การทำนายปัจจัยที่มีผลกระทบต่อการออกกลางคันของนักศึกษา โดยใช้เทคนิคดาต้า ไมน์นิง กรณีศึกษา มหาวิทยาลัยราชภัฏกำแพงเพชร

Predicting Factors Influencing Student Dropout Using Data Mining Techniques: A Case Study of Kamphaeng Phet Rajabhat University

จินดาพร อ่อนเกตุ[1*], ฆัมภิชา ตันติสันติสม[1] , พรหมเมศ วีระพันธ์[1],
นรุตม์ บุตรพลอย[2], กนกวรรณ เขียววัน[2] และ คมกริช กลิ่นอาจ[3]
Jindaporn Ongate[1*], Khumphicha Tuntisantisom[1], Phrommate Veerapan[1],
Narut Butploy[2], Kanokwan Khiewwan[2] and Komkrit Klinart[3]

[1]สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏกำแพงเพชร
[2]สาขาวิชาเทคโนโลยีคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏกำแพงเพชร
[3]สำนักส่งเสริมวิชาการและงานทะเบียน มหาวิทยาลัยราชภัฏกำแพงเพชร
[1] Department of Information Technology, Faculty of Science and Technology, Kamphaeng Phet Rajabhat University
[2] Department of Computer Technology, Faculty of Science and Technology, Kamphaeng Phet Rajabhat University
[3]The office of Academic Promotion and Registration, Kamphaeng Phet Rajabhat University
*Email: jindaporn_o@kpru.ac.th
Received: December 17, 2024; Revised: March 6, 2025; Accepted: April 4, 2025

บทคัดย่อ

งานวิจัยนี้ มีวัตถุประสงค์ คือ 1) เพื่อศึกษาสถานการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยราชภัฏกำแพงเพชร 2) เพื่อสร้างโมเดลในการทำนายปัจจัยที่มีผลกระทบต่อการออกกลางคันของนักศึกษาโดยใช้เทคนิคดาต้า ไมน์นิง 3) เพื่อทดสอบประสิทธิภาพของโมเดลที่พัฒนา โดยใช้ข้อมูลจากฐานข้อมูลงานทะเบียนของมหาวิทยาลัยราชภัฏกำแพงเพชร ของนักศึกษาระดับปริญญาตรี ภาคปกติ ซึ่งเป็นนักศึกษาหมู่เรียน 59, 60, 61, 62 และ 63 จำนวน 18 แอตทริบิวต์ และ 2,418 ชุดข้อมูล ผลการศึกษาพบว่า 1) แต่ละหมู่เรียนมีนักศึกษาออกกลางคัน โดย หมู่เรียน 59 ออกกลางคันร้อยละ 23.59 หมู่เรียน 60 ออกกลางคันร้อยละ 26.52 หมู่เรียน 61 ออกกลางคันร้อยละ 25.33 และหมู่เรียน 62 ออกกลางคันร้อยละ 25.35 2) แบ่งกลุ่มข้อมูล Training :Testing ด้วยอัตราส่วน 70:30 นำปัจจัย 10 ปัจจัยและชุดข้อมูลกลุ่ม Training มาสร้างเป็นโมเดลโดยใช้เทคนิค Classification เปรียบเทียบ 3 วิธี คือ Decision Tree, Naive Bays และ K-Nearest Neighbors 3) ทดสอบประสิทธิภาพของโมเดล ด้วยวิธีการ 10-Fold Cross Validation และวัดประสิทธิภาพด้วยค่า Accuracy พบว่าโมเดลที่สร้างด้วยเทคนิควิธี K-Nearest Neighbors มีค่าความถูกต้อง 92.19% เทคนิควิธี Decision Tree มีค่าความถูกต้อง 91.47% และเทคนิควิธี Naive Bays มีค่าความถูกต้อง 90.16% ตามลำดับ และเมื่อนำโมเดลไปใช้กับข้อมูลทดสอบ พบว่าโมเดลที่สร้างด้วยเทคนิควิธี Naive Bays มีค่าความถูกต้อง 94.20% เทคนิควิธี K-Nearest Neighbors มีค่าความถูกต้อง 93.52% และเทคนิควิธี Decision Tree มีค่าความถูกต้อง 92.84% ตามลำดับ และจากปัจจัยทั้งหมด ปัจจัยที่ทำให้นักศึกษาออกกลางคันสูงสุด 5 อันดับ ได้แก่ เกรดเฉลี่ยกลุ่มวิชาแกน/เอกบังคับ/ครูบังคับ เกรดเฉลี่ยกลุ่มวิชาภาษาและการสื่อสาร เกรดเฉลี่ยกลุ่มวิชาสังคมศาสตร์ เกรดเฉลี่ยกลุ่มวิชามนุษยศาสตร์ และสาขาวิชาที่เรียน จะเห็นได้ว่าปัจจัยส่วนใหญ่เป็นเรื่องของเกรดเฉลี่ยของกลุ่มรายวิชาต่าง ๆ และสาขาวิชาที่เรียน ดังนั้น อาจารย์ที่ปรึกษาและอาจารย์ผู้รับผิดชอบหลักสูตรควรติดตามประเมินผลการเรียนรู้ระหว่างการเรียนในรายวิชาเพื่อให้เห็นถึงแนวโน้มผลการเรียนที่คาดว่าจะได้รับเพื่อจัดกิจกรรมสนับสนุนการเรียนรู้ให้กับนักศึกษาให้ได้รับผลการเรียนตามเกณฑ์ที่เหมาะสม เพื่อลดอัตราการออกกลางคันของนักศึกษา

คำสำคัญ : การออกกลางคัน, การทำเหมืองข้อมูล, ต้นไม้ตัดสินใจ, เคเนียร์เรสเนเบอร์, การเรียนรู้นาอีฟเบย์

Abstract

The objectives of this research were: (1) to study the dropout situation of students at Kamphaeng Phet Rajabhat University, (2) to develop a predictive model for factors influencing student dropout using data mining techniques, and (3) to evaluate the model's performance. The study utilized undergraduate student records from the university's registration database, focusing on cohorts 59, 60, 61, 62, and 63, comprising 18 attributes and 2,418 data entries.

The findings revealed that: (1) the dropout rates for each cohort were 23.59% for cohort 59, 26.52% for cohort 60, 25.33% for cohort 61, and 25.35% for cohort 62. (2) The dataset was split into training and testing sets in a 70:30 ratio, using 10 key factors to build predictive models through Classification techniques, comparing three methods: Decision Tree, Naive Bayes, and K-Nearest Neighbors (KNN). (3) Model performance was evaluated using 10-Fold Cross-Validation and measured by Accuracy. The results showed that KNN achieved an Accuracy of 92.19%, Decision Tree 91.47%, and Naive Bayes 90.16% on the training data. When tested with the testing set, Naive Bayes yielded an Accuracy of 94.20%, KNN 93.52%, and Decision Tree 92.84%.

The top five factors contributing to student dropout were: average grades in core/major/required education courses, average grades in language and communication courses, average grades in social science courses, average grades in humanities courses, and the field of study. It can be observed that most of the factors are related to the students' grade point average (GPA) in different subject groups and their chosen field of study. Therefore, academic advisors and curriculum lecturers should monitor and evaluate students' learning progress during courses to estimate academic performance, enabling the implementation of targeted support activities to help students achieve the required academic standards, thereby reducing the dropout rate.

