ข้อมูลนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ ตั้งแต่ปีการศึกษา 2558 – 2563 จำนวน 614 คน และปีการศึกษา 2564 จำนวน 105 คน ซึ่งได้ทำการแบ่งชุดข้อมูลการฟื้นฟูสภาพของนักศึกษาที่ใช้สำหรับการทดลองออกเป็น 3 กรณี หลังจากนั้นนำข้อมูลแต่ละกรณีมาผ่านกระบวนการทำความสะอาดข้อมูล (Clean Data) ทำการคัดเลือกข้อมูล (Selection Data) ปรับสมดุลข้อมูล (Balancing Data) ปรับปรุงขอบเขตข้อมูล (Data Scaling) และแปลงข้อมูลให้เป็นตัวแปรหุ่น (Dummy) ดังนี้

1) ข้อมูลการฟื้นฟูสภาพกรณีที่ 1

สถานการณ์การฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากลาออก ฟื้นฟูสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา ฟื้นฟูสภาพเนื่องจากตกออก และฟื้นฟูสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด โดยมีข้อมูลของนักศึกษาปีการศึกษา 2558 – 2563 มีจำนวน 586 รายการ 39 คุณลักษณะ และมีข้อมูลของนักศึกษาปีการศึกษา 2564 มีจำนวน 103 รายการ 38 คุณลักษณะ แสดงในตาราง 9

ตารางที่ 9 ชื่อตัวแปรและความหมายของข้อมูลการฟื้นฟูสภาพกรณีที่ 1

ลำดับ	ตัวแปร	ความหมาย
1	STATUSTEXT	สถานการณ์การฟื้นฟูสภาพ
2	GPA	เกรดเฉลี่ย
3	ENTRYGPA	เกรดเฉลี่ยระดับมัธยมศึกษา
4	STUDENTSEX	เพศ
5	FATHERINCOME	รายได้บิดา
6	YEARFATHER	อายุบิดา
7	MOTHERINCOME	รายได้มารดา
8	YEARMOTHER	อายุมารดา
9	NumberOfSon	บุตรคนที่
10	NumberOfSiblings	จำนวนพี่น้อง
11	BASICS OF STAGE ACTING_F	เกรด F รายวิชา BASICS OF STAGE ACTING
12	BIOLOGY FOR PHYSICAL SCIENCE LABORATORY_F	เกรด F รายวิชา BIOLOGY FOR PHYSICAL SCIENCE LABORATORY

ลำดับ	ตัวแปร	ความหมาย
13	BIOLOGY FOR PHYSICAL SCIENCE_F	เกรด F รายวิชา BIOLOGY FOR PHYSICAL SCIENCE
14	CALCULUS FOR PHYSICAL SCIENCE II_F	เกรด F รายวิชา CALCULUS FOR PHYSICAL SCIENCE II
15	CALCULUS FOR PHYSICAL SCIENCE I_F	เกรด F รายวิชา CALCULUS FOR PHYSICAL SCIENCE I
16	CALCULUS_F	เกรด F รายวิชา CALCULUS
17	COMPUTER PROGRAMMING I_F	เกรด F รายวิชา COMPUTER PROGRAMMING I
18	ELEMENTARY PHYSICS_F	เกรด F รายวิชา ELEMENTARY PHYSICS
19	ELEMENTARY TO BUSINESS AND ENTREPRENEURSHIP_F	เกรด F รายวิชา ELEMENTARY TO BUSINESS AND ENTREPRENEURSHIP
20	ENERGY AND ENVIRONMENT_F	เกรด F รายวิชา ENERGY AND ENVIRONMENT
21	ENGLISH FOR ACADEMIC PURPOSE I (EAP I) F	เกรด F รายวิชา ENGLISH FOR ACADEMIC PURPOSE I (EAP I)
22	ENGLISH FOR ACADEMIC PURPOSES I_F	เกรด F รายวิชา ENGLISH FOR ACADEMIC PURPOSES I
23	ENGLISH FOR COMMUNICATION_F	เกรด F รายวิชา ENGLISH FOR COMMUNICATION
24	ENGLISH II_F	เกรด F รายวิชา ENGLISH II
25	ENGLISH I_F	เกรด F รายวิชา ENGLISH I
26	FUNDAMENTAL CHEMISTRY_F	เกรด F รายวิชา FUNDAMENTAL CHEMISTRY
27	GENERAL CHEMISTRY LABORATORY_F	เกรด F รายวิชา GENERAL CHEMISTRY LABORATORY
28	GENERAL MATHEMATICS_F	เกรด F รายวิชา GENERAL MATHEMATICS
29	GENERAL PHYSICS LABORATORY I_F	เกรด F รายวิชา GENERAL PHYSICS LABORATORY I
30	HAPPINESS OF LIFE_F	เกรด F รายวิชา HAPPINESS OF LIFE

ลำดับ	ตัวแปร	ความหมาย
31	INFORMATION LITERACY SKILLS_F	เกรด F รายวิชา INFORMATION LITERACY SKILLS
32	INTRODUCTION TO INFORMATION AND COMMUNICATION	เกรด F รายวิชา INTRODUCTION TO INFORMATION AND COMMUNICATION
33	LEARNING SKILLS_F	เกรด F รายวิชา LEARNING SKILLS
34	LOCAL WISDOM_F	เกรด F รายวิชา LOCAL WISDOM
35	MEDITATION FOR LIFE DEVELOPMENT_F	เกรด F รายวิชา MEDITATION FOR LIFE DEVELOPMENT
36	MULTICULTURALISM_F	เกรด F รายวิชา MULTICULTURALISM
37	OPERATIONS RESEARCH_F	เกรด F รายวิชา OPERATIONS RESEARCH
38	STATISTICAL ANALYSIS I_F	เกรด F รายวิชา STATISTICAL ANALYSIS I
39	STATISTICAL MODEL_F	เกรด F รายวิชา STATISTICAL MODEL

2) ข้อมูลการฟื้นสภาพกรณีที่ 2

สถานการณ์ฟื้นสภาพของนักศึกษาประกอบไปด้วย ฟื้นสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา และฟื้นสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด โดยมีข้อมูลของนักศึกษาปีการศึกษา 2558 – 2563 มีจำนวน 500 รายการ 27 คุณลักษณะ และมีข้อมูลของนักศึกษาปีการศึกษา 2564 มีจำนวน 99 รายการ 26 คุณลักษณะ แสดงในตาราง 10

ตารางที่ 10 ชื่อตัวแปรและความหมายของข้อมูลการฟื้นสภาพกรณีที่ 2

อันดับ	ตัวแปร	ความหมาย
1	STATUSTEXT	สถานการณ์ฟื้นสภาพ
2	GPA	เกรดเฉลี่ย
3	ENTRYGPA	เกรดเฉลี่ยระดับมัธยมศึกษา
4	FATHERINCOME	รายได้บิดา
5	YEARFATHER	อายุบิดา

อันดับ	ตัวแปร	ความหมาย
6	MOTHERINCOME	รายได้มารดา
7	YEARMOTHER	อายุมารดา
8	NumberOfSon	บุตรคนที่
9	NumberOfSiblings	จำนวนพี่น้อง
10	BIOLOGY FOR PHYSICAL SCIENCE LABORATORY_F	เกรด F รายวิชา BIOLOGY FOR PHYSICAL SCIENCE LABORATORY
11	BIOLOGY FOR PHYSICAL SCIENCE_F	เกรด F รายวิชา BIOLOGY FOR PHYSICAL SCIENCE
12	CALCULUS FOR PHYSICAL SCIENCE I_F	เกรด F รายวิชา CALCULUS FOR PHYSICAL SCIENCE I
13	CALCULUS_F	เกรด F รายวิชา CALCULUS
14	ELEMENTARY PHYSICS_F	เกรด F รายวิชา ELEMENTARY PHYSICS
15	ENGLISH FOR SCIENCES_F	เกรด F รายวิชา ENGLISH FOR SCIENCES
16	ENGLISH FOR ACADEMIC PURPOSES I_F	เกรด F รายวิชา ENGLISH FOR ACADEMIC PURPOSES I
17	ENGLISH FOR COMMUNICATIONS_F	เกรด F รายวิชา ENGLISH FOR COMMUNICATIONS
18	ENGLISH FOR COMMUNICATION_F	เกรด F รายวิชา ENGLISH FOR COMMUNICATION
19	ENGLISH I_F	เกรด F รายวิชา ENGLISH I
20	GENERAL MATHEMATICS_F	เกรด F รายวิชา GENERAL MATHEMATICS
21	GENERAL PHYSICS LABORATORY I_F	เกรด F รายวิชา GENERAL PHYSICS LABORATORY I
22	HAPPINESS OF LIFE_F	เกรด F รายวิชา HAPPINESS OF LIFE
23	INTRODUCTION TO INFORMATION AND COMMUNICATION TECHNOLOGY_F	เกรด F รายวิชา INTRODUCTION TO INFORMATION AND COMMUNICATION TECHNOLOGY

อันดับ	ตัวแปร	ความหมาย
24	MULTICULTURALISM_F	เกรด F รายวิชา MULTICULTURALISM
25	OPERATIONS RESEARCH_F	เกรด F รายวิชา OPERATIONS RESEARCH
26	STATISTICAL ANALYSIS I_F	เกรด F รายวิชา STATISTICAL ANALYSIS I
27	STATISTICAL INFORMATION PROJECT II_F	เกรด F รายวิชา STATISTICAL INFORMATION PROJECT II

3) ข้อมูลการฟื้นสภาพกรณีที่ 3

สถานะการฟื้นสภาพของนักศึกษาประกอบไปด้วย ฟื้นสภาพเนื่องจากลาออก และฟื้นสภาพเนื่องจาก ตกออก โดยมีข้อมูลของนักศึกษาปีการศึกษา 2558 – 2563 มีจำนวน 555 รายการ 49 คุณลักษณะ และมีข้อมูล ของนักศึกษาปีการศึกษา 2564 มีจำนวน 103 รายการ 48 คุณลักษณะ แสดงในตาราง 11

ตารางที่ 11 ชื่อตัวแปรและความหมายของข้อมูลการฟื้นสภาพกรณีที่ 3

อันดับ	ตัวแปร	ความหมาย
1	STATUSTEXT	สถานะการฟื้นสภาพ
2	GPA	เกรดเฉลี่ย
3	ENTRYGPA	เกรดเฉลี่ยระดับมัธยมศึกษา
4	STUDENTSEX	เพศ
5	FATHERINCOME	รายได้บิดา
6	YEARFATHER	อายุบิดา
7	MOTHERINCOME	รายได้มารดา
8	YEARMOTHER	อายุมารดา
9	NumberOfSon	บุตรคนที่
10	NumberOfSiblings	จำนวนพี่น้อง
11	BASICS OF STAGE ACTING_F	เกรด F รายวิชา BASICS OF STAGE ACTING
12	BIOLOGY FOR PHYSICAL SCIENCE LABORATORY_F	เกรด F รายวิชา BIOLOGY FOR PHYSICAL SCIENCE LABORATORY
13	BIOLOGY FOR PHYSICAL SCIENCE_F	เกรด F รายวิชา BIOLOGY FOR PHYSICAL SCIENCE

