การศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล และเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล โดยใช้เทคนิคการทำเหมืองข้อมูล
The Study of Features Affecting the Digital Literacy Test and Comparative Efficiency of Data Classification Using Data Mining Techniques

นรินทร์ จิวิตัน (Narin Jiwitan)*, วรการ ใจดี (Worakarn Jaidee)*,
และ วรรณพร ทีเก่ง (Wannaporn Teekeng)*

Received: December 19, 2022
Revised: April 24, 2023
Accepted: August 16, 2023

*ผู้นิพนธ์ประสานงาน: นรินทร์ จิวิตัน (Narin Jiwitan) อีเมล: narin@rmutl.ac.th

DOI:10.14416/j.it.2024.v1.003

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล โดยใช้เทคนิคการคัดเลือกคุณลักษณะ และเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล โดยใช้เทคนิคการทำเหมืองข้อมูล กระบวนการวิจัยเป็นไปตามขั้นตอนของ CRISP- DM ศึกษาและรวบรวมข้อมูลผลการสำรวจและประเมินสถานภาพการเข้าใจดิจิทัลของสำนักงานคณะกรรมการดิจิทัลเพื่อเศรษฐกิจและสังคมแห่งชาติ ระหว่างวันที่ 1 มกราคม 2564 ถึง 30 กันยายน 2564 จำนวน 12,374 ระเบียน 67 คุณลักษณะ ดำเนินการคัดเลือกข้อมูล กลั่นกรองข้อมูล แปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม ทำการเพิ่มความสมดุลของข้อมูลด้วยวิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE) คัดเลือกคุณลักษณะด้วยเทคนิคการคำนวณหาค่าน้ำหนัก ได้แก่ 1) Chi-Square Statistic 2) Gini Index 3) Gain Ratio 4) Information Gain และ 5) Correlation based นำผลค่าน้ำหนักสูงสุดที่คำนวณได้ 10 ลำดับแรกของแต่ละเทคนิคไปคำนวณความถี่และคัดเลือกคุณลักษณะที่มีความถี่ตั้งแต่ 2 ไปออกแบบและสร้างแบบจำลองการพยากรณ์ จำนวน 5 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคเกรเดียนท์บูตทรีส เทคนิคแรนดอมฟอร์เรสต์ เทคนิคนาอีฟเบย์ และเทคนิคการเรียนรู้เชิงลึก การวัดประสิทธิภาพแบบจำลองการพยากรณ์ด้วยวิธี K-Fold Cross Validation Test แบ่งข้อมูลออกเป็น 10 ส่วน 20 ส่วน และ 30 ส่วน วัดค่าความแม่นยำ ค่าเรียกกลับ ค่าความจำเพาะ ค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม และค่าเฉลี่ยเรขาคณิต ผลการศึกษาคุณลักษณะของผู้ทดสอบที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล พบว่าคุณลักษณะที่มีความถี่ระดับสูงสุดความถี่ 5 มีจำนวน 6 คุณลักษณะ ได้แก่ การใช้สื่อประกาศ/ป้ายฉลากสินค้า การใช้สื่อดิจิทัลเพื่อโซเชียลมีเดีย การใช้สื่อดิจิทัลเพื่อเข้าถึงเว็บไซต์ ปัญหาค่าอินเทอร์เน็ตแพง ปัญหาการเข้าถึงพื้นที่บริการอินเทอร์เน็ต และ ปัญหาข้อความสแปม/โฆษณา ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล พบว่าเทคนิคแรนดอมฟอร์เรสต์มีประสิทธิภาพการจำแนกข้อมูลมากสุดเมื่อแบ่งชุดข้อมูลออกเป็น 30 ส่วน มีค่าความถูกต้อง 76.29% ค่าประสิทธิภาพโดยรวม 76.00% และค่าเฉลี่ยเรขาคณิต 76.28%

คำสำคัญ: ทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล การคัดเลือกคุณลักษณะ การจำแนกข้อมูล การทำเหมืองข้อมูล

Abstract

The purpose of this research is to study the features that affect the results of the test for Digital Literacy by using the Feature Selection technique to compare the efficiency of Data Classification. The research process follows the steps of CRISP-DM, studying and collecting 12,374 records of 67 features of the Office of the National Digital Economy and Society Commission's survey and assessment of the state

* หลักสูตรระบบสารสนเทศทางธุรกิจ คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา
* Business Information Systems department, Faculty of Business Administration and Liberal Arts, Rajamangala University of Technology Lanna.

of digital literacy. Carry out Data Cleaning and Data Transformation into an appropriate format. SMOTE (Synthetic Minority Oversampling Technique) was used to improve the data balance. Features were selected using weighting calculation techniques: 1) Chi-Square Statistic 2) Gini Index 3) Gain Ratio 4) Information Gain and 5) Correlation-based. Using the results of the first 10 calculated highest weights of each technique to calculate the frequency and Features with frequencies higher than 2 were selected to design and create forecasting models of 5 techniques: Decision Tree, Gradient Boosted Trees, Random Forest, Naïve Bayes, and Deep Learning. Evaluation of the forecasting model using the K-Fold Cross Validation Test method, dividing the data into 10 folds, 20 folds, and 30 folds, measuring precision, recall, specificity, accuracy, F1-measure, and G-means. The results of the Study Feature Affecting the Digital Literacy Test revealed that there were 6 features with the highest frequency, frequency 5, as follows: the use of advertising media/product labels, using digital media for social media, using digital media to access websites, the problem of expensive internet service fees, problems accessing internet service areas and problems with spam/advertising messages. The comparative results of data classification efficiency test results of digital literacy found that the random forest technique was the most effective in data classification. When dividing the dataset into 30 parts, the accuracy was 76.29%, the overall efficiency was 76.00%, and the geometric mean was 76.28%.

Keywords: Digital Literacy, Feature Selection, Data Classification, Data Mining.

1. บทนำ

ในศตวรรษที่ 21 โลกได้เข้าสู่ยุคระบบเศรษฐกิจและสังคมดิจิทัล ส่งผลกระทบให้เกิดการเปลี่ยนแปลงโครงสร้างรูปแบบกิจกรรมทางเศรษฐกิจ กระบวนการผลิต การค้า การบริการ และกระบวนการทางสังคมอย่างรุนแรง [1] ทิศทางการพัฒนาประเทศไทย ตามนัยของรัฐธรรมนูญแห่งราชอาณาจักรไทย พุทธศักราช 2560 ร่างกรอบยุทธศาสตร์ชาติระยะ 20 ปี (พ.ศ. 2560 – 2579) แผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ ฉบับที่สิบสอง พ.ศ. 2560 – 2564 ได้กำหนดกรอบยุทธศาสตร์การพัฒนา 6 ด้าน คือ ยุทธศาสตร์ที่ 1 พัฒนาโครงสร้างพื้นฐานดิจิทัลประสิทธิภาพสูงให้ครอบคลุมทั่วประเทศ ยุทธศาสตร์ที่ 2 ขับเคลื่อนเศรษฐกิจด้วยเทคโนโลยีดิจิทัล ยุทธศาสตร์ที่ 3 สร้างสังคมคุณภาพที่ทั่วถึงเท่าเทียมด้วยเทคโนโลยีดิจิทัล ยุทธศาสตร์ที่ 4 ปรับเปลี่ยนภาครัฐสู่การเป็นรัฐบาลดิจิทัล ยุทธศาสตร์ที่ 5 พัฒนากำลังคนให้พร้อมเข้าสู่ยุคเศรษฐกิจและสังคมดิจิทัล และยุทธศาสตร์ที่ 6 สร้างความเชื่อมั่นในการใช้เทคโนโลยีดิจิทัล ดังนั้นเทคโนโลยีสารสนเทศจึงมีส่วนสำคัญที่จะทำให้สังคมไทยเปลี่ยนจากสังคมอุตสาหกรรมเป็นสังคมสารสนเทศ

