การจำแนกการตัดสินใจซื้อจักรยานของลูกค้าโดยใช้เทคนิควิธีเอนเซมเบิล

อนุสรา กุลสุทธิ์[1], ภูริชญ์ มหาวงศ์[1] และ มงคล แสนสุข[1*]
[1]หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลัยราชภัฏสกลนคร

บทคัดย่อ(ภาษาไทย)

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเกี่ยวกับการวิเคราะห์ปัจจัยและจำแนกการตัดสินใจซื้อจักรยานของลูกค้าโดยใช้เทคนิควิธีเอนเซมเบิล โดยการสร้างโมเดลสำหรับการจำแนกและเปรียบเทียบประสิทธิภาพของเทคนิควิธีการจำแนกประเภทแบบเดี่ยว ได้แก่ เทคนิควิธีต้นไม้ตัดสินใจ นาอีฟเบย์ และเพื่อนบ้านใกล้เคียง และเทคนิควิธีการจำแนกประเภทแบบเชิงกลุ่ม ได้แก่ เทคนิควิธี Boosting (XGBoost) โดยใช้ข้อมูล Bike Buying Prediction For Adventure Works Cycles ที่ถูกเผยแพร่โดย Rahul จากเว็บไซต์ Kaggle ในการทดลองได้ทำการเตรียมข้อมูลโดยการคัดเลือกข้อมูลเบื้องต้นประกอบด้วย 13 คุณลักษณะ จากนั้นทำการตรวจสอบข้อมูลสูญหายพบว่าชุดข้อมูลนี้ไม่พบข้อมูลที่สูญหาย แล้วทำการแปลงข้อมูลให้อยู่ในรูปแบบตัวเลขแทนความหมาย คัดเลือกคุณลักษณะที่ส่งผลกับคลาสด้วยเทคนิควิธี Information Gain และคัดเลือกค่าถ่วงน้ำหนัก จากนั้นทำการแบ่งข้อมูลด้วยเทคนิควิธี Holdout Method จากการทดลองคัดเลือกค่าถ่วงน้ำหนักพบว่าควรกำจัดคุณลักษณะ HomeOwnerFlag ซึ่งจะทำให้โมเดล Boosting (XGBoost) มีค่าประสิทธิภาพถ่วงดุลสูงที่สุดร้อยละ 75.27

***ผู้เขียนหลัก: Mongkolsaensuk@gmail.com**

คำสำคัญ: จักรยาน, เหมืองข้อมูล, การวิเคราะห์ปัจจัย, การจำแนกประเภทข้อมูล

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY

Classification of Customer Decision for Bicycle Purchase by Using the Ensemble Technique

Anutsara Kunlasut[1], Bhurich Mahawong[1] and Mongkol Saensuk[1*]

[1]Bachelor of Science Data Science program, Faculty of Science and Technology,
Sakon Nakhon Rajabhat University.

Abstract

The purpose of this research is to study factor analysis and classification of customer decisions for bicycle purchase decisions by using the ensemble technique. Single classification techniques are Decision Tree, Nave Bayes, and the K-Nearest Neighbors technique, which are constructed by building a model for classification and comparing performance. The Boosting (XGBoost) technique is used for cluster classification. Using information from "Bike Buying Prediction For Adventure Works Cycles" published by Rahul on the Kaggle website, A preliminary selection of 13 features was used to prepare the data for the experiment. After checking for lost data, it was found that this dataset did not contain any lost data. Then convert the data into numeric form instead of meaning by assigning a value to meaning. and selecting the features that conform to the class in data with information gain and adjustment with weights. Then divide the data by the holdout method. Weighted experiments revealed that the HomeOwnerFlag feature should be removed. This gives the Boosting (XGBoost) model the highest f-measure of 75.27

