การคาดการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ ด้วยเทคนิคการเรียนรู้ของเครื่อง

Prince of Songkla University Students' Dropout Prediction Using Machine Learning

กฤตกร อินแพง

Kittakorn Inpang

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science
Prince of Songkla University
2565

ลิขสิทธิ์ของมหาวิทยาลัยสงขลานครินทร์

การคาดการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ ด้วยเทคนิคการเรียนรู้ของเครื่อง

Prince of Songkla University Students' Dropout Prediction Using Machine Learning

กฤตกร อินแพง

Kittakorn Inpang

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science
Prince of Songkla University
2565

ลิขสิทธิ์ของมหาวิทยาลัยสงขลานครินทร์

ชื่อวิทยานิพนธ์ การคาดการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ ด้วยเทคนิคการเรียนรู้ของเครื่อง

ผู้เขียน นายกฤตกร อินแพง

สาขาวิชา วิทยาการข้อมูล

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

.....

(ดร. นฤบาล ยมะคุปต์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

.....

(ผู้ช่วยศาสตราจารย์ ดร. จุไรรัตน์ พุทธรักษ์)

คณะกรรมการสอบ

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร. อภิรดี แซ่ลิ่ม)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. สุภาภรณ์ ชัยอารยะเลิศ)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. ณิชนันทน์ กิตติพัฒนบวร)

.....กรรมการ

(ดร. นฤบาล ยมะคุปต์)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. จุไรรัตน์ พุทธรักษ์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

.....

(ศาสตราจารย์ ดร. ดำรงศักดิ์ ฟ้ารุ่งสาง)

คณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

ลงชื่อ .....
(ดร. นฤบาล ยมะคุปต์)
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ลงชื่อ .....
(ผู้ช่วยศาสตราจารย์ ดร. จุไรรัตน์ พุทธรักษ์)
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

ลงชื่อ .....
(นายกฤตกร อินแพง)
นักศึกษา

ข้าพเจ้าขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ .....

(นายกฤตกร อินแพง)

นักศึกษา

ชื่อวิทยานิพนธ์ การคาดการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ ด้วยเทคนิคการเรียนรู้ของเครื่อง
ผู้เขียน นายกฤตกร อินแพง
สาขาวิชา วิทยาการข้อมูล
ปีการศึกษา 2564

บทคัดย่อ

อัตราการคงอยู่ของนักศึกษาเป็นส่วนสำคัญและเป็นตัวชี้วัดหนึ่งในการวัดความสำเร็จของสถาบันการศึกษา อย่างไรก็ตามผู้ที่เข้ามาศึกษาไม่สามารถสำเร็จการศึกษาในระบบได้ทั้งหมด เนื่องจากส่วนหนึ่งต้องออกจากการศึกษากลางคัน เช่นเดียวกับมหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ โดยตั้งแต่ปีการศึกษา พ.ศ. 2556 ถึง 2560 พบว่าอัตราการออกกลางคันของนักศึกษาเพิ่มสูงขึ้นอยู่ที่ร้อยละ 19.18 ด้วยเหตุนี้ผู้วิจัยจึงได้นำเสนอการนำเทคนิคเหมืองข้อมูลและการเรียนรู้ของเครื่องมาวิเคราะห์เพื่อค้นหาคุณลักษณะที่สำคัญและสร้างแบบจำลองการเรียนรู้ของเครื่องประเภทต้นไม้ 5 แบบ เพื่อคาดการณ์การออกกลางคันของนักศึกษา โดยใช้ข้อมูลนักศึกษามหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ 6 รุ่นปีการศึกษา คือในช่วง พ.ศ. 2558 ถึง 2563 จำนวน 33,930 ราย 39 ตัวแปร ขอบเขตของงานวิจัยนี้คือข้อมูลผลลัพธ์ทางการศึกษา ข้อมูลพื้นฐานของนักศึกษา และข้อมูลครอบครัวของนักศึกษา โดยแบ่งข้อมูลเป็นสองชุดคือ ชุดข้อมูลระดับปริญญาตรี พบว่าแบบจำลองไลท์กาเดียนบูทติ้งแมชชีนเป็นวิธีที่ดีที่สุดให้ค่าพื้นที่ใต้กราฟสูงที่สุดร้อยละ 93.03 และค่าความถูกต้องร้อยละ 89.99 และปัจจัยสำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษา คือ ผลการเรียนเฉลี่ยสะสม รองลงมาคือชั้นปี ผลการเรียนเฉลี่ยปัจจุบัน ภาคการศึกษา ผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษา และคะแนนภาษาอังกฤษก่อนเข้าศึกษา ตามลำดับ และชุดข้อมูลระดับบัณฑิตศึกษา พบว่าแบบจำลองแรนดอมฟอเรสต์เป็นวิธีที่ดีที่สุดให้ค่าพื้นที่ใต้กราฟสูงที่สุดร้อยละ 78.86 และค่าความถูกต้องร้อยละ 85.28 และปัจจัยที่สำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษา คือ ผลการเรียนเฉลี่ยสะสม รองลงมาคือชั้นปี ภาคการศึกษา กลุ่มสาขาวิชาสังคมศาสตร์และมนุษยศาสตร์ ผลการเรียนเฉลี่ยปัจจุบัน ประเภทภาคสมทบ และแผนการศึกษาแผน ก แบบ ก2 ตามลำดับ โดยขั้นตอนสุดท้ายผู้วิจัยนำแบบจำลองที่ได้ไปทำการคาดการณ์กับข้อมูลจริงและแสดงผลการวิเคราะห์นำเสนอรายงานแดชบอร์ดเพื่อติดตามความเสี่ยง ซึ่งจะช่วยให้เจ้าหน้าที่ที่เกี่ยวข้องสามารถเข้าช่วยเหลือนักศึกษาที่มีความเสี่ยงได้ทันที และเพื่อช่วยผู้บริหารในการสนับสนุนการตัดสินใจและวางแผนการบริหารงานเพื่อลดอัตราการออกกลางคันในมหาวิทยาลัยให้ต่ำลงได้

คำสำคัญ : นักศึกษาออกกลางคัน, อุดมศึกษา, การเรียนรู้ของเครื่อง, เหมืองข้อมูล, การคาดการณ์

Thesis Title Prince of Songkla University Students' Dropout Prediction Using Machine Learning
Author Mr. Kittakorn Inpang
Major Program Data Science
Academic Year 2021

ABSTRACT

Student retention rate plays a critical role and serves as an essential indicator of a tertiary institution's success. However, not all first-time students complete their program at the same institution within a specified period of time: some students drop out of the program. Prince of Songkla University Hatyai Campus is no exception. From Academic Years 2013-2017, the student dropout rates rose by 19.18%. This research study adopted data mining and machine learning techniques to explore factors that predict the likelihood of a student dropping out, and to create a learning model of five-decision trees, which will be used for the prediction of the student dropouts. Data were collected from 33,930 students of Prince of Songkla University Hatyai Campus, from 6 intakes ranging from Academic Years 2015-2020, and with 39 variables. Collected data cover students' learning achievements, students' basic information, and students' family background. Data were classified into two categories: Undergraduate and Postgraduate. As for undergraduate category, the study found that Light Gradient Boosting Machine is the most appropriate methodology, as it yielded the highest value of the area under the curve of 93.03%, and the accuracy value of 89.99%. The top factors that predict the likelihood for student dropouts include the accumulated (overall) grade point average (GPAX); academic year; Grade Point Average (GPA); semester; pre-university GPAX; and pre-university English scores, respectively. As for postgraduate category, the study found that the Random Forest is the most appropriate methodology, as it yielded the highest value of the area under the curve of 78.86%, and the accuracy value of 85.28%. The top factors that predict the likelihood for student dropouts include GPAX; academic year; semester; social and humanity science; GPA; supplementary class; and Plan A, A2-Type, respectively. In the final procedure, the researcher implemented the obtained models for making a prediction with

the actual data, and visually presented the results of the analysis in the dashboard report, which can be used for monitoring possible risks. This will enable respective staff to give immediate assistance to the students who are in needs or show the likelihood to drop out, and help the management board in making decisions and devising management plans to minimize the dropout rate in their institution.

Keywords : Student Dropout, Higher Education, Machine Learning, Data Mining, Prediction

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ ดร. นฤบาล ยมะคุปต์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลักและผู้ช่วยศาสตราจารย์ ดร. จุไรรัตน์ พุทธรักษ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ช่วยแนะนำแนวทางการจัดทำวิทยานิพนธ์และช่วยตรวจทานแก้ไขวิทยานิพนธ์จนสำเร็จ

นอกจากนี้ข้าพเจ้าขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. อภิรดี แซ่ลิ่ม ประธานกรรมการ ผู้ช่วยศาสตราจารย์ ดร. สุภาภรณ์ ชัยอารยะเลิศ อาจารย์ประจำหลักสูตร และผู้ช่วยศาสตราจารย์ ดร. ณิชนันทน์ กิตติพัฒนบวร ผู้ทรงคุณวุฒิภายนอก ที่ให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ และได้กรุณาตรวจสอบและให้คำแนะนำการแก้ไขวิทยานิพนธ์ฉบับนี้ให้มีความถูกต้องครบถ้วนสมบูรณ์ และขอขอบพระคุณอาจารย์ในสาขาวิชาวิทยาการข้อมูลทุกท่านที่ถ่ายทอดความรู้และให้คำแนะนำในการจัดทำวิทยานิพนธ์ฉบับนี้ และเจ้าหน้าที่ทุกท่านที่ช่วยอำนวยการและประสานงานต่าง ๆ ในระหว่างการจัดทำวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงไปด้วยดี

สุดท้ายนี้ข้าพเจ้าขอขอบพระคุณทุนอุดหนุนการศึกษาระดับบัณฑิตศึกษาภายในประเทศ ประจำปีการศึกษา 2563 มหาวิทยาลัยสงขลานครินทร์ที่ให้ความอนุเคราะห์ทุนอุดหนุนการศึกษา รวมถึงครอบครัว และเพื่อนร่วมงานที่สนับสนุนและให้กำลังใจเสมอมาตลอดจนผู้เขียนบทความ หนังสือ ตำรา วารสารต่าง ๆ ที่ให้ความรู้แก่ข้าพเจ้าจนสามารถจัดทำวิทยานิพนธ์สำเร็จได้ด้วยดี

กฤตกร อินแพง

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญตาราง (ต่อ)

หน้า

สารบัญตาราง (ต่อ)

หน้า

สารบัญภาพ

หน้า

สารบัญภาพ (ต่อ)

หน้า

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหา

การศึกษาระดับอุดมศึกษามีความสำคัญในด้านการพัฒนาประเทศ สังคม และเศรษฐกิจ มีบทบาทในการสร้างขีดความสามารถในการแข่งขัน โดยสามารถผลิตกำลังคนที่มีทักษะการทำงาน สร้างวิทยาการ นวัตกรรม องค์ความรู้ที่มีความต้องการของประเทศในภาคอุตสาหกรรม ภาคธุรกิจ และในระดับชุมชน และท้องถิ่น [1] อย่างไรก็ตามผู้ที่เข้ามาศึกษาไม่สามารถสำเร็จการศึกษาในระบบได้ทั้งหมด เนื่องจากส่วนหนึ่งต้องออกจากการศึกษากลางคัน ส่งผลให้เกิด “ความสูญเปล่าในการลงทุนเพื่อการศึกษา” สถานศึกษาเสียเวลาและทรัพยากรในการลงทุนบริหารจัดการส่วนผู้เรียนเสียเวลาและเงินทองรวมถึงเสียขวัญและกำลังใจในการกลับไปเริ่มต้นใหม่ [2] ความกังวลที่สำคัญในการศึกษาระดับอุดมศึกษาคืออัตราการคงอยู่ของนักศึกษา ซึ่งเป็นตัวชี้วัดความสำเร็จของสถาบันการศึกษา และเป็นเกณฑ์ชี้วัดคุณภาพการเรียนการสอนที่รัฐบาลใช้เป็นเกณฑ์การจัดสรรเงินทุนงบประมาณของสถาบัน [3] การคงอยู่ของนักศึกษาที่เข้าศึกษาตั้งแต่ปีการศึกษาแรกจนสำเร็จการศึกษาเป็นหนึ่งความสำคัญไม่เฉพาะต่อภาพลักษณ์ ชื่อเสียง สวัสดิภาพทางการเงินเท่านั้น แต่ยังเป็นความท้าทายที่แสดงถึงประสิทธิภาพและความน่าเชื่อถือของสถาบัน วิธีที่จัดการความท้าทายนี้ได้อย่างมีประสิทธิภาพคือการวิเคราะห์และการนำเสนอข้อมูล หรือการทำเหมืองข้อมูลเพื่อค้นหารูปแบบที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ [4] และแปลงเป็นความรู้เพื่อปรับปรุงกระบวนการตัดสินใจวางแผนการบริหารจัดการ [5] การคาดการณ์ผลลัพธ์ที่แม่นยำของการออกกลางคันของนักศึกษาเพื่อให้สถาบันการศึกษาสามารถใช้ข้อมูลเพื่อช่วยเหลือด้านวิชาการแก่นักศึกษาที่มีความเสี่ยง [4]

เทคนิคการทำเหมืองข้อมูลมีความสำคัญต่อกระบวนการเรียนรู้และผลลัพธ์เพื่อก้าวสู่มหาวิทยาลัยยุคใหม่ เครื่องมือที่จะช่วยให้สถาบันอุดมศึกษาสามารถแก้ไขปัญหาและสนับสนุนการตัดสินใจได้คือ การทำเหมืองข้อมูลเพื่อการศึกษา (Educational Data Mining) และการวิเคราะห์การเรียนรู้ (Learning Analytics) [6] การทำเหมืองข้อมูลเพื่อการศึกษาเป็นการใช้หลักทางสถิติเหมืองข้อมูล (Data Mining) และการเรียนรู้ของเครื่อง (Machine Learning) วัตถุประสงค์เพื่อวิเคราะห์ข้อมูลเพื่อแก้ไขปัญหาการวิจัยทางการศึกษา [7] การจำแนกประเภทและประเมินการใช้

แบบจำลองที่ไม่มีผู้สอน (Unsupervised) หรือแบบมีผู้สอน (Supervised) เหมืองข้อมูลแบบไม่มีผู้สอนใช้ในสถานการณ์ที่ไม่รู้การจัดกลุ่มหรือรูปแบบ เช่น ข้อมูลกลุ่มหลักสูตรหรือประเภทหลักสูตรใดเกี่ยวข้องกับนักศึกษาประเภทใด การทำเหมืองข้อมูลแบบไม่มีผู้สอนมักใช้เพื่อศึกษารูปแบบและค้นหารูปแบบที่ซ่อนไว้ก่อนหน้าเพื่อทำความเข้าใจและจำแนก ส่วนเหมืองข้อมูลแบบมีผู้สอนใช้กับข้อมูลที่ทราบผลลัพธ์ เช่น ข้อมูลผู้สำเร็จการศึกษาที่รวมข้อมูลนักศึกษาที่ตกออก ใช้เพื่อศึกษาพฤติกรรมของสองกลุ่มเพื่อเชื่อมโยงพฤติกรรมกับประวัติการเรียนและข้อมูลอื่น ๆ ที่บันทึกไว้ เหล่านี้เรียกว่าการเรียนรู้ของเครื่อง ซึ่งเป็นส่วนหนึ่งของปัญญาประดิษฐ์เพื่อกำหนดรูปแบบที่นักวิเคราะห์ข้อมูลสามารถนำไปใช้งานกับข้อมูลใหม่ได้ เช่น นักศึกษาใหม่และแบบจำลองที่ทำนายการสำเร็จการศึกษาได้ ขั้นตอนทั้งหมดทำงานอย่างแม่นยำ รวดเร็ว โดยอัตโนมัติ ประหยัดเวลาเมื่อเทียบกับการทำนายแบบดั้งเดิม [4]

การทำเหมืองข้อมูลเพื่อการศึกษาแบบการจัดประเภท (Classification) ได้รับความนิยมมากที่สุดในการสร้างแบบจำลองการเรียนรู้ของเครื่อง มีการใช้อัลกอริทึมประเภทต้นไม้ตัดสินใจ (Decision Tree) ได้แก่ Decision Tree, Random Forest, และ Gradient Boosting ในการคาดการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี พบว่าแบบจำลองทั้ง 3 ผลลัพธ์ไม่แตกต่างอย่างมีนัยสำคัญ ซึ่ง Gradient Boosting มีความแม่นยำสูงที่สุด และพบว่าปัจจัยที่ส่งผลต่อการออกกลางคันคือ รุ่นปีการศึกษา ผลการเรียนก่อนหน้าและประเภทการเข้ารับการศึกษาตามลำดับ [8] นอกจากนี้ยังมีการนำแบบจำลองอื่นที่นิยมใช้ได้แก่ โครงข่ายประสาทเทียม (Neural networks) [9], [10] และ Decision Tree, K-Nearest Neighbors, Logistics Regression, Naïve Bayes, Random Forest และ Support Vector Machine ในการคาดการณ์การออกกลางคันของนักศึกษาในประเทศชิลี ในช่วงชั้นปีที่ 1-3 พบว่าแบบจำลอง Random Forest มีประสิทธิภาพสูงสุด และปัจจัยผลการเรียน การเข้าศึกษา และดัชนีความยากจนเป็นตัวแปรที่สำคัญในการคาดการณ์การออกกลางคันของนักศึกษา [11]

ข้อมูลนักศึกษาย้อนหลัง 5 รุ่นปีการศึกษา พ.ศ. 2556 ถึง 2560 มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ พบว่าในปีการศึกษา 2556 นักศึกษารับเข้าทั้งหมดอยู่ที่ 5,309 คน มีนักศึกษาออกกลางคันทั้งหมด 1,066 คน คิดเป็นร้อยละ 20.08 ปีการศึกษา 2557 นักศึกษารับเข้าทั้งหมดอยู่ที่ 5,248 คน มีนักศึกษาออกกลางคันทั้งหมด 981 คน คิดเป็นร้อยละ 18.69 ปีการศึกษา 2558 นักศึกษารับเข้าทั้งหมดอยู่ที่ 5,256 คน มีนักศึกษาออกกลางคันทั้งหมด 1011 คน คิดเป็นร้อยละ 19.24 ปีการศึกษา 2559 นักศึกษารับเข้าทั้งหมดอยู่ที่ 5,218 คน มีนักศึกษาออกกลางคันทั้งหมด 1,009 คน คิดเป็นร้อยละ 19.34 และปีการศึกษา 2560 นักศึกษารับเข้าทั้งหมดอยู่ที่ 4,951 คน มีนักศึกษาออกกลางคันทั้งหมด 917 คน คิดเป็นร้อยละ 18.52 ตามลำดับ รวมห้าปีย้อนหลังมีนักศึกษาออกกลางคันทั้งหมด 4,984 คน คิดเป็นร้อยละ 19.18 [12]

งานวิจัยนี้เสนอรูปแบบคุณลักษณะที่สำคัญและการคาดการณ์การออกกลางคันของนักศึกษา จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าส่วนใหญ่ศึกษาข้อมูลปัจจัยด้านผลลัพธ์ทางการศึกษาและข้อมูลส่วนตัวของนักศึกษา ซึ่งปัจจัยข้อมูลครอบครัว เช่น อาชีพบิดา-มารดา รายได้บิดา-มารดา รวมถึงปัจจัยทางด้านสุขภาพ ปัจจัยทางการเงิน และการได้รับทุนการศึกษาของนักศึกษาก็ล้วนเป็นปัจจัยสำคัญ งานวิจัยนี้จึงได้นำตัวแปรปัจจัยดังกล่าวมาวิเคราะห์โดยใช้ข้อมูลจากระบบข้อมูลพื้นฐานนักศึกษา และระบบข้อมูลนักศึกษา 6 รุ่นปีการศึกษา พ.ศ. 2558 ถึง 2563 มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ผ่านเทคนิคเหมืองข้อมูลและแบบจำลองการเรียนรู้ของเครื่องประเภทต้นไม้ 5 แบบ นำมาทดสอบและเปรียบเทียบประสิทธิภาพแบบจำลองและความเหมาะสมในการนำไปใช้งาน และเพื่อระบุคุณลักษณะปัจจัยที่สำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษา ซึ่งสิ่งนี้สามารถช่วยให้เจ้าหน้าที่ คณาจารย์ และผู้บริหารในการตัดสินใจการบริหารงานและให้ความช่วยเหลือแก่นักศึกษาที่มีความเสี่ยง เพื่อลดอัตราการออกกลางคันในมหาวิทยาลัยได้

1.2 วัตถุประสงค์

เป้าหมายของการวิจัยครั้งนี้คือหาปัจจัยที่มีผลต่ออัตราการออกกลางคันของนักศึกษาโดยใช้เทคนิคการเรียนรู้ของเครื่องเพื่อนำผลลัพธ์มาปรับปรุงประสิทธิภาพการเรียนในสถาบันอุดมศึกษา โดยมีวัตถุประสงค์ ดังนี้

1.2.1 เพื่อค้นหาปัจจัยที่มีผลต่ออัตราการออกกลางคันของนักศึกษาโดยใช้เทคนิคการเรียนรู้ของเครื่อง

1.2.2 เพื่อเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องประเภทต้นไม้ตัดสินใจ 5 แบบ และเลือกวิธีที่มีประสิทธิภาพที่สุดในการคาดการณ์การออกกลางคันของนักศึกษา

1.3 ประโยชน์ที่คาดว่าจะได้รับ

การศึกษาปัจจัยที่มีผลต่ออัตราการออกกลางคันของนักศึกษาโดยใช้เทคนิคการเรียนรู้ของเครื่อง ครั้งนี้มีประโยชน์ดังนี้

1.3.1 คณาจารย์หรือมหาวิทยาลัยสามารถใช้ข้อมูลเพื่อช่วยเหลือนักศึกษาที่มีความเสี่ยง เพื่อลดอัตราการออกกลางคันในมหาวิทยาลัยได้

1.3.2 นำข้อมูลปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาไปปรับใช้กับสถาบันอุดมศึกษาอื่น ๆ ได้

1.3.3 นำเสนอการพัฒนาและทดสอบอัลกอริทึมการเรียนรู้ของเครื่องเพื่อคาดการณ์การออกกลางคันของนักศึกษาที่แม่นยำ และรายงานแดชบอร์ดที่มีประสิทธิภาพ

1.4 ขอบเขตการวิจัย

ศึกษาปัจจัยที่มีผลต่อการออกกลางคันของนักศึกษา มหาวิทยาลัยสงขลานครินทร์ โดยใช้เทคนิคเหมืองข้อมูลและแบบจำลองการเรียนรู้ของเครื่องเพื่อนำผลลัพธ์มาปรับปรุงประสิทธิภาพการเรียนในสถาบันอุดมศึกษา ผู้วิจัยได้กำหนดขอบเขตของการวิจัยดังต่อไปนี้

1.4.1 ขอบเขตประชากร

การวิจัยนี้เป็นการศึกษาปัจจัยที่มีผลต่ออัตราการออกกลางคันของนักศึกษา มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ช่วงรุ่นปีการศึกษา พ.ศ. 2558 – 2563 โดยวิเคราะห์ปัจจัยต่าง ๆ ดังนี้

1.4.1.1 ปัจจัยผลลัพธ์ทางการศึกษา

1.4.1.2 ปัจจัยข้อมูลพื้นฐานของนักศึกษา

1.4.1.3 ปัจจัยด้านครอบครัวของนักศึกษา

1.4.2 ขอบเขตพื้นที่

มหาวิทยาลัยสงขลานครินทร์

1.4.3 ขอบเขตระยะเวลา

ระยะเวลาในการดำเนินงานวิจัย 2 ปี ตั้งแต่เดือนกรกฎาคม พ.ศ. 2563 ถึง เดือนพฤษภาคม พ.ศ. 2565

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 เหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลเป็นกระบวนการรวบรวมข้อมูล ทำความสะอาดข้อมูล ประมวลผลข้อมูล และวิเคราะห์ข้อมูล เพื่อค้นหาข้อมูลเชิงลึกที่เป็นประโยชน์ [13] เป็นการผสมผสานระหว่างความรู้และทักษะการวิเคราะห์เพื่อค้นพบความสัมพันธ์ รูปแบบ และแนวโน้มข้อมูล ผ่านข้อมูลที่จัดเก็บไว้จำนวนมาก โดยใช้เทคโนโลยีและสถิติทางคณิตศาสตร์ [4] เพื่อค้นหาและตีความเพื่อแก้ปัญหาทางธุรกิจ เช่น การทำเหมืองข้อมูลเพื่อสร้างแบบจำลองเชิงพรรณนา การสร้างแบบจำลองคาดการณ์ การเพิ่มประสิทธิภาพการดำเนินงานวิจัย การวิเคราะห์ข้อความและเหมืองข้อมูลทางสถิติ [14] กระบวนการเหมืองข้อมูลประกอบด้วยหลายขั้นตอน ได้แก่

1. การรวบรวมข้อมูล (Data Collection) การรวบรวมข้อมูลเป็นขั้นตอนที่มีความสำคัญ การรวบรวมข้อมูลมักถูกจัดเก็บไว้ในฐานข้อมูล หรือคลังข้อมูลสำหรับการประมวลผลซึ่งต้องใช้ฮาร์ดแวร์ และเครือข่ายในการประมวลผล

2. การแยกคุณลักษณะและการทำความสะอาดข้อมูล (Feature Extraction and Data Cleaning) เมื่อมีการจัดเก็บรวบรวมข้อมูล ข้อมูลมักไม่อยู่ในรูปแบบที่เหมาะสมสำหรับการประมวลผล จำเป็นต้องทำการแปลงข้อมูล หรือโครงสร้างข้อมูลให้เหมาะสม รวมถึงขั้นตอนการแยกคุณลักษณะและการทำความสะอาดข้อมูล เช่น ข้อมูลสูญหาย หรือข้อมูลผิดพลาด ซึ่งขั้นตอนนี้เป็นอีกขั้นตอนที่สำคัญต่อกระบวนการทำเหมืองข้อมูล

3. การประมวลผลเชิงวิเคราะห์และอัลกอริทึม (Analytical Processing and Algorithms) เป็นขั้นตอนการวิเคราะห์ที่มีประสิทธิภาพจากกระบวนการประมวลผลข้อมูล [13]

2.1.1.1 ประเภทของแบบจำลองเหมืองข้อมูล

กระบวนการค้นพบรูปแบบจากชุดข้อมูล เรียกว่า การสร้างแบบจำลองเชิงวิเคราะห์ เพื่อสร้างแบบจำลอง เป็นการระบุความสัมพันธ์ของตัวแปรในข้อมูลและใช้ความสัมพันธ์นั้น

เพื่อสร้างแบบจำลองการคาดการณ์หรือแบบจำลองเชิงพรรณนา ประเภทของแบบจำลองเหมืองข้อมูลประกอบด้วย 2 ประเภทได้แก่

1. Predictive Model คือแบบจำลองที่สร้างขึ้นเพื่อคาดการณ์ผลลัพธ์ที่เฉพาะเจาะจงหรือตัวแปรเป้าหมาย เช่น เทคนิค Multiple Regression, Logistic Regression และ Decision Trees

2. Description Model คือแบบจำลองที่ช่วยให้เข้าใจข้อมูลได้ดีขึ้น โดยไม่มีตัวแปรเป้าเหมายเฉพาะ ซึ่งเทคนิคแบบจำลองเชิงพรรณนาที่ใช้กันทั่วไป ได้แก่ การวิเคราะห์ปัจจัย การวิเคราะห์คลัสเตอร์ และการวิเคราะห์ความสัมพันธ์ต่าง ๆ [14]

2.1.1.2 กระบวนการมาตรฐานสำหรับเหมืองข้อมูล CRISP-DM (Cross Industry Standard Process for Data Mining)

แบบจำลองเหมืองข้อมูลถูกสร้างแบบปรับใช้ต่าง ๆ มากมาย ภายหลังได้เกิดข้อตกลงร่วมกันเกี่ยวกับกระบวนการมาตรฐานสำหรับเหมืองข้อมูล ซึ่งกระบวนการนี้เรียกว่า CRISP-DM (Cross Industry Standard Process for Data Mining) [4] ประกอบด้วย 6 ขั้นตอน ได้แก่

1. ความเข้าใจทางธุรกิจ (Business Understanding) มุ่งเน้นการทำความเข้าใจวัตถุประสงค์ของโครงการในมุมมองธุรกิจ และแปลงความรู้นี้เป็นคำจำกัดความของปัญหาการทำเหมืองข้อมูล

2. ความเข้าใจข้อมูล (Data Understanding) ทำความเข้าใจข้อมูลด้วยการรวบรวมข้อมูลและกำหนดปัญหา

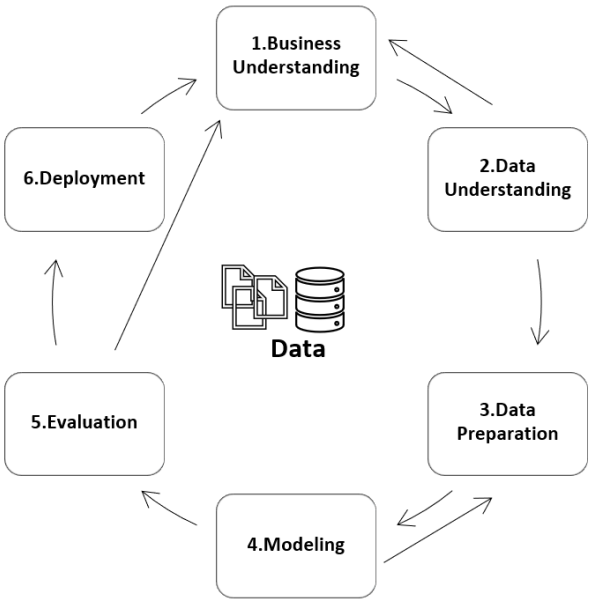
3. การเตรียมข้อมูล (Data Preparation) เตรียมข้อมูลให้ครอบคลุม และแปลงข้อมูลเตรียมสร้างแบบจำลอง

4. การสร้างแบบจำลอง (Modeling) สร้างแบบจำลองต่าง ๆ ที่เหมาะสมที่สุดที่จะนำไปใช้งาน

5. การประเมินผลแบบจำลอง (Evaluation) ประเมินแบบจำลองที่จะนำไปใช้งานจริงตามวัตถุประสงค์

6. และการปรับใช้ (Deployment) การใช้งานแบบจำลองที่สร้างขึ้น [14]

ซึ่งกระบวนการมาตรฐานสำหรับเหมืองข้อมูล CRISP-DM สามารถแสดงได้ดังภาพที่ 2.1



ภาพที่ 2.1 กระบวนการ CRISP-DM (Cross Industry Standard Process for Data Mining)

2.1.2 เหมืองข้อมูลเพื่อการศึกษา (Educational Data Mining) และการวิเคราะห์การเรียนรู้ (Learning Analytics)

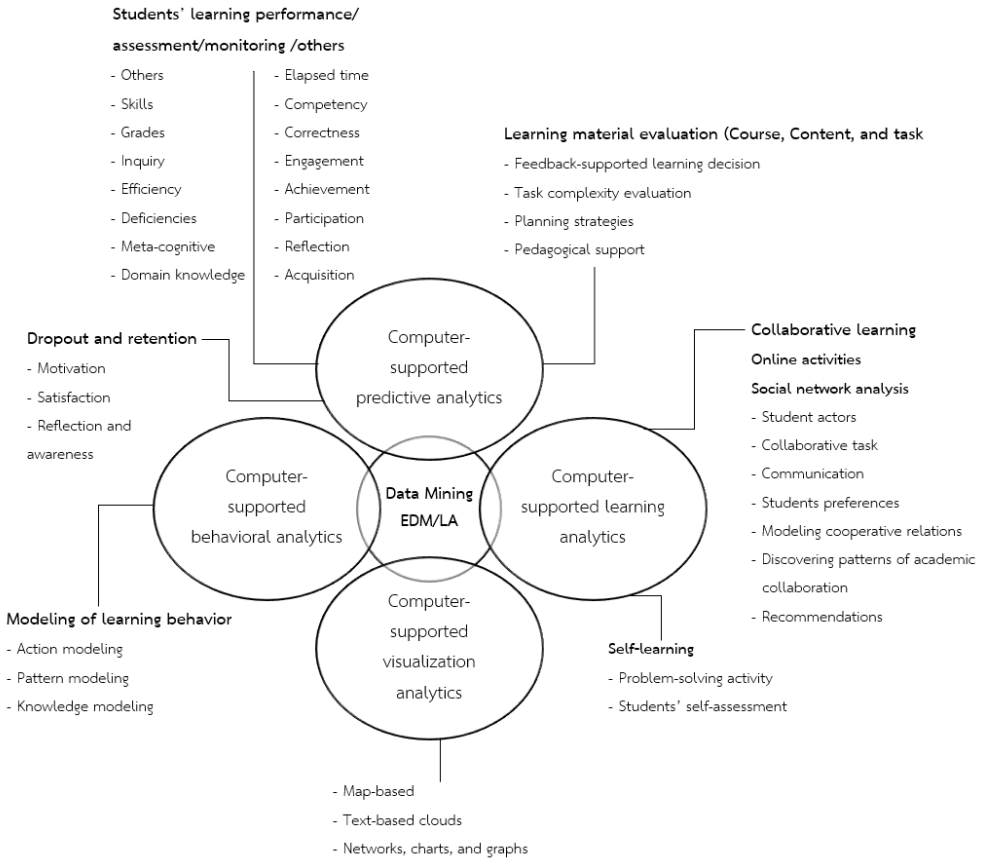
การทำเหมืองข้อมูลในระดับอุดมศึกษา (Data Mining in Higher Education) เป็นเครื่องมือที่ช่วยให้มหาวิทยาลัยสามารถคาดการณ์ข้อมูลทางวิชาการได้อย่างมีประสิทธิภาพ [4] การทำเหมืองข้อมูลเพื่อการศึกษาและการวิเคราะห์การเรียนรู้ เป็นเทคนิคการทำเหมืองข้อมูลมีความสำคัญต่อกระบวนการเรียนรู้และผลลัพธ์เพื่อก้าวสู่มหาวิทยาลัยยุคใหม่ เครื่องมือที่จะช่วยแก้ไขปัญหาและสนับสนุนการตัดสินใจ สนับสนุนการเรียนรู้ด้วยตนเอง กระบวนการที่เกี่ยวข้องกับการเรียนรู้ร่วมกัน การติดตาม การประเมินผล เช่น วัดประสิทธิภาพการเรียนและอัตราการออกกลางคันของนักศึกษาโดยใช้เทคนิค เช่น การจำแนกประเภท (Classification) การจัดกลุ่ม (Clustering) กฎการเชื่อมโยง (Association Rules) สามารถแบ่งการวิเคราะห์ออกเป็นสี่มิติหลัก ได้แก่

1. การวิเคราะห์การเรียนรู้ที่สนับสนุนด้วยคอมพิวเตอร์ (Computer-Supported Learning Analytics)

2. การวิเคราะห์เชิงคาดการณ์ที่สนับสนุนด้วยคอมพิวเตอร์ (Computer-Supported Predictive Analytics)

3. การวิเคราะห์พฤติกรรมที่สนับสนุนด้วยคอมพิวเตอร์ (Computer-Supported Behavioral Analytics)

4. และการวิเคราะห์การแสดงภาพที่รองรับด้วยคอมพิวเตอร์ (Computer-Supported Visualization Analytics) [6] การวิเคราะห์สามารถแสดงได้ดังภาพที่ 2.2



ภาพที่ 2.2 เหมืองข้อมูลเพื่อการศึกษาและการวิเคราะห์การเรียนรู้ (Data Mining EDM/LA) [6]

2.1.3 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่องเป็นรูปแบบการวิเคราะห์ข้อมูลด้วยแบบจำลองอัตโนมัติ ซึ่งเป็นสาขาหนึ่งของเทคโนโลยีด้าน AI (Artificial intelligence) บนแนวคิดที่ว่าระบบต่าง ๆ สามารถที่จะเรียนรู้และมีปฏิสัมพันธ์กับชุดข้อมูลต่าง ๆ รวมถึงสามารถระบุรูปแบบต่าง ๆ ที่เกิดขึ้นนำไปสู่การตัดสินใจได้เองโดยไม่จำเป็นต้องพึ่งพามนุษย์ [15] ตามคำจำกัดความอย่างเป็นทางการของนักวิทยาศาสตร์คอมพิวเตอร์ Tom M. Mitchell ให้คำจำกัดความการเรียนรู้ของเครื่องว่า การเรียนรู้ได้โดยไม่ต้องเขียนโปรแกรมเพิ่มเติม โดยกล่าวว่าเครื่องสามารถเรียนรู้จากประสบการณ์เพื่อปรับปรุงประสิทธิภาพในอนาคตตามประสบการณ์ที่คล้ายคลึงกัน [16] ปัจจุบันมีการนำการเรียนรู้ของเครื่องมาปรับใช้กับการทำงานร่วมกับแพทย์ในการรักษาโรคมะเร็ง ช่วยนักวิทยาศาสตร์และวิศวกรในการออกแบบสร้างองค์ความรู้ที่เกี่ยวข้อง รวมไปถึงภาคธุรกิจ องค์กร ภาครัฐ และโรงพยาบาล ตัวอย่างการนำการเรียนรู้ของเครื่องไปใช้งาน เช่น การพยากรณ์สภาพอากาศ การแบ่งกลุ่มพฤติกรรมของลูกค้าเพื่อการโฆษณาที่ตรงเป้าหมาย การระบุข้อความสแปมในอีเมล การทำนายผลการเลือกตั้ง การ

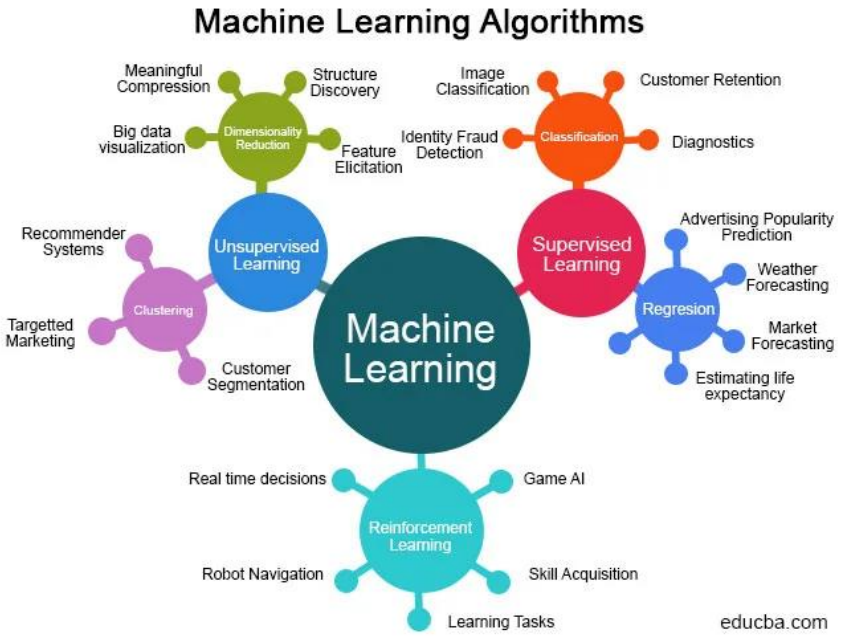
พัฒนาอัลกอริทึมสำหรับขับโดรนอัตโนมัติและรถยนต์ไร้คนขับ การฉายภาพบริเวณที่มีโอกาสก่ออาชญากรรม รวมไปถึงการค้นพบลำดับพันธุกรรมที่เชื่อมโยงกับโรคต่าง ๆ [17]

2.1.3.1 ประเภทของอัลกอริทึมการเรียนรู้ของเครื่องสามารถแบ่งได้เป็น

1. การเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการเรียนรู้ภายใต้การดูแลจากข้อมูลการฝึกอบรม ที่มีค่าป้ายกำกับ โดยใช้คุณลักษณะดังกล่าวเพื่อจำแนกข้อมูลในอนาคต เช่น การจำแนกกลุ่ม (Classification) การถดถอย (Regression) หรือการคาดการณ์ (Prediction)

2. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการเรียนรู้แบบไม่มีผู้ดูแล เป็นวิธีการที่ข้อมูลการฝึกอบรมซึ่งใช้ในการสร้างแบบจำลองไม่มีคุณลักษณะที่บอกค่าประเภทตัวอย่าง ดังนั้นจึงไม่สามารถระบุได้ว่าค่าข้อมูลใดในแต่ละแถวถูกจัดประเภทหรือมีกี่กลุ่ม แบบจำลองวิธีการเรียนรู้ดังกล่าวจะใช้คุณลักษณะที่มีอยู่ในข้อมูลตัวอย่างเพื่อระบุรูปแบบของแต่ละกลุ่มและกำหนดเอกลักษณ์กลุ่มให้กับข้อมูล ด้วยวิธีนี้ โมเดลจะแบ่งข้อมูลออกเป็นกลุ่ม ๆ ตัวอย่างการเรียนรู้แบบไม่มีผู้ดูแลคือการจัดกลุ่ม (Clustering) และการเรียนรู้เพื่อค้นหากฎการเชื่อมโยง (Association Rules)

3. การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) เป็นการเรียนรู้ของเครื่องจากการลองผิดลองถูกภายใต้สถานการณ์ต่าง ๆ โดยโมเดลเรียนรู้จากการกระทำ การตัดสินใจที่ถูกต้องและการตัดสินใจที่ผิดพลาด ซึ่งช่วยให้เรียนรู้รูปแบบและตัดสินใจได้แม่นยำยิ่งขึ้นสำหรับข้อมูลที่ไม่รู้จัก [17], [18] ประเภทของอัลกอริทึมสามารถแสดงได้ดังภาพที่ 2.3



ภาพที่ 2.3 ประเภทของอัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning Algorithms) [19]

2.1.3.2 หมวดหมู่ของอัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning Algorithm) แบ่งออกเป็น

1. การถดถอย (Regression) มีความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ ตัวแปรเป้าหมายมีลักษณะเป็นตัวเลข ในขณะที่ตัวแปรอิสระอาจเป็นหมวดหมู่หรือตัวเลขก็ได้

2. การจัดประเภท (Classification) เป็นการจัดประเภทข้อมูลแบ่งออกได้เป็น Binary คือตัวแปรเป้าหมายมีเพียงสองผลลัพธ์ เช่น 0/1, ใช่/ไม่ใช่, จริง/เท็จ, Multinomial คือตัวแปรเป้าหมายที่มีหลายผลลัพธ์ เช่น Mango, Orange, Apple และอื่น ๆ และ Ordinal คือตัวแปรเป้าหมายจะถูกเรียงลำดับ เช่น ผลการเรียน [19]

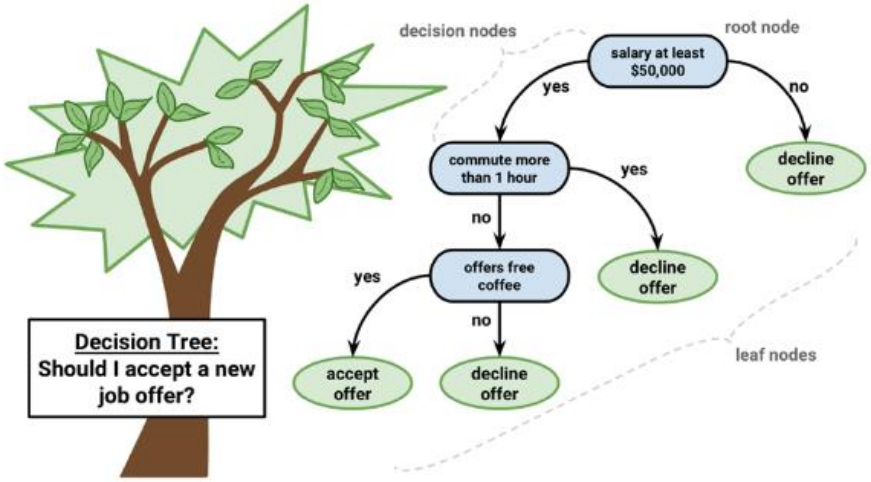
2.1.4 ต้นไม้การตัดสินใจ (Decision Tree)

ต้นไม้การตัดสินใจ (Decision Tree) คืออัลกอริทึมการเรียนรู้ของเครื่องประเภทการเรียนรู้แบบมีผู้สอน (Supervised Learning) เป้าหมายของอัลกอริทึมคือการจำแนกข้อมูลเป็นกลุ่มคลาสโดยใช้คุณลักษณะที่เกี่ยวข้องมากที่สุดของข้อมูล ซึ่งโครงสร้างการตัดสินใจประกอบด้วย

1. รูทโหนด (Root Node) คือ โหนดภายในซึ่งเป็นจุดเริ่มต้นของแผนผังการตัดสินใจต่าง ๆ ตามค่าของคุณลักษณะ

2. โหนดการตัดสินใจ (Decision Nodes) คือ ค่าที่เป็นไปได้ของการตัดสินใจตามคุณสมบัติของงาน เป็นผลลัพธ์ใช่หรือไม่ใช่ บางกรณีอาจมีความเป็นไปได้มากกว่าสองอย่าง

3. ใบ (Leaf Nodes) คือ การตัดสินใจขั้นสุดท้ายได้ ต้นไม้จะสิ้นสุดลงโดยโหนดปลายสุด (หรือที่เรียกว่าโหนดปลายทาง) ซึ่งแสดงถึงการดำเนินการที่จะเกิดขึ้นอันเป็นผลมาจากชุดของการตัดสินใจ [17] สามารถแสดงแผนผังการตัดสินใจได้ดังภาพที่ 2.4



ภาพที่ 2.4 แผนผังการตัดสินใจของต้นไม้การตัดสินใจ (Decision Tree) [17]

ประโยชน์ของอัลกอริทึมการตัดสินใจคือ โครงสร้างแบบแผนผังลำดับงานที่สามารถอ่านและตีความได้ สิ่งนี้ให้ข้อมูลเชิงลึกเกี่ยวกับวิธีการและสาเหตุที่โมเดลทำงานหรือไม่ทำงานได้ดี ตัวอย่างของการนำไปใช้งาน เช่น การศึกษาการตลาดที่เกี่ยวกับพฤติกรรมของลูกค้า ความพึงพอใจของลูกค้า และการที่ลูกค้าจะกลับมาใช้บริการ หรือการศึกษาทางการแพทย์ และการวินิจฉัยอาการของโรค เป็นต้น [17]

