การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง

อุบลรัตน์ ศิริสุขโภคา[1*] และ จรัญ แสนราช[2]

บทคัดย่อ

บทความนี้นำเสนอผลการประเมินและเปรียบเทียบประสิทธิภาพการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาเมื่อเรียนรู้ในสภาพแวดล้อมการเรียนรู้เสมือนจริง 3 เทคนิค คือ เทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes โดยการนำข้อมูลเกี่ยวกับการลงทะเบียนเรียนและผลการเรียนของผู้เรียนที่เรียนรู้ด้วยการโต้ตอบกับสภาพแวดล้อมการเรียนรู้เสมือนจริง จำนวน 8 แอททริบิวต์ มีข้อมูลจำนวน 32,950 ชุด ซึ่งผ่านการตรวจสอบและปรับให้ข้อมูลมีความสมบูรณ์ โดยแบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธี Split Test ประกอบด้วย ข้อมูลส่วนที่หนึ่ง จำนวน 26,360 ชุด คิดเป็น 80% ของข้อมูลทั้งหมด เป็นข้อมูลที่ใช้ในการสร้างโมเดลและข้อมูลส่วนที่สอง จำนวน 6,590 ชุด คิดเป็น 20% ของข้อมูลทั้งหมด เป็นข้อมูลที่ใช้ในการทดลองความถูกต้องและแม่นยำของโมเดลที่สร้าง ผลการวิจัย พบว่า เทคนิคที่ใช้ในการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริงที่มีประสิทธิภาพสูงสุด คือ k-NN (เมื่อกำหนดให้ k = 3) โดยมีค่าความแม่นตรง 97.87% ค่าความระลึก 97.90% ค่าความแม่นยำ 97.93% และค่าถ่วงดุล 0.352

คำสำคัญ: การจำแนกข้อมูล, การแบ่งข้อมูลเพื่อทดสอบโมเดล, ต้นไม้ตัดสินใจ, k-NN, เนอีฟเบย์

[1] ผู้ช่วยศาสตราจารย์ สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[2] ผู้ช่วยศาสตราจารย์ ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
[*] ผู้นิพนธ์ประสานงาน โทร. +669 4469 1651 อีเมล: ubonrat76@gmail.com

_20-0293(001-230).indd 131 8/5/2563 BE 14:19

Technical Education Journal : King Mongkut's University of Technology North Bangkok
Vol. 10 No. 3 September – December, 2019

A Comparison of the Efficiency of Data Classification in Learning Factors of Tertiary Level Students in a Virtual Learning Environment

Ubonrat Sirisukpoca[1*] and Charan Sanrach[2]

Abstract

This research presents the result of evaluation and comparison of the efficiency of data classification in learning support factors of tertiary level students in a virtual learning environment. Three primary techniques were compared in this research, including Decision Tree technique, k-NN technique, and Naive Bayes technique. The were 8 attributes with 32,950 datasets of the data which were included student registration and the learning outcome of the students who interactively learned in a virtual learning environment. The data was divided into 2 parts by Split Test method. The first part which was comprised of 26,360 datasets (approximately 80 percent of the total data) was used to generate a model while the second part which was included 6,590 datasets (approximately 20 percent of the total data) was utilized to examine the precision and accuracy of the model. The results revealed that the most efficient technique, used to classify the learning factors of the tertiary students in a virtual learning environment, was k-NN technique (k = 3) with 97.87 percent of precision, 97.90 percent of recall, 97.93 percent of accuracy, and 0.352 F-measure.

Keywords: Classification, Split Test, Decision Tree, k-NN, naive bayes

[1] Assistance Professor, Department of Computer Science Faculty of Science and Technology, Nakhon Pathom Rajabhat University