จากผลสำรวจการมีการใช้เทคโนโลยีสารสนเทศและการสื่อสารในครัวเรือน พ.ศ. 2563 [2] พบว่าประชากรไทยมีประมาณ 65.5 ล้านคน พบว่า มีผู้ใช้อินเทอร์เน็ต 55.7 ล้านคน (ร้อยละ 85.0) ใช้โทรศัพท์มือถือ 61.9 ล้านคน (ร้อยละ 94.6) ปัญหาและอุปสรรคที่พบในการใช้งานอุปกรณ์ดิจิทัล พบว่าไม่เข้าใจหน้าจอคำสั่งที่เป็นภาษาอังกฤษมากที่สุด ร้อยละ 68.2 รองลงมาคือ ไม่รู้วิธีการใช้งาน ร้อยละ 49.8 และไม่มั่นใจในความปลอดภัยในการใช้งานและภัยออนไลน์ต่าง ๆ ร้อยละ 44.9 ค่าใช้จ่ายในการใช้งานสูงเกินไป เช่น ค่าบริการ ค่าอุปกรณ์ เป็นต้น ร้อยละ 32.1% ไม่รู้ประโยชน์จากการใช้ ร้อยละ 31.1% ไม่รู้กฎ กติกา มารยาทในโลกออนไลน์ ร้อยละ 30.2% ไม่มีคนให้สอบถามเมื่อมีคำถาม ร้อยละ 27.1% และใช้เวลาในโลกออนไลน์มากเกินไป ร้อยละ 8.20% จากปัญหาดังกล่าวตามกรอบยุทธศาสตร์การพัฒนาทั้ง 6 ด้านข้างต้น โดยเฉพาะเป้าหมายประเด็นยุทธศาสตร์ที่ 3 มุ่งเน้นให้ประชาชนทุกคนมีความตระหนัก ความรู้ ความเข้าใจ ทักษะในการใช้เทคโนโลยีดิจิทัลให้เกิดประโยชน์ และสร้างสรรค์ (Digital Literacy) และเป้าหมายประเด็นยุทธศาสตร์ที่ 5 บุคลากรผู้ทำงานทุกสาขามีความรู้และทักษะด้านดิจิทัล สำนักงานปลัดกระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคมจึงได้ดำเนินกิจกรรมศึกษากรอบแนวทางการพัฒนาทักษะดิจิทัล (Digital Literacy) [3] สำรวจและประเมินสภาพการณ์รู้เท่าทันสื่อและสารสนเทศ เพื่อจัดเก็บข้อมูลที่สำคัญและจำเป็นต่อการส่งเสริม สนับสนุน พัฒนาการรู้เท่าทันสื่อและสารสนเทศในประเทศไทย ให้ประชาชนในประเทศสามารถเข้าถึง (Access) ค้นคืน (Retrieve) เข้าใจ (Understand) ประเมิน (Evaluation) ใช้ (Use) สร้าง (Create) ตลอดจนแบ่งปัน (Share) สารสนเทศและ

เนื้อหาสื่อในทุกรูปแบบ การใช้เครื่องมือต่าง ๆ อย่างมีประสิทธิภาพ มีการคิดวิเคราะห์ และมีจริยธรรม เพื่อการมีส่วนร่วมและปฏิสัมพันธ์ในกิจกรรมทั้งส่วนตัว อาชีพ และสังคม ซึ่งจะนำไปสู่การลดปัญหาความขัดแย้งทางด้านความคิดอย่างรุนแรง รวมถึงป้องกันการหลงเชื่อและถูกล่อลวงผ่านสื่อ การถูกชี้นำและการใช้สื่ออย่างรู้เท่าไม่ถึงการณ์ อีกทั้งได้ดำเนินการสำรวจและประเมินข้อมูลสถานภาพการรู้เท่าทันสื่อและสารสนเทศ (Media and Information Literacy: MIL) ให้มีข้อมูลมหภาคเชิงลึกเกี่ยวกับพฤติกรรมการบริโภคสื่อและสารสนเทศของประชากรไทย

จากที่มาดังกล่าวจึงได้ดำเนินการวิจัยเรื่องการศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัลและเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลผลการทดสอบ โดยใช้เทคนิคการทำเหมืองข้อมูล คาดว่างานวิจัยนี้จะสามารถนำมาใช้ประโยชน์กำหนดนโยบายการดำเนินงานและบทบาทของภาคส่วนต่าง ๆ ในการส่งเสริมทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัลของประชากรไทย เน้นเรื่องการรู้เท่าทันดิจิทัล (Digital Literacy) [4] โดยเฉพาะเด็กและเยาวชนร่วมถึงกลุ่มเปราะบางที่มีคุณลักษณะที่สอดคล้องกับผลการวิจัย ผ่านกิจกรรมการให้ความรู้และการฝึกอบรม เพื่อช่วยลดช่องว่างทางดิจิทัล สร้างโอกาสและความเท่าเทียมในการเข้าถึงข้อมูลข่าวสาร สารสนเทศและใช้ประโยชน์จากเทคโนโลยีดิจิทัลในชีวิตประจำวันได้อย่างรู้เท่าทัน และเตรียมพร้อมสู่การเป็นพลเมืองดิจิทัลในอนาคต

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล (Digital literacy) หมายถึง ทักษะในการนำเครื่องมือ อุปกรณ์และเทคโนโลยีดิจิทัลที่มีอยู่ในปัจจุบัน อาทิ คอมพิวเตอร์ โทรศัพท์ แท็บเล็ต โปรแกรมคอมพิวเตอร์ และสื่อออนไลน์ มาใช้ให้เกิดประโยชน์สูงสุด ในการสื่อสาร การปฏิบัติงานและการทำงานร่วมกัน หรือใช้เพื่อพัฒนากระบวนการทำงานหรือระบบงานในองค์กรให้มีความทันสมัยและมีประสิทธิภาพ [5] ทักษะดังกล่าวครอบคลุมความสามารถ 4 มิติ ได้แก่ การใช้ (Use) เข้าใจ (Understand) การสร้าง (Create) เข้าถึง (Access) เทคโนโลยีดิจิทัล ได้อย่างมีประสิทธิภาพ

2.1.2 การจัดการความไม่สมดุลของข้อมูล (Imbalanced data) เกิดขึ้นเมื่อข้อมูลที่ใช้งานมีจำนวนข้อมูลในแต่ละคลาสแตกต่างกันมาก ทำให้ผลลัพธ์การจำแนกข้อมูลมีความโน้มเอียงไปทางข้อมูลกลุ่มมาก เทคนิคการจัดการความไม่สมดุลของข้อมูล (Imbalanced data) ได้แก่ วิธีการสุ่มเกิน (over sampling) วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด (under sampling) และวิธีการสุ่มผสมผสาน (hybrid method) [6] สำหรับการสุ่มเกินโดยเทคนิค SMOTE เป็นหนึ่งในเทคนิคที่ใช้ในการแก้ปัญหาการจำแนกข้อมูลที่ไม่สมดุล โดยการสุ่มเพิ่มจำนวนข้อมูลกลุ่มน้อยให้มีจำนวนเพิ่มขึ้นใกล้เคียงหรือเท่ากับจำนวนข้อมูลกลุ่มมาก

2.1.3 เทคนิคการจำแนกข้อมูล เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคประเภทการจำแนกข้อมูลที่ได้รับความนิยมเนื่องจากอัลกอริทึมไม่ซับซ้อน สามารถแสดงผลลัพธ์เป็นผังต้นไม้ ที่ตีความและเข้าใจลักษณะของรูปแบบข้อมูลได้ง่าย การพยากรณ์ในรูปของต้นไม้ตัดสินใจเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) หมายถึงการสร้างแบบจำลองจากข้อมูลที่รู้กลุ่มที่แท้จริงก่อนล่วงหน้า รูปแบบของต้นไม้ประกอบด้วยโหนด โหนดแรกเรียกว่ารูทโหนด จากรูทโหนดจะแยกออกเป็นโหนดลูก ที่ระดับสุดท้ายจะเรียกว่าโหนดใบ ซึ่งจะแสดงคลาสที่เป็นผลลัพธ์ที่สามารถแยกแยะได้ ข้อดีคือเข้าใจได้ง่าย ดูลำดับขั้นตอนการตัดสินใจแล้ววิเคราะห์ได้ ข้อเสียคืออาจแตกกิ่งมากไปทำให้เกิดการเรียนรู้ได้ง่าย การแบ่งพิจารณาตัวแปรทีละตัวแยกกันตลอด ไม่สามารถใช้ตัวแปรสองชนิดขึ้นไปเป็นเกณฑ์ในการแบ่งพร้อมกันได้ [7]

เทคนิคเกรเดียนท์บูตทรีส (Gradient Boosted Trees) วิธีที่มีพื้นฐานมาจากเทคนิคต้นไม้ตัดสินใจ ซึ่งเป็นการปรับปรุงประสิทธิภาพของแบบจำลองให้มีค่าสูงขึ้น โดยการสุ่มสร้างต้นไม้ตัดสินใจหลายร้อยแบบจำลอง [8] และประเมินผลแต่ละแบบจำลองจนกว่าจะได้ต้นไม้ตัดสินใจที่สมบูรณ์ ข้อดีคือความอคติ (Bias) และความแปรปรวน (Variance) ลดลง [8]

