การเปรียบเทียบประสิทธิภาพแบบจำลองการทำนายผลการเรียนของนิสิตที่ใช้งานระบบการจัดการเรียนรู้ออนไลน์ด้วยการเรียนรู้ของเครื่อง

Comparison of Student Learning Outcome Prediction Models in an Online Learning Management System Using Machine Learning

ปริยานุช ประเสริฐสิริกุล* ศิริสรรพ เหล่าหะเกียรติ เรืองศักดิ์ ตระกูลพุทธิรักษ์ และ ศศิวิมล สุขพัฒน์

Pariyanuch Prasertisirikul*, Sirisup Laohakiat, Ruangsak Trakunphutthirak, and Sasivimon Sukaphat

สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

Master of Science in Data Science (MSDS) program, Srinakharinwirot University

Email: pariyanuch.prasertsirikul@g.swu.ac.th

บทคัดย่อ

งานวิจัยนี้นำเสนอการเปรียบเทียบแบบจำลองเพื่อทำนายผลการเรียนของนิสิตที่ใช้งานระบบการจัดการเรียนรู้ออนไลน์ โดยใช้การเรียนรู้ของเครื่อง ผู้วิจัยใช้ข้อมูลการเรียนผ่านทางระบบจัดการเรียนรู้ออนไลน์ของนิสิตมาวิเคราะห์เบื้องต้น เพื่อเปรียบเทียบแบบจำลองการทำนายผลการเรียน และศึกษาการใช้เทคนิคการเลือกคุณสมบัติ เพื่อเปรียบเทียบว่า แบบจำลองที่ใช้คุณลักษณะทั้งหมด และแบบจำลองที่ทำการการคัดเลือกคุณลักษณะ มีประสิทธิภาพแตกต่างกันเช่นไร โดยผู้วิจัยแบ่งกลุ่มของคุณลักษณะออกเป็น 2 ประเภท ได้แก่ คุณลักษณะที่สร้างจากพฤติกรรมการใช้งานระบบและคุณลักษณะจากคะแนนเก็บระหว่างภาคการศึกษาของนิสิต จากนั้นผู้วิจัยทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพในการทำนายของแบบจำลอง เพื่อวิเคราะห์ประสิทธิภาพของคุณลักษณะทั้งสองประเภทในการทำนายผลการเรียน โดยผู้วิจัยได้เปรียบเทียบและประเมินประสิทธิภาพของแบบจำลองของเครื่องที่ได้รับความนิยมในปัจจุบันหลากหลายแบบจำลอง ได้แก่ Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Support Vector Machines, Decision Tree, Random Forest และ XGBoost แบบจำลองเหล่านี้ร่วมกับการใช้เทคนิคการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มตัวอย่างเพิ่มกลุ่มน้อยด้วยการสร้างตัวอย่างสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) รวมทั้งการคัดเลือกคุณลักษณะ (Feature Selection) จากการทดลองพบว่าแบบจำลอง XGBoost ให้ประสิทธิภาพในการทำนายสูงสุด มีความแม่นยำถึง 83.95% ส่วนในกลุ่มของคุณลักษณะนั้น พบว่าคุณลักษณะด้านคะแนน ให้ผลการทำนายที่ดีกว่าการใช้คุณลักษณะด้านพฤติกรรมการใช้งานระบบ แต่การใช้คุณลักษณะทั้งสองด้านร่วมกัน ให้ผลการทำนายดีที่สุด ในส่วนของการคัดเลือกคุณลักษณะ พบว่าการคัดเลือกคุณลักษณะช่วยเพิ่มประสิทธิภาพในแบบจำลอง

ที่มีความซับซ้อนต่ำ แต่ในแบบจำลองที่มีความซับซ้อนสูง การคัดเลือกคุณลักษณะไม่ได้ช่วยเพิ่มประสิทธิภาพการทำนาย

คำสำคัญ: ระบบการจัดการเรียนรู้ออนไลน์ การทำนาย ผลการเรียน การเรียนรู้ของเครื่อง

ABSTRACT

This study presents the comparison of machine learning models for predicting student learning outcome in an online learning management system. Based on student activities stored in the log file of the system, we identify and create relevant features for predicting student performance. In this study, we divide the features into two groups, i.e., features related to the system usage behavior of the students and features related to the score the student obtained. Then, we analyze the importance of these two sets of features by creating models with different sets of the features. We compare the prediction performances of several well-known machine learning models including Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Support Vector Machines, Decision Tree, Random Forest and XGBoost. Several preprocessing methods are employed to improve the performance prediction including handling imbalanced data with Synthetic Minority Oversampling Technique (SMOTE) and choosing relevant features by XGBoost model. We found that XGBoost model yields highest prediction accuracy at 83.95%. The models trained by features related to score alone yield better performance than those trained by features related to the system usage behavior, but using the features from both groups together yields the best performance. Moreover, we found that feature selection can improve the performance of low-complexity models, while impair the performance of high-complexity models.

Keywords: Online Learning Management System, Prediction, Learning Outcome, Machine Learning

บทนำ

ปัจจุบันสถาบันการศึกษาหลายแห่งได้นำระบบการจัดการเรียนรู้ (Learning Management System: LMS) มาใช้ในการจัดการเรียนการสอน โดยระบบดังกล่าวมีเทคโนโลยีที่ช่วยอำนวยความสะดวกในการจัดการสอนให้แก่ผู้สอน และช่วยสนับสนุนการเรียนรู้ของผู้เรียนได้เป็นอย่างดี สร้างบรรยากาศเหมือนห้องเรียนเสมือน (Virtual Classroom) ที่ผู้สอนและผู้เรียนสามารถมีปฏิสัมพันธ์กันได้ ระบบการจัดการเรียนรู้นี้โดยทั่วไปจะสามารถบันทึกข้อมูลการเข้าใช้ระบบของผู้เรียน รวมทั้งเป็นแพลตฟอร์มสำหรับให้ผู้สอนสร้างหัวข้อการสอน การทดสอบ รวมทั้งสร้างช่องทางในการตอบคำถามของผู้เรียนอีกด้วย ระบบการจัดการเรียนการสอนแบบออนไลน์นี้กลายเป็นสิ่งจำเป็นอย่างยิ่ง ภายใต้สภาวะการแพร่ระบาดของโรคโควิด 19 ซึ่งทำให้สถาบันการศึกษาไม่สามารถจัดการเรียนการสอนในชั้นเรียนแบบเดิมได้ หากแต่จำเป็นต้องจัดการเรียนการสอนแบบออนไลน์เต็มรูปแบบภายใต้การสอนแบบออนไลน์ ผู้เรียนและผู้สอนจะมีปฏิสัมพันธ์กันน้อยลง รวมทั้งผู้สอนไม่อาจเห็นพฤติกรรมของผู้เรียน เป็นรายบุคคลได้ (Qiu *et al.*, 2022) เหมือนการเรียนในห้องเรียน นำไปสู่การติดตามความก้าวหน้าการเรียนรู้ของผู้เรียนโดยผู้สอนได้ยาก และผลการเรียนของผู้เรียนมีประสิทธิภาพน้อยลง ในการนี้ จึงมีการศึกษาจำนวนมาก ที่นำเอาเทคนิคการเก็บข้อมูลและการทำนายประสิทธิภาพการเรียนรู้ของผู้เรียนมาใช้เพื่อดูแลและเตือนการล่วงหน้า (Early Warning System) เพื่อให้ผู้สอนสามารถปรับปรุงนโยบายการสอนให้มีประสิทธิภาพและทันกาลมากขึ้น (Gasevic *et al.*, 2017)

ในการศึกษานี้ ผู้วิจัยมุ่งจะสร้างแบบจำลอง เพื่อทำนายผลการเรียนของนิสิตในเชิงรุก นั่นคือการทำนายผลการเรียนก่อนจะสิ้นภาคการศึกษา โดยอาศัยพฤติกรรมการเรียนของนิสิตแต่ละรายบุคคลที่มีบันทึกอยู่ในระบบการจัดการเรียนรู้ ร่วมกับคะแนนในบางส่วนของนิสิตในระหว่างภาคการศึกษา โดยผู้วิจัยใช้วิธีการทำเหมืองข้อมูลเพื่อการศึกษา (Educational Data Mining) มาวิเคราะห์ข้อมูลบันทึกการใช้งานกิจกรรมที่หลากหลายของผู้ใช้งาน รวมทั้งสถิติการเข้าใช้งานระบบ คะแนนงานมอบหมาย คะแนนแบบทดสอบ เป็นต้น โดยเทคนิคหนึ่งที่กล่าวถึงกันมากในการทำเหมืองข้อมูล เรียกว่า “การเรียนรู้ของเครื่อง (Machine Learning)” ซึ่งมีนักวิจัยหลายท่านได้ศึกษาและเผยแพร่หัวข้อการทำเหมืองข้อมูลเพื่อการศึกษาโดยใช้การเรียนรู้ของเครื่อง ได้แก่ การทำนายผลการเรียนของผู้เรียนจากพฤติกรรมการส่งงานมอบหมายของผู้เรียนในระบบการจัดการเรียนรู้ (Hooshyar and Pedaste, 2019) การทำนายการลาออกกลางคันของผู้เรียนจากข้อมูลทั่วไปและผลการเรียนของผู้เรียน (Tan and Shao, 2015) และการทำนายผลการเรียนเฉลี่ยสะสม (GPA) จากผลการเรียนของรายวิชาและข้อมูลทั่วไป (Almasri *et al.*, 2020)

วัตถุประสงค์การวิจัย

1. เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง และเลือกแบบจำลองที่ให้ความแม่นยำในการทำนายผลการเรียนของผู้เรียนสูงสุดในชุดข้อมูลที่ใช้ศึกษา รวมทั้งศึกษาประสิทธิภาพของการใช้เทคนิคการคัดเลือกคุณลักษณะว่า จะช่วยปรับปรุงประสิทธิภาพการทำนายหรือไม่

