การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้การศึกษาระบบเปิดด้วยสื่อการสอนอิเล็กทรอนิกส์ในระดับอุดมศึกษา

ไพศาล สิมาเลาเต่า[1*] และ จรัญ แสนราช[2]

บทคัดย่อ

งานวิจัยนี้นำเสนอผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา โดยเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของ 3 เทคนิค คือ เทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes โดยการนำข้อมูลเกี่ยวกับการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา จำนวน 10 แอททริบิวต์ มีข้อมูลจำนวนที่สามารถใช้ในงานวิจัยได้ 152,850 ชุด ซึ่งทำการแบ่งข้อมูลด้วยวิธี Cross-validation Test ออกเป็น 5 ส่วน หรือ 5-fold cross-validation โดยการสุ่มข้อมูลเพื่อแบ่งข้อมูลในแต่ละส่วนประกอบด้วยข้อมูล จำนวน 30,570 ชุด โดยทำการเลือกข้อมูล 4 ส่วน สำหรับใช้ในการสร้างโมเดล และเลือกข้อมูล 1 ส่วน สำหรับใช้ในการทดสอบประสิทธิภาพของระบบ ผลการวิจัยพบว่า เทคนิคที่ใช้ในการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา ที่มีประสิทธิภาพสูงสุด คือ Deep Learning โดยมีค่าความแม่นตรง 90.60% ค่าความระลึก 98.39% ค่าความแม่นยำ 99.01% และค่าถ่วงดุล 0.321 ซึ่งเป็นระดับการประเมินที่สามารถยอมรับได้

คำสำคัญ: การจำแนกข้อมูล, การแบ่งข้อมูลเพื่อทดสอบโมเดล, การสุ่มป่าไม้, การเรียนรู้เชิงลึก, นาอีฟเบย์

[1] อาจารย์ สาขาวิชาวิทยาการคอมพิวเตอร์ สาขาวิชาคอมพิวเตอร์ศึกษา คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[2] อาจารย์ ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
[*] ผู้นิพนธ์ประสานงาน โทร. +668 5295 9997 อีเมล: paisan.smlt@gmail.com

A Comparison of the Efficiency of Data Classification in Learning Factors through Open Educational System with Electronic Teaching Aids in Tertiary Level

Paisan Simalaotao[1*] and Charan Sanrach[2]

Abstract

This research presented the result of a comparison of the efficiency of data classification in learning factors through open educational system with electronic teaching aids of tertiary level students. There were 3 primary techniques which were compared in this research as follows: 1) Random Forest technique, 2) Deep Learning technique, and 3) Naive Bayes technique. There were 10 attributes with 152,850 datasets of the data which were provided in electronic teaching aids of open educational system for tertiary students. The data was divided into 5 parts by Cross-validation Test method, called 5-fold cross-validation. The data was randomly sampled for each section which consisted of 30,570 dataset. There were 4 parts used to generate a model and the single part was utilized to examine the precision and accuracy of the model. The results revealed that the most efficient technique, used to classify learning factors through open educational system with electronic teaching aids of tertiary level students, was Deep Learning technique with 90.60 percent of precision, 98.39 percent of recall, 99.01percent of accuracy, and 0.321 F-measure at the acceptable level of evaluation.

Keywords: Classification, Cross-validation Test, Random Forest, Deep Learning, Naive Bayes

[1] Lecturers, Department of Computer Science, Department of Computer Education, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

[2] Lecturers, Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok

[*] Corresponding Author Tel. +668 5295 9997 email: paisan.smlt@gmail.com

1. บทนำ

ความก้าวหน้าทางด้านเทคโนโลยีคอมพิวเตอร์และการสื่อสาร ทำให้เกิดการเปลี่ยนแปลงในหลายด้าน แม้แต่ทางด้านการศึกษาก็กำลังประสบปัญหาการลดจำนวนของประชากร และการเปลี่ยนมุมมองทางด้านการศึกษาอย่างรวดเร็ว เนื่องจากมีระบบการสอนออนไลน์เกิดขึ้นมากมาย ทำให้มีหลายมหาวิทยาลัยในต่างประเทศหลายเริ่มเปลี่ยนรูปแบบการเรียนรู้โดยให้ความสำคัญกับการเรียนรู้ในระบบเปิดหรือการเรียนรู้แบบไม่มีห้องเรียนหรือการศึกษาทางไกลเป็นอย่างมาก มหาวิทยาลัยของไทยจึงควรเตรียมความพร้อมในการเปลี่ยนแปลงโดยยอมรับการนำเทคโนโลยีที่ทันสมัยเข้ามาใช้ในการเรียนรู้ และยอมรับการเรียนรู้แบบไม่ใช้ห้องเรียนหรือการใช้สื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา [1] จากความสำคัญของปัญหาดังกล่าวผู้วิจัยได้ค้นคว้าและทบทวนงานวิจัยทีเกี่ยวข้อง ได้แก่