เทคนิคแรนดอมฟอร์เรสต์ (Random forest) เป็นชุดของการจำแนกประเภทแบบไม่ตัดแต่งกิ่ง (unpruned) หรือต้นไม้ถดถอย (Regression Trees) สร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูลและคุณลักษณะข้อมูลแล้วนำมาสร้างเป็นต้นไม้ตัดสินใจซึ่งมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือกเรียกข้อมูลส่วนนี้ว่า Out-of-Bag (OOB) จะถูกนำมาใช้

ในการทดสอบต้นไม้ตัดสินใจ และนำผลการทำนายที่ได้มาโหวต (Vote) เพื่อหาคำตอบสุดท้าย ข้อดีคือใช้ได้กับปัญหา Classification และ Regression ใช้ได้ทั้งกับข้อมูลมีโครงสร้างและไม่มีโครงสร้าง และ ป้องกันการเกิด overfitting ข้อเสียคือผลลัพธ์ที่ได้แปรผลเพื่อนำไปประยุกต์ใช้ทำได้ค่อนข้างยาก [9]

เทคนิคนาอีฟเบย์ (Naïve Bayes) เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น (Probability) ตามทฤษฎีของเบย์ (Bayes's theorem) เป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ มีสมมติฐานว่าปริมาณของความสนใจขึ้นอยู่กับการกระจายความน่าจะเป็น (Probability Distribution) เป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ข้อดีคือมีความง่ายต่อการนำไปใช้งาน เนื่องจากรูปแบบการคำนวณนั้นง่ายและผลลัพธ์ที่ได้สามารถประยุกต์ใช้ได้ดีเหมาะกับกรณีเซตตัวอย่างมีจำนวนมาก ข้อเสียคือจะให้ผลลัพธ์ที่ดีก็ต่อเมื่อคุณลักษณะของข้อมูลที่นำมาทดสอบเป็นอิสระต่อกัน [10]

เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เป็นชุดคำสั่งที่สร้างขึ้นมาเพื่อการเรียนรู้ของเครื่องคอมพิวเตอร์ โดยชุดคำสั่งนี้จะทำให้เครื่องสามารถประมวลผลข้อมูลจำนวนมากที่พยายามเรียนรู้วิธีการแทนข้อมูลอย่างมีประสิทธิภาพ หลักการของการเรียนรู้เชิงลึกเป็นโครงข่ายประสาทเทียม (Artificial Neural Network) [11] ที่เป็นโหนดหลาย ๆ ชั้น และใช้การประมวลผลแบบขนานทำให้สามารถประมวลผลได้ครั้งละจำนวนมากช่วยให้การเรียนรู้ของเครื่องสามารถให้ผลลัพธ์ในการตัดสินใจและคาดการณ์ได้ดีมากยิ่งขึ้น ข้อดี คือ สามารถหาความสัมพันธ์ของข้อมูลต่างรูปแบบกันได้ไม่จำเป็นต้องจัดโครงสร้างข้อมูล ไม่จำเป็นต้องจัดหมวดหมู่ข้อมูล ไม่จำเป็นต้องกำหนดการจับคู่ข้อมูลล่วงหน้า ข้อเสียคือความแม่นยำของ Deep Learning แปรผันตรงกับปริมาณข้อมูล การอธิบายกระบวนการของ Deep Learning เป็นสิ่งที่ซับซ้อนอย่างมากยากที่จะอธิบายได้ [12]

2.1.4 การวัดประสิทธิภาพ คอนฟิวชันเมทริกซ์ (Confusion Matrix) เป็นตารางประเมินผลลัพธ์การทำนายของแบบจำลอง สามารถวัดจากผลลัพธ์การจำแนกข้อมูล (Classification) รูปแบบของคอนฟิวชันเมทริกซ์แสดงดังตารางที่ 1

ตารางที่ 1 *แสดงเมทริกซ์วัดประสิทธิภาพ*

	Actual Positive	Actual Negative
Predict Positive	True Positive (TP)	False Positive (FP)
Predict Negative	False Negative (FN)	True Negative (TN)

ค่าที่ได้จากการทำนาย (Prediction) อธิบายได้ ดังนี้

ค่า TP คือจำนวนข้อมูลทดสอบที่เป็นคลาส Positive และแบบจำลองจำแนกได้ถูกต้องว่าเป็น Positive

ค่า FP คือจำนวนข้อมูลทดสอบที่ไม่ใช่คลาส Positive แต่แบบจำลองทำนายผิดว่าเป็นคลาส Positive

ค่า TN คือจำนวนข้อมูลทดสอบที่เป็นคลาส Negative และแบบจำลองทำนายได้ถูกต้องว่าเป็นคลาส Negative

ค่า FN คือจำนวนข้อมูลทดสอบที่เป็นคลาส Negative แต่แบบจำลองทำนายผิดว่าเป็นคลาส Negative

ค่าความถูกต้อง (Accuracy) คำนวณจากสมการที่ 1

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

ค่าความแม่นยำ (Precision) คำนวณจากสมการที่ 2

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)$$

ค่าเรียกกลับ (Recall หรือ Sensitivity) คำนวณจากสมการที่ 3

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{(TP+FN)} \quad (3)$$

ค่าความจำเพาะ (Specificity) คำนวณจากสมการที่ 4

$$\text{Specificity} = \frac{TN}{(TN+FP)} \quad (4)$$

ค่าประสิทธิภาพโดยรวม (F-measure) คำนวณจากสมการที่ 5

$$\text{F1 - measure} = 2\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

ค่าเฉลี่ยเรขาคณิต (G-mean) คำนวณจากสมการที่ 6

$$\text{G-mean} = \sqrt{(\text{Sensitivity} * \text{Specificity})} \quad (6)$$

2.2 งานวิจัยที่เกี่ยวข้อง

กาญจนา หฤหรรษพงศ์ [13] ได้ศึกษาการรู้ดิจิทัล

ของนักศึกษาระดับปริญญาตรี มหาวิทยาลัยวลัยลักษณ์ มีวัตถุประสงค์เพื่อ 1) ประเมินผลการรู้ดิจิทัลของนักศึกษาปริญญาตรี มหาวิทยาลัยวลัยลักษณ์ โดยจำแนกตามหลักสูตร กลุ่มสาขาวิชา และสำนักวิชา 2) วิเคราะห์ปัจจัยที่มีความสัมพันธ์ต่อผลการสอบวัดการรู้ดิจิทัล 3) หาความสัมพันธ์ของระดับการรู้ดิจิทัลกับผลสัมฤทธิ์ในการเรียน 4) ศึกษาอุปสรรคและข้อเสนอแนะในการจัดการเรียนรู้ด้านเทคโนโลยีดิจิทัลแก่ผู้เกี่ยวข้อง เครื่องมือที่ใช้ในการวิจัย คือ แบบทดสอบวัดการรู้ดิจิทัล และแบบสอบถามของกลุ่มตัวอย่าง คือ นักศึกษาชั้นปีที่ 1 จำนวน 1,220 คน โดยนำผลการสอบวิเคราะห์ร่วมกับแบบสอบถามพฤติกรรมการใช้งานเทคโนโลยีสารสนเทศของผู้สอบ ทำการคำนวณค่าร้อยละ ค่าเฉลี่ย ค่าไคว์สแควร์และค่าสถิติสหสัมพันธ์ ผลการวิจัยพบว่า ผู้เข้าสอบร้อยละ 73.77 สอบผ่านที่คะแนน 50 เปอร์เซ็นต์ขึ้นไป การจำแนกผู้ที่สอบผ่านตามช่วงคะแนนพบว่า ส่วนใหญ่มีระดับการรู้ดิจิทัลต่ำที่คะแนน 50-54 เปอร์เซ็นต์ (ร้อยละ 30.92) ผลการวิเคราะห์ความสัมพันธ์ของพฤติกรรมในการใช้งานคอมพิวเตอร์ที่มีความสัมพันธ์กับผลการสอบอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 พบปัจจัยที่มีความสัมพันธ์ ได้แก่ หลักสูตร สำนักวิชา เกรดเฉลี่ยของวิชากลุ่มคอมพิวเตอร์ จำนวนชั่วโมงในการใช้งานคอมพิวเตอร์ต่อวัน และลักษณะกิจกรรมในการใช้คอมพิวเตอร์ นอกจากนี้ยังพบว่าระดับการรู้ดิจิทัลของผู้เรียนมีความสัมพันธ์กับเกรดเฉลี่ยของผู้เรียนในทิศทางเดียวกันในระดับต่ำ ($r = 0.393$) อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.01 ส่วนอุปสรรคต่อการรู้ดิจิทัลที่สำคัญ คือ ปัญหาอินเทอร์เน็ตมหาวิทยาลัยช้า (ร้อยละ 76.70) และการเชื่อมต่ออินเทอร์เน็ตมีปัญหา (ร้อยละ 74.78)