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

[*] Corresponding Author Tel. +669 4469 1651 email: ubonrat76@gmail.com

_20-0293(001-230).indd 132 8/5/2563 BE 14:19

1. บทนำ

การศึกษาในระดับอุดมศึกษามีความสำคัญต่อการพัฒนาบุคลากรในระดับประเทศ การจัดการเรียนรู้ที่มีประสิทธิภาพจึงเป็นสิ่งสำคัญ เนื่องจากการเรียนรู้ในระดับอุดมศึกษาแตกต่างจากการเรียนรู้ในระดับอื่น ๆ ซึ่งมีผู้เรียนที่มีประสบการณ์และพื้นฐานความรู้แตกต่างกัน การจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนในระดับอุดมศึกษาจึงเป็นสิ่งจำเป็นเพื่อใช้ในการวางแผนให้ผู้เรียนในระดับอุดมศึกษาสามารถสำเร็จการศึกษาได้อย่างมีประสิทธิภาพ อีกทั้งมหาวิทยาลัยหลายแห่งมีการจัดสภาพแวดล้อมการเรียนรู้แบบเสมือนจริงเพื่อให้ผู้เรียนเกิดความเข้าใจเนื้อหาและเข้าร่วมกิจกรรมตรงตามความต้องการได้อย่างแท้จริง การเปรียบเทียบประสิทธิภาพการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริงจึงมีความสำคัญเพื่อนำไปใช้ในการวิจัยระดับสูงต่อไป

จากความสำคัญดังกล่าวผู้วิจัยได้ศึกษาค้นคว้าและทบทวนงานวิจัยที่เกี่ยวข้อง ได้แก่ งานวิจัยของ P. Dunkuntod [1] พัฒนาระบบช่วยแนะนำสำหรับการดูแลช่วยเหลือนักเรียน ด้วยการทดสอบและวัดประสิทธิภาพการจำแนกข้อมูล โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคต้นไม้ตัดสินใจ และเทคนิคเนอีฟเบย์ พบว่า เทคนิคซัพพอร์ตเวกเตอร์แมชชีน มีค่าความถูกต้องในการจำแนกข้อมูลดีที่สุด คือ 99.75% ซึ่งสามารถนำมาพัฒนาระบบช่วยแนะนำ สำหรับการดูแลช่วยเหลือนักเรียนและแก้ไขปัญหานักเรียนในด้านต่าง ๆ ได้อย่างมีประสิทธิภาพโดยผู้ใช้มีความพึงพอใจเฉลี่ยเท่ากับ 4.02 มีค่าส่วนเบี่ยงเบน มาตรฐาน 0.08 งานวิจัยของ D. Thammasiri and P. Meesad [2] ได้นำเสนอวิธีจำแนกข้อมูลด้วยเทคนิคการร่วมกันตัดสินใจจากหลายโมเดล โดยเลือกตัวแทนที่เหมาะสมด้วยขั้นตอนเชิงพันธุกรรม เพื่อให้ความถูกต้องในการจำแนกข้อมูลและตัดสินใจร่วมกันด้วยการโหวตเสียงข้างมาก พบว่า ประสิทธิภาพการเลือกใช้ตัวจำแนกร่วมกันให้ประสิทธิภาพสูงสุดเกิดจากเทคนิคต้นไม้ตัดสินใจ เทคนิคโครงข่ายประสาทเทียม และเทคนิคซัพพอร์ตเวกเตอร์แมชชีน โดยมีสัดส่วนของจำนวนโมเดลแตกต่างกันตามข้อมูลที่ใช้ในการจำแนก งานวิจัยของ N. Chirawichitchai and N. Panawas [3] ได้นำเสนอผลการจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคต้นไม้ตัดสินใจ เทคนิคเนอีฟเบย์ และเทคนิคเคเนียเรสเนเบอร์ พบว่า เทคนิคซัพพอร์ตเวกเตอร์แมชชีนมีความเหมาะสมและมีประสิทธิภาพสูงสุด เท่ากับ 86.30 งานวิจัยของ N. Nai-arun and P. Sittidech [4] นำเสนอการจำแนกข้อมูลผู้ป่วยโรคหัวใจขาดเลือดออกจากโรคหัวใจแบบอื่น โดยเปรียบเทียบประสิทธิภาพของเทคนิคซัพพอร์ตเวกเตอร์แมชชีนและเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ พบว่า เทคนิคซัพพอร์ตเวกเตอร์แมชชีนมีความเหมาะสมในการจำแนกข้อมูลดีกว่า โดยมีค่าความถูกต้องเท่ากับ 96.46 และงานวิจัยของ P. Budake and S. Boonlue [5] ทำการสร้างตัวแบบพยากรณ์เพื่อเปรียบเทียบประสิทธิภาพด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคต้นไม้ตัดสินใจ และเทคนิคโครงข่ายประสาทเทียมแบบย้อนกลับเพื่อคัดเลือกตัวแบบพยากรณ์ที่มีประสิทธิภาพสูงสุด โดยใช้ชุดข้อมูลระเบียน 18 คุณลักษณะ พบว่า เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเหมาะสมที่สุดสำหรับการพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษา โดยมีความแม่นยำสูงที่สุดเท่ากับ 89.13 มีผลการประเมินค่าความแม่นเท่ากับ 0.878 ค่าความระลึกเท่ากับ 0.891 และค่าประสิทธิภาพโดยรวมเท่ากับ 0.860