เดช ธรรมศิริ และพยุง มีสัจ [2] ทำการวิจัยเกี่ยวกับการจำแนกข้อมูลด้วยเทคนิคการร่วมกันตัดสินใจโดยใช้โมเดลในการทำงานหลายโมเดล เพื่อเลือกตัวแทนที่เหมาะสมด้วยขั้นตอนเชิงพันธุกรรม และใช้การโหวตเสียงข้างมาก ทำให้การตัดสินใจมีความถูกต้องมากขึ้น ผลการวิจัยพบว่า การเลือกใช้ Decision tree, Neuron network, Support vector machine ให้ค่าประสิทธิภาพการเลือกใช้ตัวจำแนกร่วมกันที่มีประสิทธิภาพสูงสุด ทั้งนี้ การทำงานร่วมกันจะเกิดขึ้นตามสัดส่วนของจำนวนโมเดลที่แตกต่างกันตามข้อมูลที่ใช้ในการจำแนกสอดคล้องกับ ศุภเทพ สติมั่น, พยุง มีสัจ และปิยนุช วรบุตร [3] วิจัยเกี่ยวกับการรวมกลุ่มเพื่อจำแนกประเภทข้อความไม่สมดุลสูงด้วยเทคนิค Deep learning ร่วมกับการคัดเลือกคุณลักษณะและการปรับสมดุลข้อมูล โดยการแก้ไขที่ระดับข้อมูล การแก้ปัญหาด้วยการคัดเลือกคุณลักษณะข้อมูล และการแก้ปัญหาที่ระดับขั้นตอนวิธีโดยใช้วิธีแบบรวมกลุ่มซึ่งรวมเทคนิค Deep learning ร่วมกับเทคนิคการคัดเลือกโดยกรองคุณลักษณะด้วยค่าน้ำหนักอัตราส่วนมาตรฐานเกน สามารถจัดการปัญหาและให้ประสิทธิภาพการจำแนกประเภทดีกว่าขั้นตอนวิธีมาตรฐานทั่วไป

นิเวศ จิระวิชิตชัย และนรินทร์ พนาวาส [4] ทำการวิจัยเกี่ยวกับผลการจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยเลือกใช้เทคนิค Support vector machine, Decision tree, Naive Bayes และ K-Nearest Neighbor ในการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล ผลการวิจัยพบว่า เทคนิคซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพสูงสุด คือ 86.30

Jiansheng Wu, et al. [5] ทำการวิจัยเพื่อทำนายการผสม DNA ในโปรตีนจากลำดับกรดอะมิโน โดยใช้เทคนิค Random Forest เนื่องจากมีความรวดเร็วและมีประสิทธิภาพที่ดีสำหรับข้อมูลที่มีค่าพารามิเตอร์ที่แตกต่างกัน ผลการวิจัยพบว่า เทคนิค Random Forest มีความแม่นยำโดยรวม 91.41%

Dayong Wang, Aditya Khosla, et al. [6] ทำการวิจัย โดยการพัฒนาระบบ AI ที่ใช้เทคโนโลยี Deep Learning ด้วย Caffe Framework ซึ่งประมวลผลโดยใช้ NVIDIA Tesla K80 GPU เพื่อวินิจฉัยโรคมะเร็งเต้านมได้ โดยมีความผิดพลาดเพียง 2.9% และหากนำไปใช้ร่วมกับแพทย์เฉพาะทางจะทำให้อัตราความผิดพลาดในการวินิจฉัยลดลงเหลือเพียง 0.5% เท่านั้น ซึ่งสามารถลดความผิดพลาดในการตรวจหามะเร็งเต้านมได้ถึง 85%

จากการค้นคว้าและทบทวนงานวิจัยที่เกี่ยวข้อง ผู้วิจัยจึงมีแนวคิดในการวิจัยเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา โดยใช้ข้อมูลจากฐานข้อมูลการลงทะเบียนเรียนและกิจกรรมการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ระบบเปิด ประกอบด้วยข้อมูลผู้เรียนในระดับอุดมศึกษา ข้อมูลหลักสูตร ข้อมูลการใช้สื่อการสอนอิเล็กทรอนิกส์ และข้อมูลผลการเรียน เป็นต้น เพื่อทำนายผลการเรียนรู้จากพฤติกรรมการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ด้วยการจำแนกประเภทข้อมูล (Classification) โดยเลือกใช้เทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes ด้วยการแบ่งข้อมูลสำหรับการสร้างโมเดลและทดสอบโมเดลด้วยวิธี Cross-validation Test สร้าง Training data และ Testing data สำหรับทดสอบประสิทธิภาพของโมเดล ตามลำดับ

2. วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ ด้วยสื่อการสอนอิเล็กทรอนิกส์

ในระบบเปิดของผู้เรียนระดับอุดมศึกษา ระหว่างเทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes โดยการทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test ด้วยการแยกข้อมูลสำหรับสร้าง Training data และ Testing data

3. สมมติฐานการวิจัย

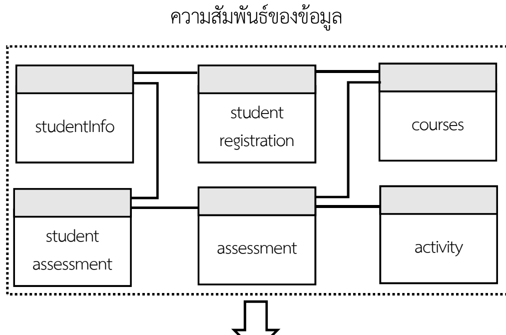
ได้เทคนิคที่เหมาะสมและมีประสิทธิภาพในการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา จากการเปรียบเทียบประสิทธิภาพระหว่างเทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes โดยการทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test ด้วยการแยกข้อมูลสำหรับสร้าง Training data และ Testing data

4. วิธีดำเนินการวิจัย

ผู้วิจัยดำเนินการวิจัยตามกระบวนการทำงาน Cross-Industry Standard Process for Data Mining หรือ CRISP-DM ซึ่งเป็นกระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านดาต้าไมน์นิง โดยแบ่งขั้นตอนการเนินการวิจัยออกเป็น 6 ขั้นตอน [7] ดังนี้

4.1 Business Understanding เป็นขั้นตอนการเข้าใจปัญหา โดยเริ่มจากการค้นคว้าและศึกษาชุดข้อมูลตัวอย่างที่ใช้ในการทดสอบ ซึ่งเป็นข้อมูลจากฐานข้อมูลการเรียนรู้ในระดับมหาวิทยาลัย ที่มีการจัดเก็บข้อมูลเกี่ยวกับข้อมูลผู้เรียน ข้อมูลหลักสูตรในระบบเปิด ข้อมูลการลงทะเบียนเรียน ข้อมูลพฤติกรรมการใช้งานสื่ออิเล็กทรอนิกส์ เป็นต้น

4.2 Data Understanding เป็นขั้นตอนทำความเข้าใจข้อมูล โดยการรวบรวมข้อมูลจากตารางต่าง ๆ มาหาความสัมพันธ์ของข้อมูลและตัดข้อมูลที่ไม่เกี่ยวข้องกับการวิจัยออก จากนั้นนำข้อมูลที่มีความสัมพันธ์กันมารวมเป็นชุดข้อมูลเดียวกันดังรูปที่ 1



assessment_type	date_submitted	date_registration	id_site	date	activity_type	sum_click	weight	score	final_result
TMA	18	-159	546652	19	resource	4	10	78	Fail
TMA	22	-53	546652	54	oucontent	1	20	70	Pass
TMA	17	-92	546652	117	resource	11	20	72	Pass
TMA	26	-52	546614	166	url	1	20	69	Fail
TMA	19	-176	546714	215	resource	1	30	79	Pass
Exam	20	-110	546652	215	homepage	8	100	70	Pass
TMA	18	-67	546876	19	url	2	10	72	Withdrawn

รวบรวมข้อมูล

รูปที่ 1 ทำความเข้าใจและรวบรวมข้อมูล

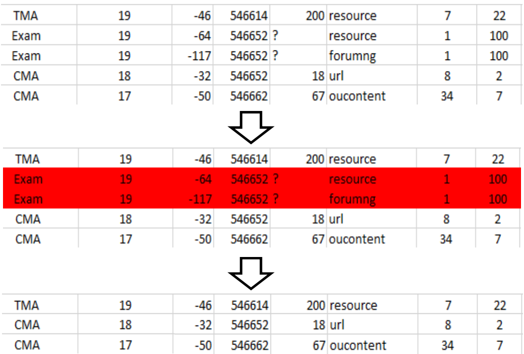
จากรูปที่ 1 ทำความเข้าใจเกี่ยวกับความสัมพันธ์ของข้อมูลและรวบรวมข้อมูลเป็นชุดเดียวกัน เพื่อตรวจสอบและวิเคราะห์ความสมบูรณ์ของข้อมูล พบว่าชุดข้อมูลที่สามารถใช้ในการจำแนกข้อมูลด้านปัจจัยสนับสนุนการเรียนรู้ในระบบเปิดด้วยสื่อการสอนอิเล็กทรอนิกส์ ประกอบด้วยข้อมูลที่มีความสัมพันธ์กันจำนวน 10 แอททริบิวต์ ดังตารางที่ 1