2. เพื่อศึกษาเปรียบเทียบประสิทธิภาพในการทำนายของคุณลักษณะ 2 กลุ่ม ได้แก่ คุณลักษณะ กลุ่มของคะแนน และคุณลักษณะกลุ่มของพฤติกรรมการเข้าใช้ระบบว่า จะมีประสิทธิภาพในการทำนายแตกต่างอย่างไร

เอกสารและงานวิจัยที่เกี่ยวข้อง

1. การทำเหมืองข้อมูล

การทำเหมืองข้อมูล เปรียบได้กับการขุดเจาะเข้าไปในข้อมูลลงสู่ระดับลึก เพื่อค้นหาสิ่งที่มีคุณค่า การทำเหมืองข้อมูลเป็นเทคนิคหนึ่งที่กล่าวถึงมากในการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data) ซึ่งเป็นสิ่งที่สำคัญที่จะทำให้องค์กรได้รับประโยชน์จากข้อมูลมหาศาล สามารถนำไปใช้ประโยชน์ในหลายด้าน เช่น ด้านธุรกิจ, ด้านการศึกษา, ด้านระบบการจราจรและการขนส่ง, ด้านสาธารณสุข (Chen *et al*., 2013) การทำเหมืองข้อมูล มีหนึ่งในกระบวนการตอบโจทย์ธุรกิจที่สำคัญ เรียกว่า “Cross-Industry Standard Process for Data Mining (CRISP-DM)” ประกอบด้วย 6 ขั้นตอน ได้แก่ การทำความเข้าใจในธุรกิจ, การทำความเข้าใจข้อมูล, การเตรียมข้อมูล, การสร้างแบบจำลอง, การวัดประสิทธิภาพของแบบจำลอง และการนำไปใช้จริง (Awlla, 2019)

2. การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง เป็นเทคนิคหนึ่งในการทำเหมืองข้อมูล และเป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence) โดยมุ่งเน้นศึกษาการทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ได้ โดยไม่ต้องเขียนโปรแกรมกำกับไว้อย่างชัดเจน กระบวนการเรียนรู้ของเครื่องมีการแบ่งชุดข้อมูลออกเป็น 2 ส่วน คือ ชุดข้อมูลการเรียนรู้ (Training Set) เป็นข้อมูลที่ถูกนำไปเรียนรู้ให้กับเครื่องและชุดข้อมูลทดสอบ (Test Set) เป็นข้อมูลที่ใช้สำหรับการประเมินประสิทธิภาพการทำนาย การเรียนรู้ของเครื่องนั้นมีการเรียนรู้หลายแบบที่ขึ้นอยู่กับลักษณะของข้อมูล โดยการเรียนรู้ของเครื่อง ที่เราใช้งานนี้ เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) แบบการจำแนกประเภทของข้อมูล (Classification) (Awad and Khanna, 2015)

3. เทคนิคการเพิ่มประสิทธิภาพการทำนายของแบบจำลอง

เทคนิคในการปรับปรุงผลลัพธ์การทำนายให้ดีขึ้นสำหรับการวิจัยนี้ 3 แบบ ประกอบด้วย 1) วิศวกรรมคุณลักษณะ (Feature Engineering) เป็นกระบวนที่ใช้ความรู้เฉพาะทางในการสร้างคุณลักษณะที่จะทำความสะอาด และแปลงข้อมูลดิบในรูปแบบที่สามารถเรียนรู้ด้วยแบบจำลอง (Mohamad *et al*., 2020) 2) การจัดการความไม่สมดุลของข้อมูล (Handling Imbalanced Data) ด้วยการสุ่มตัวอย่างเพิ่มกลุ่มน้อยด้วยการสร้างตัวอย่างสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) เป็นการสุ่มเพิ่มจำนวนข้อมูลในคลาสที่มีข้อมูลอยู่น้อยให้มากขึ้น โดยสร้างจุดเทียมระหว่างจุดที่สุ่มในคลาสที่มีอยู่น้อยและเพื่อนบ้านที่อยู่ใกล้เคียงจุดที่สุ่มนี้ เพื่อให้ข้อมูลแต่ละคลาสมีความสมดุลมากขึ้น (Sun *et al*., 2018) 3) การคัดเลือกคุณลักษณะ (Feature Selection) ด้วยวิธี Wrapped method โดยเลือกกลุ่มคุณลักษณะย่อยจากคุณลักษณะทั้งหมดไปเรียนรู้กับชุดข้อมูลการเรียนรู้ด้วยแบบจำลองที่สนใจ (งานวิจัยนี้เลือกใช้

XGBoost) แล้วนำกลุ่มคุณลักษณะ ที่ให้ผลลัพธ์ดีที่สุดนั้นไปเรียนรู้ด้วยแบบจำลองต่าง ๆ ที่นำไปใช้จริง (Jalota and Agrawal, 2021)

4. การสร้างแบบจำลองการจำแนกประเภท

แบบจำลองการจำแนกประเภทที่จะสร้างในการวิจัยนี้จำนวน 7 แบบ ซึ่งมีการใช้งานอย่างแพร่หลายในการนำมาทำนายผลการเรียนของนิสิต ประกอบด้วย

1) Logistic Regression อาศัยหลักการความน่าจะเป็นของผลลัพธ์การทำนายว่าจะเกิดขึ้นหรือไม่ โดยขึ้นอยู่กับคุณลักษณะ สมการของ Logistic Regression เป็นฟังก์ชัน Sigmoid ที่มีลักษณะเป็นเส้นโค้งรูป S โดยมีค่าระหว่าง 0 ถึง 1 ซึ่งแสดงสมการ (1) แบบจำลองนี้ไม่ได้ใช้พลังการประมวลผลในการคำนวณข้อมูลสูง ตีความได้ง่าย แต่อาจจะเกิด Overfitting (Awad and Khanna, 2015)

$$P(class) = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \qquad (1)$$

โดยที่ X_i คือ คุณลักษณะตัวที่ i เมื่อ $i = 1, 2, 3, \ldots, N$

β_0 คือ ค่าไบแอส *(Bias)*

β_i คือ ค่าสัมประสิทธิ์ของ X_i เมื่อ $i = 1, 2, 3, \ldots, N$

e คือ ค่าคงที่ทางคณิตศาสตร์ที่เป็นฐานของลอการิทึมธรรมชาติ มีค่าโดยประมาณ *2.71828*

2) Naïve Bayes อาศัยหลักการความน่าจะเป็นตามทฤษฎีของเบย์ที่บ่งบอกถึงความสัมพันธ์ระหว่างสิ่งที่กำลังเกิดขึ้นและสิ่งที่เกิดขึ้นแล้ว เพื่อทำนายคลาสจากคุณลักษณะที่เป็นอิสระต่อกันโดยกำหนดตัวแปรต่าง ๆ และสามารถเขียนดังสมการ (2) แบบจำลองนี้ทำนายข้อมูลได้ง่ายและเร็ว สามารถใช้ในการจำแนกคลาสจำนวนมาก แต่หากไม่พบข้อมูลทดสอบในชุดข้อมูลการเรียนรู้ จะได้จำนวนข้อมูลที่เกิดขึ้นเป็นศูนย์ ทำให้ไม่มีค่าความน่าจะเป็น จึงใช้วิธีการประมาณค่า Laplace แทน (Amra and Maghari, 2017)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (2)$$

โดยกำหนดตัวแปร ดังนี้

$P(c|x)$ เป็นความน่าจะเป็นที่ได้คลาสใดคลาสหนึ่ง เมื่อใช้คุณลักษณะแล้ว เรียกว่า *“Posterior Probability”*

$P(c)$ เป็นความน่าจะเป็นที่ได้คลาสใดคลาสหนึ่ง เรียกว่า *“Prior Probability”*

$P(x|c)$ เป็นความน่าจะเป็นที่ได้คุณลักษณะ เมื่อเลือกคลาสใดคลาสหนึ่งแล้ว เรียกว่า *“Likelihood”*

$P(x)$ เป็นความน่าจะเป็นที่ได้คุณลักษณะ เรียกว่า *“Prior Probability”*

3) *Support Vector Machine* ใช้สมการเส้นแบ่งคลาสออกจากกัน โดยให้มีระยะขอบของ 2 คลาสให้มากที่สุด สามารถจำแนกข้อมูลที่แบ่งแยกออกจากกันได้ทั้งแบบเชิงเส้นและแบบไม่เชิงเส้น แต่อาจจะไม่ได้มีประสิทธิภาพที่ดีที่สุดสำหรับการวิเคราะห์ข้อความ (*Awad and Khanna,* 2015) การคำนวณค่าของเส้นของทั้ง 2 คลาส ในการจำแนกประเภทข้อมูลนั้น โดยใช้สมการที่กำหนดตัวแปรต่าง ๆ ได้แก่ w^T คือ ค่าน้ำหนัก, x_i คือ คุณลักษณะของรายการที่ i, b คือ ค่าไบแอส *(Bias)* และ $i = 1, 2, \ldots, N$ เป็นของรายการที่ i ของชุดข้อมูล ดังสมการ (3)-(5)

1. สมการ Hyperplane คือ $w^T x_i + b = 0$ (3)
2. หาก $y_i = class\ 1$ จะได้ว่า $w^T x_i + b \geq 1$ (4)
3. หาก $y_i = class\ 2$ จะได้ว่า $w^T x_i + b \leq -1$ (5)

4) K-Nearest Neighbor คำนวณระยะทางแบบยุคลิดระหว่างจุดใหม่และจุดที่ใกล้เคียงที่มีคลาสอยู่แล้วจำนวน K จุด โดยพิจารณาระยะทางที่น้อยที่สุดในการเลือกคลาสให้กับจุดใหม่ เป็นวิธีการที่เข้าใจง่าย เรียนรู้ข้อมูลได้เร็ว มีความทนทานต่อข้อมูลรบกวนได้ดี แต่ค่อนข้างประมวลผลช้า (Amra and Maghari, 2017)