จากการค้นคว้าและทบทวนงานวิจัยที่เกี่ยวข้อง ผู้วิจัยจึงมีแนวคิดในการวิจัยเกี่ยวกับการเปรียบเทียบประสิทธิภาพการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง โดยใช้ข้อมูลจากฐานข้อมูลการเรียนรู้ที่ประกอบด้วยข้อมูลเกี่ยวกับหลักสูตร ผู้เรียนและการโต้ตอบกับสภาพแวดล้อมเสมือนจริง เพื่อวางแผนการทำนายผลจากเรียนรู้ของผู้เรียนในอนาคต ด้วยการจำแนกประเภทข้อมูล (Classification) โดยเลือกใช้เทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes และใช้วิธี Split Test ในการแยกข้อมูลสำหรับสร้าง Training data และ Testing data เพื่อใช้ในการทดสอบประสิทธิภาพของโมเดล ตามลำดับ

2. วัตถุประสงค์ของการวิจัย

เพื่อประเมินและเปรียบเทียบประสิทธิภาพการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริงระหว่างเทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes และ 2) เพื่อหาเทคนิคในการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริงที่เหมาะสมที่สุด

3. สมมติฐานการวิจัย

ได้แนวทางจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริงที่มีประสิทธิภาพ หลังการเปรียบเทียบระหว่างเทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes โดยแบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธี Split Test ตัวแปรซึ่งเป็นแอททริบิวต์ที่ใช้ในการจำแนกข้อมูลจำนวน 8 แอททริบิวต์ ทำให้พบเทคนิคในการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับ อุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริงที่เหมาะสมที่สุด ที่มีความถูกต้อง

4. วิธีดำเนินการวิจัย

กลุ่มตัวอย่างที่ใช้ในการวิจัย คือ ข้อมูลที่ใช้ในการวิจัยจำนวน 32,950 ชุด เครื่องมือที่ใช้ในการวิจัย ได้แก่ เครื่องมือในการรวบรวมข้อมูลเบื้องต้น คือ Microsoft Excel และเครื่องมือในการวิเคราะห์ข้อมูล คือ Rapid miner โดยผู้วิจัยเลือกใช้กระบวนการทำงาน Cross-Industry Standard Process for Data Mining หรือ CRISP-DM [6] แบ่งขั้นตอนดำเนินการวิจัยเป็น 6 ขั้นตอน ประกอบด้วย 1) Business Understanding 2) Data Understanding 3) Data Preparation 4) Modeling 5) Evaluation และ 6) Deployment ดังรูปที่ 1

รูปที่ 1 แผนภาพกระบวนการ CRISP-DM [7]

จากรูปที่ 1 งานวิจัยดำเนินการตามกระบวนการ CRISP-DM [7] ดังนี้

4.1 Business Understanding เป็นขั้นตอนการเข้าใจปัญหาและแปลงปัญหา เริ่มจากชุดข้อมูลตัวอย่างที่ใช้ในการทดสอบเป็นข้อมูลจากฐานข้อมูลการเรียนรู้ของมหาวิทยาลัยเปิดที่แบ่งการจัดการเรียนการสอนออกเป็น 2 ภาคเรียน ซึ่งเริ่มเรียนในช่วงเดือนกุมภาพันธ์และเดือนตุลาคมตามลำดับ

4.2 Data Understanding เป็นขั้นตอนทำความเข้าใจข้อมูล โดยการเก็บรวบรวมข้อมูล ตรวจสอบและวิเคราะห์ความสมบูรณ์ของข้อมูล พบว่า ชุดข้อมูลที่สามารถใช้ในการวิจัย ประกอบด้วยส่วนข้อมูลสำคัญ ได้แก่ ข้อมูลหลักสูตร ข้อมูลผู้เรียน ข้อมูลการเรียนแบบโต้ตอบกับสภาพแวดล้อมเสมือนจริง โดยผู้วิจัยเลือกแอททริบิวต์ข้อมูลสำหรับใช้ในการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง จำนวน 7 แอททริบิวต์ และเลือกข้อมูลผลการประเมินเป็นผลลัพธ์การจำแนกข้อมูล 1 แอททริบิวต์ รวม 8 แอททริบิวต์ ดังตารางที่ 1