Keywords : student dropout, data mining, decision tree, K-Nearest Neighbors, Naive Bays

1. บทนำ

มหาวิทยาลัยราชภัฏกำแพงเพชรเป็นสถาบันการศึกษาระดับอุดมศึกษา ที่มีภารกิจในการดำเนินงานของมหาวิทยาลัยตามพระราชบัญญัติมหาวิทยาลัยราชภัฏ พ.ศ. 2547 มาตรา 7 และ มาตรา 8 [1] มหาวิทยาลัยราชภัฏกำแพงเพชรจึงได้กำหนดภารกิจในการดำเนินงานของมหาวิทยาลัย ออกเป็น 6 ด้าน ได้แก่ 1) ด้านการผลิตบัณฑิต 2) ด้านการวิจัย 3) ด้านการบริการวิชาการ 4) ด้านศิลปวัฒนธรรม 5) ด้านการผลิตและพัฒนาครู และ 6) ด้านการบริหารจัดการ ซึ่งภารกิจ 2 ด้านจาก 6 ด้าน คือ ด้านการผลิตบัณฑิต และด้านการผลิตและพัฒนาครู มุ่งเน้นในการผลิตบัณฑิต ซึ่งมหาวิทยาลัยจะต้องดูแลนักศึกษาให้สามารถเล่าเรียนได้จนจบหลักสูตรหรือจบการศึกษาจนเป็นบัณฑิต และนำสิ่งที่ได้เรียนรู้ไปประยุกต์ใช้ในการทำงานและการดำรงชีวิต

มหาวิทยาลัยราชภัฏกำแพงเพชรมีบทบาทสำคัญในการมอบโอกาสทางการศึกษาแก่ชุมชนท้องถิ่น แต่ต้องเผชิญกับปัญหาในการผลิตบัณฑิต ซึ่งจำนวนบัณฑิตน้อยกว่าจำนวนนักศึกษารับเข้าเนื่องมากจากอัตราการออกกลางคันที่สูง ซึ่งอาจส่งผลกระทบต่อภารกิจของมหาวิทยาลัยในการพัฒนาทรัพยากรมนุษย์ในพื้นที่ ทำให้มหาวิทยาลัยจึงต้องเผชิญกับความท้าทายในการรักษานักศึกษาให้อยู่จนสำเร็จการศึกษา ดังนั้น การออกกลางคันของนักศึกษาจึงกลายเป็นปัญหาที่สำคัญที่จะต้องหาแนวทางในการแก้ไข การแก้ปัญหา และจะต้องหาสาเหตุของปัญหาเพื่อที่จะแก้ปัญหาได้ถูกต้อง การศึกษาถึงปัจจัยที่ทำให้นักศึกษาออกกลางคันจึงเป็นสิ่งจำเป็น

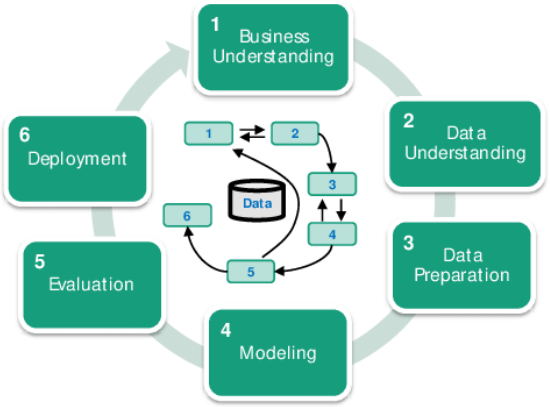
มหาวิทยาลัยมีการจัดเก็บข้อมูลนักศึกษาตั้งแต่เข้าเรียนในมหาวิทยาลัย และเก็บผลการเรียนของนักศึกษาตลอดระยะเวลาที่ศึกษาอยู่ในมหาวิทยาลัย ซึ่งการทำดาต้า ไมน์นิง (data mining) เป็นกระบวนการหนึ่งที่สามารถกระทำกับข้อมูลเหล่านั้นเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล หนึ่งในวิธีการของดาต้า ไมน์นิง คือ การจำแนกประเภทข้อมูล (Classification) จะสามารถนำข้อมูลที่มีในอดีตมาจำแนกและสร้างเป็นโมเดล ผู้วิจัยได้นำวิธีนี้มาใช้ในงานวิจัย ซึ่งจะสามารถทำนายปัจจัยที่มีผลกระทบต่อการการออกกลางคันของนักศึกษาได้ หากนำโมเดลดังกล่าวไปใช้งาน

หรือนำผลที่ได้ไปใช้งานจะช่วยให้ อาจารย์สามารถดูแลช่วยเหลือนักศึกษาได้ และมหาวิทยาลัยสามารถนำไปใช้ในการบริหารจัดการมหาวิทยาลัยได้อีกด้วย

2. ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

2.1 กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM

CRISP-DM หรือ CRoss Industry Standard Process for Data Mining เป็นกระบวนการมาตรฐานในกาวิเคราะห์ข้อมูลด้านดาต้า ไมน์นิง พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือของ 3 บริษัทคือ Daimler Chrysler, SPSS และ NCR ประกอบด้วย 6 ขั้นตอน แต่ละขั้นตอนจะเป็นขั้นตอนที่ต่อเนื่องกัน ขั้นตอนถัดไปจะรอผลลัพธ์จากขั้นตอนก่อนหน้า ขั้นตอนในกระบวนการ CRISP-DM ดังแสดงในรูปที่ 1 [2] โดยมีรายละเอียดของแต่ละขั้นตอนดังนี้ [3]



รูปที่ 1 Cross-Industry Standard Process for Data Mining

1. Business understanding เป็นขั้นตอนแรกในกระบวนการ CRISP-DM ซึ่งเน้นไปที่การเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิง โดยการกำหนดวัตถุประสงค์ของการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิง การวางแผนเบื้องต้นสำหรับการดำเนินงานด้านเหมืองข้อมูล และระบุปัจจัยที่ต้องการวิเคราะห์ เช่น ปัจจัยที่มีผลต่อการออกกลางคันของนักศึกษา

2. Data understanding เริ่มจากการเก็บรวบรวมข้อมูล การตรวจสอบข้อมูลที่รวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์

3. Data preparation ขั้นตอนนี้เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (raw data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (data cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป เป็นต้น โดยขั้นตอนนี้จะเป็นขั้นตอนที่ใช้เวลามากที่สุดของกระบวนการ CRISP-DM

4. Modeling ขั้นตอนนี้จะเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้า ไมน์นิง เช่น การจำแนกประเภทข้อมูล หรือ การแบ่งกลุ่มข้อมูล ซึ่งในขั้นตอนนี้หลายเทคนิคจะถูกนำมาใช้เพื่อให้ได้คำตอบที่ดีที่สุด ดังนั้นในบางครั้งอาจจะต้องมีการย้อนกลับไปที่ขั้นตอนที่ (3) Data Preparation เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย

5. Evaluation ในขั้นตอนนี้จะได้ผลการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้า ไมน์นิง และวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรก หรือ มีความน่าเชื่อถือมากน้อยเพียงใด ซึ่งอาจจะย้อนกลับไปยังขั้นตอนก่อนหน้าเพื่อเปลี่ยนแปลงแก้ไขเพื่อให้ได้ผลลัพธ์ตามที่ต้องการได้

6. Deployment ในกระบวนการทำงานของ CRISP-DM นั้นไม่ได้หยุดเพียงแค่ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้า ไมน์นิงเท่านั้น แม้ว่าผลลัพธ์ที่ได้จะแสดงถึงองค์ความรู้ที่มีประโยชน์ แต่จะต้องนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริงในองค์กรหรือบริษัท ตัวอย่างเช่น การสร้างรายงานเพื่อให้ผู้บริหารหรือนักการตลาดเข้าใจได้ง่ายและสามารถนำไปออกโปรโมชั่นได้ เป็นต้น

