บทความวิจัย (Research Article)

การประยุกต์ใช้ความเปลี่ยนแปลงของข้อมูลเพื่อสุ่มลดข้อมูลสำหรับการแก้ปัญหาการจัดประเภทข้อมูลที่ไม่สมดุล

A Data Change Based Under-Sampling Approach for Solving Imbalanced Data Classification

วราพรรณ ธิยานันท์ และ กฤษณะ ไวยมัย*

ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเกษตรศาสตร์
* ผู้รับผิดชอบบทความ
fengknw@ku.ac.th

Received: 3 Aug 2023
Revised: 22 Sep 2023
Accepted: 22 Sep 2023

บทคัดย่อ

ปัจจัยที่สำคัญที่สุดอย่างหนึ่งในการพัฒนาความแม่นยำในการจำแนกประเภทโดยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) คือคุณภาพของข้อมูลที่ใช้ในการเรียนรู้ อย่างไรก็ตามข้อมูลในโลกของความเป็นจริงโดยมากนั้นไม่สมดุล กล่าวคือข้อมูลส่วนใหญ่จัดอยู่ในกลุ่มข้อมูลหลัก (Majority Class) และส่วนน้อยจัดอยู่ในกลุ่มข้อมูลย่อย (Minority Class) บทความนี้นำเสนอแนวทางสำหรับการสุ่มลดข้อมูลของกลุ่มข้อมูลตัวอย่างโดยการคงไว้เฉพาะตัวแทนของกลุ่มข้อมูลนั้น ความเปลี่ยนแปลงของข้อมูลมาจากการใช้เทคนิคในการเลือกข้อมูลโดยมีเป้าหมายเพื่อลดกลุ่มข้อมูลหลักให้มีขนาดเล็กลง ผลการศึกษาแสดงให้เห็นว่ากลไกการเลือกข้อมูลจากความเปลี่ยนแปลงของข้อมูลสามารถเพิ่มความแม่นยำของข้อมูลกลุ่มย่อยได้ทั้งเทคนิคการสุ่มลดข้อมูล (Under-Sampling) และเทคนิคการสุ่มแบบผสม (Mixed Sampling)

คำสำคัญ: ความเปลี่ยนแปลงข้อมูล ปัญหาข้อมูลไม่สมดุล ต้นไม้ตัดสินใจ การเรียนรู้ของเครื่อง

Abstract

One of the most important factors for improving the accuracy of machine learning classification techniques is the quality of the training data. However, real-world data are mostly imbalanced, that is, most of the data are in majority class and little data are in minority class. This paper introduces an approach for under-sampling samples of the majority class by keeping only its representative data. A data change based selection technique is proposed to reduce the majority class data. The experimental results show that our data change based selection mechanism is able to improve the accuracy of the minority class for both under sampling and mixed sampling techniques.

Keywords: Data Change, Imbalanced Data, Decision Tree, Machine Learning

1. บทนำ

ในปัจจุบันมีหลายองค์กรนำเทคโนโลยีมาปรับใช้ในการเก็บข้อมูล ซึ่งจะนำข้อมูลที่ได้รับมาวิเคราะห์ เพื่อเพิ่มโอกาสทางธุรกิจ รวมถึงนำข้อมูลที่ได้มาปรับใช้ประกอบการตัดสินใจไปจนถึงการวางกลยุทธ์ทางธุรกิจขององค์กรซึ่งเป็นปัจจัยสำคัญที่ต้องพิจารณาเรื่องการเพิ่มประสิทธิภาพการดำเนินงาน และเพิ่มประสิทธิภาพในการทำการตลาดที่ขึ้นอยู่กับพฤติกรรมของลูกค้า การนำการเรียนรู้ของเครื่อง (Machine Learning) เข้ามาประยุกต์ใช้ในการวิเคราะห์ข้อมูล เพื่อให้สามารถคาดการณ์คุณลักษณะของลูกค้าที่มีผลต่อการเลือกผลิตภัณฑ์หรืองานบริการที่ลูกค้าพึงพอใจจากชุดข้อมูลที่มีอยู่นั้น คือการใช้เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) เพื่อนำไปทำโมเดลรวมทั้งตรวจสอบประสิทธิภาพของโมเดล โดยใช้ชุดข้อมูลจริงที่มีอยู่ [1]

การทำโมเดลให้ได้ประสิทธิภาพที่ดีนั้น นอกจากการเลือกเทคนิคที่เหมาะสมในการทำโมเดลแล้ว สิ่งสำคัญคือคุณภาพของข้อมูลที่นำมาใช้ทำโมเดลก็เป็นส่วนสำคัญที่จะเพิ่มประสิทธิภาพของโมเดล หากข้อมูลนำมาสร้างโมเดลเป็นข้อมูลที่มีคุณภาพดีจะส่งผลต่อโมเดลโดยตรงและทำให้ประสิทธิภาพของโมเดลสูงเช่นกัน ดังนั้นความท้าทายของการจัดการข้อมูลจึงเป็นสิ่งน่าสนใจที่จะต้องแก้ไขก่อนการทำโมเดล อุปสรรคสำคัญที่มักพบคือความไม่สมดุลของข้อมูล (Imbalanced Data) [2] ซึ่งมีลักษณะของข้อมูลที่มีจำนวนข้อมูลในคลาสที่แตกต่างกันมากโดยแบ่งเป็น 2 คลาส คือ คลาสข้อมูลใหญ่ (Majority Class) และคลาสข้อมูลเล็ก (Minority Class) ซึ่งการทำนายในคลาสข้อมูลเล็กมีผลทำให้ประสิทธิภาพต่ำและไม่ดีเท่าที่ควรส่งผลให้ข้อมูลที่ต้องการทำนายเพื่อนำไปใช้ในการตัดสินใจมีความผิดพลาด และโดยส่วนมากในธุรกิจคลาสข้อมูลเล็กมักจะเป็นคลาสที่ทางกลุ่มธุรกิจให้ความสนใจ ดังนั้นปัญหาความไม่สมดุลของข้อมูลจึงเป็นสิ่งที่ท้าทายในการจัดการ

ข้อมูลจริงที่ใช้ในทางธุรกิจมักจะพบปัญหาเรื่องคุณภาพของข้อมูล หนึ่งปัญหาที่พบมากที่สุดคือปัญหาความไม่สมดุลของข้อมูล มีแนวทางการแก้ไขที่หลากหลาย หนึ่งในวิธีการนั้นคือการจัดการความไม่สมดุลในระดับข้อมูล (Data Approach) โดยการเพิ่มหรือลดจำนวนข้อมูล วิธีการแรกคือ การเพิ่มจำนวนตัวอย่างข้อมูลในกลุ่มข้อมูลคลาสเล็ก เรียกว่า “Over-sampling” วิธีการที่สองคือ การลดจำนวนข้อมูลลงในกลุ่มข้อมูลคลาสใหญ่เรียกว่า “Under-sampling” และวิธีการสุดท้ายคือการผสมทั้ง 2 วิธีเข้าด้วยกันเรียกว่า “Mixed Sampling” [3] วิธีการนี้จะทำทั้งการสุ่มเพิ่มและสุ่มลด ซึ่งแต่ละวิธีมีข้อดีและข้อเสียที่แตกต่างกันออกไปตามวิธีการจัดการ