ตารางที่ 1 แอททริบิวต์ที่ใช้ในการจำแนกข้อมูล

แอททริบิวต์	ประเภทข้อมูล
highest_education	ระดับการศึกษาสูงสุดของผู้เรียน
age_band	ช่วงอายุของผู้เรียน
assessment_type	ประเภทการประเมิน
date_submitted	จำนวนครั้งในการเข้าใช้งานสื่อ
sum_click	จำนวนการคลิกใช้งานสื่อแบบโต้ตอบ
weight	ค่าน้ำหนักของการประเมิน
score	ระดับคะแนนจากการประเมิน
final_result	ผลการประเมิน

4.3 Data Preparation เป็นขั้นตอนการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมให้เป็นข้อมูลที่สามารถนำไปวิเคราะห์ต่อไปได้ เริ่มจากผู้วิจัยทำความเข้าใจรายละเอียดของแต่ละแอททริบิวต์ ดังตารางที่ 2

ตารางที่ 2 ประเภทข้อมูลของแต่ละแอททริบิวต์

แอททริบิวต์	ประเภทข้อมูล
highest_education	polynominal
age_band	polynominal
assessment_type	polynominal
date_submitted	integer
sum_click	integer
weight	integer
score	integer
final_result	polynominal

จากตารางที่ 2 เมื่อทำการตรวจสอบรายละเอียดของแต่ละเรคคอร์ด พบว่า ระดับการศึกษา ช่วงอายุของผู้เรียน ประเภทการประเมิน และผลการประเมิน เป็นข้อมูลประเภท polynomial จำนวนครั้งในการเข้าใช้งานสื่อ จำนวนการคลิกใช้งานสื่อแบบโต้ตอบ ค่าน้ำหนักของการประเมิน และระดับคะแนนที่ได้รับการประเมินเป็นข้อมูลประเภท integer

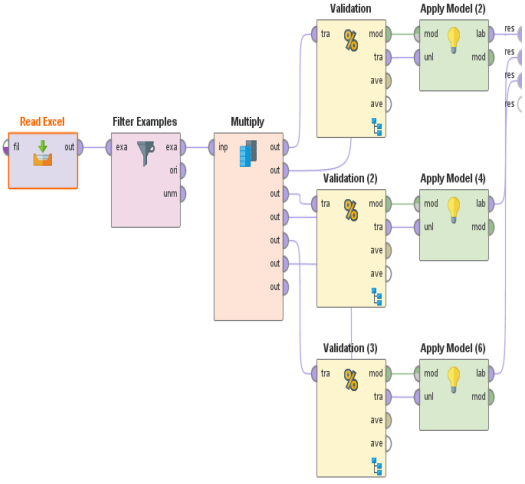
จากนั้น ทำการตรวจสอบความสมบูรณ์ของข้อมูลในแต่ละระเบียน และทำข้อมูลให้ถูกต้อง (data cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป เป็นต้น และกำหนดการแทนค่าข้อมูลของแต่ละแอททริบิวต์ให้ถูกต้องตามประเภทของข้อมูล โดยมีตัวอย่างข้อมูล ดังแสดงในรูปที่ 2