***Corresponding Author**: Mongkolsaensuk@gmail.com

Keywords: Bicycle, Data Mining, Factor Analysis, Data Classification

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY

1. บทนำ

ปัจจุบันการแพร่ระบาดของสถานการณ์โควิด-19 ทำให้อุตสาหกรรมบางประเภทต้องขาดทุนหรือปิดตัวลง ซึ่งสวนทางกับรายงานยอดขายจักรยานทั่วโลกเพิ่มขึ้นกว่าร้อยละ 200 เนื่องจากประชาชนลดความเสี่ยงจากการเดินทางด้วยขนส่งสาธารณะด้วยการเลือกที่จะเดินทางโดยจักรยานมากขึ้น ส่งผลให้ยอดขายเพิ่มขึ้นอย่างรวดเร็ว [1] แนวโน้มความต้องการที่มากขึ้นทำให้เกิดการแข่งขันเพื่อครอบครองส่วนแบ่งทางการตลาดของธุรกิจจักรยาน ซึ่งธุรกิจใดที่สามารถครอบครองส่วนแบ่งทางการตลาดเป็นส่วนใหญ่และถือครองส่วนแบ่งในระยะยาวจะส่งผลให้ธุรกิจนั้นประสบความสำเร็จ [2] จากการศึกษาพฤติกรรมการใช้จักรยานเพื่อทำความเข้าใจความต้องการของลูกค้า โดยศึกษาจากข้อมูลพฤติกรรมการซื้อจักรยานเพื่อนำมาสร้างเป็นกลยุทธ์ทางธุรกิจจะนำไปสู่การเพิ่มส่วนแบ่งทางการตลาด โดยเฉพาะธุรกิจจักรยานที่เป็นธุรกิจขนาดย่อมจะเห็นได้ว่าส่วนใหญ่ยังไม่มีการนำข้อมูลพฤติกรรมการซื้อจักรยานของลูกค้ามาวิเคราะห์ใช้ประโยชน์ ส่งผลให้เกิดปัญหาการโฆษณา/ประชาสัมพันธ์ไม่ตรงกลุ่มเป้าหมาย เสียเวลาและต้นทุนในการโฆษณาประชาสัมพันธ์ หรืออาจจะเกิดปัญหาด้านการขยายสาขาที่ต้องแบกรับต้นทุนสูง ซึ่งหากไม่ศึกษาข้อมูลพฤติกรรมการซื้อของลูกค้าเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการตัดสินใจซื้อจักรยาน อาจทำให้สาขาขยายเปิดใหม่ดังกล่าวปิดกิจการไปในที่สุด Prada [3] นำเสนองานวิจัยเกี่ยวกับการจำแนกการตัดสินใจซื้อจักรยาน เพื่อให้เข้าใจพฤติกรรมการใช้จักรยานของลูกค้า โดยใช้ชุดข้อมูล Bike Buyer ประกอบไปด้วย 13 คุณลักษณะ ID, Marital Status, Gender, Income Children, Education, Occupation, Home Owner, Cars, Commute, Distance, Region, Age และ คลาส Purchased Bike และทำการเปรียบเทียบเทคนิคการจำแนกประเภท จากผลการทดลองพบว่า Support Vector Machine และ Random Forest มีค่าความถูกต้องสูงที่สุดคือ ร้อยละ 86 Kabir, Ashraf และ Ajwad [4] ได้นำเสนองานวิจัยเกี่ยวกับพฤติกรรมของลูกค้าออนไลน์ โดยวิเคราะห์ข้อมูลการตัดสินใจซื้อสินค้าของลูกค้าที่เยี่ยมชมหน้าเว็บของร้านค้าออนไลน์ ได้รวบรวมข้อมูลจากเซสชันของผู้ใช้งานในร้านค้าออนไลน์แห่งหนึ่งจำนวน 12,330 ข้อมูล และประกอบด้วย 10 คุณลักษณะ กำหนดคลาสข้อมูลออกเป็น 2 คลาส ได้แก่ คนที่ไม่ได้ซื้อสินค้า และคนที่ซื้อสินค้า จากการทดลองพบว่าเทคนิควิธี Gradient Boosting มีค่าความถูกต้องร้อยละ 90.34 ธีรยุทธ คูณสุข และจารี ทองคำ [5] นำเสนองานวิจัยเกี่ยวกับกระบวนการคัดเลือกคุณลักษณะสำหรับเพิ่มประสิทธิภาพในการจำแนกความคิดเห็นของลูกค้า งานวิจัยนี้ได้ทำการรวบรวมข้อมูลจากเว็บไซต์ wongnai.com เป็นความคิดเห็นที่มีคะแนนจากผู้แสดงความคิดเห็นในระดับที่ 1 - 5 จะใช้เฉพาะความคิดเห็นระดับ 1, 2, 4 และ 5 กำหนดให้ข้อมูลความคิดเห็นในระดับที่ 1 และ 2 เป็นความคิดเห็นเชิงลบ ความคิดเห็น ในระดับที่ 4 และ 5 เป็นความคิดเห็นเชิงบวก หลังจากนั้นทำการตัดคำ ผลการทดลองพบว่าโมเดลนาอีฟเบย์เมื่อทำการแบ่งข้อมูลเป็น 10-fold Cross Validation ใช้การคัดเลือกคุณลักษณะด้วยเทคนิควิธี Information Gain ให้ผลลัพธ์ดีที่สุด มีค่าความถูกต้องร้อยละ 89.08 มีค่าความแม่นยำร้อยละ 80.12 และมีค่าความระลึกร้อยละ 89.10 ปริยานุช ประเสริฐสิริกุล ศิริสรรพ เหล่าหะเกียรติ เรืองศักดิ์ ตระกูลพุทธิรักษ์ และ ศศิวิมล สุขพัฒน์ [6] นำเสนอวิจัยเกี่ยวกับการเปรียบเทียบประสิทธิภาพโมเดลการทำนายผลการเรียนของนิสิตที่ใช้งานระบบการจัดการเรียนรู้ออนไลน์ด้วยการเรียนรู้ของเครื่อง ได้รวบรวมข้อมูลของนิสิตระดับปริญญาตรีชั้นปีที่ 1 มหาวิทยาลัยศรีนครินทรวิโรฒ จำนวน 405 คน ที่ลงทะเบียนรายวิชา SWU 141 ชีวิตในโลกดิจิทัล ประกอบไปด้วย 4 คุณลักษณะ ได้แก่ ด้านการเข้าใช้งานระบบ ด้านการชมสื่อการเรียนรู้ ด้านการทำแบบทดสอบ และด้านการส่งงานมอบหมาย คลาสในการทำนายคือ ผลการเรียน ได้ทำการแบ่งข้อมูลด้วยวิธี Holdout Method ปรับสมดุลข้อมูลด้วยเทคนิค SMOTE โดยปรับจำนวนข้อมูลในส่วนนิสิตที่มีผลการเรียนระดับดีมาก และระดับต้องปรับปรุงให้มีความสมดุลโดยใช้วิธีสุ่มเพิ่ม จากการทดลองพบว่าเทคนิควิธี XGBoost มีค่าความถูกต้องสูงที่สุดร้อยละ 83.95

SCIENCE AND TECHNOLOGY
UTARADIT RAJABHAT UNIVERSITY

จากปัญหาที่กล่าวข้างต้น คณะผู้วิจัยจึงเกิดแนวคิดในการศึกษาปัจจัยที่ส่งผลต่อการตัดสินใจซื้อจักรยาน และจำแนกข้อมูลการตัดสินใจซื้อจักรยานของลูกค้าจากผลการวิเคราะห์ปัจจัยด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งจะส่งผลให้ธุรกิจสามารถ

ทราบถึงปัจจัยที่ส่งผลต่อการตัดสินใจซื้อของลูกค้า และสามารถใช้โมเดลในการจำแนกการการตัดสินใจซื้อจักรยานของลูกค้า เพื่อนำไปสู่การโฆษณา/ประชาสัมพันธ์ที่ตรงกลุ่มเป้าหมาย สร้างความได้เปรียบคู่แข่งขัน และมีโอกาสในการเพิ่มส่วนแบ่งทางการตลาดและช่วยให้ลูกค้าสามารถตัดสินใจในการซื้อจักรยานได้สะดวกมากขึ้น

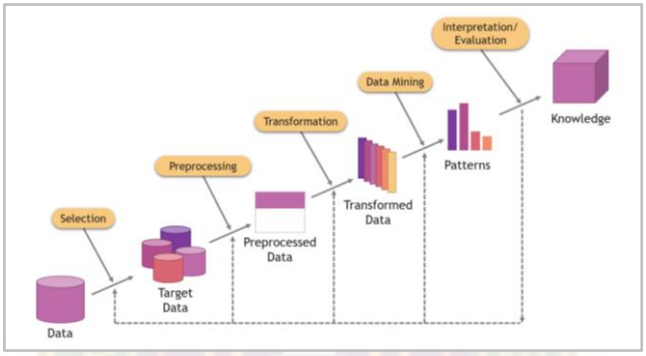
2. ความรู้พื้นฐาน

2.1 จักรยาน (Bicycle)