อัลกอริทึมการตัดสินใจ C5.0 มีการใช้งานต้นไม้การตัดสินใจมากมาย แต่หนึ่งในการใช้งานที่เป็นที่รู้จักมากที่สุดคืออัลกอริทึม C5.0 เป็นอัลกอริทึมที่ได้รับการพัฒนาโดยนักวิทยาศาสตร์คอมพิวเตอร์ J. Ross Quinlan เป็นเวอร์ชั่นปรับปรุงของอัลกอริทึม C4.5 ซึ่งเป็นการปรับปรุงให้ดีขึ้นกว่าอัลกอริทึม Iterative Dichotomiser 3 (ID3) อัลกอริทึม C5.0 ได้กลายเป็นมาตรฐานของต้นไม้การตัดสินใจ เนื่องจากสามารถใช้งานได้ดี ง่ายต่อการใช้งานและการตีความเข้าใจ เมื่อเทียบกับโมเดลการเรียนรู้ของเครื่องขั้นสูงอื่น ๆ สำหรับการวัดเพื่อระบุตัวเลือกการแบ่งของต้นไม้การตัดสินใจ C5.0 จะใช้เอนโทรปี (Entropy) ที่เป็นแนวคิดทฤษฎีการสุ่มหรือการหาความผิดปกติในชุดของคลาส โดยเอนโทรปีที่สูงจะมีความหลากหลายมากซึ่งจุดประสงค์ของต้นไม้การสินใจเพื่อหาการแบ่งและลดเอนโทรปี ทั่วไปเอนโทรปีมีค่าเพียงสองค่าคือ 0 ถึง 1 สำหรับ n คลาส เอนโทรปีช่วงระหว่าง 0 ถึง $log_2(n)$ ในแต่ละกรณี โดยค่าต่ำสุดบ่งชี้ว่าตัวอย่างมีความหลากหลายน้อย ในขณะที่มีค่าสูงหมายถึงตัวอย่างมีความหลากหลายมาก [17] คำนวณได้ตามสมการที่ (1)

$$Entropy(S) = \sum_{i=1}^{c} -p_i log2(p_i) \quad (1)$$

S คือ ชุดของข้อมูล
C คือ จำนวนของกลุ่มตัวแปรตาม
p_i คือ ความน่าจะเป็นของตัวแปรตาม I ต่อตัวแปรตามทั้งหมด
i คือ กลุ่มของตัวแปรตามทั้งหมด c กลุ่ม

ในการใช้เอนโทรปีเพื่อกำหนดคุณลักษณะที่เหมาะสม จะคำนวณการเปลี่ยนแปลงให้มีความหลากหลายลดลง ซึ่งเป็นการวัดที่เรียกว่าเกนสารสนเทศ (Information gain) สำหรับคุณลักษณะ F คำนวณจากความแตกต่างระหว่างเอนโทรปีในส่วนก่อนการแบ่งส่วน (S_1) และพาร์ติชันที่เกิดจากการแยก (S_2) [17] คำนวณได้ตามสมการที่ (2)

$$InfoGain(F) = Entropy(S_1) - Entropy(S_2) \quad (2)$$

ความซับซ้อนอย่างหนึ่งคือหลังจากแบ่งส่วนข้อมูลแล้ว ข้อมูลจะถูกแบ่งออกเป็นพาร์ติชันมากกว่าหนึ่งพาร์ติชัน ดังนั้น ฟังก์ชันในการคำนวณเอนโทรปี (S_2) จึงต้องพิจารณาเอนโทรปีรวมจากพาร์ติชันทั้งหมด ทำได้โดยชั่งน้ำหนักเอนโทรปีของแต่ละพาร์ติชันตามสัดส่วนของเร็ก-คอร์ดในพาร์ติชัน [17] สามารถระบุเป็นสมการที่ (3)

$$Entropy(S) = \sum_{i=1}^{n} w_i \, \text{Entropy}(p_i) \tag{3}$$

เอนโทรปีทั้งหมดที่เกิดจากการแบ่งส่วนคือผลรวมของเอนโทรปีของแต่ละพาร์ติชัน n พาร์ติชันที่ถ่วงน้ำหนักตามสัดส่วนของตัวอย่างที่อยู่ในพาร์ติชัน (w_i) [17]

2.1.5 การเรียนรู้หลายแบบร่วมกัน (Ensemble Learning)

เป็นการเรียนรู้ที่ใช้รูปแบบการเรียนรู้หลายแบบร่วมกัน ในการจำแนกประเภท โดยใช้เสียงข้างมากเพื่อเลือกคำตอบ ด้วยวิธีนี้ความผิดพลาดจากตัวแยกประเภทฐานบางตัวจะไม่ส่งผลต่อการจัดประเภทตราบใดที่ตัวแยกประเภทฐานส่วนใหญ่ยังสามารถจำแนกประเภทได้อย่างถูกต้อง [20]

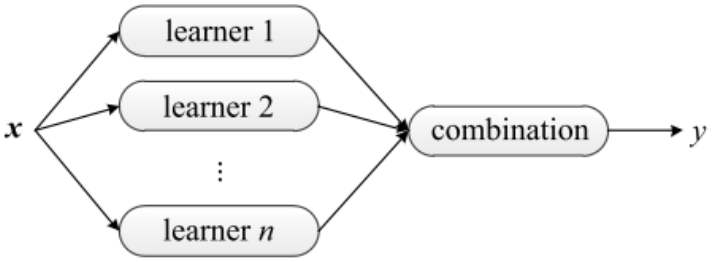
การเรียนรู้หลายแบบร่วมกันทำงานได้ดีเมื่อตัวแยกประเภทฐานมีความหลากหลายหรือตัวแยกประเภทฐานแต่ละตัวมีความสัมพันธ์กันน้อยกว่า เนื่องจากการใช้ตัวแยกประเภทฐานที่ไม่มีความหลากหลายจะมีอัลกอริทึมและโครงสร้างที่เหมือนกันหรือคล้ายคลึงกันที่จะให้แนวโน้มที่จะทำนายผลลัพธ์เดียวกันในการจัดหมวดหมู่ ดังนั้นผลที่ได้จะไม่แตกต่างจากการใช้ตัวแยกประเภทเพียงตัวเดียว การเรียนรู้หลายแบบร่วมกันสามารถแบ่งได้ 4 ประเภท ได้แก่

1. Bootstrap Aggregating or Bagging เป็นการสุ่มตัวอย่างย่อยของข้อมูลการฝึกอบรมและป้อนลงในเวอร์ชั่นต่าง ๆ ของแบบจำลองเดียวกัน และจากนั้นทั้งหมดจะมารวมกันในตอนท้ายเพื่อโหวตผลลัพธ์สุดท้าย

2. Boosting เป็นการสร้างแบบจำลองที่นำเอาจุดอ่อนหรือข้อผิดพลาดของโมเดลก่อนหน้าที่ใช้ข้อมูลการฝึกและทดสอบโมเดลมาหาว่าคุณลักษณะใดที่ทำให้เกิดข้อพลาด จากนั้นจะเพิ่มคุณลักษณะเหล่านั้นในโมเดลต่อ ๆ ไป ซึ่งช่วยลดข้อผิดพลาดของโมเดลได้

3. Bucket of Models เป็นกลุ่มโมเดลที่แตกต่างไปจากเดิมโดยที่พยายามจะคาดการณ์บางสิ่ง โดยอาจจะใช้แบบจำลอง เช่น k-mean, Decision Tree, และ Regression ทั้งสามร่วมกันในชุดข้อมูลการฝึกอบรม และทำการลงคะแนนให้กับผลการจัดหมวดหมู่ขั้นสุดท้าย

4. Stacking เป็นการใช้แบบจำลองหลายตัวในข้อมูล แล้วรวมผลลัพธ์เข้าด้วยกันด้วยวิธีใดวิธีหนึ่ง เพื่อให้ได้ผลลัพธ์สุดท้าย [21] ซึ่งสถาปัตยกรรมของการเรียนรู้หลายแบบร่วมกันสามารถแสดงได้ดังภาพที่ 2.5



ภาพที่ 2.5 สถาปัตยกรรมของการเรียนรู้หลายแบบร่วมกัน [20]

2.1.6 แรนดอมฟอเรสต์ (Random Forest)

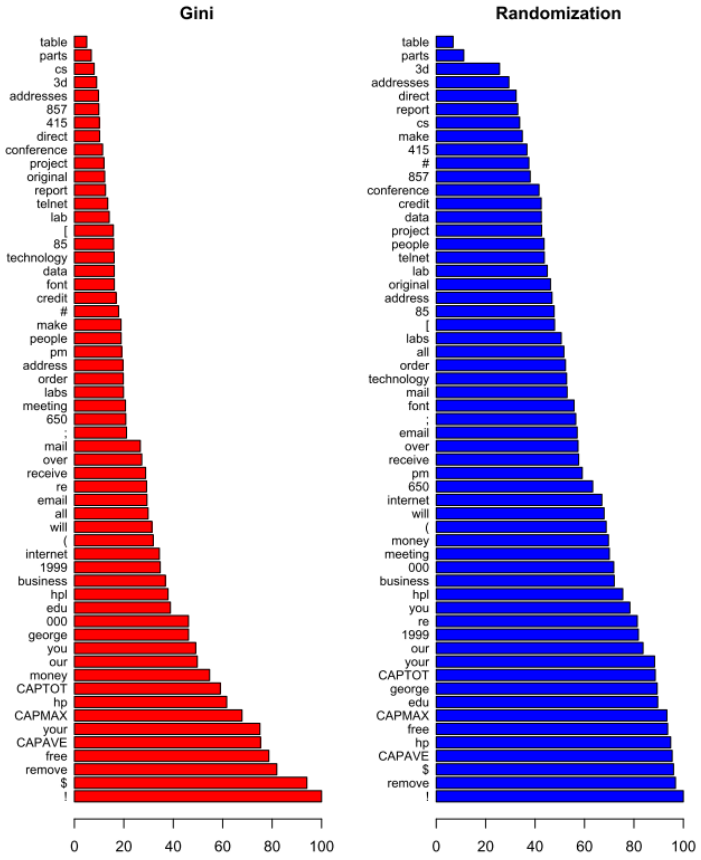
แรนดอมฟอเรสต์ (Random Forest) เป็นอัลกอริทึมการเรียนรู้ที่เรียกว่าการเรียนรู้หลายแบบร่วมกัน มีพื้นฐานจากต้นไม้การตัดสินใจ เนื่องจากต้นไม้การตัดสินใจมีปัญหาเมื่อทำงานกับข้อมูลความผิดปกติ ความผิดปกติดังกล่าวทำให้เกิด Overfitting แรนดอมฟอเรสต์ถูกใช้เพื่อแก้ไขข้อจำกัดนี้ โดยใช้วิธีการเฉลี่ยสำหรับคลาสตัวเลข และการลงคะแนนเสียงข้างมากสำหรับคลาสหมวดหมู่ ซึ่งข้อมูลที่ใช้เป็นข้อมูลการฝึกอบรมในแรนดอมฟอเรสต์จะถูกสุ่มจากข้อมูลทั้งหมดหรือแต่ละแผนผังการตัดสินใจโดยใช้ข้อมูลที่แตกต่างกันในการสร้างและสามารถแก้ปัญหาความผิดปกติของข้อมูลได้ [22], [23]

Out of Bag Samples ในการสร้างแบบจำลองการเรียนรู้ของเครื่อง ข้อมูลจะถูกแบ่งออกเป็นสองส่วนข้อมูลการฝึกอบรมและข้อมูลการทดสอบ หลังจากฝึกอัลกอริทึมการเรียนรู้ของเครื่องด้วยข้อมูลการฝึก ข้อมูลการทดสอบจะใช้เพื่อวัดประสิทธิภาพของแบบจำลอง กระบวนการนี้เป็นกระบวนการตรวจสอบไขว้ ในกรณีที่ข้อมูลขนาดเล็กส่งผลกระทบต่อกระบวนการตรวจสอบข้ามตั้งแต่กระบวนการต้องแบ่งข้อมูลออกเป็นสองส่วน ทำให้ข้อมูลหลังจากแบ่งน้อยเกินไปที่จะใช้สำหรับฝึกหรือทดสอบโมเดลได้อย่างมีประสิทธิภาพ [23]

แรนดอมฟอเรสต์จะใช้วิธี Bootstrapping เพื่อสร้างข้อมูลสำหรับแต่ละรายการต้นไม้ตัดสินใจสุ่ม จากวิธีนี้จะมีข้อมูลบางส่วนที่ไม่ได้ถูกเลือก เรียกว่าข้อมูลเหล่านี้ ซึ่งจากการสร้างแผนผังการตัดสินใจและแรนดอมฟอเรสต์จะใช้ข้อมูลเหล่านี้เป็นข้อมูลทดสอบของแผนผังการตัดสินใจนั้น และรวบรวมข้อผิดพลาดจากแผนผังการตัดสินใจแต่ละชุดของแรนดอมฟอเรสต์ผลลัพธ์ของค่าเฉลี่ยข้อผิดพลาดคือข้อผิดพลาด Out of Bag ซึ่งใช้เป็นตัวบ่งชี้ประสิทธิภาพของแรนดอม-

ฟอเรสต์ โดยหากค่าของข้อผิดพลาด Out of Bag เพิ่มขึ้น แสดงว่าประสิทธิภาพของแรนดอม-ฟอเรสต์ลดลง [23]

แผนภาพความสำคัญ (Variable Importance) แรนดอมฟอเรสต์สามารถระบุคุณสมบัติตัวแปรได้ว่าคุณสมบัติใดมีความสำคัญโดยจะใช้ตัวอย่าง Out of Bag Samples เพื่อสร้างการวัดความสำคัญตัวแปรที่แตกต่างกัน เป็นการวัดของการทำนายของตัวแปรแต่ละตัว เมื่อต้นไม้โตขึ้น ตัวอย่าง Out of Bag Samples จะถูกส่งต่อไปยังต้นไม้ และบันทึกความแม่นยำในการทำนายไว้ จากนั้น ค่าสำหรับตัวแปรจะถูกจัดเรียงแบบสุ่มในตัวอย่าง Out of Bag Samples และคำนวณความแม่นยำอีกครั้ง ความแม่นยำที่ลดลงเป็นผลมาจากการเรียงสับเปลี่ยนนี้จะนำมาเฉลี่ยบนต้นไม้ทั้งหมด และใช้เป็นตัวชี้วัดความสำคัญของตัวแปรในแรนดอมฟอเรสต์ [23] แผนภาพความสำคัญสามารถแสดงได้ดังภาพที่ 2.6



ภาพที่ 2.6 แผนภาพความสำคัญ (Variable Importance) ของแรนดอมฟอเรสต์ [23]

จากภาพที่ 2.6 คือ แผนภาพของแรนดอมฟอเรสต์ แผนภาพด้านซ้ายคือคุณลักษณะสำคัญของ Gini Splitting Index ส่วนแผนภาพด้านขวาใช้การสุ่ม Out of Bag Samples เพื่อคำนวณความสำคัญของตัวแปร สังเกตว่ามีแนวโน้มการกระจายความสำคัญอย่างสม่ำเสมอ [23]

2.1.7 กาเดียนบูทติ้ง (Gradient Boosting)

กาเดียนบูทติ้ง [18] เป็นอีกเทคนิคการเรียนรู้หลายแบบร่วมกันที่ช่วยสร้างชุดต้นไม้ทีละขั้นตอนโดยมีเป้าหมายเพื่อลดฟังก์ชันการสูญเสียเป้าหมายให้น้อยที่สุด เอาต์พุตทั่วไปสามารถคำนวณได้ตามสมการที่ (4)

$$y_E = \sum_i \alpha_i f_i(\bar{x}) \tag{4}$$

$f_i(x)$ เป็นฟังก์ชันที่แสดงถึงการเรียนรู้ที่มีความแม่นยำต่ำ (Weak Learner) อัลกอริทึมนี้ใช้แนวคิดในการเพิ่มโครงสร้างการตัดสินใจใหม่ในแต่ละขั้นตอนโดยเรียนรู้จากข้อผิดพลาดก่อนหน้า เพื่อลดฟังก์ชันการสูญเสีย โดยใช้วิธีลงไปทางที่ชันที่สุด โดยการหาจุดสูงสุดต่ำสุดของฟังก์ชัน (Method of Steepest Descent) คำนวณได้ตามสมการที่ (5)

$$y_E^{n+1} = y_E^n + \alpha_{n+1} f_{n+1}(\bar{x}) \tag{5}$$

หลังจากจากนั้นสมการจะเปลี่ยนเป็นตามสมการที่ (6)

$$y_E^{n+1} = y_E^n + \alpha_{n+1} \sum_i \nabla L(y_{T_i}, y_{E_i}) \; where \; y_{T_i} \; is \; a \; target \; class \tag{6}$$

2.1.8 ไลท์กาเดียนบูทติ้งแมชชีน LightGBM (Light Gradient Boosting Machine)

ไลท์กาเดียนบูทติ้งแมชชีน เป็นเฟรมเวิร์กที่มีประสิทธิภาพสูงของการเรียนรู้หลายแบบร่วมกันโดยใช้ต้นไม้ตัดสินใจ โดยไลท์กาเดียนบูทติ้งแมชชีนจะใช้อัลกอริทึม Leaf-wise คือเป็นการค้นหาตามค่าดีสุดดีที่สุดก่อน (Best-first) เพื่อลดฟังก์ชันการสูญเสียได้มากกว่าแบบ Level-wise ที่เป็นการค้นหาตามแนวลึก (Depth-first) ซึ่งเป็นวิธีที่อยู่ในต้นไม้ตัดสินใจทั่วไป ทำให้อัลกอริทึม Leaf-wise มีความถูกต้อง ความแม่นยำและความเร็วที่มากกว่า อย่างไรก็ตามอัลกอริทึม Leaf-wise ก็มีโอกาสที่จะเกิด Overfitting ได้เมื่อใช้กับข้อมูลที่มีขนาดเล็ก [24], [25] อัลกอริทึมสามารถแสดงได้ดังภาพที่ 2.7

ภาพที่ 2.7 อัลกอริทึม Level-wise และ อัลกอริทึม Leaf-wise [25]

2.1.9 เอ็กซ์ตรีมกาเดียนบูทติ้ง XGBoost (Extreme Gradient Boosting)

เอ็กซ์ตรีมกาเดียนบูทติ้ง [26] เป็นการเรียนรู้หลายแบบร่วมกันที่นำเอาต้นไม้การตัดสินใจหลาย ๆ ต้นที่มีพื้นฐานจากอัลกอริทึมกาเดียนบูทติ้งที่ออกแบบมาให้สามารถปรับขนาดได้สูง วัตถุประสงค์เพื่อลดฟังก์ชันการสูญเสียให้น้อยที่สุด โดยจะเรียนรู้จากความผิดพลาดก่อนหน้าทำให้มีความถูกต้อง ความแม่นยำมากขึ้น ฟังกชันการลดการสูญเสียโดยการควบคุมความซับซ้อนของต้นไม้แสดงได้ตามสมการที่ (7)

$$L_{xgb} = \sum_{i=1}^{N} L(y_{i,}F(\mathbf{x}_i)) + \sum_{m=1}^{M} \Omega(h_m) \tag{7}$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda\|w\|^2 , \tag{8}$$

ตามสมการที่ (8) โดย T คือจำนวนใบของต้นไม้และ w คือคะแนนของผลลัพธ์ของใบ โดยค่าของ γ เป็นค่าที่ควบคุมการเพิ่มลดการสูญเสียในการแยกโหนดภายใน (Internal Node) ซึ่งเทคนิคการสุ่มตัวอย่าง Randomization ของเอ็กซ์ตรีมกาเดียนบูทติ้งสามารถช่วยเพิ่มความเร็วในการฝึกแบบจำลองและช่วยลด Overfitting ของแบบจำลองได้

2.1.10 อคติและความแปรปรวน (Bias and Variance)

อคติและความแปรปรวน [27] เป็นข้อผิดพลาดหลักของการเรียนรู้ของเครื่อง อคติคืออัตราความผิดพลาดของอัลกอริทึมในชุดข้อมูลการฝึกอบรม เมื่อมีชุดข้อมูลที่มีขนาดใหญ่มาก ส่วนความแปรปรวน คืออัตราความผิดพลาดของอัลกอริทึมเมื่อเทียบข้อมูลชุดทดสอบกับชุดข้อมูลฝึกอบรม ดังนั้นการจะนิยามคำจำกัดความเพื่อระบุการเกิด Overfitting และ Underfitting ได้ว่าการเกิด Overfitting จะเกิดจากผลลัพธ์จากอัลกอริทึมที่มีความแปรปรวนสูงแต่อคติต่ำ หมายความว่าอัลกอริทึมนี้ทำงานได้ดีกับข้อมูลชุดฝึกอบรมแต่ไม่สามารถนำไปใช้งานจริงได้ ขณะที่ Underfitting จะเกิดจากผลลัพธ์ที่มีความแปรปรวนต่ำแต่อคติสูง หมายความว่าอัลกอริทึมนี้ทำงานได้ดีกับชุดข้อมูล

การทดสอบมากกว่าชุดข้อมูลฝึกอบรม ซึ่งนิยามทั้งสองนี้เราสามารถระบุได้ว่าแบบจำลองที่ได้เหมาะสมหรือไม่ ส่วนกรณีที่อัลกอริทึมมีผลลัพธ์ที่มีความแปรปรวนต่ำและอคติต่ำ หมายความว่าอัลกอริทึมนี้มีประสิทธิภาพที่ดีสามารถทำงานได้ดีกับทั้งชุดข้อมูลการฝึกอบรมและชุดข้อมูลการทดสอบ การทำความเข้าใจสิ่งนี้ช่วยให้สามารถตัดสินใจได้ว่าการเพิ่มข้อมูลหรือกระบวนการวิธีอื่น ๆ จะช่วยปรับปรุงประสิทธิภาพอย่างเหมาะสมหรือไม่ เราสามารถลดอคติและความแปรปรวนได้โดยใช้เทคนิคต่อไปนี้

2.1.10.1 เทคนิคการลดอคติ หากอัลกอริทึมมีอคติที่สูงสามารถลดอคติได้โดยใช้เทคนิคเหล่านี้

1. เพิ่มขนาดแบบจำลอง เช่น การเพิ่มเลเยอร์หรือเซลล์ประสาทในโครงข่ายประสาทเทียม แม้ว่าวิธีการนี้สามารถลดอคติได้ แต่ก็เป็นสาเหตุของความแปรปรวนที่เพิ่มขึ้นนำไปสู่ปัญหา Overfitting เนื่องจากแบบจำลองพยายามปรับข้อมูลการฝึกให้เหมาะสม ดังนั้น เราต้องใช้การทำให้เป็นมาตรฐาน ซึ่งกำจัดความแปรปรวนที่เพิ่มขึ้นพร้อมกับแนวทางนี้

2. เพิ่มคุณสมบัติเพิ่มเติมที่ช่วยขจัดอคติของอัลกอริทึม เช่นเดียวกับวิธีแรกด้วยวิธีนี้ เราสามารถกำจัดอคติได้ แต่จะเพิ่มความแปรปรวน ดังนั้นควรใช้การทำให้เป็นมาตรฐานเมื่อเราพบว่าการเพิ่มคุณสมบัติเพิ่มเติมทำให้เกิด Overfitting

3. ลดหรือขจัดการทำให้เป็นมาตรฐาน ด้วยวิธีนี้ เราสามารถลดอคติแต่เพิ่มความแปรปรวนได้พร้อม ๆ กัน

4. ปรับเปลี่ยนสถาปัตยกรรมแบบจำลอง ทำให้โมเดลมีความเหมาะสมกับปัญหา

5. เพิ่มข้อมูลการฝึกอบรมเพิ่มเติม วิธีการนี้อาจช่วยลดความแปรปรวนได้มากกว่าผลกระทบต่ออคติ

2.1.10.2 ส่วนเทคนิคการลดความแปรปรวน สามารถลดความแปรปรวนโดยใช้เทคนิคเหล่านี้

1. เพิ่มข้อมูลการฝึกอบรมเพิ่มเติม เป็นวิธีที่ง่ายที่สุดแต่เชื่อถือได้ ข้อเสียของวิธีนี้คือส่งผลต่อพลังประมวลผลในการประมวลผลข้อมูล

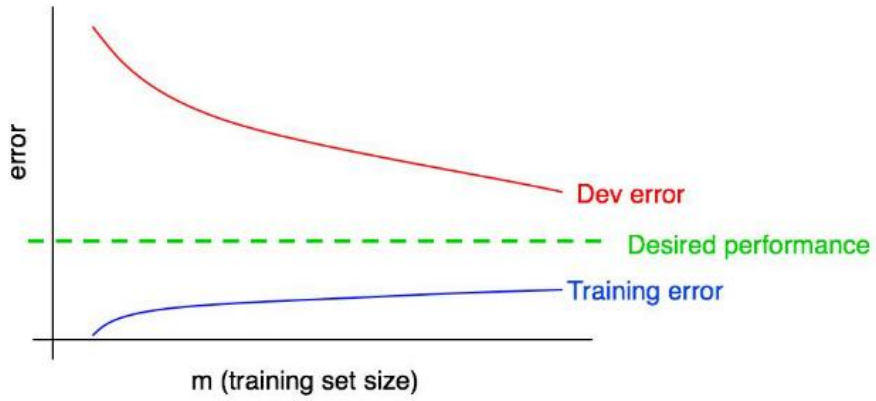
2. เพิ่มการทำให้เป็นมาตรฐาน (L2 regularization, L1 regularization, dropout) แม้ว่าวิธีนี้จะลดความแปรปรวนได้ แต่ก็เป็นสาเหตุให้อคติเพิ่มขึ้น

3. การเลือกคุณสมบัติหรือลดจำนวนคุณสมบัติ วิธีการนี้อาจช่วยลดความแปรปรวนได้ แต่จะเพิ่มอคติในเวลาเดียวกัน

4. ลดขนาดโมเดล วิธีการนี้ลดความซับซ้อนของแบบจำลอง และส่งผลต่อการลดความแปรปรวน แต่เมื่อความซับซ้อนของแบบจำลองลดลง หมายความว่าอคติอาจเพิ่มขึ้น

2.1.11 กราฟวิเคราะห์ประสิทธิภาพการเรียนรู้ (Learning Curves)

กราฟวิเคราะห์ประสิทธิภาพการเรียนรู้ [27] คือกราฟที่แสดงความสัมพันธ์ระหว่างข้อผิดพลาดจากแบบจำลองกับจำนวนชุดการฝึก ในการวาดเส้นการเรียนรู้ เราต้องทดสอบอัลกอริทึมด้วยขนาดชุดการฝึกที่แตกต่างกัน เส้นโค้งการเรียนรู้แสดงให้เห็นว่าเมื่อขนาดชุดการฝึกเพิ่มขึ้น ความแปรปรวนหรือข้อผิดพลาดจากชุดทดสอบควรลดลง หากเราพบว่าเส้นโค้งจากชุดการฝึกมีเส้นที่ราบหมายความว่าการเพิ่มข้อมูลจะไม่ช่วยปรับปรุงอัลกอริทึมการเรียนรู้ การเพิ่มขนาดชุดการฝึกหมายถึงมักจะลดความแปรปรวน แต่จะเพิ่มอคติ ดังนั้น หากเราพลอตข้อผิดพลาดของชุดการฝึกหรืออคติในกราฟการเรียนรู้ มันจะเพิ่มเมื่อขนาดชุดการฝึกเพิ่มขึ้น นอกจากนี้ อัลกอริทึมการเรียนรู้มักจะทำงานได้ดีกับข้อมูลการฝึกอบรมมากกว่าข้อมูลทดสอบ ดังนั้นเส้นโค้งของอคติมักจะอยู่เหนือเส้นโค้งของความแปรปรวน กราฟวิเคราะห์ประสิทธิภาพการเรียนรู้สามารถแสดงได้ดังภาพที่ 2.8



ภาพที่ 2.8 กราฟวิเคราะห์ประสิทธิภาพการเรียนรู้ (Learning Curves) [27]

2.1.12 การวิเคราะห์ข้อมูล (Data Analytics)

การวิเคราะห์ข้อมูล (Data Analytics) [28] คือการนำข้อมูลจากแหล่งต่าง ๆ มาวิเคราะห์ร่วมกันเพื่อปรับปรุงธุรกิจ การตลาด หรือตามวัตถุประสงค์อื่น ๆ ที่ต้องการ แบ่งออกได้เป็น 4 ขั้นตอนดังนี้

1. การวิเคราะห์ข้อมูลแบบพื้นฐาน (Descriptive Analytics) วิเคราะห์เพื่อแสดงผลรายการทางธุรกิจ เหตุการณ์ หรือกิจกรรมต่าง ๆ ที่ได้เกิดขึ้น หรือกำลังจะเกิดขึ้น

2. การวิเคราะห์แบบเชิงวินิจฉัย (Diagnostic Analytics) วิเคราะห์สาเหตุของสิ่งที่เกิดขึ้น ปัจจัยความสัมพันธ์ของตัวแปรต่าง ๆ ที่มีความสัมพันธ์ต่อกันของสิ่งที่เกิดขึ้น

3. การวิเคราะห์แบบพยากรณ์ (Predictive Analytics) การวิเคราะห์เพื่อพยากรณ์สิ่งที่กำลังจะเกิดขึ้นหรือน่าจะเกิดขึ้น โดยใช้ข้อมูลที่เกิดขึ้นแล้วกับแบบจำลองทางสถิติ

4. การวิเคราะห์แบบให้คำแนะนำ (Prescriptive Analytics) การวิเคราะห์ข้อมูลที่ซับซ้อนที่สุด เป็นการพยากรณ์สิ่งต่าง ๆ ที่จะเกิดขึ้น สาเหตุ ข้อดี ข้อเสีย และระยะเวลา พร้อมให้คำแนะนำทางเลือกต่าง ๆ และผลของแต่ละทางเลือก

2.1.13 แดชบอร์ด (Dashboard)

การจัดทำรายงานเพื่อนำเสนอและหน้าจอการติดตาม (BI Report & Dashboard) เป็นชุดเครื่องมือในการจัดทำรายงาน โดยนำข้อมูลที่มีความหลากหลายมาวิเคราะห์ด้วยชุดคำสั่งตามผู้ใช้ต้องการ จะจัดทำการนำเสนอในรูปแบบต่าง ๆ มาช่วยในการตัดสินใจได้ เช่น Dashboard แสดงข้อมูลภาพรวม ในรูปแบบแผนภาพ หรือแสดงสถานะตัวชี้วัดผลการปฏิบัติงานต่าง ๆ ในองค์กรได้ [29] ช่วยให้ผู้ใช้สามารถตรวจสอบประสิทธิภาพทางธุรกิจได้อย่างรวดเร็ว แดชบอร์ดจะดึงและสื่อสารข้อมูลเชิงลึกระดับสูง เช่น วิเคราะห์ความผิดปกติ วิเคราะห์ปัญหา วิเคราะห์แนวโน้ม และวิเคราะห์ข้อมูลขั้นสูง แดชบอร์ดสามารถให้คำตอบในภาพรวมได้อย่างรวดเร็วสำหรับคำถามทางธุรกิจที่สำคัญและเป็นประโยชน์ต่อการตัดสินใจในหลายวิธี เช่น

1. การสื่อสารการดำเนินงานของธุรกิจตามเป้าหมายที่กำหนดไว้
2. การปรับปรุงการรับรู้ข้อมูลสำหรับทุกคน
3. การจัดระเบียบข้อมูลการปฏิบัติงานให้อยู่ในรูปแบบที่มีการจัดการที่ดี
4. การแสดงภาพความสัมพันธ์ที่ซับซ้อนด้วยวิธีที่เข้าใจง่าย

2.1.13.1 ประเภทของแดชบอร์ด

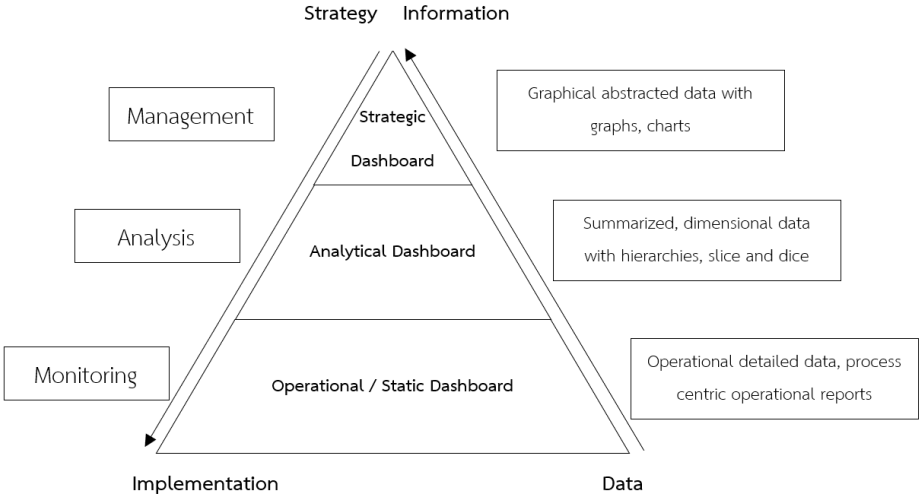
รูปลักษณ์และการประยุกต์ใช้แดชบอร์ดอาจมีรูปแบบที่แตกต่างกันออกไปขึ้นอยู่กับแพลตฟอร์มที่เลือกใช้ กรณีการใช้งานสำหรับแดชบอร์ดแต่ละประเภทยังขึ้นอยู่กับบทบาทหรือชุดทักษะเฉพาะของผู้ใช้ โดยทั่วไปแดชบอร์ดจะอยู่ภายใต้สามประเภทที่แตกต่างกัน ได้แก่

1. แผงควบคุมการทำงาน (Operational Dashboard) เป็นแดชบอร์ดการปฏิบัติงานช่วยให้ผู้ใช้ปลายทางสามารถตรวจสอบกิจกรรมทางธุรกิจ เหตุการณ์ หรือกระบวนการต่าง ๆ ทางธุรกิจในแต่ละวันที่ข้อมูลมีการเปลี่ยนแปลงอยู่ตลอดเวลา ทำให้ผู้ใช้สามารถดำเนินการกับการเปลี่ยนแปลงที่พบได้ทันที และทำการตัดสินใจในระยะสั้นเพื่อเพิ่มประสิทธิภาพ ข้อมูลที่รวบรวมในแดชบอร์ดการปฏิบัติงานนั้นเป็นแบบเรียลไทม์มากกว่า และสะท้อนถึงสิ่งที่เกิดขึ้นในธุรกิจในขณะที่มีการบริโภค

2. แดชบอร์ดเชิงกลยุทธ์ (Strategic Dashboard) หรือที่เรียกว่าแดชบอร์ดสำหรับผู้บริหาร ซึ่งจะให้ภาพรวมโดยย่อของตัวชี้วัดที่ผู้ใช้ต้องการเพื่อตรวจสอบประสิทธิภาพทางธุรกิจ วัตถุประสงค์เพื่อสนับสนุนผู้มีอำนาจตัดสินใจด้วยข้อมูลเชิงลึกเกี่ยวกับความท้าทายหรือโอกาสที่ธุรกิจอาจเผชิญโดยมุ่งเน้นไปที่รายงานสรุประดับสูงของประสิทธิภาพและการคาดการณ์การ

เปลี่ยนแปลงในมาตรการเหล่านั้น โดยทั่วไปแล้วจะสร้างขึ้นจากข้อมูลที่รวบรวมในระยะเวลาต่าง ๆ เช่น มุมมองรายสัปดาห์ รายเดือน หรือรายไตรมาส เพื่อเน้นแนวโน้มและรูปแบบในระยะยาว

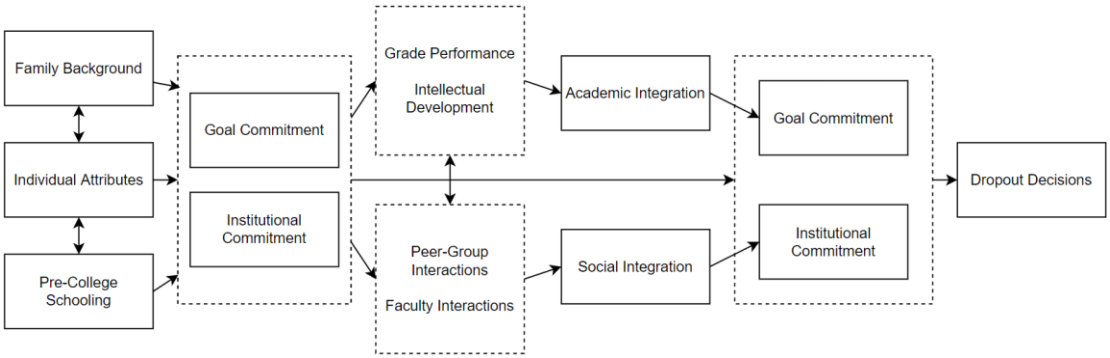
3. แดชบอร์ดวิเคราะห์ (Analytical Dashboard) เป็นแดชบอร์ดการวิเคราะห์ช่วยให้ผู้ใช้สามารถเจาะลึกรายละเอียดของข้อมูลและสนับสนุนการวิเคราะห์สาเหตุที่แท้จริง โดยทั่วไปจะใช้โดยนักวิเคราะห์ธุรกิจและผู้บริหารสายงาน เนื่องจากมีข้อมูลเชิงบริบท การวิเคราะห์เปรียบเทียบ และแนวโน้มในอดีตมากกว่าประเภทอื่น ๆ แดชบอร์ดการวิเคราะห์มีประโยชน์สำหรับการวิเคราะห์การเปลี่ยนแปลงด้วยตนเอง แม้ว่าการค้นพบข้อมูลเชิงลึกจะขึ้นอยู่กับระดับทักษะของผู้ใช้ นอกจากนี้ยังสนับสนุนความสามารถในการวิเคราะห์ขั้นสูงเพิ่มเติม เช่นการตรวจสอบธุรกิจแบบอัตโนมัติและการวิเคราะห์เสริม [30] ประเภทของแดชบอร์ดสามารถแสดงได้ดังภาพที่ 2.9



ภาพที่ 2.9 ประเภทของแดชบอร์ด [31]

2.1.14 การคงอยู่ของนักศึกษาและการออกกลางคัน (Student Retention and Dropout)

การคงอยู่ของนักศึกษา (Student Retention) คือการที่นักศึกษายังคงอยู่และมีการลงทะเบียนในแต่ละภาคการศึกษาจนกว่าจะสำเร็จการศึกษา [32], [33] นอกจากนี้มีแบบจำลองตามทฤษฎีที่มีอิทธิพลของ Vincent Tinto [32] ที่อธิบายปัจจัยที่ส่งผลต่อการคงอยู่ของนักศึกษาคือ ปัจจัยด้านครอบครัว ปัจจัยด้านคุณลักษณะส่วนบุคคล และปัจจัยด้านการศึกษา ดังภาพที่ 2.10



ภาพที่ 2.10 แบบจำลองทฤษฎี Vincent Tinto [34]

การออกกลางคัน (Dropout) สำนักงานคณะกรรมการการศึกษาแห่งชาติได้ให้ความหมายการออกกลางคันว่า คือการที่ผู้เรียนถูกจำหน่ายชื่อออกจากสถานศึกษาในขณะที่ยังไม่สำเร็จการศึกษา โดยไม่ใช่สาเหตุจากการย้ายสถานศึกษา [35]

2.2 งานวิจัยที่เกี่ยวข้อง

ปัญหาการคงอยู่ของนักศึกษาในสถาบันอุดมศึกษาของประเทศไทยจากงานวิจัยพบว่าอัตราการออกกลางคันของนักศึกษาส่วนใหญ่เกิดขึ้นในช่วงชั้นปีแรกที่เข้าศึกษา [36], [37] ปัญหานี้ถูกนำไปสู่การวิจัยเพื่อหาสาเหตุของการออกกลางคันเพื่อลดจำนวนการตกออกของนักศึกษาที่ปัจจุบันประสบปัญหาจำนวนมาก อิทธิพลที่เกิดขึ้นของการวิเคราะห์ทำเหมืองข้อมูลเพื่อการศึกษาและการวิเคราะห์การเรียนรู้ พบว่าเทคนิคการทำเหมืองข้อมูลที่สำคัญคือสถิติ การวิเคราะห์การถดถอย กฎการเชื่อมโยง และการจัดกลุ่ม เป็นเครื่องมือที่เหมาะสมและมีประโยชน์มากในการนำมาพัฒนาใช้สำหรับการคาดการณ์ข้อมูลที่เน้นนักศึกษาเป็นหลักในระดับอุดมศึกษา

2.2.1 Educational data mining and learning analytics for 2 1 st century higher education: A review and synthesis

Aldowah [6] นำเสนอการศึกษางานวิจัยกว่า 402 งานวิจัยที่เกี่ยวกับอิทธิพลของการวิเคราะห์การทำเหมืองข้อมูลในกระบวนการเรียนรู้และผลลัพธ์ของนักเรียนในระดับอุดมศึกษาและทบทวนการทำเหมืองข้อมูลเพื่อการศึกษาและการวิเคราะห์การเรียนรู้โดยครอบคลุมมิติทั้ง 4 ด้าน ได้แก่ การวิเคราะห์การเรียนรู้ที่สนับสนุนด้วยคอมพิวเตอร์ การวิเคราะห์เชิงคาดการณ์ที่สนับสนุนด้วยคอมพิวเตอร์ การวิเคราะห์พฤติกรรมที่สนับสนุนด้วยคอมพิวเตอร์ และการวิเคราะห์การแสดงภาพที่รองรับด้วยคอมพิวเตอร์ พบว่าเป็นเทคนิคที่สามารถแก้ปัญหาการเรียนรู้ได้ และเป็น

ประโยชน์ในการนำไปใช้ในการศึกษาระดับอุดมศึกษาที่เน้นผู้เรียนเป็นหลัก โดยเฉพาะอย่างยิ่งมิติด้านการวิเคราะห์เชิงคาดการณ์ที่สนับสนุนด้วยคอมพิวเตอร์เป็นเทคนิคที่มีประสิทธิภาพในการคาดการณ์รูปแบบที่น่าสนใจเพื่อสร้างแบบจำลองการเรียนรู้ที่เฉพาะเจาะจงได้

2.2.2 A Study of Factors Influency Student Dropout Rate Using Data Mining

Mahatthanachai และคณะ [38] นำเสนอเทคนิคการทำเหมืองข้อมูลที่ศึกษาปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาของคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเชียงใหม่ โดยแบ่งข้อมูลออกเป็นสองชุด โดยข้อมูลชุดแรกเป็นข้อมูลเกี่ยวกับสถาบันการศึกษาเดิมก่อนเข้าศึกษา วิเคราะห์ด้วยเทคนิคต้นไม้การตัดสินใจและอัลกอริทึม C4.5 จำนวน 1,433 คน ผลการวิจัยปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาสามอันดับแรกคือ ผลการเรียนก่อนเข้าศึกษา สาขาก่อนเข้าศึกษา และภูมิหลังทางการศึกษา โดยเทคนิคต้นไม้การตัดสินใจและอัลกอริทึม C4.5 ให้ค่าความถูกต้อง (Accuracy) ร้อยละ 72.02, 70.11 และ 68.13 ตามลำดับ และข้อมูลชุดที่สองจำนวน 2,568 คน ที่ศึกษาเกี่ยวกับผลการเรียนในปัจจุบัน เพื่อกำหนดหลักสูตรที่ส่งผลต่อการออกกลางคันของนักศึกษาโดยใช้เทคนิคกฎความสัมพันธ์และอัลกอริทึม Apriori พบว่าหลักสูตรคอมพิวเตอร์ หลักสูตรภาษาอังกฤษ หลักสูตรคณิตศาสตร์ และหลักสูตรฟิสิกส์เป็นหลักสูตรที่ส่งต่อการอออกลางคันของนักศึกษามากที่สุด

2.2.3 Determinants of University Dropout: A Case of Thailand

Tentsho และคณะ [33] นำเสนอการใช้แบบจำลอง Logistic Regression เพื่ออธิบายปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี ทั้งหมด 7 คณะ จำนวน 10,377 คน ที่ลงทะเบียนเรียนตั้งแต่ปีการศึกษา พ.ศ. 2550 ถึง 2554 พบว่ามีอัตราการออกกลางคันของนักศึกษาอยู่ที่ร้อยละ 23.9 และอัตราการออกกลางคันลดลงตั้งแต่ภาคการศึกษาที่ 2 เป็นต้นไป โดยศึกษาปัจจัยข้อมูลส่วนตัวและปัจจัยด้านวิชาการของนักศึกษา ได้แก่ ปีที่รับสมัคร คณะ เพศ ศาสนา ผลการเรียนเฉลี่ยของภาคเรียนที่ 1 และประเภทการเข้ารับ ผลวิจัยสรุปว่า ปีการศึกษา เพศ ศาสนา คณะ และผลการเรียนเฉลี่ยภาคเรียนแรกมีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติกับการออกกลางคันของนักศึกษา

2.2.4 The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining: A Case Study of Faculty of Science, Prince of Songkla University

Theppalak และคณะ [39] นำเสนอการใช้เทคนิคเหมืองข้อมูลในการวิเคราะห์และคาดการณ์การออกกลางคันของนักศึกษาคณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ ตั้งแต่ปีการศึกษา พ.ศ. 2556 ถึง 2560 ทั้งหมด 4,238 ข้อมูล โดยมี 7 คุณลักษณะ ได้แก่ วิธีการรับเข้าเรียน

วิชาเอก สถานภาพการศึกษา เงื่อนไขการลงทะเบียน ผลการเรียนเฉลี่ยของมหาวิทยาลัย จังหวัดของโรงเรียนมัธยม และผลการเรียนเฉลี่ยระดับมัธยมศึกษาตอนปลาย โดยการเปรียบเทียบแบบจำลองต้นไม้ 3 แบบ ได้แก่ C4.5 (J48), RandomTree, และ REPTree และกฎการอุปนัยได้แก่ OneR, ZeroR, และ Rule-Based Learner (JRip) พบว่า Rule-Based Learner (JRip) ให้ความถูกต้องสูงสุดในการทำนายผลลัพธ์ที่ 77.30% โดยคุณลักษณะปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาส่วนใหญ่เกี่ยวกับผลลัพธ์ทางการศึกษา ได้แก่ ผลการเรียนเฉลี่ยที่ต่ำ รวมถึงวิธีการเข้ารับเข้าเรียนและสาขาวิชา