2.2 การจำแนกประเภทข้อมูล (Classification)

การจำแนกประเภทข้อมูล มีเป้าหมายเพื่อจัดประเภทหรือกลุ่มของข้อมูลโดยอาศัยข้อมูลที่มีอยู่ในอดีต (Training Data) เพื่อสร้างแบบจำลอง (Model) สำหรับการทำนายข้อมูลใหม่ในอนาคต [2] การจำแนกประเภทนี้สามารถนำไปใช้ได้ในหลายบริบท เช่น การพยากรณ์ผลการเรียนของนักศึกษา การจัดกลุ่มลูกค้าในธุรกิจ หรือการวิเคราะห์ปัจจัยที่ส่งผลต่อการออกกลางคันของนักศึกษา

กระบวนการจำแนกประเภทข้อมูลประกอบด้วย 3 ขั้นตอนสำคัญ ได้แก่:

1. การสร้างโมเดล (Model Building)

ในขั้นตอนแรกของการจำแนกประเภทข้อมูลจะนำข้อมูลในอดีตที่มีการระบุคลาสเป้าหมาย (Target Class) มาสร้างโมเดล โดยโมเดลนี้ใช้สำหรับเรียนรู้ความสัมพันธ์ระหว่างแอตทริบิวต์ทั่วไป (Attributes) และแอตทริบิวต์เป้าหมาย ตัวอย่างเทคนิคที่นิยมในการสร้างโมเดล ได้แก่ Decision Tree, Naive Bayes และ Neural Network [4]

2. การทดสอบประสิทธิภาพของโมเดล (Model Evaluation)

เมื่อสร้างโมเดลเรียบร้อยแล้ว จะต้องทดสอบประสิทธิภาพของโมเดลกับข้อมูลใหม่ที่ไม่เคยใช้ในการฝึกฝน (Test Data) โดยมีการใช้ตัวชี้วัด (Metrics) หลายรูปแบบ เช่น ความแม่นยำ (Accuracy) ความครอบคลุม (Recall) และค่า F1-Score เพื่อตรวจสอบว่าโมเดลสามารถทำงานได้ดีเพียงใด

วิธีการทดสอบที่นิยม ได้แก่:

Cross Validation: แบ่งข้อมูลเป็นหลายส่วนและสลับกันใช้สำหรับฝึกฝนและทดสอบ เพื่อให้ได้ผลลัพธ์ที่น่าเชื่อถือ

10-Fold Cross Validation: แบ่งข้อมูลเป็น 10 ส่วน ใช้ 9 ส่วนสำหรับการฝึกฝน และอีก 1 ส่วนสำหรับการทดสอบในแต่ละรอบ

3. การนำโมเดลไปใช้งาน (Model Deployment)

โมเดลที่ผ่านการประเมินประสิทธิภาพและได้รับการปรับปรุงให้เหมาะสมแล้ว สามารถนำไปใช้ทำนายข้อมูลใหม่ในบริบทที่ต้องการ

2.3 เทคนิคการจำแนกประเภทข้อมูลด้วยวิธี Decision Tree

Decision Tree เป็นหนึ่งในเทคนิคที่นิยมใช้สำหรับการจำแนกประเภทข้อมูล (Classification) เนื่องจากมีความสามารถในการสร้างแบบจำลองที่เข้าใจง่ายและแปลผลได้โดยตรง [5] วิธีนี้ใช้โครงสร้างต้นไม้ในการตัดสินใจ โดยเริ่มจากโหนดราก (Root Node) ที่เป็นแอตทริบิวต์ที่สำคัญที่สุด และแบ่งกิ่งออกไปตามเงื่อนไขที่กำหนด จนกระทั่งถึงโหนดใบ (Leaf Node) ซึ่งแสดงผลลัพธ์ของการจำแนกประเภท

กระบวนการทำงานของ Decision Tree

1. การเลือกแอตทริบิวต์หลัก

ในการสร้างโครงสร้าง Decision Tree จะเริ่มจากการเลือกแอตทริบิวต์ที่สำคัญที่สุดสำหรับการแบ่งกลุ่มข้อมูล โดยใช้ตัวชี้วัดความสัมพันธ์ เช่น Information Gain (IG) หรือ Gini Index

ตัวชี้วัด Information Gain (IG) คำนวณจากค่าความเอนโทรปี (Entropy) ของข้อมูล:

$$IG(parent, child) = Entropy(parent)-[p(c1)x\ Entropy(c1)+p(c2)xEntropy(c2)+...] \quad (1)$$

โดยที่ Entropy(c1) = -p(c1)log p(c1) และ p(c1) คือ ค่าความน่าจะเป็นของค่า c1 ซึ่งคือ Entropy นี้จะใช้ในการวัดความแตกต่างกันของข้อมูล ถ้าข้อมูลมีความแตกต่างกันน้อยค่า Entropy จะมีค่าต่ำ แต่ถ้าข้อมูลมีค่าแตกต่างกันมากค่า Entropy จะมีค่าสูง ดังนั้นถ้าข้อมูล Entropy ของโหนดลูก (child) สามารถแบ่งแยกข้อมูลได้ดีจะมีค่า Entropy ต่ำ และจะทำให้ค่า IG มีค่าสูงเมื่อเทียบกับโหนดบน (parent)

2. การสร้างโครงสร้างต้นไม้

หลังจากเลือกแอตทริบิวต์หลัก จะสร้างโหนดและกิ่งสำหรับแอตทริบิวต์ย่อย โดยทำกระบวนการเลือกแอตทริบิวต์ซ้ำสำหรับข้อมูลที่ถูกแบ่งออกไป การแบ่งข้อมูลจะดำเนินต่อไปจนกว่าจะถึงโหนดใบ ซึ่งข้อมูลทั้งหมดในโหนดนั้นมีคลาสเดียวกัน หรือเงื่อนไขการหยุด (Stopping Condition) ถูกกำหนด เช่น ข้อมูลในโหนดมีจำนวนน้อยเกินไป

3. การตัดแต่งต้นไม้ (Tree Pruning)

เพื่อลดความซับซ้อนของโมเดลและป้องกันการเกิด Overfitting จะมีการตัดแต่งโครงสร้าง Decision Tree โดยการลบโหนดที่ไม่จำเป็นหรือโหนดที่เพิ่มความซับซ้อนโดยไม่มีประโยชน์ วิธีการตัดแต่ง ได้แก่ Pre-Pruning (ตัดแต่งขณะสร้างต้นไม้) และ Post-Pruning (ตัดแต่งหลังสร้างต้นไม้)

เทคนิค Decision Tree เป็นวิธีการที่มีประสิทธิภาพในการจำแนกประเภทข้อมูล โดยเน้นการเลือกแอตทริบิวต์ที่เหมาะสมที่สุดสำหรับการตัดสินใจในแต่ละขั้นตอน กระบวนการนี้ช่วยให้ได้โมเดลที่ใช้งานได้จริงและมีความสามารถในการทำนายที่แม่นยำ