จักรยานคันแรกของโลกคิดค้นขึ้นในปี พ.ศ. 2382 โดยชาวสกอตแลนด์ชื่อ เคริกพาทริก แมกมิลลัน จักรยานที่ได้ประดิษฐ์ขึ้นมีชื่อเรียกทั่วไปว่า “MacMillan Velocipede” เป็นพาหนะที่ขับเคลื่อนด้วยล้อหลัง โดยการวางเท้าไว้ที่บันไดและยกเท้าขึ้น-ลง เพื่อให้บันไดที่เชื่อมติดกับข้อเหวี่ยงของล้อหลัง เกิดการเคลื่อนที่ทำให้ล้อหมุนและเคลื่อนตัวไปข้างหน้าได้ ทั้งนี้ตัวจักรยานยังมีน้ำหนักมากถึง 26 กิโลกรัม จึงไม่ได้รับความนิยมมากนัก ต่อมาจักรยานได้รับการพัฒนาอย่างต่อเนื่องเพื่อให้มีรูปทรงที่ทันสมัยขับขี่ง่ายขึ้น จนกระทั่งในปี พ.ศ. 2403 เอเนสต์ มีโชซ์ และปีแยร์ ลาลเลอมอง ชาวฝรั่งเศสได้พัฒนาจักรยานรุ่น “Velocipede” โดยติดข้อเหวี่ยงและบันไดเข้ากับล้อหน้าเพื่อใช้ในการขับเคลื่อน ซึ่งล้อทำขึ้นจากไม้หุ้มด้วยแผ่นเหล็ก โครงรถทำด้วยท่อเหล็กกลวงทำให้สามารถผลิตได้ง่าย และรวดเร็วกว่ารุ่นก่อน การพัฒนาดังกล่าวได้รับความนิยมเป็นอย่างมาก เป็นเหตุให้มีโชซ์ก่อตั้งโรงงานประกอบจักรยานแห่งแรกของโลกขึ้นในชื่อ “Michaux และ Lallement” สำหรับประเทศไทยเริ่มมีจักรยานใช้เป็นครั้งแรกในสมัยรัชกาลที่ 5 โดยเจ้าพระยาภาสกรวงศ์ (พร บุนนาค) ซึ่งได้นำจักรยานเข้ามาถวายแก่พระบาทสมเด็จพระจุลจอมเกล้าเจ้าอยู่หัวหลังจากการเดินทางกลับจากการปฏิบัติหน้าที่พิเศษที่กรุงลอนดอน จักรยานที่นำเข้ามามีลักษณะล้อหน้าใหญ่ล้อหลังเล็กก่อนหน้าที่จะพัฒนามาเป็นจักรยานล้อเท่ากัน [7] จักรยานมีหลากหมายประเภท เช่น จักรยานเสือภูเขา (Mountain Bike) จักรยานเสือหมอบ (Road Bike) จักรยานไซโครคลอส (CycloCross) จักรยานไฮบริด (Hybrid Bike) จักรยานล้อโต (Fat Bike) จักรยานทั่วร์ริ่ง (Touring Bike) จักรยานพับได้ (Folding Bike) จักรยานฟิกเกียร์ (Fixed Gear) จักรยานซิงเกิล สปีด (Single Speed) จักรยานมินิ (Mini Bike) จักรยานบีเอ็มเอ็กซ์ (BMX) จักรยานแม่บ้าน เป็นต้น

2.2 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลเป็นเทคนิควิธีเพื่อค้นหาองค์ความรู้ที่มีประโยชน์จากแหล่งข้อมูลจำนวนมาก การทำเหมืองข้อมูลจะใช้ศาสตร์ด้านสถิติ (Statistics) การเรียนรู้ของเครื่องจักร (Machine learning) และการรู้จำรูปแบบ (Pattern recognition) หรือในอีกนิยามหนึ่งการทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูลเพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น เทคนิคการทำเหมืองข้อมูลมีหลายวิธี เช่น กฎความสัมพันธ์ (Association Rules) คือ การค้นหาความสัมพันธ์ของเหตุการณ์หรือวัตถุที่เกิดขึ้นพร้อมกัน เทคนิคการแบ่งกลุ่มข้อมูล (Data clustering) คือ การแบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่ม เช่น จัดกลุ่มลูกค้าจากพฤติกรรมการซื้อสินค้าของลูกค้าที่มีลักษณะคล้ายกัน สำหรับเทคนิคการจำแนกประเภทข้อมูล (Data classification) คือ การสร้างโมเดลสำหรับการจำแนกข้อมูล ซึ่งโมเดลที่ได้จะเกิดจากการเรียนรู้รูปแบบของข้อมูลแต่ละคลาส เช่น การจำแนกประเภทลูกค้าที่ยังไม่ตัดสินใจซื้อสินค้า และลูกค้าที่ตัดสินใจซื้อสินค้า ซึ่งโมเดลที่ได้จะช่วยจำแนกคลาสของข้อมูล โดยโมเดลที่ใช้จำแนกข้อมูลจะสร้างโมเดลจากข้อมูลชุดสอน (Training data) โดยข้อมูลชุดสอนประกอบด้วยคุณลักษณะ (Features) และคลาส และนำข้อมูลชุดทดสอบ (Test data) ซึ่งเป็นข้อมูลที่ทราบคลาสมาทดลองจำแนกผลลัพธ์ ถ้าผลลัพธ์ตรงกับค่าความเป็นจริงแสดงให้เห็นถึงโมเดลมีความแม่นยำ แต่ถ้าโมเดลจำแนกคลาสผลผิดพลาดก็จะทำการปรับปรุงขั้นตอนต่าง ๆ ให้มีความถูกต้องมากขึ้นจนได้โมเดลที่สามารถทำนายผลในระดับที่ยอมรับได้ [8] จึงจะสามารถนำเข้าสู่ขั้นตอนการทำเหมืองข้อมูลดังภาพที่ 1



ภาพที่ 1 ขั้นตอนการทำเหมืองข้อมูล [9]

จากภาพที่ 1 ขั้นตอนการทำเหมืองข้อมูล โดยทั่วไปการทำเหมืองข้อมูลประกอบด้วยขั้นตอนหลักดังนี้ รวบรวมข้อมูลและคัดเลือกข้อมูล จากฐานข้อมูลซึ่งจะเป็นข้อมูลเป้าหมายที่จะนำมาวิเคราะห์ตามวัตถุประสงค์ แล้วนำข้อมูลที่ผ่านการคัดเลือกมาเข้าสู่ขั้นตอนการเตรียมข้อมูล โดยการตรวจสอบและแก้ไขปัญหาด้านคุณภาพของข้อมูล จัดการแปลงข้อมูล ให้เหมาะสมกับการวิเคราะห์จากนั้นวิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูล โดยการเรียนรู้รูปแบบของข้อมูลเพื่อให้ได้โมเดล ที่บ่งบอกถึงรูปแบบของข้อมูลต่าง ๆ ๆขั้นตอนสุดท้ายคือการประเมินผลลัพธ์จากการวิเคราะห์ และนำเสนอข้อมูล [9][10]

2.2.1 การเตรียมข้อมูล (Data Preprocessing)

การเตรียมข้อมูล คือ การทำให้ข้อมูลมีคุณภาพและอยู่ในรูปแบบที่สามารถนำไปวิเคราะห์ได้ส่งผลให้ผลลัพธ์จากการทำเหมืองข้อมูลมีความถูกต้องน่าเชื่อถือและไม่เกิดความเสียหายจากการนำผลลัพธ์ไปใช้ ข้อมูลที่นำมาทำเหมืองข้อมูลอาจมีปัญหาด้านคุณภาพ [10][11] เช่น ข้อมูลบางส่วนมีค่าตัวแปรบางตัวที่ไม่ถูกต้อง (Erroneous data) เนื่องจากค่าตัวแปรอาจเกิดจากความผิดพลาดในขั้นตอนการเก็บหรือบันทึกข้อมูล โดยทั่วไปการเตรียมข้อมูลแบ่งออกเป็น 4 ขั้นตอน ได้แก่