2.2.5 Student Dropout Prediction: A KMUTT Case Study

นอกจากนี้ยังมีงานวิจัยอื่นที่ใช้การเรียนรู้ของเครื่องประเภทต้นไม้การตัดสินใจ ได้แก่ Decision Tree, Random Forest, และ Gradient Boosting งานวิจัยของ Tenpipat และคณะ [8] นำเสนอโดยใช้อัลกอริทึมต้นไม้ตัดสินใจดังกล่าวในการวิเคราะห์ปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาระดับชั้นปริญญาตรี มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี ตั้งแต่ปีการศึกษา พ.ศ. 2555 ถึง 2562 จำนวน 13,714 คน แบ่งเป็นนักศึกษาที่ออกกลางคันจำนวน 1,436 คน และที่กำลังศึกษาอยู่จำนวน 12,278 ซึ่งแสดงถึงความไม่สมดุลของข้อมูล ผู้วิจัยได้แก้ไขปัญหาดังกล่าวโดยใช้เทคนิค Synthetic Minority Oversampling (SMOTE) ในการสุ่มชุดตัวอย่างข้อมูลเพื่อให้แบบจำลองมีความถูกต้องมากขึ้น และทำการแบ่งกลุ่มตัวแปรข้อมูลเป็น 5 กลุ่มได้แก่ ครอบครัว, โรงเรียน, การเข้ารับการศึกษา, คณะสาขา และกลุ่มข้อมูลส่วนตัว พบว่าแบบจำลองทั้ง 3 ผลลัพธ์ไม่แตกต่างอย่างมีนัยสำคัญ ซึ่งแบบจำลอง Gradient Boosting มีความถูกต้องสูงที่สุด รองลงมาคือแบบจำลอง Decision Tree และ Random Forest โดยให้ค่าความถูกต้อง (Accuracy) ที่ร้อยละ 93, 92 และ 92 ตามลำดับ ซึ่งแบบจำลอง Gradient Boosting มีค่าความถูกต้องและค่าระลึก (Recall) มากที่สุด ส่วนแบบจำลอง Random Forest ให้การคาดการณ์สถานะการออกกลางคันของนักศึกษาได้มากที่สุด และพบว่าปัจจัยที่ส่งผลต่อการออกกลางคัน 5 อันดับแรก ได้แก่ ปีการศึกษา ผลการเรียนเฉลี่ยของโรงเรียนมัธยม ช่องทางการรับเข้ามหาวิทยาลัย คณะ และเพศ

2.2.6 Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques

Yaacob และคณะ [40] นำเสนอการเปรียบเทียบแบบจำลองการจำแนกประเภท 5 อัลกอริทึมได้แก่ k-NN, Decision Tree, Neural Network, Logistics Regression และ Random Forest โดยใช้เทคนิคและกระบวนการทางเหมืองข้อมูลทำการศึกษาข้อมูลนักศึกษาจำนวน 64 คน ภาคการศึกษาที่ 1 และ 2 ปีการศึกษา พ.ศ. 2559 ที่ลงทะเบียนเรียนในหลักสูตรวิทยาการคอมพิวเตอร์ มหาวิทยาลัย Universiti Teknologi MARA Cawangan Kelantan ประเทศมาเลเซีย โดยเน้นศึกษาปัจจัยผลลัพธ์ทางการศึกษา เช่น ผลการเรียนในรายวิชาต่าง ๆ ผลการเรียนสะสม เป็น

ต้น โดยพิจารณาความถูกต้องและค่าพื้นที่ใต้กราฟ พบว่าแบบจำลองส่วนใหญ่มีความถูกต้องสูงกว่าร้อยละ 80 ซึ่งแบบจำลอง Logistic Regression มีค่าความถูกต้องสูงที่สุดคือร้อยละ 90.8 และค่าพื้นที่ใต้กราฟร้อยละ 87.6 ส่วนคุณสมบัติปัจจัยที่สำคัญของผลการเรียนคือ รายวิชา Discrete Mathematics, รายวิชา Object Oriented Programming, รายวิชา Calculus I และ รายวิชา Fundamentals of Data Structures หมายความว่าหลักสูตรเหล่านี้มีผลกระทบมากที่สุดในการทำนายการออกกลางคัน

2.2.7 Dropout Prediction System to Reduce Discontinue Study Rate of Information Technology Students

Limsathitwong และคณะ [41] นำเสนอการใช้เทคนิคเหมืองข้อมูลในการคาดการณ์การออกกลางคันของนักศึกษาของสถาบันเทคโนโลยีไทย-ญี่ปุ่น ตั้งแต่ปี พ.ศ. 2551 ถึง 2558 จำนวน 28,801 ข้อมูล และ 14 คุณลักษณะ โดยเตรียมข้อมูลทำความสะอาดและปรับปรุงประสิทธิภาพของข้อมูล รวมถึงใช้เทคนิคการเลือกคุณสมบัติปัจจัยที่สำคัญ (Feature Selection) ในการคาดการณ์ คือผลการเรียนเฉลี่ยสะสมเนื่องจากมีอิทธิผลต่อการออกกลางคันของนักศึกษามากที่สุด รวมถึงผลการเรียนตามรายวิชาอีกด้วย ผู้วิจัยใช้อัลกอริทึมการจำแนกประเภทคือ ต้นไม้การตัดสินใจอัลกอริทึม C4.5 และทำการเปรียบเทียบประสิทธิภาพแบบจำลองและใช้เทคนิค 10-Fold Cross Validation จากนั้นทำการเปรียบเทียบประสิทธิภาพแบบจำลองจากชุดข้อมูล 5 วิธี ได้แก่ ชุดข้อมูลปกติ ชุดข้อมูลที่ทำการเลือกคุณสมบัติปัจจัยที่สำคัญ ชุดข้อมูลที่จัดการความไม่สมดุลของข้อมูล ชุดข้อมูลที่ใช้เทคนิค Tree Pruning Method และชุดข้อมูลที่ใช้อัลกอริทึม Random Forest ที่เป็นการเรียนรู้หลายแบบร่วมกัน และพิจารณาประสิทธิภาพแบบจำลองด้วยค่าความแม่นยำ (Precision) ค่าระลึก (Recall) และค่าความถ่วงดุล (F1-Measure) พบว่าตัวแบบจำลอง Decision Tree ด้วยชุดข้อมูลที่จัดการความไม่สมดุลของข้อมูลและชุดข้อมูลที่ใช้เทคนิค Tree Pruning Method มีค่าความแม่นยำ ค่าระลึก และค่าความถ่วงดุลเท่ากันที่ 0.80, 0.92 และ 0.85 ตามลำดับ และพบว่าข้อมูลรายวิชาด้านภาษาสำคัญต่อการออกกลางคันมากที่สุด และผู้วิจัยแนะนำว่ากระบวนการเตรียมข้อมูลและคุณภาพความสมบูรณ์ของข้อมูลเป็นสิ่งสำคัญที่ส่งผลต่อประสิทธิภาพของแบบจำลองด้วย

2.2.8 Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile

Palacios และคณะ [11] นำเสนอการศึกษาภาพรวมของการทำเหมืองข้อมูลและการเรียนรู้ของเครื่องต่าง ๆ ได้แก่ Decision Tree, k-NN, Logistics Regression, Naive Bayes, Random Forest และ Support Vector Machine ในการคาดการณ์การออกกลางคันของนักศึกษาในสามระดับได้แก่ ช่วงชั้นปีแรก ชั้นปีที่สอง และชั้นปีที่สาม โดยแบบจำลองส่วนใหญ่ให้ความถูกต้องมากกว่าร้อยละ 80 โดยพบว่าแบบจำลอง Random Forest มีประสิทธิภาพสูงสุด ส่วนคุณลักษณะ

ปัจจัยสำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษาคือ ผลการเรียน การเข้าศึกษา และดัชนีความยากจนเป็นตัวแปรที่สำคัญในการคาดการณ์การออกกลางคันของนักศึกษา

2.2.9 Risk Management Models for Prediction of Dropout Students in Thailand Higher Education

Nuankaew และคณะ [42] นำเสนอการศึกษาการใช้เทคนิคเหมืองข้อมูลด้วยกระบวนการ CRISP-DM (Cross-Industry Standard Process for Data Mining) ในการดำเนินงานวิจัยที่ศึกษาข้อมูลเชิงลึกเกี่ยวกับพฤติกรรมของนักศึกษาจำนวน 2,042 คนที่ลงทะเบียนในหลักสูตรคอมพิวเตอร์ธุรกิจ คณะเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยพะเยา โดยแบ่งการวิเคราะห์ออกเป็น 3 ขั้นตอนคือ ขั้นตอนแรกเป็นการวิเคราะห์พื้นฐานของหลักสูตร ประกอบด้วย จำนวนหน่วยกิต จำนวนรายวิชา เกณฑ์การสำเร็จการศึกษา ขั้นตอนที่สองการสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) และใช้เทคนิคการเลือกปัจจัยในชุดข้อมูล (Feature Selection) และขั้นตอนที่สามคือการทดสอบและประเมินผลแบบจำลอง โดยทำการศึกษาหลักสูตรทั้งหมด 5 หลักสูตร ระหว่างปีการศึกษา พ.ศ. 2544 ถึง 2563 ผลการวิจัยพบว่านักศึกษาส่วนใหญ่ออกกลางคันในปีการศึกษาแรกเนื่องจากผลสัมฤทธิ์ทางการเรียนที่ต่ำ โดยเกี่ยวข้องกับหลักสูตรที่ลงทะเบียนในชั้นปีที่ 1 ซึ่งพบว่ารายวิชา Business Mathematics มีความสำคัญต่อการวางแผนพัฒนาหลักสูตรในอนาคตของหลักสูตรคอมพิวเตอร์ธุรกิจ และพบว่าแบบจำลองที่ใช้เทคนิคเลือกปัจจัยใจชุดข้อมูลและใช้เทคนิค 10-Fold Cross Validation ช่วยให้แบบจำลองมีความถูกต้อง (Accuracy) มากขึ้น โดยแบบจำลองส่วนใหญ่ให้ผลลัพธ์การคาดการณ์กว่าร้อยละ 90

2.2.10 Early Dropout Prediction Model: A Case Study of University Leveling Course Students

นอกจากนี้ยังมีการนำตัวแปรปัจจัยทางวิชาการและด้านเศรษฐกิจและสังคมที่มีผลต่อการออกกลางคัน Sandoval-Palis [10] ได้นำเสนอการศึกษานี้ โดยนำปัจจัยทางดัชนีเศรษฐกิจและสังคม ปัจจัยผลการเรียน และประเภทหลักสูตรทางวิชาการ มาเปรียบเทียบผ่านแบบจำลองการถดถอย (Logistics Regression) และโครงข่ายประสาทเทียม (Neural Network) พบว่าแบบจำลองโครงข่ายประสาทเทียมมีประสิทธิภาพดีกว่า โดยให้ค่าความถูกต้อง (Accuracy) และค่าพื้นที่ใต้กราฟ (AUC Scores) สูงที่สุดคือ 0.768 และ 0.795 ตามลำดับ ซึ่งค่าพื้นที่ใต้กราฟมีค่าสูงกว่าแบบจำลองการถอดถอยที่มีค่า 0.475 โดยแบบจำลองโครงข่ายประสาทเทียมนี้ใช้สถาปัตยกรรม 4-7-1 (4 neurons input layer, 7 neurons hidden layer และ 1 neuron output layer) ในการทดสอบ ซึ่งพบว่าปัจจัยที่สำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษาคือ ผลการเรียนที่ต่ำ และดัชนีเศรษฐกิจและสังคม

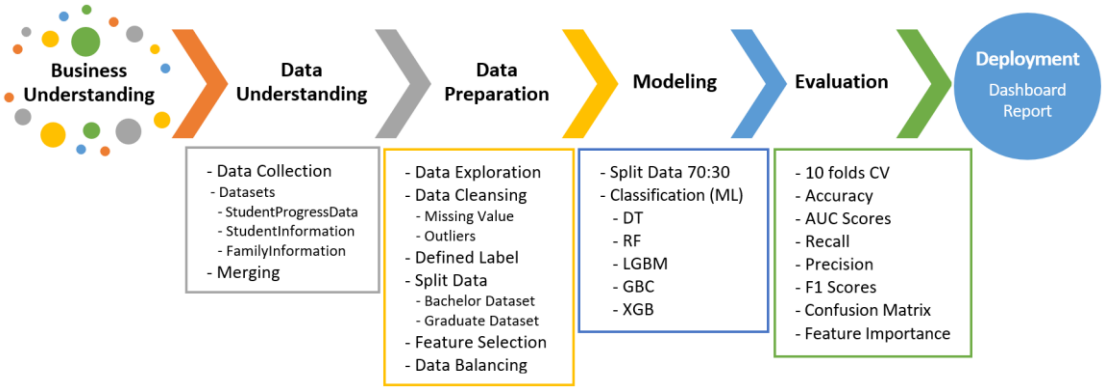
งานวิจัยส่วนใหญ่ศึกษาข้อมูลปัจจัยด้านผลลัพธ์ทางการศึกษาและข้อมูลส่วนตัวของนักศึกษา ซึ่งปัจจัยข้อมูลครอบครัว เช่น อาชีพบิดา-มารดา รายได้บิดา-มารดา รวมถึงปัจจัยทางด้านสุขภาพ ปัจจัยทางการเงิน และการได้รับทุนการศึกษาของนักศึกษาก็ล้วนเป็นปัจจัยสำคัญ งานวิจัยนี้จึงได้นำตัวแปรปัจจัยดังกล่าวมาวิเคราะห์ตามรูปแบบทฤษฎีโมเดลที่มีอิทธิพลของ Vincent Tinto ที่เกี่ยวกับปัจจัยที่ส่งผลต่อการคงอยู่ของนักศึกษา [32] ผ่านการเปรียบเทียบอัลกอริทึมการเรียนรู้ของเครื่องประเภทต้นไม้ 5 แบบ นำมาทดสอบและเปรียบเทียบประสิทธิภาพและความเหมาะสมในการนำไปใช้งาน และนำแบบจำลองที่ได้ไปทำการคาดการณ์กับข้อมูลจริงและแสดงผลการวิเคราะห์นำเสนอรายงานแดชบอร์ดในรูปแบบจินตทัศน์เพื่อติดตามความเสี่ยง ซึ่งจะช่วยให้เจ้าหน้าที่ที่เกี่ยวข้องสามารถเข้าช่วยเหลือนักศึกษาที่มีความเสี่ยงได้ทันที และเพื่อช่วยผู้บริหารในการสนับสนุนการตัดสินใจและวางแผนการบริหารงานเพื่อลดอัตราการออกกลางคันในแต่ละปีการศึกษาของมหาวิทยาลัยให้ต่ำลงได้

บทที่ 3

วิธีดำเนินงานวิจัย

งานวิจัยนี้เป็นการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล และอัลกอริทึมการเรียนรู้ของเครื่องประเภทต้นไม้การตัดสินใจ 5 แบบ ในการค้นหาปัจจัยที่ส่งผลต่อการออกกลางคันและสร้างแบบจำลองคาดการณ์การออกกลางคันของนักศึกษา และนำเสนอรายงานแดชบอร์ดในรูปแบบจินตทัศน์ ในระดับปริญญาตรี และระดับบัณฑิตศึกษา มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ รุ่นปีการศึกษา พ.ศ. 2558 ถึง 2563

ในการศึกษาวิจัยนี้ใช้เทคนิคการทำเหมืองข้อมูล คือการนำข้อมูลมาเพื่อค้นหารูปแบบความสัมพันธ์เพื่อหาผลลัพธ์หรือประโยชน์ โดยใช้กระบวนการที่เรียกว่า Cross Industry Standard Process for Data Mining (CRISP-DM) ประกอบด้วย 6 ขั้นตอนได้แก่ 1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modeling 5. Evaluation และ 6. Deployment [43] โดยใช้ภาษาโปรแกรมไพทอน (Python Programming Language) และเครื่องมือ Google Colab Notebook ในการนำเข้าข้อมูล จัดการข้อมูล สร้างแบบจำลอง และวัดประสิทธิภาพแบบจำลองเพื่อนำไปใช้งาน กระบวนการทั้งหมดแสดงตามกรอบแนวคิดการวิจัยดังภาพที่ 3.1



ภาพที่ 3.1 กรอบแนวคิดการวิจัย

3.1 การทำความเข้าใจปัญหาและความต้องการ (Business Understanding)

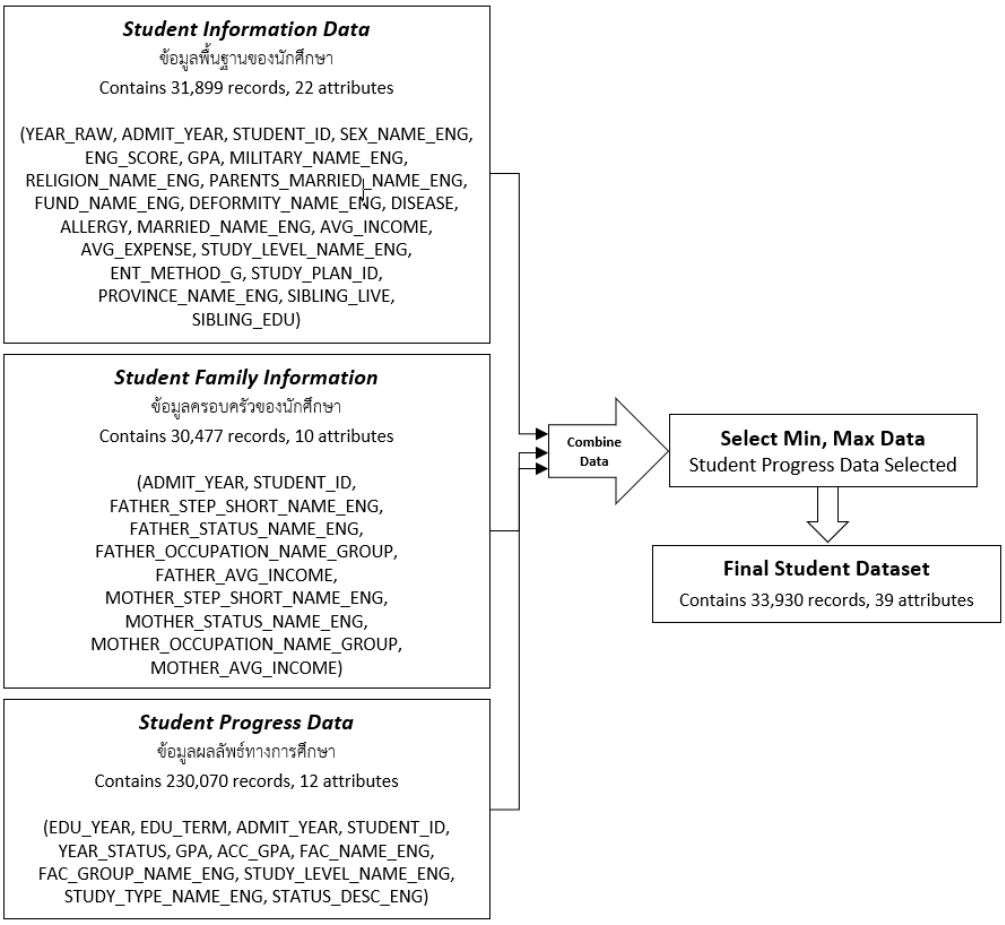
อัตราการออกกลางคันของนักศึกษายังคงมีจำนวนมาก ซึ่งส่วนใหญ่เกิดขึ้นในช่วงชั้นปีแรก ส่งผลให้เกิด “ความสูญเปล่าในการลงทุนเพื่อการศึกษา” สถานศึกษาเสียเวลาและทรัพยากรในการลงทุนบริหารจัดการ ส่วนผู้เรียนเสียเวลาและเงินทองรวมถึงเสียขวัญและกำลังใจในการกลับไปเริ่มต้นใหม่ การคงอยู่ของนักศึกษาเป็นหนึ่งความสำคัญต่อชื่อเสียง ความน่าเชื่อถือของสถาบัน และสวัสดิภาพทางการเงิน วิธีที่จัดการปัญหาดังกล่าวอย่างมีประสิทธิภาพคือการวิเคราะห์และการนำเสนอข้อมูล โดยใช้เทคนิคการทำเหมืองข้อมูลและการเรียนรู้ของเครื่องเพื่อปรับปรุงกระบวนการตัดสินใจวางแผนการบริหารจัดการเพื่อก้าวสู่มหาวิทยาลัยยุคใหม่ การคาดการณ์ผลลัพธ์ที่แม่นยำของการคงอยู่ของนักศึกษาเพื่อให้สถาบันการศึกษาสามารถใช้ข้อมูลเพื่อช่วยเหลือด้านวิชาการแก่นักศึกษาที่มีความเสี่ยง ซึ่งในงานวิจัยนี้ผู้วิจัยได้ทำการศึกษาและทำความเข้าใจทฤษฎีที่เกี่ยวข้องของการออกกลางคันของนักศึกษา ได้แก่ ทฤษฎีโมเดลของ Vincent Tinto ที่เกี่ยวกับปัจจัยที่ส่งผลต่อการคงอยู่ของนักศึกษา [32] คือ คุณลักษณะส่วนบุคคล การศึกษา และครอบครัว และศึกษาการใช้เทคนิคการทำเหมืองข้อมูลและการเรียนรู้ของเครื่อง จุดมุ่งหมายเพื่อหาคุณลักษณะปัจจัยที่ส่งผลต่อการออกกลางคัน สร้างแบบจำลองคาดการณ์ และออกแบบรายงานแดชบอร์ดในรูปแบบจินตทัศน์เพื่อรายงานติดตามความเสี่ยง ซึ่งจะช่วยให้เจ้าหน้าที่ที่เกี่ยวข้องสามารถเข้าช่วยเหลือนักศึกษาที่มีความเสี่ยงได้ทันที และเพื่อช่วยผู้บริหารในการสนับสนุนการตัดสินใจและวางแผนการบริหารงานเพื่อลดอัตราการออกกลางคันในแต่ละปีการศึกษาของมหาวิทยาลัยให้ต่ำลงได้

3.2 การทำความเข้าใจข้อมูล (Data Understanding)

งานวิจัยนี้ได้ขอความอนุเคราะห์ข้อมูลความร่วมมือทางวิชาการและวิจัยจากกองนโยบาย ยุทธศาสตร์ และแผน มหาวิทยาลัยสงขลานครินทร์ โดยเลือกตัวแปรที่ใช้ในการวิจัยจากการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง คือจากปัจจัยผลลัพธ์ทางการศึกษา ปัจจัยข้อมูลพื้นฐานของนักศึกษา และปัจจัยด้านครอบครัวของนักศึกษา ช่วงรุ่นปีการศึกษา พ.ศ. 2558 ถึง พ.ศ. 2563 ส่วนข้อมูลหลักมาจากฐานข้อมูลนักศึกษาใหม่รุ่นปีการศึกษาพ.ศ. 2558 ถึง พ.ศ. 2563 ที่ได้จากฝ่ายทะเบียนและประมวลผล สำนักการศึกษาและนวัตกรรมการเรียนรู้ และกองนโยบาย ยุทธศาสตร์ และแผน ในการศึกษาวิจัยเพื่อการทำวิทยานิพนธ์ครั้งนี้จะไม่นำข้อมูลที่สามารถระบุตัวบุคคลได้ตาม

พ.ร.บ. คุ้มครองส่วนบุคคล เช่น ชื่อ สกุล หมายเลขบัตรประชาชน เบอร์โทรศัพท์ ที่อยู่ อีเมล์มาใช้ในการวิเคราะห์ศึกษาวิจัย

ข้อมูลดิบ (Raw Data) ที่ได้จากกองนโยบาย ยุทธศาสตร์ และแผน เป็นรูปแบบข้อมูล Microsoft Excel ทั้งหมด 3 ไฟล์ ตามขอบเขตเนื้อหาของงานวิจัย คือ 1. ปัจจัยผลลัพธ์ทางการศึกษา (StudentProgressData.xlsx) ประกอบด้วยข้อมูลจำนวน 230,070 ชุด (Observations) และ 12 (Attributes) คุณลักษณะ 2. ปัจจัยข้อมูลพื้นฐานของนักศึกษา (StudentInformation.xlsx) ประกอบด้วยข้อมูลจำนวน 31,899 ชุด และ 22 คุณลักษณะ และ 3. ปัจจัยด้านครอบครัวของนักศึกษา (StudentFamilyInformation.xlsx) ประกอบด้วยข้อมูลจำนวน 30,477 ชุด และ 10 คุณลักษณะ ซึ่งผู้วิจัยใช้โปรแกรม Microsoft Access 2019 ในการนำเข้าข้อมูล สร้างคำสั่งในการเชื่อมโยงมิติของข้อมูล คัดเลือกข้อมูลนักศึกษาที่มีสถานะล่าสุดที่กำลังศึกษาอยู่และพ้นสภาพการเป็นนักศึกษา และส่งออกข้อมูลเป็นไฟล์ Comma Separated Value (CSV) สามารถแจกแจงข้อมูลได้จำนวนทั้งสิ้น 33,930 ชุด และ 39 คุณลักษณะ กระบวนการนำเข้าและเชื่อมโยงมิติของข้อมูลเพื่อให้ได้ข้อมูลสำหรับจัดการในขั้นตอนถัดไป ดังภาพที่ 3.2 และรายละเอียดข้อมูลตามตารางที่ 3.1 และ 3.2



ภาพที่ 3.2 กระบวนการเชื่อมโยงมิติของข้อมูล

ตารางที่ 3.1 ลักษณะข้อมูลประชากรตัวแปรจัดกลุ่ม

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
1	ADMIT_YEAR ปีที่เข้าศึกษา	17.9% 2558 17.6% 2559 17.0% 2560 16.0% 2561 15.9% 2563 15.5% 2562	ปี พ.ศ. ที่นักศึกษาเข้ารับการศึกษา เช่น 2558 = เข้ารับการศึกษาปี พ.ศ. 2558
2	STUDY_STATUS สถานภาพนักศึกษา	84.4% OK 5.1% E 4.4% B 3.6% R 1.5% H 0.6% D 0.3% T 0.1% v 0.0% S 0.0% I 0.0% 7 0.0% AY	สถานภาพนักศึกษา สถานะระบุถึงสถานภาพการเป็นนักศึกษา และพ้นสภาพการเป็นนักศึกษา สถานภาพการเป็นนักศึกษา OK = กำลังศึกษา H = กำลังทำวิทยานิพนธ์ D = ลาพักการศึกษา S = ให้พักการศึกษา AY = รอยืนยันสิทธิ์เข้าศึกษา (นักศึกษาใหม่) พ้นสภาพการเป็นนักศึกษา E = ลาออก (พ้นสภาพฯ) B = ไม่มาลงทะเบียน (พ้นสภาพฯ) R = ตกออก (พ้นสภาพฯ) T = ย้ายคณะหรือสาขาวิชา v = ไม่ผ่านเงื่อนไข (พ้นสภาพ)

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
			I = ถึงแก่กรรม (พ้นสภาพฯ) 7 = ใช้ระยะเวลาการศึกษาเกินกำหนด (พ้นสภาพ)
3	EDU_TERM ภาคการศึกษา	83.7% 2 16.3% 1	ภาคการศึกษาที่นักศึกษากำลังศึกษาอยู่ 1 = ภาคการศึกษาที่ 1, 2 = ภาคการศึกษาที่ 2
4	EDU_YEAR ปีการศึกษา	56.6% 2563 15.2% 2561 14.2% 2562 6.8% 2560 4.7% 2559 2.5% 2558	ปีการศึกษา พ.ศ. รุ่นปีการศึกษาที่นักศึกษากำลังศึกษาอยู่ เช่น 2563 = รุ่นปีการศึกษา พ.ศ. 2563
5	FAC_GROUP_NAME_ENG กลุ่มสาขาวิชา	45.3% Science and Technology Group 33.3% Social Sciences and Humanities Group 21.4% Health Sciences Group	กลุ่มสาขาวิชา จัดกลุ่มตามคณะที่ศึกษา Science and Technology Group = กลุ่มสาขาวิชาวิทยาศาสตร์และเทคโนโลยี Social Sciences and Humanities Group = กลุ่มสาขาวิชาสังคมศาสตร์และมนุษยศาสตร์ Health Sciences Group = กลุ่มสาขาวิชาวิทยาศาสตร์สุขภาพ
6	FAC_NAME_ENG คณะ	17.7% Faculty of Science 17.7% Faculty of Management Sciences 15.8% Faculty of Engineering 7.0% Faculty of Natural Resources	คณะที่สังกัด ตัวอย่างเช่น Faculty of Science = คณะวิทยาศาสตร์ Faculty of Management Sciences = คณะวิทยาการจัดการ

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
		6.6% Faculty of Liberal Arts	Faculty of Engineering = คณะวิศวกรรมศาสตร์
		6.3% Faculty of Medicine	
		5.9% Faculty of Law	
		5.1% Faculty of Nursing	
		3.5% Faculty of Pharmaceutical Sciences	
		3.0% Faculty of Economics	
		2.9% Faculty of Traditional Thai Medicine	
		2.9% Faculty of Agro-Industry	
		1.6% Faculty of Dentistry	
		1.4% Faculty of Medical Technology	
		0.8% Faculty of Environmental Management	
		0.6% International College	
		0.5% Faculty of Veterinary Science	
		0.2% Sino-Thai International Rubber College	
		0.1% Institute for Peace Studies	
		0.1% Public Policy Institute	
		0.1% Marine and Coastal Resources Institute	
		0.1% Graduate School	
		0.1% Health System Management Institute (HSMI)	
		0.0% Faculty of Science and Technology	
7	STUDY_LEVEL_NAME_ENG	84.5% BACHELOR DEGREE	ระดับการศึกษา

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
	ระดับการศึกษา	11.5% MASTER DEGREE 2.8% DOCTOR DEGREE 1.2% HIGHER GRADUATE DIPLOMA	BACHELOR DEGREE = ปริญญาตรี MASTER DEGREE = ปริญญาโท DOCTOR DEGREE = ปริญญาเอก HIGHER GRADUATE DIPLOMA = ประกาศนียบัตรบัณฑิตชั้นสูง
8	STUDY_TYPE_NAME_ENG ประเภทการศึกษา	90.6% Full Time Program 6.5% Part Time Program 2.4% Full Time Program (International) 0.2% Full Time Program (Special) 0.2% Part Time Program (International)	ประเภทการศึกษา Full Time Program = ภาคปกติ Part Time Program = ภาคสมทบ Full Time Program (International) = ภาคปกติ (นานาชาติ) Full Time Program (Special) = ภาคปกติ (พิเศษ) Joint degree = หลักสูตร 2+2
9	YEAR_STATUS ชั้นปีที่	31.2% 1 28.8% 4 25.6% 2 12.3% 3 1.1% 6 1.1% 5	สถานชั้นปีที่กำลังศึกษา 1 = ชั้นปีที่ 1 2 = ชั้นปีที่ 2 3 = ชั้นปีที่ 3 4 = ชั้นปีที่ 4 5 = ชั้นปีที่ 5 6 = ชั้นปีที่ 6
10	STUDENT_ID	Format: Alphanumeric; maximum length 10	รหัสนักศึกษาจำนวน 10 หลักไม่ซ้ำกัน

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
	รหัสนักศึกษา		
11	ALLERGY แพ้ยา	95.2% No 4.8% Yes	มีการแพ้ยา Yes = ใช่ No = ไม่ใช่
12	DEFORMITY_NAME_ENG ความพิการ	99.8% Not disabled 0.1% Physical disability 0.1% Visual disability 0.0% Disability hearing 0.0% Learning Disability	ความพิการ Not disabled = ไม่พิการ Physical disability = ความบกพร่องทางร่างกาย หรือสุขภาพ Learning Disability = มีปัญหาทางการเรียนรู้ Visual disability = ความบกพร่องทางการเห็น Disability hearing = ความบกพร่องทางการได้ยิน
13	DISEASE โรคประจำตัว	83.9% No 16.1% Yes	มีโรคประจำตัว Yes = ใช่ No = ไม่ใช่
14	ENT_METHOD_G ประเภทการเข้ารับ	68.2% Direct system 30.8% Admission 1.0% Other	ประเภทการเข้ารับการศึกษา Direct system = รับตรง Admission = รับแบบ Admission Other = อื่น ๆ
15	FUND_NAME_ENG ได้รับทุน	91.9% No 8.1% Yes	การได้รับทุน Yes = ใช่ No = ไม่ใช่

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
16	MARRIED_NAME_ENG สถานภาพสมรส	97.6% Single 2.0% Married 0.2% Divorced 0.1% Widowed 0.0% Separated	สถานภาพสมรส Single = โสด Married = สมรส Divorced = หย่าร้าง Widowed = หม้าย Separated = แยกกันอยู่
17	MILITARY_NAME_ENG การเกณฑ์ทหาร	64.2% No 35.8% Yes	ผ่านการเกณฑ์ทหาร Yes = ใช่ No = ไม่ใช่
18	PARENTS_MARRIED_NAME_ENG สถานภาพบิดา-มารดา	74.5% Live together 9.0% Divorce 6.9% Father deceased 3.9% Separated for other reasons 2.0% Separated due to career obligation 1.7% Mother deceased 0.7% Father and mother both remarried 0.5% Both father and mother deceased 0.4% Father remarried 0.3% Other 0.2% Mother remarried	สถานภาพบิดา-มารดา Live together = อยู่ด้วยกัน Divorce = หย่าร้าง Father deceased = บิดาถึงแก่กรรม Separated for other reasons = แยกกันอยู่เพราะสาเหตุอื่น Separated due to career obligation = แยกกันอยู่เพราะความจำเป็นเกี่ยวกับอาชีพ Mother deceased = มารดาถึงแก่กรรม Father and mother both remarried = มารดาถึงแก่กรรม

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
			Both father and mother deceased = บิดาและมารดาถึงแก่กรรม Father remarried = บิดาแต่งงานใหม่ Mother remarried = มารดาแต่งงานใหม่ Other = อื่น ๆ
19	PREV_STUDY_LEVEL_NAME_ENG ระดับการศึกษาก่อนหน้า	82.3% Senior High School 14.4% Bachelor Degree 1.8% Master Degree 0.6% High Vocational Certificate 0.3% Vocational Certificate 0.1% Higher Graduate Diploma 0.1% Postgrad Diploma 0.1% Diploma 0.1% Junior High School 0.0% Doctoral Degree 0.0% Graduate Diploma 0.0% Primary Education	ระดับการศึกษาก่อนเข้าศึกษา Senior High School = มัธยมศึกษาตอนปลาย Bachelor Degree = ปริญญาตรี Master Degree = ปริญญาโท High Vocational Certificate = ปวส. Vocational Certificate = ปวช. Higher Graduate Diploma = ประกาศนียบัตรบัณฑิตชั้นสูง Postgrad Diploma = ประกาศนียบัตรสูงกว่าปริญญาตรี Diploma = อนุปริญญา Doctoral Degree = ปริญญาเอก Graduate Diploma = ประกาศนียบัตรบัณฑิต Junior High School = มัธยมศึกษาตอนต้น Primary Education = ประถมศึกษา
20	RELIGION_NAME_ENG	79.2% Buddhism	การนับถือศาสนา

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
	ศาสนา	18.8% Islam 0.9% Christian 0.6% Undefined 0.3% Hinduism 0.1% Other 0.0% Sikhism	Buddhism = พุทธ Islam = อิสลาม Christian = คริสต์ Undefined = ไม่ระบุ Hinduism = ฮินดู Sikhism = ซิกซ์ Other = อื่น ๆ
21	SEX_NAME_ENG เพศ	63.8% Female 36.2% Male	เพศ Female = หญิง Male = ชาย
22	STUDY_PLAN_ID แผนการศึกษา	82.5% H 8.7% F 3.1% G 1.6% A 1.3% I 0.8% E 0.8% C 0.6% D 0.5% J 0.1% B 0.0% L	แผนการศึกษา H = แผนการศึกษา ป.ตรี F = แผน ก แบบ ก2 G = แผน ข A = แบบ 1.1 I = แผนการศึกษาบัณฑิต E = แผน ก แบบ ก1 C = แบบ 2.1 D = แบบ 2.2 J = แผน ป.ตรี(สหกิจ) B = แบบ 1.2

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
			L = ไม่ระบุ
23	PROVINCE_NAME_ENG จังหวัดที่เกิด	36.8% SONGKHLA 9.6% NAKHON SI THAMMARAT 7.3% YALA 6.8% TRANG 4.7% NARATHIWAT 4.7% PHATTHALUNG 4.7% PATTANI 4.5% SURAT THANI 3.9% BANGKOK 2.5% SATUN 2.4% PHUKET 2.1% KRABI 2.0% OTHER 1.3% CHUMPHON 1.2% PHANG-NGA 5.3% Others	จังหวัดที่เกิด เช่น SONGKHLA = สงขลา NAKHON SI THAMMARAT = นครศรีธรรมราช YALA = ยะลา TRANG= ตรัง NARATHIWAT = นราธิวาส Others = จังหวัดอื่น ๆ
24	FATHER_OCCUPATION_NAME_GROUP อาชีพบิดา	23.4% Private Business 20.1% Government Sector 18.9% Agricultural Sector 14.2% Contractor	กลุ่มอาชีพของบิดา Private Business = ธุรกิจส่วนตัว Government Sector = ข้าราชการ Agricultural Sector = เกษตรกร/ประมง

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
		11.8% Not specified 4.3% Others 3.9% Corporate Employee 3.3% Public Enterprise Employee	Contractor = รับจ้าง Not specified = ไม่ระบุ/ว่างงาน Corporate Employee = พนักงานบริษัท Public Enterprise Employee = รัฐวิสาหกิจ Others = อื่น ๆ
25	FATHER_STATUS_NAME_ENG สถานภาพบิดา	91.7% Alive 7.5% Deceased 0.5% Others 0.2% Disabled 0.1% Disability	สถานภาพของบิดา Alive = มีชีวิต Deceased = ถึงแก่กรรม Disabled = พิการ Disability = ทุพพลภาพ Others = อื่น ๆ
26	FATHER_STEP_SHORT_NAME_ENG ระดับการศึกษาบิดา	54.7% Less than bachelor's degree. 23.6% Bachelor Degree 15.2% Unknown 6.2% Master Degree 0.6% Doctoral Degree	ระดับการศึกษาสูงสุดของบิดา Less than bachelor's degree. = ต่ำกว่าปริญญาตรี Bachelor Degree = ปริญญาตรี Master Degree = ปริญญาโท Doctoral Degree = ปริญญาเอก Unknown = ไม่ระบุ
27	MOTHER_OCCUPATION_NAME_GROUP อาชีพมารดา	27.4% Private Business 17.4% Government Sector 17.1% Others	กลุ่มอาชีพของมารดา Private Business = ธุรกิจส่วนตัว Government Sector = ข้าราชการ

ลำดับ	ชื่อคุณลักษณะ	ข้อมูล	ความหมาย
		15.8% Agricultural Sector 10.9% Contractor 6.7% Not specified 3.3% Corporate Employee 1.5% Public Enterprise Employee	Agricultural Sector = เกษตรกร/ประมง Contractor = รับจ้าง Not specified = ไม่ระบุ/ว่างงาน Corporate Employee = พนักงานบริษัท Public Enterprise Employee = รัฐวิสาหกิจ Others = อื่น ๆ
28	MOTHER_STATUS_NAME_ENG สถานภาพมารดา	97.3% Alive 2.3% Deceased 0.2% Others 0.1% Disabled 0.1% Disability	สถานภาพของมารดา Alive = มีชีวิต Deceased = ถึงแก่กรรม Disabled = พิการ Disability = ทุพพลภาพ Others = อื่น ๆ
29	MOTHER_STEP_SHORT_NAME_ENG ระดับการศึกษามารดา	57.5% Less than bachelor's degree. 27.8% Bachelor Degree 9.7% Unknown 4.8% Master Degree 0.3% Doctoral Degree	ระดับการศึกษาสูงสุดของมารดา Less than bachelor's degree. = ต่ำกว่าปริญญาตรี Bachelor Degree = ปริญญาตรี Master Degree = ปริญญาโท Doctoral Degree = ปริญญาเอก Unknown = ไม่ระบุ

ตารางที่ 3.2 ลักษณะข้อมูลประชากรของตัวแปรต่อเนื่อง

ลำดับ	ชื่อคุณลักษณะ	ความหมาย	mean	std	min	25%	50%	75%	max
30	GPA	ผลการเรียนเฉลี่ยปัจจุบัน	2.28	1.43	0.00	1.00	2.80	3.47	4.00
31	ACC_GPA	ผลการเรียนเฉลี่ยสะสม	2.48	1.23	0.00	2.13	2.85	3.37	4.00
32	HSC_GPA	ผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษา	3.11	0.49	0.00	2.80	3.14	3.47	4.00
33	ENG_SCORE	คะแนนภาษาอังกฤษก่อนเข้าศึกษา	31.51	18.89	0.00	21.25	30.00	42.50	97.50
34	AVG_INCOME	รายรับเฉลี่ยนักศึกษา	6,606.97	14,842.77	0.00	1,800.00	4,000.00	6,000.00	1,000,000.00
35	AVG_EXPENSE	รายจ่ายเฉลี่ยนักศึกษา	5,342.37	9,414.69	0.00	2,000.00	3,800.00	6,000.00	999,999.00
36	FATHER_AVG_INCOME	รายได้เฉลี่ยบิดา	33,471.35	58,658.89	0.00	10,000.00	20,000.00	40,000.00	5,000,000.00
37	MOTHER_AVG_INCOME	รายได้เฉลี่ยมารดา	23,963.75	49,800.97	0.00	8,000.00	15,000.00	30,000.00	3,500,000.00
38	SIBLING_LIVE	จำนวนพี่น้อง	2.48	1.25	0.00	2.00	2.00	3.00	33.00
39	SIBLING_EDU	จำนวนพี่น้องที่กำลังศึกษา	1.82	0.94	0.00	1.00	2.00	2.00	21.00

3.3 การจัดการเตรียมข้อมูล (Data Preparation)

งานวิจัยนี้ใช้ภาษาโปรแกรมไพทอน (Python Programming Language) และเครื่องมือ Google Colab Notebook การนำเข้าข้อมูลไฟล์ Comma Separated Value (CSV) ที่ได้ทำการเชื่อมโยงและคัดเลือกข้อมูลแล้ว โดยในขั้นตอนนี้จะทำการจัดเตรียมข้อมูล ตรวจสอบและปรับปรุงคุณภาพความสมบูรณ์ของข้อมูลด้วยวิธีที่เหมาะสมก่อนนำไปทำแบบจำลอง สามารถแบ่งกระบวนการได้ ดังนี้

3.3.1 การทำความสะอาดข้อมูล (Data Cleansing) แบ่งได้เป็น 2 ขั้นตอนคือ

1. การแทนค่าข้อมูลสูญหาย (Imputing Missing Values) ของข้อมูลตัวแปรจัดกลุ่ม (Categorical Variable) โดยให้แทนค่าที่สูญหายด้วยค่า “Unknown” และข้อมูลตัวแปรต่อเนื่อง (Numerical Variable) แทนที่ค่าสูญหายด้วยเทคนิคการเรียนรู้เชิงลึก (Deep Learning) ด้วย DataWig Libraries ที่พัฒนาโดย Amazon ที่อ้างอิงจาก Apache MXNet ด้วยการฝึกอัลกอริทึมการเรียนรู้ของเครื่องและจัดการข้อมูลสูญหายจากการคาดการณ์ค่าที่เป็นไปได้จากตัวแปรทั้งหมดหรือคอลัมน์ทั้งหมดที่ต้องการซึ่งเหมือนกับการแทนค่าสูญหายด้วยเทคนิค Multiple Imputation by Chained Equations (MICE) [44] ดังรายละเอียดตามตารางที่ 3.3 และ 3.4

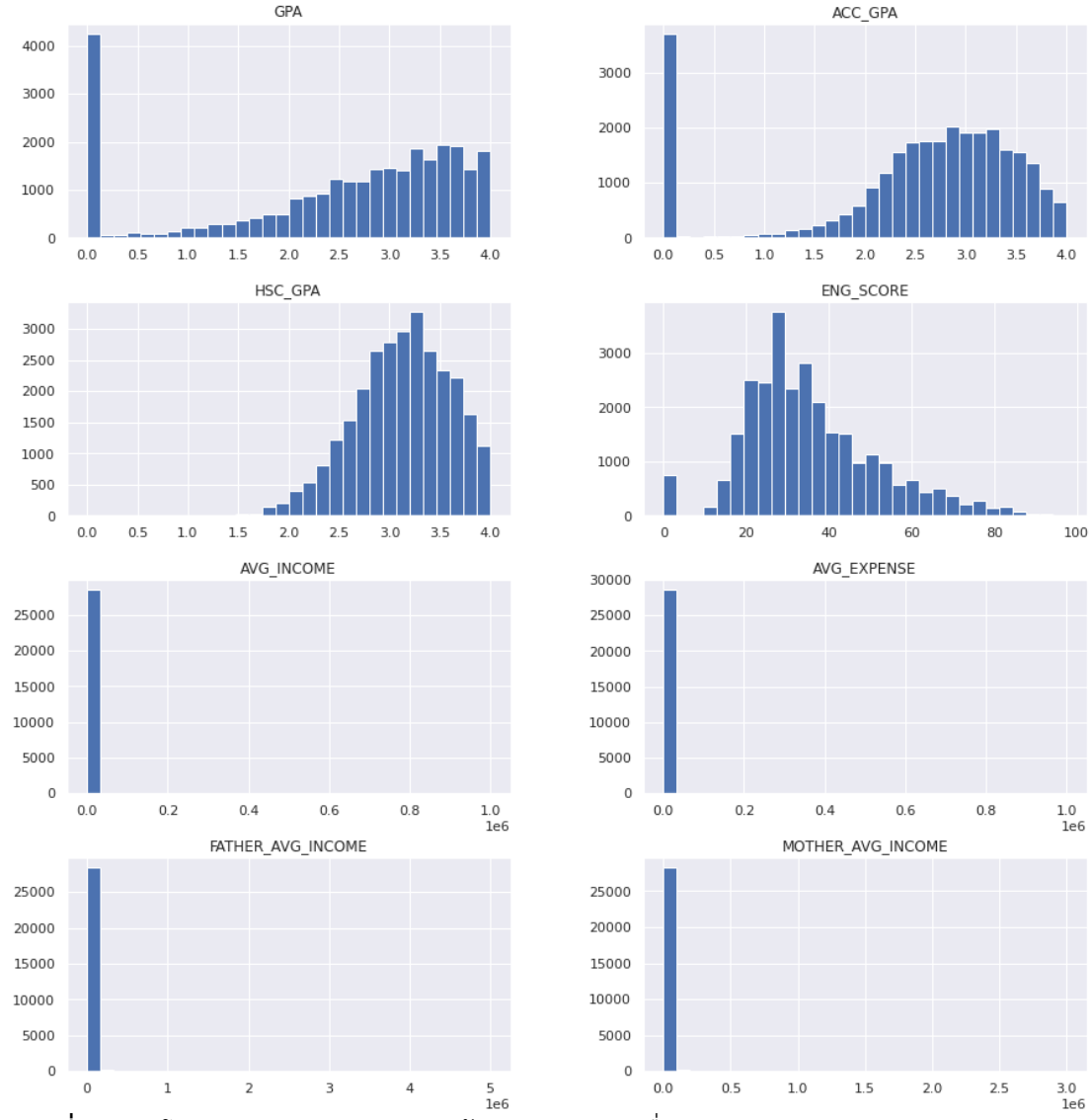
ตารางที่ 3.3 ตรวจสอบค่าสูญหายและกระบวนการแก้ไข