วิธีที่นิยมที่สุดคือ การสุ่มเพิ่มตัวอย่างข้อมูล ซึ่งต้องใช้ระยะเวลาในการฝึก และข้อเสียที่สำคัญคือ ทำให้ข้อมูลมีความเอนเอียง (Biased) และพบเจอเหตุการณ์ Over-fitting [4-5] จนกระทั่งได้มีการพัฒนาวิธีการสุ่มเพิ่มเพื่อการสร้างตัวอย่างสังเคราะห์ (Synthetic Minority Over-sampling Technique: SMOTE) [6-7] โดยใช้อัลกอริทึมเข้ามาช่วยสร้างตัวอย่างข้อมูลใหม่โดยใช้ข้อมูลใกล้เคียง (Nearest Neighbours) [8] เพื่อลดโอกาสเกิด Over-fitting ในทางกลับกันนั้นการสุ่มลดก็ถูกนำไปใช้เช่นกันแต่ไม่เป็นที่นิยมมากนัก เนื่องจากการลดจำนวนข้อมูลนั้นจะส่งผลให้สูญเสียจำนวนข้อมูลทำให้เราเสียพฤติกรรม (Records หรือ Transactions) ในการเรียนรู้ให้กับโมเดล ซึ่งทำให้มีการพัฒนาเทคนิคต่างๆ มาช่วยแก้ปัญหานี้ช่วยลดโอกาสที่ข้อมูลจะสูญหายได้และแก้ปัญหาข้อมูลที่ไม่มีประสิทธิภาพเข้ามารบกวนโมเดล แต่อย่างไรก็ตามปัญหาการกระจายคลาสที่ไม่สมดุลยังคงมีอยู่ [9]

จุดมุ่งหมายในงานวิจัยนี้คือการนำเสนอแนวทางการปรับปรุงเทคนิคการแก้ปัญหาข้อมูลไม่สมดุลโดยการประยุกต์ใช้ความคิดเห็นทางธุรกิจ โดยการนำความคิดเห็นทางธุรกิจมาเป็นส่วนช่วยในการตัดสินใจลดจำนวนข้อมูลในคลาสใหญ่โดยใช้พื้นฐานความรู้และประสบการณ์ของธุรกิจนั้น ๆ เป็นส่วนช่วยให้สามารถตัดข้อมูลที่ไม่จำเป็นออกจากชุดข้อมูลได้อย่างมีประสิทธิภาพ เพื่อเพิ่มประสิทธิภาพการทำนาย ทำให้ผู้วิจัยสนใจในวิธีการเพิ่มจำนวนข้อมูล และการลดจำนวนข้อมูลซึ่งจะนำเสนอการลดจำนวนข้อมูลในคลาสใหญ่ด้วย 2 วิธีคือ 1) การลดจำนวนข้อมูลโดยความเปลี่ยนแปลงของข้อมูล และ 2) การลดจำนวนข้อมูลโดยความเปลี่ยนแปลงของข้อมูลตามค่าถ่วงน้ำหนักที่มาจากความคิดเห็นทางธุรกิจ และเปรียบเทียบประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ(Precision), ค่าความครบถ้วน (Recall) และค่าประสิทธิภาพโดยรวม (F-Measure)

2. การสุ่มสร้างข้อมูลและสุ่มลดข้อมูลเพื่อแก้ปัญหาข้อมูลไม่สมดุล

การสุ่มสร้างข้อมูลสามารถแบ่งออกเป็น 3 ประเภท คือการสุ่มเพิ่มข้อมูล (Over-sampling) การสุ่มลดข้อมูล (Under-sampling) และการผสม (Mixed sampling) ซึ่ง 3 วิธีนี้ มีข้อดีและข้อเสียที่แตกต่างกัน

2.1 การสุ่มเพิ่มข้อมูล (Over-sampling)

เป็นวิธีการเพิ่มจำนวนข้อมูลในคลาสเล็กและเป็นวิธีที่นำมาใช้กันอย่างแพร่หลายเนื่องจากไม่ส่งผลให้สูญเสียจำนวนข้อมูลและพฤติกรรม ซึ่งวิธีการนี้เป็นการสุ่มจากข้อมูลเดิม (Random Oversampling: ROS) แต่อาจพบปัญหา Over-fitting ซึ่งปัญหานี้ได้มีการเสนอวิธีการเพิ่มจำนวนข้อมูลขึ้นมา เพื่อเอาชนะข้อบกพร่องนี้ จึงได้มีการพัฒนาเทคนิค SMOTE (Synthetic Minority Over-sampling Technique) เป็นวิธีการสร้างข้อมูลตัวใหม่จากข้อมูลเดิมด้วยการใช้หลักการของข้อมูลเพื่อนบ้านใกล้เคียง (Nearest Neighbors) โดยใช้การคำนวณระยะทางเพื่อเลือกค่าที่ใกล้เคียงที่สุดมาสร้างข้อมูลใหม่ ซึ่งสามารถให้ค่าความถูกต้องได้ดี [10]

2.2 การสุ่มลดข้อมูล (Under-sampling)

เป็นการลดจำนวนข้อมูลในคลาสใหญ่ให้ใกล้เคียงกับจำนวนข้อมูลในคลาสเล็กส่งผลให้สามารถลดเวลาการรันโมเดลและลดปัญหาขนาดข้อมูลทำให้มีการจัดเก็บข้อมูลที่ดีขึ้น อย่างไรก็ตามวิธีนี้ ยังคงพบปัญหาการกระจายคลาสที่ไม่สมดุล อาจทำให้มีการลบข้อมูลที่สำคัญออกไป ซึ่งส่งผลให้ข้อมูลที่เหลืออยู่เป็นตัวอย่างข้อมูลที่มีอคติและไม่สามารถให้ความถูกต้องได้ นั้นอาจทำให้สูญเสียข้อมูลสำคัญบางอย่างที่จะนำไปสู่การเรียนรู้ในโมเดล [9,11] โดยการสุ่มลดจำนวนข้อมูลจะถูกนำมาใช้ในการเตรียมข้อมูลและการทำความสะอาดข้อมูลก่อนการทำโมเดลซึ่งมีหลายงานวิจัยที่ใช้วิธีการนี้ด้วย

2.3 การผสม (Mixed sampling)

การนำวิธีการลดและเพิ่มจำนวนข้อมูลมาใช้ร่วมกันเพื่อทำให้จำนวนข้อมูลของคลาสเล็กและคลาสใหญ่ใกล้เคียงกัน ซึ่งบางงานวิจัยใช้วิธีนี้ในการรับมือกับความไม่สมดุลของข้อมูลซึ่งทำให้ประสิทธิภาพของโมเดลดีขึ้น [3,12]

3. แนวคิดการประยุกต์ใช้ความเปลี่ยนแปลงข้อมูลในการสุ่มลดข้อมูลเพื่อแก้ไขปัญหาความไม่สมดุล

ในวิจัยนี้มีผู้ทำวิจัยมีจุดมุ่งหมายที่จะแสดงแนวทางการสุ่มลดจำนวนข้อมูลในคลาสใหญ่ (Under-sampling) เพื่อใช้สำหรับการแก้ปัญหาความไม่สมดุลของข้อมูล โดยการใช้วิธีการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลเพื่อทำการลบข้อมูลที่ไม่ค่อยมีความเปลี่ยนแปลงออกไปจากชุดข้อมูล โดยพิจารณาจากคุณลักษณะของลูกค้าที่ต้องการซื้อรถยนต์ และงานวิจัยนี้เสนอสองแนวคิดดังต่อไปนี้