1) การทำความสะอาดข้อมูล (Data Cleaning) คือ กระบวนการการทำงานที่เกี่ยวกับการตรวจสอบและแก้ไขข้อมูล เพื่อให้ข้อมูลอยู่ในรูปแบบที่ถูกต้องและสมบูรณ์ ขั้นตอนเบื้องต้นในการทำความสะอาดข้อมูลมี 4 ขั้นตอน ได้แก่ การแจกแจงข้อมูล การแก้ไขข้อมูลที่ผิดพลาด การทำข้อมูลให้เป็นรูปแบบเดียวกัน และการลบชุดข้อความที่ซ้ำซ้อน

2) การผสานข้อมูล (Data Integration) คือ ขั้นตอนการรวมข้อมูลจากหมายแหล่งมารวมไว้ที่เดียวกัน โดยใช้คุณลักษณะที่มีอยู่ในแต่ละแหล่งข้อมูลเป็นตัวเชื่อมโยง

3) การแปลงข้อมูล (Data Transformation) คือ ขั้นตอนการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมและสามารถนำไปใช้ในการทำเหมืองข้อมูล การแปลงข้อมูลที่พบบ่อยในการทำเหมืองข้อมูล คือ เทคนิควิธีนอร์มัลไลเซชัน (Normalization)

4) การคัดเลือกคุณลักษณะ (Feature Selection) คือ การคัดเลือกคุณลักษณะหรือคุณสมบัติที่สอดคล้องกับคลาสของข้อมูล รวมถึงการกำจัดคุณลักษณะที่ไม่ส่งสอดคล้องกับคลาสของข้อมูลก่อนนำไปสร้างโมเดล ตัวอย่างการคำนวณค่าน้ำหนักที่สอดคล้องกับคลาสข้อมูลด้วยเทคนิควิธี Information Gain [12] ดังสมการที่ 1-2

$$H(D) = -\sum_{l \in C_D} P_D(l) log_2 P_D(l) \quad (1)$$

โดยกำหนดให้ $H(D)$ คือ มาตรวัดความเจือปนของข้อมูล

C_D คือ คลาสทั้งหมดที่เป็นไปได้บนชุดข้อมูล D

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY

$P_D(l)$ คือ สัดส่วนข้อมูลที่มีคลาสเท่ากับ l ต่อข้อมูลทั้งหมดในชุดข้อมูล D

ซึ่งค่า $H(D)$ จะมีค่าน้อยเมื่อคลาสของข้อมูลทั้งหมดมีแนวโน้มเหมือนกัน (มีความเจือปนน้อย) และจะมีค่ามากขึ้นเมื่อข้อมูลมีค่าคลาสที่หลากหลายมากขึ้น

$$IG(D,X) = H(D) - \sum_{c \in vals(X)} \frac{|D_c|}{|D|} H(D_c) \quad (2)$$

โดยกำหนดให้ $IG(D,X)$ คือ ค่าเกณฑ์ความรู้ ซึ่งจะเลือกค่ามากที่สุด

$vals(X)$ คือ เซตของค่าคุณลักษณะ X ทั้งหมดที่เป็นไปได้

D_c คือ ชุดของข้อมูลที่มีค่าตัวแปร X เท่ากับ c

$P_D(l)$ คือ สัดส่วนข้อมูลที่มีคลาสเท่ากับ l ต่อข้อมูลทั้งหมดในชุดข้อมูล D

2.2.2 การจำแนกประเภทข้อมูล (Data Classification)

การจำแนกประเภทข้อมูล เป็นเทคนิคหนึ่งในการเรียนรู้ของเครื่องจักรซึ่งเป็นศาสตร์ที่ว่าด้วยอัลกอริทึมที่ทำให้เกิดการเรียนรู้ของเครื่องจักรและเข้าใจในประเด็นที่เราสนใจจากข้อมูล โดยเทคนิคการจำแนกประเภทข้อมูลจัดอยู่ในกลุ่มการเรียนรู้แบบมีผู้สอน (Supervised learning) [8] เป็นการเรียนรู้ที่ต้องอาศัยข้อมูลชุดสอน ซึ่งประกอบไปด้วยคุณลักษณะและคลาสของข้อมูล โดยข้อมูลในแต่ละคุณลักษณะอาจอยู่ในรูปแบบข้อมูลเชิงปริมาณหรือเชิงคุณภาพแต่คลาสของข้อมูลจะอยู่ในรูปแบบเชิงคุณภาพ ตัวอย่างเทคนิควิธีการ

1) เทคนิควิธีการจำแนกประเภทแบบเดี่ยว

1.1) เทคนิควิธีต้นไม้ตัดสินใจ (Decision Tree) เป็นโมเดลการจำแนกประเภทข้อมูลออกเป็นกลุ่มหรือคลาสต่าง ๆ โดยใช้คุณลักษณะของข้อมูลในการจำแนกประเภท ซึ่งมีลักษณะเป็นโครงสร้างต้นไม้หัวกลับ ประกอบไปด้วย Root Node คือ โหนดรากหรือโหนดบนสุดของต้นไม้เป็นส่วนที่มีอำนาจในการจำแนกสูงที่สุด Internal Node คือ โหนดภายในเป็นส่วนของเงื่อนไขในการตัดสินใจ Branch คือ กิ่งเป็นส่วนเชื่อมต่อระหว่างโหนดที่แสดงถึงค่าของข้อมูลในคุณลักษณะที่เป็นไปได้ทั้งหมดของแต่ละโหนด และ Leaf Node คือ โหนดใบเป็นส่วนที่แสดงผลลัพธ์ที่ได้จาเงื่อนไขการตัดสินใจ อัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจหรือคลาสของข้อมูลในการทำนาย ได้แก่ ID3 Algorithm โดยการคำนวณค่า Entropy ดังสมการที่ 1 และการคำนวณค่า IG ดังสมการที่ 2

1.2) เทคนิควิธีเพื่อนบ้านใกล้เคียง (K-Nearest Neighbors : K-NN) เป็นเทคนิควิธีการจำแนกประเภท (Classification) ใช้หลักการเปรียบเทียบข้อมูลตัวอย่างกับข้อมูลที่ต้องการทำนายว่ามีความคล้ายคลึงมากน้อยเพียงใด หากข้อมูลที่ต้องการทำนายนั้นอยู่ใกล้ข้อมูลคลาสใดมากที่สุด ผลลัพธ์การจำแนกจะเป็นคลาสของข้อมูลที่อยู่ใกล้ที่สุด ซึ่งค่า K คือ จำนวนข้อมูลตัวอย่างที่มีระยะทางสั้นที่สุดที่สนใจพิจารณาจำนวน Kข้อมูล การกำหนดค่า K นิยมกำหนดเป็นเลขคี่เพื่อป้องกันปัญหาข้อมูลเป็นได้ทั้งสองประเภท K-NN เหมาะสำหรับข้อมูลแบบตัวเลขเพื่อหาวิธีการวัดระยะห่างของแต่ละคุณลักษณะ ซึ่งสามารถคำนวณจาก Distance Function [13] ดังสมการที่ 3

$$D_{Euclidean} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (3)$$

โดยกำหนดให้ $x_1, x_2, \ldots, x_n$ คือ คุณลักษณะลำดับที่ 1 ถึง n ของข้อมูลที่ทราบคลาส