ชื่อคุณลักษณะ	ประเภทข้อมูล	ข้อมูลสูญหาย	% ร้อยละ	วิธีการแทนที่
ENG_SCORE	float64	1,758	5.2	DataWig
MOTHER_AVG_INCOME	float64	1,473	4.3	DataWig
ACC_GPA	float64	747	2.2	DataWig
GPA	float64	747	2.2	DataWig
PREV_STUDY_LEVEL_NAME_ENG	object	621	1.8	“Unknown”
MOTHER_STEP_SHORT_NAME_ENG	object	620	1.8	“Unknown”
AVG_INCOME	float64	613	1.8	DataWig
FATHER_STEP_SHORT_NAME_ENG	object	611	1.8	“Unknown”
MILITARY_NAME_ENG	object	601	1.8	“Unknown”
FATHER_AVG_INCOME	float64	575	1.7	DataWig
MARRIED_NAME_ENG	object	536	1.6	“Unknown”
HSC_GPA	float64	274	0.8	DataWig
FUND_NAME_ENG	object	10	0.0	“Unknown”
ALLERGY	object	4	0.0	“Unknown”
DISEASE	object	3	0.0	“Unknown”
ENT_METHOD_G	object	2	0.0	“Unknown”

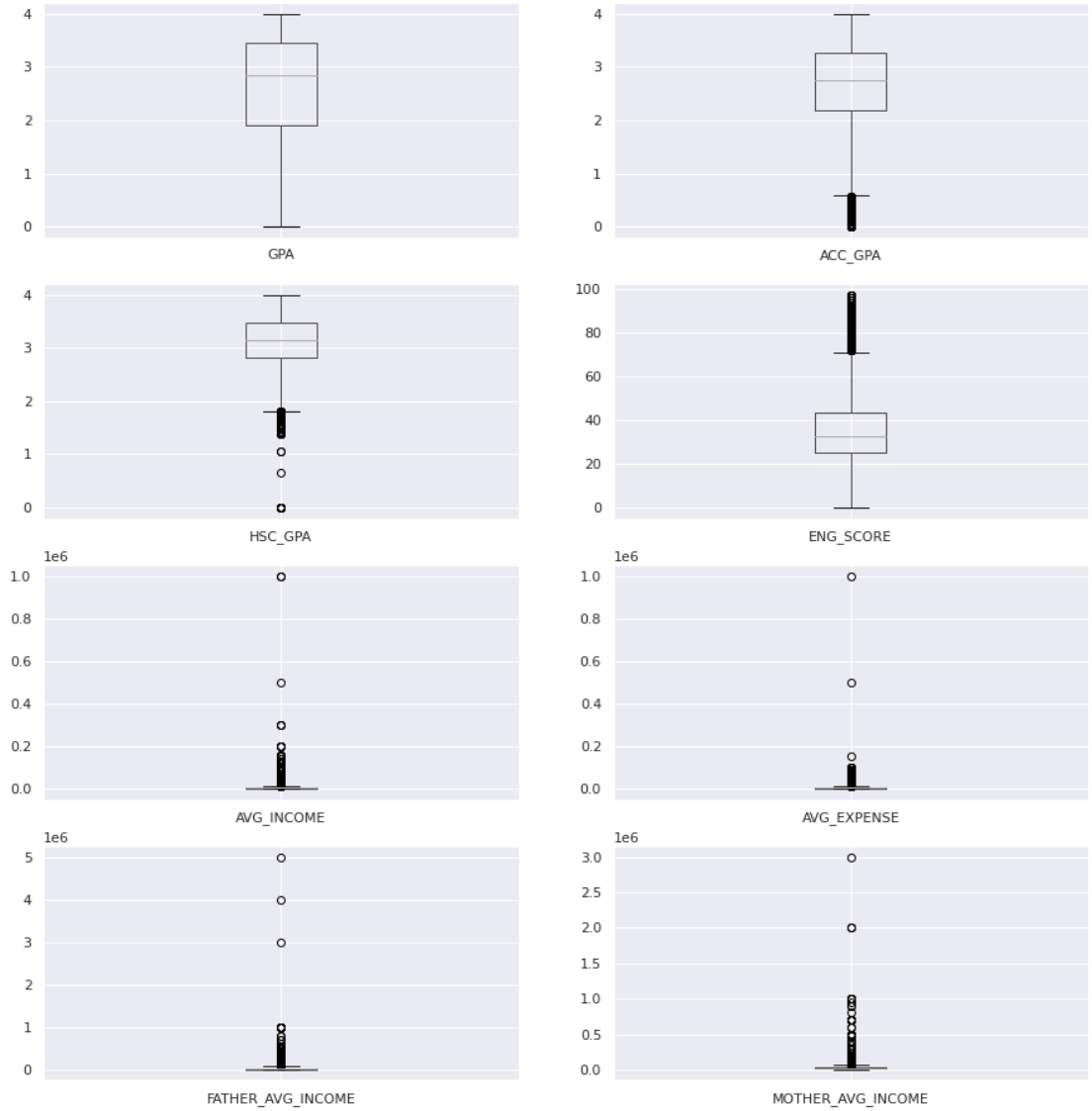
ตารางที่ 3.4 ลักษณะข้อมูลประชากรของตัวแปรต่อเนื่องหลังจากจัดการข้อมูลสูญหาย

ชื่อคุณลักษณะ	mean	std	min	25%	50%	75%	max
GPA	2.28	1.42	0.00	1.10	2.75	3.45	4.00
ACC_GPA	2.48	1.22	0.00	2.15	2.83	3.36	4.00
HSC_GPA	3.11	0.49	0.00	2.81	3.14	3.47	4.00
ENG_SCORE	29.93	19.65	0.00	20.00	29.00	41.25	97.50
AVG_INCOME	6,862.62	14,976.68	0.00	2,000.00	4,000.00	7,000.00	1,000,000.00
AVG_EXPENSE	5,342.37	9,414.69	0.00	2,000.00	3,800.00	6,000.00	999,999.00
FATHER_AVG_INCOME	33,248.48	58,236.70	0.00	10,000.00	20,000.00	40,000.00	5,000,000.00
MOTHER_AVG_INCOME	23,642.73	48,814.21	0.00	8,000.00	15,000.00	30,000.00	3,500,000.00

2. จัดการข้อมูลสุดโต่ง (Handling Outliers) ทำการสร้างฮิสโตแกรม (Histogram) และกราฟรูปกล่อง (Box Plot) เพื่อตรวจสอบข้อมูลค่าสุดโต่งของตัวแปรต่อเนื่อง ดังภาพที่ 3.3 และ 3.4 จากนั้นทำการจัดการข้อมูลสุดโต่งด้วย Interquartile Range (IQR) โดยกำหนดค่า factor k ไว้ที่ 1.5 เท่าของ IQR และแทนที่ค่าข้อมูลด้วยเทคนิคการเรียนรู้เชิงลึก (Deep Learning) ด้วย DataWig Libraries แทนการตัดข้อมูลออกไป และทำการสร้างกราฟเพื่อตรวจสอบข้อมูลอีกครั้ง ผลลัพธ์ดังภาพที่ 3.5 และ 3.6 และรายละเอียดตามตารางที่ 3.5

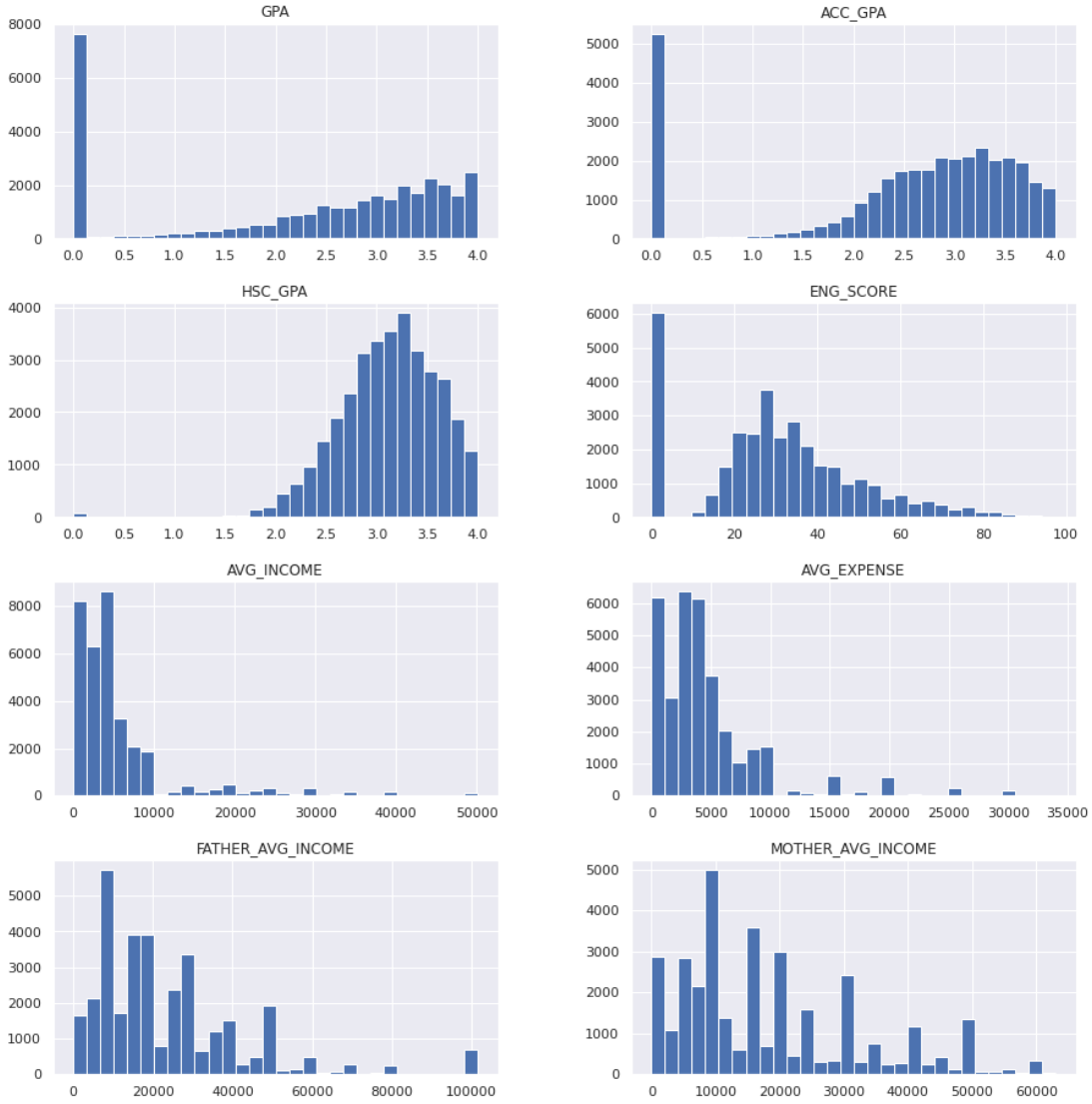


ภาพที่ 3.3 ฮิสโตแกรม (Histogram) ของข้อมูลตัวแปรต่อเนื่อง

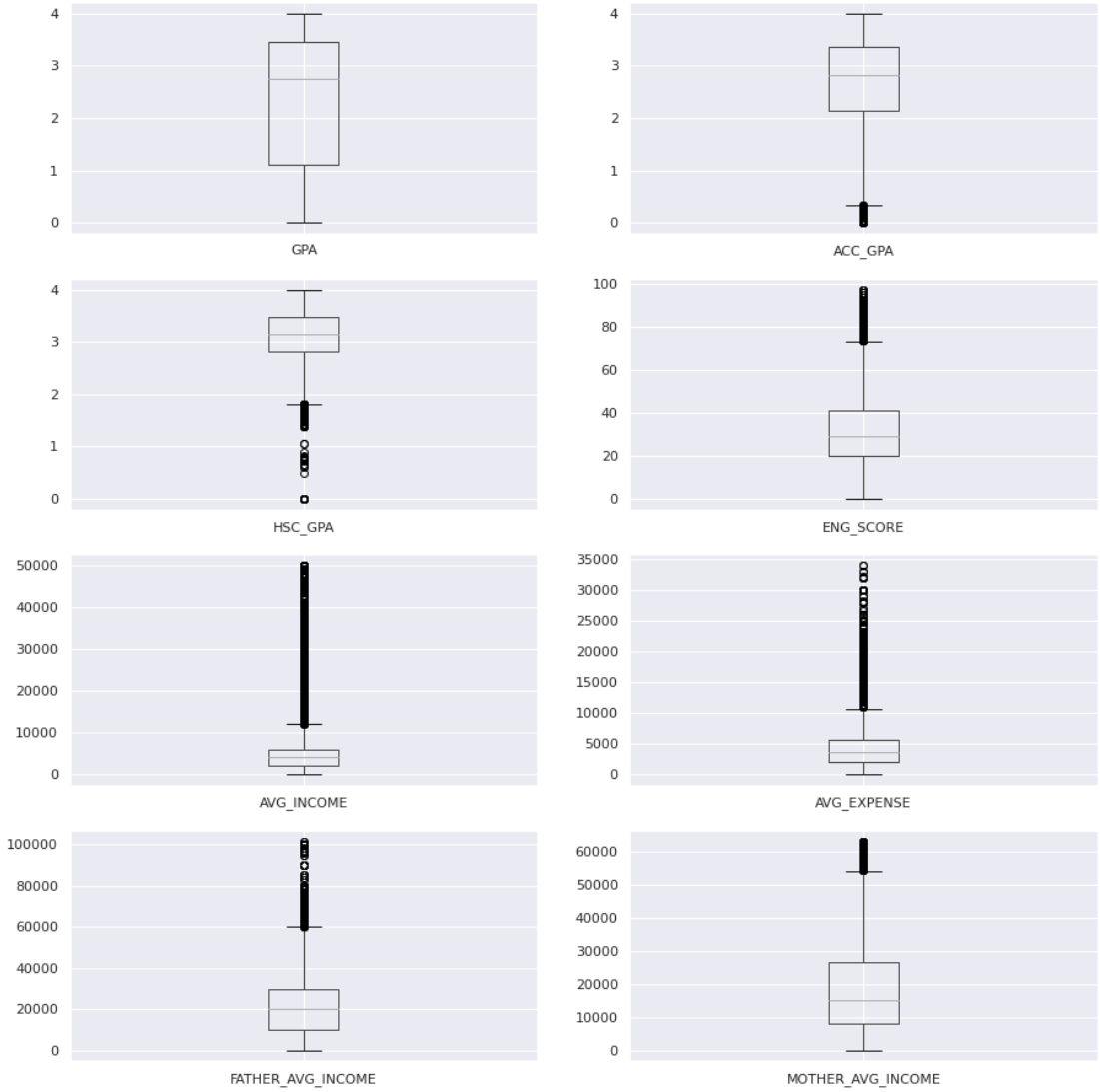


ภาพที่ 3.4 กราฟรูปกล่อง (Box Plot) ของข้อมูลตัวแปรต่อเนื่อง

จากภาพที่ 3.3 และ 3.4 พบว่าตัวแปร AVG_INCOME, AVG_EXPENSE, FATHER_AVG_INCOME และ MOTHER_AVG_INCOME มีค่าสุดโต่งดังแสดงในกราฟรูปกล่องของภาพที่ 14 ผู้วิจัยทำการจัดการค่าสุดโต่งและทำการสร้างกราฟเพื่อตรวจสอบข้อมูลอีกครั้ง ผลลัพธ์ดังภาพที่ 3.5 และ 3.6



ภาพที่ 3.5 ฮิสโตแกรม (Histogram) ของข้อมูลตัวแปรต่อเนื่องหลังจากจัดการค่าสุดโต่ง



ภาพที่ 3.6 กราฟรูปกล่อง (Box Plot) ของข้อมูลตัวแปรต่อเนื่องหลังจากจัดการค่าสุดโต่ง

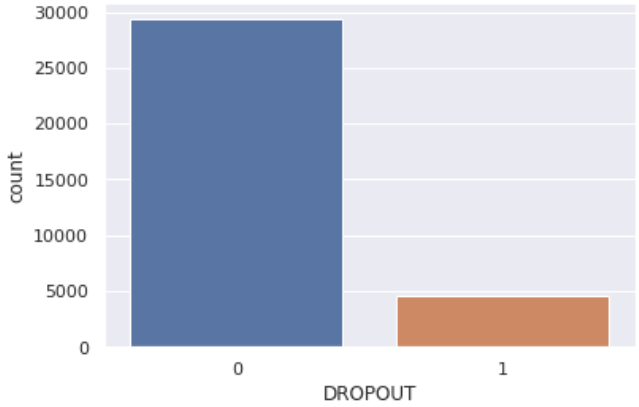
จากภาพที่ 3.6 พบว่าผลการเรียนเฉลี่ยปัจจุบันของนักศึกษาส่วนใหญ่อยู่ในช่วงระหว่าง 1.10 – 3.45 ส่วนผลการเรียนเฉลี่ยสะสมอยู่ในช่วงระหว่าง 2.15 – 3.36 และผลการเรียนเฉลี่ยก่อนเข้าศึกษาอยู่ในช่วงระหว่าง 2.81 – 3.47 และคะแนนภาษาอังกฤษก่อนเข้าศึกษานักศึกษาส่วนใหญ่มีคะแนนในช่วงระหว่าง 20 – 41.25 ส่วนปัจจัยด้านการเงินพบว่ารายรับเฉลี่ยของนักศึกษาส่วนใหญ่อยู่ระหว่าง 2,000 – 6,000 บาท และรายจ่ายเฉลี่ยของนักศึกษาส่วนใหญ่อยู่ระหว่าง 2,000 – 5,500 บาท ส่วนได้รายเฉลี่ยของมารดาส่วนใหญ่อยู่ระหว่าง 8,000 – 26,473.50 บาท และรายได้เฉลี่ยของบิดาส่วนใหญ่อยู่ระหว่าง 10,000 – 30,000 บาท ตามลำดับ ดังรายละเอียดตามตารางที่ 3.5

ตารางที่ 3.5 ลักษณะข้อมูลประชากรของตัวแปรต่อเนื่องหลังจากจัดการข้อมูลสุดโต่ง

ชื่อคุณลักษณะ	mean	std	min	25%	50%	75%	max
GPA	2.28	1.42	0.00	1.10	2.75	3.45	4.00
ACC_GPA	2.48	1.22	0.00	2.15	2.83	3.36	4.00
HSC_GPA	3.11	0.49	0.00	2.81	3.14	3.47	4.00
ENG_SCORE	29.93	19.65	0.00	20.00	29.00	41.25	97.50
AVG_INCOME	5,758.52	7,454.02	0.00	2,000.00	4,000.00	6,000.00	50,100.00
AVG_EXPENSE	4,700.81	4,843.50	0.00	2,000.00	3,500.00	5,500.00	34,000.00
FATHER_AVG_INCOME	24,314.70	19,398.63	0.00	10,000.00	20,000.00	30,000.00	101,610.00
MOTHER_AVG_INCOME	18,278.36	14,223.10	0.00	8,000.00	15,000.00	26,473.50	63,000.00

3.3.2 สร้างตัวแปรเป้าหมาย (Defined Dropout Labels)

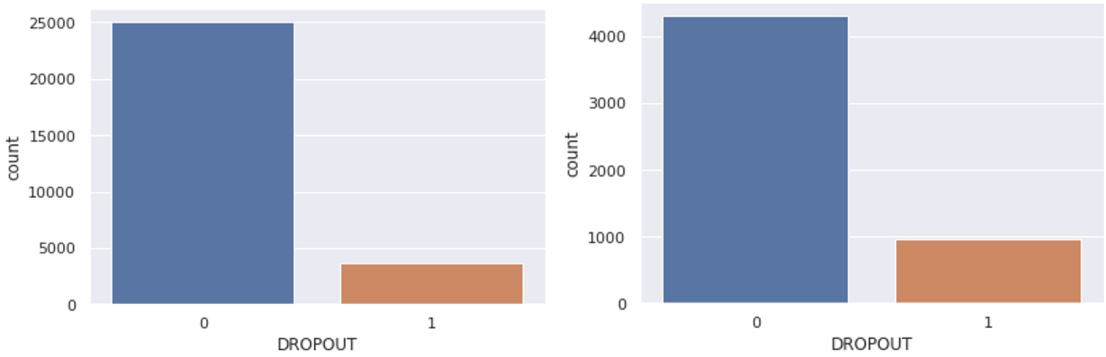
เป็นขั้นตอนการสร้างตัวแปรเป้าหมายสำหรับการคาดการณ์การออกกลางคัน “DROPOUT” โดยแบ่งตามคุณลักษณะของสถานภาพนักศึกษา (STUDY_STATUS) ที่มีค่าเท่ากับ OK, H, D, S, AY โดยกำหนดให้ตัวแปรเป้าหมาย DROPOUT = 0 หมายถึงกำลังศึกษา และ สถานภาพนักศึกษา (STUDY_STATUS) ที่มีค่าเท่ากับ E, B, R, T, v, I, 7 กำหนดให้ DROPOUT = 1 หมายถึงพ้นสภาพการเป็นนักศึกษา พบว่ามีจำนวนนักศึกษาที่กำลังศึกษาทั้งหมด 29,340 คน คิดเป็นร้อยละ 86.50 และพ้นสภาพการเป็นนักศึกษาทั้งหมดจำนวน 4,590 คน คิดเป็นร้อยละ 13.50 ตามลำดับ ดังภาพที่ 3.7



ภาพที่ 3.7 จำนวนสัดส่วนตัวแปรเป้าหมาย DROPOUT

3.3.3 แบ่งชุดข้อมูลตามกลุ่มระดับการศึกษา (Split Data for Degree)

หลังจากทำการสร้างตัวแปรเป้าหมายเรียบร้อยแล้วในขั้นตอนนี้จะทำการแบ่งกลุ่มชุดข้อมูลตามระดับการศึกษา โดยผู้วิจัยจะทำการแบ่งชุดข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลระดับปริญญาตรี และชุดข้อมูลระดับบัณฑิตศึกษา โดยแบ่งกลุ่มจากตัวแปรระดับการศึกษา (STUDY_LEVEL_NAME_ENG) พบว่าชุดข้อมูลระดับปริญญาตรีมีจำนวนนักศึกษาที่กำลังศึกษาทั้งหมด 25,031 คน คิดเป็นร้อยละ 87.30 และพ้นสภาพการเป็นนักศึกษาทั้งหมดจำนวน 3,635 คน คิดเป็นร้อยละ 12.70 และระดับบัณฑิตศึกษามีจำนวนนักศึกษาที่กำลังศึกษาทั้งหมด 4,309 คน คิดเป็นร้อยละ 81.90 และพ้นสภาพการเป็นนักศึกษาทั้งหมดจำนวน 955 คน คิดเป็นร้อยละ 18.10 ตามลำดับ ดังภาพที่ 3.8



ภาพที่ 3.8 จำนวนสัดส่วนตัวแปรเป้าหมาย DROPOUT ชุดข้อมูลระดับชั้นปริญญาตรี (ซ้าย) และชุดข้อมูลระดับชั้นบัณฑิตศึกษา (ขวา)

3.3.4 เลือกปัจจัยในชุดข้อมูล (Feature Selection)

เป็นขั้นตอนการเลือกปัจจัยในชุดข้อมูลที่มีส่วนสำคัญในการคาดการณ์ตัวแปรเป้าหมายมากที่สุด โดยพิจารณาจากค่า P-value ที่มีนัยสำคัญเชิงสถิติ ≤ 0.05 สำหรับข้อมูลตัวแปรจัดกลุ่มด้วยการทดสอบไคสแควร์ (Chi-Square Test) และข้อมูลตัวแปรแบบต่อเนื่องด้วยจากการวิเคราะห์ความแปรปรวน (Analysis of variance: ANOVA) [45] โดยนำตัวแปรที่ไม่มีผลต่อการทดสอบแบบจำลองออกไปแทนที่จะเลือกปัจจัยทั้งหมด ซึ่งจะช่วยลดความแปรปรวนหรือการเกิด Overfitting ของแบบจำลองได้ โดยผู้วิจัยนำเอาคุณลักษณะปัจจัยที่ไม่ส่งผลต่อตัวแปรเป้าหมายออกไป ได้แก่ตัวแปร EDU_YEAR, ADMIT_YEAR, STUDENT_ID, STUDY_STATUS, และตัวแปร STUDY_LEVEL_NAME_ENG โดยชุดข้อมูลระดับปริญญาตรีมีคุณลักษณะปัจจัยที่มีนัยสำคัญเชิงสถิติรวมทั้งหมด 27 คุณลักษณะ (Attributes) ไม่รวมตัวแปรที่ไม่มีผลต่อการทดสอบคือ STUDY_PLAN_ID และชุดข้อมูลระดับบัณฑิตศึกษามีคุณลักษณะปัจจัยที่มีนัยสำคัญเชิงสถิติรวมทั้งหมด 15 คุณลักษณะ (Attributes) ตามรายละเอียดตามตารางที่ 3.6 และ 3.7

ตารางที่ 3.6 คุณลักษณะปัจจัยที่เลือกเพื่อนำไปสร้างแบบจำลองของชุดข้อมูลระดับปริญญาตรี

ลำดับ	ชื่อตัวแปร	คะแนนความสำคัญ p-value
1	EDU_TERM	<0.000000*
2	FAC_NAME_ENG	<0.000000*
3	FAC_GROUP_NAME_ENG	<0.000000*
4	STUDY_TYPE_NAME_ENG	<0.000000*
5	SEX_NAME_ENG	<0.000000*
6	ENT_METHOD_G	<0.000000*
7	ACC_GPA	<0.000000*
8	GPA	<0.000000*
9	YEAR_STATUS	<0.000000*
10	HSC_GPA	<0.000000*
11	ENG_SCORE	<0.000000*
12	PREV_STUDY_LEVEL_NAME_ENG	0.000001*
13	MILITARY_NAME_ENG	0.000002*
14	RELIGION_NAME_ENG	0.000006*
15	MOTHER_AVG_INCOME	0.000030*
16	AVG_EXPENSE	0.000031*
17	MOTHER_STEP_SHORT_NAME_ENG	0.000053*
18	MOTHER_OCCUPATION_NAME_GROUP	0.000130*
19	FATHER_STEP_SHORT_NAME_ENG	0.000249*
20	FATHER_OCCUPATION_NAME_GROUP	0.000415*

ลำดับ	ชื่อตัวแปร	คะแนนความสำคัญ p-value
21	FUND_NAME_ENG	0.001383*
22	SIBLING_LIVE	0.002235*
23	MARRIED_NAME_ENG	0.002856*
24	DISEASE	0.023757*
25	ALLERGY	0.031774*
26	PROVINCE_NAME_ENG	0.039295*
27	FATHER_AVG_INCOME	0.046993*
28	DEFORMITY_NAME_ENG	0.056432
29	AVG_INCOME	0.143542
30	SIBLING_EDU	0.313038
31	FATHER_STATUS_NAME_ENG	0.499834
32	MOTHER_STATUS_NAME_ENG	0.669152
33	PARENTS_MARRIED_NAME_ENG	0.783504

*มีนัยสำคัญเชิงสถิติ 0.05

ตารางที่ 3.7 คุณลักษณะปัจจัยที่เลือกเพื่อนำไปสร้างแบบจำลองของชุดข้อมูลระดับบัณฑิตศึกษา

ลำดับ	ชื่อตัวแปร	คะแนนความสำคัญ p-value
1	ENT_METHOD_G	<0.000000*
2	STUDY_TYPE_NAME_ENG	<0.000000*
3	STUDY_PLAN_ID	<0.000000*
4	FAC_NAME_ENG	<0.000000*
5	FAC_GROUP_NAME_ENG	<0.000000*
6	ACC_GPA	<0.000000*
7	YEAR_STATUS	<0.000000*
8	GPA	<0.000000*
9	HSC_GPA	0.000005*
10	SEX_NAME_ENG	0.000033*
11	FUND_NAME_ENG	0.000671*
12	EDU_TERM	0.020414*
13	DEFORMITY_NAME_ENG	0.022060*
14	SIBLING_LIVE	0.029486*
15	MILITARY_NAME_ENG	0.035407*
16	MOTHER_STEP_SHORT_NAME_ENG	0.060501
17	PREV_STUDY_LEVEL_NAME_ENG	0.090288
18	RELIGION_NAME_ENG	0.249609
19	SIBLING_EDU	0.257473

ลำดับ	ชื่อตัวแปร	คะแนนความสำคัญ p-value
20	FATHER_STEP_SHORT_NAME_ENG	0.347664
21	MOTHER_STATUS_NAME_ENG	0.349676
22	MARRIED_NAME_ENG	0.364980
23	PARENTS_MARRIED_NAME_ENG	0.379906
24	FATHER_STATUS_NAME_ENG	0.513838
25	MOTHER_OCCUPATION_NAME_GROUP	0.538399
26	AVG_EXPENSE	0.559879
27	FATHER_OCCUPATION_NAME_GROUP	0.582225
28	DISEASE	0.593453
29	MOTHER_AVG_INCOME	0.675303
30	ALLERGY	0.770740
31	PROVINCE_NAME_ENG	0.867704
32	AVG_INCOME	0.883946
33	FATHER_AVG_INCOME	0.887056

*มีนัยสำคัญเชิงสถิติ 0.05

3.3.5 การสำรวจข้อมูล (Data Exploration)

สรุปรายละเอียดประชากรของตัวแปรเป้าหมาย DROPOUT ในชุดข้อมูลระดับปริญญาตรี และชุดข้อมูลระดับบัณฑิตศึกษา ซึ่งผู้วิจัยได้สรุปตารางข้อมูลประชากรทั้งหมดของตัวแปรเป้าหมาย DROPOUT ในภาคผนวก ก

3.3.6 การทำข้อมูลให้สมดุล (Balancing Data)

จากชุดข้อมูลระดับปริญญาตรีและระดับบัณฑิตศึกษา พบว่าอัตราส่วนของผู้ที่กำลังศึกษาต่อพ้นสภาพการเป็นนักศึกษา คือ 6.9 : 1 ในชุดข้อมูลระดับปริญญาตรี และ 4.5 : 1 ในชุดข้อมูลระดับบัณฑิตศึกษา ทำให้เกิดความไม่สมดุลของข้อมูล ดังนั้นในการศึกษานี้จึงได้ใช้เทคนิควิธีสังเคราะห์ข้อมูลเพิ่ม Synthetic Minority Oversampling Technique (SMOTE) ในการจัดการข้อมูลที่ไม่สมดุล โดยสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิมด้วยหลักการเพื่อนบ้านที่ใกล้ที่สุด เพื่อให้แบบจำลองมีความถูกต้องมากยิ่งขึ้น [46] และทำการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้และไม่ได้ใช้เทคนิค SMOTE ซึ่งรายละเอียดการทดสอบนี้อธิบายไว้ในส่วนของการประเมินผลแบบจำลอง (Evaluation Model) ในส่วนถัดไป

3.4 การสร้างแบบจำลอง (Modeling)

การสร้างแบบจำลองการเรียนรู้ของเครื่อง ผู้วิจัยใช้เครื่องมือ Google Colab Notebook และ PyCaret Machine Learning Library Version 2.8.10 [47] ซึ่งเป็นไลบรารีการเรียนรู้ของเครื่องแบบโอเพนซอร์สในไพทอนที่ใช้งานง่ายและสะดวก สามารถใช้ทำงานการเรียนรู้ของเครื่องที่ซับซ้อนได้อย่างรวดเร็วและมีประสิทธิภาพโดยใช้โค้ดเพียงไม่กี่บรรทัด [48] ในการสร้างอัลกอริทึมการจำแนกสำหรับการคาดการณ์การออกกลางคันของนักศึกษาจากชุดข้อมูลที่ผ่านการจัดเตรียมข้อมูลเรียบร้อยแล้วในขั้นตอนที่ผ่านมา ซึ่งการสร้างแบบจำลองมี 3 ขั้นตอนดังนี้

3.4.1 การตั้งค่าและกำหนดคุณสมบัติ

ทำการกำหนดตัวแปรเป้าหมาย คือ DROPOUT และแบ่งข้อมูลสำหรับเรียนรู้แบบจำลอง 80 เปอร์เซ็นต์ และสำหรับตรวจสอบแบบจำลอง 20 เปอร์เซ็นต์ โดยมีการแบ่งชุดข้อมูล 2 แบบ ได้แก่

1. ชุดข้อมูลระดับปริญญาตรี แบ่งข้อมูลสำหรับเรียนรู้แบบจำลอง 80 เปอร์เซ็นต์ (22,932 ข้อมูล) และสำหรับตรวจสอบแบบจำลอง 20 เปอร์เซ็นต์ (5,734 ข้อมูล)

2. ชุดข้อมูลระดับบัณฑิตศึกษา แบ่งข้อมูลสำหรับเรียนรู้แบบจำลอง 80 เปอร์เซ็นต์ (4,211 ข้อมูล) และสำหรับตรวจสอบแบบจำลอง 20 เปอร์เซ็นต์ (1,053 ข้อมูล)

3.4.2 การสร้างแบบจำลอง

การศึกษานี้ใช้การเรียนรู้แบบมีผู้สอน และอัลกอริทึมการจำแนกในการสร้างแบบจำลองการเรียนรู้ของเครื่องประเภทต้นไม้ตัดสินใจ 5 แบบ ได้แก่

1. ต้นไม้การตัดสินใจ (Decision Tree)
2. แรนดอมฟอเรสต์ (Random Forest)
3. ไลท์กาเดียนบูทติ้งแมชชีน (Light Gradient Boosting Machine)
4. กาเดียนบูทติ้ง (Gradient Boosting)
5. เอ็กซ์ตรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting)

3.4.3 ปรับปรุงประสิทธิภาพแบบจำลองอัลกอริทึมการเรียนรู้ของเครื่อง

เป็นการค้นหาไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุดโดยอัตโนมัติด้วยอัลกอริทึมค้นหาแบบสุ่มด้วย Optuna [49], [50] เพื่อให้ได้ผลลัพธ์แบบจำลองที่ลดปัญหาการ Overfitting และ Underfitting ซึ่งจะทำการเปรียบเทียบประสิทธิภาพแบบจำลองที่ไม่ได้ปรับปรุงประสิทธิภาพแบบจำลองและปรับปรุงประสิทธิภาพแบบจำลองในผลการวิจัยและอภิปรายผลในบทถัดไป

3.5 การประเมินผลแบบจำลอง (Evaluation)

การประเมินผลแบบจำลองที่ใช้ในงานวิจัยนี้มี 3 ขั้นตอน ได้แก่ 1. การตรวจสอบความถูกต้องของแบบจำลอง (Model Validation) 2. คุณลักษณะที่สำคัญ (Feature Importance) และ 3. แผนภาพ SHapley Additive exPlanations (SHAP)

3.5.1 ตรวจสอบความถูกต้องของแบบจำลอง (Model Validation)

ผู้วิจัยพิจารณากระบวนการประเมินประสิทธิภาพโดยการใช้ 10-Folds Cross Validation ด้วยการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling) [51] คือการแบ่งชุดข้อมูลออกเป็น 10 ชุดในการสร้างและทดสอบแบบจำลองด้วยการสุ่มข้อมูลที่มีความครบถ้วนและครอบคลุม ทำให้ผลลัพธ์มีความน่าเชื่อถือมากขึ้นและลดปัญหา Overfitting ของแบบจำลอง [8] การตรวจสอบความถูกต้องของแบบจำลองมีกระบวนหลายแบบ ซึ่งผู้วิจัยพิจารณากระบวนการเพื่อตรวจสอบแบบจำลองเพื่อประเมินประสิทธิภาพ โดยเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยตารางเมทริกซ์ความสับสน (Confusion Matrix) ตามตารางที่ 3.8

ตารางที่ 3.8 ตารางเมทริกซ์ความสับสน (Confusion Matrix)

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

ค่าความถูกต้อง (Accuracy) คือค่าที่แสดงว่าตัวแบบคาดการณ์สามารถทำนายได้ถูกต้องเป็นร้อยละเท่าไหร่ คำนวณได้จากสมการที่ (9)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

ค่าระลึก (Recall) หรือ ค่าความไว (Sensitivity) คือค่าที่วัดแบบจำลองจากตัวแปรเป้าหมายที่สนใจว่าสามารถคาดการณ์ได้ถูกต้องเป็นร้อยละเท่าไหร่ คำนวณได้จากสมการที่ (10)

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

ค่าความแม่นยำ (Precision) คือค่าที่วัดแบบบจำลองว่าสามารถคาดการณ์คำตอบได้ถูกต้องเป็นร้อยละเท่าไหร่ โดยพิจารณาจากตัวแปรเป้าหมายที่สนใจ คำนวณได้จากสมการที่ (11)

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

ค่าความถ่วงดุล (F1 Scores) คือการวัดค่าระลึกและค่าความแม่นยำที่สรุปด้วยค่าเดียวคือ F1 โดยเป็นค่าเฉลี่ยฮาร์โมนิกของค่าระลึกและค่าความแม่นยำ คำนวณได้จากสมการที่ (12)

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

ค่าพื้นที่ใต้กราฟ (AUC Scores) คือค่าพื้นที่ใต้เส้น Receiver Operating Characteristic Curve (ROC Curve) เพื่อแสดงประสิทธิภาพของแบบจำลองคาดการณ์ ซึ่งกราฟนี้ได้มาจากอัตรา True Positive Rate และ False Positive Rate เป็นวิธีการตรวจสอบที่ใช้ในการประเมินแบบจำลองคาดการณ์หากค่า AUC Scores มีค่าใกล้ 1 มากแสดงว่าแบบจำลองมีประสิทธิภาพในการจำแนกกลุ่มออกจากกันได้ถูกต้อง ซึ่งคำนวณได้จากสมการที่ (13-14)

$$True\ positive\ rate = \frac{TP}{TP + FN} \tag{13}$$

$$False\ positive\ rate = \frac{FP}{TN + FP} \tag{14}$$

3.5.2 คุณลักษณะที่สำคัญ (Feature Importance)

การคาดการณ์การออกกลางคันของนักศึกษา วิธีการประเมินนี้เป็นเทคนิคการวิเคราะห์คุณลักษณะปัจจัยที่สำคัญที่สุดในการคาดการณ์ของแบบจำลองที่สร้างขึ้น

3.5.3 แผนภาพ SHapley Additive exPlanations (SHAP)

เป็นการแสดงภาพที่ระบุคุณลักษณะปัจจัยที่สำคัญที่สุดในเชิงบวกและเชิงลบในการคาดการณ์ของแบบจำลองที่ซับซ้อน โดยการคำนวณ Mean Absolute SHAP (MAS) ในแต่ละทุก ๆ คุณลักษณะปัจจัย ซึ่ง SHAP เป็นเทคนิคที่ค่อนข้างใหม่ที่ใช้สำหรับการประเมินคุณลักษณะในแบบจำลองการเรียนรู้ของเครื่องที่มีความน่าเชื่อถือ [52]

3.6 การนำไปใช้งาน (Deployment)

ผลการวิจัยที่ได้จากการศึกษานี้สามารถนำคุณลักษณะปัจจัย และแบบจำลองที่ได้ไปประยุกต์ใช้สร้างแบบจำลองในการตัดสินใจการคงอยู่ของนักศึกษาในระดับอุดมศึกษาได้ว่า นักศึกษาจะคงอยู่ หรือ ออกกลางคัน โดยผู้วิจัยได้ขอความอนุเคราะห์ข้อมูลความร่วมมือทางวิชาการและวิจัยจากกองนโยบาย ยุทธศาสตร์ และแผน มหาวิทยาลัยสงขลานครินทร์ เพิ่มเติมในส่วนของข้อมูลนักศึกษาในปีการศึกษา พ.ศ. 2564 ภาคการศึกษาที่ 2 มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ จำนวนทั้งสิ้น 18,374 ชุด 39 คุณลักษณะ โดยแบ่งเป็นข้อมูลนักศึกษาระดับชั้นปริญญาตรี จำนวน 16,166 ข้อมูล และระดับชั้นบัณฑิตศึกษาจำนวน 2,208 ข้อมูล มาทำการคาดการณ์ผลลัพธ์จากแบบจำลองที่ได้ผ่าน Google Colab Notebook และ PyCaret ML Libraries จากนั้นทำการส่งออกข้อมูลที่ได้จากการคาดการณ์ใช้เป็นข้อมูลในการพัฒนาออกแบบรายงานแดชบอร์ดในรูปแบบจินตทัศน์เพื่อรายงานติดตามความเสี่ยงด้วย Google Data Studio ซึ่งจะช่วยให้เจ้าหน้าที่ที่เกี่ยวข้องสามารถเข้าช่วยเหลือนักศึกษาที่มีความเสี่ยงได้ทันที และเพื่อช่วยผู้บริหารในการสนับสนุนการตัดสินใจและวางแผนการบริหารงาน ซึ่งรายละเอียดผลลัพธ์นี้จะรายงานในส่วนผลการวิจัยและอภิปรายในส่วนถัดไป

บทที่ 4

ผลการวิจัยและอภิปรายผล

งานวิจัยนี้ได้นำเสนอรูปแบบคุณลักษณะที่สำคัญ และคาดการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ในช่วง 6 รุ่นปีการศึกษา พ.ศ. 2558 ถึง 2563 จากระบบข้อมูลพื้นฐานนักศึกษา และระบบข้อมูลนักศึกษา โดยใช้ข้อมูลผลลัพธ์ทางการศึกษา ข้อมูลส่วนตัวของนักศึกษา และข้อมูลครอบครัวของนักศึกษา ผ่านเทคนิคเหมืองข้อมูลและแบบจำลองการเรียนรู้ของเครื่อง 5 แบบ และนำเสนอรายงานแดชบอร์ดในรูปแบบจินตทัศน์

ในบทนี้จะนำเสนอผลการวิจัยการพัฒนาแบบจำลองเปรียบเทียบประสิทธิภาพและของความเหมาะสมของการนำแบบจำลองไปใช้งาน ระบุคุณลักษณะปัจจัยที่สำคัญ และคาดการณ์การออกกลางคันของนักศึกษาแสดงผลในรูปแบบรายงานแดชบอร์ด โดยแบ่งเป็นขั้นตอนต่าง ๆ ดังนี้

1. ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองคาดการณ์ จากการใช้เทคนิคเลือกคุณสมบัติตัวแปร
2. ผลการพัฒนาแบบจำลองคาดการณ์ชุดข้อมูลระดับชั้นปริญญาตรี
3. ผลการพัฒนาแบบจำลองคาดการณ์ชุดข้อมูลระดับชั้นบัณฑิตศึกษา
4. ผลลัพธ์การแสดงผลจากการนำแบบจำลองคาดการณ์ไปใช้งาน

4.1 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองคาดการณ์ จากการใช้เทคนิคเลือกคุณสมบัติตัวแปร

ในส่วนนี้จะเป็นการนำข้อมูลที่ทำความสะอาดแล้วสร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองคาดการณ์ จากการเลือกตัวแปรทั้งหมด และจากการใช้เทคนิคเลือกคุณสมบัติที่มีนัยสำคัญทางสถิติที่ P-value ≤ 0.05 ของตัวแปรกลุ่มด้วยวิธี Chi-Squared และตัวแปรต่อเนื่องด้วยวิธี ANOVA ของชุดข้อมูลระดับชั้นปริญญาตรีทั้งหมด 33 ตัวแปร พบว่ามีตัวแปรที่ผ่านเกณฑ์ทั้งหมด 27 ตัวแปร ส่วนระดับชั้นบัณฑิตศึกษาทั้งหมด 33 ตัวแปร พบว่ามีตัวแปรที่ผ่านเกณฑ์ทั้งหมด 15 ตัวแปร สามารถเปรียบเทียบผลลัพธ์ของแบบจำลองได้ ตามตารางที่ 4.1 และ 4.2

ตารางที่ 4.1 เปรียบเทียบผลลัพธ์ของแบบจำลองที่ใช้ตัวแปรทั้งหมด และแบบจำลองที่ใช้เทคนิคเลือกคุณสมบัติตัวแปร ของชุดข้อมูลระดับปริญญาตรี

Models	Accuracy	AUC	Recall	Precision	F1
Decision Tree	0.8725	0.7152	0.5047	0.4960	0.4999
Decision Tree (Feature Selection)	0.8724	0.7158	**0.5057**	0.4961	0.5005
Extreme Gradient Boosting	0.8796	0.9103	0.4291	0.5295	0.4739
Extreme Gradient Boosting (Feature Selection)	0.8775	0.9092	0.4219	0.5200	0.4657
Gradient Boosting	0.8937	0.9227	0.4871	0.5985	0.5368
Gradient Boosting (Feature Selection)	0.8939	0.9229	0.4871	0.5995	**0.5371**
Light Gradient Boosting Machine	**0.8941**	0.9243	0.4705	**0.6047**	0.5288
Light Gradient Boosting Machine (Feature Selection)	0.8936	**0.9248**	0.4702	0.6024	0.5277
Random Forest	0.8463	0.8853	0.1783	0.3115	0.2265
Random Forest (Feature Selection)	0.8510	0.8884	0.2128	0.3524	0.2651

ตารางที่ 4.2 เปรียบเทียบผลลัพธ์ของแบบจำลองที่ใช้ตัวแปรทั้งหมด และแบบจำลองที่ใช้เทคนิคเลือกคุณสมบัติตัวแปร ของชุดข้อมูลระดับบัณฑิตศึกษา

Models	Accuracy	AUC	Recall	Precision	F1
Decision Tree	0.7457	0.5923	0.3539	0.3155	0.3331
Decision Tree (Feature Selection)	0.7689	0.6218	**0.3908**	0.3684	0.3777
Extreme Gradient Boosting	0.8126	0.6918	0.3021	0.4663	0.3648
Extreme Gradient Boosting (Feature Selection)	0.8302	0.7338	0.3339	0.5510	0.4131
Gradient Boosting	0.8468	0.7697	0.2729	0.6863	0.3882
Gradient Boosting (Feature Selection)	**0.8494**	**0.7837**	0.2954	**0.6927**	0.4120
Light Gradient Boosting Machine	0.8350	0.7250	0.3113	0.5760	0.4028
Light Gradient Boosting Machine (Feature Selection)	0.8459	0.7627	0.3365	0.6374	**0.4377**
Random Forest	0.7642	0.6233	0.0715	0.1645	0.0991
Random Forest (Feature Selection)	0.8233	0.7453	0.2702	0.5176	0.3522