อันดับ	ตัวแปร	ความหมาย
14	CALCULUS FOR PHYSICAL SCIENCE II_F	เกรด F รายวิชา CALCULUS FOR PHYSICAL SCIENCE II
15	CALCULUS FOR PHYSICAL SCIENCE II_F	เกรด F รายวิชา CALCULUS FOR PHYSICAL SCIENCE II
16	CALCULUS FOR PHYSICAL SCIENCE I_F	เกรด F รายวิชา CALCULUS FOR PHYSICAL SCIENCE I
17	CALCULUS_F	เกรด F รายวิชา CALCULUS
18	COMPUTER PROGRAMMING I_F	เกรด F รายวิชา COMPUTER PROGRAMMING I
19	CREATIVE THINKING AND PROBLEM SOLVING_F	เกรด F รายวิชา CREATIVE THINKING AND PROBLEM SOLVING
20	DATABASE SYSTEMS AND DESIGN LABORATORY_F	เกรด F รายวิชา DATABASE SYSTEMS AND DESIGN LABORATORY
21	DATABASE SYSTEMS AND DESIGN_F	เกรด F รายวิชา DATABASE SYSTEMS AND DESIGN
22	DISCRETE MATHEMATICS AND APPLICATIONS_F	เกรด F รายวิชา DISCRETE MATHEMATICS AND APPLICATIONS
23	ELEMENTARY PHYSICS_F	เกรด F รายวิชา ELEMENTARY PHYSICS
24	ELEMENTARY TO BUSINESS AND ENTREPRENEURSHIP_F	เกรด F รายวิชา ELEMENTARY TO BUSINESS AND ENTREPRENEURSHIP
25	ENERGY AND ENVIRONMENT_F	เกรด F รายวิชา ENERGY AND ENVIRONMENT
26	ENGLISH FOR ACADEMIC PURPOSE I (EAP I)_F	เกรด F รายวิชา ENGLISH FOR ACADEMIC PURPOSE I (EAP I)
27	ENGLISH FOR ACADEMIC PURPOSES I_F	เกรด F รายวิชา ENGLISH FOR ACADEMIC PURPOSES I
28	ENGLISH FOR COMMUNICATION_F	เกรด F รายวิชา ENGLISH FOR COMMUNICATION

อันดับ	ตัวแปร	ความหมาย
29	ENGLISH II_F	เกรด F รายวิชา ENGLISH II
30	ENGLISH IV_F	เกรด F รายวิชา ENGLISH IV
31	ENGLISH I_F	เกรด F รายวิชา ENGLISH I
32	FINACIAL AND ACCOUNTING MANAGEMENT FOR EXECUTIVE_F	เกรด F รายวิชา FINACIAL AND ACCOUNTING MANAGEMENT FOR EXECUTIVE
33	FUNDAMENTAL CHEMISTRY_F	FUNDAMENTAL CHEMISTRY
34	GENERAL CHEMISTRY LABORATORY_F	เกรด F รายวิชา GENERAL CHEMISTRY LABORATORY
35	GENERAL MATHEMATICS_F	เกรด F รายวิชา GENERAL MATHEMATICS
36	GENERAL PHYSICS LABORATORY I_F	เกรด F รายวิชา GENERAL PHYSICS LABORATORY I
37	INFORMATION LITERACY SKILLS_F	เกรด F รายวิชา INFORMATION LITERACY SKILLS
38	INTRODUCTION TO INFORMATION AND COMMUNICATION TECHNOLOGY_F	เกรด F รายวิชา INTRODUCTION TO INFORMATION AND COMMUNICATION TECHNOLOGY
39	LEADERSHIP AND MANAGEMENT_F	เกรด F รายวิชา LEADERSHIP AND MANAGEMENT
40	LEARNING SKILLS_F	เกรด F รายวิชา LEARNING SKILLS
41	LOCAL WISDOM_F	เกรด F รายวิชา LOCAL WISDOM
42	MEDITATION FOR LIFE DEVELOPMENT_F	เกรด F รายวิชา MEDITATION FOR LIFE DEVELOPMENT
43	MULTICULTURALISM_F	เกรด F รายวิชา MULTICULTURALISM
44	OPERATIONS RESEARCH_F	เกรด F รายวิชา OPERATIONS RESEARCH
45	PHYSICAL EDUCATION ACTIVITY (AEROBIC DANCE)_F	เกรด F รายวิชา PHYSICAL EDUCATION ACTIVITY (AEROBIC DANCE)

อันดับ	ตัวแปร	ความหมาย
46	RISK AND INSURANCE_F	เกรด F รายวิชา RISK AND INSURANCE
47	SMALL AND MEDIUM ENTERPRISES MANAGEMENT_F	เกรด F รายวิชา SMALL AND MEDIUM ENTERPRISES MANAGEMENT
48	STATISTICAL ANALYSIS I_F	เกรด F รายวิชา STATISTICAL ANALYSIS I
49	STATISTICAL MODEL_F	เกรด F รายวิชา STATISTICAL MODEL

4.2 ผลลัพธ์อัลกอริทึมสำหรับการจำแนกประเภทแบบเดี่ยว

ผลลัพธ์จากการวิเคราะห์ข้อมูลสำหรับการจำแนกประเภทการผันสภาพของนักศึกษา ในการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยว โดยได้ทำการเปรียบเทียบ 2 อัลกอริทึม ได้แก่ DT และ SVM

4.2.1 ผลลัพธ์จากการวิเคราะห์ข้อมูลโดยใช้อัลกอริทึม DT

ผลลัพธ์การวิเคราะห์ข้อมูลในการจำแนกประเภทการผันสภาพของนักศึกษาโดยใช้อัลกอริทึม DT ในการทดลอง โดยได้กำหนดค่าพารามิเตอร์ดังตารางที่ 12 ซึ่งเป็นค่าพารามิเตอร์ที่เหมาะสมที่จะทำให้ประสิทธิภาพความถูกต้องของการจำแนกประเภทการผันสภาพของนักศึกษามีความแม่นยำมากที่สุด

ตารางที่ 12 ค่าพารามิเตอร์ที่เหมาะสม สำหรับอัลกอริทึม DT

พารามิเตอร์	ค่าพารามิเตอร์		
	การผันสภาพ กรณีที่ 1	การผันสภาพกรณี ที่ 2	การผันสภาพ กรณีที่ 3
Criterion	Entropy	Gini	GInI
Max Depth	12	18	10
Max Features	34	8	33
Min Samples Leaf	3	2	2

ตารางที่ 13 Confusion Matrix โดยใช้อัลกอริทึม DT ของการฟื้นฟูสภาพกรณีที่ 1

Class	Prediction		
	0	1	
Actual	0	86	8
	1	5	89

ตารางที่ 14 Confusion Matrix โดยใช้อัลกอริทึม DT ของการฟื้นฟูสภาพกรณีที่ 2

Class	Prediction	
	0	1
Actual	0	92
	1	94

ตารางที่ 15 Confusion Matrix โดยใช้อัลกอริทึม DT ของการฟื้นฟูสภาพกรณีที่ 3

Class		Prediction	
		0	1
Actual	0	88	6
	1	0	94

จากตารางที่ 13-15 Confusion Matrix สามารถที่จะนำมาคำนวณค่าวัดประสิทธิภาพความถูกต้องของแบบจำลองสำหรับการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม DT ดังตารางที่ 16

ตารางที่ 16 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม DT

Performance	Single Model: DT		
	การฟื้นฟูสภาพกรณี ที่ 1	การฟื้นฟูสภาพกรณี ที่ 2	การฟื้นฟูสภาพกรณี ที่ 3
Accuracy	0.93	0.99	0.97
Precision	0.93	0.99	0.97
Recall	0.93	0.99	0.97
F1-Score	0.93	0.99	0.97
AUC	0.96	0.99	0.96

จากตารางที่ 16 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม DT ซึ่งคำนวณประสิทธิภาพความถูกต้องจากตารางที่ Confusion Matrix ที่ได้กำหนดค่าพารามิเตอร์ที่เหมาะสมของอัลกอริทึม Decision Tree ดังตารางที่ 12 พบว่า การประเมินประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) อัลกอริทึม Decision Tree สามารถจำแนกประเภทข้อมูลการฟื้นฟูสภาพกรณีที่ 2 ได้ดีที่สุด โดยมีค่าเท่ากับ 99% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 3 และข้อมูลการฟื้นฟูสภาพกรณีที่ 1 โดยมีประสิทธิภาพความถูกต้องอยู่ที่ 97%, และ 93% สำหรับค่าความถ่วงดุล (F1-Score) สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา คือข้อมูลการฟื้นฟูสภาพกรณีที่ 2 อยู่ที่ 99% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 3 และข้อมูลการฟื้นฟูสภาพกรณีที่ 1 โดยมีค่าความถ่วงดุลอยู่ที่ 97%, และ 93% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา คือ ข้อมูลการฟื้นฟูสภาพกรณีที่ 2 อยู่ที่ 99% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 3 และข้อมูลการฟื้นฟูสภาพกรณีที่ 1 โดยมีค่า AUC เท่ากันอยู่ที่ 96%

4.2.2 ผลลัพธ์จากการวิเคราะห์ข้อมูลโดยใช้อัลกอริทึม SVM

ผลลัพธ์การวิเคราะห์ข้อมูลในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาโดยใช้อัลกอริทึม SVM ในการทดลอง โดยได้กำหนดค่าพารามิเตอร์ดังตารางที่ 17 ซึ่งเป็นค่าพารามิเตอร์ที่เหมาะสมที่จะทำให้ประสิทธิภาพความถูกต้องของการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษามีความแม่นยำมากที่สุด

ตารางที่ 17 ค่าพารามิเตอร์ที่เหมาะสม สำหรับอัลกอริทึม SVM

พารามิเตอร์	ค่าพารามิเตอร์		
	การปรับสภาพกรณี ที่ 1	การปรับสภาพกรณี ที่ 2	การปรับสภาพ กรณีที่ 3
C	10	100	100
Gamma	0.1	1	1
Kernel	rbf	rbf	rbf

*rbf คือ Radial Basis Function Kernel

ตารางที่ 18 Confusion Matrix โดยใช้อัลกอริทึม SVM ของการปรับสภาพกรณีที่ 1

Class	Prediction	
	0	1
Actual	0	89
	1	5
	11	83

ตารางที่ 19 Confusion Matrix โดยใช้อัลกอริทึม SVM ของการปรับสภาพกรณีที่ 2

Class	Prediction		
	0	1	
Actual	0	87	7
	1	0	94

ตารางที่ 20 Confusion Matrix โดยใช้อัลกอริทึม SVM ของการปรับสภาพกรณีที่ 3

Class	Prediction	
	0	1
Actual	0	85
	1	9
	15	79

จากตารางที่ 18-20 Confusion Matrix สามารถที่จะนำมาคำนวณค่าวัดประสิทธิภาพความถูกต้องของแบบจำลองสำหรับการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม SVM ดังตารางที่ 21