ตารางที่ 1 แอททริบิวต์ที่ใช้ในการจำแนกข้อมูล

แอททริบิวต์	ประเภทข้อมูล
assessment_type	ประเภทการประเมิน
date_registration	จำนวนวันเข้าสู่ระบบเพื่อใช้งาน
date_submitted	จำนวนวันส่งผลการร่วมกิจกรรม
id_site	ส่วนของข้อมูลที่เข้าเรียนรู้
code_module	โมดูลย่อยทีเข้าใช้งาน
activity_type	ประเภทกิจกรรมที่เข้าใช้งาน
sum_activity	จำนวนครั้งการคลิกทำกิจกรรม
weight	ค่าน้ำหนักของการร่วมกิจกรรม
score	ระดับคะแนน
final_result	ผลการประเมิน

4.3 Data Preparation เป็นขั้นตอนการแปลงข้อมูล เนื่องจากเทคนิคการจำแนกข้อมูลที่เลือกนำมาใช้ในการวิจัยมีความหลากหลาย ทำให้ข้อมูลที่เตรียมไว้ไม่สามารถใช้งานร่วมกันในบางเทคนิคได้ ผู้วิจัยจึงทำการแปลงข้อมูล โดยเริ่มจากการตรวจสอบ

ความสมบูรณ์ของข้อมูลในแต่ละระเบียน และจัดทำข้อมูลให้มีความถูกต้อง (data cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกัน หรือลบข้อมูลที่ไม่มีค่าออกไป เป็นต้น และกำหนดการแทนค่าข้อมูลของแต่ละแอททริบิวต์ให้ถูกต้องตามประเภทของข้อมูล แสดงตัวอย่างข้อมูลดังรูปที่ 2



TMA	19	-46	546614	200	resource	7	22
Exam	19	-64	546652	?	resource	1	100
Exam	19	-117	546652	?	forumng	1	100
CMA	18	-32	546652	18	url	8	2
CMA	17	-50	546662	67	oucontent	34	7

TMA	19	-46	546614	200	resource	7	22
Exam	19	-64	546652	?	resource	1	100
Exam	19	-117	546652	?	forumng	1	100
CMA	18	-32	546652	18	url	8	2
CMA	17	-50	546662	67	oucontent	34	7

TMA	19	-46	546614	200	resource	7	22
CMA	18	-32	546652	18	url	8	2
CMA	17	-50	546662	67	oucontent	34	7

รูปที่ 2 ตัวอย่างการแปลงข้อมูลสำหรับทดสอบโมเดล

จากรูปที่ 2 พบว่า มีข้อมูลที่ไม่สามารถระบุค่าได้ เช่น ค่าว่าง หรือ ?? เป็นต้น จึงทำการลบข้อมูลในบางเรคคอร์ดออกไปเพื่อให้ข้อมูลมีความสมบูรณ์ที่สุด เมื่อตรวจสอบประเภทของข้อมูล พบว่า มีข้อมูลบางแอททริบิวต์อยู่ในรูปแบบ polynomial ซึ่งไม่เหมาะกับการนำ ไปใช้ในบางเทคนิค จึงปรับข้อมูลให้อยู่ในรูปแบบ Integer เช่น ปรับข้อมูลแอททริบิวต์ activity_type โดยกำหนดให้ outcontent = 1, resource = 2, url = 3 และ homepage = 4 เป็นต้น เพื่อให้อยู่ในรูปแบบที่สามารถใช้งานร่วมกับเทคนิคในการจำแนกข้อมูลได้ ดังตารางที่ 2

ตารางที่ 2 ประเภทข้อมูลของแต่ละแอททริบิวต์

แอททริบิวต์	ประเภทข้อมูลก่อนปรับค่า	ประเภทข้อมูลหลังปรับค่า
assessment_type	polynominal	integer
date_registration	integer	integer
date_submitted	integer	integer
id_site	integer	integer
code_module	polynominal	integer
activity_type	polynominal	integer
sum_activity	integer	integer
weight	integer	integer
score	integer	integer
final_result	polynominal	polynominal

จากตารางที่ 2 พบว่า ข้อมูลของแอททริบิวต์ assessment_type, code_module และ activity_type เป็นข้อมูลประเภท polynomial จึงทำการตรวจสอบเพื่อเปรียบเทียบและแปลงข้อมูลจากประเภท polynomial เป็นข้อมูลประเภท integer ให้ถูกต้องและเหมาะสม

4.4 Modeling เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล โดยผู้วิจัยเลือกใช้การจำแนกประเภทข้อมูล (Classification) ในการทดลองประกอบด้วย