$y_1, y_2, \ldots, y_n$ คือ คุณลักษณะลำดับที่ 1 ถึง n ของข้อมูลที่ไม่ทราบคลาส

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY

n คือ จำนวนคุณลักษณะ

1.3) เทคนิควิธีนาอีฟเบย์ (Naïve Bayes) ถูกพัฒนาขึ้นโดย Thomas Bayes โดยใช้หลักการของความน่าจะเป็นแบบมีเงื่อนไขมาพัฒนาเป็นทฤษฎีบทดังกล่าว ความน่าจะเป็นของเทคนิควิธีนาอีฟเบย์จะวิเคราะห์หาความน่าจะเป็นของเหตุการณ์ที่มีโอกาสเกิดขึ้น โดยการคาดเดาจากความเป็นไปได้ทั้งหมด เช่น กำหนดให้ A และ B เป็นเหตุการณ์ใด ๆ และความน่าจะเป็นที่จะเกิดเหตุการณ์ A โดยมีเงื่อนไขว่าเหตุการณ์ B ได้เกิดขึ้นแล้ว ดังนั้นสามารถเขียนทฤษฎีเบย์ได้ดังสมการที่ 4 และสมการการทำนายผลหรือพยากรณ์ดังสมการที่ 5 [10][14]

$$P(l|X) = P(l)\prod_{i=1}^{d} P(x_i|l) \quad (4)$$

โดยกำหนดให้ $P(l)$ คือ ความน่าจะเป็นของคลาส l

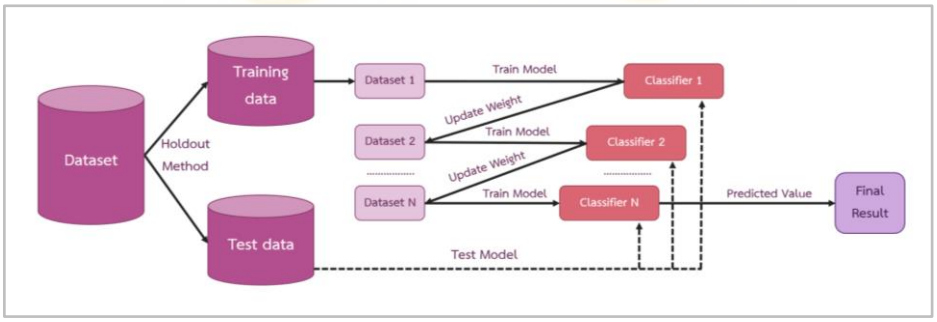
$P(x_i|l)$ คือ ความน่าจะเป็นของคุณลักษณะ x ลำดับที่ i เมื่อทราบคลาส l

$P(l|x_i)$ คือ ความน่าจะเป็นของคลาส l เมื่อทราบคุณลักษณะ X

2) เทคนิควิธีการจำแนกประเภทแบบเชิงกลุ่ม

เทคนิควิธีการจำแนกประเภทเชิงกลุ่ม หรือ เอนเซมเบิลเทคนิค คือ การสร้างโมเดลโดยอาศัยการทำงานของหลาย ๆ โมเดล ซึ่งมีเป้าหมายช่วยเพิ่มประสิทธิภาพการทำงานของโมเดล ซึ่งเพิ่มความน่าเชื่อถือได้มากกว่าโมเดลเพียงชนิดเดียว ตัวอย่างเทคนิควิธีเอนเซมเบิล ได้แก่ Bagging Stacking และ Boosting เป็นต้น

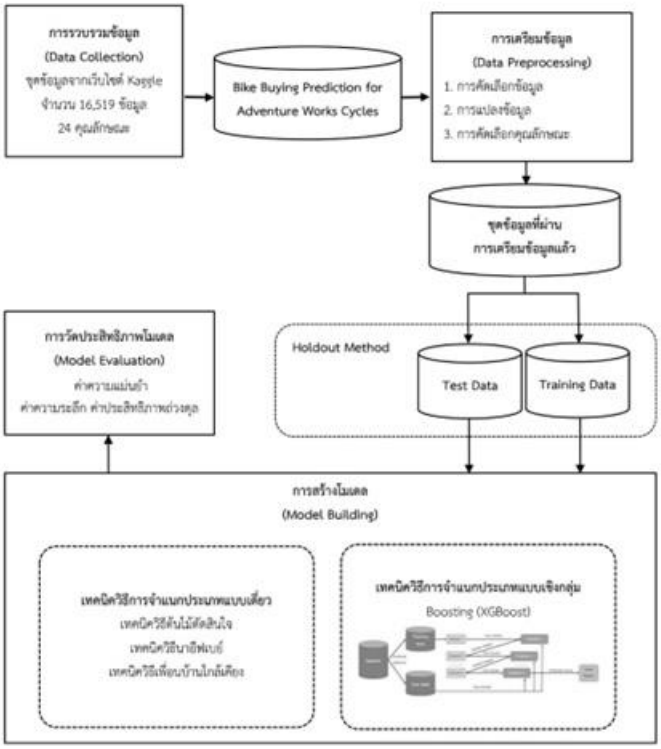
2.1) เทคนิควิธี Boosting คือ การนำข้อมูลชุดสอนไปสอนโมเดลแรก หากมีข้อผิดพลาด (Error) จะกำหนดค่าน้ำหนัก (Weight) ให้กับข้อมูล จะได้ชุดข้อมูลใหม่ ซึ่งข้อมูลผิดพลาดหาได้จาก MSE (Mean Square Error) และ Cross Entropy จากนั้นนำชุดข้อมูลใหม่ไปสอนกับโมเดลถัดไป หากยังพบข้อมูลผิดพลาดให้ทำการปรับค่าน้ำหนัก แล้วทำเช่นนี้ไปเรื่อย ๆ จนกว่าการทำนายผลจะถูกต้องดังภาพที่ 2 สำหรับโมเดลที่ใช้การเรียนรู้แบบ Boosting คือ XGBoost, AdaBoost และ Stochastic Gradient Boosting เป็นต้น [15][16]



ภาพที่ 2 กระบวนการทำงานของ Boosting

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY

3. วิธีการดำเนินการ (Materials and Methods)



ภาพที่ 3 วิธีการดำเนินงานวิจัย

3.1 การรวบรวมข้อมูล (Data Collection)