ตารางที่ 21 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม SVM

Performance	Single Model: SVM		
	การฟื้นฟูสภาพ กรณีที่ 1	การฟื้นฟูสภาพ กรณีที่ 2	การฟื้นฟูสภาพ กรณีที่ 3
Accuracy	0.91	0.96	0.87
Precision	0.91	0.96	0.87
Recall	0.91	0.96	0.87
F1-Score	0.91	0.96	0.87
AUC	0.94	0.98	0.93

จากตารางที่ 21 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม SVM ซึ่งคำนวณประสิทธิภาพความถูกต้องจากตาราง Confusion Matrix (ตารางที่ 18 - 20) กำหนดค่าพารามิเตอร์ที่เหมาะสมของอัลกอริทึม SVM ดังตารางที่ 17 พบว่า การประเมินประสิทธิภาพด้วยค่าถูกต้อง (Accuracy) อัลกอริทึม SVM สามารถจำแนกประเภทของ ข้อมูลฟื้นฟูสภาพกรณีที่ 2 ได้ดีที่สุด โดยมีค่าเท่ากับ 96% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 1 และข้อมูลการฟื้นฟูสภาพกรณีที่ 3 โดยประสิทธิภาพความถูกต้องอยู่ที่ 91% และ 87% สำหรับค่าความถ่วงดุลสูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา คือ ข้อมูลการฟื้นฟูสภาพกรณีที่ 2 อยู่ที่ 96% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 1 และข้อมูลการฟื้นฟูสภาพกรณีที่ 3 โดยมีค่าความถ่วงดุลอยู่ที่ 91% และ 87% ค่า AUC พื้นที่ใต้เส้นโค้ง ROC สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา คือ ข้อมูลการฟื้นฟูสภาพกรณีที่ 2 อยู่ที่ 98% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 1 และข้อมูลการฟื้นฟูสภาพกรณีที่ 3 โดยมีค่า AUC อยู่ที่ 94%, 93%

4.3 ผลลัพธ์อัลกอริทึมสำหรับการจำแนกประเภทแบบรวมกลุ่ม

ผลลัพธ์จากการวิเคราะห์ข้อมูลสำหรับการจำแนกประเภทการฟื้นสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม โดยได้ทำการเปรียบเทียบ 3 วิธี ได้แก่วิธี Bagging, Boosting, และ Random Forest

4.3.1 ผลลัพธ์จากการวิเคราะห์ข้อมูลด้วยวิธี Bagging

ผลการทดลองด้วยวิธี Bagging ซึ่งใช้อัลกอริทึมแบบเดี่ยวเป็นฐานในการสร้างโมเดล สำหรับการวิเคราะห์ประสิทธิภาพความถูกต้องในการจำแนกประเภทการฟื้นสภาพของนักศึกษา โดยแบ่งออกเป็น 2 รูปแบบดังต่อไปนี้

- 1) ผลลัพธ์จากการจำแนกด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

ผลลัพธ์การวิเคราะห์ข้อมูลในการจำแนกประเภทการฟื้นสภาพของนักศึกษาด้วยใช้วิธีการ Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ในการทดลอง โดยได้กำหนดค่าพารามิเตอร์ดังตารางที่ 22 ซึ่งเป็นค่าพารามิเตอร์ที่เหมาะสมที่จะทำให้ประสิทธิภาพความถูกต้องของการจำแนกประเภทการฟื้นสภาพของนักศึกษามีความแม่นยำมากที่สุดทั้ง 3 กลุ่มตัวอย่าง

ตารางที่ 22 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

พารามิเตอร์	ค่าพารามิเตอร์		
	การฟื้นสภาพ กรณีที่ 1	การฟื้นสภาพ กรณีที่ 2	การฟื้นสภาพ กรณีที่ 3
n estimators	500	100	10
Learning Rate	0.1	0.0001	1.0

ตารางที่ 23 Confusion Matrix ด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการฟื้นสภาพกรณีที่ 1

Class	Prediction	
	0	1
Actual	0	90 4
	1	0 94

ตารางที่ 24 Confusion Matrix ด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ
 พ้นสภาพกรณีที่ 2

Class	Prediction	
	0	1
Actual	0	93
	1	94

ตารางที่ 25 Confusion Matrix ด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ
 พ้นสภาพกรณีที่ 3

Class	Prediction		
	0	1	
Actual	0	90	4
	1	2	92

จากตารางที่ 23 - 25 Confusion Matrix สามารถที่จะนำมาคำนวณค่าวัดประสิทธิภาพความถูกต้อง
 ของแบบจำลองสำหรับการจำแนกประเภทการพ้นสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนกด้วย
 วิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ดังตารางที่ 9

ตารางที่ 26 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

Performance	Bagging Model (DT Base Model)		
	การฟื้นฟูสภาพ กรณีที่ 1	การฟื้นฟูสภาพ กรณีที่ 2	การฟื้นฟูสภาพ กรณีที่ 3
Accuracy	0.98	0.99	0.97
Precision	0.98	0.99	0.97
Recall	0.98	0.99	0.97
F1-Score	0.98	0.99	0.97
AUC	0.99	1.00	0.99

จากตารางที่ 26 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธี Bagging โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ซึ่งคำนวณประสิทธิภาพความถูกต้องจากตาราง Confusion Matrix (ตารางที่ 23 - 25) ที่ได้กำหนดค่าพารามิเตอร์ของที่เหมาะสมของวิธี Bagging โดยใช้ DT ดังตารางที่ 22 พบว่าข้อมูลการฟื้นฟูสภาพกรณีที่ 2 มีประสิทธิภาพค่าความถูกต้อง (Accuracy) สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาอยู่ที่ 99% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 1 และข้อมูลการฟื้นฟูสภาพกรณีที่ 3 โดยมีประสิทธิภาพความถูกต้องอยู่ที่ 98%, 97% ค่าความถ่วงดุล (F1-Score) สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา คือ ข้อมูลการฟื้นฟูสภาพกรณีที่ 2 อยู่ที่ 99% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 1 และข้อมูลการฟื้นฟูสภาพกรณีที่ 3 โดยมีค่าความถ่วงดุลอยู่ที่ 98%, 97% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาคือ ข้อมูลการฟื้นฟูสภาพกรณีที่ 2 อยู่ที่ 100% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 1 และข้อมูลการฟื้นฟูสภาพกรณีที่ 3 โดยมีค่า AUC เท่ากันอยู่ที่ 99%

2) ผลลัพธ์จากการจำแนกด้วยวิธี Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

ผลลัพธ์การวิเคราะห์ข้อมูลในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ในการทดลอง โดยได้กำหนดค่าพารามิเตอร์ดังตารางที่ 27 ซึ่งเป็นค่าพารามิเตอร์ที่เหมาะสมที่จะทำให้ประสิทธิภาพความถูกต้องของการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษามีความแม่นยำมากที่สุด

ตารางที่ 27 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

พารามิเตอร์	ค่าพารามิเตอร์		
	การผันสภาพ กรณีที่ 1	การผันสภาพ กรณีที่ 2	การผันสภาพ กรณีที่ 3
n estimators	500	100	10
Learning Rate	0.1	0.0001	1.0

ตารางที่ 28 Confusion Matrix ด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ของการผันสภาพกรณีที่ 1

Class		Prediction	
		0	1
Actual	0	84	10
	1	11	83

ตารางที่ 29 Confusion Matrix ด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ของการผันสภาพกรณีที่ 2

Class	Prediction	
	0	1
Actual	0	87
	1	94

ตารางที่ 30 Confusion Matrix ด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ของการผันสภาพกรณ์ที่ 3

Class	Prediction		
	0	1	
Actual	0	86	8
	1	13	81

จากตารางที่ 27 - 30 Confusion Matrix สามารถที่จะนำมาคำนวณค่าวัดประสิทธิภาพความถูกต้องของแบบจำลองสำหรับการจำแนกประเภทการผันสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนกประเภทด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ดังตารางที่ 31

ตารางที่ 31 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธีการ Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

Performance	Bagging Model (SVM Base Model)		
	การผันสภาพกรณ์ที่ 1	การผันสภาพกรณ์ที่ 2	การผันสภาพกรณ์ที่ 3
Accuracy	0.89	0.96	0.89
Precision	0.89	0.96	0.89
Recall	0.89	0.96	0.89
F1-Score	0.89	0.96	0.89
AUC	0.94	0.98	0.95

จากตารางที่ 31 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธี Bagging โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ซึ่งคำนวณประสิทธิภาพความถูกต้องจากตาราง Confusion Matrix (ตารางที่ 27 – 30) ที่ได้กำหนดค่าพารามิเตอร์ที่เหมาะสมของวิธี Bagging โดยใช้ SVM ดังตารางที่ 27 พบว่าข้อมูลการผันสภาพกรณ์ที่ 2 มีประสิทธิภาพค่าความถูกต้อง (Accuracy) สูงสุดในการจำแนกประเภทการผันสภาพของนักศึกษาอยู่ที่ 96% ตามด้วยข้อมูลการผันสภาพกรณ์ที่ 3 และข้อมูลการผันสภาพกรณ์ที่ 1 โดยมีประสิทธิภาพความถูกต้องเท่ากันอยู่ที่ 89% ค่าความถ่วงดุล (F1-Score) สูงสุดในการจำแนกประเภทการผัน

สภาพของนักศึกษา คือ ข้อมูลการพ้นสภาพกรณีที่ 2 อยู่ที่ 96% ตามด้วยข้อมูลการพ้นสภาพกรณีที่ 3 และข้อมูลการพ้นสภาพกรณีที่ 1 โดยมีค่าความถ่วงดุลเท่ากันอยู่ที่ 89% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC สูงสุดในการจำแนกประเภทการพ้นสภาพของนักศึกษา คือ ข้อมูลการพ้นสภาพกรณีที่ 2 อยู่ที่ 98% ตามด้วยข้อมูลการพ้นสภาพกรณีที่ 3 และข้อมูลการพ้นสภาพกรณีที่ 1 โดยมีค่าความถ่วงดุลอยู่ที่ 95%, และ 94%

4.3.2 ผลลัพธ์จากการวิเคราะห์ข้อมูลด้วยวิธี Boosting

ผลการทดลองด้วยวิธี Boosting ซึ่งใช้อัลกอริทึมแบบเดี่ยวเป็นฐานในการสร้างโมเดล สำหรับการวิเคราะห์ประสิทธิภาพความถูกต้องในการจำแนกประเภทการพ้นสภาพของนักศึกษา โดยแบ่งออกเป็น 2 รูปแบบดังต่อไปนี้

1) ผลลัพธ์จากการจำแนกด้วยวิธี Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

ผลลัพธ์การวิเคราะห์ข้อมูลในการจำแนกประเภทการพ้นสภาพของนักศึกษาด้วยวิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ในการทดลอง โดยได้กำหนดค่าพารามิเตอร์ดังตารางที่ 32 ซึ่งเป็นค่าพารามิเตอร์ที่เหมาะสมที่จะทำให้ประสิทธิภาพความถูกต้องของการจำแนกประเภทการพ้นสภาพของนักศึกษามีความแม่นยำมากที่สุด