2.4 เทคนิคการจำแนกประเภทข้อมูลด้วยวิธี Naive Bayes

เป็นการสร้างโมเดลแบบง่ายและไม่ซับซ้อน โดยอาศัยทฤษฎีความน่าจะเป็นเป็นหลัก [2]

$$P(C|A) = \frac{P(A|C)\text{-}P(C)}{P(A)} \quad (2)$$

จากสมการของ Bayes อธิบายได้ว่า ถ้าต้องการทำนายคลาส C เมื่อทราบแอตทริบิวต์ A แล้วสามารถคำนวณได้จากความน่าจะเป็นของแอตทริบิวต์ A ที่มีคลาส C ในเทรนนิงดาต้า และค่าความน่าจะเป็นของแอตทริบิวต์ A และคลาส C

$P(C|A)$ คือ ค่าความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น A และมีคลาส C

$P(A|C)$ คือ ค่าความน่าจะเป็นที่ข้อมูลเทรนนิง ดาต้าที่มีคลาส C และมีแอตทริบิวต์ A โดยที่ $A = a_1 \cap a_2 \ldots \cap a_M$ และ M คือจำนวนแอตทริบิวต์ในเทรนนิง ดาต้า

$P(C)$ คือ ค่าความน่าจะเป็นของคลาส C

แต่การที่แอตทริบิวต์ $A = a_1 \cap a_2 \ldots \cap a_M$ จะเกิดขึ้นในเทรนนิง ดาต้าอาจจะมีจำนวนน้อยมากหรือไม่มีรูปแบบของแอตทริบิวต์แบบนี้เกิดขึ้นเลย ดังนั้นจึงได้ใช้หลักการที่ว่าแต่ละแอตทริบิวต์ทุกตัวมีความเป็น independent ต่อกันทำให้สามารถเปลี่ยน $P(A|C)$ ได้เป็น

$$P(A|C) = P(a_1|C)x\ P(a_1|C)x...\ P(a_M|C) \quad (3)$$

ทำให้เวลาคำนวณสามารถคำนวณแยกทีละแอตทริบิวต์ได้ดังสมการ

$$P(C|A) = P(a_1|C)x\ P(a_1|C)x...x\ P(a_M|C)x\ P(C) \quad (4)$$

2.5 เทคนิคการจำแนกประเภทข้อมูลด้วยวิธี K-Nearest Neighbors

วิธีการ K-Nearest Neighbors (K-NN) [2] คล้ายกับการแบ่งกลุ่มข้อมูล คือทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ตัว และคำตอบที่ทำนายได้คือคลาสที่พบมากที่สุดของข้อมูลที่เป็นเพื่อนบ้างทั้ง K ตัว ในทางเทคนิคมักใช้วิธีการวัดระยะห่างแบบ Euclidean ซึ่งเกิดจากรากที่สองของผลต่างระหว่างแอตทริบิวต์ต่าง ๆ ยกกำลังสอง ดังสมการ

$$D_{Euclidean} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_L - y_L)^2} \quad (5)$$

โดยที่ x_1 คือ แอตทริบิวต์ที่ 1 ของข้อมูลจุดที่ 1 และ y_1 คือ แอตทริบิวต์ที่ 1 ของข้อมูลจุดที่ 2 โดยข้อมูลทั้งสองตัว (X และ Y) มีจำนวนแอตทริบิวต์เท่ากับ L

นอกจากนี้จะเห็นได้ว่าวิธี K-NN ไม่มีการสร้างโมเดลแต่จะใช้ข้อมูลเทรนนิง ดาต้าทั้งหมดมาช่วยในการนำนาย ทำให้วิธีการนี้ถูกจัดอยู่ในกลุ่มของ lazy model นั่นเอง

2.6 งานวิจัยที่เกี่ยวข้อง

จีระนันต์ เจริญรัตน์ [6] ได้ทำงานวิจัย เรื่อง การวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาที่มีผลการเรียนปกติโดยใช้ต้นไม้ตัดสินใจ ซึ่งงานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาระดับปริญญาตรี ที่มีผลการเรียนของเกรดเฉลี่ยสะสมตั้งแต่ 2.00 ขึ้นไป ในการวิเคราะห์ข้อมูลจะใช้ข้อมูลนักศึกษาระดับปริญญาตรี คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏสกลนคร ที่เข้าศึกษาระหว่างปี พ.ศ. 2553-2557 มีจำนวน 3,385 ชุดข้อมูล เลือกใช้เทคนิคการทำเหมืองข้อมูล แบบ Classification เลือกการทำนายข้อมูลด้วยวิธี Decision Tree และใช้อัลกอริธึมชนิด J48 การทดสอบโมเดลที่ได้จะทำการทดสอบแบบ 10-fold cross validation โดยใช้โปรแกรม WEKA ผลการวิเคราะห์พบว่าปัจจัยที่มีความสำคัญที่จะส่งผลต่อการพ้นสภาพของนักศึกษาที่มีผลการเรียนปกติแบ่งเป็น 4 กลุ่ม ตามกลุ่มเกรดเฉลี่ยสะสม ดังนี้ กลุ่ม Best (GPA>3.50 ขึ้นไป) ปัจจัยคือ วุฒิการศึกษาเดิม กลุ่ม Excellent (GPA=3.00-3.50) ปัจจัยคือ อาชีพมารดา และสาขาวิชาที่เรียน กลุ่ม Good (GPA=2.50-2.99) ปัจจัยคือ ทุนกู้ยืมเพื่อการศึกษา สถานภาพของครอบครัว รายได้ของบิดา รายได้ของมารดา และจังหวัด กลุ่ม Medium (GPA=2.00-2.49) ปัจจัยคือ ทุนกู้ยืมเพื่อการศึกษา สถานภาพครอบครัว และรายได้ของมารดา จากผลการวิจัยนี้สามารถนำกฎการจำแนกข้อมูลที่ได้มาใช้ในการพัฒนาระบบพยากรณ์การพ้นสภาพของนักศึกษาต่อไป และผู้บริหารสามารถใช้เป็นแนวทางในการบริหารจัดการหรืออาจารย์ที่ปรึกษาสามารถวางแผนการเรียน ดูแลการลงทะเบียนเรียนของนักศึกษา ให้ความช่วยเหลือ และส่งเสริมนักศึกษาได้อย่างเหมาะสม

ชนิดาภา บุญประสม และจรัญ แสนราช [7] ได้ทำงานวิจัย เรื่อง การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคการทำเหมืองข้อมูล ซึ่งการวิจัยมีวัตถุประสงค์เพื่อ 1) วิเคราะห์หาปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาระดับปริญญาตรี 2) สังเคราะห์โมเดลสำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี และ 3) เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของโมเดลด้วยเทคนิควิธี Decision Tree, K-Nearest Neighbors, Naive Bayes โดยใช้ข้อมูลจากฐานข้อมูลงานทะเบียนของมหาลัยวิทยาลัยราชภัฏอุบลราชธานี ของนักศึกษาระดับปริญญาตรี ระหว่างปีการศึกษา 2558-2560 มีจำนวน 11 แอททริบิวต์และ 13,729 ชุดข้อมูล เมื่อนำมาวิเคราะห์ค่าน้ำหนักของแอตทริบิวต์ ด้วยวิธีการ Information Theory พบว่า 1) มีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาจำนวน 8 ปัจจัย 2) นำปัจจัยที่ได้มาทำการสร้างเป็นโมเดลทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold Cross Validation และวัดประสิทธิภาพด้วย ค่า Accuracy เพื่อหาวิธีการที่มีความถูกต้องมากที่สุด 3) ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลพบว่าโมเดลที่ สร้างด้วยเทคนิควิธี Naive Bayes มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้อง 93.58 % มากกว่าเทคนิควิธี Decision Tree มีค่าเฉลี่ยความถูกต้อง 93.52 % และเทคนิควิธี K-Nearest Neighbors มีค่าเฉลี่ยความถูกต้อง 87.95 % และมี ปัจจัยที่เกี่ยวข้องสูงสุด 5 อันดับ ได้แก่ การกู้ยืมกองทุนเพื่อการศึกษา สาขาวิชา เกรดเฉลี่ย อาชีพของมารดา และอาชีพ ของบิดา

