วิธีการเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูลที่ไม่สมดุลด้วยการสุ่มแบบสองระดับ

ณัฐณัย เนติรุ่งโรจน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2562

TOP: An Efficient TwO-levels of Positive Resampling Framework for Class Imbalanced Data

Nathaniel Netirungroj

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Big Data Engineering,

College of Innovative Technology and Engineering,

Dhurakij Pundit University

2019

ใบรับรองงานวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ วิธีการเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูลที่ไม่สมดุลด้วยการสุ่มแบบสองระดับ

เสนอโดย ณัฐณัย เนติรุ่งโรจน์

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ

(ศาสตราจารย์ ดร.ธนารักษ์ ธีระมั่นคง)

.....กรรมการและอาจารย์ที่ปรึกษาหลัก

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วรพล พงษ์เพ็ชร)

.....กรรมการ

(ดร.ธนภัทร ฆังคะจิตร)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....

(ผู้ช่วยศาสตราจารย์ ดร.ณรงค์เดช กีรติพรานนท์)

คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 19 เดือน มิ.ย. พ.ศ. 2562

หัวข้อวิทยานิพนธ์	วิธีการเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูลที่ไม่สมดุลด้วยการสุ่มแบบสองระดับ
ชื่อผู้เขียน	ณัฐณัย เนติรุ่งโรจน์
อาจารย์ที่ปรึกษา	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2561

บทคัดย่อ

ในปัจจุบันเทคนิคการจำแนกประเภทข้อมูล (Classification) ได้ถูกนำมาใช้งานอย่างแพร่หลาย วัตถุประสงค์เพื่อต้องการจำแนกข้อมูลหรือพยากรณ์ข้อมูลในรูปแบบต่าง ๆ ตัวอย่างเช่น การจำแนกประเภทข้อมูลการฉ้อโกงออกจากข้อมูลปกติ (Fraud Detection) หรือการคาดการณ์ว่าลูกค้าคนใดบ้างจะยกเลิกการใช้บริการ (Churn Prediction) ปัญหาที่พบเสมอจากสองตัวอย่างนี้คือข้อมูลขาดความสมดุล (Imbalanced Data) เนื่องจากข้อมูลการฉ้อโกงและคนที่ยกเลิกการใช้บริการมีปริมาณน้อยกว่าข้อมูลปกติมาก มีหลายงานวิจัยได้นำเสนอวิธีการปรับสมดุลของตัวข้อมูลเพื่อแก้ไขปัญหาดังกล่าวด้วยวิธีการสุ่มเพื่อแก้ไขข้อมูลแบบต่าง ๆ เช่น การลดปริมาณข้อมูล (Under Sampling) การเพิ่มปริมาณข้อมูล (Over Sampling) หรือ การลดและเพิ่มปริมาณข้อมูล (Under and Over sampling)

วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการใหม่ที่เรียกว่า TOP (TwO-levels of Positive resampling framework) โดยมีวัตถุประสงค์เพื่อเพิ่มประสิทธิภาพในการจำแนกประเภทข้อมูลของโมเดล แนวคิดของวิธีการนี้คือการพิจรณาความสำคัญของข้อมูลเริ่มจากจุดที่สนใจ หากข้อมูลอยู่ไกลจะถือว่ามีความสำคัญน้อย วิธีการนี้จะสร้างขอบเขตของข้อมูลเป็นสองวงล้อมโดยเทียบกับข้อมูลที่สนใจอยู่ ข้อมูลภายในวงล้อมแรกจะถูกเพิ่มปริมาณขึ้น ข้อมูลที่อยู่ในวงล้อมที่สองจะถูกลดปริมาณข้อมูลลง วิธีการนี้ทำให้สามารถนำวิธีการปรับเปลี่ยนข้อมูลจากงานวิจัยก่อนหน้ามาประยุกต์ใช้ได้

ผู้วิจัยได้ทำการทดลองและเปรียบเทียบประสิทธิภาพโดยการใช้วิธีการจำแนกประเภทข้อมูล 11 แบบกับข้อมูลทั้งสิ้น 15 ชุดที่ทำการแก้ไขปริมาณของข้อมูลด้วยการสุ่มแบบต่าง ๆ และพบว่าวิธีการที่นำเสนอมีประสิทธิภาพที่เพิ่มจากวิธีการแก้ไขข้อมูลก่อนหน้าเป็นจำนวนสูงสุดร้อยละ 8.52

Thesis Title	TOP: An Efficient TwO-levels of Positive Resampling Framework for Class Imbalanced Data
Author	Nathaniel Netirungroj
Thesis Advisor	Dr. Eakasit Pacharawongsakda
Department	Big Data Engineering
Academic Year	2018

ABSTRACT

Classification models are widely use in real-world application. Imbalanced class situation tend to occur in many cases. It is when one of the labels has smaller portion than the other, for example, a number of customers who cancel their subscription compared to all subscriber in telecommunication industry. This problem leads to compromised model performances. There are many pre-processing techniques have been proposed such as under-sampling, over-sampling and hybrid-sampling to handle this situation. In this work, we proposed an alternative data pre-processing framework called TwO-levels of Positive resampling (TOP). The main idea of this method is to perform resampling task in two areas around each minority instance. It generates synthetic minority in an inner area which is closest to genuine minority while reduce majority class that located in an outer area. With this approach, artificial data points were created more carefully with less majority information loss. We have benchmarked TOP with 3 types of resampling techniques including over-sampling, under-sampling, and hybrid-sampling by training 11 machine learning model on 15 datasets. As a result, our technique has improved model performance up to 8.52 percent compared to other techniques.

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณในความกรุณาของอาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา เป็นอย่างสูงที่เสียสละเวลาอันมีค่าเพื่อให้คำปรึกษาและคำแนะนำในการทำวิจัย ตลอดระยะเวลาที่ผ่านมา อาจารย์ได้ให้ข้อเสนอแนะ ความคิดเห็นและทรัพยากรที่มีประโยชน์ต่องานวิจัยชิ้นนี้ รวมถึงเอาใจใส่ผู้วิจัยเป็นอย่างดี ขอขอบพระคุณอาจารย์ที่สละเวลาเพื่อมาเป็นกรรมการในการสอบวิทยานิพนธ์ และได้ให้คำแนะนำ แนวทางทราเป็นประโยชน์ต่องานวิจัย ขอขอบคุณเจ้าหน้าที่หลักสูตรวิศวกรรมศาสตร์มหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ทุกท่านสำหรับการช่วยเหลือและประสานงานเพื่อให้การดำเนินการทำวิจัยเป็นไปอย่างราบรื่น ขอขอบคุณมิตรสหายทุกท่านที่ได้ให้คำปรึกษาและให้กำลังใจในเสมอมา ตลอดจนมหาวิทยาลัยที่ให้โอกาสในการศึกษาเรียนรู้ตามความสนใจของผู้วิจัย และสุดท้ายผู้วิจัยขอขอบคุณครอบครัวที่ให้การสนับสนุน และความเข้าใจ ซึ่งทั้งหมดได้ส่งผลให้งานวิจัยสำเร็จลุล่วงไปได้ด้วยดี หากมีสิ่งใดที่ผู้วิจัยได้ทำผิดพลาดหรือบกพร่องประการใด ผู้วิจัยต้องกราบขออภัยเป็นอย่างสูงมา ณ โอกาสนี้ ผู้วิจัยหวังเป็นอย่างยิ่งว่าวิทยานิพนธ์ฉบับนี้จะเป็นพื้นฐานในการต่อยอดองค์ความรู้ของผู้ที่สนใจศึกษาในงานด้านนี้ต่อไป