5) Decision Tree ใช้แผนผังต้นไม้กลับหัวช่วยในการตัดสินใจ แบบจำลองนี้ตีความได้ง่าย เนื่องจากต้นไม้ตัดสินใจเป็นการเลียนแบบการตัดสินใจของมนุษย์ ทำงานกับคุณลักษณะได้ทุกชนิด สามารถจัดการข้อมูลสูญหาย แต่ค่อนข้างไม่ยืดหยุ่น หากเปลี่ยนแปลงข้อมูล ทำให้แผนภาพ Decision Tree เปลี่ยนได้ อาจทำให้เกิด Overfitting กับข้อมูลได้ง่าย ทำให้ส่งผลต่อการทำนายข้อมูลได้ไม่ค่อยดีนัก (Patel and Prajapati, 2018)

6) Random Forest นำ Decision Tree หลายต้นที่ไม่ซ้ำกันมาทำงานร่วมกันแล้วเลือกคลาสส่วนใหญ่ที่ได้จากการทำนายของแต่ละต้น สามารถแก้ปัญหาความไม่ยืดหยุ่นของต้นไม้ตัดสินใจและการเกิด Overfitting ง่าย (Jayaprakash *et al.*, 2020)

7) XGBoost ย่อมาจาก Extreme Gradient Boosting เรียนรู้ความผิดพลาดการทำนายของแบบจำลอง Decision Tree หลายแบบจำลอง การทำงานของ XGBoost เหล่านี้ เพื่อช่วยลดการทำนายข้อมูลผิดพลาด โดยเรียนรู้จากแบบจำลองก่อนหน้านี้เรื่อย ๆ เป็นแบบอนุกรมทำให้ช่วยเพิ่มประสิทธิภาพการทำนายในด้านความแม่นยำ ความไว และความจำเพาะของข้อมูลการทดสอบ (Bishop, 2006)

5. งานวิจัยที่เกี่ยวข้อง

มีงานวิจัยมากมายที่มุ่งเป้าพัฒนาแบบจำลอง เพื่อทำนายผลการเรียนของผู้เรียน โดยอาศัยเทคนิคการเรียนรู้ของเครื่อง เช่น ข้อมูลผลการเรียนจะมีลักษณะไม่สมดุล เนื่องจากผู้เรียนที่ได้คะแนนสูงมากและต่ำมาก จะมีจำนวนไม่มากนัก ซึ่งในบทความของ Ashfaq *et al.* (2020) ได้ศึกษาการสร้างแบบจำลองเพื่อจัดการกับข้อมูลไม่สมดุล โดยใช้เทคนิค SMOTE ส่วน Amrieh *et al.* (2016) ได้ใช้แบบจำลองแบบ Ensemble เพื่อช่วยให้เพิ่มประสิทธิภาพการทำงานของแบบจำลอง โดยแบบจำลอง Decision Tree ที่ร่วมกับการทำ Boosting มีประสิทธิภาพดีที่สุด โดยมีความแม่นยำถึง 85%

บทความของ Ramaswami *et al.* (2020) กับ Shrestha and Pokharel (2021) มีการคัดเลือกคุณลักษณะ เพื่อเพิ่มประสิทธิภาพการทำงานของแบบจำลอง พบว่าแบบจำลองที่ได้รับการคัดเลือกคุณลักษณะ มีประสิทธิภาพสูงกว่าแบบจำลองที่ไม่ผ่านการคัดเลือกคุณลักษณะ

ส่วน Riestra-González *et al.* (2021) ได้ใช้เทคนิคการจัดกลุ่ม (Clustering) ร่วมด้วย เพื่อทำให้ผลการทำนายดีขึ้น และร่วมทำการคัดเลือกคุณลักษณะวิธี Recursive Feature Elimination and Cross-Validation Selection (RFECV) ได้ผลว่าแบบจำลอง Multiple perceptron มีประสิทธิภาพการทำนายดีที่สุด มีความแม่นยำเท่ากับ 80.1%

นอกจากนี้ยังมีบทความที่ทดลองข้อมูลการเรียน 2 กลุ่ม ได้แก่ ข้อมูล 10 สัปดาห์ และข้อมูล 6 สัปดาห์ โดย Quinn and Gray (2019) พบว่า ความแม่นยำการทำนายระดับผลการเรียนมากที่สุด 60.5% ด้วยแบบจำลอง Random Forest และความแม่นยำการทำนายผลการเรียนของผู้เรียนผ่านหรือไม่ผ่าน โดยมีค่ามากที่สุด 82.18% ด้วยแบบจำลอง Random Forest เช่นเดียวกัน โดยใช้ข้อมูล 10 สัปดาห์

จากงานวิจัยที่เกี่ยวข้องเหล่านี้ ผู้วิจัยได้ทำการสรุปแนวทางการพัฒนาแบบจำลองของการศึกษาครั้งนี้ โดยใช้เทคนิค SMOTE เพื่อปรับจำนวนข้อมูลให้มีความสมดุล และทดลองใช้แบบจำลองประเภทต่าง ๆ ที่บทความที่เกี่ยวข้องเลือกมาใช้งาน เพื่อเปรียบเทียบประสิทธิภาพและเลือกแบบจำลองที่เหมาะสมกับการศึกษานี้ นอกจากนี้จากการที่ผู้วิจัยในบางบทความ รายงานว่าการคัดเลือกคุณลักษณะมีผลต่อประสิทธิภาพของการทำนาย ผู้วิจัยจึงทดลองเปรียบเทียบประสิทธิภาพแบบจำลองที่ทำการคัดเลือกคุณลักษณะร่วมด้วย กับประสิทธิภาพของแบบจำลองที่ใช้คุณลักษณะครบถ้วน

อีกทั้ง เพื่อให้เข้าใจประสิทธิภาพการทำนายของคุณลักษณะได้ชัดเจนขึ้น ผู้วิจัยจึงทำการเปรียบเทียบประสิทธิภาพของคุณลักษณะ โดยจัดคุณลักษณะออกเป็น 2 กลุ่ม ได้แก่ คุณลักษณะที่เกี่ยวข้องกับคะแนน และคุณลักษณะที่เกี่ยวข้องกับพฤติกรรมการใช้งานระบบ เพื่อศึกษาผลต่อประสิทธิภาพการทำนายของคุณลักษณะทั้งสองกลุ่มนี้

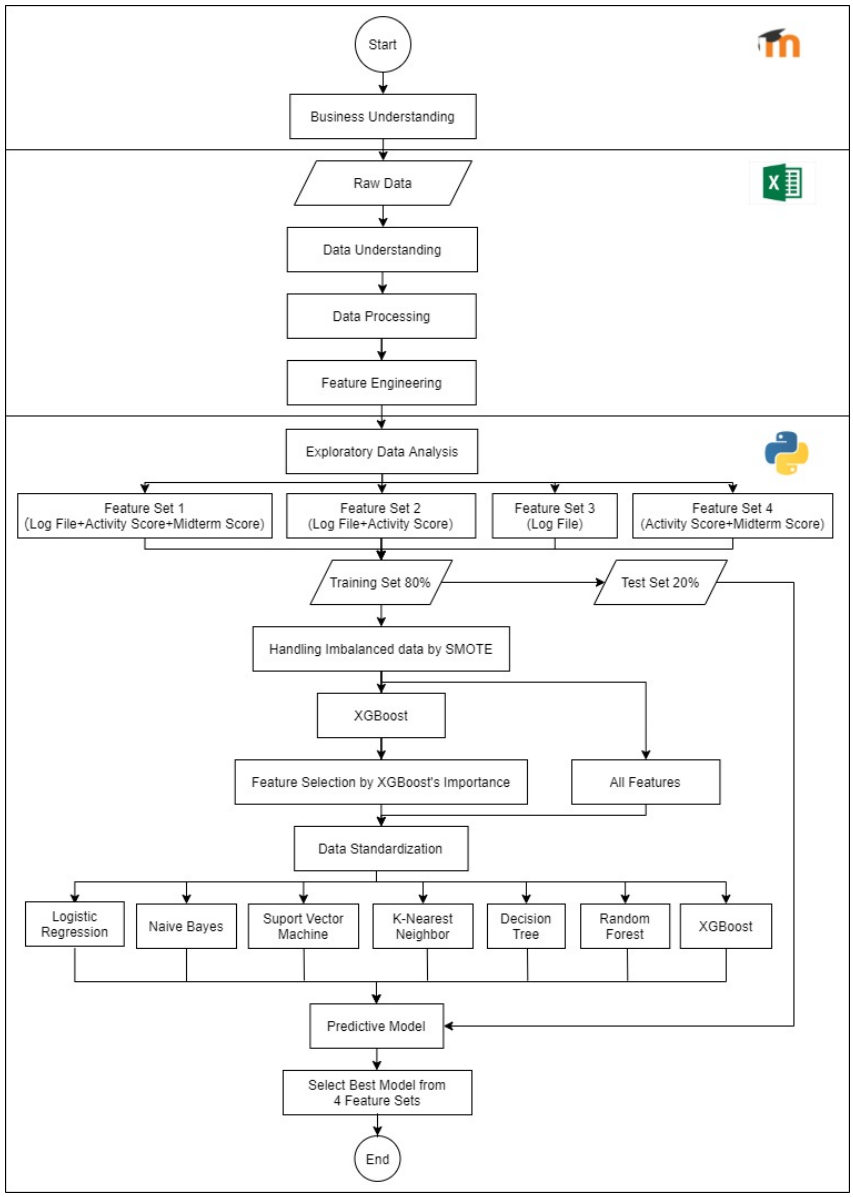
วิธีการดำเนินงานวิจัย

กลุ่มตัวอย่างที่ใช้ในการวิจัย

กลุ่มตัวอย่างที่ใช้ในการวิจัยจำนวน 405 คน มาจากจำนวนห้องเรียน 3 ห้อง โดยกลุ่มตัวอย่างที่ใช้ในการวิจัย เป็นกลุ่มของนิสิตในอาจารย์ผู้สอน 2 ท่าน ที่ได้ให้ความอนุเคราะห์ข้อมูลต่อการวิจัยชุดนี้