เทคนิค Random Forest [8] เป็นการเพิ่มความหลากหลายของโมเดลด้วยการสุ่มแอตทริบิวต์ โดยใช้เทคนิค Decision Tree ในการสร้างโมเดล โดยสุ่มข้อมูลและคุณลักษณะของ Decision Tree ซึ่งสร้างจากตัวอย่างแบบเลือกแล้วใส่กลับ (Sampling with Replacement) ซึ่งจะมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือก หรือ Out-of-Bag (OOB) โดยนำมาทดสอบ Decision Tree วิธีการดังกล่าวนี้เรียกว่า Bagging และหาผลโหวตที่มากที่สุด [9]

เทคนิค Deep Learning [10] เป็นการเรียนรู้เชิงลึก โดยแบ่งชั้นการทำงานออกเป็นชั้นนำเข้าข้อมูล (input layer) ชั้นซ่อน (hidden layer) และชั้นผลลัพธ์ (output layer) โดยมีการประมวลผลในชั้นซ่อนหลายชั้น ซึ่งข้อมูลแต่ละชั้นได้มาจากการปฏิสัมพันธ์กับชั้นอื่น โดยมี activation function ทำหน้าที่ปรับค่าของผลลัพธ์ของแต่ละโหนดในแต่ละชั้น จนได้ผลลัพธ์ตามที่ต้องการ

เทคนิค Naive Bayes [11] เป็นเทคนิคการวิเคราะห์ความน่าจะเป็น สามารถใช้ได้กับตัวแปรที่มีความหลากหลายแล้วนำมาสร้างแบบจำลอง โดยคำนวณการทำนาย [12] จากสมการ (1) และสร้างสมการคำนวณความน่าจะเป็นของการจำแนกประเภทแบบเบย์ดังสมการ (2)

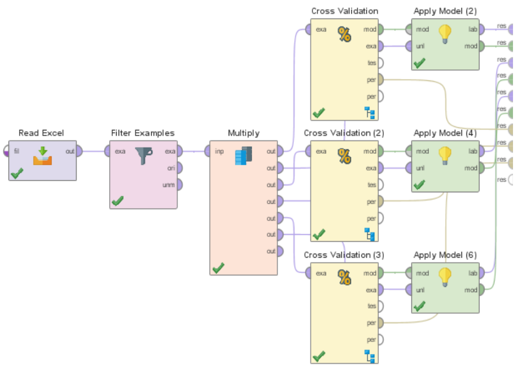
$$P(h|D) = [P(D|h) * P(h)]/P(D) \qquad (1)$$

$$P(d|h) = P(a1,..,aT|h) = \prod_t P(a|h) \qquad (2)$$

เมื่อ P(h) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ h

P(h|D) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ h เมื่อเกิดเหตุการณ์ D จากตัวแปรที่กำหนด

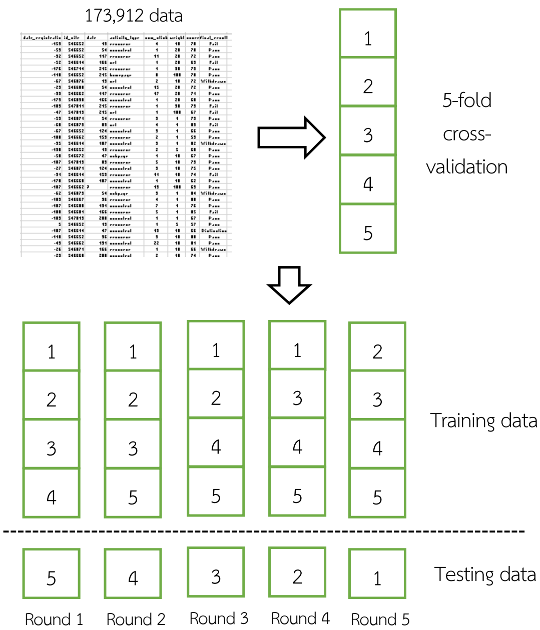
การทำ classification ดังกล่าวเป็นการหาความน่าจะเป็นจาก training data โดยทำการแยกข้อมูลที่ผ่านการเตรียมความพร้อมมาสร้าง training data และ testing data เพื่อใช้ทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test โดยมีรูปแบบการทดลอง ดังรูปที่ 3



รูปที่ 3 รูปแบบการทดลองสร้างโมเดลในการวิจัย

จากรูปที่ 3 ผู้วิจัยใช้เครื่องมือในการวิจัย คือ โปรแกรม Rapid Miner Studio โดยเริ่มการทำงานของโมเดลด้วยการนำเข้าข้อมูลในรูปแบบไฟล์ .xlsx คัดกรองข้อมูลเบื้องต้นเพื่อลบ missing data หรือกรองข้อมูลที่ไม่จำเป็นต่อการทำงานของบางโมเดลด้วยเครื่องมือ Filter Examples จากนั้นเลือกใช้เครื่องมือ Multiply ในการจำลองชุดข้อมูลเพื่อนำใช้ในการสร้างและทดสอบโมเดลต่าง ๆ โดยข้อมูลจะถูกส่งไปเข้าสู่กระบวนการแยกข้อมูลด้วยวิธี Cross-validation Test เพื่อสร้างชุดข้อมูลสำหรับการสร้างโมเดล และชุดข้อมูลสำหรับการทดสอบโมเดล โดยเลือกใช้โมเดลในการจำแนกข้อมูล ประกอบด้วยเทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes และทำการเรียกดูผลของการสร้างโมเดลด้วยเครื่องมือ Apply Model ซึ่งสามารถตรวจสอบประสิทธิภาพด้วยเครื่องมือ Performance