ณัฐณัย เนติรุ่งโรจน์

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

การจำแนกประเภทข้อมูล (Data Classification) เป็นงานหนึ่งที่สำคัญและเป็นพื้นฐานในการเรียนรู้ของเครื่อง (Machine Learning) รวมถึงวงจรชีวิตของมนุษย์และสิ่งมีชีวิตอื่นอีกด้วย ข้อมูลมากมายเกิดขึ้นและถูกประมวลผลอยู่ตลอดเวลา ซึ่งเป็นสิ่งที่ทำให้เกิดการตัดสินใจต่าง ๆ คอมพิวเตอร์ก็เช่นเดียวกัน การตัดสินใจจะดำเนินไปได้ก็ต่อเมื่อมีข้อมูลตั้งต้นและรู้ว่าข้อมูลเหล่านั้นมีความแตกต่างกันอย่างไร โดยข้อมูลจะต้องมีจำนวนลักษณะที่ระบุได้อย่างน้อยที่สุดสองแบบ (Binary) จากนั้นจึงจะเกิดการเรียนรู้จากข้อมูล แล้วจึงนำไปใช้จำแนกลักษณะได้ในลำดับถัดไป

ความสมดุลระหว่างตัวอย่างข้อมูลนั้นเกิดขึ้นได้ยากในโลกความเป็นจริง เนื่องจากปัจจัยการเกิดขึ้นของข้อมูลนั้นมีหลากหลายเหตุผล โอกาสที่ข้อมูลตัวอย่างจะมีขนาดไม่เท่ากันจึงเป็นไปได้สูงมาก เมื่อข้อมูลที่ตัวอย่างไม่สมดุลกันถูกนำมาใช้ในการฝึกโมเดลเพื่อการจำแนก (Classification Model) ผลที่จะเกิดขึ้นอย่างหลีกเลี่ยงไม่ได้คือโมเดลเหล่านั้นจะมีความเอนเอียง (Bias) ด้วยข้อมูลที่มีปริมาณมากกว่า (Negative) และในขณะเดียวกันข้อมูลที่มีปริมาณน้อยกว่า (Positive) ก็มีปริมาณไม่เพียงพอต่อการเรียนรู้ของโมเดล ดังนั้นประสบการณ์ของโมเดลจึงไม่มากพอที่จะจำแนกข้อมูลใหม่ได้อย่างแม่นยำ จึงส่งผลให้โมเดลนั้นมีประสิทธิภาพที่ต่ำในท้ายที่สุด ข้อมูลที่ขาดความสมดุลเป็นปัญหาที่พบได้บ่อยในการวิเคราะห์ข้อมูล ซึ่งส่งผลโดยตรงกับประสิทธิภาพของโมเดลการเรียนรู้ของเครื่อง มีงานวิจัยจำนวนไม่น้อยที่พยายามนำเสนอแนวทางการแก้ไขปัญหาดังกล่าวในรูปแบบต่าง ๆ ซึ่งวิธีการแต่ละวิธีสามารถรับมือกับลักษณะของข้อมูลที่แตกต่างกันออกไปในหลาย ๆ ระดับ ตลอดไปจนถึงวิธีการแก้ปัญหาที่อ้างอิงกับโจทย์เฉพาะทาง (Domain Specific) อีกด้วย

อย่างไรก็ดี เนื่องจากข้อมูลนั้นแต่ละชุดนั้นมีลักษณะที่แตกต่างกันออกไป ข้อมูลบางประเภทมีรูปแบบที่ไม่ชัดเจน (Weak Pattern) บางประเภทมีสัญญาณรบกวนมาก (Noisy) บางประเภทมีความหนาแน่นสูง (Dense) เป็นต้น ดังนั้นวิธีการปรับสมดุลข้อตัวอย่างข้อมูลควรจะเหมาะสมและสอดคล้องไปตามลักษณะดังกล่าว โดยเฉพาะอย่างยิ่งหากข้อมูลตัวอย่าง Positive ที่มีข้อมูลตัวอย่าง Negative กระจุกตัวอยู่โดยรอบอย่างใกล้ชิดจนอาจเกิดการทับซ้อน (Overlapping) จะเพิ่มระดับความยากในการตัดสินใจให้โมเดล นอกเหนือจากนี้ การใช้วิธีการปรับสมดุลต่าง ๆ อาจไม่จำเป็นต้องทำงานอย่างครอบคลุมทุกพื้นที่ของข้อมูล เพราะนั่นอาจส่งผลให้ข้อมูลที่สร้างขึ้นใหม่หรือข้อมูลที่ถูกลบออกมีความผิดพลาดได้เพื่อนำไปฝึกโมเดล นอกเหนือจากนี้ จำนวนของข้อมูลตัวอย่างอาจไม่จำเป็นต้องมีขนาดเท่ากันเสมอไปอีกด้วย

จากเหตุผลดังกล่าว ผู้วิจัยจึงได้นำเสนอแนวทางในการเพิ่มความแม่นยำในการจำแนกของโมเดล โดยการปรับปรุงความสมดุลของชุดข้อมูลด้วยวิธีการพิจารณาข้อมูล Positive ทุกตัวอย่างเพื่อพยายามลดการกระจุกตัวของข้อมูล Negative ที่ตั้งอยู่รอบ ๆ และในขณะเดียวกันจะเพิ่มจำนวน Positive ขึ้นมาอีกด้วย นอกจากนี้จะดำเนินการเปรียบเทียบผลลัพธ์ของการวัดประสิทธิภาพของโมเดลต่าง ๆ เมื่อนำมาใช้ร่วมกับวิธีการปรับสมดุลหลายประเภท

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อนำเสนอวิธีการปรับสมดุลข้อมูลเพื่อเพิ่มประสิทธิภาพของโมเดลที่ใช้ในการจำแนกข้อมูลในแง่ของความถูกต้องแม่นยำบนข้อมูลที่มีความไม่สมดุล
2. เพื่อนำเสนอวิธีการปรับสมดุลข้อมูลที่สามารถทำงานร่วมกับโมเดลพื้นฐานได้หลายประเภท
3. เพื่อเปรียบเทียบประสิทธิภาพด้านความถูกต้องแม่นยำของโมเดลในการจำแนกข้อมูลโดยการใช้หลักการแก้ไขตัวอย่างแบบต่าง ๆ