ภาพรวมของกระบวนการทำงาน

ผู้วิจัยได้แสดงภาพรวมกระบวนการดำเนินงานวิจัยในการทำนายผลการเรียนของนิสิตที่ใช้ระบบการจัดการเรียนรู้ออนไลน์ โดยใช้การเรียนรู้ของเครื่อง ดังภาพประกอบ 1 จากข้อมูลการเข้าใช้งานและผลคะแนนที่มีการบันทึกในระบบ LMS ซึ่งใช้ Moodle เป็นแพลตฟอร์มการเรียนรู้ออนไลน์ ผู้วิจัยใช้โปรแกรม Microsoft Excel เพื่อพิจารณาข้อมูลเบื้องต้นและสร้างข้อมูลตารางของคุณลักษณะ จากนั้นในส่วนการสร้างแบบจำลอง ผู้วิจัยเลือกใช้ภาษา Python โดยใช้ Scikit-learn เป็น Library หลักสำหรับสร้างแบบจำลองและทำวิศวกรรมคุณลักษณะ รวมทั้งการคัดเลือกคุณลักษณะ



ภาพประกอบ 2 แผนผังกระบวนการทำงาน

1. การทำความเข้าใจในธุรกิจ

ก่อนเปิดภาคการศึกษา สำนักนวัตกรรมการเรียนรู้ มศว ได้นำระบบการจัดการเรียนรู้ออนไลน์มาใช้การสนับสนุนการเรียนการสอน ซึ่งใช้ Moodle เป็นแพลตฟอร์มการเรียนรู้ออนไลน์ โดยผู้ดูแลระบบเปิดรายวิชา SWU 141 ชีวิตในโลกดิจิทัล และกำหนดเงื่อนไขของกิจกรรม

การเรียนการสอนที่แตกต่างกันในการตั้งค่าระบบ ดังนี้ 1) การชมสื่อการเรียนรู้ จำนวน 11 เรื่อง สามารถชมล่วงหน้าทุกเรื่องตลอดเวลา 2) การทำแบบทดสอบ จำนวน 6 เรื่อง สามารถทำซ้ำได้ 3 ครั้ง ภายในระยะเวลาที่กำหนด โดยระบบจัดเก็บคะแนนมากที่สุดเป็นคะแนนเก็บ 3) การส่งงานมอบหมาย จำนวน 2 เรื่อง สามารถส่งงานมอบหมายในระยะเวลาที่กำหนด และสำนักฯ มีการกำหนดเกณฑ์การประเมินผลการเรียนรู้ ดังตารางที่ 1

ตารางที่ 1 แสดงเกณฑ์การประเมินผลการเรียนรู้

คะแนนการประเมินผลการเรียนรู้	สัดส่วนการประเมิน
1. ครึ่งแรกของภาคการศึกษา	
1.1 ด้านคะแนนกิจกรรมในห้องเรียน	30%
1.2 ด้านคะแนนกลางภาค	30%
รวม	**60%**
2. ครึ่งหลังของภาคการศึกษา	
2.1 ด้านคะแนนโครงงาน	30%
2.2 ด้านคะแนนพฤติกรรมการมีส่วนร่วม	10%
รวม	**40%**
รวมทั้งหมด	**100%**

2. การเก็บรวบรวมข้อมูลและการเตรียมข้อมูล

การศึกษาวิจัยนี้เลือกใช้ข้อมูลบันทึกการใช้งานระบบ Moodle ของนิสิตระดับปริญญาตรี ชั้นปีที่ 1 มหาวิทยาลัยศรีนครินทรวิโรฒ จำนวน 405 คน ที่ลงทะเบียนรายวิชา SWU 141 ชีวิตในโลกดิจิทัล ภาคการศึกษาที่ 2 ปีการศึกษา 2563 ตั้งแต่วันที่ 7 มกราคม ถึง 5 มีนาคม 2564 ในช่วงการเรียนของครึ่งภาคการศึกษาเป็นระยะเวลา 9 สัปดาห์ ข้อมูลบันทึกการใช้งานระบบเหล่านี้ที่ดึงมาจากระบบ เป็นข้อมูลการเข้าถึงกิจกรรมการใช้งานที่หลากหลายของผู้ดูแลระบบ อาจารย์ และนิสิต เป็นรายบุคคลในแต่ละวันที่และเวลา จึงเลือกข้อมูลเฉพาะข้อมูลการใช้งานของนิสิต ตั้งแต่วันที่ 7 มกราคม ถึง 5 มีนาคม 2564 จำนวน 99,925 รายการ และปรับรูปแบบวันที่และเวลาการใช้งานให้แยกวันที่และเวลาออกจากกัน โดยใช้เครื่องมือต่าง ๆ ในโปรแกรม Microsoft Excel

3. การทำวิศวกรรมคุณลักษณะ

ในการศึกษานี้ เราสร้างคุณลักษณะ เพื่อสะท้อนพฤติกรรมของผู้เรียนในด้านต่าง ๆ ได้แก่ ด้านการเข้าใช้งานระบบ, ด้านการชมสื่อการเรียนรู้, ด้านการทำแบบทดสอบ และด้านการส่งงานมอบหมาย โดยคุณลักษณะที่สร้างขึ้นจะถูกนำมาใช้ในการทำนายผลการเรียน โดยได้แสดงรายละเอียดของตัวแปรที่ใช้ในการทำนายและตัวแปรที่ต้องการทำนายไว้ในตารางที่ 2 ทั้งนี้ เราจะไม่ใช้คะแนน

ในส่วนครึ่งหลังของภาคการศึกษามาใช้ในร่วมในการทำนาย ซึ่งประกอบด้วยคะแนนโครงงานและคะแนนพฤติกรรมการมีส่วนร่วม เนื่องจากผู้วิจัยต้องการสร้างแบบจำลอง เพื่อทำการทำนายผลการเรียนโดยใช้แต่ข้อมูลเฉพาะในช่วงต้นของภาคการศึกษา ซึ่งจะทำให้สามารถค้นหานิสิตที่ควรได้รับการเอาใจใส่ และดูแลได้แต่เนิ่น ๆ หากจะรอผลคะแนนในส่วนครึ่งหลังของภาคการศึกษาอาจทำให้การทำนายผลการเรียนได้ล่าช้า ส่งผลให้ผู้สอนมีเวลาไม่เพียงพอในการให้คำแนะนำและเอาใจใส่ต่อนิสิตที่ต้องการความช่วยเหลือ ทำให้นิสิตอาจไม่สามารถปรับปรุงการเรียนรู้ของตนเองได้ทันท่วงที อีกทั้งคะแนนในส่วนพฤติกรรมการมีส่วนร่วม จะเป็นส่วนที่ประเมินโดยผู้สอน จากพฤติกรรมการมีส่วนร่วมในการเข้าชั้นเรียนในตอนสิ้นภาคการศึกษา มิได้เป็นคะแนนที่สามารถประเมินได้จากการใช้งานระบบการเรียนออนไลน์ จึงมิได้นำมาใช้เป็นคุณลักษณะในที่นี้ ทั้งนี้เราจึงใช้คุณลักษณะในส่วนของคะแนนไม่เกิน 60% เท่านั้น

ตารางที่ 2 แสดงรายละเอียดคำอธิบายของชุดข้อมูล

ตัวแปร	คำอธิบายของตัวแปร
1. ตัวแปรที่ต้องการทำนาย (Target หรือ Class)	
- ผลการเรียน (1 คุณลักษณะ)	แบ่งผลการเรียนออกเป็น 3 กลุ่มคือ ผลการเรียนดีมาก ดี และ ต้องปรับปรุง
2. ตัวทำนาย (Predictor หรือ Feature)	
2.1 ข้อมูลทั่วไป	
- คณะที่นิสิตเรียน (1 คุณลักษณะ)	มนุษยศาสตร์, ศึกษาศาสตร์, สังคมศาสตร์
- เพศ (1 คุณลักษณะ)	ชาย, หญิง
2.2 ด้านพฤติกรรมการใช้งานระบบของนิสิต	
การเข้าใช้งานระบบ	
- จำนวนครั้งที่คลิกเข้าใช้งานระบบ (1 คุณลักษณะ)	จำนวนครั้งที่คลิกเข้าใช้งานระบบ
การชมสื่อการเรียนรู้	
- จำนวนครั้งที่คลิกเข้าถึงสื่อการเรียนรู้ (12 คุณลักษณะ)	จำนวนครั้งที่คลิกเข้าถึงสื่อการเรียนรู้แต่ละเรื่อง (11 เรื่อง) และรวมทุกเรื่อง
- จำนวนวันที่ชมสื่อการเรียนรู้ (12 คุณลักษณะ)	จำนวนวันที่ชมสื่อการเรียนรู้แต่ละเรื่อง (11 เรื่อง) และรวมทุกเรื่อง
- จำนวนวันระหว่างวันแรกที่เรียนและวันแรกที่ชมสื่อการเรียนรู้ (12 คุณลักษณะ)	จำนวนวันระหว่างวันแรกที่เรียนและวันแรกที่ชมสื่อการเรียนรู้แต่ละเรื่อง (11 เรื่อง) และรวมทุกเรื่อง
- จำนวนเรื่องที่ชมสื่อการเรียนรู้ (1 คุณลักษณะ)	จำนวน 1 ถึง 10 เรื่อง

ตารางที่ 3 (ต่อ)