3. วิธีดำเนินการวิจัย

ในการศึกษาสถานการณ์การออกกลางคันของนักศึกษามหาวิทยาลัยราชภัฏกำแพงเพชร ผู้วิจัยได้นำข้อมูลของนักศึกษาระดับปริญญาตรี ภาคปกติ ปีการศึกษา 2559-2563 มาจัดกลุ่มเป็นนักศึกษาที่มีสถานะออกกลางคัน ได้แก่

นักศึกษาที่ลาออก และออกตามระเบียบข้อบังคับของมหาวิทยาลัย และนักศึกษาที่มีสถานะอื่น ๆ จะอยู่ในกลุ่มสถานะคงอยู่ จากนั้นนำข้อมูลของนักศึกษาหมู่เรียน 59-62 มาสรุปเป็นร้อยละ เพื่อศึกษาสถานการณ์การออกกลางคัน และนำข้อมูลนักศึกษาหมู่เรียน 59-63 แบ่งกลุ่มข้อมูล Training :Testing ด้วยอัตราส่วน 70:30 เพื่อนำไปสร้างโมเดล

สำหรับการสร้างโมเดลในการทำนายปัจจัยที่มีผลกระทบต่อการออกกลางคันของนักศึกษาโดยใช้เทคนิคดาต้า ไมน์นิง และการทดสอบประสิทธิภาพของโมเดลที่พัฒนา ผู้วิจัยได้เสนอวิธีการทำดาต้า ไมน์นิง โดยใช้โปรแกรม WEKA โดยขั้นตอนการทำใช้กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM (CRoss Industry Standard Process for Data Mining) ซึ่งประกอบด้วย 6 ขั้นตอนการดำเนินงาน ดังนี้

1. การทำความเข้าใจธุรกิจ (Business Understanding) เข้าใจปัญหาและทำการวิเคราะห์เหมืองข้อมูล โดยวางแผนเบื้องต้นสำหรับการดำเนินงานด้านเหมืองข้อมูล และระบุปัจจัยที่ต้องการวิเคราะห์ ซึ่งปัจจัยที่ต้องการวิเคราะห์คือปัจจัยที่มีผลต่อการออกกลางคันของนักศึกษา โดยสร้างโมเดลเพื่อทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี ภาคปกติ ของนักศึกษามหาวิทยาลัยราชภัฏกำแพงเพชร โดยใช้เทคนิคการทำเหมืองข้อมูลด้วยเทคนิควิธี Decision Tree ชุดข้อมูลที่ใช้ในการวิเคราะห์ เป็นชุดข้อมูลของนักศึกษาระดับปริญญาตรี ภาคปกติ ของนักศึกษามหาวิทยาลัยราชภัฏกำแพงเพชร ตั้งแต่ปีการศึกษา 2559-2563

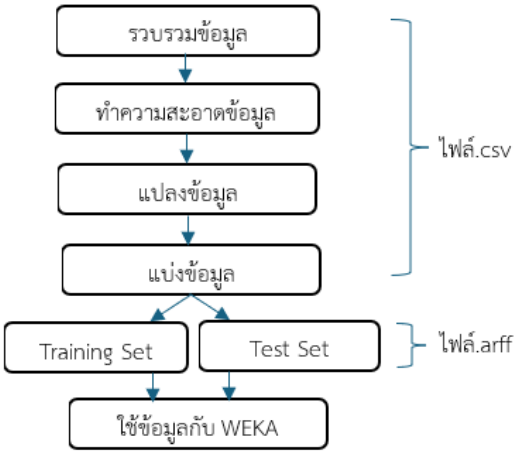
2. การทำความเข้าใจข้อมูล (Data Understanding) ทำการรวบรวมข้อมูล (Data Collection) ของนักศึกษา ทั้งข้อมูลส่วนตัว ข้อมูลด้านการศึกษาก่อนเข้าเรียนที่มหาวิทยาลัย และข้อมูลนักศึกษาขณะเรียนที่มหาวิทยาลัยจากฐานข้อมูลงานทะเบียนของมหาวิทยาลัยราชภัฏกำแพงเพชร ซึ่งมี 18 แอตทริบิวต์ สำหรับการจำแนกประเภทเพื่อหาประสิทธิภาพในการจำแนกข้อมูลและสร้างความสัมพันธ์ของแอตทริบิวต์

3. การเตรียมข้อมูล (Data Preparation) เป็นการจัดการข้อมูลเพื่อเตรียมพร้อมสำหรับการวิเคราะห์และการสร้างโมเดลในการทำนายข้อมูล โดยข้อมูลที่ได้เก็บรวบรวมในรูปแบบของตาราง Excel โดย

3.1 การคัดเลือกข้อมูล (Data Selection) เลือกใช้ข้อมูลที่เกี่ยวข้องโดยศึกษาจากงานวิจัยที่เกี่ยวกับการออกกลางคัน [6-9] และสำนักส่งเสริมวิชาการและงานทะเบียนมีการจัดเก็บข้อมูล ประกอบด้วย 18 แอตทริบิวต์ 1) แอตทริบิวต์ที่เป็น Input 17 แอตทริบิวต์ ได้แก่ เพศของนักศึกษา อาชีพของบิดาและมารดา รายได้ของบิดาและมารดา เกรดเฉลี่ยก่อนเข้าศึกษาระดับอุดมศึกษา วุฒิการศึกษาที่เรียนจบก่อนเข้าศึกษาระดับอุดมศึกษา แผนการเรียนที่จบก่อนเข้าศึกษาระดับอุดมศึกษา สาขาวิชาที่เรียนในมหาวิทยาลัย เกรดเฉลี่ยกลุ่มวิชาภาษาและการสื่อสาร เกรดเฉลี่ยกลุ่มวิชามนุษยศาสตร์ เกรดเฉลี่ยกลุ่มวิชาสังคมศาสตร์ เกรดเฉลี่ยกลุ่มวิชาวิทยาศาสตร์ เกรดเฉลี่ยกลุ่มวิชาแกน/เอกบังคับ/ครูบังคับ เกรดเฉลี่ยกลุ่มวิชาเอกเลือก การกู้ยืมกยศ. ค่าเล่าเรียนและค่าครองชีพ 2) แอตทริบิวต์ที่เป็น Output 1 แอตทริบิวต์ คือ สถานะการออกกลางคันของนักศึกษา

3.2 การกลั่นกรองข้อมูล (Data cleaning) เป็นการปรับข้อมูลให้ถูกต้อง ซึ่งอยู่ในรูปแบบของตาราง Excel โดยตัดข้อมูลที่ผิดพลาด ข้อมูลผิดรูปแบบ ข้อมูลที่มีค่าว่าง ออก คงเหลือข้อมูลที่นำไปใช้ในการประมวลผล 2,418 รายการ