วิชญ์วิสิฐ เกษรสิทธิ์ และคณะ [14] ได้ศึกษาการจำแนกโดยใช้ข้อมูลที่ไม่สมดุลเป็นปัญหาสำคัญในเทคนิคการจำแนก ซึ่งการจำแนกข้อมูลที่มีข้อมูลในกลุ่มมากและกลุ่มน้อยปะปนกันจะทำให้ข้อมูลในกลุ่มมากจะมีคุณสมบัติบางประการที่บดบังคุณสมบัติของกลุ่มน้อย ทำให้การจำแนกข้อมูลในกลุ่มน้อยไม่สามารถจำแนกได้อย่างมีประสิทธิภาพ การวิจัยครั้งนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลผู้ป่วยโรคเบาหวาน โดยการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูลจำนวน 4 วิธี คือวิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) โดยใช้เทคนิคการจำแนกคือวิธีการถดถอยโลจิสติกแบบมัลติโนเมียลและวิธีต้นไม้การตัดสินใจในการจำแนกผู้ป่วยโรคเบาหวาน จากการเปรียบเทียบประสิทธิภาพของสถิติและอัลกอริทึมในการจำแนก พบว่าข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลใหม่ สามารถจำแนกผู้ป่วยโรคเบาหวานด้วยวิธีต้นไม้การตัดสินใจมีผลลัพธ์ดีที่สุด

วรการ ใจดี และ นพณัฐ วรรณภีร์ [15] ได้ศึกษาปัจจัยที่ส่งผลต่อการสำเร็จการศึกษาตามแผนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคการคัดเลือกคุณลักษณะบนชุดข้อมูลที่ไม่สมดุล ผู้วิจัยได้เก็บรวบรวมข้อมูลนักศึกษาจากสาขาวิชาระบบสารสนเทศทางคอมพิวเตอร์ คณะบริหารธุรกิจและคณะศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา จำนวน 358 คน พบว่า ชุดข้อมูลมีความไม่สมดุลกัน จึงได้ทำการแก้ปัญหาโดยใช้วิธีการปรับสมดุลให้กับชุดข้อมูลโดยใช้เทคนิคการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย (SMOTE) การวิเคราะห์ปัจจัยใช้เทคนิคการคัดเลือกคุณลักษณะ 3 เทคนิค คือ 1) Chi-Square Feature Selection 2) Information Gain Feature Selection และ 3) Correlation Based Feature Selection ดำเนินการคัดเลือกคุณลักษณะ 10 อันดับแรกที่มีความสำคัญมากที่สุดของแต่ละเทคนิค มาคำนวณความถี่ทางสถิติของแต่ละปัจจัยใน 10 อันดับแรกจากการคัดเลือกคุณลักษณะทั้งสามเทคนิค ผลการวิจัยพบว่า ทั้งสามเทคนิคที่ใช้ให้ผลลัพธ์ที่ใกล้เคียงกันและมีจำนวนวิชาที่ติดมาเป็นอันดับ 1-10 มากถึงหกวิชาที่เหมือนกัน ปัจจัยที่มีความสำคัญมากที่สุดในทุกเทคนิคการคัดเลือกคุณลักษณะคือ วิชาการเขียนโปรแกรมคอมพิวเตอร์ 1 ซึ่งเป็นวิชาหลักของสาขาวิชาฯ เป็นปัจจัยที่ส่งผลต่อการสำเร็จการศึกษาตามแผนของนักศึกษาอย่างมีนัยสำคัญ

รัชพล กลัดชื่น และ จรัญ แสนราช [16] ได้ศึกษาการเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษาโดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายและคุณลักษณะที่มีต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา โดยใช้เทคนิคการจำแนกข้อมูล 3 เทคนิค ได้แก่ Decision Tree: J48graft, Naïve Bayes และ Rule Induction ทำการเปรียบเทียบประสิทธิภาพแบบจำลองการทำนายผลการศึกษาพบว่าการใช้เทคนิค Decision Tree: J48graft ด้วยการเลือกคุณลักษณะแบบ Forward Selection และการเลือกคุณลักษณะทั้งหมด มีค่าความถูกต้องเท่ากับ 83.08%

และ 81.71% ตามลำดับ

อรรถพล กิตติธนาชัย [17] ได้ศึกษาความสัมพันธ์ระหว่างสถานภาพที่มีความสัมพันธ์ต่อสมรรถนะของนักเรียนโรงเรียนสาธิตมหาวิทยาลัยศรีนครินทรวิโรฒประสานมิตร (ฝ่ายมัธยม) กลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้คือ นักเรียนชั้นมัธยมศึกษาชั้นปีที่ 1 - 6 ภาคเรียนที่ 1 ประจำปีการศึกษา 2555 ที่กำลังศึกษาอยู่ในโรงเรียนสาธิตมหาวิทยาลัยศรีนครินทรวิโรฒประสานมิตร (ฝ่ายมัธยม) จำนวน 326 คน เครื่องมือที่ใช้ในการรวบรวมข้อมูลคือ แบบสอบถาม โดยสถิติที่ใช้วิเคราะห์ข้อมูล ได้แก่ ค่าร้อยละ (Percentage) หาค่าเฉลี่ย ($\bar{x}$) ค่าเบี่ยงเบนมาตรฐาน (S.D.) ไคว์สแควร์ (Chi – Square) และสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Person 's Product Moment Correlation Coefficient) ผลการวิจัยพบว่า พฤติกรรมการใช้เทคโนโลยีสารสนเทศ ด้านความรู้ ด้านเจตคติ และด้านการปฏิบัติ มีสัมพันธ์กันทางบวกทุกค่าอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.001 โดยพฤติกรรมการใช้เทคโนโลยีสารสนเทศด้านการปฏิบัติกับสมรรถนะโดยรวม มีความสัมพันธ์สูงสุด และพฤติกรรมการใช้เทคโนโลยีสารสนเทศด้านเจตคติกับสมรรถนะโดยรวมมีความสัมพันธ์ต่ำสุด

จากงานวิจัยข้างต้น ผู้วิจัยได้แนวคิดการวิเคราะห์ปัจจัยที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล แต่เนื่องด้วยข้อมูลมีความไม่สมดุลกันจึงแก้ปัญหาโดยวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) ซึ่งเป็นการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อยให้ใกล้เคียงหรือเท่ากับกลุ่มมากก่อนนำไปคำนวณหาค่าน้ำหนักและคัดเลือกคุณลักษณะจากความถี่ที่เกิดขึ้นก่อนนำไปออกแบบและสร้างแบบจำลองการพยากรณ์ด้วยเทคนิคต่าง ๆ ต่อไป

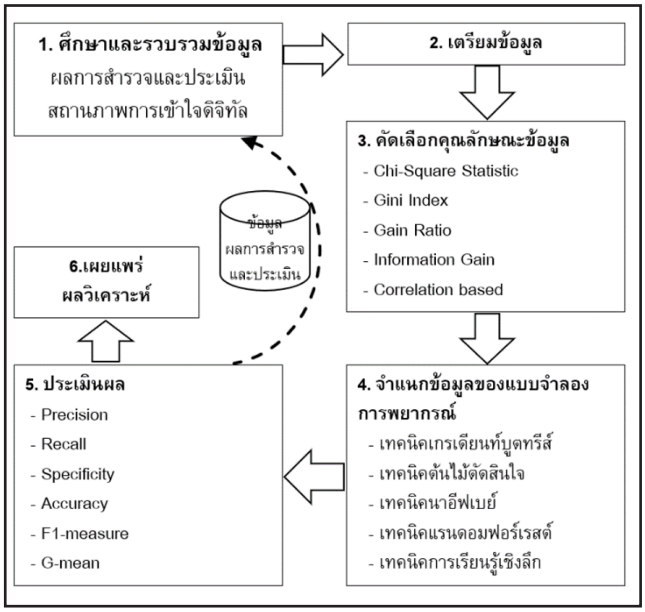
3. วิธีดำเนินการวิจัย

เครื่องมือที่ใช้ในการวิจัยใช้โปรแกรม MS Excel ในการรวบรวมและเตรียมข้อมูล ใช้โปรแกรม RapidMiner Studio 10.1 (Educational Edition) ในการสร้างแบบจำลองด้วยโมเดลต่าง ๆ และประเมินผลประสิทธิภาพของโมเดล

กระบวนการวิจัยมีขั้นตอนที่เป็นไปตามกรอบแนวคิดของ CRISP – DM ดังภาพที่ 1

3.1 ขั้นตอนการศึกษาและรวบรวมข้อมูล (Data Gathering)