ตารางที่ 1 แนวคิดการสุ่มลดข้อมูลในคลาสใหญ่โดยการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

Customer ID	Transaction Date	1st Car's Attributes (รถคนท 1)				Target	Transaction Date	2nd Car's Attributes (รถคนท 2)				Target
		Color	Ton	Usage	Occupation			Color	Ton	Usage	Occupation	
คนท 1	2020 Jan	Pink	3-6	Commercial	Police	Truck	2021 Apr	Pink	3-6	Private	Police	Truck
คนท 2	2021 Mar	Red	1-3	Commercial	Salesman	Truck	2022 Oct	Red	1-3	Commercial	Salesman	Truck
คนท 3	2021 Jan	Red	6-9	Commercial	Salesman	Truck	2021 Aug	Blue	6-9	Private	Trainer	Sedan
คนท 4	2020 Jul	Green	1-3	Private	Police	Sedan	2022 Feb	Green	3-6	Private	Salesman	Truck

ตารางที่ 2 แนวคิดการสุ่มลดข้อมูลในคลาสใหญ่โดยการถ่วงน้ำหนักข้อมูลก่อนการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

Customer ID	Transaction Date	1st Car's Attributes (รถคนท 1)				Target	Transaction Date	2nd Car's Attributes (รถคนท 2)				Target
		Color	Ton	Usage	Occupation			Color	Ton	Usage	Occupation	
คนท 1	2020 Jan	Pink	3-6	Commercial	Police	Truck	2021 Apr	Pink	3-6	Private	Police	Truck
คนท 2	2021 Mar	Red	1-3	Commercial	Salesman	Truck	2022 Oct	Red	1-3	Commercial	Salesman	Truck
คนท 3	2021 Jan	Red	6-9	Commercial	Salesman	Truck	2021 Aug	Blue	6-9	Private	Trainer	Sedan
คนท 4	2020 Jul	Green	1-3	Private	Police	Sedan	2022 Feb	Green	3-6	Private	Salesman	Truck
Weight		40%	5%	45%	5%	5%		40%	5%	45%	5%	5%
		100%						100%				

3.1 การสุ่มลดจำนวนข้อมูลในคลาสใหญ่โดยการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

วิธีการนี้ คือ การนำข้อมูลของลูกค้าเปรียบเทียบความเปลี่ยนของข้อมูลในแต่ละคุณลักษณะของลูกค้าที่ต้องการซื้อรถยนต์ (Attributes) เพื่อลบรายการข้อมูล (Transactions) ออกจากชุดข้อมูล ซึ่งเป็นรายการข้อมูลที่ไม่มีการเปลี่ยนแปลงมากนัก ในการวิจัยนี้ทางผู้วิจัยได้นำข้อมูลก่อนการตัดสินใจซื้อรถยนต์คันที่ 1 และ คันที่ 2 มาทำการทดลองโดยนำคุณลักษณะของลูกค้าที่ต้องการซื้อรถยนต์และเป้าหมายประเภทรถยนต์ที่ลูกค้าต้องการของรถคันที่ 1 และ คันที่ 2 มาเปรียบเทียบกันดังตารางที่ 1 จะเห็นว่าเมื่อทำการเปรียบเทียบแล้วจะมีข้อมูลที่เหมือนและ

แตกต่างกัน จากนั้นทำการเลือกลบรายการข้อมูลมีการความเปลี่ยนแปลงน้อยออกจากชุดข้อมูลเพื่อเป็นการลดข้อมูลที่ซ้ำซ้อนออก หลักการนี้ทำให้ข้อมูลมีการกระจายตัวดีดังเดิม และไม่ได้เสียความสำคัญของข้อมูลไป เนื่องจากข้อมูลที่ลบไปนั้นเป็นข้อมูลที่ไม่ค่อยมีความเปลี่ยนแปลง

3.2 การสุ่มลดจำนวนข้อมูลในคลาสใหญ่โดยการถ่วงน้ำหนักข้อมูลและการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

วิธีการนี้คือ การถ่วงน้ำหนักข้อมูลก่อนนำไปเปรียบเทียบความเปลี่ยนของข้อมูลในแต่ละคุณลักษณะของลูกค้าที่ต้องการซื้อรถยนต์เพื่อลบรายการข้อมูลออกจากชุดข้อมูล ซึ่งเป็นรายการข้อมูลที่ไม่ค่อยมีการเปลี่ยนแปลง โดยเริ่มจากการถ่วงน้ำหนักข้อมูลซึ่งผู้วิจัยได้ให้เจ้าของข้อมูลหรือเจ้าของผลิตภัณฑ์เลือกคุณลักษณะของลูกค้าที่ต้องการซื้อรถยนต์ ที่ลูกค้ามีแนวโน้มให้ความสนใจก่อนการตัดสินใจเลือกซื้อรถยนต์เนื่องจากเป็นสิ่งที่ลูกค้าให้ความสำคัญมากสุด จากนั้นเจ้าของข้อมูลจะเป็นผู้ให้ค่าถ่วงน้ำหนักข้อมูลในแต่ละ คุณลักษณะเป็นเปอร์เซ็นต์ที่แตกต่างกันออกไป โดยให้เป็นเปอร์เซ็นต์ในแต่ละคุณลักษณะและต้องสามารถรวมกันได้ 100 เปอร์เซ็นต์ตามตัวอย่างในตารางที่ 2 เห็นได้ว่าทุกคุณลักษณะของลูกค้าที่ต้องการซื้อรถยนต์มีการจัดสรรเปอร์เซ็นต์เพื่อการถ่วงน้ำหนักไว้สำหรับเตรียมทำการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลในขั้นตอนถัดไป

เมื่อได้ค่าถ่วงน้ำหนักข้อมูลแล้ว ผู้วิจัยจะทำการเปรียบเทียบข้อมูลโดยนำค่าถ่วงน้ำหนักข้อมูลเข้าไปคูณ จากนั้นจะทำการเลือกลบรายการข้อมูลที่มีการความเปลี่ยนแปลงน้อยออกจากชุดข้อมูลเพื่อเป็นการลดข้อมูลที่ซ้ำซ้อนออก หลักการนี้ทำให้ข้อมูลไม่ได้เสียความสำคัญของข้อมูลไปและที่สำคัญสามารถเก็บข้อมูลที่มีนัยยะสำคัญต่อการตัดสินใจเลือกซื้อรถยนต์ของลูกค้า เนื่องจากข้อมูลที่ลบไปนั้นเป็นข้อมูลที่ไม่มีความเปลี่ยนแปลงมากนักและไม่ได้เป็นปัจจัยสำคัญต่อการเลือกซื้อรถยนต์

ดังนั้นในงานวิจัยนี้จะเสนอการลดจำนวนข้อมูลในคลาสใหญ่ (Under-Sampling) แบ่งออกเป็น 2 วิธีคือ 1) การสุ่มลดจำนวนข้อมูลในคลาสใหญ่โดยการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล และ 2) การสุ่มลดจำนวนข้อมูลในคลาสใหญ่โดยเทคนิคการถ่วงน้ำหนักข้อมูลก่อนการนำไปเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