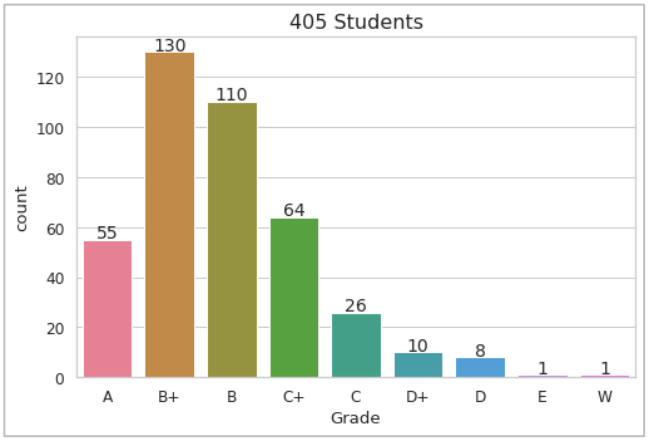
ตัวแปร	คำอธิบายของตัวแปร
การทำแบบทดสอบ	
- จำนวนครั้งที่คลิกเข้าถึงแบบทดสอบ (7 คุณลักษณะ)	จำนวนครั้งที่คลิกเข้าถึงแบบทดสอบแต่ละเรื่อง (6 เรื่อง) และรวมทุกเรื่อง
- จำนวนวันระหว่างวันที่ทำแบบทดสอบครั้งแรกและวันสุดท้ายการทำแบบทดสอบ (7 คุณลักษณะ)	จำนวนวันระหว่างวันที่ทำแบบทดสอบครั้งแรกและวันสุดท้ายการทำแบบทดสอบแต่ละเรื่อง (6 เรื่อง) และรวมทุกเรื่อง
- จำนวนครั้งที่ทำแบบทดสอบซ้ำ (7 คุณลักษณะ)	จำนวน 1 ถึง 3 ครั้ง (6 เรื่อง) และรวมทุกเรื่อง
- จำนวนเรื่องที่ทำแบบทดสอบ (1 คุณลักษณะ)	จำนวน 1 ถึง 6 เรื่อง
- ระยะเวลาการทำแบบทดสอบ ครั้งที่ 1 (6 คุณลักษณะ)	ระยะเวลาการทำแบบทดสอบแต่ละเรื่อง (นาที) (6 เรื่อง)
- ระยะเวลาการทำแบบทดสอบ ครั้งที่ 2 (6 คุณลักษณะ)	ระยะเวลาการทำแบบทดสอบแต่ละเรื่อง (นาที) (6 เรื่อง)
- ระยะเวลาการทำแบบทดสอบ ครั้งที่ 3 (6 คุณลักษณะ)	ระยะเวลาการทำแบบทดสอบแต่ละเรื่อง (นาที) (6 เรื่อง)
- ระยะเวลารวมการทำแบบทดสอบรวมทุกครั้ง (7 คุณลักษณะ)	ระยะเวลารวมการทำแบบทดสอบรวมทุกครั้งแต่ละเรื่อง (นาที) (6 เรื่อง) และรวมทุกเรื่อง
- ระยะเวลาเฉลี่ยการทำแบบทดสอบรวมทุกครั้ง (7 คุณลักษณะ)	ระยะเวลาเฉลี่ยการทำแบบทดสอบรวมทุกครั้งแต่ละเรื่อง (นาที) (6 เรื่อง) และเฉลี่ยทุกเรื่อง
การส่งงานมอบหมาย	
- จำนวนครั้งที่คลิกเข้าถึงงานมอบหมาย (4 คุณลักษณะ)	จำนวนครั้งที่คลิกเข้าถึงงานมอบหมายแต่ละเรื่อง (3 เรื่อง) และรวมทุกเรื่อง
- จำนวนวันระหว่างวันแรกที่ส่งงานมอบหมายและวันสุดท้ายการส่งงานอบหมาย (1 คุณลักษณะ)	จำนวนวันระหว่างวันแรกที่ส่งงานมอบหมายและวันสุดท้ายการส่งงานอบหมาย (1 เรื่อง)

ตารางที่ 4 (ต่อ)

ตัวแปร	คำอธิบายของตัวแปร
2.3 ด้านคะแนนการประเมินผลการเรียนรู้	
คะแนนกิจกรรมในห้องเรียน (30%)	
- คะแนนกิจกรรมในห้องเรียนออนไลน์ (4 คุณลักษณะ)	คะแนนกิจกรรมในห้องเรียนออนไลน์แต่ละเรื่อง (3 เรื่อง) และรวมทุกเรื่อง
- คะแนนการเข้าเรียน (1 คุณลักษณะ)	คะแนนเฉลี่ยการเข้าเรียน
- คะแนนแบบทดสอบ (7 คุณลักษณะ)	คะแนนแบบทดสอบแต่ละเรื่อง (6 เรื่อง) และรวมทุกเรื่อง
คะแนนกลางภาค (30%)	
- คะแนนสอบกลางภาค (1 คุณลักษณะ)	คะแนนสอบกลางภาค
- คะแนนงานมอบหมาย (3 คุณลักษณะ)	คะแนนงานมอบหมาย (2 เรื่อง) และรวมทุกเรื่อง

4. การสำรวจข้อมูล

จากการสำรวจข้อมูล โดยใช้ภาษา Python พบว่าไม่มีข้อมูลที่ขาดหายไปในคุณลักษณะใด ๆ และนำคุณลักษณะเหล่านั้นไปศึกษาความสัมพันธ์กับระดับผลการเรียนด้วยวิธีการกระจายของข้อมูลแบบฮิสโตแกรม (Histogram) ที่แสดงความถี่ของข้อมูล ดังแสดงภาพประกอบ 2

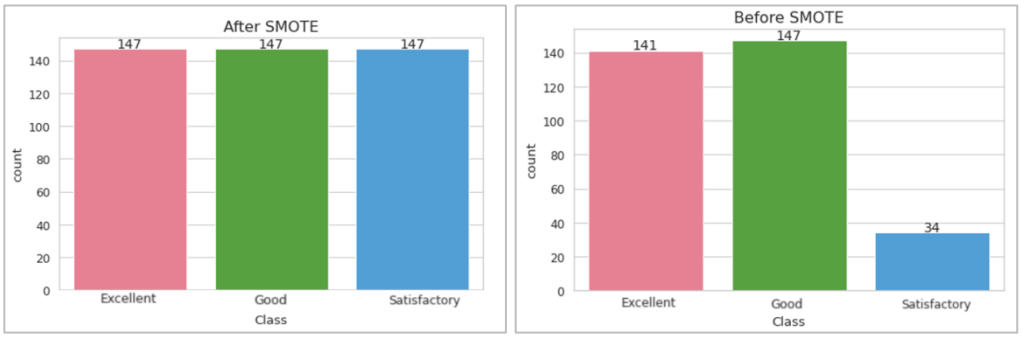


ภาพประกอบ 2 จำนวนนิสิตจำแนกเกรด

จากภาพประกอบ 2 พบว่าจำนวนนิสิตที่ได้เกรด E 1 คน และเกรด W 1 คน ดังนั้น คัดข้อมูลของนิสิตที่ได้เกรด E และ W ออก จึงเหลือจำนวนนิสิต 403 คน ดังภาพประกอบ 2 ทั้งนี้ ตัวแปรที่ต้องการทำนายจะเป็นผลการเรียนดีมาก ดี และต้องปรับปรุง ซึ่งได้จากการแปลงผลเกรดที่ได้รับ โดยให้นิสิตที่ได้รับเกรดเป็น A, B+ ถือเป็นผลการเรียนดีมากมีจำนวน 185 คน เกรด B, C+ เป็นผลการเรียนดี มีจำนวน 174 คน และ D+ ลงไป เป็นผลการเรียนต้องปรับปรุงมีจำนวน 44 คน

5. การจัดการความไม่สมดุลของข้อมูล

ผู้วิจัยทำการแบ่งชุดข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลสำหรับใช้ฝึกสอนแบบจำลอง (Training set) และข้อมูลสำหรับทดสอบ (Test set) ในอัตราส่วน 80 ต่อ 20 โดยมีข้อมูลสำหรับฝึกสอนแบบจำลอง จำนวน 322 คน และข้อมูลการทดสอบ จำนวน 81 คน เนื่องจากในข้อมูลสำหรับฝึกสอนแบบจำลอง ประกอบด้วยนิสิตได้ผลการเรียนดีมาก 141 คน, ผลการเรียนดี 147 คน และผลการเรียนต้องปรับปรุง 34 คน มีลักษณะของข้อมูลที่ไม่สมดุล จึงทำการปรับความสมดุลของข้อมูล โดยใช้เทคนิค SMOTE โดยปรับจำนวนข้อมูลในส่วนนิสิตที่มีผลการเรียนระดับดีมาก และระดับต้องปรับปรุง ให้มีความสมดุล โดยใช้วิธีสุ่มเพิ่ม ทำให้ข้อมูลทั้งสามกลุ่มมีจำนวนเท่ากัน คือ 147 คน ดังแสดงภาพประกอบ 2



ภาพประกอบ 2 แสดงจำนวนนิสิตแต่ละระดับผลการเรียน ก่อนและหลังการทำ SMOTE

6. การสร้างแบบจำลองการจำแนกประเภท

หลังจากได้ข้อมูลสำหรับฝึกสอนแบบจำลองที่มีความสมดุลด้วยการทำ SMOTE โดยใช้ SMOTE(k_neighbors=13, random_state=42) แล้ว เราจะนำข้อมูลชุดนี้มาฝึกสอนแบบจำลอง โดยในการศึกษานี้ เราทดลองใช้แบบจำลองที่ได้รับความนิยมและมีประสิทธิภาพสูงในการเรียนรู้ของเครื่องทั้งสิ้น 7 แบบจำลอง ซึ่งมีการกำหนดพารามิเตอร์ที่เหมาะสม ได้แก่ Logistic Regression โดยใช้ LogisticRegression(random_state=42), Naïve Bayes โดยใช้ GaussianNB(), K-Nearest Neighbor โดยใช้ KNeighborsClassifier(), Support Vector Machines โดยใช้ SVC(kernel='linear'), Decision Tree โดยใช้ DecisionTreeClassifier(random_state=42), Random Forest โดยใช้ RandomForestClassifier