ข้อมูลที่นำมาใช้วิจัยเป็นข้อมูลผลการสำรวจและประเมินสถานภาพการเข้าใจดิจิทัล (Digital Literacy) ของสำนักงานคณะกรรมการดิจิทัลเพื่อเศรษฐกิจและสังคมแห่งชาติ ระหว่างวันที่ 1 มกราคม 2564 – 30 กันยายน 2564 จำนวน 12,374 ระเบียน จำนวน 67 คุณลักษณะ ซึ่งใช้กระบวนการสำรวจตามกรอบการประเมินการเข้าใจสื่อและสารสนเทศระดับโลก ตามเอกสารการเตรียมความพร้อมและสมรรถนะของประเทศของยูเนสโก มาใช้เป็นเอกสารหลักและดำเนินการสำรวจจากประชาชนทั่วประเทศไทย



ภาพที่ 1 *กรอบแนวคิดการวิจัย*

ตารางที่ 2 *แสดงรายละเอียดคุณลักษณะข้อมูล*

ลำดับ	คุณลักษณะ	ตัวอย่างข้อมูล
ข้อมูลเชิงประชากรศาสตร์		
1	เพศ	ชาย
2	การแบ่งเขตปกครอง	ในเขตเทศบาล
3	อาชีพ	พนักงานเอกชน
4	รายได้เฉลี่ยต่อเดือน	10,001-15,000 บาท
5	ระดับการศึกษาสูงสุด	มัธยมศึกษาตอนปลาย
6	สถานะครองคู่	โสด
7	ศาสนา	พุทธ
8	จำนวนสมาชิกในครัวเรือน	4
9	สถานะความพิการ	0

ตารางที่ 2 *แสดงรายละเอียดคุณลักษณะข้อมูล (ต่อ)*

ลำดับ	คุณลักษณะ	ตัวอย่างข้อมูล
ข้อมูลเชิงพฤติกรรมศาสตร์		
11	มีคอมพิวเตอร์ส่วนตัว	1
12	เวลาเฉลี่ยใช้สื่อสังคมออนไลน์ต่อวัน	น้อยกว่า 1 ชั่วโมง/วัน
13	เคยผ่านการอบรมอบรม MIL/DL	0
ข้อมูลเชิงพฤติกรรมศาสตร์ (ประสบการณ์การใช้สื่อดิจิทัลตามวัตถุประสงค์)		
14	ระดับการศึกษาสูงสุด	1
15	สถานะครองคู่	0
16	ศาสนา	1
...	...	การใช้สื่อดิจิทัลเพื่อวัตถุประสงค์อื่น ๆ
ข้อมูลเชิงพฤติกรรมศาสตร์ (ประสบการณ์การใช้อุปกรณ์เข้าถึงสื่อดิจิทัล)		
26	การเข้าถึงสื่อดิจิทัลด้วยโน้ตบุ๊ก	1
27	การเข้าถึงสื่อดิจิทัลด้วยคอมพิวเตอร์ตั้งโต๊ะ	0
28	การเข้าถึงสื่อดิจิทัลด้วยโทรศัพท์มือถือ	1
...	...	อุปกรณ์ที่ใช้เข้าถึงสื่อดิจิทัลอื่นๆ
ข้อมูลเชิงพฤติกรรมศาสตร์ (ประสบการณ์การใช้สื่อในชีวิตประจำวัน)		
37	หนังสือ	1
38	หนังสือพิมพ์	0
39	ประกาศ/ป้ายฉลากสินค้า	0
...	...	สื่อที่เข้าถึงในชีวิตประจำวันอื่น ๆ
ข้อมูลเชิงพฤติกรรมศาสตร์ (ประสบการณ์ปัญหาเข้าถึงสื่อดิจิทัล)		
57	ปัญหาค่าอินเทอร์เน็ตแพง	1
58	ปัญหาถูกแฮ็ก/แบล็กเมล์	0
59	ปัญหาอีเมลขยะ	1
...	...	ปัญหาเข้าถึงสื่อดิจิทัลอื่นๆ

ตารางที่ 2 *แสดงรายละเอียดคุณลักษณะข้อมูล (ต่อ)*

ลำดับ	คุณลักษณะ	ตัวอย่างข้อมูล
ข้อมูลคลาสหรือลาเบล		
67	ผลการประเมินทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล	1 : ขั้นพื้นฐาน 2 : สูงกว่าขั้นพื้นฐาน

ข้อมูลผลการสำรวจและประเมินสถานภาพการเข้าใจดิจิทัล แบ่งออกเป็น 3 ส่วน คือ ข้อมูลเชิงประชากรศาสตร์ ได้แก่ เพศ การแบ่งเขตปกครอง อาชีพ รายได้เฉลี่ยต่อเดือน ระดับการศึกษาสูงสุด เป็นต้น ข้อมูลเชิงพฤติกรรมศาสตร์ ได้แก่ พฤติกรรมการใช้อินเทอร์เน็ต วัตถุประสงค์ของการใช้สื่อดิจิทัล สื่อที่เข้าถึงในชีวิตประจำวัน ปัญหาการเข้าถึงสื่อดิจิทัล และข้อมูลคลาสหรือลาเบล คือ ผลการประเมินทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล จำแนกออกเป็น 2 ระดับ คือ ระดับขั้นพื้นฐานและระดับสูงกว่าขั้นพื้นฐาน

3.2 ขั้นตอนการเตรียมข้อมูล (Data Pre-processing)

จากการรวบรวมข้อมูล พบว่าข้อมูลบางระเบียนหายไป (Missing Value) ได้แก่ รายได้เฉลี่ยต่อเดือนมีข้อมูลว่าง 47 ระเบียน คิดเป็น 0.38% ของข้อมูลทั้งหมด จำนวนสมาชิกในครัวเรือนมีข้อมูลว่าง 59 ระเบียน คิดเป็น 0.48 % ของข้อมูลทั้งหมด ส่งผลให้ไม่สามารถคัดเลือกคุณลักษณะได้ แต่ข้อมูลมีการสูญหายน้อยกว่า 5% ผู้วิจัยจึงแก้ปัญหาทั้ง 2 คุณลักษณะ โดยการแทนค่าของข้อมูลที่หายเป็นค่าเดียว (Single Imputation Methods) ด้วยการแทนที่ด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าในแต่ละตัวแปร (Mean Imputation)

หลังจากขั้นตอนการเตรียมข้อมูล พบว่าข้อมูลคลาสผลการประเมินทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล ระดับขั้นพื้นฐานมีจำนวนน้อยกว่าระดับสูงกว่าขั้นพื้นฐาน ส่งผลให้คลาสเกิดความไม่สมดุล ผู้วิจัยจึงแก้ปัญหาด้วยวิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling TEchnique: SMOTE) โดยการเพิ่มข้อมูลในกลุ่มน้อยทำให้การกระจายของข้อมูลมีความสมดุลมากขึ้น [14]

3.3 ขั้นตอนการคัดเลือกคุณลักษณะข้อมูล (Feature Selection)

ขั้นตอนการคัดเลือกคุณลักษณะ เพื่อศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล โดยใช้เทคนิคการคำนวณหาค่าน้ำหนัก (Filter approach)

ได้แก่ 1) Chi-Square Statistic 2) Gini Index 3) Gain Ratio 4) Information Gain และ 5) Correlation based

Chi-Square Statistic เป็นการเลือกคุณลักษณะแบบการคัดกรอง (Filter-Base Feature Selection) จะใช้เงื่อนไขทางสถิติในการคัดกรอง โดยค่าสถิติไคสแวร์วัดค่าความสัมพันธ์ระหว่างคุณลักษณะกับคลาสคำตอบเพื่อจัดลำดับคุณลักษณะตามค่านัยสำคัญทางสถิติ สามารถคำนวณได้จากสมการที่ 7

ตารางที่ 3 *แสดงผลการเปรียบเทียบมูลที่ผ่านกระบวนการสุ่มเพิ่มข้อมูลกลุ่มน้อย*

คลาส	ข้อมูลเดิม	การสังเคราะห์ข้อมูลเพิ่ม
ขั้นพื้นฐาน	6964	6964
สูงกว่าขั้นพื้นฐาน	5410	6964

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (7)$$

x^2 = ค่าสถิติไคสแควร์

O_i = ความถี่ที่ได้จากการสังเกต (Observed Frequency)

E_i = ความถี่ที่คาดหวัง (Expected Frequency) ซึ่งมีค่าเท่ากับจำนวนข้อมูลคูณด้วยสัดส่วนที่คาดหวัง

Gini Index (GI) เป็นวิธีคัดเลือกคุณลักษณะข้อมูลที่ขึ้นต่อคลาส โดยจะเลือกคุณลักษณะจากระดับความแตกต่างของคลาส [18] แสดงได้จากสมการที่ 8

$$GI(l_i) = \sum_{j=1}^{k} p(l_i|C_j)^2 \, p(C_j|l_i)^2 \qquad (8)$$

โดย $C = \{C_1,\ldots,C_k\}$ แทนคลาสของคุณลักษณะทั้งหมด K คลาส

Gain Ratio (GR) เป็นเทคนิคเพื่อเพิ่มประเมินความน่าเชื่อถือของมิติข้อมูล โดยการวัด Gain Ratio ในแต่ละคลาส [19] การหาค่า Gain Ratio คำนวณได้จากสมการที่ 9