ตารางที่ 3 การแปลงข้อมูลจากข้อมูลระดับรายการข้อมูลเป็นการแสดงตามระดับรหัสลูกค้า

Customer ID	Transaction Date	Attributes				Target
		Color	Ton	Usage	Occupation	
คนท 1	2020 Jan	Pink	3-6	Commercial	Police	Truck
คนท 1	2021 Apr	Pink	3-6	Private	Police	Truck
คนท 2	2021 Mar	Red	1-3	Commercial	Salesman	Truck
คนท 2	2022 Oct	Red	1-3	Commercial	Salesman	Truck
คนท 3	2021 Jan	Red	6-9	Commercial	Salesman	Truck
คนท 3	2021 Aug	Blue	6-9	Private	Trainer	Sedan
คนท 4	2020 Jul	Green	1-3	Private	Police	Sedan
คนท 4	2022 Feb	Green	3-6	Private	Salesman	Truck

Previous Latest

Customer ID	Transaction Date	1st Car's Attributes (รถคนท 1)				Target	Transaction Date	2nd Car's Attributes (รถคนท 2)				Target
		Color	Ton	Usage	Occupation			Color	Ton	Usage	Occupation	
คนท 1	2020 Jan	Pink	3-6	Commercial	Police	Truck	2021 Apr	Pink	3-6	Private	Police	Truck
คนท 2	2021 Mar	Red	1-3	Commercial	Salesman	Truck	2022 Oct	Red	1-3	Commercial	Salesman	Truck
คนท 3	2021 Jan	Red	6-9	Commercial	Salesman	Truck	2021 Aug	Blue	6-9	Private	Trainer	Sedan
คนท 4	2020 Jul	Green	1-3	Private	Police	Sedan	2022 Feb	Green	3-6	Private	Salesman	Truck

4. ขั้นตอนการดำเนินงานวิจัย

4.1 การเตรียมข้อมูล

ข้อมูลที่ใช้ทางผู้วิจัยได้ทำการขออนุญาตเจ้าของข้อมูลเพื่อนำมาทำการวิจัย ซึ่งข้อมูลได้ถูกอำพรางรายละเอียดที่จะสามารถระบุตัวตน และระบุรายละเอียดของเจ้าของข้อมูลได้ โดยข้อมูลที่ถูกอำพรางนั้นจะเป็นข้อมูลส่วนบุคคลของลูกค้า และเพื่อให้สามารถทำโมเดลได้อย่างถูกต้อง ทางผู้วิจัยต้องทำการนำข้อมูลมาเตรียมให้พร้อมในการทำโมเดล โดยมีการแบ่งขั้นตอนการเตรียมความพร้อมข้อมูลออกเป็น 5 ขั้นตอนดังนี้

4.1.1 การทำความสะอาดข้อมูล (Data Cleansing)

การจัดการจัดรูปแบบข้อมูลให้ถูกต้อง การจัดการค่าว่าง การจัดช่วงข้อมูล และการลบข้อมูลที่ไม่เกี่ยวข้องกับงานวิจัย

4.1.2 การแปลงข้อมูล (Data Transformation)

การเปลี่ยนแปลงระดับข้อมูลจากระดับรายการข้อมูล เป็นการแสดงตามระดับรหัสลูกค้า (Customer ID) ดังตารางที่ 3

4.1.3 การเลือกข้อมูล (Data Selection)

การเลือกข้อมูลต้องมีการพูดคุยกับเจ้าของข้อมูลในการเลือกคุณลักษณะของลูกค้าที่ทางเจ้าของข้อมูลให้ความสนใจ

4.1.4 การสร้าง Attribute ใหม่ (Attribute Creation)

การสร้างข้อมูลใหม่จากข้อมูลประเภทตัวเลข โดยการคำนวณตัวเลขขึ้นมาใหม่

4.1.5 การรวบรวมข้อมูล (Data Integration)

การนำคุณลักษณะของลูกค้าที่ต้องการซื้อรถยนต์ที่มีการเลือกจากเจ้าของข้อมูลและสร้างใหม่มารวมกัน

4.2 การประยุกต์ใช้การเปลี่ยนแปลงข้อมูลในการสุ่มลดข้อมูล

การประยุกต์ใช้การเปรียบเทียบความเปลี่ยนแปลงข้อมูลในการสุ่มลดจำนวนข้อมูลเพื่อแก้ไขปัญหาความไม่สมดุล โดยเริ่มจากนำแนวคิดในหัวข้อที่ 3 มาทดลองใช้กับโมเดล โดยจะอธิบายวิธีการประยุกต์แนวคิดในหัวข้อที่ 3.1 และ 3.2 ดังต่อไปนี้

วิธีการแรกคือ การสุ่มลดจำนวนข้อมูลในคลาสใหญ่โดยเปรียบเทียบความเปลี่ยนแปลงของข้อมูลเพียงอย่างเดียว ตามแนวคิดในหัวข้อที่ 3.1 คือการนำข้อมูลมาเปรียบเทียบกัน จากนั้นทำการแทนค่าความเปลี่ยนแปลงข้อมูล กรณีที่ข้อมูลเหมือนกันให้มีค่าเป็น 1 และ ข้อมูลต่างกันให้มีค่าเป็น 0 เมื่อได้ตัวเลขความเปลี่ยนแปลงข้อมูลระหว่างรถคันที่ 1 และคันที่ 2 แล้ว จากนั้นทำการหาเปอร์เซ็นต์ความเปลี่ยนแปลงของข้อมูล แล้วลบข้อมูลที่ไม่ค่อยมีความเปลี่ยนแปลงออกจากชุดข้อมูล ซึ่งในงานวิจัยนี้เราจะลบข้อมูลของผู้ซื้อรถยนต์ที่ไม่ค่อยมีความเปลี่ยนแปลงซึ่งมีค่าน้อยกว่าหรือเท่ากับ 80 เปอร์เซ็นต์ ดังตารางที่ 4 จะเห็นได้ว่าลูกค้าคนที่ 2 ข้อมูลไม่มีความเปลี่ยนแปลงเลย ดังนั้นเราจึงทำการลบข้อมูลของลูกค้าคนที่ 2 ออกจากชุดข้อมูล

ตารางที่ 4 การประยุกต์ใช้การเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

Customer ID	การเปลี่ยนแปลงระหว่างรถคันที่ 1 และ รถคันที่ 2				Target	ผลรวมความเปลี่ยนแปลง	%ของผลรวมความเปลี่ยนแปลง	การเก็บข้อมูลที่ไม่ค่อยมีความเปลี่ยนแปลงที่ ≤ 80%
	Color	Ton	Usage	Occupation				
คนที่ 1	1	1	0	1	1	4	4/5 = 80%	Keep
คนที่ 2	1	1	1	1	1	5	5/5 = 100% (ข้อมูลไม่มีการเปลี่ยนแปลง)	Deleted
คนที่ 3	0	1	0	0	0	1	1/5 = 20%	Keep
คนที่ 4	1	0	1	0	0	2	2/5 = 40%	Keep

Note 1 : ค่าของรถคันที่ 1 เหมือนกับ รถคันที่ 2
0 : ค่าของรถคันที่ 1 ไม่เหมือนกับ รถคันที่ 2