highest_education	age_band	assessment_type	date_submitted	sum_click	weight	score	final_result
A Level or Equivalent	35-55	TMA	18	4	10	78	Fail
Lower Than A Level	35-55	TMA	22	1	20	70	Pass
A Level or Equivalent	0-35	TMA	17	11	20	72	Pass
HE Qualification	0-35	TMA	26	1	20	69	Fail
A Level or Equivalent	35-55	TMA	19	1	30	79	Distinction
HE Qualification	0-35	Exam	20	8	100	70	Pass
A Level or Equivalent	35-55	TMA	18	2	10	72	Fail
A Level or Equivalent	0-35	TMA	19	15	20	72	Pass
HE Qualification	0-35	TMA	9	17	20	71	Pass
A Level or Equivalent	35-55	TMA	18	1	20	68	Pass
A Level or Equivalent	0-35	TMA	19	1	30	73	Fail
A Level or Equivalent	0-35	Exam	18	1	100	67	Fail
Lower Than A Level	35-55	CMA	17	3	1	73	Pass
A Level or Equivalent	35-55	CMA	19	4	1	83	Fail
A Level or Equivalent	35-55	CMA	17	3	1	66	Withdrawn
A Level or Equivalent	35-55	CMA	19	2	1	59	Pass
A Level or Equivalent	35-55	CMA	19	3	1	82	Withdrawn
Lower Than A Level	0-35	TMA	22	2	5	60	Pass
A Level or Equivalent	0-35	TMA	18	1	18	67	Pass
Lower Than A Level	35-55	TMA	18	5	18	73	Pass
A Level or Equivalent	0-35	TMA	21	3	18	75	Pass
HE Qualification	0-35	TMA	17	11	18	74	Fail

รูปที่ 2 ตัวอย่างข้อมูลสำหรับสร้างและทดสอบโมเดล

จากรูปที่ 2 กำหนดให้ข้อมูล highest_education, age_band, assessment_type, date_submitted, sum_click, weight, score เป็นแอททริบิวต์ และข้อมูล final_result เป็น label ที่ใช้ในการหาโมเดลที่เหมาะสม โดยข้อมูลที่อยู่ในรูปแบบ polynomial จะถูกปรับให้อยู่ในรูปแบบ Integer เช่น 0-35 ปรับค่าเป็น 1 และ 36-55 ปรับค่าเป็น 2 เป็นต้น เพื่อให้เหมาะกับโมเดลที่ใช้ในการวิจัย

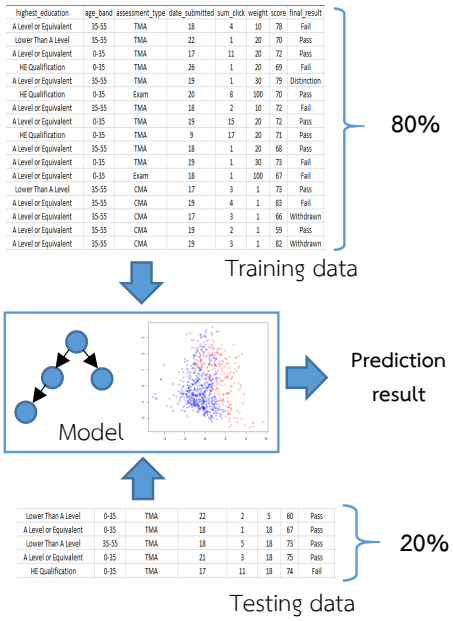
4.4 Modeling เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล จากขั้นตอน Data Preparation พบว่า ประเภทของข้อมูลที่เลือกใช้ ประกอบด้วยข้อมูล Integer และ ข้อมูล nominal ผู้วิจัยเลือกใช้การจำแนกประเภทข้อมูล(Classification) เป็นโมเดลในการทดลอง ประกอบด้วย เทคนิค Decision Tree [8] เป็นการทำนายโดยใช้การตัดสินใจเลือกด้วยเงื่อนไขต่าง ๆ ในลักษณะเหมือนต้นไม้ที่แตกเส้นทางสำหรับการตัดสินใจจนกว่าจะได้สิ่งที่ดีที่สุด เทคนิค k-NN [9] เป็นเทคนิคการแบ่งกลุ่มด้วยการเลือกค่าที่มีความใกล้เคียงกันโดยใช้ค่า K และระยะห่างของข้อมูลในการแบ่งกลุ่มข้อมูล และเทคนิค Naive Bayes [10] เป็นการหาความน่าจะเป็นในการแบ่งกลุ่มซึ่งเป็นการทำ classification จาก training data โดยมีรูปแบบการทดลองดังรูปที่ 3



รูปที่ 3 รูปแบบการทดลองสร้างโมเดลในการวิจัย

จากรูปที่ 3 นำเข้าข้อมูลในรูปแบบไฟล์ .xlsx คัดกรองข้อมูลเบื้องต้นด้วย Filter Examples เพื่อลบ missing data และจำลองชุดข้อมูลเพื่อใช้ในการทดลองในโมเดลต่าง ๆ ด้วย Multiply เมื่อเข้าสู่ขั้นตอนของโมเดลในการจำแนกข้อมูล ประกอบด้วย เทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes ในแต่ละเทคนิคจะทำการแยกข้อมูลออกเป็น 2 ส่วน ด้วยวิธี Split Validation และทำการเรียนรู้ชุดข้อมูลเพื่อจำแนกข้อมูลตามลำดับ