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \qquad (9)$$

ค่า $SplitInfo_A(D)$ หมายถึง ปริมาณข้อมูลที่ถูกพิจารณาโดยการแบ่งชุดข้อมูลในชุดข้อมูล D ออกเป็น m ชุดข้อมูลย่อยตามคุณลักษณะ A ดังสมการที่ 10

$$SplitInfo_A(D) = -\sum_{j=1}^{m} \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|} \qquad (10)$$

Information Gain (IG) เป็นการประเมินค่าเพื่อใช้แบ่งข้อมูลด้วยการคำนวณค่า Gain สำหรับแต่ละมิติข้อมูล ถ้ามิติข้อมูลใดมีค่า Gain สูงสุด จะถูกเลือกให้เป็นกลุ่มย่อยที่มีอำนาจจำแนก [20] การคำนวณค่า Entropy คำนวณได้จากสมการที่ 11

$$Entropy(p) = -\sum_{i=0}^{c-1} p(j|t)\log_2 p(j|t) \qquad (11)$$

โดย $\sum_i$ คือ ผลรวมของความน่าจะเป็นของค่า j ที่เกิดในคลาส t

$p(j|t)$ คือ ค่าความถี่ที่มีความสัมพันธ์ของกลุ่ม j กับโหนด t

การคำนวณค่า Gain คำนวณได้จากสมการที่ 12

$$Gain = Entropy(p) - \left(\sum_{i=1}^{k} \frac{ni}{n} Entropy(i)\right) \qquad (12)$$

โดย Entropy(p) คือ ค่า Entropy ของตัว Root

$\sum_{i=1}^{k} \frac{ni}{n} Entropy(i)$ คือค่า Entropy ในแต่ละโหนดย่อย

Correlation based เป็นเทคนิคที่ใช้การหากลุ่มคุณลักษณะที่ได้รับการประเมินค่าจากความสามารถในการคาดการณ์ โดยคุณลักษณะที่ถูกคัดเลือกใช้สำหรับการจำแนกประเภทของข้อมูล [20] สมการประเมินกลุ่มย่อยของมิติข้อมูลแสดงได้จากสมการที่ 13

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (13)$$

โดย M_s คือ ค่าที่ค้นหาได้ของมิติข้อมูลกลุ่มย่อย S ซึ่งประกอบด้วยมิติข้อมูล K

$\overline{r_{cf}}$ คือ ค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส ($f \in s$)

$\overline{r_{ff}}$ คือ ค่าเฉลี่ยความสัมพันธ์ระหว่างมิติข้อมูล

3.4 ขั้นตอนการสร้างแบบจำลอง (Modeling)

การสร้างแบบจำลองเป็นกระบวนการสร้างรูปแบบทางคณิตศาสตร์หนึ่งอย่างหรือมากกว่า โดยอาศัยข้อมูลจากอดีตมาสร้างเป็นแบบจำลอง (Model) การพยากรณ์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ในงานวิจัยนี้ออกแบบและสร้างแบบจำลองการพยากรณ์จำนวน 5 แบบจำลอง ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคเกรเดียนท์บูตทรีส เทคนิคแรนดอมฟอร์เรสต์ เทคนิคนาอีฟเบย์ และ เทคนิคการเรียนรู้เชิงลึก

3.5 ขั้นตอนการประเมินผล (Evaluation)

การประเมินผลเป็นการวัดประสิทธิภาพจากผลลัพธ์ของแบบจำลองการพยากรณ์ว่ามีความน่าเชื่อถือมากน้อยเพียงใด หากมีความน่าเชื่อถือไม่เพียงพอจะกลับไปทบทวนขั้นตอนที่ 1 - 4 ซ้ำอีกครั้ง

ในงานวิจัยนี้ทดสอบแบบจำลองการพยากรณ์ด้วยวิธี K-Fold Cross Validation Test แบ่งข้อมูลออกเป็น 10 ส่วน 20 ส่วน และ 30 ส่วน แต่ละส่วนมาจากการสุ่มจุดเริ่มต้นเดียวกัน (use local random seed) ข้อมูลกระจายจำนวนเท่าๆ กัน การทดสอบประสิทธิภาพของแบบจำลองได้ดำเนินการวัดค่าความแม่นยำ ค่าเรียกกลับ ค่าความจำเพาะ ค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม และค่าเฉลี่ยเรขาคณิต

3.6 เผยแพร่ผลวิเคราะห์ (Deployment)

หลังจากทำการประเมินผลแบบจำลองการพยากรณ์การจำแนกข้อมูลของชุดข้อมูลผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัลเรียบร้อยแล้ว ผู้วิจัยจะเผยแพร่ผลวิเคราะห์ เพื่อให้หน่วยงานทั้งภาครัฐและเอกชนสามารถนำผลลัพธ์ที่ได้ไปใช้ประโยชน์ในการส่งเสริมความรู้และพัฒนาทักษะทางดิจิทัลแก่ประชาชนทุกช่วงวัย ช่วยลดช่องว่างทางดิจิทัล สร้างโอกาสและความเท่าเทียมในการเข้าถึงข้อมูลข่าวสาร สารสนเทศและใช้ประโยชน์จากเทคโนโลยีดิจิทัลในชีวิตประจำวัน การอาชีพ และเตรียมพร้อมสู่การเป็นพลเมืองดิจิทัลในอนาคต

4. สรุปผลการวิจัย

4.1 ผลการศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล

ผลการศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล โดยใช้เทคนิคการคำนวณหาค่าน้ำหนัก ได้แก่ 1) Chi-Square Statistic 2) Gini Index 3) Gain Ratio 4) Information Gain และ 5) Correlation based สามารถแสดงผลค่าน้ำหนักสูงสุดที่คำนวณได้ 10 ลำดับแรกของแต่ละเทคนิค โดยเรียงตามลำดับหมายเลขในวงเล็บ และคำนวณความถี่ของแต่ละคุณลักษณะ ดังตารางที่ 4

ผลการศึกษาพบว่า คุณลักษณะ 10 ลำดับแรกจากผลการคำนวณค่าน้ำหนักสูงสุดทั้ง 5 เทคนิคให้ผลลัพธ์ที่ใกล้เคียงกัน คุณลักษณะที่มีความถี่ 5 ได้แก่ การใช้สื่อประกาศ/ป้ายฉลากสินค้า ปัญหาค่าอินเทอร์เน็ตแพง ปัญหาการเข้าถึงพื้นที่บริการอินเทอร์เน็ต การใช้สื่อดิจิทัลเพื่อโซเชียลมีเดีย ปัญหาข้อความสแปม/โฆษณา และการใช้สื่อดิจิทัลเพื่อเข้าถึงเว็บไซต์ คุณลักษณะที่มีความถี่ 4 ได้แก่ ปัญหาการไม่สามารถเชื่อมต่ออินเทอร์เน็ต และการใช้สื่อดิจิทัลเพื่อสัมมนาและอบรม คุณลักษณะที่มีความถี่ 2 ได้แก่ การใช้สื่อดิจิทัลเพื่อการเรียนรู้ผ่านสื่ออิเล็กทรอนิกส์ การใช้สื่อสิ่งพิมพ์ในชีวิตประจำวัน ระดับการศึกษาสูงสุด และอาชีพ คุณลักษณะที่มีความถี่ 1 ได้แก่ มีคอมพิวเตอร์ส่วนตัว การใช้สื่อดิจิทัลเพื่อดูละครหรือภาพยนตร์ การใช้สื่อดิจิทัลเพื่อดูทีวี และการใช้สื่อดิจิทัลเพื่อฟังวิทยุ

ผู้วิจัยคัดเลือกคุณลักษณะที่มีความถี่ตั้งแต่ 2 ไปออกแบบและสร้างแบบจำลองการพยากรณ์ ได้แก่ การใช้สื่อประกาศ/ป้ายฉลากสินค้า ปัญหาค่าอินเทอร์เน็ตแพง ปัญหาการเข้าถึงพื้นที่บริการอินเทอร์เน็ต การใช้สื่อดิจิทัลเพื่อโซเชียลมีเดีย ปัญหาข้อความสแปม/โฆษณา การใช้สื่อดิจิทัลเพื่อเข้าถึงเว็บไซต์ ปัญหาการไม่สามารถเชื่อมต่ออินเทอร์เน็ต การใช้สื่อดิจิทัลเพื่อสัมมนาและอบรม การใช้สื่อดิจิทัลเพื่อการเรียนรู้ผ่านสื่ออิเล็กทรอนิกส์ การใช้สื่อสิ่งพิมพ์ในชีวิตประจำวัน ระดับการศึกษาสูงสุด และอาชีพ