ตารางที่ 32 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

พารามิเตอร์	ค่าพารามิเตอร์		
	การพ้นสภาพ กรณีที่ 1	การพ้นสภาพกรณี ที่ 2	การพ้นสภาพ กรณีที่ 3
n estimators	500	100	100
Learning Rate	1.0	1.0	1.0

ตารางที่ 33 Confusion Matrix วิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ
 พันสภาพกรณีที่ 1

Class	Prediction	
	0	1
Actual	0	93
	1	94

ตารางที่ 34 Confusion Matrix วิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ
 พันสภาพกรณีที่ 2

Class		Prediction	
		0	1
Actual	0	93	1
	1	0	94

ตารางที่ 35 Confusion Matrix วิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ของการ
 พันสภาพกรณีที่ 3

Class	Prediction	
	0	1
Actual	0	93
	1	94

จากตารางที่ 33 - 35 Confusion Matrix สามารถที่จะนำมาคำนวณค่าวัดประสิทธิภาพความถูกต้อง
 ของแบบจำลองสำหรับการจำแนกประเภทการพันสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนก
 ประเภทด้วยวิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ดังตารางที่ 36

ตารางที่ 36 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธีการ Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

Performance	Boosting Model (DT Base Model)		
	การฟื้นฟูสภาพ กรณีที่ 1	การฟื้นฟูสภาพ กรณีที่ 2	การฟื้นฟูสภาพกรณี ที่ 3
Accuracy	0.99	0.99	0.99
Precision	0.99	0.99	0.99
Recall	0.99	0.99	0.99
F1-Score	0.99	0.99	0.99
AUC	0.99	0.98	0.95

จากตารางที่ 36 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธี Boosting โดยใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ซึ่งคำนวณประสิทธิภาพความถูกต้องจากตาราง Confusion Matrix (ตารางที่ 33 - 35) ที่ได้กำหนดค่าพารามิเตอร์ที่เหมาะสมของวิธี Boosting โดยใช้ DT ดังตารางที่ 32 พบว่าข้อมูลการฟื้นฟูสภาพกรณีที่ 1 - 3 มีประสิทธิภาพค่าความถูกต้อง (Accuracy) ในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาเท่ากันอยู่ที่ 99% ในข้อมูลการฟื้นฟูสภาพกรณีที่ 1 - 3 มีค่าความถ่วงดุล (F1-Score) ในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาเท่ากันอยู่ที่ 99% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาคือ ข้อมูลการฟื้นฟูสภาพกรณีที่ 1 อยู่ที่ 99% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณีที่ 2 และข้อมูลการฟื้นฟูสภาพกรณีที่ 3 โดยมีค่าความถ่วงดุลอยู่ที่ 98%, และ 95%

2) ผลลัพธ์จากการจำแนกด้วยวิธี Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ผลลัพธ์การวิเคราะห์ข้อมูลในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ในการทดลอง โดยได้กำหนดค่าพารามิเตอร์ดังตารางที่ 37 ซึ่งเป็นค่าพารามิเตอร์ที่เหมาะสมที่จะทำให้ประสิทธิภาพความถูกต้องของการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษามีความแม่นยำมากที่สุด

ตารางที่ 37 ค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธี Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

พารามิเตอร์	ค่าพารามิเตอร์		
	การฟื้นฟูสภาพ กรณีที่ 1	การฟื้นฟูสภาพกรณี ที่ 2	การฟื้นฟูสภาพกรณี ที่ 3
n estimators	500	50	50
Learning Rate	0.01	1.0	1.0

ตารางที่ 38 Confusion Matrix ด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ของการฟื้นฟูสภาพกรณีที่ 1

Class	Prediction	
	0	1
Actual	0	55 39
	1	9 85

ตารางที่ 39 Confusion Matrix ด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ของการฟื้นฟูสภาพกรณีที่ 2

Class	Prediction	
	0	1
Actual	0	89
	1	9

ตารางที่ 40 Confusion Matrix ด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ของการผันสภาพกรณ์ที่ 3

Class	Prediction	
	0	1
Actual	0	81
	1	10

จากตารางที่ 38 - 40 Confusion Matrix สามารถที่จะนำมาคำนวณค่าวัดประสิทธิภาพความถูกต้องของแบบจำลองสำหรับการจำแนกประเภทการผันสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนกประเภทด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ดังตารางที่ 41

ตารางที่ 41 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธีการ Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้

Performance	Boosting Model (SVM Base Model)		
	การผันสภาพกรณ์ที่ 1	การผันสภาพกรณ์ที่ 2	การผันสภาพกรณ์ที่ 3
Accuracy	0.74	0.93	0.88
Precision	0.74	0.93	0.88
Recall	0.74	0.93	0.88
F1-Score	0.74	0.93	0.88
AUC	0.86	0.98	0.95

จากตารางที่ 41 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทด้วยวิธี Boosting โดยใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ ซึ่งคำนวณประสิทธิภาพความถูกต้องจาก Confusion Matrix (ตารางที่ 38 - 40) ที่ได้กำหนดค่าพารามิเตอร์ที่เหมาะสมของวิธี Boosting โดยใช้ SVM ดังตารางที่ 37 พบว่าข้อมูลการผันสภาพกรณ์ที่ 2 มีประสิทธิภาพค่าความถูกต้อง (Accuracy) สูงสุดในการจำแนกประเภทการผัน

สภาพของนักศึกษาอยู่ที่ 93% ตามด้วยข้อมูลการฟื้นสภาพกรณีที่ 3 และข้อมูลการฟื้นสภาพกรณีที่ 1 โดยมีประสิทธิภาพความถูกต้องอยู่ที่ 88%, 74% ค่าความถ่วงดุล (F1-Score) สูงสุดในการจำแนกประเภทการฟื้นสภาพของนักศึกษา คือ ข้อมูลการฟื้นสภาพกรณีที่ 2 อยู่ที่ 93% ตามด้วยข้อมูลการฟื้นสภาพกรณีที่ 3 และข้อมูลการฟื้นสภาพกรณีที่ 1 โดยมีค่าความถ่วงดุลอยู่ที่ 88%, 74% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC สูงสุดในการจำแนกประเภทการฟื้นสภาพของนักศึกษา คือ ข้อมูลการฟื้นสภาพกรณีที่ 2 อยู่ที่ 98% ตามด้วยข้อมูลการฟื้นสภาพกรณีที่ 3 และข้อมูลการฟื้นสภาพกรณีที่ 1 โดยมีค่า ROC อยู่ที่ 95%, 86%

4.3.3 ผลลัพธ์จากการจำแนกข้อมูลด้วยวิธี Random Forest

ผลลัพธ์การวิเคราะห์ข้อมูลในการจำแนกประเภทการฟื้นสภาพของนักศึกษาด้วยวิธี Random Forest โดยได้กำหนดค่าพารามิเตอร์ดังตารางที่ 42 ซึ่งเป็นค่าพารามิเตอร์ที่เหมาะสมที่จะทำให้ประสิทธิภาพความถูกต้องของการจำแนกประเภทการฟื้นสภาพของนักศึกษามีความแม่นยำมากที่สุด

ตารางที่ 42 ค่าพารามิเตอร์ที่เหมาะสมโดยใช้อัลกอริทึม Random Forest

พารามิเตอร์	ค่าพารามิเตอร์		
	การฟื้นสภาพ กรณีที่ 1	การฟื้นสภาพ กรณีที่ 2	การฟื้นสภาพ กรณีที่ 3
Criterion	Gini	Gini	Entropy
Max Depth	30	20	30
Max Features	Log2	Log2	auto
Min Samples Split	2	2	2

ตารางที่ 43 Confusion Matrix ด้วยวิธี Random Forest ของการฟื้นสภาพกรณีที่ 1

Class	Prediction	
	0	1
Actual	0	93
	1	94

ตารางที่ 44 Confusion Matrix ด้วยวิธี Random Forest ของการฟื้นฟูสภาพกรณีที่ 2

Class		Prediction	
		0	1
Actual	0	93	1
	1	0	94

ตารางที่ 45 Confusion Matrix ด้วยวิธี Random Forest ของการฟื้นฟูสภาพกรณีที่ 3

Class	Prediction	
	0	1
Actual	0	93
	1	94

จากตารางที่ 43 - 45 Confusion Matrix สามารถที่จะนำมาคำนวณค่าวัดประสิทธิภาพความถูกต้องของแบบจำลองสำหรับการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาในการสร้างแบบจำลองการจำแนกประเภทด้วยวิธี Random Forest ดังตารางที่ 46

ตารางที่ 46 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม Random Forest

Performance	Random Forest		
	การฟื้นฟูสภาพกรณี ที่ 1	การฟื้นฟูสภาพ กรณีที่ 2	การฟื้นฟูสภาพ กรณีที่ 3
Accuracy	0.99	0.99	0.99
Precision	0.99	0.99	0.99
Recall	0.99	0.99	0.99
F1-Score	0.99	0.99	0.99
AUC	0.99	1.00	0.99

จากตารางที่ 46 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภท โดยใช้อัลกอริทึม Random Forest ซึ่งคำนวณประสิทธิภาพความถูกต้องจากตาราง Confusion Matrix (ตารางที่ 38 - 40) ที่ได้กำหนดค่าพารามิเตอร์ที่เหมาะสมของอัลกอริทึม Random Forest ดังตารางที่ 42 พบว่า ข้อมูลการฟื้นฟูสภาพกรณี 1 - 3 มีประสิทธิภาพค่าความถูกต้อง (Accuracy) สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาอยู่ที่ 99% ค่าความถ่วงดุล (F1-Score) สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา คือ ข้อมูลการฟื้นฟูสภาพกรณี 1 - 3 อยู่ที่ 99% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC สูงสุดในการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา คือ ข้อมูลการฟื้นฟูสภาพกรณี 2 อยู่ที่ 100% ตามด้วยข้อมูลการฟื้นฟูสภาพกรณี 1 และข้อมูลการฟื้นฟูสภาพกรณี 3 โดยมีค่า AUC เท่ากันอยู่ที่ 99%