ทั้งนี้ทฤษฎีที่เกี่ยวข้องจะกล่าวถึงในบทที่ 2 และรายละเอียดของระเบียบวิธีวิจัยของวิธีการที่นำเสนอจะกล่าวถึงในบทที่ 3 ในส่วนของผลการวิจัยจะกล่าวถึงในบทที่ 4 และในบทที่ 5 จะกล่าวถึงผลสรุปของการวิจัย ข้อจำกัด รวมถึงข้อเสนอแนะ

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถนำวิธีการที่นำเสนอไปใช้ปรับสมดุลข้อมูลเพื่อช่วยให้การสร้างโมเดลประเภทต่าง ๆ ให้มีประสิทธิภาพมากยิ่งขึ้น

2. สามารถนำวิธีการที่นำเสนอไปประยุกต์ใช้กับงานเฉพาะทางที่ลักษณะข้อมูลมีความขาดสมดุล และมีความทับซ้อนแฝงอยู่ เช่น ข้อมูลเซ็นเซอร์ ข้อมูลการโกง ข้อมูลพฤติกรรมผู้บริโภค หรือ ข้อมูลเครื่องจักร

1.4 ขอบเขตของงานวิจัย

งานวิจัยนี้กำหนดขอบเขตของงานออกเป็น 3 ส่วนดังนี้

1. ขอบเขตของแผนงาน อธิบายถึงงานที่จะทำการพัฒนา

เป็นการศึกษาเพื่อเพิ่มประสิทธิภาพความถูกต้องแม่นยำของโมเดลการเรียนรู้ของที่ใช้ในการจำแนกข้อมูล โดยนำวิธีการปรับสมดุลข้อมูล (Sampling) มาใช้เพื่อแก้ไขข้อมูลที่ขาดความสมดุลก่อนใช้ในการฝึกโมเดล ซึ่งจะไม่รวมถึงแนวทางการแก้ปัญหาด้วยแนวทางการให้น้ำหนัก (Cost Sensitive) และการพิจารณาเลือกตัวแปร (Feature Selection)

2. ขอบเขตของข้อมูล อธิบายถึงข้อมูลที่นำมาใช้ในงานวิจัย

ข้อมูลทั้งหมดที่ใช้นำมาจากแหล่งสาธารณะที่ถูกยอมรับอย่างเป็นวงการในการทดลองงานวิจัย โดยจะเป็นข้อมูลที่มีลาเบลจำนวน 2 ค่าเท่านั้น มีสัดส่วนความไม่สมดุลอยู่ระหว่าง 1.5 ถึง 9 และระดับความทับซ้อนอยู่ระหว่าง 0 ถึง 0.7

บทที่ 2

แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง

การวิจัยเรื่อง วิธีการสุ่มข้อมูลโดยพิจารณาตัวอย่างที่สนใจในพื้นที่โดยรอบข้อมูล Positive สองชั้นด้วย ผู้วิจัยได้ทำการศึกษา ค้นคว้า จากแหล่งความรู้ทางอินเตอร์เน็ต โดยงานวิจัยมีแนวคิดที่เกี่ยวข้อดังนี้

2.1.1 การเรียนรู้ของเครื่อง (Machine Learning)

การที่เครื่องคอมพิวเตอร์สามารถเรียนรู้งาน (Task) โดยอ้างอิงจากความรู้ของชุดข้อมูลที่เกิดขึ้น เพื่อที่จะสามารถทำงานได้เองอย่างมีประสิทธิภาพ โดยการเรียนรู้ของเครื่องนั้นมีองค์ประกอบที่สำคัญ 3 ส่วนดังนี้

2.1.1.1 ชุดข้อมูล (Dataset)

คือผลของเหตุการณ์ที่เกิดขึ้นหลาย ๆ ครั้ง โดยข้อมูลเหล่านี้จะถูกเก็บไว้ในฐานข้อมูลเพื่อเตรียมนำมาใช้ในการวิเคราะห์เพื่อสกัดหาความรู้ ภายในชุดข้อมูลจะประกอบไปด้วยข้อมูลตัวอย่าง (Example) จำนวนหนึ่ง ที่มีตัวบ่งชี้คุณสมบัติ หรือตัวแปร (Feature) ต่าง ๆ ของข้อมูล ซึ่งคุณสมบัติของข้อมูลตัวอย่างแต่ละตัวสามารถมีลักษณะที่เหมือนหรือต่างกันก็ได้

ชุดข้อมูลจะเป็นสิ่งที่กำหนดของเขตความสามารถของการเรียนรู้ของเครื่อง รวมไปถึงแนวทางการที่จะใช้ในการฝึกเพราะเนื่องจากชุดข้อมูลนั้นอาจจะมีรูปแบบที่ไม่เหมือนกัน รูปแบบของข้อมูลแบบออกเป็น 2 ประเภทคือ ข้อมูลที่ตัวอย่างที่มีป้ายกำกับไว้ (Labeled) หรือเรียกว่า “คำตอบ” และข้อมูลที่

ตัวอย่างข้อมูลไม่มีป้ายกำกับไว้ (Unlabeled) ยกตัวอย่างเช่นข้อมูลขนาดเสื้อยืด ข้อมูลของเสื้อแต่ละตัวจะประกอบไปด้วย ความยาว ความกว้าง และสี ถ้าหากข้อมูลชุดนี้มีคุณสมบัติกำกับไว้ก็จะรู้ว่าเสื้อแต่ละตัวนั้นอยู่ในกลุ่ม (Class) ขนาดแบบใด เล็ก ปานกลาง หรือใหญ่ แต่ถ้าหากไม่มีคุณสมบัติกำกับไว้ ก็จะไม่สามารถรู้กลุ่มที่ของขนาดที่แน่ชัด

2.1.1.2 โมเดล (Model)

คือขั้นตอนทางคณิตศาสตร์ที่นำไปใช้ในการประมวลผลข้อมูล โดยแนวทางในการนำโมเดลไปใช้เรียนรู้ข้อมูลนั้นสามารถแบ่งออกได้เป็น 2 ประเภทหลักดังนี้

2.1.1.2.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning)

การเรียนรู้ประเภทนี้จะเป็นการสร้างโมเดลจากชุดข้อมูลสำหรับการสอน (Training dataset) เพื่อสร้างโมเดล (Model) ออกมา แล้วจึงนำโมเดลนั้นไปใช้เพื่อช่วยให้เครื่องคอมพิวเตอร์สามารถประมวลผลข้อมูลชุดใหม่ได้เอง โมเดลนั้นมีอยู่หลายประเภทเช่น โมเดลที่มีลักษณะเป็นกฎ (Rule Based) ความน่าจะเป็น (Probabilistic Based) ระยะห่าง (Distance Based) หรือโครงข่ายประสาท (Neural Network Based) เป็นต้น ขั้นตอนการสร้างโมเดลนั้นเรียกว่าการฝึกโมเดล (Train model) แล้วจึงทดสอบประสิทธิภาพโมเดล (Test model) โดยโมเดลได้จะมีความแตกต่างกันขึ้นอยู่กับชุดข้อมูลฝึก ประเภทของโมเดลที่ใช้ และแนวทางในการฝึก และที่สำคัญคือโมเดลที่ได้ออกมานั้นจะตั้งอยู่บนพื้นฐานของชุดข้อมูลฝึกที่ใช้ในการสอนโมเดล