4.2 ผลการศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล

การคำนวณเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคเกรเดียนท์บูตทรีส เทคนิค

ตารางที่ 4 *แสดงคุณลักษณะ 10 อันดับแรกที่มีความสำคัญมากที่สุดของแต่ละเทคนิค*

คุณลักษณะ	Chi-Square	Gini Index	Gain Ratio	Information Gain	Correlation based	ความถี่
การใช้สื่อประกาศ/ป้ายฉลากสินค้า	758.43519 (1)	0.02161 (2)	0.03386 (1)	0.03151 (1)	0.18698 (3)	5
ปัญหาค่าอินเทอร์เน็ตแพง	691.31027 (4)	0.02007 (4)	0.03018 (7)	0.02920 (4)	0.17871 (5)	5
ปัญหาการเข้าถึงพื้นที่บริการอินเทอร์เน็ต	656.33207 (6)	0.01943 (5)	0.03147 (4)	0.02833 (5)	0.18187 (4)	5
การใช้สื่อดิจิทัลเพื่อโซเชียลมีเดีย	729.94252 (3)	0.02162 (1)	0.03143 (5)	0.03142 (2)	0.18731 (2)	5
ปัญหาข้อความสแปม/โฆษณา	617.00276 (7)	0.01821 (6)	0.02795 (8)	0.02649 (6)	0.17760 (6)	5
การใช้สื่อดิจิทัลเพื่อเข้าถึงเว็บไซต์	657.53430 (5)	0.02065 (3)	0.03265 (2)	0.03010 (3)	0.19118 (1)	5
ปัญหาการไม่สามารถเชื่อมต่ออินเทอร์เน็ต		0.01692 (9)	0.03149 (3)	0.02478 (8)	0.17216 (7)	4
การใช้สื่อดิจิทัลเพื่อสัมมนาและอบรม		0.01563 (10)	0.03131 (6)	0.02294 (10)	0.15929 (8)	4
การใช้สื่อดิจิทัลเพื่อการเรียนรู้ผ่านสื่ออิเล็กทรอนิกส์	559.93311 (10)		0.02317 (10)			2
การใช้สื่อสิ่งพิมพ์ในชีวิตประจำวัน			0.02606 (9)		0.15183 (9)	2
ระดับการศึกษาสูงสุด		0.01810 (7)		0.02644 (7)		2
อาชีพ		0.01693 (8)		0.02468 (9)		2
มีคอมพิวเตอร์ส่วนตัว					0.14368 (10)	1
การใช้สื่อดิจิทัลเพื่อดูละครหรือภาพยนตร์	565.73283 (9)					1
การใช้สื่อดิจิทัลเพื่อดูทีวี	737.89788 (2)					1
การใช้สื่อดิจิทัลเพื่อฟังวิทยุ	584.78202 (8)					1

แรนดอมฟอร์เรสต์ เทคนิคนาอีฟเบย์ และเทคนิคการเรียนรู้เชิงลึก จากขั้นตอนการสร้างแบบจำลองการพยากรณ์ และขั้นตอนการประเมินผล แสดงผลการคำนวณค่าความแม่นยำ ค่าเรียกกลับ ค่าความจำเพาะ ค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม และค่าเฉลี่ยเรขาคณิต ดังตารางที่ 5 และตารางที่ 6

จากตารางที่ 6 พบว่าผลการคำนวณค่าความถูกต้อง (Accuracy) เทคนิคต้นไม้ตัดสินใจ มีค่ามากสุด 70.87% เมื่อแบ่งข้อมูลออกเป็น 20 ส่วน เทคนิคเกรเดียนท์บูตทรีส มีค่ามากสุด 69.72% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคแรนดอมฟอร์เรสต์ มีค่ามากสุด 76.29% เมื่อแบ่งชุดข้อมูลออกเป็น 30 ส่วน เทคนิคนาอีฟเบย์ มีค่ามากสุด 66.64% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคการเรียนรู้เชิงลึก มีค่ามากสุด 69.72% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน ผลการคำนวณประสิทธิภาพโดยรวม (F1-measure) เทคนิคต้นไม้ตัดสินใจ มีค่ามากสุด 72.05% เมื่อแบ่งข้อมูลออกเป็น 20 ส่วน เทคนิคเกรเดียนท์บูตทรีส มีค่ามากสุด 65.22% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคแรนดอมฟอร์เรสต์ มีค่ามากสุด 76.00% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคนาอีฟเบย์ มีค่ามากสุด 64.88% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคการเรียนรู้เชิงลึก มีค่ามากสุด 65.22% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน

ตารางที่ 5 *ผลการคำนวณค่าความแม่นยำ ค่าเรียกกลับ และค่าความจำเพาะ*

Classification	Precision			Recall			Specificity		
	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน
Decision Tree	0.6883	0.6924	0.6881	0.7399	0.7510	0.7507	0.6648	0.6664	0.6597
Gradient Boosted Tree	0.7609	0.7724	0.7659	0.5691	0.5531	0.5679	0.8212	0.8370	0.8264
Random Forest	0.7641	0.7684	0.7695	0.7457	0.7471	0.7507	0.7698	0.7748	0.7751
Naïve Bay	0.6842	0.6847	0.6850	0.6159	0.6153	0.6162	0.7157	0.7167	0.7167
Deep Learning	0.7609	0.7724	0.7659	0.5691	0.5531	0.5679	0.8212	0.8370	0.8264

ตารางที่ 6 *ผลการคำนวณค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม และค่าเฉลี่ยเรขาคณิต*

Algorithm	Accuracy			F1-measure			G-mean		
	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน
Decision Tree	0.7024	0.7087	0.7052	0.7132	0.7205	0.7180	0.7014	0.7075	0.7037
Gradient Boosted Tree	0.6951	0.6951	0.6972	0.6512	0.6446	0.6522	0.6836	0.6804	0.6851
Random Forest	0.7578	0.7610	0.7629	0.7548	0.7576	0.7600	0.7577	0.7609	0.7628
Naïve Bay	0.6658	0.6660	0.6664	0.6482	0.6482	0.6488	0.6639	0.6641	0.6645
Deep Learning	0.6951	0.6951	0.6972	0.6512	0.6446	0.6522	0.6836	0.6804	0.6851

ผลการคำนวณค่าเฉลี่ยเรขาคณิต (G-mean) เทคนิคต้นไม้ตัดสินใจ มีค่ามากสุด 70.75% เมื่อแบ่งข้อมูลออกเป็น 20 ส่วน เทคนิคเกรเดียนท์บูตทรีส มีค่ามากสุด 68.51% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคแรนดอมฟอร์เรสต์ มีค่ามากสุด 76.28% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคนาอีฟเบย์ มีค่ามากสุด 66.45% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน เทคนิคการเรียนรู้เชิงลึก มีค่ามากสุด 68.51% เมื่อแบ่งข้อมูลออกเป็น 30 ส่วน

5. อภิปรายผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาคุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล โดยใช้เทคนิคการคัดเลือกคุณลักษณะและเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล โดยใช้เทคนิคการทำเหมืองข้อมูล กระบวนการวิจัยเป็นไปตามขั้นตอนของ CRISP-DM ผลการศึกษาพบว่า

5.1 คุณลักษณะที่ส่งผลต่อผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัล ใช้เทคนิคการคัดเลือกคุณลักษณะคำนวณหาค่าน้ำหนัก จำนวน 5 เทคนิค ซึ่งให้ผลลัพธ์ที่ใกล้เคียงกัน พบว่าคุณลักษณะที่มีความถี่ระดับสูงสุดความถี่ 5 มีจำนวน 6 คุณลักษณะ ได้แก่ การใช้สื่อประกาศ/ป้ายฉลากสินค้า ปัญหาค่าอินเทอร์เน็ตแพง ปัญหาการเข้าถึงพื้นที่บริการอินเทอร์เน็ต การใช้สื่อดิจิทัลเพื่อโซเชียลมีเดีย ปัญหาข้อความสแปมหรือโฆษณา และการใช้สื่อดิจิทัลเพื่อเข้าถึงเว็บไซต์ทั้ง 6 คุณลักษณะเป็นคุณลักษณะข้อมูลเชิงพฤติกรรมศาสตร์ที่มีความสำคัญมากที่สุดต่อการใช้ การเข้าใจ การสร้าง และเข้าถึงเทคโนโลยีดิจิทัลได้อย่างมีประสิทธิภาพ ซึ่งสอดคล้องกับงานวิจัยของอรรถพล กิตติธนาชัย ที่ได้ศึกษาพฤติกรรมการใช้เทคโนโลยีสารสนเทศที่สัมพันธ์ต่อสมรรถนะของนักเรียนโรงเรียนสาธิตมหาวิทยาลัยศรีนครินทรวิโรฒประสานมิตร (ฝ่ายมัธยม) [17] ผลการวิจัยผลการหาความสัมพันธ์ตามสมมติฐานพบว่าพฤติกรรมการใช้เทคโนโลยีสารสนเทศ ด้านความรู้ ด้านเจตคติ