จากตารางที่ 4.1 และ 4.2 พบว่าการใช้เทคนิคเลือกคุณสมบัติตัวแปรที่มีนัยสำคัญทางสถิติช่วยให้แบบจำลองคาดการณ์มีค่าความถูกต้องและค่าพื้นที่ใต้กราฟมากขึ้น ผู้วิจัยจึงได้นำคุณสมบัติตัวแปรเหล่านี้ไปเปรียบเทียบประสิทธิภาพและความเหมาะสมของการนำไปใช้งานต่อไป

4.2 ผลการพัฒนาแบบจำลองคาดการณ์ชุดข้อมูลระดับชั้นปริญญาตรี

ในส่วนนี้จะนำแบบจำลองการเรียนรู้ของเครื่องประเภทต้นไม้การตัดสินใจทั้ง 5 แบบมาทำการคาดการณ์การออกกลางคันของนักศึกษาระดับชั้นปริญญาตรี โดยใช้เทคนิคเลือกคุณสมบัติตัวแปร การปรับจูนไฮเปอร์พารามิเตอร์ และการใช้เทคนิควิธีสังเคราะห์ข้อมูลเพิ่มของแต่ละแบบจำลอง เพื่อเปรียบเทียบประสิทธิภาพแบบจำลอง และความเหมาะสมของการนำไปใช้งาน โดยแบ่งชุดข้อมูลออกเป็น 10 ชุดในการสร้างและทดสอบแบบจำลอง K-Fold Cross Validation 10 folds จากชุดข้อมูลเรียนรู้ 80% (22,932) และชุดข้อมูลตรวจสอบอีก 20% (5,734)

4.2.1 เทคนิคต้นไม้การตัดสินใจ (Decision Tree)

ในการสร้างแบบจำลอง Decision Tree จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.3

ตารางที่ 4.3 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Decision Tree

Classifier	Hyperparameter
Decision Tree Tuned	ccp_alpha=0.0, class_weight=None, criterion='entropy', max_depth=8, max_features=0.8113419998867496, max_leaf_nodes=None, min_impurity_decrease=6.267288942902784e-05, min_impurity_split=None, min_samples_leaf=5,min_samples_split=5, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=634, splitter='best'
Decision Tree Tuned (SMOTE)	ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=15, max_features=0.7784512253408779, max_leaf_nodes=None, min_impurity_decrease=7.534123633494594e-05, min_impurity_split=None, min_samples_leaf=6, min_samples_split=10, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=634, splitter='best'

ผลการสร้างแบบจำลอง Decision Tree โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง

(Accuracy) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุด คือ 0.8942 และ 0.6017 ตามลำดับ ส่วนแบบจำลองที่ใช้เทคนิค SMOTE มีค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) และค่าความถ่วงดุล (F1 Scores) สูงที่สุดคือ 0.9049, 0.5564 และ 0.5543 ตามลำดับ ตามตารางที่ 4.4

ตารางที่ 4.4 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Decision Tree

Models	Accuracy	AUC	Recall	Precision	F1
Decision Tree Tuned	**0.8942**	0.8960	0.4895	**0.6017**	0.5384
Decision Tree Tuned (SMOTE)	0.8869	**0.9049**	**0.5564**	0.5527	**0.5543**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 91.37 ตามตารางที่ 4.5

ตารางที่ 4.5 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Decision Tree ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Decision Tree	**Studying**	19202	831	19048	985
		Dropout	1235	1664	995	1904
		Accuracy		90.99%		**91.37%**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 89.75 ตามตารางที่ 4.6

ตารางที่ 4.6 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Decision Tree ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Decision Tree	**Studying**	4750	248	4687	311
		Dropout	340	396	308	428
		Accuracy		**89.75%**		89.20%

4.2.2 เทคนิคต้นไม้การตัดสินใจ (Gradient Boosting)

สร้างแบบจำลอง Gradient Boosting จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.7

ตารางที่ 4.7 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Gradient Boosting

Classifier	Hyperparameter
Gradient Boosting Tuned	criterion='friedman_mse', init=None, learning_rate=0.11995895471442317, loss='deviance', max_depth=3, max_features=0.759605214073612, max_leaf_nodes=None, min_impurity_decrease=7.015593498552102e-05, min_impurity_split=None, min_samples_leaf=2, min_samples_split=5, min_weight_fraction_leaf=0.0, n_estimators=87, n_iter_no_change=None, presort='deprecated', random_state=634, subsample=0.8178998325268436, tol=0.0001, validation_fraction=0.1, warm_start=False
Gradient Boosting Tuned (SMOTE)	criterion='friedman_mse', init=None, learning_rate=2.274080158987668e-06, loss='deviance', max_depth=10, max_features=0.48049171102542365, max_leaf_nodes=None, min_impurity_decrease=8.208321540305571e-06, min_impurity_split=None, min_samples_leaf=1, min_samples_split=8, min_weight_fraction_leaf=0.0, n_estimators=241, n_iter_no_change=None, presort='deprecated', random_state=634, subsample=0.696542314452461, tol=0.0001, validation_fraction=0.1, warm_start=False

ผลการสร้างแบบจำลอง Gradient Boosting โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่า

ความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8946, 0.9289, 0.6933 และ 0.6247 ตามลำดับ ส่วนแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความแม่นยำ (Precision) สูงที่สุดคือ 0.6003 ตามตารางที่ 4.8

ตารางที่ 4.8 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Gradient Boosting

Models	Accuracy	AUC	Recall	Precision	F1
Gradient Boosting Tuned	0.8938	0.9222	0.4839	**0.6003**	0.5353
Gradient Boosting Tuned (SMOTE)	**0.8946**	**0.9289**	**0.6933**	0.5688	**0.6247**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 92.09 ตามตารางที่ 4.9

ตารางที่ 4.9 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Gradient Boosting ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Gradient Boosting	**Studying**	19179	854	18849	1184
		Dropout	1391	1508	63	2269
		Accuracy		90.21%		**92.09%**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 89.83 ตามตารางที่ 4.10

ตารางที่ 4.10 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Gradient Boosting ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Gradient Boosting	**Studying**	4757	241	4637	361
		Dropout	366	370	222	514
		Accuracy		89.41%		**89.83%**

4.2.3 เทคนิคต้นไม้การตัดสินใจ (Light Gradient Boosting Machine)

สร้างแบบจำลอง Light Gradient Boosting Machine จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.11

ตารางที่ 4.11 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Light Gradient Boosting Machine

Classifier	Hyperparameter
Light Gradient Boosting Machine Tuned	bagging_fraction=0.9718376845285519, bagging_freq=3, boosting_type='gbdt', class_weight=None, colsample_bytree=1.0, feature_fraction=0.9245315743763174, importance_type='split', learning_rate=0.0063243324427668295, max_depth=-1,min_child_samples=25, min_child_weight=0.001, min_split_gain=0.11971047617659925, n_estimators=294, n_jobs=-1, num_leaves=162, objective=None, random_state=634, reg_alpha=0.00010167628769265196, reg_lambda=6.950895343691504e-09, silent='warn', subsample=1.0, subsample_for_bin=200000, subsample_freq=0
Light Gradient Boosting Machine Tuned (SMOTE)	bagging_fraction=0.5656608696510287, bagging_freq=0, boosting_type='gbdt', class_weight=None, colsample_bytree=1.0, feature_fraction=0.8449653146546208, importance_type='split', learning_rate=0.024861137054985785, max_depth=-1, min_child_samples=45, min_child_weight=0.001, min_split_gain=0.843530636485069, n_estimators=95, n_jobs=-1, num_leaves=67, objective=None, random_state=634, reg_alpha=4.351608464426655e-06, reg_lambda=0.07601458359804605, silent='warn', subsample=1.0, subsample_for_bin=200000, subsample_freq=0

ผลการสร้างแบบจำลอง Light Gradient Boosting Machine โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8999, 0.9303, และ 0.6838 ตามลำดับ ส่วนแบบจำลองที่ใช้เทคนิค SMOTE มีค่าระลึก (Recall) และค่าถ่วงดุล (F1 Scores) สูงที่สุดคือ 0.6054 และ 0.5984 ตามลำดับ ตามตารางที่ 4.12

ตารางที่ 4.12 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Light Gradient Boosting Machine

Models	Accuracy	AUC	Recall	Precision	F1
Light Gradient Boosting Machine Tuned	**0.8999**	**0.9303**	0.3891	**0.6838**	0.4957
Light Gradient Boosting Machine Tuned (SMOTE)	0.8973	0.9283	**0.6054**	0.5921	**0.5984**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 93.89 ตามตารางที่ 4.13

ตารางที่ 4.13 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Light Gradient Boosting Machine ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Light Gradient Boosting Machine	**Studying**	19869	164	19082	951
		Dropout	1236	1663	927	1972
		Accuracy		**93.89%**		91.81%

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 90.53 ตามตารางที่ 4.14

ตารางที่ 4.14 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Light Gradient Boosting Machine ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Light Gradient Boosting Machine	**Studying**	4874	124	4714	284
		Dropout	419	317	266	470
		Accuracy		**90.53%**		90.41%

4.2.4 เทคนิคต้นไม้การตัดสินใจ (Random Forest)

สร้างแบบจำลอง Random Forest จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.15

ตารางที่ 4.15 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Random Forest

Classifier	Hyperparameter
Random Forest Tuned	bootstrap=True, ccp_alpha=0.0, class_weight={}, criterion='gini', max_depth=9, max_features=0.8827501912186471, max_leaf_nodes=None, max_samples=None, min_impurity_decrease=7.01156001505306e-08, min_impurity_split=None, min_samples_leaf=4, min_samples_split=6, min_weight_fraction_leaf=0.0, n_estimators=221, n_jobs=-1, oob_score=False, random_state=634, verbose=0, warm_start=False
Random Forest Tuned (SMOTE)	bootstrap=False, ccp_alpha=0.0, class_weight='balanced', criterion='gini', max_depth=9, max_features=0.8766737313400876, max_leaf_nodes=None, max_samples=None, min_impurity_decrease=9.509229719584798e-07, min_impurity_split=None, min_samples_leaf=4, min_samples_split=9, min_weight_fraction_leaf=0.0, n_estimators=226, n_jobs=-1, oob_score=False, random_state=634, verbose=0, warm_start=False

ผลการสร้างแบบจำลอง Random Forest โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.9018, 0.9300, และ 0.6467 ตามลำดับ ส่วนแบบจำลองที่ใช้เทคนิค SMOTE มีค่าระลึก (Recall) และค่าถ่วงดุล (F1 Scores) สูงที่สุดคือ 0.6733 และ 0.6056 ตามลำดับ ตามตารางที่ 4.16

ตารางที่ 4.16 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Random Forest

Models	Accuracy	AUC	Recall	Precision	F1
Random Forest Tuned	**0.9018**	**0.9300**	0.4912	**0.6467**	0.5581
Random Forest Tuned (SMOTE)	0.8892	0.9230	**0.6733**	0.5507	**0.6056**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 93.41 ตามตารางที่ 4.17

ตารางที่ 4.17 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Random Forest ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Random Forest	**Studying**	19576	457	18678	1355
		Dropout	1054	1845	718	2181
		Accuracy		**93.41%**		90.96%

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 90.70 ตามตารางที่ 4.18

ตารางที่ 4.18 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Random Forest ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Random Forest	**Studying**	4813	185	4630	368
		Dropout	348	388	222	514
		Accuracy		**90.70%**		89.71%

4.2.5 เทคนิคต้นไม้การตัดสินใจ (Extreme Gradient Boosting)

สร้างแบบจำลอง Extreme Gradient Boosting จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.19

ตารางที่ 4.19 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Extreme Gradient Boosting

Classifier	Hyperparameter
Extreme Gradient Boosting Tuned	base_score=0.5, booster='gbtree', callbacks=None, colsample_bylevel=1, colsample_bynode=1, colsample_bytree=0.8802288917202505, early_stopping_rounds=None, enable_categorical=False, eval_metric=None, gamma=0, gpu_id=0, grow_policy='depthwise', importance_type=None, interaction_constraints='', learning_rate=4.89600331032762e-06, max_bin=256, max_cat_to_onehot=4, max_delta_step=0, max_depth=3, max_leaves=0, min_child_weight=2, missing=nan, monotone_constraints='()', n_estimators=115, n_jobs=-1, num_parallel_tree=1, objective='binary:logistic', predictor='auto', random_state=634, reg_alpha=0.004109878925992609 reg_lambda=2.5947235925069237e-06, sampling_method=uniform, scale_pos_weight=1.270503475873832, subsample=0.42613997810498, tree_method='gpu_hist', use_label_encoder=True, validate_parameters=1, verbosity=0
Extreme Gradient Boosting Tuned (SMOTE)	base_score=0.5, booster='gbtree', callbacks=None, colsample_bylevel=1, colsample_bynode=1, colsample_bytree=0.8850877055995618, early_stopping_rounds=None, enable_categorical=False, eval_metric=None, gamma=0, gpu_id=0, grow_policy='depthwise', importance_type=None,

Classifier	Hyperparameter
	interaction_constraints='', learning_rate=0.1098112143506425, max_bin=256, max_cat_to_onehot=4, max_delta_step=0, max_depth=5, max_leaves=0, min_child_weight=3, missing=nan, monotone_constraints='()', n_estimators=198, n_jobs=-1, num_parallel_tree=1,objective='binary:logistic', predictor='auto', random_state=634, reg_alpha=0.0007616496701947522 reg_lambda=1.6562708449602826e-07, sampling_method=uniform, scale_pos_weight=1.0611030349269899, subsample=0.9266198529773333, tree_method='gpu_hist', use_label_encoder=True, validate_parameters=1, verbosity=0

ผลการสร้างแบบจำลอง Extreme Gradient Boosting โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) ค่าระลึก (Recall) และค่าถ่วงดุล (F1 Scores) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8849, 0.6023, และ 0.5655 ตามลำดับ ส่วนแบบจำลองที่ใช้เทคนิค SMOTE มีค่าพื้นที่ใต้กราฟ (AUC Scores) และค่าความแม่นยำ (Precision) สูงที่สุดคือ 0.9167 และ 0.5512 ตามลำดับ ตามตารางที่ 4.20

ตารางที่ 4.20 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Extreme Gradient Boosting

Models	Accuracy	AUC	Recall	Precision	F1
Extreme Gradient Boosting Tuned	**0.8849**	0.9147	**0.6023**	0.5451	**0.5655**
Extreme Gradient Boosting Tuned (SMOTE)	0.8848	**0.9167**	0.4819	**0.5512**	0.5141

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 92.28 ตามตารางที่ 4.21

ตารางที่ 4.21 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Extreme Gradient Boosting ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Extreme Gradient Boosting	**Studying**	18975	1058	19305	728
		Dropout	1407	1492	1042	1857
		Accuracy		89.25%		**92.28%**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 89.29 ตามตารางที่ 4.22

ตารางที่ 4.22 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Extreme Gradient Boosting ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Extreme Gradient Boosting	**Studying**	4728	270	4717	281
		Dropout	344	392	352	384
		Accuracy		**89.29%**		88.96%

4.2.6 ผลการเปรียบเทียบประสิทธิภาพแบบจำลองคาดการณ์ชุดข้อมูลระดับปริญญาตรี

ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาระดับปริญญาตรี ด้วยเทคนิคการเรียนรู้ของเครื่องประเภทต้นไม้การตัดสินใจ 5 แบบ ที่ได้ทำการปรับจูนไฮเปอร์พารามิเตอร์ ประกอบด้วยชุดข้อมูลที่หนึ่งเป็นชุดข้อมูลปกติที่ไม่ได้ปรับความสมดุล และชุดข้อมูลที่สองที่ใช้เทคนิค SMOTE เพื่อปรับความสมดุลของข้อมูล โดยเมื่อพิจารณาประสิทธิภาพของแบบจำลองในชุดข้อมูลที่หนึ่งพบว่าแบบจำลอง Random Forest ให้ค่าความถูกต้อง (Accuracy) มีค่าสูงที่สุดคือ 0.9018 รองลงมาคือแบบจำลอง Light Gradient Boosting Machine มีค่า 0.8999 แบบจำลอง Decision Tree มีค่า 0.8942 แบบจำลอง Gradient Boosting

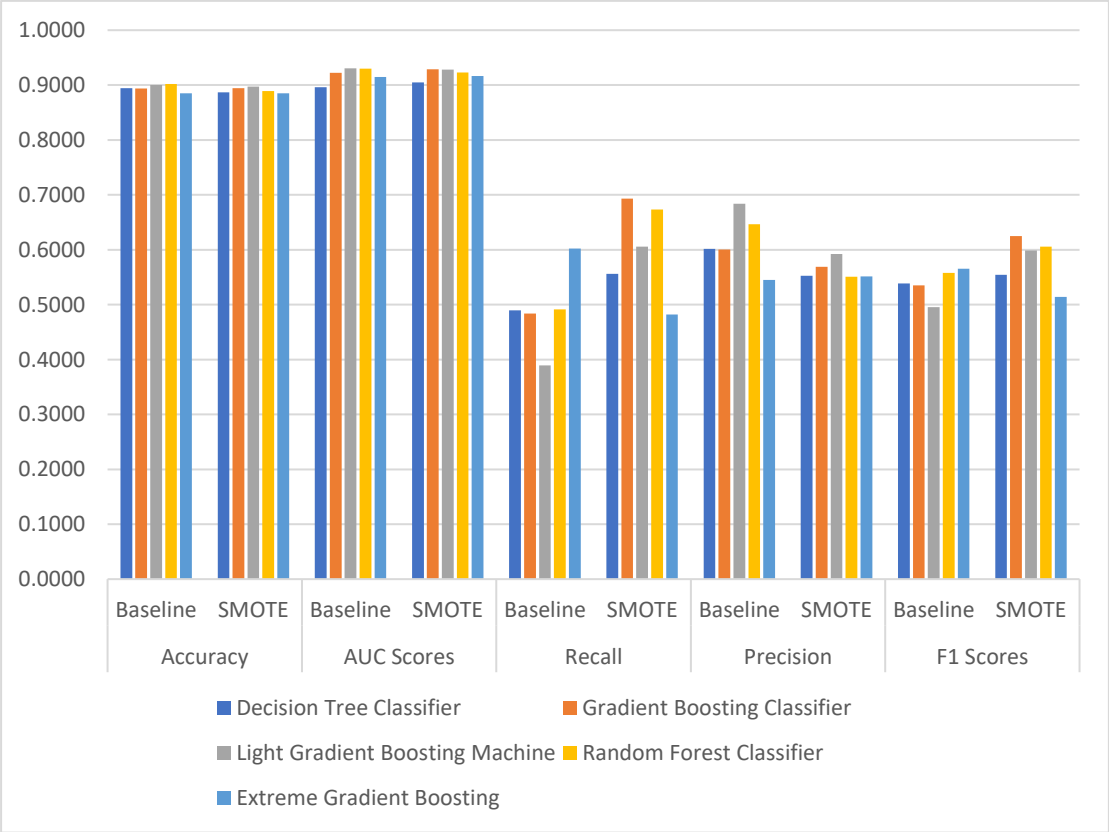
มีค่า 0.8938 และแบบจำลอง Extreme Gradient Boosting มีค่า 0.8849 ตามลำดับ และเมื่อพิจารณาค่าพื้นที่ใต้กราฟ (AUC Scores) พบว่าแบบจำลอง Light Gradient Boosting Machine มีค่าสูงที่สุดคือ 0.9303 รองลงมาคือแบบจำลอง Random Forest มีค่า 0.9300 แบบจำลอง Gradient Boosting มีค่า 0.9222 แบบจำลอง Extreme Gradient Boosting มีค่า 0.9147 และแบบจำลอง Decision Tree มีค่า 0.8960 ตามลำดับ และเมื่อพิจารณาค่าความถ่วงดุล (F1 Scores) พบว่าแบบจำลอง Extreme Gradient Boosting มีค่าสูงที่สุดคือ 0.5655 รองลงมาคือแบบจำลอง Random Forest มีค่า 0.5581 แบบจำลอง Decision Tree มีค่า 0.5384 แบบจำลอง Gradient Boosting มีค่า 0.5353 และแบบจำลอง Light Gradient Boosting มีค่าต่ำที่สุดคือ 0.4957 ตามลำดับ

และเมื่อพิจารณาในข้อมูลชุดที่สองที่ใช้เทคนิค SMOTE ปรับความสมดุลของข้อมูล พบว่าค่าระลึก (Recall) ของแบบจำลองส่วนใหญ่มีค่าเพิ่มขึ้นและค่าความแม่นยำ (Precision) ลดลง แต่เมื่อเทียบสัดส่วนการคาดการณ์ของหมวดหมู่จริงและเท็จตามค่าระลึก (Recall) และค่าความแม่นยำ (Precision) ที่รวมกันคือค่า F1 Scores พบว่าแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น โดยแบบจำลอง Light Gradient Boosting Machine ให้ค่าความถูกต้อง (Accuracy) สูงที่สุดคือ 0.8973 รองลงมาคือแบบจำลอง Gradient Boosting มีค่า 0.8946 แบบจำลอง Random Forest มีค่า 0.8892 แบบจำลอง Decision Tree มีค่า 0.8869 และแบบจำลอง Extreme Gradient Boosting มีค่าต่ำที่สุดคือ 0.8848 ส่วนค่าพื้นที่ใต้กราฟ (AUC Scores) พบว่าแบบจำลอง Gradient Boosting มีค่าสูงที่สุดคือ 0.9289 รองลงมาคือแบบจำลอง Light Gradient Boosting Machine มีค่า 0.9283 แบบจำลอง Random Forest มีค่า 0.9230 แบบจำลอง Extreme Gradient Boosting มีค่า 0.9167 และแบบจำลอง Decision Tree มีค่าต่ำที่สุดคือ 0.9049 และเมื่อพิจารณาค่า F1 Scores พบว่าแบบจำลอง Gradient Boosting มีค่าสูงที่สุดคือ 0.6247 รองลงมาคือ แบบจำลอง Random Forest มีค่า 0.6056 แบบจำลอง Light Gradient Boosting Machine มีค่า 0.5984 แบบจำลอง Decision Tree มีค่า 0.5543 และแบบจำลอง Extreme Gradient Boosting มีค่าต่ำที่สุดคือ 0.5141 ตามลำดับ ตามตารางที่ 4.23 และภาพที่ 4.1

ตารางที่ 4.23 ตารางเปรียบเทียบประสิทธิภาพของแบบจำลอง ระดับชั้นปริญญาตรี

Models	Accuracy	AUC	Recall	Precision	F1
Decision Tree	0.8942	0.8960	0.4895	0.6017	0.5384
Decision Tree (SMOTE)	0.8869	0.9049	0.5564	0.5527	0.5543
Extreme Gradient Boosting	0.8849	0.9147	0.6023	0.5451	0.5655
Extreme Gradient Boosting (SMOTE)	0.8848	0.9167	0.4819	0.5512	0.5141

Models	Accuracy	AUC	Recall	Precision	F1
Gradient Boosting	0.8938	0.9222	0.4839	0.6003	0.5353
Gradient Boosting (SMOTE)	0.8946	0.9289	**0.6933**	0.5688	**0.6247**
Light Gradient Boosting Machine	0.8999	**0.9303**	0.3891	**0.6838**	0.4957
Light Gradient Boosting Machine (SMOTE)	0.8973	0.9283	0.6054	0.5921	0.5984
Random Forest	**0.9018**	0.9300	0.4912	0.6467	0.5581
Random Forest (SMOTE)	0.8892	0.9230	0.6733	0.5507	0.6056

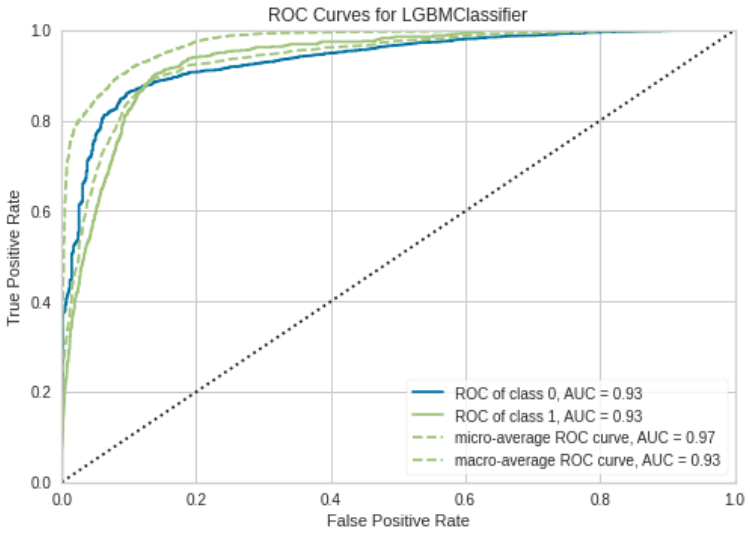


ภาพที่ 4.1 กราฟเปรียบเทียบประสิทธิภาพของแบบจำลอง ระดับชั้นปริญญาตรี

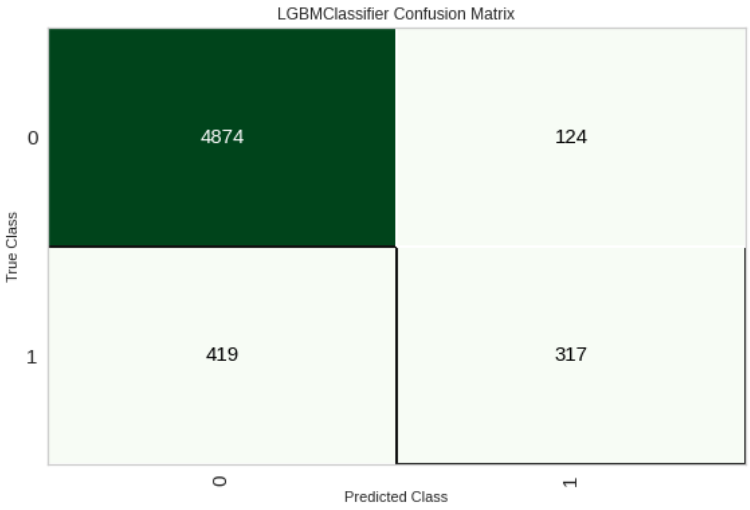
ผู้วิจัยประเมินประสิทธิภาพแบบจำลองจากค่าพื้นที่ใต้กราฟ (AUC Scores) เป็นสำคัญ เนื่องจากเป็นวิธีการตรวจสอบที่ใช้ในการประเมินแบบจำลองคาดการณ์หากค่า AUC Scores มีค่าใกล้ 1 มาก แสดงว่าแบบจำลองมีประสิทธิภาพในการจำแนกกลุ่มออกจากกันได้ถูกต้อง ซึ่งกราฟนี้ได้มาจากอัตรา True Positive Rate และ False Positive Rate ผลจากการทดลองแสดงให้เห็นว่าแบบจำลอง Light Gradient Boosting Machine ที่ไม่ได้ปรับปรุงความสมดุลของข้อมูล มีประสิทธิภาพสูงสุด โดยให้ค่าพื้นที่ใต้กราฟ (AUC Scores) สูงที่สุด มีค่า 0.9303 รองลงมาคือแบบจำลอง Random Forest ที่มีค่าใกล้เคียงกันคือ 0.9300 และแบบจำลอง Gradient Boosting

มีค่า 0.9222 โดยแบบจำลองทั้ง 3 แบบให้ค่าความถูกต้อง (Accuracy) กว่าร้อยละ 90 โดยผู้วิจัยนำแบบจำลอง Light Gradient Boosting Machine ที่ได้ดังกล่าวไปใช้งานต่อไป

จากผลลัพธ์ข้างต้นสามารถแสดงค่าพื้นที่ใต้กราฟ (AUC Scores) และตารางเมทริกซ์ความสับสน (Confusion Matrix) ของแบบจำลอง Light Gradient Boosting Machine จากชุดข้อมูลตรวจสอบ 20% ได้ดังภาพที่ 4.2 และ 4.3 ตามลำดับ

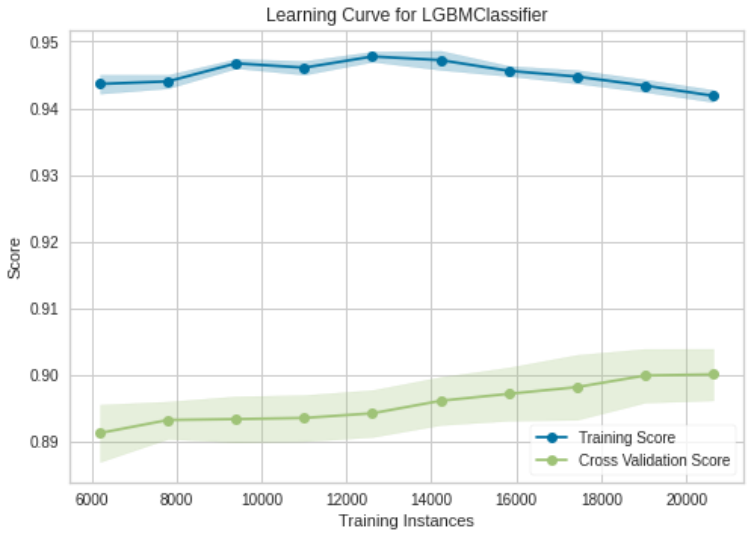


ภาพที่ 4.2 ROC Curves ของแบบจำลอง Light Gradient Boosting Machine จากชุดข้อมูลตรวจสอบ



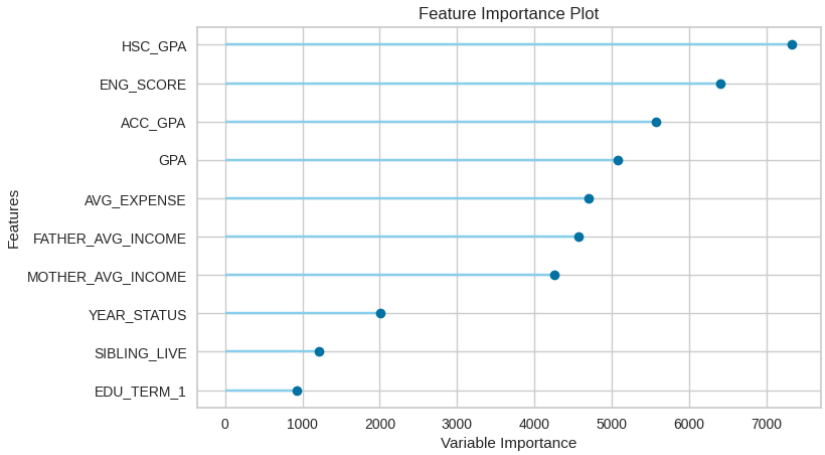
ภาพที่ 4.3 Confusion Matrix ของแบบจำลอง Light Gradient Boosting Machine จากชุดข้อมูลตรวจสอบ

จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง Light Gradient Boosting Machine ของชุดข้อมูลระดับปริญญาตรี ด้วย Learning Curve พบว่าเมื่อใช้ชุดการฝึกขนาดเล็ก คะแนนการฝึกจะสูงหรือมีอคติต่ำ แต่คะแนนการทดสอบต่ำหรือมีความแปรปรวนสูง กล่าวคือ แบบจำลองนั้น Overfitting จากนั้นเมื่อทำการเพิ่มขนาดชุดข้อมูลการฝึกพบว่ามีอคติสูงขึ้นแต่ความแปรปรวนลดลงซึ่งหมายความว่าการเพิ่มชุดข้อมูลการฝึกแบบจำลองจะช่วยลดปัญหาการ Overfitting ลงได้ ดังภาพที่ 4.4



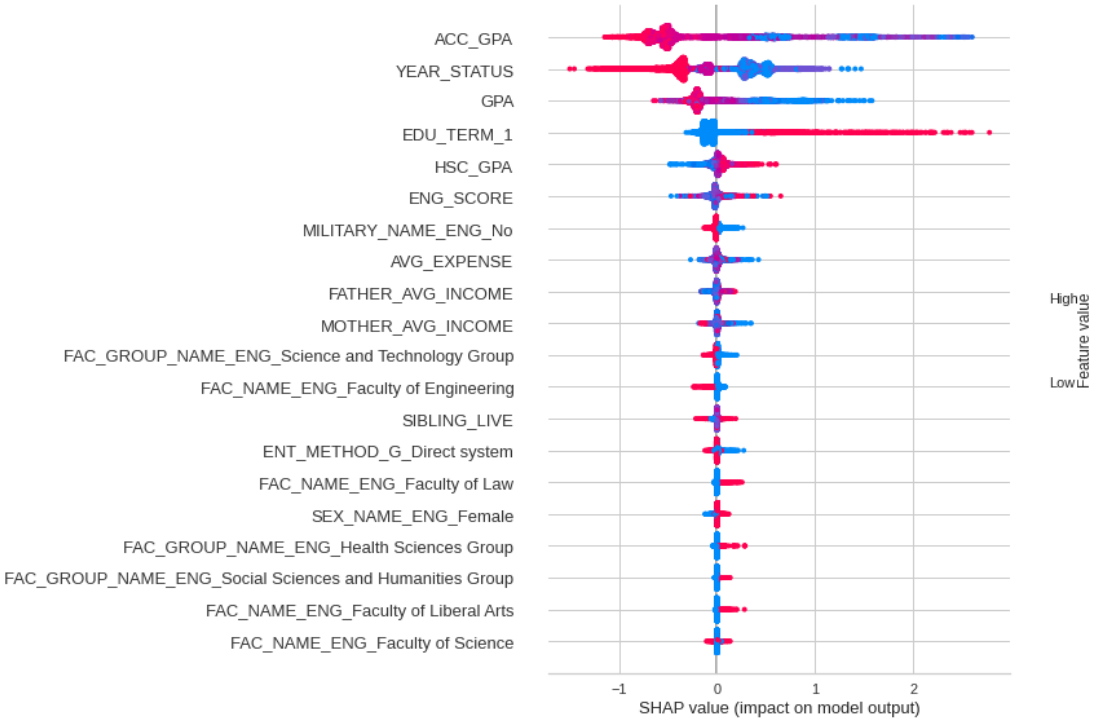
ภาพที่ 4.4 Learning Curve ของแบบจำลอง Light Gradient Boosting Machine

ส่วนผลลัพธ์คุณสมบัติปัจจัยที่สำคัญ 10 อันดับแรกของแบบจำลอง Light Gradient Boosting Machine ได้แก่ ผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษาเป็นปัจจัยสำคัญที่สุด รองลงมาคือ คะแนนภาษาอังกฤษก่อนเข้าศึกษา ผลการเรียนเฉลี่ยสะสม ผลการเรียนเฉลี่ยปัจจุบัน รายจ่ายเฉลี่ยนักศึกษา รายได้เฉลี่ยมารดา รายได้เฉลี่ยบิดา สถานะชั้นปีที่ จำนวนพี่น้อง และภาคการศึกษาที่ 1 ตามลำดับ ดังภาพที่ 4.5



ภาพที่ 4.5 คุณสมบัติปัจจัยที่สำคัญ 10 อันดับแรกของแบบจำลอง Light Gradient Boosting Machine

ผลลัพธ์คุณสมบัติปัจจัยที่สำคัญจากแผนภาพ (SHAP) ที่แสดงปัจจัยเชิงบวกและเชิงลบที่สำคัญที่สุดของแต่ละปัจจัยของแบบจำลอง Random Forest โดยค่าที่สูงแทนด้วยสีแดง และค่าที่ต่ำแทนด้วยสีฟ้า พบว่าปัจจัยที่สำคัญที่สุดคือ ผลการเรียนเฉลี่ยสะสมโดยผลการเรียนที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษาและผลการเรียนที่สูงส่งผลต่อการคงอยู่ของนักศึกษา รองลงมาคือชั้นปีซึ่งพบว่าในช่วงชั้นปีแรกนักศึกษาจะออกกลางคันมากกว่า และผลการเรียนเฉลี่ยปัจจุบันที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษา และส่วนใหญ่นักศึกษาจะออกกลางคันในภาคการศึกษาแรก ส่วนผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษาและคะแนนภาษาอังกฤษก่อนเข้าศึกษาไม่ได้ส่งผลต่อตัวแปรทั้งสองคลาสอย่างมีนัยสำคัญ ส่วนปัจจัยด้านการเงินพบว่ารายจ่ายเฉลี่ยของนักศึกษาไม่ได้ส่งผลต่อตัวแปรเป้าหมายอย่างมีนัยสำคัญ แต่พบว่าบิดาที่มีรายได้เฉลี่ยสูงนักศึกษาจะออกกลางคันมากกว่า ซึ่งตรงกันข้ามกับรายได้เฉลี่ยของมารดาที่พบว่ารายได้เฉลี่ยที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษามากกว่ารายได้เฉลี่ยที่สูงกว่า และพบว่านักศึกษาประเภทรับตรงมีอัตราการออกกลางคันน้อยกว่าประเภท Admission และพบว่าส่วนใหญ่นักศึกษาในกลุ่มสาขาวิชามนุษยศาสตร์และสังคมศาสตร์จะมีอัตราการออกกลางคันมากกว่ากลุ่มวิทยาศาสตร์สุขภาพที่มีอัตราการออกกลางคันน้อยกว่า ส่วนตัวแปรปัจจัยอื่น ๆ ไม่ส่งผลต่อตัวแปรเป้าหมายอย่างมีนัยสำคัญ ดังภาพที่ 4.6



ภาพที่ 4.6 คุณสมบัติปัจจัยที่สำคัญจากแผนภาพ (SHAP) ของแบบจำลอง Light Gradient Boosting Machine

4.3 ผลการพัฒนาแบบจำลองคาดการณ์ชุดข้อมูลระดับชั้นบัณฑิตศึกษา

ในส่วนนี้จะนำแบบจำลองการเรียนรู้ของเครื่องประเภทต้นไม้การตัดสินใจทั้ง 5 แบบ มาทำการคาดการณ์การออกกลางคันของนักศึกษาระดับชั้นบัณฑิตศึกษา โดยใช้เทคนิคเลือกคุณสมบัติตัวแปร การปรับจูนไฮเปอร์พารามิเตอร์ และการใช้เทคนิควิธีสังเคราะห์ข้อมูลเพิ่มของแต่ละแบบจำลอง เพื่อเปรียบเทียบประสิทธิภาพแบบจำลอง และความเหมาะสมของการนำไปใช้งาน โดยแบ่งชุดข้อมูลออกเป็น 10 ชุดในการสร้างและทดสอบแบบจำลอง K-Fold Cross Validation 10 folds จากชุดข้อมูลเรียนรู้ 80% (4,211) และชุดข้อมูลตรวจสอบอีก 20% (1,053)

4.3.1 เทคนิคต้นไม้การตัดสินใจ (Decision Tree)

ในการสร้างแบบจำลอง Decision Tree จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.24

ตารางที่ 4.24 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Decision Tree

Classifier	Hyperparameter
Decision Tree Tuned	ccp_alpha=0.0, class_weight=None, criterion='entropy', max_depth=5, max_features=0.7097575197717978, max_leaf_nodes=None, min_impurity_decrease=1.2176674734370814e-06, min_impurity_split=None, min_samples_leaf=4, min_samples_split=6, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=634, splitter='best'
Decision Tree Tuned (SMOTE)	ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=14, max_features=0.5299597777749986, max_leaf_nodes=None, min_impurity_decrease=3.3286524135805355e-09, min_impurity_split=None, min_samples_leaf=4, min_samples_split=7, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=634, splitter='best'

ผลการสร้างแบบจำลอง Decision Tree โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8411, 0.7591 และ 0.6635 ตามลำดับ ส่วนแบบจำลองที่ใช้เทคนิค SMOTE มีค่าระลึก (Recall) และค่าความถ่วงดุล (F1 Scores) สูงที่สุดคือ 0.4691 และ 0.4427 ตามตารางที่ 4.25

ตารางที่ 4.25 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Decision Tree

Models	Accuracy	AUC	Recall	Precision	F1
Decision Tree Tuned	**0.8411**	**0.7591**	0.2607	**0.6635**	0.3601
Decision Tree Tuned (SMOTE)	0.7891	0.7112	**0.4691**	0.4233	**0.4427**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 87.96 ตามตารางที่ 4.26

ตารางที่ 4.26 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Decision Tree ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Decision Tree	**Studying**	3287	169	3204	252
		Dropout	485	270	255	500
		Accuracy		84.47%		**87.96%**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 83.19 ตามตารางที่ 4.27

ตารางที่ 4.27 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Decision Tree ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Decision Tree	**Studying**	806	47	740	113
		Dropout	130	70	114	86
		Accuracy		**83.19%**		78.44%

4.3.2 เทคนิคต้นไม้การตัดสินใจ (Gradient Boosting)

สร้างแบบจำลอง Gradient Boosting จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.28

ตารางที่ 4.28 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Gradient Boosting

Classifier	Hyperparameter
Gradient Boosting Tuned	ccp_alpha=0.0, criterion='friedman_mse', init=None, learning_rate=0.014240765680578197, loss='deviance', max_depth=9, max_features=0.6445675491172762, max_leaf_nodes=None, min_impurity_decrease=1.6623248340662306e-09, min_impurity_split=None, min_samples_leaf=4, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=107, n_iter_no_change=None, presort='deprecated', random_state=634, subsample=0.9297093575172972, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False
Gradient Boosting Tuned (SMOTE)	ccp_alpha=0.0, criterion='friedman_mse', init=None, learning_rate=0.02024595699168006, loss='deviance', max_depth=6, max_features=0.5292854327228281, max_leaf_nodes=None, min_impurity_decrease=0.009015363936207022, min_impurity_split=None, min_samples_leaf=3, min_samples_split=9, min_weight_fraction_leaf=0.0, n_estimators=155, n_iter_no_change=None, presort='deprecated', random_state=634, subsample=0.8348581066012426, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False

ผลการสร้างแบบจำลอง Gradient Boosting โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8509 และ 0.7183 ตามลำดับ ส่วนค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.7842, 0.4492 และ 0.4875 ตามลำดับ ตามตารางที่ 4.29

ตารางที่ 4.29 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Gradient Boosting

Models	Accuracy	AUC	Recall	Precision	F1
Gradient Boosting Tuned	**0.8509**	0.7820	0.2822	**0.7183**	0.4025
Gradient Boosting Tuned (SMOTE)	0.8319	**0.7842**	**0.4492**	0.5413	**0.4875**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 89.03 ตามตารางที่ 4.30

ตารางที่ 4.30 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Gradient Boosting ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Gradient Boosting	**Studying**	3430	26	3229	227
		Dropout	436	319	315	440
		Accuracy		**89.03%**		87.13%

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 84.62 ตามตารางที่ 4.31

ตารางที่ 4.31 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Gradient Boosting ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Gradient Boosting	**Studying**	832	21	768	85
		Dropout	141	59	106	94
		Accuracy		**84.62%**		81.86%

4.3.3 เทคนิคต้นไม้การตัดสินใจ (Light Gradient Boosting Machine)

สร้างแบบจำลอง Light Gradient Boosting Machine จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.32

ตารางที่ 4.32 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Light Gradient Boosting Machine

Classifier	Hyperparameter
Light Gradient Boosting Machine Tuned	bagging_fraction=0.4930653502543217, bagging_freq=7, boosting_type='gbdt', class_weight=None, colsample_bytree=1.0, feature_fraction=0.5343498506182052, importance_type='split', learning_rate=0.03608092217951852, max_depth=-1, min_child_samples=26, min_child_weight=0.001, min_split_gain=0.6786859715976972, n_estimators=80, n_jobs=-1, num_leaves=187, objective=None, random_state=634, reg_alpha=1.296290442338103e-07, reg_lambda=0.031101847314716723, silent='warn', subsample=1.0, subsample_for_bin=200000, subsample_freq=0
Light Gradient Boosting Machine Tuned (SMOTE)	bagging_fraction=0.6373183531590281, bagging_freq=7, boosting_type='gbdt', class_weight=None, colsample_bytree=1.0, feature_fraction=0.5984439918153773, importance_type='split', learning_rate=0.025145420020642312, max_depth=-1, min_child_samples=3, min_child_weight=0.001, min_split_gain=0.8789874151353874, n_estimators=247, n_jobs=-1, num_leaves=153, objective=None, random_state=634, reg_alpha=4.689789223484412e-05, reg_lambda=5.705118598889439e-07, silent='warn', subsample=1.0, subsample_for_bin=200000, subsample_freq=0

ผลการสร้างแบบจำลอง Light Gradient Boosting Machine โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8485, 0.7445 และ 0.7545 ตามลำดับ ส่วนค่าระลึก (Recall) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.3776 และ 0.4572 ตามลำดับ ตามตารางที่ 4.33

ตารางที่ 4.33 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Light Gradient Boosting Machine

Models	Accuracy	AUC	Recall	Precision	F1
Light Gradient Boosting Machine Tuned	**0.8485**	**0.7745**	0.2345	**0.7545**	0.3560
Light Gradient Boosting Machine Tuned (SMOTE)	0.8409	0.7709	**0.3776**	0.5907	**0.4572**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 91.47 ตามตารางที่ 4.34

ตารางที่ 4.34 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Light Gradient Boosting Machine ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Light Gradient Boosting Machine	**Studying**	3411	45	3379	77
		Dropout	566	189	282	473
		Accuracy		85.49%		**91.47%**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 83.48 ตามตารางที่ 4.35

ตารางที่ 4.35 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Light Gradient Boosting Machine ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Light Gradient Boosting Machine	**Studying**	832	21	801	52
		Dropout	154	46	122	78
		Accuracy		83.38%		**83.48%**

4.3.4 เทคนิคต้นไม้การตัดสินใจ (Random Forest)

สร้างแบบจำลอง Random Forest จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.36

ตารางที่ 4.36 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Random Forest

Classifier	Hyperparameter
Random Forest Tuned	bootstrap=False, ccp_alpha=0.0, class_weight={}, criterion='gini', max_depth=7, max_features=0.7907046396479619, max_leaf_nodes=None, max_samples=None, min_impurity_decrease=3.6815809059953807e-06, min_impurity_split=None, min_samples_leaf=4, min_samples_split=8, min_weight_fraction_leaf=0.0, n_estimators=267, n_jobs=-1, oob_score=False, random_state=634, verbose=0, warm_start=False
Random Forest Tuned (SMOTE)	bootstrap=True, ccp_alpha=0.0, class_weight='balanced', criterion='gini', max_depth=11, max_features=0.4793543132516757, max_leaf_nodes=None, max_samples=None, min_impurity_decrease=9.25987373470026e-08, min_impurity_split=None, min_samples_leaf=3, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=200, n_jobs=-1, oob_score=False, random_state=634, verbose=0, warm_start=False