(random_state=42) และ XGBoost โดยใช้ XGBClassifier(randomm_state=42) ก่อนเข้าสู่แบบจำลอง เหล่านี้มีการปรับขนาดของข้อมูลโดยใช้ StandardScaler() ในการศึกษานี้ผู้วิจัยสร้างแบบจำลองขึ้น 4 ชุด เพื่อศึกษาและเปรียบเทียบประสิทธิภาพการทำนายของคุณลักษณะ 2 กลุ่ม ได้แก่ คุณลักษณะกลุ่มคะแนน และคุณลักษณะกลุ่มพฤติกรรมการเข้าใช้งานระบบ โดยแบบจำลองชุดที่ 1 จะใช้ข้อมูลคุณลักษณะครบทั้งสองกลุ่ม ซึ่งจะประกอบด้วยคุณลักษณะทั้งสิ้น 115 คุณลักษณะ ส่วนในแบบจำลองชุดที่ 2 จะไม่นำคะแนนสอบกลางภาคมาคิดด้วย โดยนำมาเฉพาะคะแนนส่วนของกิจกรรมในห้องเรียน ทำให้เหลือเพียง 111 คุณลักษณะ แบบจำลองชุดที่ 3 จะไม่นำคุณลักษณะกลุ่มคะแนนมาใช้ อาศัยแต่เพียงคุณลักษณะของพฤติกรรมการใช้งานระบบของนิสิตเท่านั้น มีจำนวนทั้งสิ้น 99 คุณลักษณะ ในส่วนแบบจำลอง 4 จะให้แต่เพียงคุณลักษณะกลุ่มคะแนน ไม่ใช้คุณลักษณะกลุ่มพฤติกรรมการใช้งานระบบของนิสิต มีจำนวนทั้งสิ้น 18 คุณลักษณะ ดังแสดงตารางที่ 3

ตารางที่ 3 รายละเอียดการใช้คุณลักษณะสำหรับทำนาย แบบจำลองทั้ง 4 ชุด

คุณลักษณะที่ใช้สำหรับทำนาย	**แบบจำลอง**			
	ชุดที่ 1 115 คุณลักษณะ	**ชุดที่ 2 111 คุณลักษณะ**	**ชุดที่ 3 99 คุณลักษณะ**	**ชุดที่ 4 18 คุณลักษณะ**
ข้อมูลทั่วไป				
- คณะที่นิสิตเรียน (1 คุณลักษณะ)	√	√	√	√
- เพศ (1 คุณลักษณะ)	√	√	√	√
คุณลักษณะด้านพฤติกรรมการใช้งานระบบของนิสิต				
- การเข้าใช้งานระบบ (1 คุณลักษณะ)	√	√	√	X
- การชมสื่อการเรียนรู้ (37 คุณลักษณะ)	√	√	√	X
- การทำแบบทดสอบ (54 คุณลักษณะ)	√	√	√	X
- การส่งงานมอบหมาย (5 คุณลักษณะ)	√	√	√	X
คุณลักษณะด้านคะแนนการประเมินผลการเรียนรู้				
- คะแนนกิจกรรมในห้องเรียน (30%) (12 คุณลักษณะ)	√	√	X	√
- คะแนนกลางภาค (30%) (4 คุณลักษณะ)	√	X	X	√

ดังนั้น ในการศึกษานี้ เราจะอาศัยคะแนนไม่เกิน 60% ร่วมกับข้อมูลพฤติกรรมการใช้งานระบบของนิสิต มาใช้ในการทำนาย และข้อมูลรายบุคคล ได้แก่ เพศ และคณะ มาร่วมทำนายผลการเรียน ทั้งนี้การที่ผู้วิจัยนำข้อมูลในส่วนของเพศและคณะมาใช้งานด้วย เพื่อเพิ่มคุณลักษณะของข้อมูลให้สมบูรณ์มากขึ้น

นอกจากนี้ในแบบจำลองในแต่ละชุด ผู้วิจัยยังได้เปรียบเทียบประสิทธิภาพระหว่างแบบจำลองที่ใช้คุณลักษณะของข้อมูลในแต่ละชุดครบถ้วน กับแบบจำลองที่ใช้ข้อมูลที่ผ่านการเลือกคุณลักษณะ โดยใช้แบบจำลอง XGBoost เป็นตัวคัดเลือกคุณลักษณะซึ่งใช้ SelectFromModel(XGBClassifier (random_state=42)) โดยผลการทดลองที่ได้ ได้แสดงไว้ในตารางที่ 4

7. การวัดประสิทธิภาพของแบบจำลอง

หลังจากสร้างแบบจำลองประเภทเหล่านั้นแล้ว สำหรับการวิจัยนี้จะวัดความถูกต้อง (Accuracy), ความแม่นยำ (Precision) เพื่อประเมินแบบจำลองว่า แบบจำลองมีประสิทธิภาพในระดับดีมากหรือน้อยแค่ไหน โดยแสดงสมการ (6) - (7) (Almasri และคนอื่น ๆ , 2020)

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

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

โดยตัวแปรต่าง ๆ จากสมการ (6)-(7) มีความหมาย ดังนี้

1. True Positive (TP) คือ ผลการทำนายว่าเป็นคลาสที่สนใจ เช่นเดียวกับข้อมูลแท้จริงเป็นคลาสที่สนใจมีความหมายว่าเป็นการทำนายถูกต้อง

2. True Negative (TN) คือ ผลการทำนายว่าไม่ได้เป็นคลาสที่สนใจ เช่นเดียวกับข้อมูลแท้จริงไม่ได้เป็นคลาสที่สนใจ มีความหมายว่าเป็นการทำนายถูกต้อง

3. False Positive (FP) คือ ผลการทำนายว่าเป็นคลาสที่สนใจ แต่ข้อมูลแท้จริงไม่ได้เป็นคลาสที่สนใจมีความหมายว่าเป็นการทำนายผิดพลาด

4. False Negative (FN) คือ ผลการทำนายว่าไม่ได้เป็นคลาสที่สนใจ แต่ข้อมูลแท้จริงเป็นคลาสที่สนใจมีความหมายว่าเป็นการทำนายผิดพลาด

ผลการทดลองและการอภิปรายผล

ในการวัดประสิทธิภาพแบบจำลอง เราใช้ความแม่นยำเป็นตัววัดประสิทธิภาพ เพื่อเปรียบเทียบหลังจากนำชุดข้อมูลการทดสอบ 4 ชุด ไปทำนายข้อมูลด้วยแบบจำลองการจำแนกประเภท ผลการทำนายผลการเรียนที่ได้จากการใช้คุณลักษณะทั้งหมด และการคัดเลือกคุณลักษณะของชุดข้อมูล

ที่แตกต่างกัน 4 ชุด โดยแบบจำลองในแต่ละชุดที่มีประสิทธิภาพดีที่สุด จะแสดงด้วยตัวเลขเข้ม (Bold Font) ดังตารางที่ 4

ตารางที่ 4 การเปรียบเทียบความแม่นยำสำหรับชุดทดสอบที่แตกต่างกัน 4 ชุด ของแบบจำลอง 7 แบบ

แบบจำลอง		ชุดที่ 1 ข้อมูลบันทึกการใช้งาน คะแนนกิจกรรมในห้องเรียนออนไลน์ และคะแนนกลางภาค		ชุดที่ 2 ข้อมูลบันทึกการใช้งาน และคะแนนกิจกรรมในห้องเรียนออนไลน์		ชุดที่ 3 ข้อมูลบันทึกการใช้งาน		ชุดที่ 4 คะแนนกิจกรรมในห้องเรียนออนไลน์ และคะแนนกลางภาค	
		ความแม่นยำ (%)	เปลี่ยนแปลง*	ความแม่นยำ (%)	เปลี่ยนแปลง*	ความแม่นยำ (%)	เปลี่ยนแปลง*	ความแม่นยำ (%)	เปลี่ยนแปลง*
1. Logistic Regression	ใช้คุณลักษณะทั้งหมด	62.96	11.11	65.43	3.71	56.79	9.88	74.07	-14.81
	คัดเลือกคุณลักษณะ	74.07		69.14		66.67		59.26	
2. Naïve Bayes	ใช้คุณลักษณะทั้งหมด	65.43	-	60.49	6.18	48.15	-	64.20	-3.71
	คัดเลือกคุณลักษณะ	65.43		66.67		48.15		60.49	
3. K-Nearest Neighbor	ใช้คุณลักษณะทั้งหมด	62.96	9.88	62.96	1.24	55.56	8.64	**77.78**	-7.41
	คัดเลือกคุณลักษณะ	72.84		64.20		64.20		70.37	
4. Support Vector Machine	ใช้คุณลักษณะทั้งหมด	71.60	1.24	58.02	8.65	55.56	7.40	71.60	-12.34
	คัดเลือกคุณลักษณะ	72.84		66.67		62.96		59.26	
5. Decision Tree	ใช้คุณลักษณะทั้งหมด	65.43	-11.11	58.02	-	46.91	-	69.14	-9.88
	คัดเลือกคุณลักษณะ	54.32		58.02		46.91		59.26	
6. Random Forest	ใช้คุณลักษณะทั้งหมด	69.14	4.93	69.14	8.64	**67.90**	-7.41	76.54	-17.28
	คัดเลือกคุณลักษณะ	74.07		77.78		60.49		59.26	
7. XGBoost	ใช้คุณลักษณะทั้งหมด	**83.95**	-9.88	**79.01**	-	65.43	1.24	76.54	-14.81
	คัดเลือกคุณลักษณะ	74.07		**79.01**		66.67		61.73	

หมายเหตุ เปลี่ยนแปลง* หมายถึง ร้อยละการเปลี่ยนแปลงของความแม่นยำ หลังจากผ่านการคัดเลือกคุณลักษณะ

1) ผลการเปรียบเทียบแบบจำลองการทำนายผลการเรียนที่ใช้ข้อมูลพฤติกรรมการใช้งานระบบร่วมกับคะแนน และแบบจำลองที่ใช้เพียงคะแนน

จากตารางที่ 4 จะเห็นว่า เมื่อเปรียบเทียบผลของแบบจำลองแต่ละชนิด เมื่อนำมาใช้กับข้อมูลแต่ละชุด ข้อมูลชุดที่ 1 จะให้ประสิทธิภาพสูงสุด ซึ่งเป็นไปตามที่คาดการณ์ เนื่องจากการที่เราใช้คุณลักษณะอย่างครบถ้วน ควรได้ผลการทำนายดีแม่นยำที่สุด โดยแบบจำลองที่ทำนายได้แม่นยำที่สุด