4.4 การเปรียบเทียบประสิทธิภาพความถูกต้องในการจำแนกประเภท

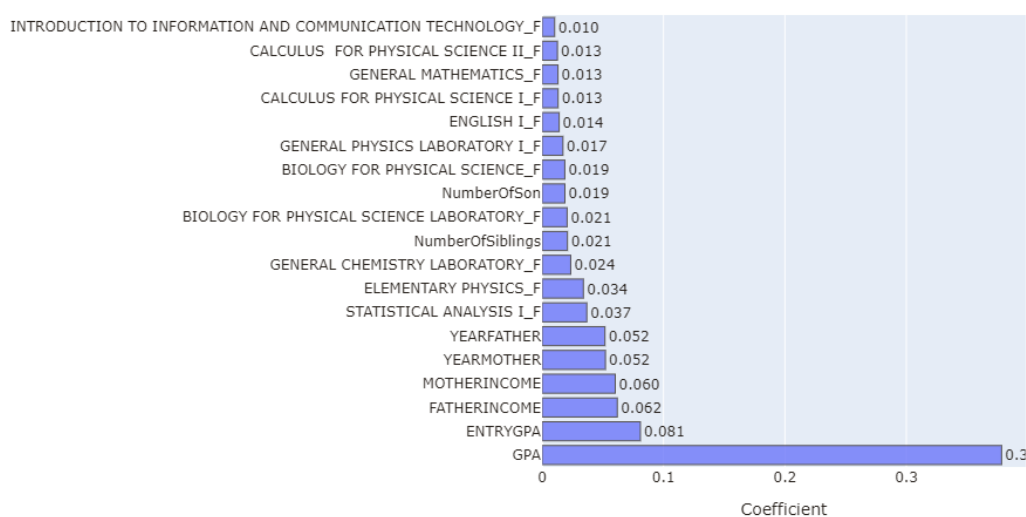
4.4.1 ข้อมูลการฟื้นฟูสภาพกรณี 1

ข้อมูลการฟื้นฟูสภาพกรณี 1 โดยคณะกรรมการฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากลาออก ฟื้นฟูสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา ฟื้นฟูสภาพเนื่องจากตกออก และฟื้นฟูสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด จากการทดลองในการสร้างการจำแนกประเภทแบบเดี่ยว และการจำแนกประเภทแบบรวมกลุ่ม การใช้การจำแนกประเภทแบบรวมกลุ่มที่ใช้ DT เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ พบว่า “ให้ประสิทธิภาพสูงกว่าการใช้การจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม DT” และการใช้การจำแนกประเภทแบบรวมกลุ่มที่ใช้ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ พบว่า “ให้ประสิทธิภาพน้อยกว่าการใช้การจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม SVM”

อัลกอริทึมที่มีประสิทธิภาพในการจำแนกประเภทการฟื้นฟูสภาพดีที่สุด คือ Random Forest และ Boosting Model (DT Base Model) ซึ่งให้ค่าประสิทธิภาพไม่แตกต่างกันโดยมีค่าความถูกต้อง (Accuracy) อยู่ที่ 99% ค่าความแม่นยำ (Precision) อยู่ที่ 99% ค่าความครบถ้วน (Recall) อยู่ที่ 99% ค่าความถ่วงดุล (F1-Score) อยู่ที่ 99% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC อยู่ที่ 99% สำหรับการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา โดยเลือกใช้ Random Forest เนื่องจากมีความซับซ้อนที่น้อย มีกฎในการจำแนกข้อมูลทำให้สามารถอธิบายข้อมูล และสามารถนำไปประยุกต์ใช้งานได้หลากหลายแพลตฟอร์ม แสดงในตารางที่ 47

ตารางที่ 47 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทการผันสภาพของข้อมูลการผันสภาพกรณี
ที่ 1

Performance	Single Model		Ensemble Model				
	DT	SVM	Bagging		Boosting		Random Forest
			DT	SVM	DT	SVM	
Accuracy	0.93	0.91	0.98	0.89	<u>0.99</u>	0.74	<u>0.99</u>
Precision	0.93	0.91	0.98	0.89	<u>0.99</u>	0.74	<u>0.99</u>
Recall	0.93	0.91	0.98	0.89	<u>0.99</u>	0.74	<u>0.99</u>
F1-Score	0.93	0.91	0.98	0.89	<u>0.99</u>	0.74	<u>0.99</u>
AUC	0.96	0.94	0.99	0.94	<u>0.99</u>	0.86	<u>0.99</u>



ภาพที่ 19 ปัจจัยที่ส่งผลต่อการผันสภาพของนักศึกษาของการผันสภาพกรณีที่ 1

จากการทดลองในการสร้างตัวจำแนกประเภทการผันสภาพของการผันสภาพกรณีที่ 1 โดยใช้ อัลกอริทึม Random Forest พบว่า ปัจจัยที่สำคัญต่อการผันสภาพของนักศึกษา 10 อันดับแรกคือ เกรดเฉลี่ย (GPA), เกรดเฉลี่ยระดับมัธยม (ENTRYGPA), รายได้บิดา (FATHERINCOME), รายได้มารดา (MOTHERINCOME), อายุมารดา (YEARMOTHER), อายุบิดา (YEARFATHER), การได้เกรด F ในรายวิชา STATISTICAL ANALYSIS I, ELEMENTARY PHYSICS, GENERAL CHEMISTRY LABORATORY, จำนวนพี่น้อง ตามลำดับ แสดงในภาพที่ 19

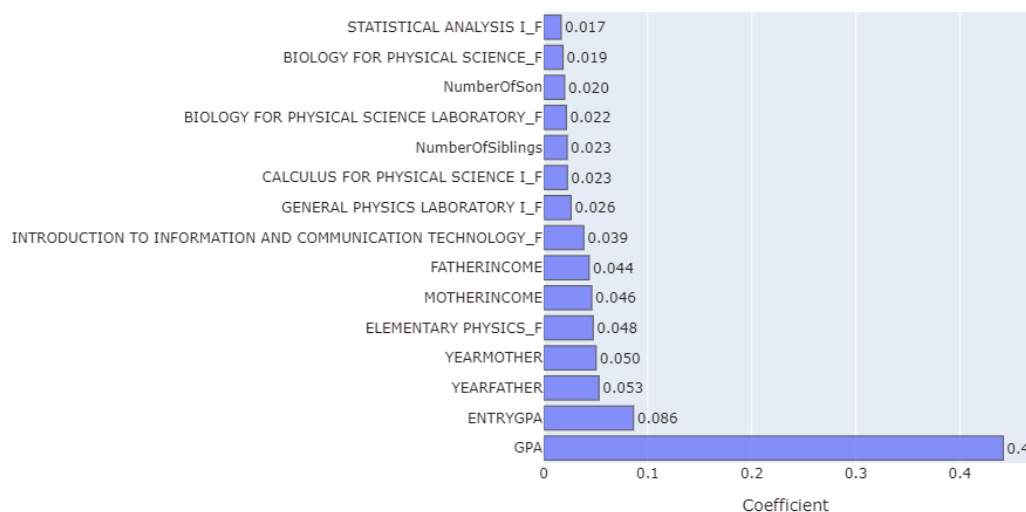
4.4.2 ข้อมูลการฝึกสภาพครั้งที่ 2

ข้อมูลการฝึกสภาพครั้งที่ 2 โดยสถานะการฝึกสภาพของนักศึกษาประกอบไปด้วย ฝึกสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา และฝึกสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด จากการทดลองในการสร้างการจำแนกประเภทแบบเดี่ยว และการจำแนกประเภทแบบรวมกลุ่ม พบว่า การใช้การจำแนกประเภทแบบรวมกลุ่มที่ใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐานสำหรับการเรียนรู้ให้ประสิทธิภาพไม่แตกต่างจากการใช้การจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม DT และ SVM

อัลกอริทึมที่มีประสิทธิภาพในการจำแนกประเภทการฝึกสภาพดีที่สุด คือ Random Forest ซึ่งให้ค่าประสิทธิภาพมีค่าความถูกต้อง (Accuracy) อยู่ที่ 99% ค่าความแม่นยำ (Precision) อยู่ที่ 99% ค่าความครบถ้วน (Recall) อยู่ที่ 99% ค่าความถ่วงดุล (F1-Score) อยู่ที่ 99% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC อยู่ที่ 100% สำหรับการจำแนกประเภทการฝึกสภาพของนักศึกษา แสดงในตารางที่ 48

ตารางที่ 48 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทการฝึกสภาพของข้อมูลการฝึกสภาพครั้งที่ 2

Performance	Single Model		Ensemble Model				
	DT	SVM	Bagging		Boosting		Random Forest
			DT	SVM	DT	SVM	
Accuracy	0.99	0.96	0.99	0.96	0.99	0.92	0.99
Precision	0.99	0.96	0.99	0.96	0.99	0.92	0.99
Recall	0.99	0.96	0.99	0.96	0.99	0.92	0.99
F1-Score	0.99	0.96	0.99	0.96	0.99	0.92	0.99
AUC	0.99	0.98	0.99	0.98	0.98	0.98	1.00



ภาพที่ 20 ปัจจัยที่ส่งผลต่อการฟื้นสภาพของนักศึกษาของการฟื้นสภาพกรณีที่ 2

จากการทดลองในการสร้างตัวจำแนกประเภทการฟื้นสภาพของการฟื้นสภาพกรณีที่ 2 โดยใช้ อัลกอริทึม Random Forest พบว่า ปัจจัยที่สำคัญต่อการฟื้นสภาพของนักศึกษา 10 อันดับแรก คือ เกรดเฉลี่ย (GPA), เกรดเฉลี่ยระดับมัธยม (ENTRYGPA), อายุบิดา (YEARFATHER), อายุมารดา (YEARMOTHER), การได้เกรด F ในรายวิชา ELEMENTARY PHYSICS, รายได้มารดา (MOTHERINCPME), รายได้บิดา (FATHERINCOME), การได้เกรด F ในรายวิชา INTRODUCTION TO INFORMATION AND COMMUNICATION TECHNOLOGY, GENERAL PHYSICAL LABORATORY, CALCULUS FOR PHYSICAL SCIENCE I ตามลำดับ แสดงในภาพที่ 20

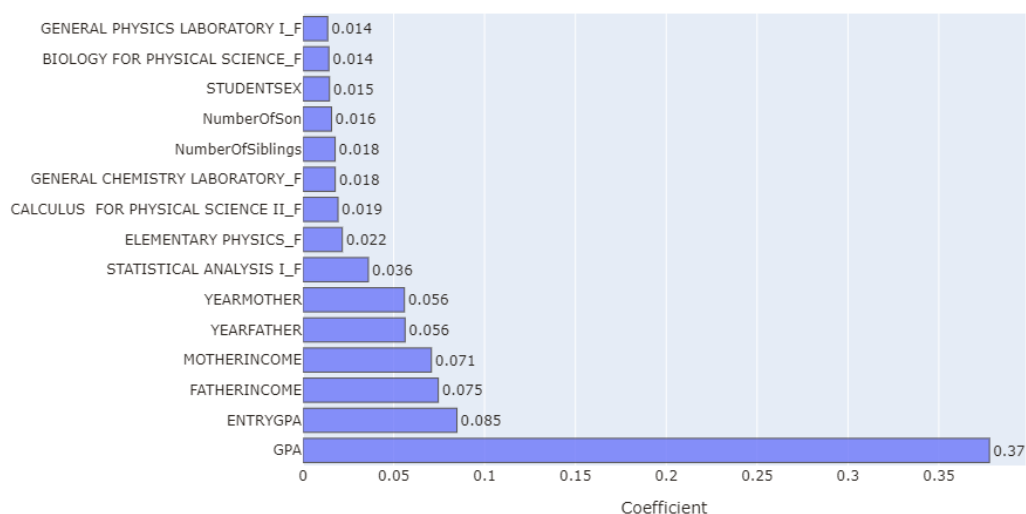
4.4.3 ข้อมูลการฟื้นสภาพกรณีที่ 3

ข้อมูลการฟื้นสภาพกรณีที่ 3 โดยสถานะการฟื้นสภาพของนักศึกษาประกอบไปด้วย ฟื้นสภาพเนื่องจากลาออก และฟื้นสภาพเนื่องจากตกรอก จากการทดลองในการสร้างการจำแนกประเภทแบบเดี่ยว และการจำแนกประเภทแบบรวมกลุ่ม พบว่า การใช้การจำแนกประเภทแบบรวมกลุ่มที่ใช้ DT และ SVM เป็น อัลกอริทึมพื้นฐานสำหรับการเรียนรู้ให้ประสิทธิภาพไม่แตกต่างจากการใช้การจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม DT และ SVM