ผลการสร้างแบบจำลอง Random Forest โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความ

ถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8528, 0.7886, และ 0.6983 ตามลำดับ ส่วนแบบจำลองที่ใช้เทคนิค SMOTE มีค่าระลึก (Recall) และค่าถ่วงดุล (F1 Scores) สูงที่สุดคือ 0.4491 และ 0.4794 ตามลำดับ ตามตารางที่ 4.37

ตารางที่ 4.37 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Random Forest

Models	Accuracy	AUC	Recall	Precision	F1
Random Forest Tuned	**0.8528**	**0.7886**	0.3193	**0.6983**	0.4358
Random Forest Tuned (SMOTE)	0.8257	0.7877	**0.4491**	0.5211	**0.4794**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 90.14 ตามตารางที่ 4.38

ตารางที่ 4.38 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Random Forest ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Random Forest	**Studying**	3402	54	3268	188
		Dropout	484	271	227	528
		Accuracy		86.13%		**90.14%**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 84.14 ตามตารางที่ 4.39

ตารางที่ 4.39 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Random Forest ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Random Forest	**Studying**	824	29	768	85
		Dropout	138	62	104	96
		Accuracy		**84.14%**		82.05%

4.3.5 เทคนิคต้นไม้การตัดสินใจ (Extreme Gradient Boosting)

สร้างแบบจำลอง Extreme Gradient Boosting จะทำการปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่มเพื่อให้แบบจำลองมีความแม่นยำมากที่สุด ดังตารางที่ 4.40

ตารางที่ 4.40 การปรับจูนไฮเปอร์พารามิเตอร์ของแบบจำลอง Extreme Gradient Boosting

Classifier	Hyperparameter
Extreme Gradient Boosting Tuned	base_score=0.5, booster='gbtree', callbacks=None, colsample_bylevel=1, colsample_bynode=1, colsample_bytree=0.8850877055995618, early_stopping_rounds=None, enable_categorical=False, eval_metric=None, gamma=0, gpu_id=0, grow_policy='depthwise', importance_type=None, interaction_constraints='', learning_rate=0.1098112143506425, max_bin=256, max_cat_to_onehot=4, max_delta_step=0, max_depth=5, max_leaves=0, min_child_weight=3, missing=nan, monotone_constraints='()', n_estimators=198, n_jobs=-1, num_parallel_tree=1, objective='binary:logistic', predictor='auto', random_state=634, reg_alpha=0.0007616496701947522, reg_lambda=1.6562708449602826e-07, scale_pos_weight=1.0611030349269899, subsample= 0.9266198529773333, tree_method='gpu_hist', use_label_encoder=True, validate_parameters=1, verbosity=0
Extreme Gradient Boosting Tuned (SMOTE)	base_score=0.5, booster='gbtree', callbacks=None, colsample_bylevel=1, colsample_bynode=1, colsample_bytree=0.8850877055995618, early_stopping_rounds=None, enable_categorical=False, eval_metric=None, gamma=0, gpu_id=0, grow_policy='depthwise', importance_type=None, interaction_constraints='', learning_rate=0.1098112143506425, max_bin=256, max_cat_to_onehot=4, max_delta_step=0, max_depth=5, max_leaves=0,

Classifier	Hyperparameter
	min_child_weight=3, missing=nan, monotone_constraints='()', n_estimators=198, n_jobs=-1, num_parallel_tree=1, objective='binary:logistic', predictor='auto', random_state=634, reg_alpha=0.0007616496701947522, reg_lambda=1.6562708449602826e-07, scale_pos_weight=1.0611030349269899, subsample=0.9266198529773333, tree_method='gpu_hist', use_label_encoder=True, validate_parameters=1, verbosity=0

ผลการสร้างแบบจำลอง Extreme Gradient Boosting โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) ของแบบจำลองที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ที่ใช้เทคนิค SMOTE และที่ไม่ได้ใช้เทคนิค SMOTE ในการสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของคลาส พบว่าค่าความถูกต้อง (Accuracy) และค่าความแม่นยำ (Precision) ของแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าสูงที่สุดคือ 0.8418 และ 0.6104 ส่วนแบบจำลองที่ใช้เทคนิค SMOTE มีค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) และค่าถ่วงดุล (F1 Scores) สูงที่สุดคือ 0.7606, 0.3974, และ 0.4654 ตามลำดับ ส่วน ตามตารางที่ 4.41

ตารางที่ 4.41 ตารางแสดงผลประสิทธิภาพแบบจำลองคาดการณ์ Extreme Gradient Boosting

Models	Accuracy	AUC	Recall	Precision	F1
Extreme Gradient Boosting Tuned	**0.8418**	0.7551	0.3418	**0.6104**	0.4349
Extreme Gradient Boosting Tuned (SMOTE)	0.8376	**0.7606**	**0.3974**	0.5703	**0.4654**

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลเรียนรู้ 80% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 89.65 ตามตารางที่ 4.42

ตารางที่ 4.42 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Extreme Gradient Boosting ของชุดข้อมูลเรียนรู้ 80%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Extreme Gradient Boosting	**Studying**	3387	69	3346	110
		Dropout	367	388	341	414
		Accuracy		**89.65%**		89.29%

เมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน ของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาที่ปรับจูนไฮเปอร์พารามิเตอร์ และเปรียบเทียบการใช้เทคนิค SMOTE และไม่ได้ใช้ จากชุดข้อมูลตรวจสอบ 20% พบว่าแบบจำลองที่ไม่ได้ใช้เทคนิค SMOTE มีค่าความถูกต้องสูงที่สุด ร้อยละ 83.29 ตามตารางที่ 4.43

ตารางที่ 4.43 ผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาจากแบบจำลอง Extreme Gradient Boosting ของชุดข้อมูลตรวจสอบ 20%

Models			Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Extreme Gradient Boosting	**Studying**	807	46	792	61
		Dropout	130	70	121	179
		Accuracy		**83.29%**		82.72%

4.3.6 ผลการเปรียบเทียบประสิทธิภาพแบบจำลองคาดการณ์ชุดข้อมูลระดับบัณฑิตศึกษา

ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองในการคาดการณ์การออกกลางคันของนักศึกษาระดับบัณฑิตศึกษา ด้วยเทคนิคการเรียนรู้ของเครื่องประเภทต้นไม้การตัดสินใจ 5 แบบที่ได้ทำการปรับจูนไฮเปอร์พารามิเตอร์ ประกอบด้วยชุดข้อมูลที่หนึ่งเป็นชุดข้อมูลปกติที่ไม่ได้ปรับความสมดุล และชุดข้อมูลที่สองที่ใช้เทคนิค SMOTE เพื่อปรับความสมดุลของข้อมูล โดยเมื่อพิจารณาประสิทธิภาพของแบบจำลองในชุดข้อมูลที่หนึ่งพบว่าแบบจำลอง Random Forest ให้ค่าความถูกต้อง (Accuracy) มีค่าสูงที่สุดคือ 0.8528 รองลงมาคือแบบจำลอง Gradient Boosting มีค่า 0.8509 แบบจำลอง Light Gradient Boosting Machine มีค่า 0.8485 แบบจำลอง Extreme

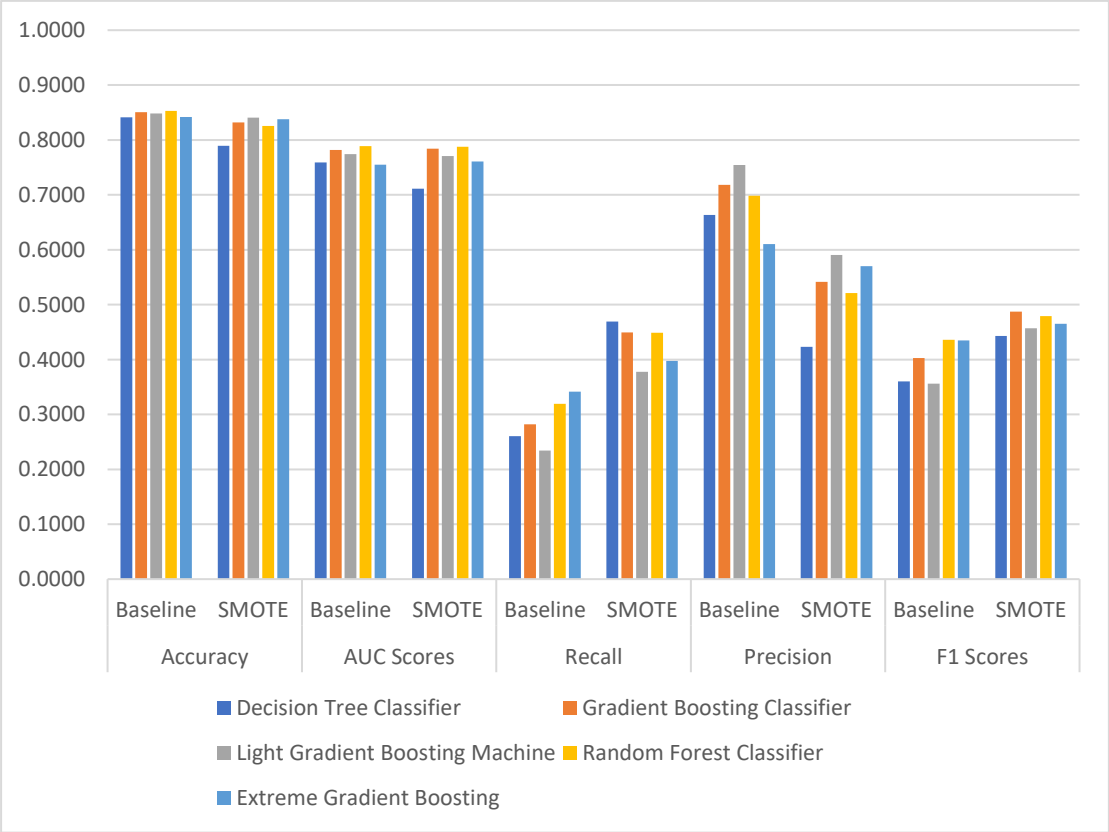
Gradient Boosting มีค่า 0.8418 และแบบจำลอง Decision Tree มีค่า 0.8411 ตามลำดับ และเมื่อพิจารณาค่าพื้นที่ใต้กราฟ (AUC Scores) พบว่าแบบจำลอง Random Forest มีค่าสูงที่สุดคือ 0.7886 รองลงมาคือแบบจำลอง Gradient Boosting มีค่า 0.7820 แบบจำลอง Light Gradient Boosting Machine มีค่า 0.7745 แบบจำลอง Decision Tree มีค่า 0.7591 และแบบจำลอง Extreme Gradient Boosting มีค่า 0.7551 ตามลำดับ และเมื่อพิจารณาค่าความถ่วงดุล (F1 Scores) พบว่าแบบจำลอง Random Forest มีค่าสูงที่สุดคือ 0.4358 รองลงมาคือแบบจำลอง Extreme Gradient Boosting มีค่า 0.4349 แบบจำลอง Gradient Boosting มีค่า 0.4025 แบบจำลอง Decision Tree มีค่า 0.3601 และแบบจำลอง Light Gradient Boosting มีค่าต่ำที่สุดคือ 0.3560 ตามลำดับ

และเมื่อพิจารณาในข้อมูลชุดที่สองที่ใช้เทคนิค SMOTE ปรับความสมดุลของข้อมูล พบว่าค่าระลึก (Recall) ของแบบจำลองส่วนใหญ่มีค่าเพิ่มขึ้นและค่าความแม่นยำ (Precision) ลดลง แต่เมื่อเทียบสัดส่วนการคาดการณ์ของหมวดหมู่จริงและเท็จตามค่าระลึก (Recall) และค่าความแม่นยำ (Precision) ที่รวมกันคือค่า F1 Scores พบว่าแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น โดยแบบจำลอง Light Gradient Boosting Machine ให้ค่าความถูกต้อง (Accuracy) สูงที่สุดคือ 0.8409 รองลงมาคือแบบจำลอง Extreme Gradient Boosting มีค่า 0.8376 แบบจำลอง Gradient Boosting มีค่า 0.8319 แบบจำลอง Random Forest มีค่า 0.8257 และแบบจำลอง Decision Tree มีค่าต่ำที่สุดคือ 0.7891 ส่วนค่าพื้นที่ใต้กราฟ (AUC Scores) พบว่าแบบจำลอง Random Forest มีค่าสูงที่สุดคือ 0.7877 รองลงมาคือแบบจำลอง Gradient Boosting มีค่า 0.7842 แบบจำลอง Light Gradient Boosting Machine มีค่า 0.7709 แบบจำลอง Extreme Gradient Boosting มีค่า 0.7606 และแบบจำลอง Decision Tree มีค่าต่ำที่สุดคือ 0.7112 และเมื่อพิจารณาค่า F1 Scores พบว่าแบบจำลอง Gradient Boosting มีค่าสูงที่สุดคือ 0.4875 รองลงมาคือแบบจำลอง Random Forest มีค่า 0.4794 แบบจำลอง Extreme Gradient Boosting มีค่า 0.4654 แบบจำลอง Light Gradient Boosting Machine มีค่า 0.4572 และแบบจำลอง Decision Tree มีค่าต่ำที่สุดคือ 0.4427 ตามลำดับ ตามตารางที่ 4.44 และภาพที่ 4.7

ตารางที่ 4.44 ตารางเปรียบเทียบประสิทธิภาพของแบบจำลอง ระดับชั้นบัณฑิตศึกษา

Models	Accuracy	AUC	Recall	Precision	F1
Decision Tree	0.8411	0.7591	0.2607	0.6635	0.3601
Decision Tree (SMOTE)	0.7891	0.7112	**0.4691**	0.4233	0.4427
Extreme Gradient Boosting	0.8418	0.7551	0.3418	0.6104	0.4349
Extreme Gradient Boosting (SMOTE)	0.8376	0.7606	0.3974	0.5703	0.4654

Models	Accuracy	AUC	Recall	Precision	F1
Gradient Boosting	0.8509	0.7820	0.2822	0.7183	0.4025
Gradient Boosting (SMOTE)	0.8319	0.7842	0.4492	0.5413	**0.4875**
Light Gradient Boosting Machine	0.8485	0.7745	0.2345	**0.7545**	0.3560
Light Gradient Boosting Machine (SMOTE)	0.8409	0.7709	0.3776	0.5907	0.4572
Random Forest	**0.8528**	**0.7886**	0.3193	0.6983	0.4358
Random Forest (SMOTE)	0.8257	0.7877	0.4491	0.5211	0.4794

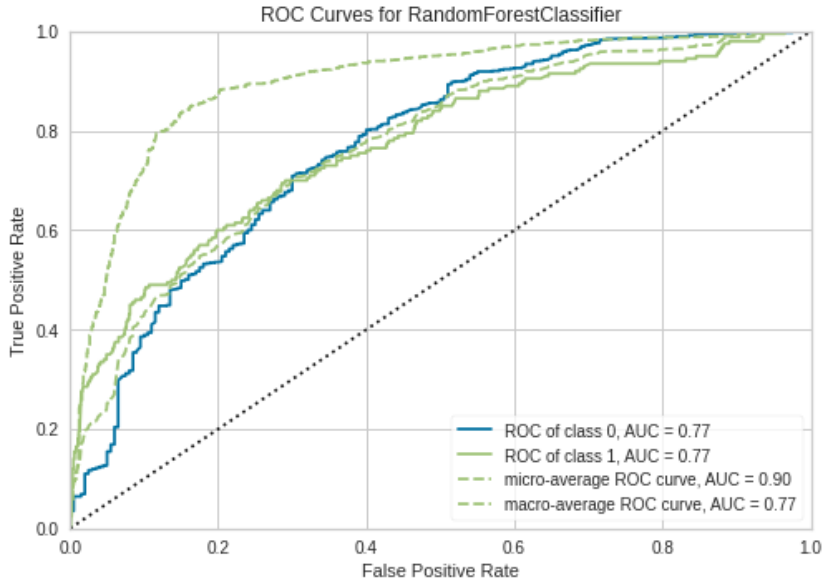


ภาพที่ 4.7 กราฟเปรียบเทียบประสิทธิภาพของแบบจำลอง ระดับชั้นบัณฑิตศึกษา

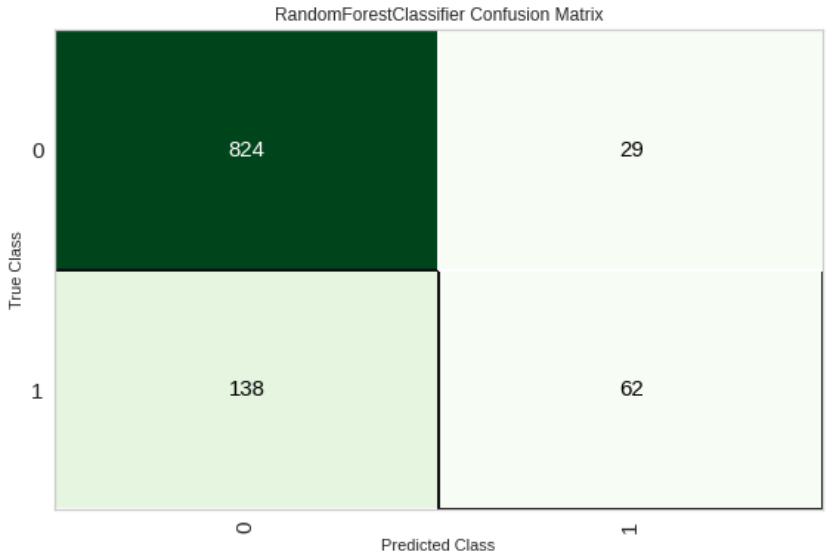
ผู้วิจัยประเมินประสิทธิภาพแบบจำลองจากค่าพื้นที่ใต้กราฟ (AUC Scores) เป็นสำคัญ เนื่องจากเป็นวิธีการตรวจสอบที่ใช้ในการประเมินแบบจำลองคาดการณ์หากค่า AUC Scores มีค่าใกล้ 1 มาก แสดงว่าแบบจำลองมีประสิทธิภาพในการจำแนกกลุ่มออกจากกันได้ถูกต้อง ซึ่งกราฟนี้ได้มาจากอัตรา True Positive Rate และ False Positive Rate ผลจากการทดลองแสดงให้เห็นว่าแบบจำลอง Random Forest ที่ไม่ได้ปรับปรุงความสมดุลของข้อมูล มีประสิทธิภาพสูงสุด โดยให้ค่าพื้นที่ใต้กราฟ (AUC Scores) สูงที่สุด มีค่า 0.7886 รองลงมาคือแบบจำลอง Random Forest

(SMOTE) และแบบจำลอง Gradient Boosting (SMOTE) มีค่า 0.7877 และ 0.7842 โดยส่วนใหญ่แบบจำลองให้ค่าความถูกต้อง (Accuracy) กว่าร้อยละ 80 ตามลำดับ

จากผลลัพธ์ข้างต้นสามารถแสดงค่าพื้นที่ใต้กราฟ (AUC Scores) และตารางเมทริกซ์ความสับสน (Confusion Matrix) ของแบบจำลอง Random Forest จากชุดข้อมูลตรวจสอบ 20% ได้ดังภาพที่ 4.8 และ 4.9 ตามลำดับ

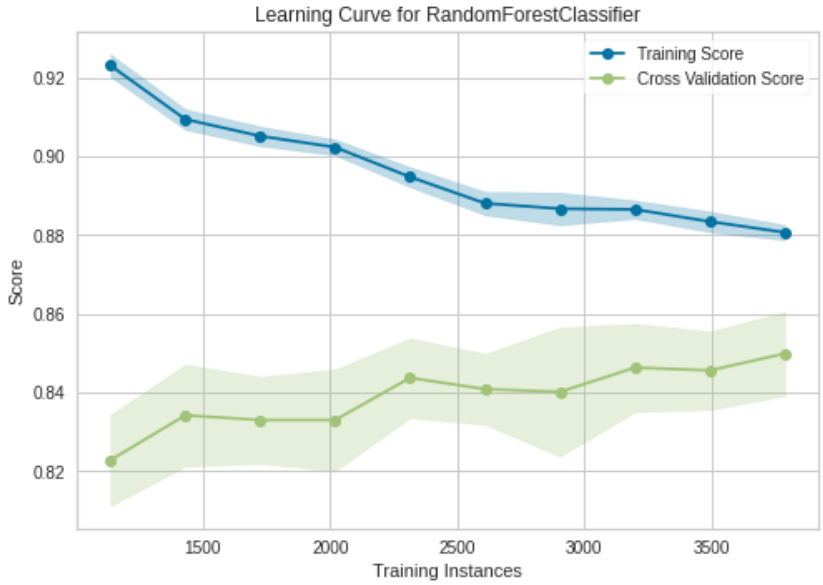


ภาพที่ 4.8 ROC Curves ของแบบจำลอง Random Forest จากชุดข้อมูลตรวจสอบ



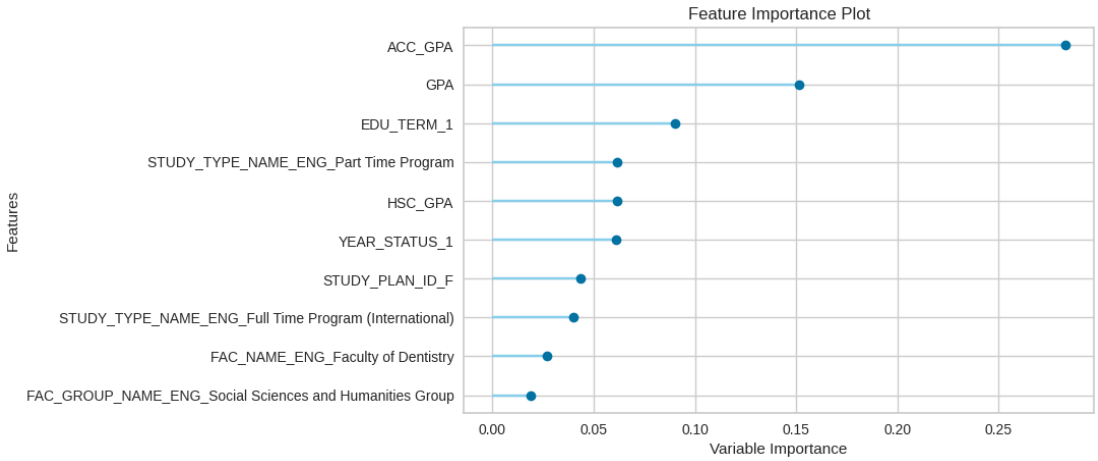
ภาพที่ 4.9 Confusion Matrix ของแบบจำลอง Random Forest จากชุดข้อมูลตรวจสอบ

จากนั้นทำการวัดประสิทธิภาพของแบบจำลอง Random Forest ของชุดข้อมูลระดับบัณฑิตศึกษา ด้วย Learning Curve พบว่าเมื่อใช้ชุดการฝึกขนาดเล็กคะแนนการฝึกจะสูงหรือมีอคติต่ำ แต่คะแนนการทดสอบต่ำหรือมีความแปรปรวนสูง กล่าวคือแบบจำลองนั้น Overfitting จากนั้นเมื่อทำการเพิ่มขนาดชุดข้อมูลการฝึกพบว่ามีอคติสูงขึ้นแต่ความแปรปรวนลดลงซึ่งหมายความว่าการเพิ่มชุดข้อมูลการฝึกแบบจำลองจะช่วยลดปัญหาการ Overfitting ลงได้ ดังภาพที่ 4.10



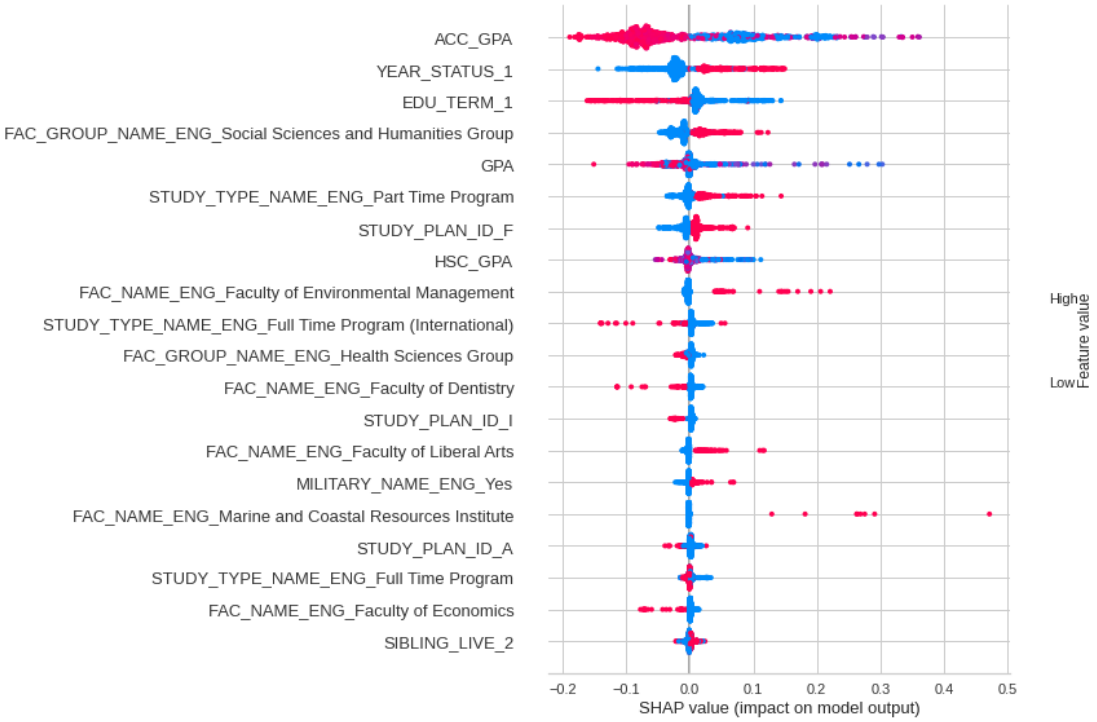
ภาพที่ 4.10 Learning curve ของแบบจำลอง Random Forest

ส่วนผลลัพธ์คุณสมบัติปัจจัยที่สำคัญ 10 อันดับแรกของแบบจำลอง Random Forest ได้แก่ ผลการเรียนเฉลี่ยสะสมเป็นปัจจัยสำคัญที่สุด รองลงมาคือผลการเรียนเฉลี่ยปัจจุบัน ภาคการศึกษาที่ 1 ประเภทภาคสมทบ ผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษา ชั้นปีที่ 1 แผนการศึกษาแผน ก แบบ ก2 ประเภทการศึกษาภาคปกติ (นานาชาติ) คณะทันตแพทยศาสตร์ กลุ่มสาขาวิชาสังคมศาสตร์และมนุษยศาสตร์ ตามลำดับ ดังภาพที่ 4.11



ภาพที่ 4.11 คุณสมบัติปัจจัยที่สำคัญ 10 อันดับแรกของแบบจำลอง Random Forest

ผลลัพธ์คุณสมบัติปัจจัยที่สำคัญจากแผนภาพ (SHAP) ที่แสดงปัจจัยเชิงบวกและเชิงลบที่สำคัญที่สุดของแต่ละปัจจัยของแบบจำลอง Random Forest โดยค่าที่สูงแทนด้วยสีแดง และค่าที่ต่ำแทนด้วยสีฟ้า พบว่าปัจจัยที่สำคัญที่สุดคือ ผลการเรียนเฉลี่ยสะสม โดยผลการเรียนที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษามากกว่า และผลการเรียนที่สูงส่งผลต่อการคงอยู่ของนักศึกษา รองลงมาคือชั้นปีที่ พบว่านักศึกษาที่ออกกลางคันในระดับบัณฑิตศึกษาส่วนใหญ่จะเกิดในช่วงชั้นปีแรกของการศึกษาเช่นเดียวกันกับนักศึกษาระดับปริญญาตรี และส่วนใหญ่นักศึกษาจะออกกลางคันในภาคการศึกษาที่สองมากกว่า โดยส่วนใหญ่กลุ่มสาขาวิชาสังคมศาสตร์และมนุษยศาสตร์มีนักศึกษาออกกลางคันมากกว่ากลุ่มสาขาวิชาวิทยาศาสตร์สุขภาพที่นักศึกษาออกกลางคันน้อยกว่า และผลการเรียนเฉลี่ยปัจจุบันที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษา และนักศึกษาประเภทภาคสมทบมีการออกกลางคันของนักศึกษามากกว่านักศึกษาภาคปกติ ส่วนปัจจัยแผนการศึกษาแผน ก แบบ ก2 จะมีการออกกลางคันมากกว่าแผนอื่น ๆ ส่วนผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษาที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษา โดยคณะที่มีนักศึกษาออกกลางคันจำนวนมากส่วนใหญ่ ได้แก่ คณะการจัดการสิ่งแวดล้อม คณะศิลปะศาสตร์ และสถาบันทรัพยากรทะเลและชายฝั่ง ซึ่งตรงกันข้ามกับคณะทันตแพทยศาสตร์ และคณะเศรษฐศาสตร์ที่ส่วนใหญ่มีนักศึกษาออกกลางคันน้อยกว่า ส่วนปัจจัย อื่น ๆ ไม่ได้ส่งผลต่อตัวแปรเป้าหมายทั้งสองค่าอย่างมีนัยสำคัญ ดังภาพที่ 4.12



ภาพที่ 4.12 คุณสมบัติปัจจัยที่สำคัญจากแผนภาพ (SHAP)

4.4 ผลลัพธ์การแสดงผลจากการนำแบบจำลองคาดการณ์ไปใช้งาน

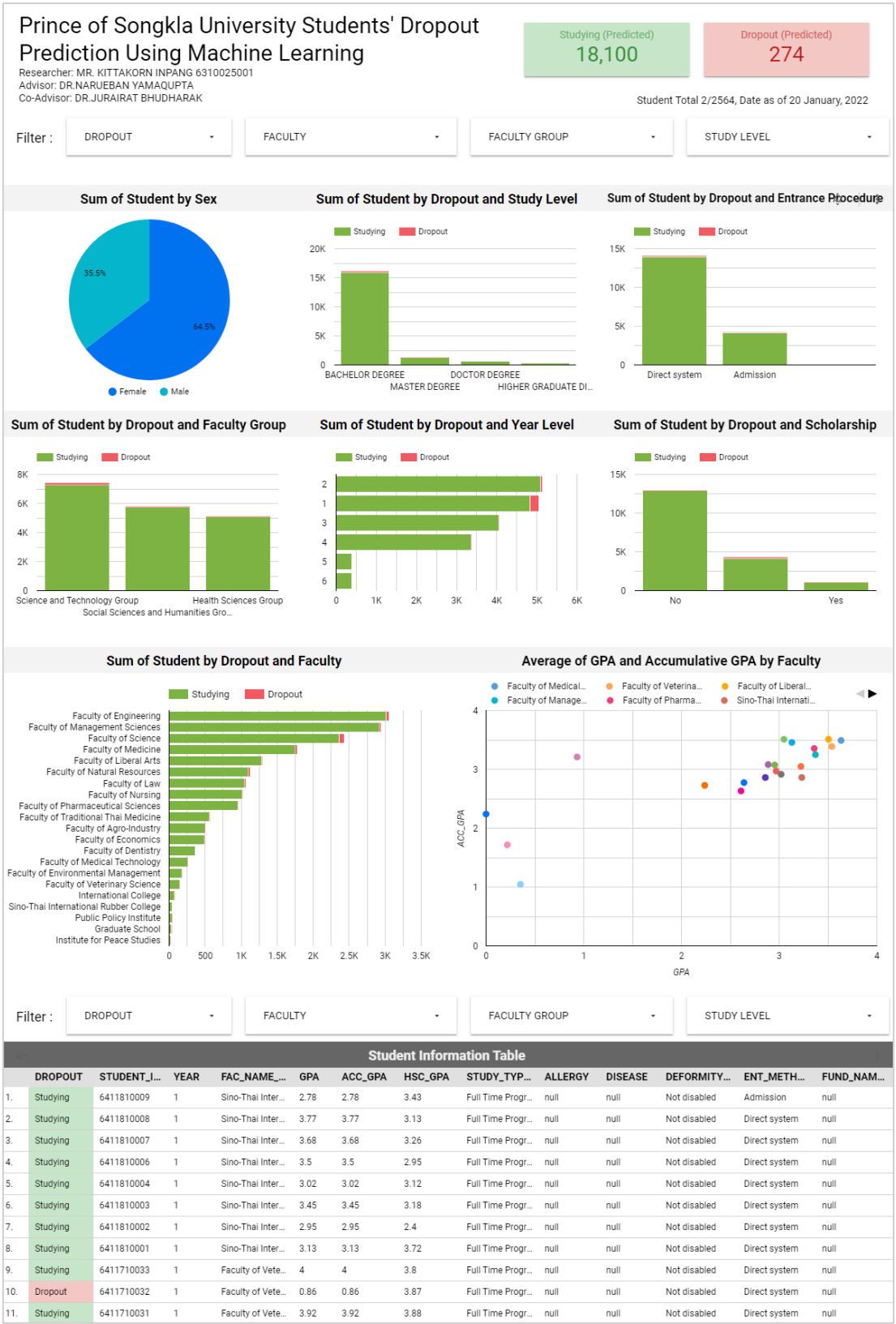
ในส่วนนี้ผู้วิจัยนำเสนอการแสดงผลจากการนำแบบจำลองคาดการณ์ไปใช้งานในรูปแบบแดชบอร์ดสำหรับการวิเคราะห์และติดตามการออกกลางคันของนักศึกษา ระดับปริญญาตรีและระดับบัณฑิตศึกษา มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ โดยมีรายละเอียดดังนี้

4.4.1 รายงานและผลการวิเคราะห์นักศึกษาคงอยู่และนักศึกษาที่ออกกลางคัน มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ สามารถแบ่งมิติการวิเคราะห์ออกเป็น 8 ด้าน คือ มิติด้านเพศ มิติด้านระดับการศึกษา มิติด้านประเภทการเข้ารับการศึกษา มิติด้านกลุ่มสาขาวิชา มิติด้านชั้นปีที่ มิติด้านการได้รับทุนการศึกษา มิติด้านคณะ และมิติด้านความสัมพันธ์ของผลการเรียนเฉลี่ยปัจจุบันและผลการเรียนเฉลี่ยสะสมตามคณะ

4.4.2 แดชบอร์ดของรายงานประกอบไปด้วย 1) แผนภูมิวงกลมแสดงสัดส่วนเพศของนักศึกษา 2) แผนภูมิแท่งแสดงจำนวนนักศึกษาที่คงอยู่และออกกลางคันตามระดับการศึกษา 3) แผนภูมิแท่งแสดงจำนวนนักศึกษาที่คงอยู่และออกกลางคันตามประเภทการเข้ารับการศึกษา 4) แผนภูมิแท่งแสดงจำนวนนักศึกษาที่คงอยู่และออกกลางคันตามกลุ่มสาขาวิชา 5) แผนภูมิแท่ง

แสดงจำนวนนักศึกษาที่คงอยู่และออกกลางคันตามชั้นปีที่ 6) แผนภูมิแท่งแสดงจำนวนนักศึกษาที่คงอยู่และออกกลางคันตามสถานะการได้รับทุนการศึกษา 7) แผนภูมิแท่งแสดงจำนวนนักศึกษาที่คงอยู่และออกกลางคันตามคณะ 8) กราฟบับเบิ้ลแสดงข้อมูลความสัมพันธ์ของผลการเรียนเฉลี่ยปัจจุบันและผลการเรียนเฉลี่ยสะสมตามคณะ และ 9) ตารางแสดงรายละเอียดข้อมูล ประกอบด้วย สถานะการออกกลางคัน รหัสนักศึกษา ชั้นปีที่ ชื่อคณะ ผลการเรียนเฉลี่ยปัจจุบัน ผลการเรียนเฉลี่ยสะสม ผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษา ประเภทการศึกษา การแพ้ยา โรคประจำตัว ความพิการ ประเภทการเข้ารับการศึกษา และการได้รับทุนการศึกษา

ผู้วิจัยแสดงตัวอย่างการวิเคราะห์ข้อมูลโดยพัฒนาและออกแบบรายงานแสดงผลด้วย Google Data Studio โดยนำข้อมูลนักศึกษาทั้งหมด ภาคการศึกษาที่ 2 ปีการศึกษา พ.ศ. 2564 มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ มาทำการคาดการณ์ด้วยแบบจำลองที่ดีที่สุดจากผลการพัฒนาแบบจำลอง และพัฒนาออกแบบรายงานแดชบอร์ดด้วย Google Data Studio พบว่าแบบจำลองคาดการณ์จำนวนนักศึกษาทั้งหมด 18,374 คน แบ่งเป็นนักศึกษาคงอยู่จำนวน 18,100 คน ร้อยละ 98.5 ออกกลางคัน 274 คน ร้อยละ 1.5 สามารถแบ่งตามระดับปริญญาตรีทั้งหมด 16,166 คน แบ่งเป็นนักศึกษาคงอยู่จำนวน 15,936 คน ร้อยละ 98.6 ออกกลางคัน 230 คน ร้อยละ 1.4 และระดับบัณฑิตศึกษาทั้งหมด 2,208 คน แบ่งเป็นนักศึกษาคงอยู่จำนวน 2,164 คน ร้อยละ 98.0 ออกกลางคัน 44 คน ร้อยละ 2.0 ตามลำดับ แสดงดังภาพที่ 4.13



	DROPOUT	STUDENT_I...	YEAR	FAC_NAME_...	GPA	ACC_GPA	HSC_GPA	STUDY_TYP...	ALLERGY	DISEASE	DEFORMITY...	ENT_METH...	FUND_NAM...
1.	Studying	6411810009	1	Sino-Thai Inter...	2.78	2.78	3.43	Full Time Progr...	null	null	Not disabled	Admission	null
2.	Studying	6411810008	1	Sino-Thai Inter...	3.77	3.77	3.13	Full Time Progr...	null	null	Not disabled	Direct system	null
3.	Studying	6411810007	1	Sino-Thai Inter...	3.68	3.68	3.26	Full Time Progr...	null	null	Not disabled	Direct system	null
4.	Studying	6411810006	1	Sino-Thai Inter...	3.5	3.5	2.95	Full Time Progr...	null	null	Not disabled	Direct system	null
5.	Studying	6411810004	1	Sino-Thai Inter...	3.02	3.02	3.12	Full Time Progr...	null	null	Not disabled	Direct system	null
6.	Studying	6411810003	1	Sino-Thai Inter...	3.45	3.45	3.18	Full Time Progr...	null	null	Not disabled	Direct system	null
7.	Studying	6411810002	1	Sino-Thai Inter...	2.95	2.95	2.4	Full Time Progr...	null	null	Not disabled	Direct system	null
8.	Studying	6411810001	1	Sino-Thai Inter...	3.13	3.13	3.72	Full Time Progr...	null	null	Not disabled	Direct system	null
9.	Studying	6411710033	1	Faculty of Vete...	4	4	3.8	Full Time Progr...	null	null	Not disabled	Direct system	null
10.	Dropout	6411710032	1	Faculty of Vete...	0.86	0.86	3.87	Full Time Progr...	null	null	Not disabled	Direct system	null
11.	Studying	6411710031	1	Faculty of Vete...	3.92	3.92	3.88	Full Time Progr...	null	null	Not disabled	Direct system	null

ภาพที่ 4.13 แดชบอร์ดรายงานและผลการวิเคราะห์นักศึกษาคงอยู่และนักศึกษาที่ออกกลางคัน มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

บทที่ 5

สรุปและข้อเสนอแนะ

ในบทนี้ประกอบด้วยสองส่วนหลัก ได้แก่ ส่วนแรกเป็นการสรุปผลการดำเนินงานวิจัยนี้ และส่วนสุดท้ายเป็นข้อจำกัดและข้อเสนอแนะงานวิจัยในอนาคตเกี่ยวกับการคาดการณ์ปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษา ด้วยการเรียนรู้ของเครื่อง ดังรายละเอียดต่อไปนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอรูปแบบคุณลักษณะที่สำคัญ และคาดการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ในช่วง 6 รุ่นปีการศึกษา พ.ศ. 2558 ถึง 2563 จากระบบข้อมูลพื้นฐานนักศึกษา และระบบข้อมูลนักศึกษา โดยใช้ข้อมูลผลลัพธ์ทางการศึกษา ข้อมูลส่วนตัวของนักศึกษา และข้อมูลครอบครัวของนักศึกษา ผ่านเทคนิคเหมืองข้อมูลและแบบจำลองการเรียนรู้ของเครื่อง 5 แบบ และนำเสนอรายงานแดชบอร์ดในรูปแบบจินตทัศน์

ในการศึกษาวิจัยนี้ใช้เทคนิคการทำเหมืองข้อมูลและเครื่องมือ Google Colab Notebook ในการนำเข้าข้อมูล จัดการข้อมูล สร้างแบบจำลอง และวัดประสิทธิภาพแบบจำลองเพื่อนำไปใช้งาน โดยทำความเข้าใจข้อมูลที่ได้ขอความอนุเคราะห์ข้อมูลจากกองนโยบาย ยุทธศาสตร์ และแผน มหาวิทยาลัยสงขลานครินทร์ จากระบบลงทะเบียนและระบบข้อมูลพื้นฐานนักศึกษา มีขอบเขตเนื้อหาของงานวิจัย คือ 1. ปัจจัยผลลัพธ์ทางการศึกษา 2. ปัจจัยข้อมูลพื้นฐานของนักศึกษา และ 3. ปัจจัยด้านครอบครัวของนักศึกษา โดยทำการเชื่อมโยงมิติของข้อมูล คัดเลือกข้อมูลนักศึกษาที่มีสถานะล่าสุดที่กำลังศึกษาอยู่และพ้นสภาพการเป็นนักศึกษา และส่งออกข้อมูลจำนวนทั้งสิ้น 33,930 ราย 39 ตัวแปร จากนั้นทำการจัดเตรียมข้อมูลให้สมบูรณ์ก่อนสร้างแบบจำลองและทำการแบ่งชุดข้อมูลออกเป็นสองชุด คือ ชุดข้อมูลระดับปริญญาตรีและชุดข้อมูลระดับบัณฑิตศึกษา จากนั้นทำการเลือกคุณสมบัติที่มีนัยสำคัญทางสถิติที่ P-value ≤ 0.05 ของตัวแปรกลุ่มด้วยวิธี Chi-Squared และตัวแปรต่อเนื่องด้วยวิธี ANOVA ของข้อมูลทั้งสองชุดพบว่าช่วยให้แบบจำลองคาดการณ์มีความแม่นยำมากขึ้น รวมถึงจัดการข้อมูลที่ไม่สมดุลของข้อมูลทั้งสองชุดด้วยเทคนิควิธีสังเคราะห์ข้อมูลเพิ่ม

ในขั้นตอนการสร้างแบบจำลองผู้วิจัยใช้อัลกอริทึมการจำแนกสำหรับการคาดการณ์การออกกลางคันของนักศึกษา โดยใช้อัลกอริทึมการเรียนรู้ของเครื่องประเภทต้นไม้ตัดสินใจ 5 แบบได้แก่ 1. ต้นไม้การตัดสินใจ (Decision Tree) 2. แรนดอมฟอเรสต์ (Random Forest) 3. ไลท์กาเดียนบูทติ้งแมชชีน (Light Gradient Boosting Machine) 4. กาเดียนบูทติ้ง (Gradient Boosting) และ 5. เอ็กซ์ตรีม-กาเดียนบูทติ้ง (Extreme Gradient Boosting) โดยทำการแบ่งข้อมูลสำหรับเรียนรู้แบบจำลอง 80% และสำหรับตรวจสอบแบบจำลอง 20% จากนั้นทำการประเมินประสิทธิภาพแบบจำลองด้วยเทคนิค 10-folds Cross Validation ด้วยการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ และประเมินผลแบบจำลองด้วยตารางเมทริกซ์ความสับสน (Confusion Matrix) ค่าความถูกต้อง (Accuracy) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) ค่าความถ่วงดุล (F1 Scores) และค่าพื้นที่ใต้กราฟ (AUC Scores) โดยผู้วิจัยเลือกการวัดประสิทธิภาพของแบบจำลองด้วยค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าพื้นที่ใต้เส้น Receiver Operating Characteristic Curve (ROC Curve) เพื่อแสดงประสิทธิภาพของแบบจำลองคาดการณ์ในการจำแนกกลุ่มออกจากกันได้ถูกต้อง

ผลการพัฒนาแบบจำลองคาดการณ์ชุดข้อมูลระดับชั้นปริญญาตรีแสดงให้เห็นว่าแบบจำลอง Light Gradient Boosting Machine ที่ปรับปรุงประสิทธิภาพแบบจำลองและไม่ได้ปรับความสมดุลของข้อมูลมีประสิทธิภาพสูงสุด โดยให้ค่าพื้นที่ใต้กราฟ (AUC Scores) สูงที่สุด มีค่า 0.9303 ค่าความถูกต้อง (Accuracy) ส่วน มีค่า 0.8999 และค่าความถ่วงดุล (F1 Scores) มีค่า 0.4957 ตามลำดับ

ส่วนผลลัพธ์คุณลักษณะที่สำคัญต่ออัตราการออกกลางคันของนักศึกษาระดับปริญญาตรี พบว่าส่วนใหญ่เกี่ยวข้องกับปัจจัยทางด้านผลลัพธ์ทางการศึกษาโดยปัจจัยที่สำคัญที่สุดได้แก่ ผลการเรียนเฉลี่ยสะสม และผลการเรียนเฉลี่ยปัจจุบัน โดยผลการเรียนที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษามากกว่าผลการเรียนที่สูง และส่วนใหญ่นักศึกษาจะออกกลางคันในภาคการศึกษาแรก เช่นเดียวกับชั้นปีที่นักศึกษาส่วนใหญ่ออกกลางคันในช่วงชั้นปีแรกมากกว่า ซึ่งสอดคล้องกับงานวิจัยที่ได้ศึกษาที่ส่วนใหญ่พบว่าปัจจัยสำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษาคือผลลัพธ์ทางการศึกษา ได้แก่ ผลการเรียนเฉลี่ยที่ต่ำ [11], [39] และผลการเรียนเฉลี่ยก่อนเข้าศึกษา [8], [38] ภาคการศึกษาแรก [33] และช่วงชั้นปีแรกที่เข้าศึกษา [36], [37] ส่วนปัจจัยทางด้านการเงินพบว่าปัจจัยส่วนใหญ่ไม่ได้ส่งผลต่อตัวแปรเป้าหมายอย่างมีนัยสำคัญ แต่พบว่ารายได้เฉลี่ยของมารดาที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษา ส่วนปัจจัยอื่น ๆ และปัจจัยด้านสุขภาพและการได้รับทุนการศึกษาของนักศึกษาไม่ส่งผลต่อตัวแปรเป้าหมายอย่างมีนัยสำคัญทางสถิติ