และด้านการปฏิบัติมีความสัมพันธ์ต่อระดับสมรรถนะของนักเรียน พฤติกรรมการใช้เทคโนโลยีสารสนเทศกับสมรรถนะโดยรวมสัมพันธ์กันทางบวกทุกค่า อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.001 โดยพฤติกรรมการใช้เทคโนโลยีสารสนเทศด้านการปฏิบัติกับสมรรถนะโดยรวม มีความสัมพันธ์สูงสุด และพฤติกรรมการใช้เทคโนโลยีสารสนเทศด้านเจตคติกับสมรรถนะโดยรวมมีความสัมพันธ์ต่ำสุด

5.2 ผลการคำนวณประสิทธิภาพการทำนายผลของวิธีการจำแนกข้อมูล ประสิทธิภาพการจำแนกข้อมูล ด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคเกรเดียนท์บูตทรีส เทคนิคแรนดอมฟอร์เรสต์ เทคนิคนาอีฟเบย์ และเทคนิคการเรียนรู้เชิงลึก โดยคำนวณค่าความแม่นยำ ค่าเรียกกลับ ค่าความจำเพาะ ค่าความถูกต้อง ค่าประสิทธิภาพโดยรวมและค่าเฉลี่ยเรขาคณิต ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์พบว่าเทคนิคแรนดอมฟอร์เรสต์มีประสิทธิภาพการจำแนกข้อมูลมากสุด เมื่อแบ่งชุดข้อมูลออกเป็น 30 ส่วน มีค่าความถูกต้อง 76.29% ค่าประสิทธิภาพโดยรวม 76.00% และค่าเฉลี่ยเรขาคณิต 76.28% จากผลการวิจัยพบว่าค่าความถูกต้องมีค่าไม่สูงมากนัก อาจจะปรับพารามิเตอร์บางตัวของแต่ละแบบจำลองการพยากรณ์หรือออกแบบ และพัฒนาแบบจำลองการพยากรณ์ด้วยขั้นตอนวิธีที่ผู้วิจัยพัฒนาขึ้นเอง เพื่อให้เหมาะสมและสอดคล้องกับข้อมูลผลการทดสอบทักษะความเข้าใจและใช้เทคโนโลยีดิจิทัลที่ได้จัดเก็บไว้

6. ข้อเสนอแนะ

6.1 ควรเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกข้อมูลและสังเคราะห์ผลข้อมูลที่จำแนกถูกและข้อมูลที่จำแนกไม่ถูก หลังจากการแก้ปัญหาความไม่สมดุลข้อมูลแต่ละวิธี คือ วิธีการสุ่มเพิ่ม (Oversampling) การสุ่มลด (Undersampling) วิธีผสมผสาน (Hybrid method) และการสังเคราะห์ข้อมูลใหม่ (Synthetic Minority Oversampling TEchnique: SMOTE)

6.2 ควรเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกข้อมูลที่ไม่มีการคัดเลือกคุณลักษณะ และ/หรือแสดงผลการคำนวณประสิทธิภาพของโมเดลที่เกิดจากจำนวนการใช้คุณลักษณะที่แตกต่างกันหรือกำหนดระดับ Threshold ขึ้นมาในแต่ละเทคนิคการคัดเลือกคุณลักษณะ เพื่อกำจัดคุณลักษณะที่มีค่าดัชนีของความสำคัญน้อยกว่าระดับที่อ้างอิงออกไป

7. เอกสารอ้างอิง

[1] R. Ratanabanchuen. *The digital economy and the readiness of Thai households through the study of "Digital literacy"*. Available Online at https://www.pier.or.th/abridged/2021/04/, accessed on 15 October 2022.

[2] National Statistical Office, Ministry of Digital Economy and Society. *The 2022 Household Survey on the Use of Information and Communication Technology, National Statistical Office,* 2022.

[3] Office of National Digital Economy and Society Commission, *Media and Information Literacy Summary Survey Report Thailand 2019,* Ministry of Digital Economy and Society, 2022.

[4] P. Kularb. "The Promotion of Media, Information, and Digital Literacy for Children and Youth: Case Studies from South Korea, Singapore, and the United Kingdom." *Journalism,* Vol. 13, No. 2, pp. 130, May, 2020.

[5] A. Boonyoo. "What is Digital Literacy." *Journal of the Department of Science Service, Ministry of Science and Technology,* Vol. 66, No. 207, pp. 28-29, May, 2018.

[6] P. Thanathamathee, and Y. Sirisathitkul. "Improved Classification Techniques for Imbalanced Data Sets of Elderly's Knee Osteoarthritis." *Science and Technology Journal,* Vol. 27, No. 6, pp.1164 - 1178, November - December, 2019.

[7] P. Laopilai, and C. Sanrach. "Analysis for Student Dropout in Undergraduate Using Data Mining Technique." *The Science Journal of Phetchaburi Rajabhat University,* Vol. 16, No. 2, pp. 61-71, 2019.

[8] S. Sinlapasorn. "Modeling used to predict osteoarthritis scores WOMAC of Post-Knee Replacement Surgery Patients with Characteristic Engineering and Machine Learning Techniques." *Proceedings of the 25th Annual Meeting in Mathematics (AMM 2021)*, 2021.

[9] B. Leo. "Random Forests." *Machine Learning,* vol. 45, No. 1, pp. 5 - 32, 2001.

[10] A. Chutipascharoen, and C. Sanrach. "A Comparison of the Efficiency of Algorithms and Feature Selection Methods for Predicting the Success of Personal Overseas Money Transfer." *KKU Research Journal of Humanities and Social Sciences,* Vol. 3, No. 6, pp. 105 - 113, 2018.

[11] ABB Corporate Research Center. *What is Deep Learning?*. Available Online at www.new.abb.com/news/detail/58004/deep-learning, accessed on 1 November 2021.

[12] S. Chayatummagoon, and P. Chongstitvatana. "Image classification of sugar crystal with deep learning." *2021 13th International Conference on Knowledge and Smart Technology (KST),* Bangsaen, Chonburi, Thailand, pp. 118 - 122, 2022.

[13] K. Haruehansapong. "Digital Literacy of Undergraduate Students at Walailak University." *Walailak Journal of Learning Innovations,* Vol. 5, No. 2, pp. 27 - 40, December, 2019.

[14] W. Kesornsit, J. Jitthavech and V. Lorchirachoonkul. "Reducing Multiclass to Binary for Classification Techniquesof Re-Hospitalization of Diabetes Patients." *Thai Science and Technology Journal (TSTJ),* Vol. 28, No. 1, January, 2020.

[15] W. Jaidee, and N. Wannapee. "The Study of Factors Affecting for On-time Graduation of Ungraduated Student Using Feature Selection Technique on Imbalanced Datasets." *Journal of Information Science and Technology (JIST),* Vol. 10, No. 1, pp. 75 - 84, June, 2020.

[16] S. Suepitak, B. Nissaidee, and W. Buathong. "Classification Techniques Comparison for Predicting Graduate Trend." *PKRU SciTech J.,* Vol. 5, No. 2, pp. 42 - 50, December, 2021.

[17] A. Kittitanachai. *The using Information and Technology behavior which Relates for capability of Srinakarinwirot University Prasarnmit Demonstration School secondary,* Presented in Partial Fulfillment of the Requirements for the Master of Education Degree in Educational Technology at Srinakharinwirot University, 2012.

[18] T. Phiasai, and N. Chinpanthana. "Comparison of Feature Selection Method with ReliefF base Multi Label Algorithm to Improve Semantic Image Classification." *Journal of Information Science and Technology (JIST),* Vol. 11, No. 1, pp. 88 - 96, June, 2021.

[19] V. Nuipian, and P. Meesut. "Comparison of filtering and consolidation techniques of text mining for text classification." *The Journal of Industrial Technology,* Vol. 9, No. 3, pp.118 - 129, September - December, 2013.

[20] K. Rothjanawan, and W. Phetjirachotkul. "Feature Selection Methods for Imputation Missing Values of Time Series Data using Data Mining." *Princess of Naradhiwas University Journal (PNUJR),* Vol. 13, No. 2, pp. 326 - 341, May, 2021.