ประเภทของโมเดลแบ่งออกได้เป็น 2 ประเภทหลักคือ (1) โมเดลที่ทำงานกับข้อมูลที่ตัวอย่างข้อมูลถูกกำกับด้วยค่าตัวเลขจำนวนจริงซึ่งเป็นค่าเชิงปริมาณ (Ordinal) หรือแบบต่อเนื่อง (Continuous) เรียกว่าโมเดลเชิงการถดถอย (Regression Algorithm) (2) โมเดลที่ทำงานกับข้อมูลที่ตัวอย่างข้อมูลถูกกำกับด้วยค่าเชิงคุณภาพ (Categorical) หรือไม่ต่อเนื่อง (Discrete) เรียกว่าโมเดลสำหรับจำแนก (Classification Algorithm) ซึ่งโมเดลทั้งสองประเภทนี้สามารถนำไปใช้ในการประมวลผลข้อมูลตัวอย่างใหม่ที่จะเกิดขึ้นในอนาคตอันใกล้เพื่อหาคำตอบได้

2.1.1.2.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

การเรียนรู้ประเภทนี้จะไม่นำชุดข้อมูลมาใช้ในการสอนโมเดล และข้อมูลจะต้องไม่มีคำตอบกำกับไว้ การทำงานจะเป็นการนำชุดข้อมูลส่งไปให้โมเดลอธิบายลักษณะของข้อมูลนั้นเพื่อทำการจัดกลุ่มข้อมูล (Clustering) โดยอ้างอิงจากความคล้ายคลึงหรือความต่างของตัวอย่างข้อมูลของชุดข้อมูลที่ใช้ ณ ขณะนั้น การเรียนรู้ประเภทนี้จะไม่สามารถนำไปใช้เพื่อพยากรณ์คำตอบของข้อมูลชุดใหม่ได้อย่างตรงไปตรงมา

2.1.1.3 การวัดประสิทธิภาพของโมเดล (Evaluation Metric)

การวัดประสิทธิภาพของโมเดลนั้นช่วยตรวจสอบค่าความคลาดเคลื่อนของโมเดลว่ามีความแม่นยำมากน้อยเพียงใด เครื่องมือที่ใช้วัดนั้นสามารถแบ่งออกได้ตามประเภทของโมเดล สำหรับโมเดลที่ใช้เพื่อจำแนกประเภทข้อมูลโดยวิธีการเรียนรู้แบบมีผู้สอนนั้นมีเครื่องมือวัดผลที่สำคัญดังนี้

2.1.1.3.1 ตารางแจกแจงผลลัพธ์ (Confusion Matrix)

คือ เครื่องมือที่ใช้สำหรับแสดงข้อมูลผลการทำนายของโมเดลเมื่อเทียบกับข้อมูลผลความเป็นจริง จำนวนแถวและคอลัมน์ในตารางนั้นขึ้นอยู่กับจำนวนของกลุ่มข้อมูลภายในป้ายกำกับคูณกับจำนวนกลุ่มที่จะสามารถเป็นไปได้ ตารางฝั่งซ้ายคือ ข้อมูลเปรียบเทียบระหว่างผลของข้อมูลจริงและผลการทำนาย

ตารางเปรียบเทียบผลของโมเดล	
ผลความเป็นจริง	ผลการทำนาย
กลุ่ม ก	กลุ่ม ข
กลุ่ม ก	กลุ่ม ข
กลุ่ม ข	กลุ่ม ข
กลุ่ม ข	กลุ่ม ก
กลุ่ม ข	กลุ่ม ข
กลุ่ม ก	กลุ่ม ก
กลุ่ม ก	กลุ่ม ก
กลุ่ม ข	กลุ่ม ข
กลุ่ม ก	กลุ่ม ข
กลุ่ม ก	กลุ่ม ข

กำหนดให้ข้อมูลมีป้ายกำกับมีจำนวน 2 กลุ่ม ประกอบด้วย กลุ่ม ก เป็นข้อมูลเชิงบวก และ กลุ่ม ข เป็นข้อมูลเชิงลบ

จำนวนข้อมูลทั้งหมด 10 ตัวอย่าง		ผลความเป็นจริง	
		กลุ่ม ก	กลุ่ม ข
ผลการทำนาย	กลุ่ม ก	2	4
	กลุ่ม ข	1	3

ภาพที่ 2.1 ลักษณะตารางเทียบระหว่างผลความจริงของข้อมูลและผลการทำนาย

จำนวนตัวเลขการแจงแจงภายในตารางแจกแจงผลลัพธ์สามารถอธิบายได้ดังนี้

กรณีที่ 1 โมเดลทำนายข้อมูลกลุ่ม ก ไปเป็นกลุ่ม ก โดยที่กำหนดกลุ่ม ก คือ กลุ่มที่พิจารณาเป็นหลัก เรียกว่า ถูกต้องในเชิงบวก (True Positive)

กรณีที่ 2 โมเดลทำนายข้อมูลกลุ่ม ข ไปเป็นกลุ่ม ข โดยที่กำหนดกลุ่ม ก คือ กลุ่มที่พิจารณาเป็นหลัก เรียกว่า ถูกต้องในเชิงลบ (True Negative)

กรณีที่ 3 โมเดลทำนายข้อมูลกลุ่ม ข ไปเป็นกลุ่ม ก โดยที่กำหนดกลุ่ม ก คือ กลุ่มที่พิจารณาเป็นหลัก เรียกว่า ผิดพลาดในเชิงบวก (False Positive)

กรณีที่ 4 โมเดลทำนายข้อมูลกลุ่ม ก ไปเป็นกลุ่ม ข โดยที่กำหนดกลุ่ม ก คือ กลุ่มที่พิจารณาเป็นหลัก เรียกว่า ผิดพลาดในเชิงลบ (False Negative)

ตารางที่ 2.1 ลักษณะตารางแจกแจงผลลัพธ์โดยแบ่งตามชื่อเรียก

		ผลความเป็นจริง	
		กลุ่ม ก	กลุ่ม ข
ผลการทำนาย	กลุ่ม ก	ถูกต้องในเชิงบวก (True Positive)	ผิดพลาดในเชิงบวก (False Positive)
	กลุ่ม ข	ผิดพลาดในเชิงลบ (False Negative)	ถูกต้องในเชิงลบ (True Negative)

(2)

2.1.1.3.2 ค่าร้อยละความถูกต้อง (Accuracy)

คือ ค่าที่อธิบายถึงประสิทธิภาพความถูกต้องโดยรวมของโมเดลที่มีต่อข้อมูลทั้งหมด สามารถคำนวนได้จากการนำค่าที่ได้จากตารางการแจกแจงผลลัพธ์มาพิจารณาดังสมการต่อไปนี้

$$accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

2.1.1.3.3 ค่าร้อยละความผิดพลาด (Error)