4.5 Evaluation เป็นขั้นตอนการทดสอบประสิทธิภาพโมเดล เมื่อตรวจสอบความสมบูรณ์ของข้อมูลในขั้นตอน Data Preparation พบว่า มีข้อมูลที่ใช้ในการวิจัยจำนวน 32,950 ชุด เมื่อเลือกโมเดลในการจำแนกข้อมูลตามขั้นตอน Modeling จะทำการแบ่งข้อมูลเพื่อนำไปใช้ในการทดสอบประสิทธิภาพโมเดลด้วยวิธี Split Test ซึ่งเป็นวิธีที่ได้รับความนิยมเหมาะสำหรับชุดข้อมูลที่มีขนาดใหญ่ เนื่องจากใช้เวลาในการสร้างโมเดลน้อย โดยการสุ่มข้อมูลเพื่อแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลส่วนที่หนึ่ง จำนวน 26,360 ชุด คิดเป็น 80% ของข้อมูลทั้งหมดสำหรับใช้ในการสร้างโมเดลและข้อมูลส่วนที่สอง จำนวน 6,590 ชุด คิดเป็น 20% ของข้อมูลทั้งหมด โดยการแบ่งข้อมูลด้วยวิธี Split Test แสดงได้ดังรูปที่ 4



รูปที่ 4 การแบ่งข้อมูลด้วยวิธี Split Test

จากรูปที่ 4 การสุ่มเพื่อแบ่งข้อมูลสำหรับสร้างโมเดลและทดสอบโมเดลในสัดส่วน 80 : 20 ประกอบด้วยข้อมูลสำหรับใช้เป็น Training data จำนวน 80% และข้อมูลสำหรับใช้เป็น Testing data จำนวน 20%

4.6 Deployment เป็นขั้นตอนการปรับใช้ เมื่อทดสอบโมเดลจนได้ค่าความแม่นตรง ค่าความระลึก และค่าความแม่นยำ จนสามารถยอมรับได้แล้ว สามารถนำผลลัพธ์จากโมเดลไปใช้ในการทำนายและบอกถึงความเป็นไปได้ในการผ่านการเรียนของผู้เรียนที่ลงทะเบียนเรียนในรายวิชาที่เรียนในสภาพแวดล้อมการเรียนรู้เสมือนจริงได้

5. ผลการวิจัย

5.1 ผลการศึกษาประสิทธิภาพการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง สามารถประเมินและเปรียบเทียบประสิทธิภาพของตัวแบบระหว่างเทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes โดยใช้เกณฑ์การวัดประสิทธิภาพของตัวแบบรู้จำด้วยวิธี Predictive Modeling [11] ซึ่งประกอบ ด้วยค่าความแม่นตรง (Precision) ค่าความระลึก (Recall) ค่าความแม่นยำ

(Accuracy) และค่าถ่วงดุล (F-Measure) พบว่า เทคนิคในการจำแนกข้อมูลที่ใช้ให้ผลลัพธ์ดังตารางที่ 3

ตารางที่ 3 ผลการวัดประสิทธิภาพ

Decision tree: Accuracy: 63.84%

	true Fail	true Pass	true Withdrawn	class precision
pred. Fail	8860	0	0	55.57%
pred. Pass	0	9980	0	62.61%
pred. Withdrawn	0	0	19708	98.73%
Class recall	98.04%	72.92%	35.96%	

k-NN: Accuracy: 97.93%

	true Fail	true Pass	true Withdrawn	class precision
pred. Fail	19082	0	0	97.69%
pred. Pass	0	19323	0	98.37%
pred. Withdrawn	0	0	18903	97.56%
Class recall	97.72%	97.91%	98.07%	

*k = 3

Naive Bayes: Accuracy: 59.86%

	true Fail	true Pass	true Withdrawn	class precision
pred. Fail	6742	0	0	47.78%
pred. Pass	0	11504	0	64.09%
pred. Withdrawn	0	0	14237	71.19%
Class recall	73.83%	57.70%	60.04%	