วิธีการที่สอง คือ การสุ่มลดจำนวนข้อมูลในคลาสใหญ่โดยการถ่วงน้ำหนักข้อมูลและการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล เราจะนำเทคนิคการถ่วงน้ำหนักข้อมูลเข้ามาประยุกต์ใช้เพิ่มเติมโดยใช้แนวคิดในหัวข้อที่ 3.2 คือ นำข้อมูลมาเปรียบเทียบกัน จากนั้นทำการแทนค่าความเปลี่ยนแปลงข้อมูล กรณีที่ข้อมูลเหมือนกันให้มีค่าเป็น 1 และ ข้อมูลต่างกันให้มีค่าเป็น 0 จากนั้นนำค่าถ่วงน้ำหนักที่ได้รับจากเจ้าของข้อมูลคูณด้วยค่าการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลนั้น ทำให้ได้ตัวเลขความเปลี่ยนแปลงข้อมูลระหว่างรถคันที่ 1 และคันที่ 2 จากนั้นทำการเปลี่ยนตัวเลขความเปลี่ยนแปลงของลูกค้าที่ได้เป็นเปอร์เซ็นต์ และลบข้อมูลของผู้ซื้อรถยนต์ที่ไม่ค่อยมีความเปลี่ยนแปลงซึ่งมีค่าน้อยกว่าหรือเท่ากับ 80 เปอร์เซ็นต์ในขั้นตอนสุดท้าย ตามตารางที่ 5 จะเห็นได้ว่าลูกค้าคนที่ 2 ข้อมูลไม่มีความเปลี่ยนแปลงเลย และลูกค้าที่ 4 มีความเปลี่ยนแปลงเพียง 15 เปอร์เซ็นต์ (ข้อมูลมีความเหมือนกัน 85 เปอร์เซ็นต์) ดังนั้นจึงทำการลบข้อมูลของลูกค้าคนที่ 2 และ 4 ออกจากชุดข้อมูล

ตารางที่ 5 การประยุกต์ใช้เทคนิคการถ่วงน้ำหนักข้อมูลก่อนการนำไปเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

Customer ID	การเปลี่ยนแปลงระหว่างรถคันที่ 1 และ รถคันที่ 2				Target	ผลรวมความเปลี่ยนแปลง	%ของผลรวมความเปลี่ยนแปลง	การเก็บข้อมูลที่ไม่ค่อยมีความเปลี่ยนแปลงที่ ≤ 80%
	Color	Ton	Usage	Occupation				
Weight	40%	5%	45%	5%	5%	100%		
คนที่ 1	1 x 40%	1 x 5%	0	1 x 5%	1 x 5%	0.55	55%	Keep
คนที่ 2	1 x 40%	1 x 5%	1 x 45%	1 x 5%	1 x 5%	1	100%	Deleted
คนที่ 3	0	1 x 5%	0	0	0	0.05	5%	Keep
คนที่ 4	1 x 40%	0	1 x 45%	0	0	0.85	85%	Deleted

Note 1 : ค่าของรถคันที่ 1 เหมือนกับ รถคันที่ 2
0 : ค่าของรถคันที่ 1 ไม่เหมือนกับ รถคันที่ 2

เพื่อเปรียบเทียบประสิทธิภาพได้อย่างชัดเจนมากขึ้น ในงานวิจัยจะแบ่งกลุ่มโมเดลต่างๆ เป็น 2 กลุ่มดังนี้

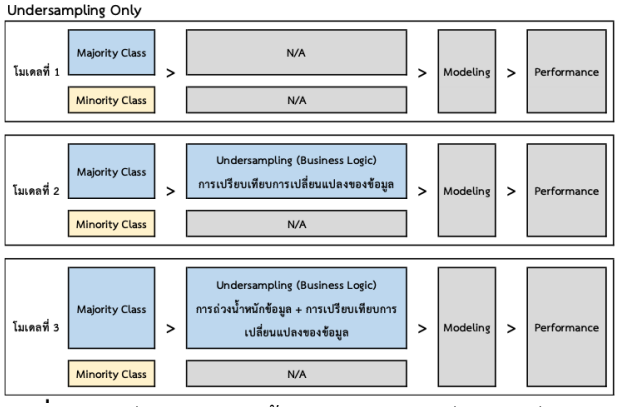
4.2.1 การสุ่มลดจำนวนข้อมูลในคลาสใหญ่เพียงอย่างเดียว (Under-Sampling)

กลุ่มนี้จะประยุกต์การสุ่มลดจำนวนข้อมูลในคลาสใหญ่ตามที่กล่าวข้างต้น เริ่มที่โมเดลที่ 1 ไม่มีการสุ่มเพิ่มและสุ่มลดข้อมูล โมเดลที่ 2 ทำการสุ่มลดข้อมูลที่คลาสใหญ่ด้วยการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล และโมเดลที่ 3 จะทำการสุ่มลดข้อมูลที่คลาสใหญ่ด้วยการถ่วงน้ำหนักข้อมูลก่อนการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลตามลำดับ ดังรูปที่ 1

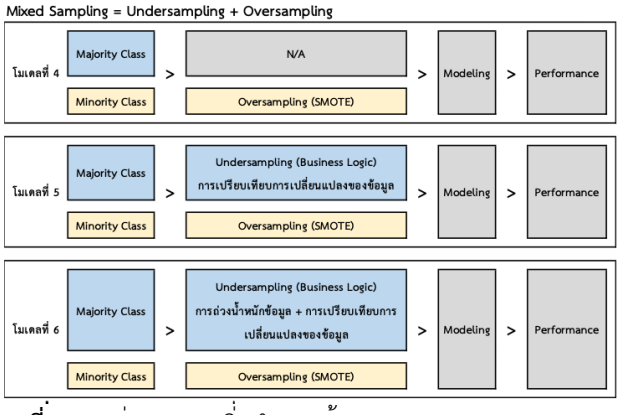
4.2.2 การผสม (Mixed Sampling)

กลุ่มนี้จะทำการสุ่มลดจำนวนข้อมูลในคลาสใหญ่ตามโมเดลในกลุ่มที่ 1 และผสมการสุ่มเพิ่มจำนวนข้อมูลในคลาสเล็กด้วยเทคนิค SMOTE ซึ่งโมเดลที่ 4 จะทำการสุ่มเพิ่มจำนวนข้อมูลที่คลาสเล็กด้วยเทคนิค SMOTE เพียงอย่างเดียว ส่วนโมเดลที่ 5

ทำการสุ่มลดจำนวนข้อมูลที่คลาสใหญ่ด้วยเทคนิคการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลพร้อมกับการสุ่มเพิ่มจำนวนข้อมูลที่คลาสเล็กด้วยเทคนิค SMOTE สำหรับโมเดลที่ 6 ทำการสุ่มลดจำนวนข้อมูลที่คลาสใหญ่ด้วยเทคนิคการถ่วงน้ำหนักข้อมูลก่อนการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลพร้อมกับทำการสุ่มเพิ่มจำนวนข้อมูลที่คลาสเล็กด้วยเทคนิค SMOTE ตามลำดับดังรูปที่ 2



รูปที่ 1 การสุ่มลดจำนวนข้อมูลในคลาสใหญ่เพียงอย่างเดียว (Under-Sampling)



รูปที่ 2 การสุ่มลดและเพิ่มจำนวนข้อมูล (Mixed Sampling)

5. ผลการวิจัย

5.1 การประเมินผล

การประเมินผลในการวิจัยนี้จะใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และการวัดประสิทธิภาพการทดลองของโมเดลด้วย Confusion Matrix หรือ ตารางการวัดความสามารถในการแก้ปัญหาวิธีการจำแนกประเภทข้อมูล (Classification) โดยมีค่าประสิทธิภาพโดยรวมจากค่าความครบถ้วนและค่าความแม่นยำ ดังรูปที่ 3 ซึ่งค่า TP, TN, FP และ FN มีความหมายดังนี้