คือ ค่าที่อธิบายถึงความผิดพลาดโดยรวมของโมเดลที่มีต่อข้อมูลทั้งหมด สามารถคำนวนได้จากการนำค่าที่ได้จากตารางการแจกแจงผลลัพธ์มาพิจารณาดังสมการต่อไปนี้

$$Error = 1 - accuracy$$

2.1.1.3.4 ค่าร้อยละของจำนวนของผลการทำนายที่ถูกต้องของกลุ่มข้อมูลที่สนใจ

คือ ค่าที่อธิบายถึงความถูกต้องของกลุ่มข้อมูลที่กำลังพิจารณาเมื่อเทียบกับผลลัพธ์ของการทำนาย สามารถคำนวณได้สมการดังต่อไปนี้

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

2.1.1.3.5 ค่าร้อยละจำนวนความถูกต้องของกลุ่มข้อมูลที่สนใจ (Recall)

คือ ค่าที่อธิบายถึงความถูกต้องของผลการทำนายของกลุ่มข้อมูลที่กำลังพิจารณาอยู่เมื่อเทียบกับผลของความเป็นจริง

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

2.1.1.3.6 ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score)

คือ ค่าเฉลี่ยของค่ากลางของผลจากการหารจำนวนข้อมูลทั้งหมด

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.1.1.3.7 ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean)

คือ ค่าเฉลี่ยของค่ากลางของผลจากการหารจำนวนข้อมูลทั้งหมด

$$GM = \sqrt{\frac{True\ Positive}{True\ Positive+False\ Negative} \times \frac{True\ Positive}{True\ Positive+Positive}}$$

2.1.1.3.8 ค่าพื้นที่ใต้กราฟแสดงความถูกต้อง (Area Under the Receiver Operating Characteristics Curve)

$$AU_{ROC} = \int_0^1 \frac{TP}{P}\, d\frac{FP}{N}$$

2.1.1.3.9 ค่าพื้นที่ใต้กราฟแสดงความถูกต้อง (Area Under the Precision Recall Curve)

$$AU_{PRC} = \int_0^1 p(r)\, dr,$$

2.1.1.3.10 การตรวจสอบความถูกต้องของโมเดลโดยการแบ่งข้อมูลฝึกแบบหลายชุด (Cross Validation)

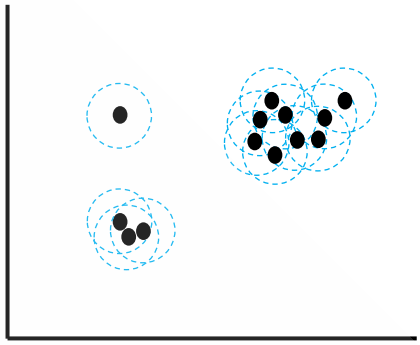
คือ วิธีการที่ใช้ทดสอบความถูกต้องของโมเดลโดยการแบ่งข้อมูลออกหลายส่วนเพื่อสอนโมเดล แล้วจึงใช้ข้อมูลบางส่วนจากข้อมูลเหล่านั้นมาใช้ทดสอบบประสิทธิภาพความถูกต้องของโมเดล จำนวนการแบ่งข้อมูลสามารถแทนได้ด้วยค่า k เช่นหากต้องการแบ่งข้อมูลเพื่อทดสอบ วน ดังนั้น k จะมีค่าเท่ากับ 5 โดยโมเดลจะถูกฝึกด้วยข้อมูล 4 ส่วน แล้วทดสอบด้วยข้อมูล 1 ส่วนที่เหลือ ขั้นตอนนี้จะทำงานวนไปจนกระทั่งสามารถใช้ข้อมูลทุกส่วนมาฝึกและทดสอบโมเดล จากนั้นจึงนำผลการทดสอบทั้งหมด 5 ครั้งมาหาค่าเฉลี่ยเพื่อแสดงค่าความถูกต้องเฉลี่ยของโมเดล

2.1.2 วิธีการจัดกลุ่มข้อมูล (Clustering Method)

2.1.2.1 การจัดกลุ่มตามความหนาแน่น (DBSCAN)

วิธีนี้ถูกออกบบมาเพื่อใช้ในการจัดกลุ่มข้อมูลโดยอ้างอิงจากความหนาแน่นของข้อมูลในแต่ละพื้นที่ โดยสมมติฐานของลักษณะความหนาแน่นของข้อมูลนั้นจะต้องเหมือนหรือคล้ายคลึงกันมาก การทำงานขั้นแรกคำนวณระยะห่างระหว่างข้อมูล โดยฟังก์ชั่นการวัดระยะห่างที่นิยมใช้คือระยะห่างยูคลิด จากนั้นทำการพิจารณาข้อมูลทุกตัวโดยนำค่าระยะห่าง (epsilon) มาใช้ในการกำหนดพื้นที่ที่ใช้ใน

การตัดสินใจว่ามีความหนาแน่นเกิดขึ้นภายในชุดข้อมูล โดยจะนับจากจำนวนข้อมูลที่อยู่รอบ ๆ ข้อมูลที่กำลังพิจารณาจำนวน n ตัว (minPts) จากนั้นจึงจะติดป้ายกำกับว่าข้อมูลตัวอย่างใดอยู่ในกลุ่มข้อมูลใด



ภาพที่ 2.2 ลักษณะการทำงานของการแบ่งกลุ่มข้อมูล DBSCAN

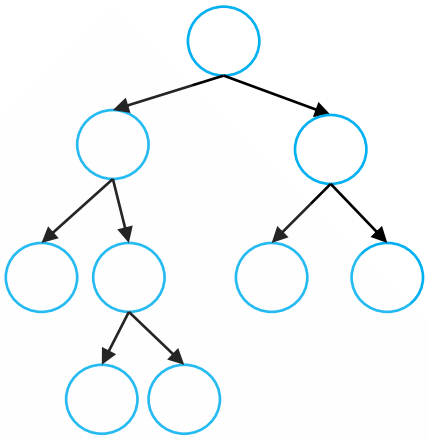
ข้อดีข้อวิธีการนี้คือสามารถแบ่งกลุ่มข้อมูลตามความหนาแน่นได้ดี ข้อมูลไม่จำเป็นต้องสามารถอธิบายได้ด้วยสมการเส้นตรง รวมถึงทนทานต่อข้อมูลที่ผิดแปลก (Outlier) ได้ นอกจากนี้หากผู้ใช้มีความคุ้นเคยกับข้อมูลมากพอก็จะสามารถกำหนดระยะห่างได้ง่าย ตลอดจนไม่จำเป็นต้องกำหนดจำนวนกลุ่มที่ต้องการแบ่ง อย่างไรก็ดีข้อเสียของวิธีการนี้คือมีความอ่อนไหวสูงจึงต้องกำหนดระยะห่างอย่างระมัดระวัง และหากลักษณะของความหนาแน่นภายในชุดข้อมูลมีหลายชนิด และมีความแตกต่างกันมาก ก็จะทำให้ค่าของระยะห่างที่กำหนดไม่สามารถใช้ร่วมกันได้ จึงส่งผลให้การแบ่งกลุ่มมีความคลาดเคลื่อนได้