จากตารางที่ 3 ผลการวัดประสิทธิภาพการจำแนกข้อมูลระหว่างเทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes พบว่า เทคนิค k-NN (k=3) ให้ค่าความแม่นตรงสูงที่สุด (pred. Fail = 97.69%, pred. Pass = 98.37%, pred. Withdrawn = 97.56%) ค่าความระลึกสูงที่สุด (true Fail = 97.72%, true Pass = 97.91%, true Withdrawn = 98.07%) และค่าความแม่นยำสูงที่สุด (Accuracy : 97.93%) สามารถสรุปได้ ดังตารางที่ 4

ตารางที่ 4 สรุปผลการวัดประสิทธิภาพ

model	Accuracy	Recall	Precision	MAE
Decision tree	63.84%	68.97%	72.30%	0.323
k-NN	97.93%	97.90%	97.87%	0.352
Naive Bayes	59.86%	63.85%	61.02%	0.252

5.2 จากตารางที่ 4 สรุปได้ว่า เทคนิคที่ใช้ในการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง ที่มีประสิทธิภาพสูงสุด คือ k-NN โดยมีค่าความแม่นตรง 97.87% ค่าความระลึก 97.90% ค่าความแม่นยำ 97.93% และค่าถ่วงดุล 0.352 ซึ่งสามารถแสดงผลการทำนายด้วยการจำแนกข้อมูล ดังรูปที่ 5

Row No.	final_result	prediction(final_result)	date_submitted	sum_click	weight	score
1	Fail	Fail	18	4	10	78
2	Pass	Pass	22	1	20	70
3	Pass	Pass	17	11	20	72
4	Fail	Fail	26	1	20	69
5	Distinction	Distinction	19	1	30	79
6	Pass	Pass	20	8	100	70
7	Fail	Fail	18	2	10	72
8	Pass	Pass	19	15	20	72
9	Pass	Pass	9	17	20	71
10	Pass	Pass	18	1	20	68

รูปที่ 5 ผลทำนายด้วยการจำแนกข้อมูลเทคนิค k-NN

จากรูปที่ 5 การทำนาย prediction (final_result) ด้วยการจำแนกข้อมูลเทคนิค k-NN ได้ผลตรงกับผลลัพธ์จริง (final_result) ในระดับมาก สามารถยอมรับผลลัพธ์ได้

6. สรุปและอภิปรายผลการวิจัย

การหาประสิทธิภาพการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง ด้วยการประเมินและเปรียบเทียบประสิทธิภาพของตัวแบบระหว่าง เทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes พบว่า

ผลประเมินและเปรียบเทียบประสิทธิภาพจากการวิเคราะห์ข้อมูลผู้เรียนที่ลงทะเบียนเรียนในระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง ผู้วิจัยเลือกใช้ข้อมูล ระดับการศึกษา ช่วงอายุของผู้เรียน ประเภทการประเมิน จำนวนครั้งในการเข้าใช้งานสื่อ จำนวนการคลิกใช้งานสื่อแบบโต้ตอบ ค่าน้ำหนักของการประเมิน ระดับคะแนนที่ได้รับการประเมิน และผลการประเมิน เป็นข้อมูลในการใช้สร้างโมเดลและทดสอบโมเดล โดยข้อมูลที่เลือกใช้มีประเภทข้อมูลที่หลากหลาย เช่น ประเภท integer เป็นข้อมูลประเภท polynomial เป็นต้น จึงจำเป็นต้องมีการปรับและทำข้อมูลให้ถูกต้อง (data cleaning) และกำหนดการแทนค่าข้อมูลที่สำคัญ

และลบ เรคคอร์ดที่มีข้อมูลไม่สมบูรณ์ ทำให้การทดลองมีประสิทธิภาพ ซึ่งสอดคล้องกับผลการวิจัยของ P. Simalaotao and J. Sanrach [12] ซึ่งมีการปรับและทำข้อมูลให้ถูกต้องก่อนการทดลองเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล

ข้อมูลที่ใช้ในการวิจัยเมื่อผ่านการตรวจสอบและปรับให้ข้อมูลมีความสมบูรณ์แล้ว พบว่ามีข้อมูลจำนวน 32,950 ชุด โดยผู้วิจัยทำการแบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธี Split Test ซึ่งเป็นที่นิยมในการจำแนกข้อมูลที่มีปริมาณมาก ประกอบด้วย