การทดลองวิเคราะห์ปัจจัยและสร้างโมเดลเพื่อจำแนกข้อมูลการตัดสินใจซื้อจักรยาน ได้นำชุดข้อมูล Bike Buying Prediction For Adventure Works Cycles [18] มาประยุกต์ใช้ในการทำเหมืองข้อมูล ชุดข้อมูลดังกล่าวได้ถูกเผยแพร่โดย Rahul จากเว็บไซต์ Kaggle เป็นข้อมูลเชิงคุณภาพและข้อมูลเชิงปริมาณจำนวน 16,519 ข้อมูล ประกอบด้วย 24 คุณลักษณะ ได้แก่ CustomerID, Title, PhoneNumber, FirstName, MiddleName, LastName, Suffix, Gender, AddressLine, AddressLine2, City, StateProvinceName, CountryRegionName, PostalCode, BirthDate, Education, MaritalStatus, HomeOwnerFlag, NumberCarsOwned, NumberChildrenAtHome, AveMonthSpend, TotalChildren และ YearlyIncome และคลาสของข้อมูลประกอบไปด้วย 2 คลาส ได้แก่คลาส 1 และ 0 โดยที่คลาส 1 แทนซื้อจักรยานจำนวน 5,489 ข้อมูล และคลาส 0 แทนไม่ซื้อจักรยานจำนวน 11,030 ข้อมูล โดยชุดข้อมูลดังกล่าวถูกเก็บรวบรวมจากบริษัท Adventure Works Cycles เมื่อปี พ.ศ. 2541 จากนั้นได้นำเข้าสู่ขั้นตอนการเตรียมข้อมูลโดยมีขั้นตอนดังต่อไปนี้

3.2 การเตรียมข้อมูล (Data Preprocessing)

3.2.1 การคัดเลือกข้อมูล

เมื่อทำการรวบรวมข้อมูลเสร็จสมบูรณ์ ได้ทำการคัดเลือกข้อมูลเบื้องต้นประกอบด้วย 13 คุณลักษณะ จาก 24 คุณลักษณะ ได้แก่ City, StateProvinceName, CountryRegionName, Education, Occupation, Gender, MaritalStatus, HomeOwnerFlag, NumberCarsOwned, NumberChildrenAtHome, TotalChildren, AveMonthSpend และ YearlyIncome เนื่องจาก 11 คุณลักษณะที่เหลือเป็นคุณลักษณะที่เฉพาะไม่ซ้ำกันจึงไม่นำมาวิเคราะห์ ได้แก่คุณลักษณะ

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY

CustomerID, Title, FirstName, MiddleName, LastName, Suffix, AddressLine1, AddressLine2, PostalCode, PhoneNumber และ BirthDate จากนั้นทำการตรวจสอบข้อมูลสูญหายพบว่าชุดข้อมูลนี้ไม่พบข้อมูลที่สูญหาย

3.2.2 การแปลงข้อมูล

ทำการแปลงข้อมูลคุณลักษณะ StateProvinceName CountryRegionName Gender City Education Occupation และ MaritalStatus ให้อยู่ในรูปแบบตัวเลขแทนความหมาย ส่วนคุณลักษณะ TotalChildren NumberChildrenAtHome HomeOwnerFlag NumberCarsOwned AveMonthSpend และ YearlyIncome เป็นข้อมูลตัวเลขซึ่งไม่ต้องทำการแปลงข้อมูล

3.2.3 การคัดเลือกคุณลักษณะ

คัดเลือกคุณลักษณะหรือปัจจัยที่สอดคล้องกับคลาสของข้อมูลการคำนวณค่าถ่วงน้ำหนักด้วยเทคนิควิธี Information Gain และกำจัดคุณลักษณะที่ไม่ส่งผลกับคลาสของข้อมูลตามช่วงค่า Threshold

3.3 การสร้างโมเดล (Model Building)

การสร้างโมเดลสำหรับการจำแนกการตัดสินใจซื้อจักรยานของลูกค้าและทำการแบ่งข้อมูลด้วยวิธี Holdout Method จะแบ่งข้อมูลเป็นข้อมูลชุดสอนร้อยละ 80 และข้อมูลชุดทดสอบร้อยละ 20 จากนั้นสร้างโมเดลการจำแนกประเภทแบบเดี่ยว ได้แก่ เทคนิควิธีต้นไม้ตัดสินใจ นาอีฟเบย์ และเพื่อนบ้านใกล้เคียง และสร้างโมเดลการจำแนกประเภทแบบเชิงกลุ่ม ได้แก่ เทคนิควิธี Boosting (XGBoost)

3.4 การวัดประสิทธิภาพโมเดล (Model Performance Measurement)

การวัดประสิทธิภาพโมเดล คือ การวัดประสิทธิภาพผลลัพธ์ที่ได้จากการทำนายข้อมูลทดสอบ ถ้าข้อมูลทดสอบมีผลการทำนายตรงกับค่าความจริงแสดงให้เห็นว่าโมเดลมีประสิทธิภาพในการทำนายผล โดยสามารถวัดได้ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าประสิทธิภาพถ่วงดุล (F-Measure) [17] ซึ่งสามารถคำนวณได้จากตารางที่ 3.1 Confusion Matrix ดังสมการที่ 6-10

ตารางที่ 3.1 Confusion Matrix

		Predicted Value	
		0	1
Actual Value	0	$True_0$	$False_1$
	1	$False_0$	$True_1$

เมื่อ

$True_0$ คือ จำนวนข้อมูลที่ทำนายถูกต้องว่าเป็นคลาส 0

$True_1$ คือ จำนวนข้อมูลที่ทำนายถูกต้องว่าเป็นคลาส 1

$False_0$ คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาส 0 แต่ค่าความจริงคือคลาส 1

$False_1$ คือ จำนวนข้อมูลที่ทำนายผิดว่าเป็นคลาส 1 แต่ค่าความจริงคือคลาส 0

การวัดประสิทธิภาพค่าความแม่นยำของคลาส 1 และคลาส 0 ดังสมการที่ 6 และ 7

$$Precision_1 = \frac{True_1}{True_1 + False_1} \quad (6)$$

$$Precision_0 = \frac{True_0}{True_0 + False_0} \quad (7)$$

การวัดประสิทธิภาพค่าความระลึกของคลาส 1 และคลาส 0 ดังสมการที่ 8 และ 9

$$Recall_1 = \frac{True_1}{True_1 + False_0} \quad (8)$$

$$Recall_0 = \frac{True_0}{True_0 + False_1} \quad (9)$$

การวัดประสิทธิภาพค่าประสิทธิภาพถ่วงดุลดังสมการที่ 10

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (10)$$

4. ผลการวิจัยและอภิปรายผล

งานวิจัยนี้ได้ทำการทดลองศึกษาปัจจัยและจำแนกการตัดสินใจซื้อจักรยานของลูกค้าโดยใช้เทคนิควิธีเอนเซมเบิล สร้างโมเดลสำหรับจำแนกการตัดสินใจซื้อจักรยานของลูกค้าจากปัจจัยที่ได้จากการวิเคราะห์ด้วยเทคนิควิธีต้นไม้ตัดสินใจ นาอีฟเบย์ เพื่อนบ้านใกล้เคียง และ Boosting (XGBoost) และเปรียบเทียบประสิทธิภาพของโมเดลด้วย 4 ตัวชี้วัด ได้แก่ ค่าความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพถ่วงดุล ทำการรวบรวมข้อมูล Bike Buying Prediction For Adventure Works Cycles ที่ได้ถูกเผยแพร่โดย Rahul จากเว็บไซต์ Kaggle เป็นข้อมูลเชิงคุณภาพและข้อมูลเชิงปริมาณจำนวน 16,519 ข้อมูล