2.1.3 วิธีการจำแนกข้อมูล (Classification Model)

2.1.3.1 การเรียนรู้แบบเดี่ยว (Single Classifier)

2.1.3.1.1 ต้นไม้ตัดสินใจ (Decision Tree)

โมเดลประเภทนี้มีความสามารถในการสร้างสูตรสำเร็จออกมาเป็นกฎที่มีลักษณะเป็นโครงสร้างตามแนวดิ่ง (Hierarchy Structure) โดยอ้างอิงจากลักษณะของข้อมูลที่ใช้ฝึก ภายในของโมเดลประกอบไปด้วย 2 ส่วนสำคัญดังนี้



ภาพที่ 2.3 ลักษณะการแบ่งโหนดของต้นไม้ตัดสินใจ

2.1.3.1.2 โหนด (Node)

เป็นสิ่งที่บรรจุค่าของข้อมูลไว้ ต้นไม้จะนำคุณสมบัติของข้อมูลมาพิจารณาแล้วประมวลผลออกมาเป็นค่าความน่าจะเป็นของกลุ่มข้อมูลทุกกลุ่มในคุณสมบัติดังกล่าว ประเภทของโหนดมี 3 ประเภทดังนี้ (1) โหนดราก (Root Node) คือกฎชั้นแรกสุดของต้นไม้ตัดสินใจที่ถูกสร้างขึ้นมาโดยอ้างอิงจากคุณสมบัติที่มีความสามารถในการจำแนกข้อมูลตามลักษณะได้ชัดเจนที่สุด (2) โหนดตัดสินใจ (Decision Node) คือกฎที่ต่อเนื่องมาจากโหนด สำหรับกฎในชั้นนี้จะนำคุณสมบัติของข้อมูลที่มีความสำคัญน้อยกว่าคุณสมบัติที่อยู่ในชั้นรากมาพิจารณา โดยโหนดประเภทนี้เกิดขึ้นได้หลายโหนด และแต่ละโหนดสามารถเชื่อมต่อกันได้ตามแนวดิ่ง (3) โหนดใบ (Leaf Node) คือกฎชั้นสุดท้ายที่เกิดขึ้นของโหนดตัดสินใจแต่ละโหนด ซึ่งเป็นโหนดที่เก็บป้ายกำกับคำตอบไว้ โหนดใบจะเกิดขึ้นได้ต่อเมื่อต้นไม้ไม่สามารถสร้างกฎเพื่อจำแนกข้อมูลได้อีก

คุณสมบัติของแต่ละโหนดนั้นจะมีความสำคัญต่างกัน โหนดที่อยู่ชั้นบนสุดจะมีความสำคัญมากที่สุด เพราะนำคุณสมบัติที่มีความสัมพันธ์กับคำตอบมากที่สุดมาใช้ โดยความสำคัญของคุณสมบัติต้นไม้เลือกมาพิจารณานั้นจะลดน้อยลงมาตามลำดับของชั้นในต้นไม้

2.1.3.1.3 การแบ่งโหนด (Split)

คือการคำนวณเพื่อที่จะแบ่งแยกโหนดออกไปสร้างเป็นโหนดใหม่ภายในโหนดที่กำลังพิจารณา โดยการหาค่าของการแบ่งแยกนั้นสามารถทำได้โดยเริ่มจากหาค่าที่บ่งบอกถึงระดับของกลุ่ม

ข้อมูลภายในคุณสมบัตินั้น (Entropy) เมื่อเทียบกับผลรวมของกลุ่มข้อมูลทั้งหมดและเมื่อเทียบรายกลุ่มข้อมูล ตามลำดับ

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

โดยกำหนดให้ E คือค่า Entropy และ p หมายถึงค่าความน่าจะเป็น (0 ถึง 1) ของสิ่งที่เกิดขึ้นในเหตุการณ์ที่สนใจ (T, X) เมื่อนำมาเทียบกับคลาสคำตอบ c แล้วจึงนำผลลัพธ์ไปใช้คำนวณหาค่าความสำคัญของคุณสมบัตินั้น (Information Gain) ดังสมการต่อไปนี้

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

ผลลัพธ์ที่ได้จะบ่งบอกถึงระดับความสำคัญของคุณสมบัติที่สนใจในชุดข้อมูลนั้น โดยมีค่าตั้งแต่ 0 ถึง 1 หากค่ายิ่งเข้าใกล้ 1 หมายความว่าคุณสมบัตินั้นมีความสำคัญมาก และจะถูกหยิบมาใช้ในการแบ่งข้อมูลในชั้นบน จากนั้นจะสามารถนำไปใช้ประมวลผลข้อมูลชุดใหม่ได้

ต้นไม้ตัดสินใจที่นิยมใช้งานอย่างแพร่หลายโดยเฉพาะ C4.5 และ C5.0 ซึ่งความแตกต่างระหว่างสองแบบนี้คือ ความเร็วในการทำงานและการใช้ทรัพยากรของคอมพิวเตอร์ โดยที่ C5.0 มีความสามารถที่ดีกว่า C4.5 รวมถึงความแม่นยำในการจำแนก เพราะ C5.0 มีการนำค่าความผิดพลาดในการจำแนก (Variable Misclassification Costs) มาใช้แบบแยกส่วนไปตามแต่ละตัวอย่างข้อมูล ในขณะที่ C4.5 มองว่าค่าความผิดพลาดของแต่ละตัวอย่างนั้นมีลักษณะเหมือนกัน

ข้อดีของต้นไม้ตัดสินใจที่ชัดเจนที่สุดคือเป็นโมเดลที่มีความซับซ้อนต่ำ จึงทำให้ผลลัพธ์ที่ได้จากการทำงานสามารถแปรผลได้ง่าย สามารถทำงานกับข้อมูลที่มีตัวแปรมีความสัมพันธ์กันแบบไม่เป็นเส้นตรงได้ดี นอกจากนี้ยังสามารถฝึกได้อย่างรวดเร็ว อย่างไรก็ดี เนื่องจากโมเดลนี้ใช้หลักการแบ่งแยกข้อมูลโดยอ้างอิงตามค่าของข้อมูลเป็นหลัก ทำให้บางครั้งประสิทธิภาพความแม่นยำของโมเดลไม่สามารถเพิ่มขึ้นตามขนาดของข้อมูลได้ ถ้าหากตัวแปรที่ใช้มีค่าความสัมพันธ์ระหว่างกันต่ำเป็นจำนวนมากเกินไปหรือตัวแปรที่ใช้เป็นค่าต่อเนื่องก็อาจทำให้ประสิทธิภาพของโมเดลนี้ลดลงได้

2.1.3.2 การถดถอยแบบโลจิสติกส์ (Logistics Regression)