คือ แบบจำลอง XGBoost โดยมีความแม่นยำถึง 83.95% แสดงให้เห็นได้ว่า แบบจำลอง XGBoost มีศักยภาพในการนำมาใช้ในการทำนายผลการเรียนของนิสิตได้อย่างมีประสิทธิภาพ

ตารางที่ 5 ค่า Prediction และค่า Recall ของแบบจำลอง XGBoost ของข้อมูลชุดที่ 1

ระดับผลการเรียน	ค่า Precision	ค่า Recall
ดีมาก	0.93	0.86
ดี	0.76	0.81
ต้องปรับปรุง	0.73	0.80

นอกจากนี้ เมื่อพิจารณาจากค่า Prediction และค่า Recall ของ XGBoost ในข้อมูลชุดที่ 1 ดังแสดงในตารางที่ 5 จะเห็นว่า แบบจำลองจะมีค่า Recall ในกลุ่มผลการเรียนต้องปรับปรุงอยู่ที่ 0.80 แม้ว่าก่อนหน้านี้ ในกลุ่มข้อมูลของเรา นิสิตในกลุ่มนี้มีปริมาณค่อนข้างน้อย แต่แบบจำลองนี้ยังมีความสามารถค้นหานิสิตในกลุ่มนี้ออกมาได้ เทียบเท่ากับนิสิตในกลุ่มผลการเรียนดีมาก และดี ซึ่งนิสิตในกลุ่มนี้ จะเป็นนิสิตกลุ่มที่เราต้องการแยกแยะออกมา เพื่อให้ผู้สอนให้ความเอาใจใส่เป็นพิเศษ และนิสิตได้มีโอกาสปรับปรุงผลการเรียนให้ดีขึ้นได้

ในส่วนของการทำนายผลการเรียนโดยไม่ใช้คะแนนสอบกลางภาคในชุดที่ 2 นั้น จะเห็นได้ว่า โดยรวมแล้ว ประสิทธิภาพของแบบจำลองในข้อมูลชุดที่ 1 สูงกว่าแบบจำลองในข้อมูลชุดที่ 2 โดยเฉลี่ย 4.06% แสดงให้เห็นว่าการทำนายล่วงหน้าก่อนที่นิสิตจะได้ผลสอบกลางภาคออกมานั้น มีความแม่นยำลดลงไม่มากนัก นอกจากนี้แม้ว่าค่าความแม่นยำของแบบจำลอง XGBoost จะตกลงบ้าง แต่ค่า Recall ในกรณีข้อมูลชุดที่ 2 ในกลุ่มนิสิตผลการเรียนต้องปรับปรุงก็ยังมีค่าอยู่ที่ 0.80 เช่นเดิม

ทว่า หากเราไม่พิจารณาคะแนน ดังในข้อมูลชุดที่ 3 จะเห็นว่า ประสิทธิภาพของแบบจำลองทุกชนิดลดลงอย่างเห็นได้ชัดเจน โดยเฉลี่ยแล้วลดลงถึง 12.17% เมื่อเทียบกับประสิทธิภาพของแบบจำลองในข้อมูลชุดที่ 1 ทว่า ในข้อมูลชุดนี้ แบบจำลอง XGBoost ก็ยังมีประสิทธิภาพสูงอยู่ เมื่อเทียบกับแบบจำลองชนิดอื่น อีกทั้ง ในข้อมูลชุดที่ 3 นี้ ค่า Recall ในกลุ่มผลการเรียนต้องปรับปรุงของทุกแบบจำลอง จะมีค่าต่ำกว่าหรือเท่ากับ 0.60 ทั้งสิ้น (โดยแบบจำลอง XGBoost มีค่า Recall เท่ากับ 0.50) ดังนั้น ในการทำนายผลการเรียนโดยอาศัยแต่เพียงข้อมูลพฤติกรรมการใช้งานระบบ ของนิสิตอย่างเดียว โดยไม่อาศัยคะแนนในชั้นเรียนอาจไม่เพียงพอต่อการทำนายผลการเรียนของนิสิตได้อย่างแม่นยำ

ในส่วนของประสิทธิภาพแบบจำลองที่ใช้ข้อมูลชุดที่ 4 จะมีความแม่นยำเพิ่มขึ้นไม่มากโดยเฉลี่ยถึง 4.06% เมื่อเปรียบเทียบกับความแม่นยำของข้อมูลชุดที่ 1 แต่แบบจำลอง XGBoost ในชุดที่ 1 มีค่าความแม่นยำสูงกว่าแบบจำลองนี้ในชุดที่ 4 ถึง 7.41% แสดงให้เห็นว่า แบบจำลอง XGBoost ที่ใช้แต่เพียงคะแนนที่ได้อย่างเดียว โดยไม่พิจารณาข้อมูลพฤติกรรมการใช้งานระบบของนิสิต ให้ผลการทำนายแม่นยำที่น้อยกว่าเมื่อเทียบกับแบบจำลองนี้ที่ใช้ข้อมูลพฤติกรรมการใช้งานระบบของนิสิตร่วมกับคะแนน

จากการศึกษาในส่วนนี้ จะเห็นว่า การใช้คุณลักษณะครบถ้วนทั้งสองกลุ่มคือ คุณลักษณะด้านคะแนน และคุณลักษณะด้านพฤติกรรมการเข้าใช้ระบบร่วมกัน ในชุดข้อมูลที่ 1 จะให้ประสิทธิภาพของทุกแบบจำลองสูงสุด ในขณะที่ในชุดข้อมูลที่ 3 เราใช้คุณลักษณะด้านพฤติกรรมการเข้าใช้ระบบแต่เพียงอย่างเดียว จะให้ผลการทำนายที่ต่ำกว่าชุดข้อมูลที่ 4 ที่ใช้แต่เพียงคุณลักษณะด้านคะแนนอย่างเดียว จึงทำให้เห็นได้ว่า อย่างไรก็ตาม คุณลักษณะด้านคะแนน ช่วยให้ทำนายผลการเรียนได้แม่นยำกว่า คุณลักษณะด้านพฤติกรรมแต่เพียงอย่างเดียว แต่ทั้งนี้ แบบจำลองจะให้ความแม่นยำสูงสุดเมื่อใช้คุณลักษณะทั้งสองกลุ่มนี้ร่วมกัน

ในส่วนของประสิทธิภาพแบบจำลองชุดข้อมูลแต่ละชุดแล้ว จะเห็นว่า แบบจำลอง XGBoost ให้ประสิทธิภาพการทำนายสูงสุด ทั้งในข้อมูลชุดที่ 1 และ 2 ซึ่งเป็นข้อมูลส่วนที่ใช้คุณลักษณะมากที่สุด และมีผลการทำนายที่แม่นยำที่สุด ส่วนในชุดข้อมูลที่ 3 แบบจำลองที่ให้ประสิทธิภาพดีที่สุดจะเป็น Random forest ได้ค่า 67.9 ส่วนแบบจำลอง XGBoost จะให้ค่าความแม่นยำเป็นอันดับสอง ที่ค่า 65.43 ส่วนในชุดข้อมูลที่ 4 แบบจำลอง XGBoost จะให้ความแม่นยำ 76.54 ซึ่งมีค่าเป็นอันดับสอง ในขณะที่แบบจำลอง K-Nearest Neighbor ให้ค่าความแม่นยำสูงสุดที่ 77.78 จะเห็นได้ว่า โดยรวมแล้ว แบบจำลอง XGBoost สามารถทำนายข้อมูลชุดต่าง ๆ ได้อย่างแม่นยำอย่างสม่ำเสมอ แม้ว่าในบางชุดข้อมูล แบบจำลอง XGBoost อาจมิได้ให้ค่าประสิทธิภาพที่สูงที่สุด แต่ค่าความแม่นยำที่ได้ ก็มิได้แย่กว่าแบบจำลองที่ดีที่สุดมากนัก ทั้งในชุดข้อมูลที่ 3 และ 4 แบบจำลอง XGBoost ยังให้ความแม่นยำเป็นอันดับสอง จึงสรุปได้ว่า แบบจำลอง XGBoost เป็นแบบจำลองที่เหมาะกับการนำมาใช้งานกับในข้อมูลชุดนี้มากที่สุด เนื่องจากให้ประสิทธิภาพโดยรวมสูงที่สุดอย่างสม่ำเสมอในชุดข้อมูลทุกประเภท

ในส่วนของการศึกษาชุดข้อมูลที่ 3 และ 4 นั้น เราต้องการศึกษาประสิทธิภาพการทำนายของการใช้คุณลักษณะกลุ่มของพฤติกรรมการเข้าใช้งานระบบ เทียบกับการใช้คุณลักษณะกลุ่มของคะแนนสอบ ซึ่งในการศึกษานี้ เราพบว่าในการสร้างแบบจำลองที่โดยรวมแล้ว การทำนายผลการเรียนด้วยคุณลักษณะกลุ่มของคะแนนสอบแต่เพียงอย่างเดียวจะให้ความแม่นยำสูงกว่า การใช้คุณลักษณะกลุ่มของพฤติกรรม การเข้าใช้งานระบบแต่เพียงอย่างเดียว

แต่ในการสร้างแบบจำลองเพื่อใช้งานจริง เราจะสร้างแบบจำลองที่ใช้ชุดข้อมูลที่ 1 เป็นหลัก จึงได้ข้อสรุปว่า แบบจำลอง XGBoost เป็นแบบจำลองที่เหมาะสมกับข้อมูลชุดนี้ที่สุด

2) ผลการเปรียบเทียบแบบจำลองการทำนายผลการเรียนที่ใช้คุณลักษณะทั้งหมดและแบบจำลองที่มีการคัดเลือกคุณลักษณะ

ตารางที่ 6 จำนวนคุณลักษณะก่อนและหลังจากทำการคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะ	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4
ก่อนการคัดเลือกคุณลักษณะ	115	111	99	18
หลังการคัดเลือกคุณลักษณะ	36	33	41	5