ในการทดลองศึกษาปัจจัยที่ส่งผลต่อการตัดสินใจซื้อจักรยานของลูกค้า ได้ใช้วิธีถ่วงน้ำหนักของคุณลักษณะที่มีผลต่อคลาสการตัดสินใจซื้อจักรยานของลูกค้าด้วยเทคนิควิธี Information Gain โดยมีผลลัพธ์ค่าน้ำหนักของแต่ละคุณลักษณะดังตารางที่ 4.1 ผลลัพธ์ค่าถ่วงน้ำหนักของคุณลักษณะด้วยเทคนิควิธี Information Gain

ตารางที่ 4.1 ผลลัพธ์ค่าถ่วงน้ำหนักของคุณลักษณะด้วยเทคนิควิธี Information Gain

คุณลักษณะ	ค่าถ่วงน้ำหนัก
HomeOwnerFlag	0.000, T1
CountryRegionName	0.004, T2
Education	0.004, T2
StateProvinceName	0.007, T3
Gender	0.010
Occupation	0.022
City	0.024
MaritalStatus	0.031
TotalChildren	0.042
NumberCarsOwned	0.048
AveMonthSpend	0.142
NumberChildrenAtHome	0.158
YearlyIncome	0.893

จากตารางที่ 4.1 ผลลัพธ์ค่าถ่วงน้ำหนักของคุณลักษณะต่าง ๆ ด้วยเทคนิควิธี Information Gain แสดงให้เห็นถึงค่าถ่วงน้ำหนักของคุณลักษณะต่าง ๆ โดยมีช่วงค่าระหว่าง 0 ถึง 1 ซึ่งค่าถ่วงน้ำหนักเข้าใกล้ 0 แสดงให้เห็นถึงคุณลักษณะไม่มีความ

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY

สอดคล้องกับคลาส และเมื่อค่าถ่วงน้ำหนักเข้าใกล้ 1 แสดงให้เห็นถึงคุณลักษณะมีความสอดคล้องกับคลาสหรือมีอำนาจในการจำแนกคลาส ดังนั้นจากตารางที่ 4.1 ดังกล่าว แสดงให้เห็นคุณลักษณะ HomeOwnerFlag, CountryRegionName, Education และ StateProvinceName มีค่าถ่วงน้ำหนักระหว่าง 0.000 ถึง 0.007 ซึ่งมีค่าค่อนข้างต่ำหรือมีความสัมพันธ์กับคลาสค่อนข้างน้อย เนื่องจากจักรยานมีหลากหลายประเภทและมีลักษณะที่เหมาะสมครอบคลุมทุกระดับการศึกษา ดังนั้นระดับการศึกษาจึงไม่ส่งผลต่อการตัดสินใจซื้อจักรยานกรรมสิทธิ์บ้าน ประเทศและจังหวัดก็ไม่ส่งผลในการตัดสินใจซื้อจักรยานเช่นเดียวกัน เนื่องจากผู้ที่มีกรรมสิทธิ์ในบ้านหรือเป็นผู้เช่าอาศัยก็สามารถตัดสินใจซื้อหรือไม่ซื้อจักรยานได้ ดังนั้นจึงได้กำหนดให้คุณลักษณะ HomeOwnerFlag อยู่ในช่วงค่า Threshold ที่ T1 คุณลักษณะ CountryRegionName และ Education อยู่ในช่วงค่า Threshold ที่ T2 และคุณลักษณะ StateProvinceName อยู่ในช่วงค่า Threshold ที่ T3 ซึ่งจะทำการทดลองกำจัดคุณลักษณะและนำคุณลักษณะที่เหลือไปสร้างโมเดล ซึ่งได้ผลลัพธ์ประสิทธิภาพการจำแนกข้อมูล ดังตารางที่ 4.2 ผลการทดลองที่ 2 การเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกการตัดสินใจซื้อจักรยานของลูกค้า

ตารางที่ 4.2 การเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกการตัดสินใจซื้อจักรยานของลูกค้า

การคัดเลือกคุณลักษณะ	โมเดล	ค่าความแม่นยำคลาส 1	ค่าความแม่นยำคลาส 0	ค่าความระลึกคลาส 1	ค่าความระลึกคลาส 0	ค่าประสิทธิภาพถ่วงดุล
13 คุณลักษณะ	DecisionTree	57.56	79.52	58.94	78.58	68.65
	Naïve Bayes	50.76	88.66	69.57	77.70	71.67
	K-NN	54.16	76.95	54.55	76.67	65.59
	Boosting (XGBoost)	**57.74**	**89.36**	**73.55**	**80.55**	**75.27**
T1	DecisionTree	56.31	79.74	58.68	78.14	68.22
	Naïve Bayes	50.76	**88.71**	69.66	77.91	71.70
	K-NN	53.98	76.73	54.23	76.55	65.37
	Boosting (XGBoost)	**57.65**	88.39	**71.71**	**80.34**	**74.49**
T2	DecisionTree	56.04	80.89	59.96	78.27	68.79
	Naïve Bayes	51.03	**88.98**	70.28	78.06	72.03
	K-NN	52.91	75.86	52.82	75.93	64.38
	Boosting (XGBoost)	**57.03**	88.16	**71.09**	**80.07**	**74.06**
T3	DecisionTree	**57.92**	80.29	60.02	78.89	69.28
	Naïve Bayes	51.66	88.84	70.28	78.25	72.20
	K-NN	51.39	77.05	53.35	75.63	64.35
	Boosting (XGBoost)	55.68	**88.98**	**72.07**	**79.72**	**74.07**

จากตารางที่ 4.2 การเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกการตัดสินใจซื้อจักรยานของลูกค้า พบว่า ด้านประสิทธิภาพความแม่นยำ โดยมีค่าความแม่นยำของคลาส 0 สูงที่สุด คือ เทคนิควิธี Boosting (XGboost) โดยใช้ทั้ง 13 คุณลักษณะ อยู่ที่ร้อยละ 89.36 โดยที่ค่าความแม่นยำของคลาส 1 สูงที่สุด คือ เทคนิควิธีต้นไม้ตัดสินใจร่วมกับการคัดเลือกคุณลักษณะด้วย T3 มีค่าความแม่นยำร้อยละ 57.92 ด้านค่าความระลึก โดยมีค่าความระลึกของคลาส 0 สูงที่สุด คือ เทคนิควิธี Boosting (XGboost) โดยใช้ทั้ง 13 คุณลักษณะ อยู่ที่ร้อยละ 80.55 และค่าความระลึกของคลาส 1 อยู่ที่ร้อยละ 73.55 ในเทคนิควิธีเดียวกัน ด้านประสิทธิภาพถ่วงดุล คือ เทคนิควิธี Boosting (XGboost) โดยใช้ทั้ง 13 คุณลักษณะ 75.27