4.5 Evaluation เป็นขั้นตอนทดสอบประสิทธิภาพโมเดล ข้อมูลที่ใช้ในการวิจัยเมื่อผ่านขั้นตอน Business Understanding และ Data Understanding พบว่ามีข้อมูล จำนวน 173,912 ชุด หลังผ่านตรวจสอบความสมบูรณ์ของข้อมูลในขั้นตอน Data Preparation พบว่ามีข้อมูลที่สามารถใช้ในการวิจัยได้ จำนวน 152,850 ชุด ทำการแบ่งข้อมูลที่ใช้ในการวิจัยเพื่อนำไปใช้ในการทดสอบประสิทธิภาพโมเดลด้วยวิธี Cross-validation Test ซึ่งเป็นวิธีที่ได้รับความนิยมเนื่องจากมีความน่าเชื่อถือในการวิจัย โดยผู้วิจัยกำหนดให้มีการแบ่งข้อมูลออกเป็น 5 ส่วน หรือ 5-fold cross-validation โดยการสุ่มข้อมูลเพื่อแบ่งข้อมูลออกเป็น 5 ส่วน ในแต่ละส่วนประกอบด้วยข้อมูล จำนวน 30,570 ชุด และทำการเลือกข้อมูล 4 ส่วน สำหรับใช้ในการสร้างโมเดล และเลือกข้อมูล 1 ส่วน สำหรับใช้ในการทดสอบประสิทธิภาพของระบบ โดยการแบ่งข้อมูลด้วยวิธี Cross-validation Test ที่แบ่งข้อมูลแบบ 5-fold cross-validation ดังรูปที่ 4



รูปที่ 4 การแบ่งข้อมูลด้วยวิธี 5-fold cross-validation

จากรูปที่ 4 แบ่งข้อมูลออกเป็น 5 ส่วน ด้วย Cross-validation Test หรือ 5-fold cross-validation พบว่า

รอบที่ 1 ใช้ข้อมูลส่วนที่ 2,3,4 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 1 เพื่อทดสอบประสิทธิภาพ

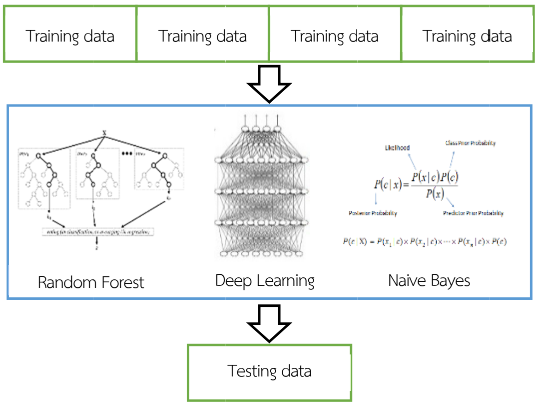
รอบที่ 2 ใช้ข้อมูลส่วนที่ 1,3,4 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 2 เพื่อทดสอบประสิทธิภาพ

รอบที่ 3 ใช้ข้อมูลส่วนที่ 1,2,4 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 3 เพื่อทดสอบประสิทธิภาพ

รอบที่ 4 ใช้ข้อมูลส่วนที่ 1,2,3 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 4 เพื่อทดสอบประสิทธิภาพ

รอบที่ 5 ใช้ข้อมูลส่วนที่ 1,2,3 และ 4 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 5 เพื่อทำการทดสอบประสิทธิภาพ

ทั้งนี้ ในการจำแนกข้อมูลด้วยเทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes ทุกครั้งที่มีการสร้างโมเดลและทดสอบโมเดล จะมีการเลือกข้อมูล 4 ส่วน สำหรับใช้เป็น Training data และเลือกข้อมูล 1 ส่วน สำหรับใช้เป็น Testing data เสมอ โดยทำการสลับให้ข้อมูลในแต่ละส่วนสามารถถูกนำมาเป็นส่วนของ Testing data ได้ โดยมีรูปแบบการทดลองดังรูปที่ 5