อัลกอริทึมที่มีประสิทธิภาพในการจำแนกประเภทการฟื้นสภาพดีที่สุด คือ Random Forest ซึ่งให้ ประสิทธิภาพค่าความถูกต้อง (Accuracy) อยู่ที่ 99% ค่าความแม่นยำ (Precision) อยู่ที่ 99% ค่าความครบถ้วน (Recall) อยู่ที่ 99% ค่าความถ่วงดุล (F1-Score) อยู่ที่ 99% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC อยู่ที่ 99% สำหรับการจำแนกประเภทการฟื้นสภาพของนักศึกษา แสดงในตารางที่ 49

ตารางที่ 49 การวัดประสิทธิภาพความถูกต้องในการจำแนกประเภทการฟื้นฟูสภาพของข้อมูลการฟื้นฟูสภาพกรณี
ที่ 3

Performance	Single Model		Ensemble Model				
	DT	SVM	Bagging		Boosting		Random Forest
			DT	SVM	DT	SVM	
Accuracy	0.97	0.87	0.97	0.89	<u>0.99</u>	0.88	<u>0.99</u>
Precision	0.97	0.87	0.97	0.89	<u>0.99</u>	0.88	<u>0.99</u>
Recall	0.97	0.87	0.97	0.89	<u>0.99</u>	0.88	<u>0.99</u>
F1-Score	0.97	0.87	0.97	0.89	<u>0.99</u>	0.88	<u>0.99</u>
AUC	0.96	0.93	<u>0.99</u>	0.89	0.95	0.95	<u>0.99</u>



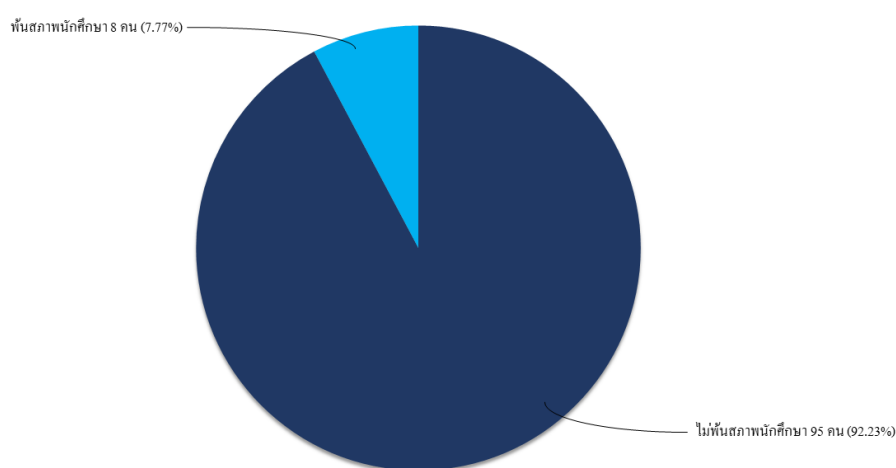
ภาพที่ 21 ปัจจัยที่ส่งผลต่อการฟื้นฟูสภาพของนักศึกษาของการฟื้นฟูสภาพกรณีที่ 3

จากการทดลองในการสร้างตัวจำแนกประเภทการฟื้นฟูสภาพของการฟื้นฟูสภาพกรณีที่ 3 โดยใช้ อัลกอริทึม Random Forest พบว่า ปัจจัยที่สำคัญต่อการฟื้นฟูสภาพของนักศึกษา 10 อันดับแรก คือ เกรดเฉลี่ย (GPA), เกรดเฉลี่ยระดับมัธยม (ENTRYGPA), รายได้บิดา (FATHERINCOME), รายได้มารดา (MOTHERINCOME), อายุบิดา (YEARFATHER), อายุมารดา (YEARMOTHER), การได้เกรด F ในรายวิชา STATISTICAL ANALYSIS I, ELEMENTARY PHYSICS, CALCULUS FOR PHYSICAL SCIENCE II, GENERAL CHEMISTRY LABORATORY ตามลำดับ แสดงในภาพที่ 21

4.5 การทำนายการฟื้นสภาพของนักศึกษาปีการศึกษา 2564

ผลลัพธ์จากการเลือกแบบจำลองที่ดีที่สุดโดยใช้อัลกอริทึม Random Forest สำหรับการเรียนรู้ ข้อมูลการฟื้นสภาพของนักศึกษาปีการศึกษา 2564 โดยทำการแบ่งชุดข้อมูลออกเป็น 3 กลุ่มตัวอย่าง เพื่อทำการจำแนกประเภทหรือทำนายการฟื้นสภาพของนักศึกษาปีการศึกษา 2564 ดังนี้

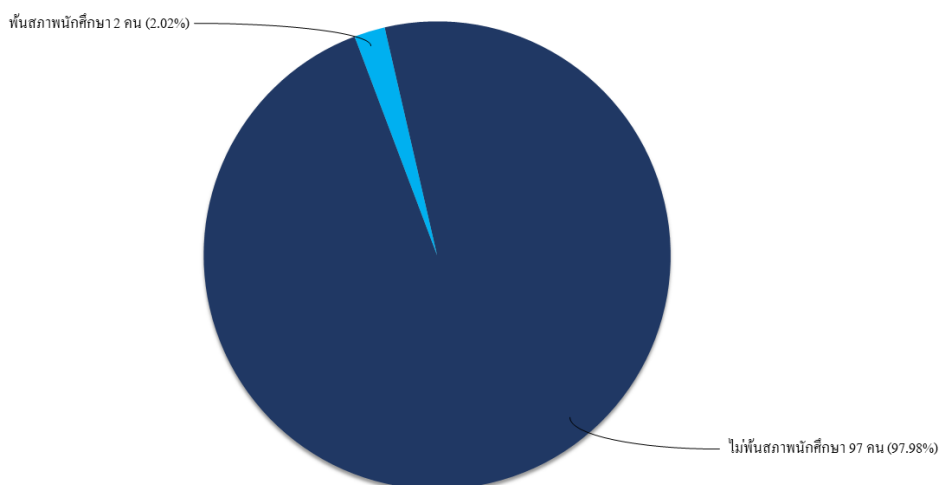
4.5.1 ข้อมูลการฟื้นสภาพกรณีที่ 1



ภาพที่ 22 แผนภูมิวงกลมของการทำนายการฟื้นสภาพนักศึกษากรณีที่ 1

ผลลัพธ์จากการใช้งานใช้อัลกอริทึม Random Forest สำหรับการเรียนรู้ข้อมูลการฟื้นสภาพของนักศึกษาปีการศึกษา 2564 ของข้อมูลการฟื้นสภาพกรณีที่ 1 มีจำนวน 103 คน โดยที่สถานะการฟื้นสภาพของนักศึกษาประกอบไปด้วย ฟื้นสภาพเนื่องจากลาออก ฟื้นสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา ฟื้นสภาพเนื่องจากตกออก และฟื้นสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด พบว่า นักศึกษาที่ถูกทำนายว่าไม่ฟื้นสภาพมีจำนวน 95 คนคิดเป็น 92.23% และนักศึกษาที่ถูกทำนายว่าฟื้นสภาพนักศึกษามีจำนวน 8 คนคิดเป็น 7.77% แสดงในภาพที่ 22

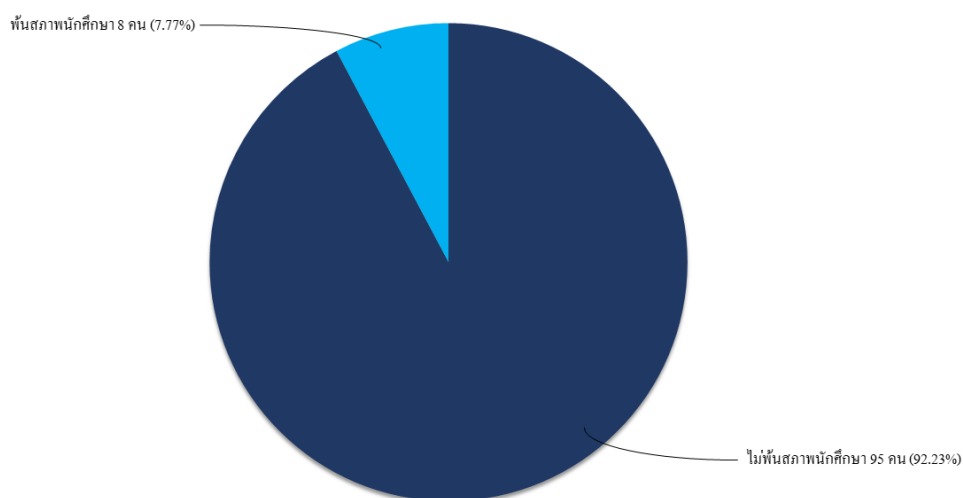
4.5.2 ข้อมูลการพ้นสภาพกรณีที่ 2



ภาพที่ 23 แผนภูมิวงกลมของการทำนายการพ้นสภาพนักศึกษากรณีที่ 2

ผลลัพธ์จากการใช้งานใช้อัลกอริทึม Random Forest สำหรับการเรียนรู้ข้อมูลการพ้นสภาพของนักศึกษาปีการศึกษา 2564 ของข้อมูลการพ้นสภาพกรณีที่ 2 มีจำนวน 99 คน โดยที่สถานะการพ้นสภาพของนักศึกษาประกอบไปด้วย พ้นสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา และพ้นสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด พบว่า นักศึกษาที่ถูกทำนายว่าไม่พ้นสภาพมีจำนวน 97 คนคิดเป็น 97.98% และนักศึกษาที่ถูกทำนายว่าพ้นสภาพนักศึกษามีจำนวน 2 คนคิดเป็น 2.02% แสดงในภาพที่ 23

4.5.3 ข้อมูลการพ้นสภาพกรณีที่ 3



ภาพที่ 24 แผนภูมิวงกลมของการทำนายการพ้นสภาพนักศึกษากรณีที่ 3

ผลลัพธ์จากการใช้งานใช้อัลกอริทึม Random Forest สำหรับการเรียนรู้ข้อมูลการพ้นสภาพของนักศึกษาปีการศึกษา 2564 ของข้อมูลการพ้นสภาพกรณีที่ 3 มีจำนวน 103 คน โดยที่สถานะการพ้นสภาพของนักศึกษาประกอบไปด้วย พ้นสภาพเนื่องจากลาออก และพ้นสภาพเนื่องจากตกรอก พบว่า นักศึกษาที่ถูกทำนายว่าไม่พ้นสภาพมีจำนวน 95 คนคิดเป็น 92.23% และนักศึกษาที่ถูกทำนายว่าพ้นสภาพนักศึกษามีจำนวน 8 คนคิดเป็น 7.77% แสดงในภาพที่ 24