3.3 การแปลงรูปแบบของข้อมูล (Data transformation) ข้อมูลที่นำมาวิเคราะห์มีรูปแบบในการจัดเก็บทั้งประเภท Nominal และ Numeric จึงต้องแปลงให้อยู่ในรูปแบบเดียวกัน เช่น เพศ (1) ชาย (2) หญิง รายได้ต่อปี แบ่งออกเป็นช่วง (11) รายได้ ≤ 80,000 บาท/ปี (12) รายได้ 80,001-100,000 บาท/ปี (13) รายได้ 100,001-120,000 บาท/ปี (14) รายได้ 120,001 - 130,000 บาท/ปี (15) รายได้ 130,001 - 149,999 บาท/ปี (20) รายได้ 150,000 - 300,000 บาท/ปี (30) รายได้ > 300,000 บาท/ปี เกรดเฉลี่ยรายวิชากลุ่มต่าง ๆ แบ่งออกเป็นช่วง (1) มีเกรดเฉลี่ย <1.00 (2) มีเกรดเฉลี่ย 1.00-1.49 (3) มีเกรดเฉลี่ย 1.50-1.99 (4) มีเกรดเฉลี่ย 2.00-2.49 (5) มีเกรดเฉลี่ย 2.50-2.99 (6) มีเกรดเฉลี่ย 3.00-3.49 (7) มีเกรดเฉลี่ย 3.50-3.99 (8) มีเกรดเฉลี่ย 4.00 การกู้ยืมกยศ. ค่าเล่าเรียนและกู้ยืมกยศ.ค่าครองชีพ (0) ไม่กู้ (1) กู้ เป็นต้น ในส่วนของการจัดกลุ่มสถานะการออกกลางคันของนักศึกษาจะจัดกลุ่มเป็น 2 กลุ่ม คือ Y ออกกลางคัน และ N คงอยู่ และกำหนดเป็นคลาสเป้าหมาย



รูปที่ 2 การเตรียมข้อมูลก่อนทำ Data Mining ด้วย WEKA

4. การสร้างโมเดล (Modeling) การสร้างและทดสอบความถูกต้องของโมเดล ด้วยเทคนิคทาง Data Mining สังเคราะห์โมเดลจากข้อมูลที่มีอยู่ เพื่อทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี ภาคปกติ ซึ่งผู้วิจัยใช้โปรแกรม Weka ในการสร้างโมเดลการทำนายด้วยวิธี Decision Tree, Naive Bays และ K-Nearest Neighbors

5. การประเมินผล (Evaluation) ทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold cross validation ในการวัดประสิทธิภาพของการจำแนกประเภทข้อมูล ได้แก่ การหาค่าความระลึก (Recall) ค่าความถ่วงดุล (F-measure) ความแม่นยำ (Precision) และค่าความถูกต้อง (Accuracy)

6. การนำแบบจำลองไปใช้งาน (Deployment) นำโมเดลที่พัฒนาไปใช้ในการทำนายการออกกลางคันของนักศึกษาเพื่อใช้เป็นแนวทางในการป้องกันและแก้ไขปัญหาการออกกลางคันของนักศึกษาต่อไป

4. ผลการวิจัย

4.1 ผลการศึกษาสถานการณ์การออกกลางคันของนักศึกษา มหาวิทยาลัยราชภัฏกำแพงเพชร โดยใช้ข้อมูลของนักศึกษาระดับปริญญาตรี ภาคปกติ หมู่เรียน 59-62 นำมาสรุปเป็นร้อยละ มีผลดังตารางที่ 1

ตารางที่ 1 ร้อยละการออกกลางคันของนักศึกษาระดับปริญญาตรี ภาคปกติ หมู่เรียน 59-62

หมู่เรียน	ร้อยละการออกกลางคัน
59	23.59
60	26.52
61	25.33
62	25.35

จากตารางที่ 1 จะเห็นได้ว่ามีการออกกลางคันของนักศึกษาอย่างต่อเนื่อง โดยมีอัตราการออกกลางคันของนักศึกษาทุกหมู่เรียนใกล้เคียงกัน ซึ่งคิดเป็น 1 ใน 4 ของนักศึกษาของแต่ละหมู่เรียน

4.2 ผลการสร้างโมเดลในการทำนายปัจจัยที่มีผลกระทบต่อการออกกลางคันของนักศึกษาโดยใช้เทคนิคดาต้า ไมน์นิง

ผู้วิจัยได้แบ่งกลุ่มข้อมูล Training :Testing ด้วยอัตราส่วน 70:30 และนำไปสร้างโมเดลโดยนำแอตทริบิวต์ 18 แอตทริบิวต์มาใช้ในการสร้างโมเดล โดยใช้เทคนิค Decision Tree ด้วยอัลกอริทึม C4.5 (J48) หลังจากสร้างโมเดล พบว่า ปัจจัยที่มีผลต่อการออกกลางคันของนักศึกษามีทั้งหมด 10 ปัจจัย เรียงลำดับจากมากไปน้อย ได้แก่

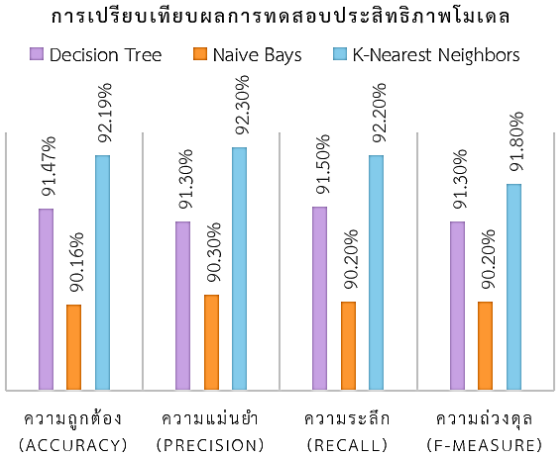
1) เกรดเฉลี่ยกลุ่มวิชาแกน/เอกบังคับ/ครูบังคับ
2) เกรดเฉลี่ยกลุ่มวิชาภาษาและการสื่อสาร
3) เกรดเฉลี่ยกลุ่มวิชาสังคมศาสตร์
4) เกรดเฉลี่ยกลุ่มวิชามนุษยศาสตร์
5) สาขาวิชาที่เรียนในมหาวิทยาลัย
6) เกรดเฉลี่ยกลุ่มวิชาวิทยาศาสตร์
7) เกรดเฉลี่ยกลุ่มวิชาเอกเลือก
8) เกรดเฉลี่ยก่อนเข้าศึกษาระดับอุดมศึกษา
9) วุฒิการศึกษาที่เรียนจบก่อนเข้าศึกษาระดับอุดมศึกษา
10) แผนการเรียนที่จบก่อนเข้าศึกษาระดับอุดมศึกษา

จากนั้นจึงนำปัจจัย 10 ปัจจัยและชุดข้อมูลกลุ่ม Training มาสร้างเป็นโมเดลด้วยเทคนิควิธี Decision Tree, Naive Bays และ K-Nearest Neighbors

4.3 ผลการทดสอบประสิทธิภาพโมเดล สำหรับการทำนายการออกกลางคันของนักศึกษาระดับปริญญาตรี ด้วยวิธีการ 10-Fold Cross Validation มีผลดังตารางที่ 2

ตารางที่ 2 ผลการทดสอบประสิทธิภาพโมเดล

เทคนิควิธีดาต้าไมน์นิง	ความถูกต้อง (Accuracy)	ความแม่นยำ (Precision)	ความระลึก (Recall)	ความถ่วงดุล (F-measure)
Decision Tree	91.47%	91.30%	91.50%	91.30%
Naive Bays	90.16%	90.30%	90.20%	90.20%
K-Nearest Neighbors	92.19%	92.30%	92.20%	91.80%