ข้อเสนอแนะจากงานวิจัย

งานวิจัยนี้เป็นงานวิจัยที่ได้ทำการทดลองจำแนกข้อมูลการตัดสินใจซื้อจักรยานโดยใช้ชุดข้อมูลจากต่างประเทศ ซึ่งในการตัดสินใจเลือกซื้อจักรยานในประเทศไทยอาจต้องวิเคราะห์ปัจจัยแวดล้อมด้านอื่น ๆ ประกอบการพิจารณาการตัดสินใจซื้อ และงานวิจัยนี้เป็นการทดลองกับข้อมูลที่มีความไม่สมดุล ดังนั้นในการทดลองครั้งถัดไปคณะผู้วิจัยได้วางแผนปรับปรุงรูปแบบการสุ่มข้อมูลของแต่ละคลาสในมีจำนวนเท่ากันและทำการวัดประสิทธิภาพของแบบจำลอง รวมถึงการปรับแต่งพารามิเตอร์ของแบบจำลองเพื่อเพิ่มประสิทธิภาพของการทำนายผลลัพธ์

กิตติกรรมประกาศ (Acknowledgment)

คณะผู้วิจัยขอขอบพระคุณ อาจารย์มงคล แสนสุข และผู้ที่มีส่วนร่วมในการให้ความรู้และให้แนวทางในการดำเนินรายงานการวิจัยไม่ว่าจะมากหรือน้อยทุกท่าน ขอขอบพระคุณสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏสกลนคร ที่ให้ทุนสนับสนุนการวิจัยในครั้งนี้ให้ลุล่วงไปด้วยดี และขอขอบพระคุณสาขาวิชาคณิตศาสตร์และสถิติ มหาวิทยาลัยราชภัฏสกลนครที่ให้ความอนุเคราะห์สถานที่และอุปกรณ์ที่จำเป็นต่อการทำวิจัยในครั้งนี้ด้วย หากมีข้อแนะนำหรือข้อผิดพลาดประการใด คณะผู้วิจัยจึงขออภัยมา ณ ที่นี้ด้วย

เอกสารอ้างอิง

[1] ศูนย์ข้อมูลธุรกิจไทยในไต้หวัน. จักรยานและเมืองสีเขียว : อุตสาหกรรมจักรยานโตขึ้นอย่างต่อเนื่อง ท่ามกลางสถานการณ์การแพร่ระบาดของโรคโควิด-19. 2564. เข้าถึงวันที่ 14 พฤศจิกายน 2565. จาก https://thaibiztaiwan.thaiembassy.org/th/content/จักรยานและเมืองสีเขียว-อุตสาหกรรมจักรยานโตขึ้นอย่า?cate=5d846bc315e39c10e0002853

[2] กฤษณ์ พิโรจน์. Market Share คืออะไร? ส่วนแบ่งทางการตลาดสำคัญอย่างไร. 2564. เข้าถึงวันที่ 7 ตุลาคม 2565. จาก https://greedisgoods.com/market-share-คือ/

[3] Prada S. J, Bike Buyer Prediction. International Journal of Computer Engineering and Technology (IJCET). No. 11, pp. 45 – 51, 2020

[4] MR. Kabir FB. Ashraf and R. Ajwad, Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data. 22nd International Conference on Computer and Information Technology (ICCIT 2019). pp. 1 – 6, 2019.

[5] ธีรยุทธ คูณสุข และจารี ทองคำ. กระบวนการคัดเลือกคุณลักษณะสำหรับเพิ่มประสิทธิภาพในการจำแนกความคิดเห็นของลูกค้า. วารสารมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ฉบับวิทยาศาสตร์และเทคโนโลยี, ปีที่ 13 ฉบับที่ 1. 129 - 143, 2563.

[6] ปริยานุช ประเสริฐสิริกุล ศิริสรรพ เหล่าหะเกียรติ เรืองศักดิ์ ตระกูลพุทธิรักษ์ และศศิวิมล สุขพัฒน์. การเปรียบเทียบประสิทธิภาพแบบจำลองการทำนายผลการเรียนของนิสิตที่ใช้งานระบบการจัดการเรียนรู้ออนไลน์ด้วยการเรียนรู้ของเครื่อง. วารสารวิทยาศาสตร์และเทคโนโลยี, ปีที่ 6 ฉบับที่ 1. 79 - 99, 2565.

[7] ปาริชาต ปาวิชัย. ย้อนรอย "จักรยาน" ของเห่อใหม่สมัย ร.5 ราคาคันละร้อย!. 2561. เข้าถึงวันที่ 28 ตุลาคม 2565. จาก https://www.silpa-mag.com/history/article23994

[8] ณัฐวัฒน์ พงษ์จิตร. มาทำความรู้จัก Machine Learning เบื้องต้น. 2561. เข้าถึงวันที่ 16 ตุลาคม 2565. จาก https://medium.com/@natthawatphongchit/machine-learningbasics-2b38 700cb10b

[9] ณัฐดนัย วังประธรรม. DATA MINING. 2562. เข้าถึงวันที่ 29 กรกฎาคม 2565. จาก https://nutdnuy.medium.com/data-mining-8bb5398d885c

[10] สุรพงศ์ เอื้อวัฒนามงคล. การทำเหมืองข้อมูล. สถาบันบัณฑิตพัฒนบริหารศาสตร์. 2562

[11] อนุพงศ์ สุขประเสริฐ. Data Pre-Preprocession. 2562. เข้าถึงวันที่ 26 กรกฎาคม 2565. จาก https://slideplayer.in.th/slide/15935722/

[12] สมจิตต์ รัตนอุดมโชค ดวงฤดี อิ่มบุญสุ และสุวิวรรธน์ วิชกูล. การวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษา ด้วยเทคนิคการทำเหมืองข้อมูล. วารสารวิชาการหลักสูตรและการสอน มหาวิทยาลัยราชภัฏสกลนคร, ปีที่ 13 ฉบับที่ 36. 79 - 90, 2564.

[13] 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.

[14] โกเมศ อัมพวัน. การจำแนกประเภทและการทำนายข้อมูล. 2559. เข้าถึงวันที่ 27 กรกฎาคม 2565. จาก https://staff.informatics.buu.ac.th/ komate/886464/tion.pdf

[15] อริสรา โชติวิจิตร. Botnoi Classroom มาทำให้ผลลัพธ์ดีขึ้นด้วย Ensemble Method กันเถอะ!. 2562. เข้าถึงวันที่ 23 สิงหาคม 2565. จาก https://medium.com/botnoiclassroom/bot noi-classroom-มาทำให้ผลลัพธ์ดีขึ้นด้วย-ensemble-method-กันเถอะb5ac9acfa8d3

[16] บัญชา ปะสีละเตสัง. สร้างการเรียนรู้สำหรับ AI ด้วย Python Machine Learning. ซีเอ็ดยูเคชั่น. 2564

[17] เอกสิทธิ์ พัชรวงศ์ศักดา. การวิเคราะห์ข้อมูล ด้วยเทคนิคดาต้าไมน์นิงเบื้องต้น. เอเชียดิจิตอลการพิมพ์. 2557

[18] Rahul. Bike Buying Prediction For Adventure Works Cycles. 2562. เข้าถึงวันที่ 24 พฤษภาคม 2566. จาก https://www.kaggle.com/datasets/rahulsah06/bike-buying-prediction-for-adventure-works-cycles

SCIENCE AND TECHNOLOGY
UTTARADIT RAJABHAT UNIVERSITY