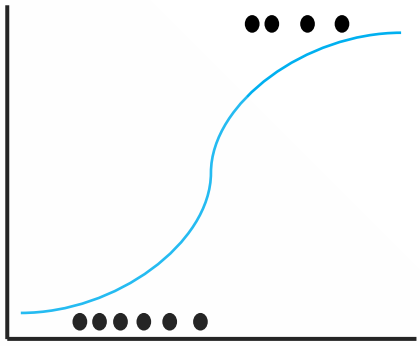
เป็นโมเดลทางสถิติศาสตร์ที่สามารถคำนวนความน่าจะเป็นของกลุ่มข้อมูลของตัวแปรตามที่มีลักษณะเป็นกลุ่มข้อมูลจำนวนสองกลุ่ม (Binary) เช่น ใช่หรือไม่ใช่ ทำหรือไม่ทำ โมเดลมีสมมติฐานว่าความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามเป็นแบบเส้นตรง โมเดลจะประมาณการจากตัวแปรต้นทั้งหมดในชุดข้อมูลเพื่อคำนวณหาค่าสัมประสิทธิ์ (Coefficient) ของตัวแปรต้นแต่ละตัวด้วยสมการกำลังสองน้อยที่สุด (Least Squares) แล้วจึงนำค่าที่ได้ไปใช้ในการแบ่งข้อมูลออกจากกันด้วยสมการเส้นตรงที่ถูกแปลงให้มีลักษณะคล้ายคลึงกับตัวเอสที่เป็นพยัญชนะภาษาอังกฤษดังภาพที่ 2.4 โดยสมการของโมเดลสามารถเขียนได้ดังนี้

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

กำหนดให้ X คือค่าของตัวแปรต้น B คือค่าสัมประสิทธิ์ Ln คือลอกาลิทึมฐานธรรมชาติของความน่าจะเป็น P/1-P โดยที่ P คือ ค่าความน่าจะเป็นของเหตุการณ์ที่สนใจ (Logit Link Function) สมการข้างต้นสามารถปรับให้อยู่ในรูปของสมการ Sigmoid ด้วยการผกผัน เพราะผลลัพธ์สุดท้ายที่สนใจคือ y ซึ่งหมายถึงค่าความน่าจะเป็นที่มีค่าอยู่ระหว่าง 0 ถึง 1 โดยสมการสุดท้ายสามารถเขียนได้ดังนี้

$$y = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

สมการและค่าสัมประสิทธิ์ที่คำนวนได้จากชุดข้อมูลฝึกจะถูกบันทึกและนำมาใช้ในการประมาณค่าผลลัพธ์ของข้อมูลใหม่ต่อไป



ภาพที่ 2.4 การจำแนกข้อมูลด้วยสมการโลจิสติกส์

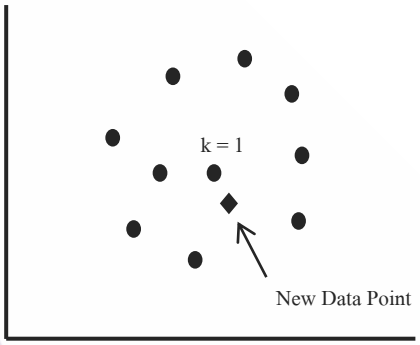
โมเดลนี้สามารถทำงานได้รวดเร็วและมีประสิทธิภาพความแม่นยำค่อนข้างสูง สามารถนำสมการผลลัพธ์ที่ได้ไปใช้ในการเลือกตัวแปรที่สำคัญได้ และยังทนทานต่อข้อมูลที่แปลกแยก (Outlier) ในระดับหนึ่ง แต่ข้อเสียคือไม่สามารถทำงานได้ดีหากตัวแปรต้นมีความสัมพันธ์เชิงเส้นตรงระหว่างกันเองสูงและจำนวนมากเกินไป (Multicollinearity)

2.1.3.3 เพื่อนบ้านที่ใกล้ที่สุด (k-Nearest Neighbor)

เป็นโมเดลที่นำลักษณะความคล้ายคลึงของข้อมูลที่ใช้ฝึกมาใช้ตัดสินใจเพื่อจำแนกข้อมูลใหม่ โมเดลจะพิจารณาว่าข้อมูลใหม่นั้นมีตำแหน่งอยู่ใกล้เคียงกับข้อมูลชุดฝึกตัวใดเป็นจำนวนเท่าใด แล้วจึงตัดสินใจว่าข้อมูลที่ใหม่นั้นเป็นกลุ่มเดียวกับข้อมูลที่ใกล้เคียงที่สุด โดยการวัดระยะห่างระหว่างข้อมูลนั้น โมเดลจะใช้หลักการวัดระยะห่างบนปริภูมิยูคลิด เพื่อคำนวณระยะห่างจากข้อมูลที่สนใจไปยังข้อมูลทั้งหมด โดยกำหนดค่า k คือจำนวนข้อมูลที่ใกล้ที่สุดที่พิจารณา วิธีการวัดระยะห่างสามารถเขียนเป็นสมการได้ดังนี้

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

กำหนดให้ d คือ ฟังก์ชั่นในการหาระยะห่างระหว่างข้อมูล (Distance) โดยที่ p คือค่าของตัวแปรใด ๆ ในชุดข้อมูลฝึก และ q คือค่าของตัวแปรใด ๆ ในชุดข้อมูลใหม่ที่ต้องการวัดระยะ



ภาพที่ 2.5 การจำแนกด้วยเพื่อนบ้านที่ใกล้ที่สุด ที่ค่า k = 1

ข้อดีของโมเดลนี้คือสามารถจำแนกข้อมูลได้ตามความคล้ายคลึง อย่างไรก็ดีจุดอ่อนหนึ่งของโมเดลนี้มีความแตกต่างกันออกไปขึ้นอยู่กับข้อจำกัดของหลักการวัดระยะห่าง สำหรับการวัดระยะแบบยูคลิดนั้นมีข้อสังเกตหนึ่งคือวิธีการนี้ไม่ได้นำความสัมพันธ์ระหว่างตัวแปรมาพิจารณา การที่มีตัวแปรที่ไม่เกี่ยวข้องมากเกินไปอาจทำให้โมเดลเกิดความผิดพลาดให้การตัดสินใจได้ในบางกรณี รวมถึงการที่ข้อมูลมีค่าที่แปลกแยก (Outlier) หรือหากมีข้อมูลสูญหาย (Missing Value) ก็จะส่งผลต่อประสิทธิภาพในการเทียบเคียงความคล้ายคลึงได้ และมากไปกว่านั้นหากข้อมูลที่ขนาดใหญ่จะส่งผลให้โมเดลทำงานได้ช้า

2.1.3.4 ความน่าจะเป็นแบบเบย์ (Naïve Bayes)

โมเดลนี้ได้นำหลักการความน่าจะเป็นมาประยุกต์ใช้เพื่อเรียนรู้ลักษณะของชุดข้อมูลฝึกเพื่อนำไปใช้จำแนกข้อมูลใหม่ โดยทฤษฎีของเบย์ (Bayes Theorem) สามารถเขียนเป็นสมการได้ดังนี้