ส่วนผลการพัฒนาแบบจำลองคาดการณ์ชุดข้อมูลระดับชั้นบัณฑิตศึกษาแสดงให้เห็นว่าแบบจำลอง Random Forest ที่ปรับปรุงประสิทธิภาพแบบจำลองและไม่ได้ปรับความสมดุลของข้อมูลมีประสิทธิภาพสูงสุด โดยให้ค่าพื้นที่ใต้กราฟ (AUC Scores) และค่าความถูกต้อง (Accuracy)

สูงที่สุด มีค่า 0.7886 และ 0.8528 และค่าความถ่วงดุล (F1 Scores) ของแบบจำลอง Random Forest มีค่า 0.4358 ตามลำดับ

ส่วนผลลัพธ์ส่วนคุณลักษณะที่สำคัญต่ออัตราการออกกลางคันของนักศึกษาระดับบัณฑิตศึกษา พบว่าส่วนใหญ่เกี่ยวข้องกับปัจจัยทางด้านผลลัพธ์ทางการศึกษาเช่นเดียวกับระดับปริญญาตรี โดยปัจจัยที่สำคัญที่สุด ได้แก่ ผลการเรียนเฉลี่ยสะสม โดยผลการเรียนที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษามากกว่าผลการเรียนที่สูง รองลงมาคือพบว่าส่วนใหญ่นักศึกษาระดับบัณฑิตศึกษาจะออกกลางคันในช่วงชั้นปีแรกของการศึกษาเช่นเดียวกับระดับปริญญาตรี และส่วนใหญ่อัตราการออกกลางคันจะเกิดขึ้นในภาคการศึกษาที่สองมากกว่าภาคการศึกษาแรก ซึ่งส่วนใหญ่นักศึกษาในกลุ่มสาขาวิชาสังคมศาสตร์และมนุษยศาสตร์มีอัตราการออกกลางคันมากกว่ากลุ่มสาขาวิชาวิทยาศาสตร์สุขภาพที่นักศึกษาออกกลางคันน้อยกว่า และผลการเรียนเฉลี่ยปัจจุบันที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษา และนักศึกษาประเภทภาคสมทบมีอัตราการออกกลางคันของนักศึกษาสูงกว่านักศึกษาภาคปกติ ส่วนปัจจัยแผนการศึกษาแผน ก แบบ ก2 จะมีการออกกลางคันมากกว่าแผนอื่น ๆ และผลการเรียนเฉลี่ยสะสมก่อนเข้าศึกษาที่ต่ำส่งผลต่อการออกกลางคันของนักศึกษา โดยคณะที่มีอัตราการออกกลางคันของนักศึกษาส่วนใหญ่ ได้แก่ คณะการจัดการสิ่งแวดล้อม คณะศิลปะศาสตร์ และสถาบันทรัพยากรทะเลและชายฝั่ง ซึ่งตรงกันข้ามกับคณะทันตแพทยศาสตร์ และคณะเศรษฐศาสตร์ที่ส่วนใหญ่มีนักศึกษาออกกลางคันน้อยกว่า ส่วนปัจจัย อื่น ๆ ไม่ได้ส่งผลต่อตัวแปรเป้าหมายทั้งสองค่าอย่างมีนัยสำคัญ ซึ่งสอดคล้องกับงานวิจัยที่ได้ศึกษาที่พบว่าปัจจัยสำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษาคือผลลัพธ์ทางการศึกษา ได้แก่ ผลการเรียนเฉลี่ยที่ต่ำ [11], [39] ผลการเรียนเฉลี่ยก่อนเข้าศึกษา [8], [38] ช่วงชั้นปีแรกที่เข้าศึกษา [36], [37] รวมถึงปัจจัยด้านคณะ [8], [33] ส่วนปัจจัยอื่น ๆ ได้แก่ ปัจจัยทางครอบครัว ปัจจัยทางด้านสุขภาพ และการได้รับทุนการศึกษาของนักศึกษาส่วนใหญ่ไม่ส่งผลต่อตัวแปรเป้าหมายอย่างมีนัยสำคัญทางสถิติ

ในส่วนของการนำเสนอการแสดงผลจากการนำแบบจำลองคาดการณ์ไปใช้งานในรูปแบบแดชบอร์ดสำหรับการวิเคราะห์และติดตามการออกกลางคันของนักศึกษา ผู้วิจัยพัฒนาและออกแบบรายงานแสดงผลด้วย Google Data Studio โดยนำข้อมูลนักศึกษาทั้งหมด ภาคการศึกษาที่ 2 ปีการศึกษา พ.ศ. 2564 มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ มาทำการคาดการณ์ด้วยแบบจำลองที่ดีที่สุดจากผลการพัฒนาแบบจำลอง พบว่าแบบจำลองสามารถคาดการณ์จำนวนนักศึกษาทั้งหมด 18,374 คน แบ่งเป็นนักศึกษาคงอยู่จำนวน 18,100 คน ร้อยละ 98.5 ออกกลางคัน 274 คน ร้อยละ 1.5 สามารถแบ่งตามระดับปริญญาตรีทั้งหมด 16,166 คน แบ่งเป็นนักศึกษาคงอยู่จำนวน 15,936 คน ร้อยละ 98.6 ออกกลางคัน 230 คน ร้อยละ 1.4 และระดับบัณฑิตศึกษาทั้งหมด 2,208 คน แบ่งเป็นนักศึกษาคงอยู่จำนวน 2,164 คน ร้อยละ 98.0 ออกกลางคัน 44 คน ร้อยละ 2.0

ตามลำดับ และสามารถแบ่งมิติการวิเคราะห์ออกเป็น 8 ด้าน คือ มิติด้านเพศ มิติด้านระดับการศึกษา มิติด้านประเภทการเข้ารับการศึกษา มิติด้านกลุ่มสาขาวิชา มิติด้านชั้นปีที่ มิติด้านการได้รับทุนการศึกษา มิติด้านคณะ และมิติด้านความสัมพันธ์ของผลการเรียนเฉลี่ยปัจจุบันและผลการเรียนเฉลี่ยสะสมตามคณะ โดยสามารถแสดงผลรายงานแดชบอร์ดในรูปแบบจินตทัศน์เพื่อติดตามความเสี่ยง ซึ่งจะช่วยให้เจ้าหน้าที่ที่เกี่ยวข้องสามารถเข้าช่วยเหลือนักศึกษาที่มีความเสี่ยงได้ทันที และเพื่อช่วยผู้บริหารในการสนับสนุนการตัดสินใจและวางแผนการบริหารงานเพื่อลดอัตราการออกกลางคันในแต่ละปีการศึกษาของมหาวิทยาลัยให้ต่ำลงได้

5.2 ปัญหาและข้อจำกัดของการวิจัย

5.2.1 ขาดข้อมูลปัจจัยในการวิเคราะห์ ได้แก่ ปัจจัยทางด้านผลการเรียนในรายวิชาและปัจจัยสภาพแวดล้อมของนักศึกษา เพื่อให้การวิเคราะห์ที่ครอบคลุมทั้งหมด

5.2.2 ข้อมูลมีค่าสูญหายและไม่ถูกต้อง ผู้วิจัยต้องใช้กระบวนการหลายวิธีในการจัดการรวมถึงต้องใช้ทรัพยากรในการประมวลผลที่สูง

5.3 ข้อเสนอแนะ

ข้อเสนอแนะจากผลการศึกษาปัจจัยที่มีผลต่ออัตราการออกกลางคันของนักศึกษามหาวิทยาลัยสงขลานครินทร์ และเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องประเภทต้นไม้ 5 แบบ เพื่อให้ผู้วิจัยและผู้พัฒนานำข้อมูลที่ได้ไปศึกษาและพัฒนาเพิ่มเติมมีรายละเอียดดังนี้

5.3.1 นำข้อมูลผลการเรียนของรายวิชามา เช่น วิชาในหมวดวิทยาศาสตร์และเทคโนโลยี หมวดสังคมศาสตร์ หรือหมวดอื่น ๆ มาวิเคราะห์เพิ่มเติมเพื่อหาผลลัพธ์ที่ซ่อนอยู่ซึ่งจะช่วยให้แบบจำลองได้เรียนรู้และวิเคราะห์ได้ละเอียดยิ่งขึ้น และช่วยให้สามารถวิเคราะห์รายวิชาที่ส่งผลต่อการออกกลางคันหรือคงอยู่ของนักศึกษาได้

5.3.2 ควรเพิ่มชุดข้อมูลสำหรับการเรียนรู้แบบจำลองให้มากขึ้น เช่น ข้อมูลนักศึกษาวิทยาเขตอื่น ๆ ของมหาวิทยาลัยสงขลานครินทร์ ซึ่งจะช่วยลดความแปรปรวนและเพิ่มประสิทธิภาพให้กับแบบจำลองรวมถึงลด Overfitting ของแบบจำลอง

บรรณานุกรม

[1] Office of the Education Council - Ministry of Education, *Education in Thailand 2018*. 2018.

[2] C. Yongsorn, “EDUCATIONAL WASTE AFFECTS THE QUAITY OF GRADUATES,” *Journal of Education Faculty of Education Srinakharinwirot University*, vol. 18, no. 1, pp. 1-9. Thai, 2017.

[3] G. M. (Glenda M. Crosling, M. Heagney, and L. (Elizabeth) Thomas, *Improving student retention in higher education : the role of teaching and learning*, 1st ed. Routledge, 2008.

[4] J. Luan, “Executive brief Data Mining Applications in Higher Education,” *Insol.Lt*, 2006, Accessed: Jul. 25, 2021. [Online]. Available: www.spss.com/downloads.

[5] S. Natek and M. Zwilling, “Student data mining solution-knowledge management system related to higher education institutions,” *Expert Systems with Applications*, vol. 41, no. 14, pp. 6400–6407, 2014, doi: 10.1016/j.eswa.2014.04.024.

[6] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, “Educational data mining and learning analytics for 21st century higher education: A review and synthesis,” *Telematics and Informatics*, vol. 37, no. April 2018, pp. 13–49, 2019, doi: 10.1016/j.tele.2019.01.007.

[7] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[8] W. Tenpipat and K. Akkarajitsakul, “Student Dropout Prediction: A KMUTT Case Study,” *2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020*, 2020, doi: 10.1109/IBDAP50342.2020.9245457.

[9] M. Alban and D. Mauricio, “Neural networks to predict dropout at the universities,” *International Journal of Machine Learning and Computing*, vol.

9, no. 2, pp. 149–153, 2019, doi: 10.18178/ijmlc.2019.9.2.779.

[10] I. Sandoval-Palis, D. Naranjo, J. Vidal, and R. Gilar-Corbi, “Early dropout prediction model: A case study of university leveling course students,” *Sustainability (Switzerland)*, vol. 12, no. 22, pp. 1–17, 2020, doi: 10.3390/su12229314.

[11] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, “Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile,” *Entropy*, vol. 23, no. 4, pp. 1–23, 2021, doi: 10.3390/e23040485.

[12] “สถิตินักศึกษา Online --- มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่.” https://reg.psu.ac.th/StatStudentHatYai/index.aspx (accessed Oct. 04, 2021).

[13] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015.

[14] B. Leventhal, “An introduction to data mining and other techniques for advanced analytics,” *Journal of Direct, Data and Digital Marketing Practice*, vol. 12, no. 2, pp. 137–153, 2010, doi: 10.1057/dddmp.2010.35.

[15] “Machine Learning: What it is and why it matters | SAS,” 2021. https://www.sas.com/en_us/insights/analytics/machine-learning.html (accessed Jul. 29, 2021).

[16] K. A. Gaurav and L. Patel, *Machine Learning With R*. 2020.

[17] B. Lantz, *Machine Learning with R Second Edition*, Second Edi. 2015.

[18] G. Bonaccorso, *Machin Learning Algorithm*, vol. 49, no. 23–6. 2017.

[19] “Machine Learning Algorithms | Know Top 8 Machine Learning Algorithms.” https://www.educba.com/machine-learning-algorithms (accessed Jul. 29, 2021).

[20] M. Sewell, *Ensemble methods - Zhou*, vol. 2, no. Schapire 1990. 2007.

[21] K. Frank, *Hands-On Data Science and Python Machine Learning*. 2017.

[22] M. Friendly, D. Meyer, and A. Zeileis, *Discrete Data Analysis with R*. 2015.

[23] T. et. all. Hastie, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction.,” *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2017, [Online]. Available: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf.

[24] G. Ke *et al.*, “LightGBM: A highly efficient gradient boosting decision tree,”

Advances in Neural Information Processing Systems, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.

[25] “Features — LightGBM 3.3.2.99 documentation.” https://lightgbm.readthedocs.io/en/latest/Features.html (accessed May 11, 2022).

[26] C. Bentéjac, A. Csörg**ő**, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.

[27] T. S. Ng, “Machine learning,” *Studies in Systems, Decision and Control*, vol. 65, pp. 121–151, 2016, doi: 10.1007/978-981-10-1509-0_9.

[28] ดร. เยาวลักษณ์ ชาติบัญชาชัย and คุณโสภณ เพิ่มศิริวัลลภ, “คำถามที่พบบ่อย (FAQ) เกี่ยวกับ Big data และ Data analytics.”

[29] อุดมธนะธีระ เกียรติพงษ์, “BI องค์ประกอบของ Business Intelligence (BI).” https://www.iok2u.com/index.php/article/information-technology/1047-bi-business-intelligence-bi (accessed Apr. 21, 2022).

[30] “What are Dashboards? - BI and Analytics Tools | Types & Examples.” https://www.yellowfinbi.com/glossary/dashboards (accessed Apr. 21, 2022).

[31] “Types of Business Dashboards You Must Know - Ubiq BI.” https://ubiq.co/analytics-blog/types-business-dashboards-must-know/ (accessed Apr. 21, 2022).

[32] J. P. Bean, “College Student Retention - Defining Student Retention, A Profile of Successful Institutions and Students, Theories of Student Departure - Factors, School, Model, and Social - StateUniversity.com,” 2013. https://education.stateuniversity.com/pages/1863/College-Student-Retention.html (accessed Jul. 29, 2021).

[33] K. Tentsho, R. McNeil, and P. Tongkumchum, “Determinants of University Dropout: A Case of Thailand,” *Asian Social Science*, vol. 15, no. 7, p. 49, 2019, doi: 10.5539/ass.v15n7p49.

[34] T. Yu and J. C. Richardson, “An exploratory factor analysis and reliability analysis of the student online learning readiness (SOLR) instrument,” *Online Learning Journal*, vol. 19, no. 5, 2015, doi: 10.24059/olj.v19i5.593.

[35] ค. เมฆขลา, “การพัฒนาระบบการช่วยเหลือนักเรียนโรงเรียนโปลิเทคนิคลานนา,” 2552.

[36] B. Hanthongchai, “Factors Affecting to Drop out and Survival Pathways of First Year Undergraduate Students of Institute of Physical Education Udonthani,” *MBU Education Journal*, vol. 7, no. 2, pp. 247–271, 2019.

[37] W. Taipjutorus, “Reducing Attrition Rate of First-year Undergraduate Students: An Implementation of RMUTP Pre-University Program,” pp. 252–258, 2017.

[38] B. Mahatthanachai, H. Ninsonti, and N. Tantranont, “A Study of Factors Influency Student Dropout Rate Using Data Mining,” *The Golden Teak : Humanity and Social Science Journal*, vol. 22, no. 4, pp. 46–55, 2016, [Online]. Available: https://www.tci-thaijo.org/index.php/tgt/article/view/88196.

[39] N. Theppalak, “The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining :,” *Journal of University of Babylon for Pure and Applied Sciences*, vol. 27, no. 1, pp. 356–368, 2019.

[40] W. F. Wan Yaacob, N. Mohd Sobri, S. A. M. Nasir, W. F. Wan Yaacob, N. D. Norshahidi, and W. Z. Wan Husin, “Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques,” *Journal of Physics: Conference Series*, vol. 1496, no. 1, 2020, doi: 10.1088/1742-6596/1496/1/012005.

[41] K. Limsathitwong, K. Tiwatthanont, and T. Yatsungnoen, “Dropout prediction system to reduce discontinue study rate of information technology students,” *Proceedings of 2018 5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science, ICBIR 2018*, pp. 110–114, 2018, doi: 10.1109/ICBIR.2018.8391176.

[42] P. Nuankaew, W. Nuankaew, and P. Nasa-Ngium, “Risk Management Models for Prediction of Dropout Students in Thailand Higher Education,” *International Journal of Innovation, Creativity and Change. www.ijicc.net*, vol. 15, no. 3, p. 2021, 2021, [Online]. Available: www.ijicc.net.

[43] R. Wirth, “CRISP-DM : Towards a Standard Process Model for Data Mining,” *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, no. 24959, pp. 29–39, 2000.

[44] F. Bießmann *et al.*, “DataWig: Missing Value Imputation for Tables,” *Journal of Machine Learning Research*, vol. 20, pp. 1–6, 2019, Accessed: May 16, 2022. [Online]. Available: http://jmlr.org/papers/v20/18-753.html.

[45] P. Dangeti, *Statistics for Machine Learning*. 1967.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. Sept. 28, pp. 321–357, 2002.

[47] “PyCaret Guide - PyCaret.” https://pycaret.org/guide/ (accessed Jul. 27, 2021).

[48] M. Ali, “Introduction to Regression in Python with PyCaret,” 2021. https://towardsdatascience.com/introduction-to-regression-in-python-with-pycaret-d6150b540fc4 (accessed Jul. 30, 2021).

[49] “Optuna: A hyperparameter optimization framework — Optuna 2.10.0 documentation.” https://optuna.readthedocs.io/en/stable/ (accessed Apr. 03, 2022).

[50] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” 2019, Accessed: Apr. 03, 2022. [Online]. Available: https://github.com/pfnet/optuna/.

[51] A. Pérez *et al.*, *Mastering Text mining in R*, vol. 5, no. 1. 2017.

[52] S. Bulathwela, M. Pérez-Ortiz, A. Lipani, E. Yilmaz, and J. Shawe-Taylor, “Predicting Engagement in Video Lectures,” *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 2020.

ภาคผนวก ก
ตารางข้อมูลประชากรของตัวแปรเป้าหมาย DROPOUT

ตารางที่ ก.1 สรุปรายละเอียดประชากรของตัวแปรเป้าหมาย DROPOUT ชุดข้อมูลระดับปริญญาตรี

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
รวมระดับปริญญาตรี	**25,031**	**87.3%**	**3,635**	**12.7%**
ภาคการศึกษาที่				
1	2,484	8.7%	1,662	5.8%
2	22,547	78.7%	1,973	6.9%
ชั้นปีที่				
1	6,820	23.8%	2,292	8.0%
2	4,440	15.5%	985	3.4%
3	3,519	12.3%	213	0.7%
4	9,532	33.3%	142	0.5%
5	353	1.2%	3	0.0%
6	367	1.3%		0.0%
คณะ				
Faculty of Agro-Industry	751	2.6%	106	0.4%
Faculty of Dentistry	281	1.0%	12	0.0%
Faculty of Economics	761	2.7%	114	0.4%
Faculty of Engineering	3,971	13.9%	616	2.2%
Faculty of Law	1,678	5.9%	312	1.1%
Faculty of Liberal Arts	1,752	6.1%	144	0.5%
Faculty of Management Sciences	4,542	15.8%	442	1.5%
Faculty of Medical Technology	396	1.4%	64	0.2%
Faculty of Medicine	1,508	5.3%	62	0.2%
Faculty of Natural Resources	1,784	6.2%	304	1.1%
Faculty of Nursing	1,264	4.4%	62	0.2%
Faculty of Pharmaceutical Sciences	916	3.2%	131	0.5%
Faculty of Science	4,262	14.9%	1,071	3.7%
Faculty of Traditional Thai Medicine	798	2.8%	123	0.4%
Faculty of Veterinary Science	152	0.5%	8	0.0%
International College	162	0.6%	51	0.2%
Sino-Thai International Rubber College	53	0.2%	13	0.1%
กลุ่มสาขาวิชา				

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Health Sciences Group	5,315	18.5%	462	1.6%
Science and Technology Group	10,983	38.3%	2,161	7.5%
Social Sciences and Humanities Group	8,733	30.5%	1,012	3.5%
ประเภทการศึกษา				
Full Time Program	24,218	84.5%	3,427	12.0%
Full Time Program (International)	215	0.8%	64	0.2%
Full Time Program (Special)	36	0.1%	23	0.1%
Joint degree	7	0.0%		0.0%
Part Time Program	555	1.9%	121	0.4%
เพศ				
Female	16,328	57.0%	2,086	7.3%
Male	8,703	30.4%	1,549	5.4%
การเกณฑ์ทหาร				
No	16,205	56.5%	2,175	7.6%
Unknown	31	0.1%	9	0.0%
Yes	8,795	30.7%	1,451	5.1%
ศาสนา				
Buddhism	20,063	70.0%	2,754	9.6%
Christian	204	0.7%	22	0.1%
Islam	4,661	16.3%	838	2.9%
Other	6	0.0%	1	0.0%
Undefined	97	0.3%	20	0.1%
โรคประจำตัว				
No	21,031	73.4%	3,019	10.5%
Unknown	1	0.0%	1	0.0%
Yes	3,999	14.0%	615	2.1%
แพ้ยา				
No	23,969	83.6%	3,481	12.1%
Unknown	1	0.0%	1	0.0%
Yes	1,061	3.7%	153	0.5%
สถานภาพสมรส				

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Divorced	28	0.1%	6	0.0%
Married	91	0.3%	29	0.1%
Separated	4	0.0%	1	0.0%
Single	24,677	86.1%	3,552	12.4%
Unknown	213	0.7%	44	0.2%
Widowed	18	0.1%	3	0.0%
ระดับการศึกษาก่อนหน้า				
Bachelor Degree	608	2.1%	155	0.5%
Diploma	34	0.1%	3	0.0%
Doctoral Degree	5	0.0%		0.0%
Graduate Diploma	1	0.0%		0.0%
High Vocational Certificate	187	0.7%	25	0.1%
Higher Graduate Diploma	17	0.1%	1	0.0%
Junior High School	22	0.1%	2	0.0%
Master Degree	19	0.1%	10	0.0%
Postgrad Diploma	7	0.0%	5	0.0%
Primary Education	6	0.0%	2	0.0%
Senior High School	24,011	83.8%	3,415	11.9%
Unknown	32	0.1%	9	0.0%
Vocational Certificate	82	0.3%	8	0.0%
ประเภทการเข้ารับ				
Admission	8,880	31.0%	1,569	5.5%
Direct system	16,151	56.3%	2,066	7.2%
จังหวัดที่เกิด				
SONGKHLA	9,445	32.9%	1,352	4.7%
NAKHON SI THAMMARAT	2,514	8.8%	289	1.0%
YALA	1,794	6.3%	323	1.1%
TRANG	1,747	6.1%	247	0.9%
PHATTHALUNG	1,195	4.2%	163	0.6%
SURAT THANI	1,186	4.1%	151	0.5%
NARATHIWAT	1,168	4.1%	210	0.7%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
PATTANI	1,161	4.1%	223	0.8%
BANGKOK	942	3.3%	130	0.5%
SATUN	647	2.3%	107	0.4%
PHUKET	644	2.2%	71	0.2%
Others	2,588	9.0%	369	1.3%
จำนวนพี่น้อง				
0	108	0.4%	18	0.1%
1	3,528	12.3%	470	1.6%
2	12,283	42.8%	1,709	6.0%
3	6,270	21.9%	934	3.3%
4	1,656	5.8%	291	1.0%
5	596	2.1%	106	0.4%
6	290	1.0%	49	0.2%
7	126	0.4%	23	0.1%
8	87	0.3%	23	0.1%
9	47	0.2%	6	0.0%
10	25	0.1%	3	0.0%
11	7	0.0%	1	0.0%
12	7	0.0%	1	0.0%
33	1	0.0%	1	0.0%
จำนวนพี่น้องที่กำลังศึกษาอยู่				
0	204	0.7%	41	0.1%
1	8,922	31.1%	1,257	4.4%
2	11,159	38.9%	1,588	5.5%
3	3,725	13.0%	570	2.0%
4	677	2.4%	130	0.5%
5	211	0.7%	32	0.1%
6	74	0.3%	9	0.0%
7	30	0.1%	7	0.0%
8	19	0.1%		0.0%
9	7	0.0%	1	0.0%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
10	2	0.0%		0.0%
21	1	0.0%		0.0%
ได้รับทุน				
No	23,628	82.4%	3,475	12.1%
Unknown	8	0.0%	1	0.0%
Yes	1,395	4.9%	159	0.6%
ความพิการ				
Disability hearing	5	0.0%	2	0.0%
Not disabled	24,992	87.2%	3,625	12.6%
Physical disability	20	0.1%	4	0.0%
Visual disability	14	0.0%	4	0.0%
ระดับการศึกษาบิดา				
Bachelor Degree	5,801	20.2%	889	3.1%
Doctoral Degree	127	0.4%	14	0.0%
Less than bachelor's degree	14,041	49.0%	1,897	6.6%
Master Degree	1,454	5.1%	234	0.8%
Unknown	3,608	12.6%	601	2.1%
สถานภาพบิดา				
Alive	23,263	81.2%	3,357	11.7%
Deceased	1,566	5.5%	251	0.9%
Disability	30	0.1%	5	0.0%
Disabled	63	0.2%	6	0.0%
Others	109	0.4%	16	0.1%
อาชีพบิดา				
Agricultural Sector	4,860	17.0%	674	2.4%
Contractor	3,988	13.9%	515	1.8%
Corporate Employee	1,091	3.8%	124	0.4%
Government Sector	4,669	16.3%	777	2.7%
Not specified	2,515	8.8%	372	1.3%
Others	1,071	3.7%	164	0.6%
Private Business	6,020	21.0%	882	3.1%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Public Enterprise Employee	817	2.9%	127	0.4%
ระดับการศึกษามารดา				
Bachelor Degree	6,923	24.2%	1,012	3.5%
Doctoral Degree	65	0.2%	7	0.0%
Less than bachelor's degree	14,603	50.9%	2,009	7.0%
Master Degree	1,122	3.9%	196	0.7%
Unknown	2,318	8.1%	411	1.4%
สถานภาพมารดา				
Alive	24,487	85.4%	3,550	12.4%
Deceased	467	1.6%	70	0.2%
Disability	13	0.0%	2	0.0%
Disabled	21	0.1%	4	0.0%
Others	43	0.2%	9	0.0%
อาชีพมารดา				
Agricultural Sector	3,988	13.9%	530	1.8%
Contractor	3,098	10.8%	389	1.4%
Corporate Employee	907	3.2%	117	0.4%
Government Sector	4,070	14.2%	678	2.4%
Not specified	1,358	4.7%	189	0.7%
Others	4,163	14.5%	620	2.2%
Private Business	7,076	24.7%	1,063	3.7%
Public Enterprise Employee	371	1.3%	49	0.2%
สถานภาพบิดา-มารดา				
Both father and mother deceased	76	0.3%	10	0.0%
Divorce	2,354	8.2%	369	1.3%
Father and mother both remarried	177	0.6%	25	0.1%
Father deceased	1,477	5.2%	240	0.8%
Father remarried	103	0.4%	14	0.0%
Live together	18,791	65.6%	2,660	9.3%
Mother deceased	372	1.3%	56	0.2%
Mother remarried	54	0.2%	8	0.0%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Other	55	0.2%	8	0.0%
Separated due to career obligation	539	1.9%	78	0.3%
Separated for other reasons	1,033	3.6%	167	0.6%
ผลการเรียนเฉลี่ยปัจจุบัน				
0.00-1.00	2,605	9.1%	2,232	7.8%
1.01-2.00	2,222	7.8%	496	1.7%
2.01-3.00	7,381	25.7%	727	2.5%
3.01-4.00	12,823	44.7%	180	0.6%
ผลการเรียนเฉลี่ย				
0.00-1.00	1,955	6.8%	1,945	6.8%
1.01-2.00	1,359	4.7%	633	2.2%
2.01-3.00	10,880	38.0%	901	3.1%
3.01-4.00	10,837	37.8%	156	0.5%
ผลการเรียนเฉลี่ยก่อนเข้าศึกษา				
0.00-1.00	8	0.0%	2	0.0%
1.01-2.00	296	1.0%	122	0.4%
2.01-3.00	8,780	30.6%	1,626	5.7%
3.01-4.00	15,947	55.6%	1,885	6.6%
คะแนนภาษาอังกฤษก่อนเข้าศึกษา				
Less than 25	5,694	19.9%	1,255	4.4%
25-50	14,658	51.1%	1,884	6.6%
51-75	3,977	13.9%	448	1.6%
76 and above	702	2.4%	48	0.2%
รายรับเฉลี่ยนักศึกษา				
Less than 5,000	19,614	68.4%	2,818	9.8%
5,001-10,000	5,404	18.9%	812	2.8%
10,001 and above	13	0.0%	5	0.0%
รายจ่ายเฉลี่ยนักศึกษา				
Less than 5,000	21,128	73.7%	2,997	10.5%
5,001 and above	3,903	13.6%	638	2.2%
รายได้เฉลี่ยบิดา				

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Less than 15,000	11,330	39.5%	1,530	5.3%
15,001-30,000	8,028	28.0%	1,270	4.4%
30,001-45,000	3,120	10.9%	458	1.6%
45,001 and above	2,553	8.9%	377	1.3%
รายได้เฉลี่ยมารดา				
Less than 15,000	14,480	50.5%	1,947	6.8%
15,001-30,000	6,336	22.1%	1,046	3.6%
30,001-45,000	2,612	9.1%	406	1.4%
45,001 and above	1,603	5.6%	236	0.8%

ตารางที่ ก.2 สรุปรายละเอียดประชากรของตัวแปรเป้าหมาย DROPOUT ชุดข้อมูลระดับบัณฑิตศึกษา

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
รวมระดับบัณฑิตศึกษา	**4,309**	**81.9%**	**955**	**18.1%**
ภาคการศึกษาที่				
1	1,121	21.3%	280	5.3%
2	3,188	60.6%	675	12.8%
ชั้นปีที่				
1	1,039	19.7%	435	8.3%
2	2,795	53.1%	458	8.7%
3	392	7.4%	54	1.0%
4	79	1.5%	8	0.2%
5	4	0.1%		0.0%
คณะ				
Faculty of Agro-Industry	94	1.8%	30	0.6%
Faculty of Dentistry	241	4.6%	17	0.3%
Faculty of Economics	124	2.4%	32	0.6%
Faculty of Engineering	645	12.3%	145	2.8%
Faculty of Environmental Management	202	3.8%	70	1.3%
Faculty of Liberal Arts	257	4.9%	99	1.9%
Faculty of Management Sciences	775	14.7%	236	4.5%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Faculty of Medicine	525	10.0%	55	1.0%
Faculty of Natural Resources	249	4.7%	49	0.9%
Faculty of Nursing	338	6.4%	54	1.0%
Faculty of Pharmaceutical Sciences	113	2.1%	17	0.3%
Faculty of Science	574	10.9%	105	2.0%
Faculty of Science and Technology		0.0%	1	0.0%
Faculty of Traditional Thai Medicine	54	1.0%	14	0.3%
Graduate School	25	0.5%	1	0.0%
Health System Management Institute (HSMI)	18	0.3%	6	0.1%
Institute for Peace Studies	29	0.6%	9	0.2%
Marine and Coastal Resources Institute	14	0.3%	15	0.3%
Public Policy Institute	32	0.6%		0.0%
กลุ่มสาขาวิชา				
Health Sciences Group	1,321	25.1%	163	3.1%
Science and Technology Group	1,803	34.3%	416	7.9%
Social Sciences and Humanities Group	1,185	22.5%	376	7.1%
ประเภทการศึกษา				
Evening Program	11	0.2%	6	0.1%
Full Time Program	2,610	49.6%	499	9.5%
Full Time Program (International)	473	9.0%	64	1.2%
Part Time Program	1,181	22.4%	361	6.9%
Part Time Program (International)	34	0.6%	25	0.5%
เพศ				
Female	2,693	51.2%	525	10.0%
Male	1,616	30.7%	430	8.2%
การเกณฑ์ทหาร				
No	2,490	47.3%	520	9.9%
Unknown	469	8.9%	92	1.7%
Yes	1,350	25.6%	343	6.5%
ศาสนา				
Buddhism	3,336	63.4%	729	13.8%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Christian	83	1.6%	13	0.2%
Hinduism	88	1.7%	18	0.3%
Islam	700	13.3%	181	3.4%
Other	29	0.6%	3	0.1%
Sikhism	1	0.0%		0.0%
Undefined	72	1.4%	11	0.2%
โรคประจำตัว				
No	3,607	68.5%	816	15.5%
Unknown	1	0.0%		0.0%
Yes	701	13.3%	139	2.6%
แพ้ยา				
No	3,953	75.1%	880	16.7%
Unknown	2	0.0%		0.0%
Yes	354	6.7%	75	1.4%
สถานภาพสมรส				
Divorced	32	0.6%	9	0.2%
Married	442	8.4%	114	2.2%
Separated	4	0.1%	2	0.0%
Single	3,586	68.1%	784	14.9%
Unknown	236	4.5%	43	0.8%
Widowed	9	0.2%	3	0.1%
ระดับการศึกษาก่อนหน้า				
Bachelor Degree	3,297	62.6%	722	13.7%
Doctoral Degree	7	0.1%	3	0.1%
Graduate Diploma	11	0.2%	1	0.0%
High Vocational Certificate	1	0.0%	1	0.0%
Higher Graduate Diploma	24	0.5%	4	0.1%
Master Degree	455	8.6%	122	2.3%
Postgrad Diploma	27	0.5%	5	0.1%
Senior High School	1	0.0%	1	0.0%
Unknown	485	9.2%	95	1.8%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Vocational Certificate	1	0.0%	1	0.0%
ประเภทการเข้ารับ				
Direct system	3,993	75.9%	938	17.8%
Other	314	6.0%	17	0.3%
Unknown	2	0.0%		0.0%
จังหวัดที่เกิด				
SONGKHLA	1,357	25.8%	316	6.0%
NAKHON SI THAMMARAT	371	7.0%	77	1.5%
YALA	297	5.6%	57	1.1%
TRANG	259	4.9%	58	1.1%
BANGKOK	208	4.0%	53	1.0%
PHATTHALUNG	196	3.7%	41	0.8%
NARATHIWAT	179	3.4%	48	0.9%
SURAT THANI	172	3.3%	25	0.5%
PATTANI	156	3.0%	42	0.8%
PHUKET	89	1.7%	13	0.2%
Others	1,025	19.5%	225	4.3%
จำนวนพี่น้อง				
0	34	0.6%	14	0.3%
1	489	9.3%	93	1.8%
2	1,845	35.0%	390	7.4%
3	1,140	21.7%	260	4.9%
4	408	7.8%	102	1.9%
5	170	3.2%	37	0.7%
6	102	1.9%	27	0.5%
7	59	1.1%	16	0.3%
8	29	0.6%	5	0.1%
9	12	0.2%	4	0.1%
10	11	0.2%	5	0.1%
11	8	0.2%	2	0.0%
12	1	0.0%		0.0%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
20	1	0.0%		0.0%
จำนวนพี่น้องที่กำลังศึกษาอยู่				
0	371	7.0%	105	2.0%
1	2,233	42.4%	477	9.1%
2	1,226	23.3%	270	5.1%
3	358	6.8%	74	1.4%
4	81	1.5%	22	0.4%
5	25	0.5%	4	0.1%
6	8	0.2%	1	0.0%
7	4	0.1%	2	0.0%
8	1	0.0%		0.0%
9	1	0.0%		0.0%
10	1	0.0%		0.0%
ได้รับทุน				
No	3,278	62.3%	780	14.8%
Unknown	1	0.0%		0.0%
Yes	1,030	19.6%	175	3.3%
ความพิการ				
Disability hearing	2	0.0%	2	0.0%
Learning Disability	1	0.0%	1	0.0%
Not disabled	4,293	81.6%	948	18.0%
Physical disability	7	0.1%	3	0.1%
Visual disability	6	0.1%	1	0.0%
แผนการศึกษา				
A	411	7.8%	95	1.8%
B	19	0.4%	4	0.1%
C	194	3.7%	52	1.0%
D	150	2.8%	29	0.6%
E	207	3.9%	57	1.1%
F	2,193	41.7%	499	9.5%
G	762	14.5%	200	3.8%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
I	373	7.1%	19	0.4%
ระดับการศึกษาบิดา				
Bachelor Degree	959	18.2%	201	3.8%
Doctoral Degree	45	0.9%	7	0.1%
Less than bachelor's degree	1,789	34.0%	402	7.6%
Master Degree	321	6.1%	67	1.3%
Unknown	1,195	22.7%	278	5.3%
สถานภาพบิดา				
Alive	3,689	70.1%	801	15.2%
Deceased	578	11.0%	143	2.7%
Disability	8	0.2%	3	0.1%
Disabled	9	0.2%	1	0.0%
Others	25	0.5%	7	0.1%
อาชีพบิดา				
Agricultural Sector	704	13.4%	162	3.1%
Contractor	269	5.1%	54	1.0%
Corporate Employee	88	1.7%	21	0.4%
Government Sector	1,149	21.8%	239	4.5%
Not specified	912	17.3%	216	4.1%
Others	193	3.7%	47	0.9%
Private Business	844	16.0%	186	3.5%
Public Enterprise Employee	150	2.8%	30	0.6%
ระดับการศึกษามารดา				
Bachelor Degree	1,099	20.9%	212	4.0%
Doctoral Degree	23	0.4%	4	0.1%
Less than bachelor's degree	2,065	39.2%	488	9.3%
Master Degree	227	4.3%	40	0.8%
Unknown	895	17.0%	211	4.0%
สถานภาพมารดา				
Alive	4,091	77.7%	896	17.0%
Deceased	200	3.8%	56	1.1%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
Disability	3	0.1%		0.0%
Disabled	6	0.1%	2	0.0%
Others	9	0.2%	1	0.0%
อาชีพมารดา				
Agricultural Sector	670	12.7%	161	3.1%
Contractor	182	3.5%	40	0.8%
Corporate Employee	63	1.2%	17	0.3%
Government Sector	971	18.4%	192	3.6%
Not specified	579	11.0%	140	2.7%
Others	835	15.9%	181	3.4%
Private Business	934	17.7%	209	4.0%
Public Enterprise Employee	75	1.4%	15	0.3%
สถานภาพบิดา-มารดา				
Both father and mother deceased	67	1.3%	20	0.4%
Divorce	273	5.2%	69	1.3%
Father and mother both remarried	22	0.4%	6	0.1%
Father deceased	490	9.3%	118	2.2%
Father remarried	8	0.2%	2	0.0%
Live together	3,174	60.3%	668	12.7%
Mother deceased	108	2.1%	29	0.6%
Mother remarried	8	0.2%	2	0.0%
Other	19	0.4%	3	0.1%
Separated due to career obligation	42	0.8%	11	0.2%
Separated for other reasons	98	1.9%	27	0.5%
ผลการเรียนเฉลี่ยปัจจุบัน				
0.00-1.00	2,694	51.2%	733	13.9%
1.01-2.00	23	0.4%	52	1.0%
2.01-3.00	51	1.0%	41	0.8%
3.01-4.00	1,541	29.3%	129	2.5%
ผลการเรียนเฉลี่ย				
0.00-1.00	1,102	20.9%	456	8.7%

ตัวแปรและข้อมูล	กำลังศึกษาอยู่		ออกกลางคัน	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
1.01-2.00	7	0.1%	22	0.4%
2.01-3.00	72	1.4%	87	1.7%
3.01-4.00	3,128	59.4%	390	7.4%
ผลการเรียนเฉลี่ยก่อนเข้าศึกษา				
0.00-1.00	54	1.0%	19	0.4%
1.01-2.00	2	0.0%	1	0.0%
2.01-3.00	1,488	28.3%	391	7.4%
3.01-4.00	2,765	52.5%	544	10.3%
รายรับเฉลี่ยนักศึกษา				
Less than 10,000	1,408	26.7%	281	5.3%
10,001-20,000	1,305	24.8%	358	6.8%
20,001-30,000	998	19.0%	205	3.9%
30,001 and above	598	11.4%	111	2.1%
รายจ่ายเฉลี่ยนักศึกษา				
Less than 10,000	2,340	44.5%	532	10.1%
10,001-20,000	1,473	28.0%	335	6.4%
20,001-30,000	488	9.3%	87	1.7%
30,001 and above	8	0.2%	1	0.0%

ภาคผนวก ข
ผลงานตีพิมพ์และเผยแพร่

การวิเคราะห์เปรียบเทียบการเรียนรู้ของเครื่องเพื่อคาดการณ์การออกกลางคันของนักศึกษาระดับปริญญาตรี: กรณีศึกษา มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

A Comparative Analysis of Machine Learning for Dropout Prediction in Undergraduate Students: A Case Study of Prince of Songkla University Hat Yai Campus

กฤตกร อินแพง Kittakorn Inpang[1]
นฤบาล ยมะคุปต์ Narueban Yamaqupta[2]
จุไรรัตน์ พุทธรักษ์ Jurairat Bhudharak[3]

บทคัดย่อ

อัตราการคงอยู่ของนักศึกษาเป็นส่วนสำคัญและเป็นตัวชี้วัดหนึ่งในการวัดความสำเร็จของมหาวิทยาลัยที่ส่งผลกระทบต่ออันดับ ชื่อเสียง และสวัสดิภาพทางการเงินของมหาวิทยาลัย เช่นเดียวกับมหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ โดยตั้งแต่ปีการศึกษา พ.ศ. 2556 - 2560 พบว่าอัตราการออกกลางคันของนักศึกษาเพิ่มสูงขึ้นอยู่ที่ร้อยละ 17.54 ด้วยเหตุนี้งานวิจัยนี้จึงนำเสนอการใช้เทคนิคเหมืองข้อมูลและการเรียนรู้ของเครื่องเพื่อวิเคราะห์คุณลักษณะที่สำคัญ และสร้างแบบจำลองเพื่อคาดการณ์การออกกลางคันของนักศึกษา ขอบเขตของงานวิจัยนี้คือข้อมูลผลลัพธ์ทางการศึกษา ข้อมูลพื้นฐานของนักศึกษา และข้อมูลครอบครัวของนักศึกษา ในระดับปริญญาตรี 6 รุ่นปีการศึกษา คือ ในช่วง พ.ศ. 2558 - 2563 รวมทั้งหมด 17 คณะของมหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ด้วยแบบจำลองอัลกอริทึมการจำแนกประเภท 5 แบบ นำมาทดสอบและเปรียบเทียบประสิทธิภาพและความเหมาะสมในการนำไปใช้งาน ผลการทดสอบพบว่าแบบจำลอง Random Forest ที่ปรับจูนไฮเปอร์พารามิเตอร์แล้วเป็นวิธีที่ดีที่สุดให้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 90.97% และคุณลักษณะปัจจัยที่สำคัญที่สุดคือปัจจัยด้านผลการเรียน ได้แก่ ผลการเรียนสะสม ผลการเรียนปัจจุบัน ผลการเรียนก่อนเข้าศึกษา รวมถึงภาคการศึกษาและชั้นปี และคะแนนภาษาอังกฤษก่อนเข้าศึกษา ตามลำดับ

คำสำคัญ: นักศึกษาออกกลางคัน การเรียนรู้ของเครื่อง อุดมศึกษา เหมืองข้อมูล การคาดการณ์

Abstract

Student retention rate is an important measure of a university's success that affects its rank, reputation, and financial well-being. For Prince of Songkla University, Hat Yai Campus as well, from 2013 – 2017, the student dropout rate rose to 17.54%. As a

[1] นักศึกษาระดับปริญญาโท หลักสูตรหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล โครงการจัดตั้งวิทยาลัยวิทยาศาสตร์ดิจิทัล มหาวิทยาลัยสงขลานครินทร์ Email: 6310025001@psu.ac.th
[2] ดร. สาขาวิชาการจัดการการท่องเที่ยว คณะพาณิชยศาสตร์และการจัดการ มหาวิทยาลัยสงขลานครินทร์ Email: narueban.y@psu.ac.th
[3] ผู้ช่วยศาสตราจารย์ ดร. สาขาวิชาการจัดการธุรกิจผ่านสื่ออิเล็กทรอนิกส์ คณะพาณิชยศาสตร์และการจัดการ มหาวิทยาลัยสงขลานครินทร์ Email: jurairat.b@psu.ac.th

result, this research presents the use of data mining and machine learning techniques to analyze important characteristics and build a model to predict student dropout. The scope of this research includes study result data, student information, and family information of bachelor's degree students for 6 academic years, during 2015 - 2020, a total of 17 faculties of Prince of Songkla University, Hat Yai Campus. Five classification algorithm models were tested and compared for performance and application suitability. The results showed that the hyperparameter tuned Random Forest model was the best method, yielding the highest accuracy at 90.97%, and attributed the most important factors to academic performance: cumulative grade point average, followed by grade point average, study results before admission in semester and academic year level, as well as English scores before admission, respectively.