โดยรวมแล้ว การคัดเลือกคุณลักษณะ มิได้ช่วยให้แบบจำลองทำงานได้มีประสิทธิภาพเพิ่มขึ้นนัก ในแบบจำลองที่มีความแม่นยำค่อนข้างสูง เช่น พวก XGBoost เป็นต้น แต่ในแบบจำลองที่มีความซับซ้อนต่ำ เช่น แบบจำลอง Naïve Bayes หรือ Logistic Regression การคัดเลือกคุณลักษณะช่วยเพิ่มความแม่นยำอย่างเห็นได้ชัด ทั้งนี้ สาเหตุอาจเป็นเพราะแบบจำลองเหล่านี้ มีความซับซ้อนไม่มากนัก เมื่อนำมาใช้กับข้อมูลที่มีจำนวนคุณลักษณะมาก ๆ จะทำให้ไม่สามารถประมวลผลข้อมูลได้ดีนักทำให้มีความแม่นยำค่อนข้างต่ำ แต่เมื่อลดจำนวนคุณลักษณะลง ทำให้แบบจำลองเหล่านี้สามารถทำงานกับข้อมูลที่มีความซับซ้อนน้อยกว่าได้อย่างมีประสิทธิภาพมากขึ้น จึงทำให้แบบจำลองเหล่านี้มีความแม่นยำสูงขึ้น เมื่อเราทำการคัดเลือกคุณลักษณะด้วย ในขณะที่แบบจำลองพวก XGBoost เดิมที มีความซับซ้อนสูงอยู่แล้ว สามารถประมวลผลข้อมูลที่มีความซับซ้อนมาก ๆ ได้ จึงทำให้ทำงานกับข้อมูลที่ใช้คุณลักษณะครบถ้วนได้อย่างมีประสิทธิภาพ ในขณะที่เมื่อเราทำการคัดเลือกคุณลักษณะ จำเป็นต้องมีการตัดข้อมูลบางส่วนออกไป ทำให้ผลการทำนายไม่แม่นยำเท่ากับในกรณีที่แบบจำลองประมวลผลข้อมูลที่ใช้คุณลักษณะครบถ้วน

สรุปงานวิจัย

งานวิจัยนี้มีการสร้างแบบจำลองการทำนายผลการเรียนของนิสิตที่ใช้ระบบการจัดการเรียนรู้ออนไลน์ โดยร่วมกับการปรับความสมดุลของข้อมูลด้วยเทคนิค SMOTE และการคัดเลือกคุณลักษณะ ผลการเรียนของนิสิตแบ่งออกเป็น 3 ระดับ ได้แก่ ดีมาก, ดี และต้องปรับปรุง ชุดข้อมูลสำหรับการทำนายผลการเรียน มีคุณลักษณะหลายประเภท ได้แก่ ข้อมูลทั่วไป พฤติกรรมการใช้งานระบบ คะแนนเก็บ และคะแนนกลางภาค ชุดข้อมูลนี้ถูกแบ่งออกเป็น 4 กรณี ตามช่วงเวลาการเรียนที่แตกต่างกัน และประเภทของคุณลักษณะ ได้แก่ ชุดที่ 1 ใช้ข้อมูลทั้งหมดทุกคุณลักษณะ, ชุดที่ 2 จะไม่นำคะแนนกลางภาคมาใช้, ชุดที่ 3 อาศัยคุณลักษณะของพฤติกรรมการใช้งานระบบของนิสิตเท่านั้น และชุดที่ 4 ใช้เพียงคะแนนทั้งหมดเท่านั้นมีการประเมินประสิทธิภาพของแบบจำลองต่าง ๆ โดยพิจารณาค่าความแม่นยำจากการเปรียบเทียบประสิทธิภาพของแบบจำลอง การทำนายผลการเรียน

ที่ใช้ข้อมูลพฤติกรรมการใช้งานระบบร่วมกับคะแนนและแบบจำลองที่ใช้เพียงคะแนน กับการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้คุณลักษณะทุกประเภท และแบบจำลองที่มีการคัดเลือกคุณลักษณะ สรุปได้ว่า

1. แบบจำลอง XGBoost ที่ใช้ข้อมูลคุณลักษณะของพฤติกรรมการใช้งานระบบร่วมกับคะแนนในชุดที่มีประสิทธิภาพการทำนายที่ดีที่สุด โดยคุณลักษณะด้านคะแนนจะมีผลต่อความแม่นยำของแบบจำลองมากกว่าคุณลักษณะด้านพฤติกรรมการใช้งานระบบ

2. การคัดเลือกคุณลักษณะ ช่วยให้แบบจำลองที่มีความซับซ้อนไม่มาก เช่น Logistic Regression หรือ Naïve Bayes มีประสิทธิภาพดีขึ้น แต่ไม่ได้ช่วยแบบจำลองที่มีความซับซ้อนสูงอยู่แล้ว เช่น XGBoost

นอกจากนี้คุณลักษณะที่สำคัญในการทำนายผลการเรียนของนิสิต คือ พฤติกรรมการใช้งานระบบของนิสิต มีหลายพฤติกรรมถึง 19 คุณลักษณะ คะแนนผลการประเมินการเรียนรู้ มีเพียง 10 คุณลักษณะ และข้อมูลคณะของนิสิต ดังนั้น สามารถนำแบบจำลอง XGBoost ไปทำนายผลการเรียนของนิสิต โดยใช้คุณลักษณะของพฤติกรรมการใช้งานระบบร่วมกับคะแนนและคณะของนิสิต เพื่อนำผลการทำนายนี้ไปคำแนะนำและช่วยเหลือนิสิตได้อย่างเจาะจง สามารถติดตามความก้าวหน้าในการเรียนของนิสิตได้เป็นรายบุคคล หรือวางแผนการสอนให้เหมาะสมกับนิสิต และนิสิตมีเวลาพัฒนาทักษะการเรียนรู้มากขึ้น

กิตติกรรมประกาศ

บทความวิจัยนี้สำเร็จลุล่วงไปด้วยความกรุณาของ รองศาสตราจารย์ ดร.ชนัตถ์ พูนเดช รองศาสตราจารย์ ดร.ธนิตา เลิศพรกุลรัตน์ ขอขอบพระคุณท่านที่ได้ให้ความอนุเคราะห์ให้ผู้วิจัยได้ใช้ข้อมูล รวมทั้งเสียสละเวลาอันมีค่าในการให้คำแนะนำที่เป็นประโยชน์อย่างยิ่ง เปิดโอกาสให้ซักถาม ช่วยให้ข้อมูลนิสิต ข้อมูลกิจกรรมการเรียนการสอน เกณฑ์การประเมินผลการเรียนรู้ และขอขอบคุณ รองศาสตราจารย์ ดร.จิตต์ภิญญา ชุมสาย ณ อยุธยา อาจารย์ ดร.สุนทรี สกุลพราหมณ์ ที่ให้ความอนุเคราะห์แก่ผู้วิจัยได้เก็บข้อมูล

เอกสารอ้างอิง

Almasri, A., Alkhawaldeh, R.S. and Çelebi, E. (2020). Clustering-Based EMT Model for Predicting Student Performance. Arabian Journal for Science and Engineering, 45(12), 10067-10078. https://doi:10.1007/s13369-020-04578-4.

Amra, I.A.A. and Maghari, A.Y.A. (2017). Students performance prediction using KNN and Naïve Bayesian. *In International Conference on Information Technology (ICIT)*.

Amrieh, E., Hamtini, T. and Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, 9, 119-136. https://doi:10.14257/ijdta.2016.9.8.13.

Ashfaq, U., Marikannan, B.P. and Raheem, M. (2020). Managing Student Performance: A Predictive Analytics using Imbalanced Data. *International Journal of Recent Technology and Engineering*, 8, 2277-2283.

Awad, M. and Khanna, R. (2015*). Efficient Leaning Machines.* Apress, USA. from https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_1.

Awlla, A. (2019). Performance Analysis and Prediction Student Performance to build effective student Using Data Mining Techniques. *UHD Journal of Science and Technology*, 3(10), 10-15.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Chen, J., Chen, Y., Du, X. et al. (2013). Big data challenge: a data management perspective. *Front. Comput. Sci*, 7, 157-164. https://doi.org/10.1007/s11704-013-3903-7

Gasevic, D., Siemens, G. and Rose, C.P. (2017). Guest editorial: Special section on learning analytics. *IEEE Trans. Learn. Technol*, 10, 3–5.

Hooshyar, D. and Pedaste, M. (2019). Mining Educational Data to Predict Students' Performance through Procrastination Behavior. *Entropy*, 22, 12. https://doi:10.3390/e22010012.

Jalota C., Agrawal R. (2021) Feature Selection Algorithms and Student Academic Performance: A Study. In: Gupta D., Khanna A., Bhattacharyya S., Hassanien A.E., and S., Jaiswal A. (eds) *International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing*, vol 1165. Singapore: Springer. https://doi.org/10.1007/978-981-15-5113-0_23.

Jayaprakash, S., Krishnan, S. and Jaiganesh, V. (2020). Predicting Students Academic Performance using an Improved Random Forest Classifier. *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*.

Mohamad, N., Ahmad, N., Jawawi, D. and Mohd Hashim, S. (2020). Feature Engineering for Predicting MOOC Performance. IOP Conference Series: *Materials Science and Engineering*, 884, 012070.

Patel, H. and Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering*, 6, 74-78.

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B. and Chen, P. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Sci Rep* 12, 453.

Quinn, R. and Gray, G. (2019). Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, 5. https://doi:10.22554/ijtel.v5i1.57.

Ramaswami, G., Susnjak, T., Mathrani, A. and Umer, R. (2020). Predicting Students Final Academic Performance using Feature Selection Approaches. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1-5.

Riestra-González, M., Paule-Ruíz, M.D.P. and Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, 163, 104108. https://doi.org/10.1016/j.compedu.2020.104108.

Shrestha, S. and Pokharel, M. (2021). Educational data mining in moodle data. *International Journal of Informatics and Communication Technology*, 10, 9. https://doi:10.11591/ijict.v10i1.pp9-18.

Sun, J., Lang, J., Fujita, H. and Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, 425, 76-91.

Tan, M. and Shao, P. (2015). Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. *International Journal of Emerging Technologies in Learning*, 10. https://doi:10.3991/ijet.v10i1.4189.