รูปที่ 5 รูปแบบการทดลอง

4.6 Deployment เป็นขั้นตอนการปรับใช้ เมื่อทดสอบจนได้โมเดลที่ให้ค่าความแม่นตรง ค่าความระลึก ค่าความแม่นยำ และค่าความถูกต้องที่สามารถยอมรับได้แล้ว สามารถนำรูปแบบการจำแนกข้อมูลที่ได้จากงานวิจัยไปใช้ในการทำนายข้อมูลการเลือกใช้ปัจจัยสนับสนุนการเรียนรู้ ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษาต่อไป

5. ผลการวิจัย

ผลการศึกษาประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา ระหว่างเทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes โดยการทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test ด้วยการแยกข้อมูลสำหรับสร้าง Training data และ Testing data ใช้เกณฑ์การวัดประสิทธิภาพของตัวแบบรู้จำด้วยวิธี Predictive Modeling [13] ซึ่งประกอบด้วยค่าความแม่นตรง (Precision) คือ การวัดความสามารถของระบบในการขจัดข้อมูลที่ไม่เกี่ยวข้องออกไป แสดงถึงความแม่นยำ ค่าความระลึก (Recall) คือ ค่าความสามารถของระบบในการเลือกข้อมูล ค่าความแม่นยำ (Accuracy) คือ ค่าประสิทธิภาพการพยากรณ์ของตัวแบบโดยรวม และค่าถ่วงดุล (F-Measure) พบว่า เทคนิคในการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษาที่เลือกใช้ ให้ผลลัพธ์ดังตารางที่ 3

ตารางที่ 3 ผลการวัดประสิทธิภาพการจำแนกข้อมูล

Random Forest: Accuracy: 86.78%				
	True Fail	true Pass	true Withdrawn	class precision
pred. Fail	80970	0	0	80.67%
pred. Pass	0	84290	0	82.61%
pred. Withdrawn	0	0	90400	86.74%
Class recall	88.04%	85.25%	82.96%	
Deep Learning: Accuracy: 99.01%				
pred. Fail	92730	0	0	88.69%
pred. Pass	0	94380	0	90.17%
pred. Withdrawn	0	0	99710	92.95%
Class recall	99.08%	98.33%	97.77%	
Naive Bayes: Accuracy: 88.55%				
pred. Fail	80182	0	0	80.09%
pred. Pass	0	81854	0	86.18%
pred. Withdrawn	0	0	95204	90.35%
Class recall	83.37%	87.92%	91.37%	

จากตารางที่ 3 ผลการวัดประสิทธิภาพการจำแนกข้อมูลระหว่างเทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes พบว่า

เทคนิค Random Forest ให้ค่าความแม่นตรง 83.34% ค่าความระลึก 85.41% ค่าความแม่นยำ 86.78% และค่าถ่วงดุล 0.367

เทคนิค Deep Learning ให้ค่าความแม่นตรง 90.60% ค่าความระลึก 98.39% ค่าความแม่นยำ 99.01% และค่าถ่วงดุล 0.321

เทคนิค Naive Bayes ให้ค่าความแม่นตรง 85.54% ค่าความระลึก 87.55% ค่าความแม่นยำ 88.55% และค่าถ่วงดุล 0.294

สามารถสรุปตามแนวคิด Predictive Modeling ดังตารางที่ 4

ตารางที่ 4 สรุปผลการวัดประสิทธิภาพ

model	Accuracy	Recall	Precision	MAE
Random Forest	86.78%	85.41%	83.34%	0.367
Deep Learning	99.01%	98.39%	90.60%	0.321
Naive Bayes	88.55%	87.55%	85.54%	0.294

จากตารางที่ 4 สรุปได้ว่า เทคนิคที่ใช้ในการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา ที่มีประสิทธิภาพสูงสุด คือ Deep Learning โดยมีค่าความแม่นตรง 90.60% ค่าความระลึก 98.39% ค่าความแม่นยำ 99.01% และค่าถ่วงดุล 0.321

6. อภิปรายผล

การหาประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา ด้วยการประเมินและเปรียบเทียบประสิทธิภาพของตัวแบบระหว่าง เทคนิค Random Forest เทคนิค Deep Learning และเทคนิค Naive Bayes โดยการทดสอบประสิทธิภาพด้วยวิธี Cross-validation Test ด้วยการแยกข้อมูลสำหรับสร้าง Training data และ Testing data เมื่อใช้เกณฑ์การวัดประสิทธิภาพของตัวแบบรู้จำด้วยวิธี Predictive Modeling พบว่า