ข้อมูลส่วนที่หนึ่ง จำนวน 26,360 ชุด คิดเป็น 80% ของข้อมูลทั้งหมด เป็นข้อมูลที่ใช้สำหรับใช้ในการสร้างโมเดลตามเทคนิค Decision Tree เทคนิค k-NN และเทคนิค Naive Bayes

ข้อมูลส่วนที่สอง จำนวน 6,590 ชุด คิดเป็น 20% ของข้อมูลทั้งหมด เป็นข้อมูลที่ใช้ในการทดลองความถูกต้องและแม่นยำของโมเดลที่สร้างจากข้อมูลในส่วนที่หนึ่ง

การแบ่งข้อมูลด้วยวิธี Split Test จะได้ข้อมูลที่ใช้ในการสร้างโมเดลและข้อมูลที่ใช้ในการทดสอบโมเดลไม่ใช่ข้อมูลชุดเดียวกัน ทำให้ผลการวิจัยมีความน่าเชื่อถือมากขึ้น

ผลการหาเทคนิคในการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริงที่เหมาะสมที่สุด พบว่า เทคนิคที่ใช้ในการจำแนกปัจจัยสนับสนุนการเรียนรู้ของผู้เรียนระดับอุดมศึกษาในสภาพแวดล้อมการเรียนรู้เสมือนจริง ที่มีประสิทธิภาพสูงสุด คือ k-NN (เมื่อกำหนดให้ k = 3) โดยมีค่าความแม่นตรง 97.87% ค่าความระลึก 97.90% ค่าความแม่นยำ 97.93% และค่าถ่วงดุล 0.352

ข้อเสนอทางการวิจัย สามารถนำรูปแบบการจำแนกข้อมูลด้วยเทคนิค k-NN ไปทดสอบประสิทธิภาพด้วยกระบวนการแบ่งข้อมูลตามวิธี Split Test ที่มีการทำซ้ำหลาย ๆ ครั้ง เพื่อเพิ่มความน่าเชื่อถือ หรือเลือกใช้กระบวนการอื่น เช่น Cross validation Test หรือ Self Consistency Test เป็นต้น

7. เอกสารอ้างอิง

[1] P. Dunkuntod, "Recommender System for Counseling Students using Support Vector Machine Classification Technique Case Study of Mattayom Wat Makutkasat School," Graduate College, King Mongkut's University of Technology North Bangkok, 2010. (in Thai)

[2] D. Thammasiri 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. 12, pp. 239-303, 2011. (in Thai)

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

[4] N. Nai-arun and P. Sittidech, "Using Support Vector Machine and Back Propagation Neural Network for Classification of Ischemic Heart Disease and other Forms of Heart Disease," in Proceedings of The 4th Science Research National Conference, Faculty of Science. Location : Naresuan University, Phitsanulok, 2012. (in Thai)

[5] P. Budake and S. Boonlue, "Graduation Forecasting using Support Vector Machine," Veridian E-Journal Science and Technology Silpakorn University, vol. 1, no. 6, pp. 40-49, 2014. (in Thai)

[6] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," Journal of Data Warehousing, vol. 5, no. 4, pp. 13–22, 2000.

[7] Wikipedia, "Cross Industry Standard Process for Data Mining," 2017. [Online]. Available: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining. [Accessed 9 July 2017].

[8] J. Fürnkranz, "Decision Tree. In: Encyclopedia of Machine Learning," Springer, pp. 263-267, 2010.

[9] A. Mucherino, P.J.Papajorgji and P.M. Pardalos, "k-Nearest Neighbor Classification," in Data Mining in Agriculture. Springer Optimization and Its Applications, New York, NY, Springer, pp. 83-106, 2009.

[10] G. I. Webb, "Naïve Bayes," in Encyclopedia of Machine Learning and Data Mining, pp. 895-896, 2017.

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

[12] P. Simalaotao and J. Sanrach, “A Comparison of the Efficiency of Data Classification in Learning Factors through Open Educational System with Electronic Teaching Aids in Tertiary Level,” Technical Education Journal: King Mongkut’s University of Technology North Bangkok, Vol. 10, No. 1, pp. 66-74, 2019. (in Thai)