- True Positive (TP) ทำนายว่า 1 มีผลลัพธ์ตามจริง 1
- True Negative (TN) ทำนายว่า 0 มีผลลัพธ์ตามจริง 0
- False Positive (FP) ทำนายว่า 1 มีผลลัพธ์ตามจริง 0
- False Negative (FN) ทำนายว่า 0 มีผลลัพธ์ตามจริง 1

สำหรับงานวิจัยนี้จะใช้ตัววัดประสิทธิภาพ 4 ตัวดังนี้

5.1.1 ค่าความถูกต้อง (Accuracy)

ค่าความถูกต้องของโมเดลโดยพิจารณารวมทุกคลาส

5.1.2 ค่าความครบถ้วน (Recall)

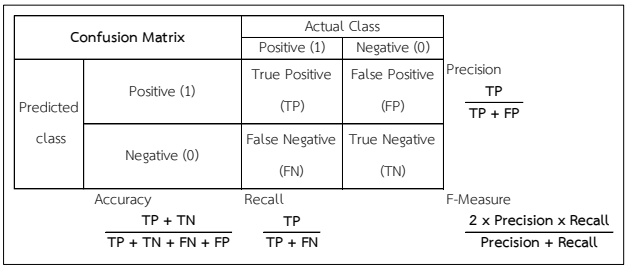
ค่าความครบถ้วนในโมเดลโดยพิจารณาแยกทีละคลาส

5.1.3 ค่าความแม่นยำ (Precision)

ค่าความแม่นยำของข้อมูลโดยพิจารณาแยกทีละคลาส

5.1.4 ค่าประสิทธิภาพโดยรวม (F-Measure)

ค่าประสิทธิภาพโดยรวมจากค่าความครบถ้วน และความแม่นยำของข้อมูลโดยพิจารณาแยกทีละคลาส



รูปที่ 3 การวัดประสิทธิภาพของโมเดลด้วย Confusion Matrix และ F-Measure

5.2 ผลการวิจัย

ในหัวข้อที่ 4 ของงานวิจัยนี้เสนอแนวทางการทดลองเพื่อวิจัยประสิทธิภาพของโมเดลที่มีการจัดการข้อมูลไม่สมดุลในคลาสใหญ่ ดังนั้นในการทดลองนี้ได้จัดโมเดลเป็นกลุ่มดังต่อไปนี้

5.2.1 โมเดลที่ 1

ไม่มีการสุ่มเพิ่มจำนวนข้อมูลในคลาสเล็กหรือการสุ่มลดจำนวนข้อมูลในคลาสใหญ่

5.2.2 โมเดลที่ 2-3

ทำการสุ่มลดจำนวนข้อมูลจากคลาสใหญ่เพียงอย่างเดียวด้วยเทคนิคการเปรียบเทียบความเปลี่ยนแปลง โดยโมเดลที่ 2 ทำการสุ่มลดจำนวนข้อมูลจากคลาสใหญ่ด้วยเทคนิคการเปรียบเทียบความเปลี่ยนแปลงข้อมูล และโมเดลที่ 3 คลาสใหญ่จะทำการถ่วงน้ำหนักข้อมูลก่อนจากนั้นทำการสุ่มลดจำนวนข้อมูลด้วยเทคนิคการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล

5.2.3 โมเดลที่ 4

ใช้การสุ่มเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE ที่คลาสเล็กเพียงอย่างเดียว

ตารางที่ 6 ประสิทธิภาพของโมเดลทั้ง 6 แบบด้วย Confusion Matrix และ F-Measure

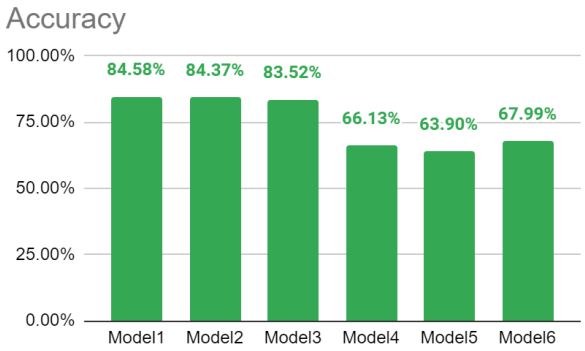
Confusion Matrix		Undersampling Only		
		Model 1	Model 2	Model 3
Technique	Undersampling (80%)	N/A	เปรียบเทียบความเปลี่ยนแปลง	ค่าถ่วงน้ำหนัก + เปรียบเทียบความเปลี่ยนแปลง
	Oversampling	N/A	N/A	N/A
Class	Accuracy	84.58%	84.37%	83.52%
Majority	Precision	84.64%	84.85%	85.36%
	Recall	99.89%	99.23%	97.18%
	F-Measure	91.64%	91.48%	90.89%
Minority	Precision	44.30%	39.35%	35.36%
	Recall	0.49%	2.74%	8.46%
	F-Measure	0.98%	5.12%	13.66%

Confusion Matrix		Mixed Sampling = Undersampling + Oversampling		
		Model 4	Model 5	Model 6
Technique	Undersampling (80%)	N/A	เปรียบเทียบความเปลี่ยนแปลง	ค่าถ่วงน้ำหนัก + เปรียบเทียบความเปลี่ยนแปลง
	Oversampling	SMOTE	SMOTE	SMOTE
Class	Accuracy	66.13%	63.90%	67.99%
Majority	Precision	87.71%	88.10%	88.00%
	Recall	69.74%	66.28%	71.97%
	F-Measure	77.70%	75.65%	79.18%
Minority	Precision	21.81%	21.54%	23.06%
	Recall	46.35%	50.83%	46.13%
	F-Measure	29.66%	30.26%	30.75%

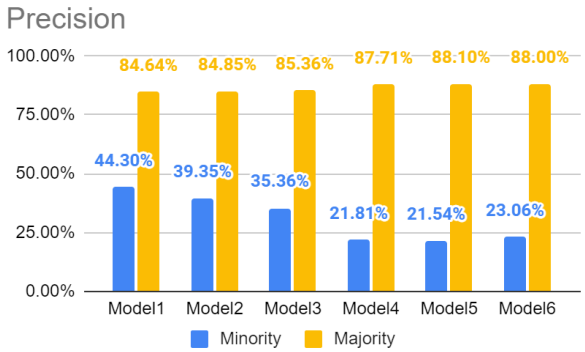
5.2.4 โมเดลที่ 5-6

ใช้วิธีการทดลองแบบผสมโดยการสุ่มลดจำนวนข้อมูลจากคลาสใหญ่ด้วยเทคนิคการเปรียบเทียบความเปลี่ยนแปลง และสุ่มเพิ่มจำนวนข้อมูลจากคลาสเล็กด้วยเทคนิค SMOTE ในโมเดลที่ 5 โดยทำการสุ่มลดจำนวนข้อมูลจากคลาสใหญ่ด้วยเทคนิคการเปรียบเทียบความเปลี่ยนแปลง และทำการสุ่มเพิ่มจำนวนข้อมูลจากคลาสเล็กโดยเทคนิค SMOTE ส่วนโมเดลที่ 6 ในคลาสใหญ่ทำการถ่วงน้ำหนักข้อมูลก่อนจากนั้นทำการสุ่มลดจำนวนข้อมูลด้วยเทคนิคการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล ในส่วนของคลาสเล็กทำการสุ่มเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE

จากการจัดโมเดลทั้ง 4 กลุ่ม ทำให้ได้ผลการทดลองดังตารางที่ 6 ผลการทดลองแสดงให้เห็นค่า Accuracy ในโมเดลที่ 2-3 มีค่าใกล้เคียงกับโมเดลที่ 1 โดยมีค่า Accuracy มากกว่า 80 เปอร์เซ็นต์ ในขณะที่โมเดลที่ 4-6 ทำให้ค่า Accuracy ลดลง แต่อย่างไรก็ตามยังคงมีค่ามากกว่า 60 เปอร์เซ็นต์ดังรูปที่ 4 และในกลุ่มการทดลองแบบการผสม (การทดลองแบบการสุ่มลดจำนวนข้อมูลและการสุ่มเพิ่มจำนวนข้อมูล) ผลการทดลองแสดงให้เห็นว่าโมเดลที่ 6 ให้ค่า Accuracy สูงที่สุดในกลุ่มนี้



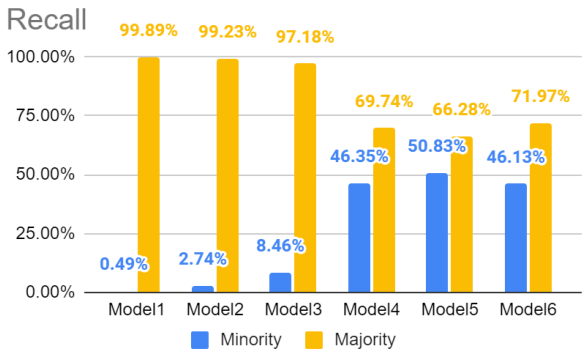
รูปที่ 4 การเปรียบค่า Accuracy ในคลาสเล็กและคลาสใหญ่



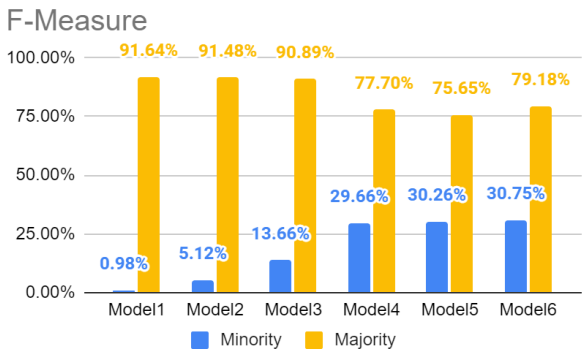
รูปที่ 5 การเปรียบค่า Precision ในคลาสเล็กและคลาสใหญ่

การเปรียบเทียบค่า Precision หรือ ค่าความแม่นยำของข้อมูลในแต่ละคลาสตามรูปที่ 5 จะเห็นได้ว่าทุกโมเดลสำหรับคลาสใหญ่ไม่ค่อยมีความเปลี่ยนแปลง และในแต่ละโมเดลยังคงมีค่ามากกว่า 80 เปอร์เซ็นต์ แต่ในคลาสเล็กค่า Precision จะมีค่า

ลดลง โดยเฉพาะโมเดลที่ 4-6 มีการใช้เทคนิค SMOTE สุ่มลดข้อมูลในคลาสใหญ่



รูปที่ 6 การเปรียบค่า Recall ในคลาสเล็กและคลาสใหญ่



รูปที่ 7 เปรียบเทียบค่า F-Measure ในคลาสเล็กและคลาสใหญ่

การเปรียบเทียบค่า Recall หรือ ค่าความครบถ้วนในโมเดลในแต่ละคลาสในรูปที่ 6 จะแสดงให้เห็นว่าคลาสใหญ่ในโมเดลที่ 1-3 ให้ค่า Recall ได้สูงมาก ขณะที่โมเดล 4-6 การทดลองแบบผสมโดยใช้เทคนิค SMOTE สุ่มลดข้อมูลในคลาสใหญ่ส่งผลให้ค่า Recall ลดลงประมาณ 30 เปอร์เซ็นต์ สำหรับคลาสเล็กของโมเดลที่ 1 ซึ่งไม่มีการสุ่มเพิ่มหรือลดจำนวนข้อมูลเลย ค่า Recall มีค่าไม่ถึง 1 เปอร์เซ็นต์ ในขณะที่โมเดล 2-3 ที่มีการใช้เทคนิคสุ่มลดในคลาสใหญ่อย่างเดียวให้ผลที่ดีขึ้น แต่กลับไม่ค่อยมีความเปลี่ยนแปลงในโมเดลที่ 4-6

การเปรียบเทียบค่า F-Measure หรือ ค่าประสิทธิภาพโดยรวมจากค่าความครบถ้วน และความแม่นยำของข้อมูลในแต่ละคลาส จากรูปที่ 7 จะเห็นว่าคลาสใหญ่ในโมเดลที่ 1-3 ค่า F-Measure ไม่ค่อยมีการเปลี่ยนแปลงและมีค่า F-Measure ที่ประมาณ 90 เปอร์เซ็นต์ ในขณะที่โมเดล 4-6 ที่ใช้เทคนิค SMOTE สุ่มลดข้อมูลในคลาสใหญ่มีค่า F-Measure ลดลงประมาณ 20 เปอร์เซ็นต์ ส่วนคลาสเล็กของโมเดลที่ 1 ซึ่งไม่มีการสุ่มเพิ่มหรือลดจำนวนข้อมูลค่า F-Measure ที่ค่าไม่ถึง 1 เปอร์เซ็นต์ แต่ในโมเดลที่ 2-3 ค่า F-Measure มีการเปลี่ยนแปลงและเพิ่มอย่างเห็นได้ชัด โดยโมเดลที่ 3 ให้ค่า F-Measure ถึง 13.66 เปอร์เซ็นต์ และสำหรับโมเดล 4-6 กลับไม่ค่อยมีความเปลี่ยนแปลงนัก

6. สรุปผลการดำเนินงานวิจัย

จากความกังวลเรื่องการขาดแคลนข้อมูล ความอคติของโมเดล และการสูญเสียข้อมูลไป ทำให้การสุ่มลดข้อมูลเพื่อแก้ปัญหาข้อมูลไม่สมดุลนั้นยังคงเป็นวิธีที่ไม่นิยมทำ แต่ในงานวิจัยทั้ง 6 โมเดล ได้นำการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลและการใช้ค่าถ่วงน้ำหนักผสมกับการเปรียบเทียบความเปลี่ยนแปลงมาประยุกต์ใช้ ซึ่งในกรณีนี้เจ้าข้อมูลต้องการเน้นผลในคลาสเล็กเนื่องจากเป็นกลุ่มเป้าหมายในทางธุรกิจ จากรูปที่ 8 จะเห็นว่าค่าประสิทธิภาพของคลาสเล็ก (Minority) มีค่าประสิทธิภาพโดยรวม (F-Measure) ดีขึ้นโดยมีผลลัพธ์ดังนี้

โมเดลที่ 1 ไม่มีการสุ่มเพิ่มจำนวนข้อมูลในคลาสเล็กและลดจำนวนข้อมูลในคลาสใหญ่ทำให้ผลการทำนายได้ไม่แม่นยำอย่างมาก