ผลการวิเคราะห์ข้อมูลที่เกี่ยวข้องกับเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา พบว่า เมื่อผ่านขั้นตอน Business Understanding และขั้นตอน Data Understanding มีข้อมูลที่มีความสัมพันธ์กันและส่งผลต่อผลการเรียนของผู้เรียน คือ ข้อมูลประเภทการประเมิน ข้อมูลจำนวนวันเข้าสู่ระบบเพื่อใช้งาน ข้อมูลจำนวนวันที่ส่งผลการร่วมกิจกรรม ข้อมูลส่วนของข้อมูลที่เข้าเรียนรู้ ข้อมูลโมดูลย่อยทีเข้าใช้งาน ข้อมูลประเภทกิจกรรมที่เข้าใช้งาน ข้อมูลจำนวนครั้งของการคลิกเพื่อทำกิจกรรม ข้อมูลค่าน้ำหนักของการร่วมกิจกรรม ข้อมูลระดับคะแนน และข้อมูลผลการประเมิน ซึ่งมีข้อมูล จำนวน 173,912 ชุด หลังผ่านตรวจสอบความสมบูรณ์ของข้อมูลในขั้นตอน Data Preparation จะมีข้อมูลที่สามารถใช้ในการวิจัยได้จำนวน 152,850 และเมื่อทำการแบ่งข้อมูลด้วยวิธี Cross-validation Test ออกเป็น 5 ส่วน หรือ 5-fold cross-validation โดยการสุ่มข้อมูลเพื่อแบ่งข้อมูลในแต่ละส่วนประกอบด้วยข้อมูล จำนวน 30,570 ชุด และทำการเลือกข้อมูล 4 ส่วน สำหรับใช้ในการสร้างโมเดล และเลือกข้อมูล 1 ส่วน สำหรับใช้ในการทดสอบประสิทธิภาพของระบบ ทำให้มีข้อมูลจำนวนมากพอสำหรับการสร้างโมเดล และใช้สำหรับทดสอบประสิทธิภาพ ซึ่งการสุ่มข้อมูลและแบ่งส่วนการทำงานดังกล่าวทำให้ผลการวิจัยมีความน่าเชื่อถือ

ผลการวิจัย พบว่า เทคนิคที่ใช้ในการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ด้วยสื่อการสอนอิเล็กทรอนิกส์ในระบบเปิดของผู้เรียนระดับอุดมศึกษา ที่มีประสิทธิภาพสูงสุด คือ Deep Learning โดยมีค่าความแม่นตรง 90.60% ค่าความระลึก 98.39% ค่าความแม่นยำ 99.01% และค่าถ่วงดุล 0.321 เป็นระดับการประเมินที่สามารถยอมรับได้

ข้อเสนอทางการวิจัย สามารถนำรูปแบบการจำแนกข้อมูลด้วยเทคนิค Deep Learning ที่มีการแบ่งกลุ่มข้อมูลด้วยวิธี Cross validation Test ไปใช้ทดลองกับข้อมูลอื่นที่มีจำนวนข้อมูลปริมาณมาก หรือนำไปเปรียบเทียบกับเทคนิคการจำแนกข้อมูลแบบอื่น

7. เอกสารอ้างอิง

[1] W. Srisa-an, The Development of Sukhothai Thammathirat Open University from a Concept to a Reality, University of Minnesota, 2006.

[2] D. Thamsiri and P. Meesad, "Ensemble Data Classification Based on Decision Tree, Artificial Neuron Network and Support Vector Machine Optimized by Genetic Algorithm," *The Journal of KMUTNB.* vol. 21, no. 2, pp. 239 - 303, 2011. (in Thai)

[3] S. Satiman, P. Meesad and P. Voraboot, "Ensemble for highly imbalanced text classification base on deep learning with feature selection and data balancing techniques," in *The 13th National Conference on Computing and Information Technology (NCCIT 2017)*, pp. 20 - 25, 2017.

[4] N. Chirawichitchai and N. Panawas, "Sentiment Classification Using Machine Learning Techniques," *Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE 2011)*, Faculty of ICT, Mahidol University, Nakhon Pathom, 2011.

[5] W. Jiansheng, et al. "Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature," *Bioinformatics*, vol. 25, no. 1, pp. 30 – 35, 2009.

[6] W. Dayong, et al., Deep Learning for Identifying Metastatic Breast Cancer, CSAIL. Massachusetts Institute of Technology, 2016.

[7] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no.4, pp. 13 – 22, 2000.

[8] L. Breiman, "Deep Learning," *Machine Learning*, vol. 45, pp. 5 – 32, 2001.

[9] L. Breiman. "Random Forest," *Machine Learning*, vol. 45, pp. 5 - 32, 2001.

[10] C. Wick, "Deep Learning," *Informatik Spektrum*, vol. 40, pp. 103 – 107, 2017.

[11] I. W. Geoffrey, "Naïve Bayes," *Encyclopedia of Machine Learning and Data Mining*, pp. 895 - 896, 2017.

[12] J. R. Richard and M. W. Geatz, Data Mining a Tutorial-Based Primer, Pearson Education Inc., 2003.

[13] M. Brame, Principles of Data Mining, Springer-Verlag London Limited: London, 2007.