บทที่ 5

สรุปผลการวิจัย

5.1 สรุปผลการวิจัย

ในงานวิจัยครั้งนี้มีจุดประสงค์เพื่อสร้างแบบจำลองเปรียบเทียบประสิทธิภาพความถูกต้องของการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาด้วยวิธีการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยว (Single Classification) กับการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม (Ensemble Classification) เพื่อได้แบบจำลองที่มี ความถูกต้อง และความแม่นยำ สำหรับการจำแนก ประเภทการฟื้นฟูสภาพของนักศึกษาที่ดีที่สุด โดยใช้ข้อมูลของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ โดยใช้ข้อมูลเฉพาะนักศึกษาปี 1 ในแต่ละปีการศึกษา 6 ปีย้อนหลังตั้งแต่ ปีการศึกษา 2558–2563 มีจำนวน 614 คน โดยได้ทำการทดลองออกเป็น 3 กรณี จากนั้นนำข้อมูลแต่ละกรณีมาผ่านกระบวนการทำความสะอาดข้อมูล (Clean Data) สร้างตัวแปรหุ่น (Dummy Variable) ทำการคัดเลือกข้อมูล (Selection Data) ปรับสมดุลข้อมูล (Balancing Data) และปรับปรุงขอบเขตข้อมูล (Data Scaling) ก่อนการเรียนรู้หรือสร้างแบบจำลองจำแนกประเภท จากผลการวิจัยพบว่า วิธีการสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่ม (Ensemble Model) มีประสิทธิภาพการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษาได้ดีกว่าวิธีการสร้างแบบจำลองการจำแนกประเภทแบบเดี่ยว (Single Model) ซึ่งมีความสอดคล้องตรงตามสมมติฐานที่ตั้งไว้ จากการเปรียบเทียบประสิทธิภาพการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา พบว่า การใช้อัลกอริทึม Random Forest สำหรับการเรียนรู้ให้ประสิทธิภาพในการจำแนกทั้ง 3 กลุ่มตัวอย่างดีที่สุด โดยประสิทธิภาพทั้ง 3 กรณี ให้ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) ค่าความถ่วงดุล (F1-Score) มากกว่า 98% และค่า AUC พื้นที่ใต้เส้นโค้ง ROC มากกว่า 99% สำหรับการจำแนกประเภทการฟื้นฟูสภาพของนักศึกษา

จากแบบจำลองจากการใช้อัลกอริทึม Random Forest พบว่า ปัจจัยที่ส่งผลต่อการฟื้นฟูสภาพของนักศึกษาของข้อมูลการฟื้นฟูสภาพกรณีที่ 1 โดยสถานะการฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากลาออก ฟื้นฟูสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา ฟื้นฟูสภาพเนื่องจากตกออก และฟื้นฟูสภาพเนื่องจากไม่ลงทะเบียนตามเวลากำหนด พบว่าปัจจัยที่สำคัญต่อการฟื้นฟูสภาพของนักศึกษา 10 อันดับแรกคือ เกรดเฉลี่ย, เกรดเฉลี่ยระดับมัธยม, รายได้บิดา, รายได้มารดา, อายุมารดา, อายุบิดา, เกรด F รายวิชา STATISTICAL ANALYSIS I, ELEMENTARY PHYSICS, GENERAL CHEMISTRY LABORATORY, จำนวนพี่น้อง ตามลำดับ

ปัจจัยที่ส่งผลต่อการฟื้นฟูสภาพของนักศึกษาของข้อมูลการฟื้นฟูสภาพกรณีที่ 2 โดยสถานะการฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากไม่ชำระค่าต่อทะเบียนนักศึกษา และฟื้นฟูสภาพ

เนื่องจากไม่ลงทะเบียนตามเวลากำหนด พบว่าปัจจัยที่สำคัญต่อการฟื้นฟูสภาพของนักศึกษา 10 อันดับแรกคือ เกรดเฉลี่ย, เกรดเฉลี่ยระดับมัธยม, อายุบิดา, อายุมารดา, เกรด F รายวิชา ELEMENTARY PHYSICS, รายได้มารดา, รายได้บิดา, เกรด F รายวิชา INTRODUCTION TO INFORMATION AND COMMUNICATION TECHNOLOGY, GENERAL PHYSICAL LABORATORY, CALCULUS FOR PHYSICAL SCIENCE I ตามลำดับ

ปัจจัยที่ส่งผลต่อการฟื้นฟูสภาพของนักศึกษาของข้อมูลการฟื้นฟูสภาพครั้งที่ 3 โดยคณะกรรมการฟื้นฟูสภาพของนักศึกษาประกอบไปด้วย ฟื้นฟูสภาพเนื่องจากลาออก และฟื้นฟูสภาพเนื่องจากตกออก พบว่าปัจจัยที่สำคัญต่อการฟื้นฟูสภาพของนักศึกษา 10 อันดับแรกคือ เกรดเฉลี่ย, เกรดเฉลี่ยระดับมัธยม, รายได้บิดา, รายได้มารดา, อายุบิดา, อายุมารดา, เกรด F รายวิชา STATISTICAL ANALYSIS I, ELEMENTARY PHYSICS, CALCULUS FOR PHYSICAL SCIENCE II, GENERAL CHEMISTRY LABORATORY ตามลำดับ

อีกทั้งเมื่อนำแบบจำลอง Random Forest มาใช้ในการทำนายการฟื้นฟูสภาพของนักศึกษาปีการศึกษา 2564 ของข้อมูลการฟื้นฟูสภาพครั้งที่ 1, 2, 3 โดยใช้อัลกอริทึม Random Forest พบว่านักศึกษาที่ถูกทำนายว่าไม่ฟื้นฟูสภาพคิดเป็น 91.23%, 97.98%, 92.23% และนักศึกษาที่ถูกทำนายว่าฟื้นฟูสภาพนักศึกษาคิดเป็น 7.77%, 2.02%, 7.77% จากจำนวน 103, 99, 103 คนตามลำดับ

5.2 อภิปรายผลการวิจัย

การเปรียบเทียบประสิทธิภาพการจำแนกประเภทแบบเดี่ยวโดยใช้อัลกอริทึม DT และ SVM และการจำแนกประเภทแบบรวมกลุ่มที่ใช้ DT และ SVM เป็นอัลกอริทึมพื้นฐาน พบว่า การใช้เทคนิคการจำแนกประเภทแบบรวมกลุ่มให้ประสิทธิภาพความแม่นยำมากกว่าการจำแนกประเภทแบบเดี่ยว และอัลกอริทึม Random Forest ให้ประสิทธิภาพความแม่นยำมากที่สุดสำหรับการจำแนกการฟื้นฟูสภาพของนักศึกษา ซึ่งสอดคล้องกับสมมติฐานที่ตั้งไว้ และงานวิจัยของ Naseem et al. (2020) ได้ศึกษาและเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกประเภทด้วยการเรียนรู้ของเครื่อง (Machine Learning) ในการฟื้นฟูสภาพของนักศึกษา ซึ่งพบว่า อัลกอริทึม Random Forest มีประสิทธิภาพในการจำแนกมากที่สุด ทั้งนี้เนื่องจากการใช้อัลกอริทึม Random Forest สามารถเพิ่มเติมในส่วนของฟังก์ชันการทำงานแบบสุ่มเลือกคุณลักษณะ ของข้อมูลที่ใช้ในการวิเคราะห์เข้ามา ทำให้ลดค่าสหสัมพันธ์ (Correlation) ของคุณลักษณะลงในการสร้างต้นไม้แต่ละต้นที่มีความเป็นอิสระต่อกัน จึงทำให้ต้นไม้ในแต่ละต้นที่ถูกสร้างขึ้น เพื่อใช้สำหรับการจำแนกประเภทข้อมูลมีโครงสร้างต้นไม้ที่มีขนาดเล็ก ซึ่งจะทำงานได้เร็วและให้ประสิทธิภาพที่ดีกว่าวิธีการแบบอื่นๆ (ปัทม์ อุปการ์, 2560)

สำหรับปัจจัยที่ส่งผลต่อการพัฒนาของนักศึกษา พบว่า เกรดเฉลี่ย (GPA) เป็นปัจจัยที่ส่งผลต่อการพัฒนาของนักศึกษามากที่สุดซึ่งได้ผลการวิเคราะห์ปัจจัยที่มีอิทธิพลต่อการพัฒนาของนักศึกษาทั้ง 3 กรณี ได้แก่ เกรดเฉลี่ย, สอดคล้องกับงานวิจัยของ Tenpipat (2020), Naseem (2020), และนนทวัฒน์ (2563) ที่พบว่าเกรดเฉลี่ยมีผลต่อการพัฒนาของนักศึกษา ซึ่งเป็นผลมาจากคะแนนเฉลี่ยสะสมไม่ถึงเกณฑ์ที่มหาวิทยาลัยกำหนด

ในทำนองเดียวกัน การวิจัยครั้งนี้ยังพบว่าผลการเรียนในรายวิชาพื้นฐานที่ต้องเรียนในชั้นปีที่ 1 มีผลต่อการพัฒนาของนักศึกษา โดยการศึกษานี้คล้ายกับการศึกษาการพัฒนาของนักศึกษาสาขาวิทยาการคอมพิวเตอร์ที่ University of the South Pacific ที่พบว่าผลการศึกษา (เกรด) ในรายวิชาพื้นฐาน มีผลต่อการพัฒนาของนักศึกษา (Naseem, 2020)

อีกทั้งยังพบว่าผลการเรียนระดับมัธยมศึกษาเป็นอีกหนึ่งปัจจัยที่มีอิทธิพลต่อการพัฒนาของนักศึกษา ซึ่งผลที่พบจากการวิจัยครั้งนี้สนับสนุนผลการศึกษาของมหาวิทยาลัยในประเทศไทย (นนทวัฒน์ (2563), ซอและ (2561), และ Tenpipat (2020)) ที่พบว่าเกรดเฉลี่ยระดับมัธยมมีผลต่อการพัฒนาของนักศึกษา เนื่องจากเกรดเฉลี่ยระดับมัธยมมีความสัมพันธ์ต่อเกรดเฉลี่ยระดับอุดมศึกษา การที่นักศึกษาที่มีเกรดระดับมัธยมค่อนข้างต่ำจะทำให้มีโอกาสที่นักศึกษาจะมีโอกาสพัฒนาค่อนข้างสูง

ผลการศึกษาในครั้งนี้ยังพบว่าปัจจัยทางด้านรายได้ของครัววงศ์ส่งผลต่อการพัฒนาของนักศึกษา ซึ่งสอดคล้องกับการศึกษาของ Hutagaol (2019) ที่พบว่ารายได้ของผู้ปกครองเป็นปัจจัยทำนายที่สำคัญในการพัฒนาของนักศึกษา