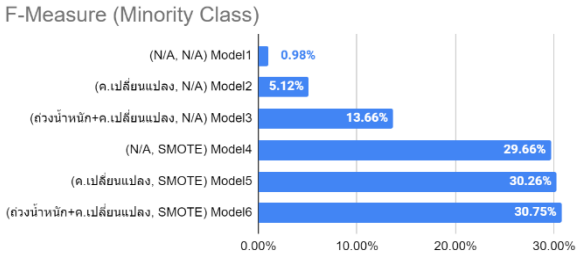
โมเดลที่ 2 ใช้การสุ่มลดที่คลาสใหญ่ด้วยการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลเข้ามาช่วยทำให้ผลดีขึ้นมาเล็กน้อยเมื่อมีจำนวนข้อมูลที่ใกล้เคียงกันกับคลาสเล็ก

โมเดลที่ 3 ใช้การสุ่มลดที่คลาสใหญ่ร่วมกับการถ่วงน้ำหนักก่อนการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลทำให้ผลดีขึ้นมากอย่างเห็นได้ชัด การให้น้ำหนักความสำคัญของข้อมูลให้ผลได้ค่อนข้างดี เมื่อข้อมูลมีความถูกต้อง และมีจำนวนข้อมูลที่ใกล้เคียงกันกับคลาสเล็ก

โมเดลที่ 4 ใช้การสุ่มเพิ่ม SMOTE ที่คลาสเล็กเพียงอย่างเดียว ทำให้เห็นว่าเมื่อข้อมูลมากขึ้นจะทำนายได้ผลดีกว่าเดิม

โมเดลที่ 5 ใช้การสุ่มเพิ่ม SMOTE ที่คลาสเล็กร่วมกับการสุ่มลดที่คลาสใหญ่ด้วยวิธีการเปรียบเทียบความเปลี่ยนแปลงของข้อมูล จะเห็นว่าเมื่อข้อมูลมีความถูกต้อง และมีจำนวนเยอะขึ้นส่งผลให้ทำนายได้ผลดีกว่าเดิม

โมเดลที่ 6 ใช้การสุ่มเพิ่ม SMOTE ที่คลาสเล็กร่วมกับการสุ่มลดที่คลาสใหญ่โดยการถ่วงน้ำหนักและนำไปเปรียบเทียบความเปลี่ยนแปลงของข้อมูล ทำให้เห็นว่าเมื่อข้อมูลมีความถูกต้อง และมีจำนวนมากขึ้น ร่วมกับการให้น้ำหนักความสำคัญจะเป็นวิธีที่ทำนายได้แม่นยำที่สุด และได้ผลลัพธ์ในระดับที่เจ้าของข้อมูลมีความพึงพอใจ



รูปที่ 8 การเปรียบค่า F-Measure ในคลาสเล็ก

การเปรียบเทียบค่าประสิทธิภาพ (F-Measure) โดยรวมจากค่าความครบถ้วน และความแม่นยำของข้อมูลในคลาสเล็กเมื่อนำการถ่วงน้ำหนักร่วมกับการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลมาใช้ในการสุ่มลดข้อมูลในคลาสใหญ่และใช้เทคนิค SMOTE ในการสุ่มเพิ่มข้อมูลในคลาสเล็ก ทำให้ผลลัพธ์ของโมเดลถูกต้องแม่นยำ นอกจากนี้ ปัจจัยที่เจ้าของข้อมูลพิจารณาว่ามีผลต่อธุรกิจและเป็นปัจจัยในทางธุรกิจที่เจ้าของข้อมูลให้ความสำคัญเป็นสิ่งที่ทำให้ข้อมูลมีความถูกต้อง และมีจำนวนเท่ากันมากขึ้น ดังนั้น ในการนำการถ่วงน้ำหนักและการเปรียบเทียบความเปลี่ยนแปลงของข้อมูลมาใช้ในการสุ่มลดข้อมูลในคลาสใหญ่ เจ้าของข้อมูลต้องมีความเข้าใจในธุรกิจเป็นอย่างดี เพื่อสามารถพิจารณาปัจจัยที่สำคัญ เนื่องจากปัจจัยที่เลือกจะส่งผลต่อความถูกต้องและแม่นยำในทำโมเดล

7. เอกสารอ้างอิง

[1] C. Romero, and S. Ventura. “Data mining in education,” Wiley Interdisciplinary Reviews Data Mining Knowledge Discovery., Vol. 3(1), pp. 12-27, 2013.

[2] วิชญ์วิสิฐ เกษรสิทธิ์ ดร.วิชิต หล่อจีระชุณห์กุล และ ดร.จิราวัลย์ จิตรถเวช, “การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน,” วารสารวิจัยมหาวิทยาลัยขอนแก่น (ฉบับบัณฑิตศึกษา)., ปีที่ 18(ฉบับที่ 3), หน้า 11-21, 2561.

[3] Y. Pristyanto, I. Pratama, and A.F. Nugraha. “Data level approach for imbalanced class handling on educational data mining multiclass classification”. International Conference on Information and Communications Technology (ICOIACT). 6-7 Mar. Yogyakarta Indonesia : pp. 310-314, 2018.

[4] E. Rendon, R. Alejo, C. Castorena, F.J. Isidro-Ortega and E.E. Granda-Gutierrez, “Data sampling methods to deal with the big data multi-class imbalance problem,” Applied Sciences., Vol. 10(4):1276, 2020.

[5] ภาณุภณ จิระอัมพร และ เอกสิทธิ์ พัชรวงศ์ศักดา. การปรับปรุงวิธีการสุ่มตัวอย่างใหม่สำหรับข้อมูลไม่สมดุลด้วยเทคนิคแบบผสม DB2SM. วิศวกรรมศาสตรมหาบัณฑิต. วิศวกรรมข้อมูลขนาดใหญ่. วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์, 2564.

[6] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves. “Data imbalance in classification: Experimental evaluation,” Information Sciences., Vol. 513, pp. 429-441, 2020.

[7] รักชนก ปิยพาณิชยกุล และ กฤษณะ ไวยมัย. การแก้ปัญหาข้อมูลไม่สมดุลแบบหลายคลาสสำหรับการทำนายการยกเลิกการใช้บริการอินเทอร์เน็ต. วิทยาศาสตรมหาบัณฑิต. เทคโนโลยีสารสนเทศ. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์, 2023.

[8] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer. “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research., Vol. 16, pp. 321-357, 2002.

[9] T. Elhassan, M. Aljurf, F.Al-Mohanna and M. Shoukri, “Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method,” Global Journal of Technology and Optimization., Special Issues, pp. 1-11, 2016.

[10] R.I. Rashu, N. Haq, and R.M. Rahman. “Data mining approaches to predict final grade by overcoming class imbalance problem”. 17th International conference on computer and information technology (ICCIT). 22-23 Dec. Dhaka Bangladesh : pp. 14-19, 2014.

[11] E.F. Swana, W. Doorsamy, and P. Bokoro, “Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset,” Sensors (Basel)., Vol. 22(9), pp. 1-21, 2022.

[12] A. Hanskunatai. “A New Hybrid Sampling Approach for Classification of Imbalanced Datasets”. 3rd International Conference on Computer and Communication Systems. 27-30 Apr. Nagoya Japan : pp. 67-71, 2018.