รูปที่ 3 กราฟแสดงผลการเปรียบเทียบการทดสอบประสิทธิภาพโมเดล

จากกราฟจะเห็นได้ว่าโมเดลที่สร้างด้วยเทคนิควิธี K-Nearest Neighbors มีประสิทธิภาพสูงสุด มีค่าความถูกต้องร้อยละ 92.19 เทคนิควิธี Decision Tree มีค่าความ

ถูกต้องร้อยละ 91.47 และเทคนิควิธี Naive Bays มีค่าความถูกต้องร้อยละ 90.16 ตามลำดับ

4.4 ผลการใช้ข้อมูลทดสอบกับโมเดล เมื่อนำโมเดลไปใช้กับข้อมูลทดสอบ พบว่า โมเดลที่สร้างด้วยเทคนิควิธี Naive Bays มีค่าความถูกต้องร้อยละ 94.20 เทคนิควิธี K-Nearest Neighbors มีค่าความถูกต้องร้อยละ 93.52 และเทคนิควิธี Decision Tree มีค่าความถูกต้องร้อยละ 92.84 ตามลำดับ

จากปัจจัยทั้งหมด ปัจจัยที่ทำให้นักศึกษาออกกลางคันสูงสุด 5 อันดับ ได้แก่ เกรดเฉลี่ยกลุ่มวิชาแกน/เอกบังคับ/ครูบังคับ เกรดเฉลี่ยกลุ่มวิชาภาษาและการสื่อสาร เกรดเฉลี่ยกลุ่มวิชาสังคมศาสตร์ เกรดเฉลี่ยกลุ่มวิชามนุษยศาสตร์ และสาขาวิชาที่เรียน

5. สรุปและอภิปรายผลการวิจัย

สถานการณ์การออกกลางคันของนักศึกษาระดับปริญญาตรี ภาคปกติ มหาวิทยาลัยราชภัฏกำแพงเพชร ของนักศึกษาหมู่เรียน 59-62 โดยรวบรวมข้อมูลจากฐานข้อมูลงานทะเบียนนักศึกษามหาวิทยาลัยราชภัฏกำแพงเพชร ทั้งข้อมูลส่วนตัว ข้อมูลด้านการศึกษาก่อนเข้าเรียนที่มหาวิทยาลัย และข้อมูลนักศึกษาขณะเรียนที่มหาวิทยาลัย พบว่ามีการออกกลางคันอย่างต่อเนื่องและมีอัตราการออก 1 ใน 4 ของจำนวนนักศึกษาในแต่ละหมู่เรียน ส่งผลกระทบต่อการผลิตบัณฑิตที่จะเป็นกำลังสำคัญในการพัฒนาท้องถิ่นและพัฒนาประเทศ รวมถึงส่งผลกระทบต่อการบริหารจัดการของมหาวิทยาลัย

จากการศึกษาวิเคราะห์ปัจจัยที่มีผลต่อการออกกลางคันของนักศึกษาระดับปริญญาตรี ภาคปกติ โดยนำข้อมูลจำนวน 18 แอตทริบิวต์ มาสร้างโมเดลด้วยเทคนิค Decision Tree ภายใต้อัลกอริทึม C4.5 (J48) พบว่า มีเพียง input 10 ปัจจัย ที่มีนัยสำคัญต่อการพยากรณ์แนวโน้มการออกกลางคัน ขณะที่ 7 ปัจจัยที่เหลือไม่มีความสัมพันธ์ที่มีนัยสำคัญเพียงพอ หรืออาจไม่มีอิทธิพลต่อกระบวนการจำแนกประเภทอย่างเด่นชัด ดังนั้น ในกระบวนการสร้างโมเดล Naive Bayes และ K-Nearest Neighbors (KNN) ได้มีการเลือกใช้เฉพาะ 10 ปัจจัยที่สำคัญที่สุด เพื่อลดความซับซ้อนของโมเดล และเพิ่มประสิทธิภาพในการพยากรณ์ ซึ่งสอดคล้องกับแนวปฏิบัติที่เป็นมาตรฐานในงานด้าน Data Mining และ Machine Learning ที่มุ่งเน้นการใช้ Feature Selection เพื่อลดปัจจัยที่ไม่เกี่ยวข้อง อันจะช่วยเพิ่มประสิทธิภาพของโมเดลและลดความเสี่ยงของปัญหา Overfitting

ปัจจัยทั้ง 10 รายการ ที่ได้รับการคัดเลือก ส่วนใหญ่เกี่ยวข้องกับผลสัมฤทธิ์ทางการศึกษา ประกอบด้วย เกรดเฉลี่ยของกลุ่มรายวิชาต่าง ๆ สาขาวิชาที่ศึกษา เกรดเฉลี่ยก่อนเข้าศึกษาระดับอุดมศึกษา และวุฒิการศึกษาที่สำเร็จก่อนเข้าศึกษาระดับอุดมศึกษา ซึ่งสอดคล้องกับผลการศึกษาของ ชนิดาภา บุญประสม และจรัญ แสนราช ในส่วนของเกรดเฉลี่ย และสาขาวิชา แต่ไม่สอดคล้องกับผลการศึกษาของ จีระนันต์ เจริญรัตน์ [6] และ ชนิดาภา บุญประสม และจรัญ แสนราช [7] ในส่วนของปัจจัยด้านรายได้ของบิดามารดา และการกู้ยืมเงินเพื่อการศึกษา ที่มีผลกระทบต่อการออกกลางคัน ทั้งนี้ความแตกต่างของผลการศึกษาอาจเกิดจากความแตกต่างของกลุ่มตัวอย่างที่ใช้ในการวิจัย โดยกลุ่มตัวอย่างในงานวิจัยนี้อาจมีสถานะทางเศรษฐกิจของครอบครัวและการสนับสนุนทางการเงินจากครอบครัวที่แตกต่างกัน อีกทั้งนักศึกษาที่กู้ยืมเงินเพื่อการศึกษา (กยศ.) ได้รับการอนุมัติทุกคนคิดเป็นร้อยละ 100 รวมทั้งนักศึกษายังได้รับมอบทุนการศึกษาผ่านมหาวิทยาลัยจากแหล่งอื่นอีกด้วย ส่งผลให้ปัจจัยด้านรายได้และการกู้ยืมเงินเพื่อการศึกษามีผลกระทบต่อการตัดสินใจลาออกจากมหาวิทยาลัยน้อยลง

เพื่อช่วยลดอัตราการออกกลางคันของนักศึกษา บทบาทของอาจารย์ที่ปรึกษาและอาจารย์ประจำสาขาวิชาจึงมีความสำคัญอย่างยิ่ง ควรมีการติดตามผลการเรียนของนักศึกษาอย่างใกล้ชิด เพื่อช่วยเหลือและให้คำแนะนำทางด้านวิชาการ โดยเฉพาะในกลุ่มรายวิชาที่นักศึกษามีแนวโน้มที่จะประสบปัญหาในการเรียนสูง นอกจากนี้มหาวิทยาลัยสามารถนำข้อมูลที่ได้จากงานวิจัยนี้ไปใช้ในการวางแผนการสนับสนุนทางวิชาการ เช่น โครงการติวเสริม การให้คำปรึกษาทางการศึกษา หรือโปรแกรมช่วยเหลือนักศึกษาที่มีผลการเรียนต่ำ เพื่อเพิ่มโอกาสในการสำเร็จการศึกษาของนักศึกษา