Keywords: Student Dropout, Machine Learning, Higher Education, Data Mining, Prediction

บทนำ

การศึกษาระดับอุดมศึกษาเป็นปัจจัยสำคัญในด้านการพัฒนาประเทศ โดยสามารถผลิตกำลังคนที่มีองค์ความรู้ ทักษะการทำงาน เป็นที่ต้องการของประเทศในภาคอุตสาหกรรม ภาคธุรกิจ และชุมชน สร้างวิทยาการ นวัตกรรม และขีดความสามารถในการแข่งขันของประเทศ (Office of the Education Council - Ministry of Education, 2018) ความกังวลที่สำคัญในการศึกษาระดับอุดมศึกษาคือการคงอยู่ของนักศึกษาที่เข้าศึกษาตั้งแต่ปีการศึกษาแรกจนสำเร็จการศึกษา อัตราการคงอยู่ของนักศึกษามีความสำคัญต่อชื่อเสียง ภาพลักษณ์ สวัสดิภาพทางการเงิน และเป็นตัวชี้วัดความสำเร็จของสถาบันการศึกษาที่แสดงถึงประสิทธิภาพและความน่าเชื่อถือของสถาบัน อีกทั้งเป็นเกณฑ์ชี้วัดคุณภาพการเรียนการสอนสำหรับการจัดสรรเงินทุนและงบประมาณของสถาบันจากรัฐบาลด้วย (Crosling et al., 2008) วิธีที่จัดการความท้าทายนี้ได้อย่างมีประสิทธิภาพคือการทำเหมืองข้อมูลเพื่อค้นหารูปแบบจากฐานข้อมูลขนาดใหญ่ (Luan, 2006) และแปลงเป็นองค์ความรู้เพื่อปรับปรุงกระบวนการตัดสินใจ วิเคราะห์และนำเสนอข้อมูลเพื่อวางแผนการบริหารจัดการ (Natek & Zwilling, 2014) การคาดการณ์การออกกลางคันของนักศึกษาที่แม่นยำสามารถนำข้อมูลนักศึกษาที่มีความเสี่ยงมาวิเคราะห์เพื่อช่วยเหลือด้านวิชาการแก่นักศึกษาศึกษาในสถาบันการศึกษา (Luan, 2006)

เทคนิคการทำเหมืองข้อมูลเพื่อการศึกษาและการวิเคราะห์การเรียนรู้เป็นเครื่องมือที่ช่วยให้สถาบันอุดมศึกษาสามารถแก้ไขปัญหาและสนับสนุนการตัดสินใจเพื่อก้าวสู่มหาวิทยาลัยยุคใหม่ได้ (Aldowah et al., 2019) เหมืองข้อมูลและการเรียนรู้ของเครื่องเป็นการใช้หลักทางสถิติเพื่อวิเคราะห์ข้อมูลและแก้ไขปัญหา (Romero & Ventura, 2010) เช่น การศึกษาผลลัพธ์ทางวิชาการของกลุ่มที่สำเร็จการศึกษาและกลุ่มที่ออกกลางคัน หรือการวิเคราะห์ข้อมูลนักศึกษาใหม่ โดยสร้างแบบจำลองเพื่อคาดการณ์การสำเร็จการศึกษา ซึ่งขั้นตอนทั้งหมดทำงานโดยอัตโนมัติ รวดเร็ว และแม่นยำ ซึ่งดีกว่าการทำนายแบบดั้งเดิม เหล่านี้เรียกว่าการเรียนรู้ของเครื่องและปัญญาประดิษฐ์ (Luan, 2006)

ข้อมูลนักศึกษาระดับปริญญาตรี มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ย้อนหลัง 5 รุ่นปีการศึกษา พ.ศ. 2556 - 2560 พบว่ามีจำนวนนักศึกษาออกกลางคันอยู่เป็นจำนวนมากรวมทั้งหมด 3,669 คน หรือคิดเป็นร้อยละ 17.54

งานวิจัยนี้ใช้ข้อมูลจากระบบข้อมูลพื้นฐานนักศึกษาและระบบข้อมูลนักศึกษา 6 รุ่นปีการศึกษา พ.ศ. 2558 - 2563 มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ และเทคนิคการทำเหมืองข้อมูลเพื่อออกแบบและพัฒนาแบบจำลองคาดการณ์การออกกลางคันของนักศึกษา ด้วยแบบจำลองอัลกอริทึมการจำแนกประเภท (Classification Algorithms) 5 แบบ โดยนำปัจจัยด้านสุขภาพ ด้านการเงิน และการรับทุนการศึกษาจากขอบเขตข้อมูลพื้นฐานของนักศึกษามาวิเคราะห์ ผ่านการเปรียบเทียบอัลกอริทึมการเรียนรู้ของเครื่องและเลือกแบบจำลองที่ดีที่สุดสำหรับการคาดการณ์การออกกลางคันของนักศึกษาเพื่อลดอัตราการออกกลางคันในมหาวิทยาลัยได้

วัตถุประสงค์

1. เพื่อค้นหาคุณลักษณะปัจจัยสำคัญที่มีผลต่ออัตราการออกกลางคันของนักศึกษา ระดับปริญญาตรี มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ โดยใช้เทคนิคการเรียนรู้ของเครื่อง
2. เพื่อเปรียบเทียบแบบจำลองอัลกอริทึมการจำแนกประเภท 5 แบบ และเลือกแบบจำลองที่มีประสิทธิภาพที่สุดไปใช้งาน

ประโยชน์ที่คาดว่าจะได้รับ

1. สถาบันอุดมศึกษาสามารถนำข้อมูลปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาไปปรับใช้เพื่อช่วยเหลือนักศึกษาที่มีความเสี่ยง เพื่อลดอัตราการออกกลางคันในมหาวิทยาลัยได้
2. นำเสนอการพัฒนาแบบจำลองคาดการณ์การออกกลางคันของนักศึกษาด้วยอัลกอริทึมการเรียนรู้ของเครื่องที่ถูกต้อง แม่นยำ

การทบทวนวรรณกรรม

การคงอยู่ของนักศึกษา (Student Retention) คือการที่นักศึกษายังคงอยู่และมีการลงทะเบียนในแต่ละภาคการศึกษาจนกว่าจะสำเร็จการศึกษา (Bean, 2013)-(Bonneau & Director, n.d.) นอกจากนี้มีแบบจำลองตามทฤษฎีที่มีอิทธิพลของ Vincent Tinto (Bean, 2013) ที่อธิบายปัจจัยที่ส่งผลต่อการคงอยู่ของนักศึกษาคือ ครอบครัว คุณลักษณะส่วนบุคคล และการศึกษา

การออกกลางคัน (Dropout) สำนักงานคณะกรรมการการศึกษาแห่งชาติได้ให้ความหมายการออกกลางคันว่า คือการที่ผู้เรียนถูกจำหน่ายชื่อออกจากสถานศึกษาในขณะที่ยังไม่สำเร็จการศึกษาโดยไม่ใช่สาเหตุจากการย้ายสถานศึกษา (เมฆขลา, 2552)

การทำเหมืองข้อมูลเพื่อการศึกษา (Educational Data Mining) และการวิเคราะห์การเรียนรู้ (Learning Analytics) เป็นเทคนิคการทำเหมืองข้อมูลมีความสำคัญต่อกระบวนการเรียนรู้และผลลัพธ์เพื่อก้าวสู่มหาวิทยาลัยยุคใหม่ เครื่องมือที่จะช่วยแก้ไขปัญหาและสนับสนุนการตัดสินใจ สนับสนุนการเรียนรู้ด้วยตนเอง กระบวนการที่เกี่ยวข้องกับการเรียนรู้ร่วมกัน การติดตาม การประเมิน เช่น การคงอยู่ของนักศึกษา ประสิทธิภาพความสำเร็จ และการออกกลางคัน (Aldowah et al., 2019)

การเรียนรู้ของเครื่อง (Machine Learning) เป็นรูปแบบการวิเคราะห์ข้อมูลด้วยแบบจำลองอัตโนมัติ บนแนวคิดที่ว่าระบบต่าง ๆ สามารถที่จะเรียนรู้และมีปฏิสัมพันธ์กับชุดข้อมูลต่าง ๆ รวมถึงสามารถระบุรูปแบบต่าง ๆ ที่เกิดขึ้นนำไปสู่การตัดสินใจได้เองโดยไม่จำเป็นต้องพึ่งพามนุษย์ (*Machine Learning: What It Is and Why It Matters | SAS*, 2021) ปัจจุบันมีอัลกอริทึมการเรียนรู้ของเครื่องจำนวนมากที่เรียนรู้และใช้งานส่วนใหญ่คือการถดถอยเชิงเส้นและลอจิสติก ตามด้วยอัลกอริทึมขั้นสูง ตัวอย่างอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised) ได้แก่ แรนดอมฟอเรสต์ (Random Forest) เป็นเทคนิคสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจหลาย ๆ ต้นรวมกัน และนำเอาคุณสมบัติในชุดข้อมูลหลายชุด ๆ ของเหล่านั้นมาสุ่มเลือกคุณสมบัติ เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) เป็นเทคนิคที่นำเอาต้นไม้การตัดสินใจหลาย ๆ ต้นมาเรียนรู้จากข้อผิดพลาดก่อนหน้า ทำให้มีความแม่นยำในการทำนายสูง ทฤษฎีของเบย์ (Naïve Bayes Theorem) เป็นเทคนิคการทำเหมืองข้อมูลด้วยใช้การคำนวณความน่าจะเป็นที่ได้รับความนิยมอย่างสูงสำหรับการจำแนกประเภท การเรียนรู้แบบไม่มีผู้สอน (Unsupervised) การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means Clustering) เป็นเทคนิคการแบ่งกลุ่มโดยการสังเกตจำนวน n สิ่งเป็น k กลุ่ม โดยแต่ละกลุ่มที่มีค่าเฉลี่ยใกล้เคียงกันที่สุด (Kilic, 2020)

ปัญหาการออกกลางคันของนักศึกษาในสถาบันอุดมศึกษาของประเทศไทยจากงานวิจัยของ (Hanthongchai, 2019) และ (Taipjutorus, 2017) พบว่าอัตราการออกกลางคันของนักศึกษาส่วนใหญ่เกิดขึ้นในช่วงชั้นปีแรกที่เข้าศึกษา ซึ่งปัญหานี้ถูกนำไปสู่การวิจัยเพื่อหาสาเหตุของการออกกลางคัน เพื่อลดจำนวนการออกกลางคันของนักศึกษาที่ปัจจุบันประสบปัญหาจำนวนมาก (Komol Chantawong, 2016) ได้ใช้เครื่องมือแบบสอบถามถึงปัจจัยสาเหตุที่เกี่ยวข้องพบว่าสาเหตุส่วนใหญ่มาจากปัจจัยด้านหลักสูตรและการเรียนการสอน ส่วนงานวิจัยของ **(**Krongkaew et al**.,** 2018**)** และ **(**Arandon, 2559**)** พบว่าสาเหตุเพราะสาขาวิชาที่เรียนไม่สอดคล้องกับความถนัดและความสามารถ

(Aldowah et al., 2019) กล่าวว่าอิทธิพลที่เกิดขึ้นของการวิเคราะห์ทำเหมืองข้อมูลเพื่อการศึกษา และการวิเคราะห์การเรียนรู้ พบว่าเทคนิคการทำเหมืองข้อมูลที่สำคัญคือสถิติ การวิเคราะห์การถดถอย กฎการเชื่อมโยง และการจัดกลุ่ม เป็นเครื่องมือที่เหมาะสมและมีประโยชน์มากในการนำมาพัฒนาใช้สำหรับการคาดการณ์ข้อมูลที่เน้นนักศึกษาเป็นหลักในระดับอุดมศึกษา

(Mahatthanachai et al., 2016) นำเสนอเทคนิคการทำเหมืองข้อมูลด้วยต้นไม้การตัดสินใจและ C4**.**5 อัลกอริทึมพบว่าปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาจำนวน 1,433 คน ของคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเชียงใหม่ ผลการวิจัยปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาคือ ผลการเรียนก่อนเข้าศึกษาถึง 72%

(Theppalak, 2019) เปรียบเทียบแบบจำลองต้นไม้ 3 แบบ ได้แก่ C4.5, RandomTree, และ REPTree และใช้กฎการอุปนัยได้แก่ OneR, ZeroR, และ rule-based learner (JRip) ของการออกกลางคันของนักศึกษาคณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์จำนวน 4,238 รายการ 7 คุณลักษณะ ได้แก่ คะแนนเฉลี่ยระดับชั้นมัธยมศึกษาตอนปลาย ผลการเรียนเฉลี่ยของมหาวิทยาลัย วิธีการรับเข้าเรียน ระยะเวลาในการลงทะเบียน สาขาวิชา สถานภาพการศึกษา และจังหวัดของโรงเรียนมัธยม พบว่า JRip ให้ความแม่นยำสูงสุดในการทำนายผลลัพธ์ที่ 77.30% โดยคุณลักษณะปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาส่วนใหญ่เกี่ยวกับผลลัพธ์ทางการศึกษา เช่น ผลการเรียนเฉลี่ยที่ต่ำ รวมถึงวิธีการเข้ารับเข้าเรียน และสาขาวิชา ตามลำดับ

นอกจากนี้ยังมีงานวิจัยอื่นที่ใช้ต้นไม้การตัดสินใจ Decision Tree, Random Forest, และ Gradient Boosting งานวิจัยของ (Tenpipat & Akkarajitsakul, 2020) นำเสนอโดยแบ่งกลุ่มตัวแปรข้อมูลเป็น 5 กลุ่ม ได้แก่ ครอบครัว, โรงเรียน, การเข้ารับการศึกษา, คณะสาขา และกลุ่มข้อมูลส่วนตัว โดยใช้เทคนิค Synthetic Minority Oversampling (SMOTE) ในการทำ Data Balancing พบว่าแบบจำลองทั้ง 3 ผลลัพธ์ไม่แตกต่างอย่างมีนัยสำคัญ ซึ่ง Gradient Boosting มีความแม่นยำสูงที่สุดที่ 93% และพบว่าปัจจัยที่ส่งผลต่อการออกกลางคันคือ รุ่นปีการศึกษา ผลการเรียนก่อนหน้า และประเภทการเข้ารับการศึกษาตามลำดับ

จากการทบทวนทฤษฎีและงานวิจัยที่เกี่ยวข้อง ผู้วิจัยได้ทำการใช้เทคนิคการทำเหมืองข้อมูลเพื่อการศึกษา และวิเคราะห์การเรียนรู้เพื่อออกแบบและพัฒนาแบบจำลองคาดการณ์การออกกลางคันของนักศึกษา ด้วยแบบจำลองอัลกอริทึมการจำแนกประเภท โดยนำปัจจัยด้านสุขภาพ ด้านการเงิน และการรับทุนการศึกษาจากขอบเขตข้อมูลพื้นฐานของนักศึกษามาวิเคราะห์ ผ่านการเปรียบเทียบอัลกอริทึมการเรียนรู้ของเครื่อง และเลือกแบบจำลองที่ดีที่สุดสำหรับการคาดการณ์การออกกลางคันของนักศึกษา

วิธีดำเนินการวิจัย

งานวิจัยนี้ใช้ภาษาโปรแกรมไพทอน (Python Programming Language) และเครื่องมือ Google Colab Notebook ในการนำเข้าและจัดการข้อมูล สร้างและวัดประสิทธิภาพแบบจำลอง โดยใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหารูปแบบความสัมพันธ์และเพื่อหาผลลัพธ์ โดยใช้กระบวนการ CRISP-DM (Cross Industry Standard Process for Data Mining) ประกอบด้วย 6 ขั้นตอน ได้แก่

1. Business Understanding อัตราการออกกลางคันยังมีจำนวนมาก ผู้วิจัยได้ทำการศึกษาและทำความเข้าใจทฤษฎีที่เกี่ยวข้อง ได้แก่ ทฤษฎีโมเดลของ Vincent Tinto ที่เกี่ยวกับปัจจัยที่ส่งผลต่อการคงอยู่นักศึกษา (Bean, 2013) และศึกษาเทคนิคการทำเหมืองข้อมูลและการเรียนรู้ของเครื่อง

2. Data Understanding ทำความเข้าใจข้อมูลจากระบบลงทะเบียนและระบบข้อมูลพื้นฐานนักศึกษาระดับปริญญาตรี มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ รุ่นปีการศึกษา 2558-2563 โดยมีขอบเขตเนื้อหา คือ 1. ผลลัพธ์ทางการศึกษา 2. ข้อมูลพื้นฐานของนักศึกษา และ 3. ข้อมูลครอบครัวของนักศึกษา จำนวนทั้งสิ้น 28,666 ข้อมูล (Observations) 41 คุณลักษณะ (Attributes)

3. Data Preparation จัดการข้อมูลเพื่อปรับปรุงคุณภาพและความสมบูรณ์ด้วยวิธีที่เหมาะสม ได้แก่ จัดการข้อมูลสูญหาย รายละเอียดตามตารางที่ 1 จัดการข้อมูลสุดโต่ง การสร้างตัวแปรเป้าหมาย DROPOUT ซึ่งพบว่ามีจำนวนนักศึกษาที่กำลังศึกษาทั้งหมด 25,031 คน ร้อยละ 87.3 แทนด้วย DROPOUT = 0 และพ้นสภาพการเป็นนักศึกษาทั้งหมดจำนวน 3,635 คน ร้อยละ 12.7 แทนด้วย DROPOUT = 1 ตามลำดับ และสุดท้ายเลือกปัจจัยในชุดข้อมูลที่ส่งผลต่อตัวแปรเป้าหมายอย่างมีนัยสำคัญทางสถิติที่ P-value < 0.05 สำหรับข้อมูลตัวแปรจัดกลุ่มด้วยวิธี Chi-Squared และข้อมูลตัวแปรแบบต่อเนื่องด้วยวิธี ANOVA พบว่ามีตัวแปรที่มีนัยสำคัญทางสถิติเพื่อนำไปสร้างแบบจำลองทั้งสิ้น 22 คุณลักษณะ (Attributes) รายละเอียดตามตารางที่ 3

ตารางที่ 1

ตรวจสอบค่าสูญหายและกระบวนการแก้ไข

ลำดับ	ชื่อตัวแปร	ข้อมูลสูญหาย	วิธีการ	ลำดับ	ชื่อตัวแปร	ข้อมูลสูญหาย	วิธีการ
1	หลักสูตร (ปี)	2,847 (9.9%)	Mode	10	ระดับการศึกษาก่อนหน้า	41 (0.1%)	Mode
2	แผนการศึกษา	2,833 (9.9%)	Mode	11	ระดับการศึกษามารดา	41 (0.1%)	Mode
3	รายได้เฉลี่ยมารดา	1,185 (4.1%)	Mode	12	ระดับการศึกษาบิดา	40 (0.1%)	Mode
4	ผลการเรียนสะสม	619 (2.2%)	Mean	13	สถานะการเกณฑ์ทหาร	40 (0.1%)	Mode
5	ผลการเรียนปัจจุบัน	619 (2.2%)	Mean	14	เป็นบุตรคนที่	40 (0.1%)	Mode
6	รายได้บิดา	453 (1.6%)	Mode	15	รายได้เฉลี่ยนักศึกษา	40 (0.1%)	Mode
7	สถานภาพสมรส	257 (0.9%)	Mode	16	การรับทุนการศึกษา	9 (0.0%)	Mode
8	ผลการเรียนก่อนเข้าศึกษา	65 (0.2%)	Mean	17	แพ้ยา	2 (0.0%)	Mode
9	คะแนนภาษาอังกฤษก่อนเข้าศึกษา	46 (0.2%)	Mean	18	โรคประจำตัว	2 (0.0%)	Mode

ตารางที่ 2

คุณลักษณะปัจจัยที่เลือกเพื่อนำไปสร้างแบบจำลองของชุดข้อมูลระดับปริญญาตรี

ลำดับ	ชื่อตัวแปร	p-value	ลำดับ	ชื่อตัวแปร	p-value
1	ภาคการศึกษา	<0.0001*	19	อาชีพบิดา	0.0072*
2	คณะ	<0.0001*	20	การรับทุนการศึกษา	0.0084*
3	กลุ่มสาขาวิชา	<0.0001*	21	สถานภาพสมรส	0.0145*
4	ประเภทการศึกษา	<0.0001*	22	อาชีพมารดา	0.0395*
5	เพศ	<0.0001*	23	รายได้เฉลี่ยมารดา	0.0589
6	สถานะการเกณฑ์ทหาร	<0.0001*	24	รายได้เฉลี่ยนักศึกษา	0.0631
7	ประเภทการเข้ารับ	<0.0001*	25	จังหวัดที่เกิด	0.0771
8	ผลการเรียนสะสม	<0.0001*	26	รายจ่ายเฉลี่ยนักศึกษา	0.0928
9	ผลการเรียนปัจจุบัน	<0.0001*	27	ความพิการ	0.1963
10	ชั้นปีที่	<0.0001*	28	โรคประจำตัว	0.2329
11	ผลการเรียนก่อนเข้าศึกษา	<0.0001*	29	แพ้ยา	0.2529
12	คะแนนภาษาอังกฤษก่อนเข้าศึกษา	<0.0001*	30	สถานมารดา	0.6314
13	หลักสูตร (ปี)	<0.0001*	31	จำนวนพี่น้องที่กำลังศึกษา	0.6758
14	ศาสนา	<0.0001*	32	สถานบิดา	0.6968
15	ระดับการศึกษาก่อนหน้า	<0.0001*	33	เป็นบุตรคนที่	0.7727
16	ระดับการศึกษามารดา	0.0013*	34	รายได้เฉลี่ยบิดา	0.7986
17	ระดับการศึกษาบิดา	0.0014*	35	สถานภาพบิดา-มารดา	0.9134
18	จำนวนพี่น้อง	0.0051*			

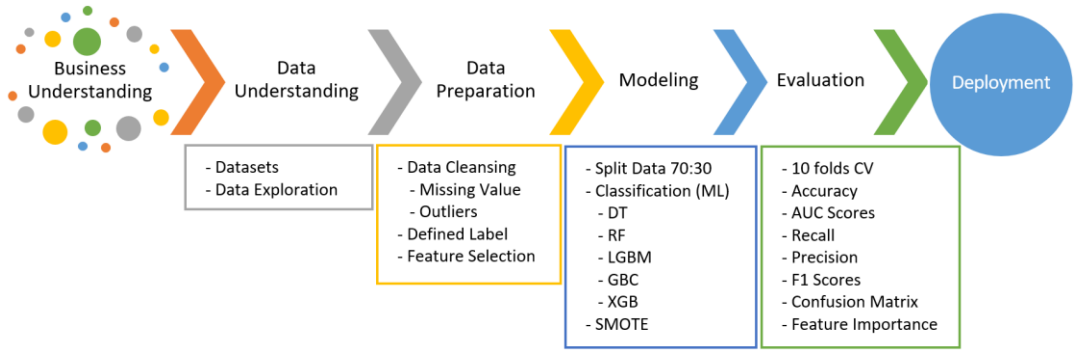
*หมายเหตุ: *มีนัยสำคัญเชิงสถิติ < 0.05, ไม่รวม 6 คุณลักษณะปัจจัยที่นำออกไปเนื่องจากไม่ส่งผลต่อตัวแปรเป้าหมาย ได้แก่ ปีข้อมูล, ปีการศึกษา, รหัส นศ., สถานภาพ นศ., ระดับการศึกษา และแผนการศึกษา*

4. Modeling สร้างแบบจำลองการเรียนรู้ของเครื่อง ผู้วิจัยใช้ PyCaret Machine Learning Library (Ali, 2021) ในการสร้างอัลกอริทึมการจำแนกประเภท (Classification Algorithms) 5 แบบ ได้แก่ 1. Decision Tree Classifier (DF) 2. Random Forest Classifier (RF) 3. Light Gradient Boosting (LGBM) 4. Gradient Boosting Classifier และ 5. Extreme Gradient Boosting (XGB) โดยทำการแบ่งข้อมูลสำหรับการฝึกแบบจำลอง 70% (19,063 ข้อมูล) และสำหรับทดสอบแบบจำลอง 30% (8,170 ข้อมูล) สร้างและปรับปรุงประสิทธิภาพแบบจำลองโดยอัตโนมัติในชุดของไฮเปอร์

พารามิเตอร์ที่เหมาะสมโดยใช้การค้นหาแบบสุ่ม และเปรียบเทียบแบบจำลองที่ใช้และไม่ได้ใช้เทคนิควิธีสังเคราะห์ข้อมูลเพิ่ม Synthetic Minority Oversampling Technique (SMOTE) ในการจัดการข้อมูลที่มีคลาสแตกต่างกัน ให้แบบจำลองมีความแม่นยำมากยิ่งขึ้น (Chawla et al., 2002)

5. ประเมินผลแบบจำลอง (Evaluation) มี 2 ขั้นตอน ได้แก่ (1) ตรวจสอบความถูกต้องของแบบจำลอง (Model Validation) ผู้วิจัยพิจารณากระบวนการประเมินประสิทธิภาพโดยการทำ K-Fold Cross Validation 10 folds คือการแบ่งชุดข้อมูลออกเป็น 10 ชุดในการสร้างและทดสอบแบบจำลอง ทำให้ผลลัพธ์มีความน่าเชื่อถือมากขึ้นและลดปัญหา Overfitting ของแบบจำลอง (Tenpipat & Akkarajitsakul, 2020) เปรียบเทียบประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้กราฟ (AUC Scores) ค่าระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าความถ่วงดุล (F1 Scores) และตารางเมทริกซ์ความสับสน (Confusion Matrix) (2) คุณลักษณะที่สำคัญ (Feature Importance) เป็นเทคนิควิเคราะห์คุณลักษณะปัจจัยที่สำคัญที่สุดในการคาดการณ์ของแบบจำลองที่

6. Deployment ผลการวิจัยที่ได้จากการศึกษานี้สามารถนำคุณลักษณะปัจจัย และแบบจำลองที่ได้ไปประยุกต์ใช้สร้างแบบจำลองในการตัดสินใจการคงอยู่และการออกกลางคันของนักศึกษาในระดับอุดมศึกษา เพื่อช่วยผู้บริหารและเจ้าหน้าที่ที่เกี่ยวข้องในการสนับสนุนการตัดสินใจและวางแผนการบริหารงาน ซึ่งกระบวนการกรอบแนวคิดการวิจัยทั้งหมดแสดงดังภาพที่ 1



ภาพที่ 1 กรอบแนวคิดการวิจัย

ผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพแบบจำลองคาดการณ์ประเภทต้นไม้การตัดสินใจทั้ง 5 แบบที่ปรับปรุงประสิทธิภาพแบบจำลองแล้ว โดยทำการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้เทคนิควิธีสังเคราะห์ข้อมูลเพิ่มและแบบจำลองที่ไม่ได้ใช้เทคนิคการสังเคราะห์ข้อมูล จากการการแบ่งชุดข้อมูลออกเป็น 10 ชุดในการสร้างและทดสอบแบบจำลอง พบว่าแบบจำลอง Random Forest ให้ค่าความถูกต้องสูงที่สุดคือ 0.9074 รองลงมาคือแบบจำลอง Light Gradient Boosting Machine มีค่า 0.9044 แบบจำลอง Decision Tree มีค่า 0.9023 แบบจำลอง Extreme Gradient Boosting มีค่า 0.8945 แบบจำลอง Extreme Gradient Boosting (SMOTE) มีค่า 0.8921 แบบจำลอง Random Forest (SMOTE) มีค่า 0.8914 แบบจำลอง Light Gradient Boosting Machine (SMOTE) มีค่า 0.8911 แบบจำลอง Gradient Boosting (SMOTE) มีค่า 0.8908 แบบจำลอง Gradient Boosting มีค่า 0.8850 และแบบจำลอง Decision Tree (SMOTE) มีค่าต่ำสุดคือ 0.8839 ตามลำดับ จากนั้นพิจารณาค่าพื้นที่ใต้กราฟ พบว่าแบบจำลอง Light Gradient Boosting Machine ให้ค่าสูงที่สุดคือ 0.9365

รองลงมาคือแบบจำลอง Random Forest มีค่า 0.9347 และแบบจำลอง Random Forest (SMOTE) มีค่า 0.9345 และเมื่อพิจารณาแบบจำลองที่ใช้เทคนิควิธีสังเคราะห์ข้อมูลเพิ่มเพื่อปรับความสมดุลของคลาสพบว่าค่าระลึกของแบบจำลองส่วนใหญ่มีค่าเพิ่มขึ้นและค่าความแม่นยำลดลง แต่เมื่อเทียบสัดส่วนการคาดการณ์ของหมวดหมู่จริงและเท็จตามค่าระลึกและค่าความแม่นยำที่รวมกันคือค่าความถ่วงดุลพบว่าแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ซึ่งผลลัพธ์การเปรียบเทียบแบบจำลองแสดงได้ตามตารางที่ 3

ตารางที่ 3

ตารางเปรียบเทียบประสิทธิภาพของแบบจำลอง

Models	Accuracy	AUC	Recall	Precision	F1
Decision Tree	0.9023	0.9214	0.5145	0.6393	0.5685
Decision Tree (SMOTE)	0.8839	0.8879	0.6849	0.5305	0.5975
Gradient Boosting	0.8850	0.9202	0.4760	0.5505	0.5098
Gradient Boosting (SMOTE)	0.8908	0.9326	0.8189	0.5439	0.6535
Light Gradient Boosting Machine	0.9044	**0.9365**	0.4400	0.6880	0.5362
Light Gradient Boosting Machine (SMOTE)	0.8911	0.9289	0.7174	0.5519	0.6236
Random Forest	**0.9074**	0.9347	0.4947	**0.6831**	0.5730
Random Forest (SMOTE)	0.8914	0.9345	**0.8189**	0.5455	**0.6547**
Extreme Gradient Boosting	0.8945	0.9285	0.5244	0.5923	0.5552
Extreme Gradient Boosting (SMOTE)	0.8921	0.9274	0.5620	0.5729	0.5669

จากนั้นเมื่อพิจารณาผลลัพธ์ตารางเมทริกซ์ความสับสน (Confusion Matrix) ของแบบจำลองที่ปรับจูนไฮเปอร์พารามิเตอร์ จากชุดข้อมูลทดสอบ 30% พบว่าแบบจำลอง Random Forest ที่ไม่ได้ใช้เทคนิคการสังเคราะห์ข้อมูลเพิ่มมีค่าความถูกต้องสูงที่สุดคือร้อยละ 90.97 รองลงมาคือแบบจำลอง Light Gradient Boosting Machine มีค่าร้อยละ 90.71 แบบจำลอง Decision Tree มีค่าร้อยละ 90.4 ซึ่งผลลัพธ์การเปรียบเทียบแบบจำลองตารางเมทริกซ์ความสับสนแสดงได้ตามตารางที่ 4

ตารางที่ 4

ตารางผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาของแบบจำลอง ชุดข้อมูลตรวจสอบ 30%

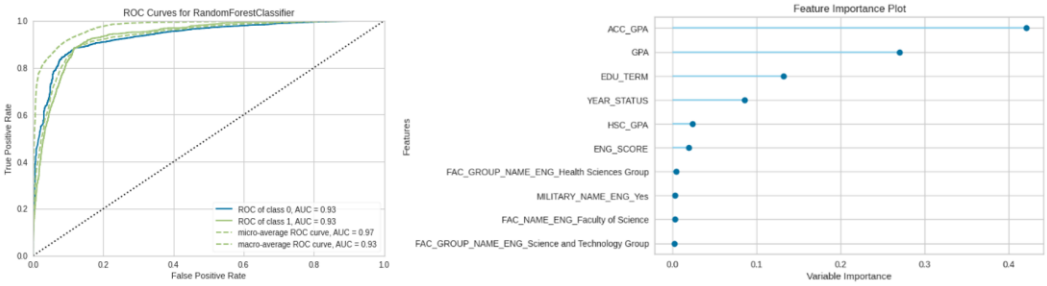
	Models		Predicted Class			
			Baseline		SMOTE	
			Studying	Dropout	Studying	Dropout
True Class	Decision Tree	**Studying**	7151	337	6729	759
		Dropout	494	618	323	789
		Accuracy		90.34%		87.42%
	Gradient Boosting	**Studying**	7080	408	6746	742
		Dropout	539	573	205	907
		Accuracy		88.99%		88.99%
	Light Gradient Boosting Machine	**Studying**	7264	224	6883	605
		Dropout	575	537	317	795
		Accuracy		90.71%		89.28%
	Random Forest	**Studying**	7235	253	6738	750
		Dropout	524	588	198	914

ตารางที่ 4

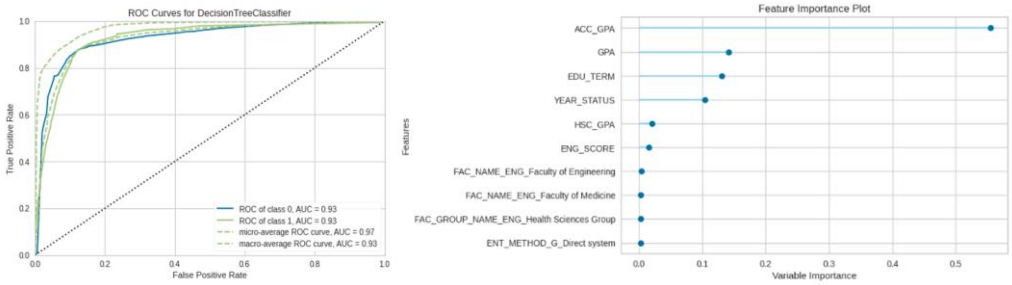
ตารางผลลัพธ์การคาดการณ์การออกกลางคันของนักศึกษาของแบบจำลอง ชุดข้อมูลตรวจสอบ 30% (ต่อ)

Models		Predicted Class			
		Baseline		SMOTE	
		Studying	Dropout	Studying	Dropout
	Accuracy		**90.97%**		88.98%
Extreme Gradient Boosting	**Studying**	7112	376	7034	454
	Dropout	501	611	451	661
	Accuracy		89.80%		89.48%

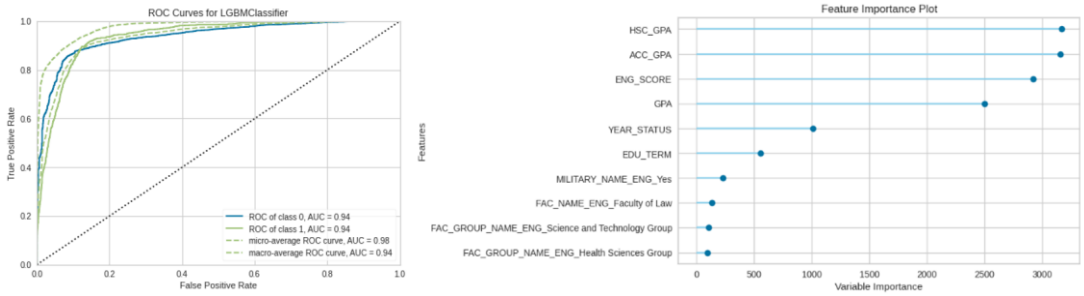
ขั้นตอนสุดท้ายทำการพิจารณาค่าพื้นที่ใต้กราฟและคุณลักษณะที่สำคัญของแบบจำลอง โดยผู้วิจัยเลือกแบบจำลองที่มีค่าความถูกต้อง 3 อันดับแรก พบว่าแบบจำลอง Random Forest และแบบจำลอง Decision Tree มีค่าพื้นที่ใต้กราฟเท่ากันคือ 0.93 และปัจจัยที่สำคัญ 6 อันดับแรกเท่ากัน คือ ผลการเรียนเฉลี่ย ACC_GPA รองลงมาคือผลการเรียนปัจจุบัน GPA ปัจจัยภาคการศึกษา EDU_TERM ชั้นปี YEAR_STATUS ผลการเรียนก่อนเข้าศึกษา HSC_GPA และคะแนนภาษาอังกฤษก่อนเข้าศึกษา ENG_SCORE ตามภาพที่ 2-3 ส่วนแบบจำลอง Light Gradient Boosting Machine มีค่าพื้นที่ใต้กราฟอยู่ที่ 0.94 และปัจจัยที่สำคัญคือ ผลการเรียนก่อนเข้าศึกษา HSC_GPA รองลงมาคือผลการเรียนเฉลี่ย ACC_GPA คะแนนภาษาอังกฤษก่อนเข้าศึกษา ENG_SCORE ผลการเรียนปัจจุบัน GPA ชั้นปี YEAR_STATUS และภาคการศึกษา EDU_TERM โดยส่วนใหญ่ปัจจัยของแบบจำลองคือผลการเรียนสะสม ACC_GPA และผลการเรียนปัจจุบันเป็นปัจจัยที่สำคัญที่สุด ตามภาพที่ 4



ภาพที่ 2 ค่าพื้นที่ใต้กราฟและปัจจัยที่สำคัญของแบบจำลอง Random Forest



ภาพที่ 3 ค่าพื้นที่ใต้กราฟและปัจจัยที่สำคัญของแบบจำลอง Decision Tree



ภาพที่ 4 ค่าพื้นที่ใต้กราฟและปัจจัยที่สำคัญของแบบจำลอง Light Gradient Boosting Machine

สรุปและอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อค้นหาปัจจัยที่มีผลต่ออัตราการออกกลางคันของนักศึกษาและเปรียบเทียบแบบจำลองอัลกอริทึมการจำแนกประเภท 5 แบบ พบว่าแบบจำลองที่มีค่าความถูกต้อง 3 อันดับแรก ได้แก่แบบจำลอง Random Forest แบบจำลอง Light Gradient Boosting Machine และ Decision Tree ให้ค่าความถูกต้องใกล้เคียงกันที่ร้อยละ 90.74%, 90.71% และร้อยละ 90.34% ตามลำดับ ส่วนค่าพื้นที่ใต้กราฟ แบบจำลอง Light Gradient Boosting Machine มีค่าสูงที่สุดร้อยละ 93.65% ค่าระลึก แบบจำลอง Random Forest (SMOTE) มีค่าสูงที่สุดร้อยละ 81.89 ค่าความแม่นยำ แบบจำลอง Random Forest มีค่าสูงที่สุดร้อยละ 68.31 และค่าความถ่วงดุล แบบจำลอง Random Forest (SMOTE) มีค่าสูงที่สุดร้อยละ 65.47 และเมื่อพิจารณาผลลัพธ์การคาดการณ์การออกกลางคันตารางเมทริกซ์ความสับสนพบว่าแบบจำลอง Random Forest ให้ค่าความถูกต้องสูงที่สุดคือร้อยละ 90.97% ดังนั้นแบบจำลอง Random Forest ที่ปรับจูนไฮเปอร์พารามิเตอร์แล้วเป็นวิธีที่ดีที่สุด

การศึกษานี้พบว่าปัจจัยที่มีอิทธิผลต่อการออกกลางคันของนักศึกษามากที่สุดของแบบจำลอง Random Forest ได้แก่ ตัวแปรผลการเรียนสะสม ACC_GPA, ผลการเรียนปัจจุบัน GPA, ภาคการศึกษา EDU_TERM, ชั้นปี YEAR_STATUS, ผลการเรียนก่อนเข้าศึกษา HSC_GPA, และคะแนนภาษาอังกฤษก่อนเข้าศึกษา ENG_SCORES ตามลำดับ ซึ่งสอดคล้องกับงานวิจัยของ (Theppalak, 2019) ที่พบว่าคุณลักษณะปัจจัยที่สำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษาคือ ผลการเรียนเฉลี่ยที่ต่ำ และสอดคล้องกับงานวิจัยของ (Mahatthanachai et al., 2016) และงานวิจัยของ (Tenpipat & Akkarajitsakul, 2020) ที่พบว่าคุณลักษณะปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษาคือ ผลการเรียนก่อนเข้าศึกษา ตามลำดับ

ส่วนปัจจัยทางด้านสุขภาพและปัจจัยทางด้านการเงิน ได้แก่ ตัวแปรโรคประจำตัว DISEASE, แพ้ยา ALLERGY, ความพิการ DEFORMITY_NAME_ENG ตัวแปรรายได้เฉลี่ยนักศึกษา AVG_INCOME, รายจ่ายเฉลี่ยนักศึกษา AVG_EXPENSE, รายได้เฉลี่ยมารดา MOTHER_AVG_INCOME และรายได้เฉลี่ยบิดา FATHER_AVG_INCOME พบว่าไม่เป็นปัจจัยสำคัญที่ส่งผลต่อการออกกลางคันของนักศึกษาอย่างมีนัยสำคัญที่ P-value < 0.05

ข้อเสนอแนะ

ข้อเสนอแนะจากผลการศึกษาปัจจัยที่มีผลต่ออัตราการออกกลางคันของนักศึกษา ระดับปริญญาตรี มหาวิทยาลัยสงขลานครินทร์ และเปรียบเทียบแบบจำลองอัลกอริทึมการจำแนกประเภท 5 แบบเพื่อให้ผู้วิจัยและผู้พัฒนานำข้อมูลที่ได้ไปศึกษาและพัฒนาเพิ่มเติมมีรายละเอียดดังนี้

1. ข้อมูลจากระบบลงทะเบียน และระบบข้อมูลพื้นฐานนักศึกษา มหาวิทยาลัยสงขลานครินทร์ควรจัดเก็บข้อมูลให้มีความถูกต้องครบถ้วนซึ่งจะช่วยลดการสูญหายของข้อมูลและช่วยให้แบบจำลองเรียนรู้ได้ดียิ่งขึ้น

2. นำข้อมูลผลการเรียนของรายวิชามาวิเคราะห์เพิ่มเติมเพื่อช่วยให้แบบจำลองได้เรียนรู้และวิเคราะห์ได้ละเอียดยิ่งขึ้น และช่วยให้สามารถวิเคราะห์รายวิชาที่ส่งผลต่อการออกกลางคันหรือคงอยู่ของนักศึกษาได้

3. ควรนำแบบจำลองที่ได้ไปพัฒนาออกแบบรายงานแดชบอร์ดเพื่อรายงานและติดตามการออกกลางคันของนักศึกษา โดยนำแบบจำลองที่ได้มาคาดการณ์ข้อมูลจริง และรายงานติดตามความเสี่ยง ซึ่งจะช่วยให้เจ้าหน้าที่ที่เกี่ยวข้องสามารถเข้าช่วยเหลือนักศึกษาที่มีความเสี่ยงได้ทันที

เอกสารอ้างอิง

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, *37*(April 2018), 13–49. https://doi.org/10.1016/j.tele.2019.01.007

Ali, M. (2021). *Introduction to Regression in Python with PyCaret*. https://towardsdatascience.com/introduction-to-regression-in-python-with-pycaret-d6150b540fc4

Arandon, D. (2559). *Causes of students ' drop out : A case study of Bachelor Degree of Business Administration in Management (English program)*. Prince of Songkla University.

Bean, J. P. (2013). *College Student Retention - Defining Student Retention, A Profile of Successful Institutions and Students, Theories of Student Departure - Factors, School, Model, and Social - StateUniversity.com*. https://education.stateuniversity.com/pages/1863/College-Student-Retention.html

Bonneau, K., & Director, A. (n.d.). *Brief ?: What is a Dropout?* Retrieved July 29, 2021, from http://www.ncpublicschools.org/research/dropout/reports/

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(Sept. 28), 321–357.

Crosling, G. M. (Glenda M., Heagney, M., & Thomas, L. (Elizabeth). (2008). *Improving student retention in higher education: the role of teaching and learning* (1st ed.). Routledge.

Hanthongchai, B. (2019). Factors Affecting to Drop out and Survival Pathways of First Year Undergraduate Students of Institute of Physical Education Udonthani. *MBU Education Journal*, *7*(2), 247–271.

Kilic, A. (2020). Artificial Intelligence and Machine Learning in Cardiovascular Health Care. *Annals of Thoracic Surgery*, *109*(5), 1323–1329. https://doi.org/10.1016/j.athoracsur.2019.09.042

Komol Chantawong. (2016). Factors Causing the College Dropouts of Valaya Alongkorn Rajabhat University Under the Royal Patronage, Sakaeo Campus. *วารสารวไลยอลงกรณ์ปริทัศน์ ปีที่5 ฉบับที่1 มกราคม-มิถุนายน 2558*, *5*(1), 127–141.

Krongkaew, A., Janin, P., Kaithong, N., Punpuch, P., & ... (2018). สาเหตุการออกกลางคันของนักศึกษา มหาวิทยาลัยราชภัฏกำแพงเพชร Causes of Undergraduate Student Dropout at Kamphaeng Phet Rajabhat University. *Arit.Kpru.Ac.Th*, *24*, 136–143. https://arit.kpru.ac.th/ap/e-dcms/contents/catalog/20191118110236.pdf

Luan, J. (2006). Executive brief Data Mining Applications in Higher Education. *Insol.Lt*. www.spss.com/downloads.

Machine Learning: What it is and why it matters | SAS. (2021). https://www.sas.com/en_us/insights/analytics/machine-learning.html

Mahatthanachai, B., Ninsonti, H., & Tantranont, N. (2016). A Study of Factors Influency Student Dropout Rate Using Data Mining. *The Golden Teak : Humanity and Social Science Journal*, *22*(4), 46–55. https://www.tci-thaijo.org/index.php/tgt/article/view/88196

Natek, S., & Zwilling, M. (2014). Student data mining solution-knowledge management system related to higher education institutions. *Expert Systems with Applications*, *41*(14), 6400–6407. https://doi.org/10.1016/j.eswa.2014.04.024

Office of the Education Council - Ministry of Education. (2018). *Education in Thailand 2018*.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Taipjutorus, W. (2017). *Reducing Attrition Rate of First-year Undergraduate Students: An Implementation of RMUTP Pre-University Program*. 252–258.

Tenpipat, W., & Akkarajitsakul, K. (2020). Student Dropout Prediction: A KMUTT Case Study. *2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020*. https://doi.org/10.1109/IBDAP50342.2020.9245457

Theppalak, N. (2019). The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining : *Journal of University of Babylon for Pure and Applied Sciences*, *27*(1), 356–368.

เมฆขลา, ค. (2552). *การพัฒนาระบบการช่วยเหลือนักเรียนโรงเรียนโปลิเทคนิคลานนา*.

ประวัติผู้เขียน

ชื่อ สกุล นายกฤตกร อินแพง

รหัสประจำตัวนักศึกษา 6310025001

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
เทคโนโลยีบัณฑิต (เทคโนโลยีสารสนเทศ)	มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ วิทยาเขตวังไกลกังวล	2554

ทุนการศึกษา

ทุนอุดหนุนการศึกษาระดับบัณฑิตศึกษาภายในประเทศ ประจำปีการศึกษา 2563 มหาวิทยาลัยสงขลานครินทร์

ตำแหน่งและสถานที่ทำงาน

พ.ศ. 2555 – 2556 ตำแหน่ง นักวิชาการคอมพิวเตอร์ คณะพยาบาลศาสตร์ มหาวิทยาลัยสงขลานครินทร์

พ.ศ. 2556 – ปัจจุบัน ตำแหน่ง นักวิชาการอุดมศึกษา กองนโยบาย ยุทธศาสตร์ และแผน มหาวิทยาลัยสงขลานครินทร์

การตีพิมพ์เผยแพร่ผลงาน

Inpang K, Yamaqupta N, Bhudharak J. A Comparative Analysis of Machine Learning for Dropout Prediction in Undergraduate Students : A Case Study of Prince of Songkla University Hat Yai Campus. The 14th National Conference on Administration and Management, Faculty of Management Sciences, Prince of Songkla University. 2022;482–493.

รางวัล

ผลงานวิจัยดีเด่นเรื่อง การวิเคราะห์เปรียบเทียบการเรียนรู้ของเครื่องเพื่อคาดการณ์การออกกลางคันของนักศึกษาระดับปริญญาตรี: กรณีศึกษา มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ในการประชุมวิชาการระดับชาติด้านการบริหารจัดการ ครั้งที่ 14 ประจำปี 2565 วันที่ 14 พฤษภาคม 2565 คณะวิทยาการจัดการ มหาวิทยาลัยสงขลานครินทร์