5.3 ประโยชน์ของสถิติที่ใช้ในการวิเคราะห์

การวิจัยเรื่องการจำแนกประเภทสำหรับการพัฒนาของนักศึกษาระดับปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ ซึ่งตามวัตถุประสงค์ได้ทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทแบบเดี่ยว (Single Classification) และแบบรวมกลุ่ม (Ensemble Classification) พบว่าการใช้อัลกอริทึม Random Forest จำแนกการพัฒนาของนักศึกษาได้ดีที่สุด และสามารถทราบปัจจัยที่สำคัญต่อการพัฒนาของนักศึกษา เพื่อที่จะทำการหาแนวทางในการป้องกัน แก้ไขปัญหา และวางกลยุทธ์ในการจัดการการพัฒนาของนักศึกษาได้อย่างทันทั่วถึง

5.4 ข้อเสนอแนะ

5.4.1 นำข้อมูลการพัฒนาของนักศึกษาแต่ละคณะให้อัลกอริทึมเรียนรู้ เพื่อที่จะเพิ่มประสิทธิภาพในการทำนายการพัฒนาของนักศึกษาได้แม่นยำมากขึ้น

5.4.2 เพิ่มการพิจารณาปัจจัยด้าน การกู้ยืมยศ. และการได้รับทุนการศึกษาเป็นต้น

- 5.4.3 สามารถนำอัลกอริทึมไปใช้ในการพัฒนาเว็บแอปพลิเคชัน (web application) ให้สามารถใช้งานได้จริง ในการทำนายการฟื้นฟูสภาพของนักศึกษา
- 5.4.4 สามารถนำแบบจำลองไปใช้งานในระบบออนไลน์ เพื่อให้ผู้ที่เกี่ยวข้องในการดูแลนักศึกษาในการเตรียมความพร้อม วางกลยุทธ์ และวางแผนรับมือกับการฟื้นฟูสภาพของนักศึกษาในระดับปริญญาตรี
- 5.4.5 การนำอัลกอริทึมการตรวจจับค่าผิดปกติ (Outlier Detection Algorithms) ต่างๆมาประยุกต์ใช้สำหรับการลบค่าที่ผิดปกติออกจากข้อมูล เช่น Isolation Forest, Minimum Covariance Determinant, Local Outlier Factor, และ One-Class SVM
- 5.4.6 งานวิจัยนี้ผู้วิจัยได้ทำการปรับขอบเขตข้อมูลโดยใช้วิธีการ Min-Max Normalization นอกเหนือจากนี้สามารถนำการทำ Feature Engineering ต่างๆมาประยุกต์ใช้สำหรับกระบวนการจัดเตรียมข้อมูลอาจทำให้ประสิทธิภาพในการจำแนกการฟื้นฟูสภาพของนักศึกษาเพิ่มสูงขึ้น ไม่มากนักน้อย เช่น Principal Component Analysis (PCA), K-Means, Encoding, และ Transform Data
- 5.4.7 งานวิจัยนี้มุ่งเน้นที่การสร้างแบบจำลองการจำแนกประเภทแบบรวมกลุ่มโดยใช้วิธี Bagging Boosting, และ Random Forest ซึ่งแบบจำลองการจำแนกประเภทแบบรวมกลุ่มมีวิธีการอื่นๆอีกมากมาย ตัวอย่างเช่น Stacking, Blending, และ Voting นอกจากนี้ยังมีโมเดลที่ถูกพัฒนามาจากวิธีการ Bagging และ Boosting เช่น XGBoost, GBM, Light GBM, CatBoost เป็นต้น
- 5.4.8 การนำอัลกอริทึมโครงข่ายประสาทเทียม (Neuron Network) หรือ การเรียนรู้เชิงลึก (Deep Learning) มาสร้างแบบจำลองการจำแนกการฟื้นฟูสภาพของนักศึกษา อาจทำให้ประสิทธิภาพในการจำแนกการฟื้นฟูสภาพของนักศึกษาเพิ่มสูงขึ้น ไม่มากนักน้อย เช่น Artificial Neural Network (ANN) และ Convolutional Neural Network (CNN)

เอกสารอ้างอิง

- ชนิดาภา บุญประสม , &จรัญ แสนราช. (2561). การวิเคราะห์การทำนายการลาออกกลางคันของ
นักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูล. **Technical Education**
Journal King Mongkut's University of Technology North Bangkok, 9(1),142-151.
- ชอและ เกป็น, พิมลพรรณ ลีลาภัทรพันธุ์, และอัจฉราพร ขกขุน. (2561). การวิเคราะห์ปัจจัยที่มีผลต่อ
การฟื้นฟูสภาพของนักศึกษาโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษา หลักสูตรวิทยาการ
คอมพิวเตอร์หลักสูตรเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏยะลา. **Veridian E-**
Journal, Science and Technology Silpakorn University, 5(4), 96-110.
- ชนันท์ จระสมบุรณ์ และ วราภรณ์ วิทยานนท์. (2561). การทำนายการซื้อซ้ำของผู้ซื้อโดยใช้เทคนิคการ
เรียนรู้ของเครื่อง (วิทยานิพนธ์ปริญญาโทบริหารบัณฑิต). กรุงเทพฯ: มหาวิทยาลัยศรีนคริน-
ทรวิโรฒ
- นนทวัฒน์ ทวีชาติ, อรยา เฟื่องประจัญ, วิไลรัตน์ ยาทองไชย, และชูศักดิ์ ยาทองไชย. (2563).
ระบบทำนายการฟื้นฟูสภาพของนักศึกษาระดับปริญญาตรี คณะวิทยาศาสตร์
มหาวิทยาลัยราชภัฏบุรีรัมย์ ด้วยเทคนิคการทำเหมืองคอม. **Journal of Science and**
Technology Buriram Rajabhat, 4(1), 47-60.
- นิตานันท์ พลอาสา. (2558). การสร้างแบบจำลองการขายผลิตภัณฑ์และพยากรณ์ยอดขายประกัน
ชีวิต โดยเทคนิคการทำเหมืองข้อมูล กรณีศึกษา บริษัทประกันชีวิตแห่งหนึ่ง
(วิทยานิพนธ์ปริญญาโทบริหารบัณฑิต). กรุงเทพฯ: มหาวิทยาลัยธรรมศาสตร์
- ปัทมธนา บุญรักษา และ จาริ ทองคำ. (2560). การเปรียบเทียบประสิทธิภาพของแบบจำลองการเกิด
อุบัติเหตุทางถนน โดยใช้เทคนิคอนุกรมเวลา. **วารสารวิชาการ การจัดการเทคโนโลยี**
สารสนเทศและนวัตกรรม, 4(2), 40-46.
- ปัทมภ์ อุปการ. (2560). การปรับปรุงประสิทธิภาพสำหรับการจำแนกประเภทชนิดของกลุ่มเมมโดยใช้
วิธีการเรียนรู้แบบรวมกลุ่ม (วิทยานิพนธ์ปริญญาโทบริหารบัณฑิต). พิษณุโลก: มหาวิทยาลัยนเรศวร
- วิชญ์วิสิฐ เกษรสิทธิ์, ดร.วิชิต หล่อจิระขุนทด, และดร.จิราวัลย์ จิตรถเวช. (2561). การแก้ปัญหาข้อมูล
ไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน. **KKU Research Journal**
(Graduate Study), 18(3), 11-21.
- สุกัญญา ทารศ. (2562). ปัจจัยจำแนกการออกกลางคันของนิสิตปริญญาตรี มหาวิทยาลัยมหาสารคาม
(วิทยานิพนธ์ปริญญาโทบริหารบัณฑิต). มหาสารคาม: มหาวิทยาลัยมหาสารคาม

- สำนักทะเบียนมหาวิทยาลัยขอนแก่น. (2562). ระเบียบมหาวิทยาลัยขอนแก่นว่าด้วย การศึกษาชั้น
ปริญญาตรี พ.ศ. ๒๕๖๑. สืบค้น 4 ตุลาคม 2564.
จาก https://home.kku.ac.th/meeting/Document/KKU_R2562-bachelorgegree.pdf
- สำนักทะเบียนและประมวลผล มหาวิทยาลัยขอนแก่น. (2564). ทะเบียนรายชื่อนักศึกษา จากระบบของ
สำนักทะเบียนมหาวิทยาลัยขอนแก่น. สืบค้น 4 ตุลาคม 2564. จาก <https://reg.kku.ac.th/>
- Gatchalee, P. (2019). **Confusion Matrix เครื่องมือสำคัญในการประเมินผลลัพธ์ของการ
ทำนายในMachine learning**. สืบค้น 15 กรกฎาคม 2564,
จาก [https://medium.com/@pagongatchalee/confusion-matrix-เครื่องมือสำคัญใน
การประเมินผลลัพธ์ของการทำนาย-ในmachine-learning-fba6e3f9508c](https://medium.com/@pagongatchalee/confusion-matrix-เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย-ในmachine-learning-fba6e3f9508c)
- Hutagaol, N., & Suharjito. (2019). Predictive Modelling of Student Dropout Using Ensemble
Classifier Method in Higher Education. **Advance in Science, Technology and Engineering
Systems Journal (ASTES)**, 4(4). 206-211.
- Naseem, M., Chaudhary, K., Sharma, B., & Lal, A., G. (2019). Using Ensemble Decision
Tree Model to Predict Student Dropout in Computing Science. **IEEE Asia-
Pacific Conference on Computer Science and Data Engineering (CSDE)**.
doi: 10.1109/CSDE48274.2019.9162389
- Tenpipat, W., & Akkarajitsakul, K. (2020). Student Dropout Prediction:
A KMUTT Case Study. **2020 1st International Conference on Big Data
Analytics and Practices (IBDAP)**. doi: 10.1109/IBDAP50342.2020.9245457
- Uddin, S., Khan, A., & Hossain, M. (2019). Comparing different supervised machine learning
algorithms for disease prediction. **BMC Med Inform Decis Mak**, 19(1), 1-16.
doi: 10.1186/s12911-019-1004-8

ภาคผนวก ก
ตารางการดำเนินโครงการวิจัย
และค่าใช้จ่ายในการดำเนินโครงการ

ดำเนินโครงการภายในระยะเวลาระหว่างเดือนกรกฎาคม พ.ศ. 2564 ถึง เดือนเมษายน พ.ศ. 2565 โดยมีขั้นตอนการดำเนินงานดังนี้

ตารางที่ 50 รายละเอียดค่าใช้จ่ายในการดำเนินงาน

[illegible]

การวิจัยครั้งนี้มีงบประมาณค่าใช้จ่ายในการดำเนินงานดังนี้

ตารางที่ 51 การดำเนินงานโครงการวิจัย

รายการ	จำนวนเงิน
จัดทำเอกสารต่าง ๆ ในงานวิจัย	0
จัดทำรูปเล่มฉบับสมบูรณ์	0
จัดทำโปสเตอร์	0
รวมเป็นเงิน	0