อย่างไรก็ตาม งานวิจัยนี้มีข้อจำกัดในเรื่องของแหล่งข้อมูลที่ใช้ในการวิเคราะห์ ซึ่งอ้างอิงจากฐานข้อมูลของสำนักส่งเสริมวิชาการและงานทะเบียนของมหาวิทยาลัยเป็นหลัก จึงอาจยังไม่ได้ครอบคลุม ปัจจัยด้านพฤติกรรมทางสังคม ความเครียด แรงกดดันทางจิตวิทยา หรือปัจจัยด้านสิ่งแวดล้อมของนักศึกษา ซึ่งอาจมีอิทธิพลต่อการตัดสินใจลาออกจากมหาวิทยาลัยในบางกรณี ดังนั้น งานวิจัยในอนาคตควรพิจารณาใช้วิธีการเก็บข้อมูลที่ครอบคลุมมากขึ้น เช่น การใช้แบบสอบถามหรือการสัมภาษณ์เชิงลึก เพื่อนำปัจจัยทางด้านจิตวิทยาและสังคมเข้ามาร่วมวิเคราะห์ เพื่อให้สามารถระบุสาเหตุของการออกกลางคันได้อย่างแม่นยำยิ่งขึ้น

ในส่วนของการทดสอบประสิทธิภาพของโมเดลที่สร้างด้วยเทคนิควิธี Decision Tree, Naive Bays และ K-Nearest Neighbors ด้วยวิธีการ 10-Fold Cross Validation พบว่าโมเดลที่สร้างจากชุดข้อมูล Training ด้วยวิธี K-Nearest

Neighbors มีประสิทธิภาพสูงสุด มีค่าความถูกต้องร้อยละ 92.19 มากกว่า Decision Tree และ Naive Bays แต่เมื่อนำโมเดลไปใช้กับชุดข้อมูลทดสอบกลับพบว่า โมเดลที่สร้างด้วยวิธี Naive Bays มีประสิทธิภาพสูงสุด มีค่าความถูกต้องร้อยละ 94.20 มากกว่า Naive Bays และ Decision Tree ที่ผลลัพธ์เป็นเช่นนี้ อาจจะเกิดจาก Overfitting โมเดล Decision Tree อาจมีความซับซ้อนเกินไป จึงสามารถจำรูปแบบของข้อมูลฝึกสอนได้ดีมาก แต่เมื่อนำไปใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน กลับไม่สามารถทำนายได้ดีเท่าที่ควร เนื่องจากโมเดลจับความผิดพลาด (noise) ในข้อมูลฝึกสอนไปด้วย หรือ Underfitting โมเดล Naive Bayes อาจมีความซับซ้อนน้อยเกินไป จึงไม่สามารถจับรูปแบบของข้อมูลได้อย่างละเอียดพอ เมื่อนำไปใช้กับข้อมูลฝึกสอน จึงให้ผลลัพธ์ที่ไม่ดีนัก แต่เมื่อนำไปใช้กับข้อมูลทดสอบ ซึ่งอาจมีรูปแบบที่เรียบง่ายกว่า กลับให้ผลลัพธ์ที่ดีขึ้น หรือ ความแตกต่างของข้อมูล ข้อมูลฝึกสอนและข้อมูลทดสอบอาจมีความแตกต่างกันในบางแง่มุม เช่น การกระจายของข้อมูล ความสมดุลของคลาส ทำให้โมเดลที่ทำงานได้ดีกับข้อมูลฝึกสอน อาจไม่ทำงานได้ดีกับข้อมูลทดสอบ ดังนั้นการเปรียบเทียบผลลัพธ์ของโมเดลต่างๆ จะช่วยให้เข้าใจถึงข้อดีข้อเสียของแต่ละโมเดล และสามารถเลือกโมเดลที่เหมาะสมกับปัญหาได้มากที่สุด ซึ่งการเลือกโมเดลที่ดีที่สุดนั้นขึ้นอยู่กับหลายปัจจัย เช่น ลักษณะของข้อมูล เป้าหมายของการทำนาย และข้อจำกัดด้านทรัพยากร อย่างไรก็ตาม สามารถนำโมเดลนี้ไปใช้ในการทำนายการออกกลางคันของนักศึกษาต่อไปได้

6. ข้อเสนอแนะ

1. ควรมีการปรับปรุงโมเดล Decision Tree, Naive Bayes, และ K-Nearest Neighbors (KNN) เพื่อลดปัญหาการ Overfitting หรือ Underfitting เพื่อเพิ่มความแม่นยำ

2. ควรใช้เทคนิคดาต้า ไมน์นิงอื่นๆ เพิ่มเติม จากนั้นเปรียบเทียบค่าความถูกต้อง (accuracy) ซึ่งจะช่วยให้ได้ภาพรวมที่ดีขึ้นเกี่ยวกับประสิทธิภาพของโมเดลและสามารถเลือกวิธีที่เหมาะสมที่สุด

3. ควรนำปัจจัยด้านพฤติกรรมทางสังคม ความเครียด แรงกดดันทางจิตวิทยา หรือปัจจัยด้านสิ่งแวดล้อมของนักศึกษา รวมถึงสถานการณ์ต่าง ๆ ในปัจจุบัน เข้ามาร่วมวิเคราะห์ เพื่อให้สามารถระบุสาเหตุของการออกกลางคันได้อย่างแม่นยำยิ่งขึ้น

7. กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณสำนักส่งเสริมวิชาการและงานทะเบียนที่ให้ข้อมูลที่นำมาใช้ในการทำวิจัย และขอขอบคุณมหาวิทยาลัยราชภัฏกำแพงเพชรที่ให้การสนับสนุนทุนวิจัยในครั้งนี้

8. เอกสารอ้างอิง

[1] Kamphaeng Phet Rajabhat University. Roles and duties of Kamphaeng Phet Rajabhat University. [Online]. [Cited November 21, 2019]. Available: https://www.kpru.ac.th/about-kpru/kpru-mission.php (in Thai).

[2] E. Pacharawongsakda. *An Introduction to Data Mining Techniques*. Bangkok: Asia Digital Co. Ltd. ,2014 (in Thai).

[3] M. Frye, D. Gyulai, J. Bergmann, and R. Schmitt, “Adaptive scheduling through machine learning-based process parameter prediction,” *MM Journal, Special Issue on HSM2019.*pp.3060-3066, Dec.2020.

[4] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. (3rd Edition). San Francisco: Morgan Kaufmann, 2011.

[5] J. R. Quinlan. *Induction of Decision Trees. Machine Learning*, 1986.

[6] J. Chareonrat. “Analysis on factors affecting normal-grade student dismissal using decision tree,” *SNRU Journal of Science and Technology*, vol. 8, no.2, pp. 256-267.May.-Aug. 2016 (in Thai).

[7] C. Boonprasom and C. Sanrach. “Predictive Analytic for Student Dropout in Undergraduate using Data Mining Technique, *Technical Education Journal : King Mongkut’s University of Technology North Bangkok*, vol. 9, no. 1, pp. 142-151, Jan.-Apr. 2018 (in Thai).

[8] P. Laopilai and C. Sanrach. “Analysis for Student Dropout in Undergraduate Using Data Mining Technique,” *The Sci J of PhetchaburiRajabhat University*,vol.16, no.2,pp. 61-71,Jul.- Dec. 2019 (in Thai).

[9] A. Paphat, W. Sriurai, and N. Ditcharoen. “Improved University Student Dropout Prediction Using Feature Selection with Multilayer Perceptron Neural Network,” *Journal of Science and Science Education*, vol. 5, no. 1, pp.39-48, Jan.-Jun. 2022 (in Thai).