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

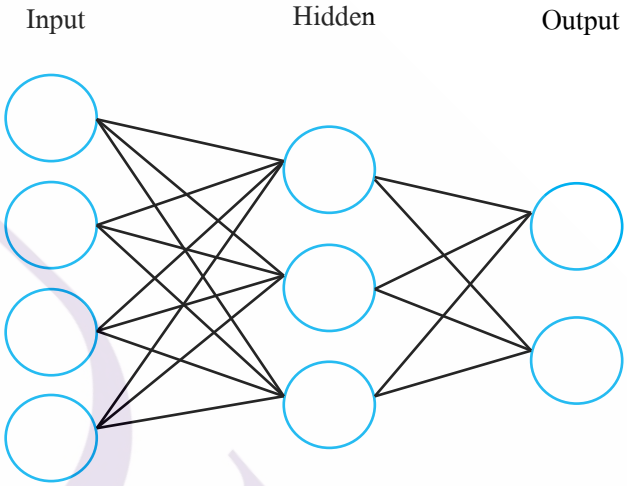
กำหนดให้ c คือป้ายกำกับของข้อมูล และ x คือ ตัวแปรต้นของข้อมูล ซึ่งจากสมการข้างต้นสามารถอธิบายได้ว่า ป้ายกำกับ c สามารถจำแนกได้ตามความน่าจะเป็นที่ตัวแปร x มีป้ายกำกับ c อยู่ในชุดข้อมูลฝึก ซึ่ง P(C|X) แสดงถึงข้อมูลที่มีตัวแปร x จะมีป้ายกำกับ c (Posterior Probability) โดยที่สามารถคำนวณได้จาก P(A|C) ที่หมายถึง ค่าความน่าจะเป็นที่ชุดข้อมูลฝึกที่มีป้ายกำกับ c และมีตัวแปร x โดยที่พิจารณาจากทุกตัวแปรในชุดข้อมูลฝึกรวมกัน (Likelihood) หารด้วย P(x) ซึ่งหมายถึงค่าความน่าจะเป็นของป้ายกำกับ c (Prior Probability) เนื่องจากสมมติฐานของโมเดลนี้ได้เชื่อว่าตัวแปรต้นทุกตัวนั้นเป็นอิสระต่อกันจึงทำให้สามารถเขียนสมการได้ดังนี้

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

ในบางครั้งป้ายกำกับ c อาจไม่พบได้ในตัวแปร x จะส่งผลให้ความน่าจะเป็นมีค่าเป็น 0 ซึ่งส่งผลให้ผลการจำแนกของป้ายกำกับ c นั้น ๆ มีค่าเป็น 0 ตามลำดับ (Zero-Frequency Problem) ดังนั้นการแก้ไขสามารถทำได้โดยการเพิ่มค่าความถี่จำนวน 1 เข้าไปยัง x ทุกตัวของป้ายกำกับ c นั้น (Laplace Smoothing) นอกเหนือจากนี้ ถ้าหากข้อมูลตัวแปรที่ใช้มีลักษณะเป็นค่าต่อเนื่อง โมเดลจะนำค่าเหล่านั้นมาทำการแบ่งช่วง (Binning) ต่างลักษณะการกระจายตัวของข้อมูลเพื่อที่จะได้นำไปนับเป็นความถี่ต่อไป

อย่างไรก็ดี เนื่องจากโมเดลนี้ได้อ้างอิงลักษณะของการกระจายตัวของข้อมูลฝึกทั้งหมดเพื่อนำมาใช้ในการจำแนกข้อมูลชุดใหม่ ดังนั้นโมเดลจะมีความอ่อนไหวต่อการเปลี่ยนแปลงของการกระจายตัวของข้อมูลชุดใหม่ ซึ่งจะส่งผลให้ประสิทธิภาพการทำงานของโมเดลนี้มีความเสี่ยงที่จะเปลี่ยนแปลงไปในทางที่แย่ลงได้อย่างง่ายดาย รวมถึงการที่โมเดลทำการแบ่งช่วงตัวแปรต่อเนื่องในขณะฝึกก็จะทำให้เกิดการสูญเสียข้อมูล (Information Loss) ได้ และมากไปกว่านั้น หากตัวอย่างของข้อมูลฝึกมีจำนวนไม่สมดุลกันก็จะส่งผลต่อประสิทธิภาพความถูกต้องในการจำแนกอีกด้วย

2.1.3.5 โครงข่ายประสาทเทียม (Neural Network)



ภาพที่ 2.5 การจำแนกด้วยเพื่อนบ้านที่ใกล้ที่สุด ที่ค่า k = 1

โมเดลนี้ได้นำแนวคิดการทำงานของสมองมนุษย์มาประยุกต์ใช้เป็นวิธีการในการเรียนรู้ องค์ประกอบหลักของโมเดลนั้นแบ่งออกได้เป็น 3 ส่วนดังนี้

1. ลำดับชั้นของเซลล์ประสาท (Layer) โมเดลนี้ประกอบไปด้วย 3 ชั้น คือ ชั้นนำเข้าข้อมูล (Input Layer) ชั้นซ่อนเร้น (Hidden Layer) และ ชั้นนำข้อมูลออก (Output Layer) ตามลำดับ

2. เซลล์ประสาท (Node) โดยในแต่ละชั้นจะประกอบด้วยโหนดที่ทำหน้าที่เหมือนเซลล์ประสาท ซึ่งในแต่ละโหนดจะประกอบไปด้วยค่าตัวเลขของข้อมูลที่ได้จากการคำนวนก่อนหน้า โหนดที่อยู่ในชั้นต่าง ๆ จะเรียกว่าโหนดข้อมูลนำเข้า โหนดข้อมูลซ่อนเร้น และโหนดนำข้อมูลออก ตามลำดับ

3. ค่าน้ำหนัก (Weight) เป็นค่าที่ได้จากการคำนวณด้วยฟังก์ชั่นภายในโหนดแต่ละชั้น ซึ่งเป็นทั้งผลลัพธ์ของโหนดก่อนหน้า และเป็นข้อมูลตั้งต้นของโหนดถัดไปโดยค่าน้ำหนักจะมีการปรับเปลี่ยนทุกครั้งโดยอ้างอิงจากอัตราการเรียนรู้ (Learning rate) เพื่อลดความคลาดเคลื่อนในการ

(18)

คำนวณ การฝึกของโมเดลจะเริ่มจากการที่รับข้อมูลตั้งต้นเข้ามายังชั้นนำเข้าข้อมูลโดยที่จำนวนเซลล์ประสาทในชั้นนี้จะมีจำนวนเท่ากับจำนวนตัวแปรต้นของข้อมูล จากนั้นข้อมูลเหล่านี้จะถูกนำมาคำนวณด้วยสมการข้างล่างนี้

$$\underset{j}{net} = \sum_{i} w_{ij} * X_i$$

โดยกำหนดให้ net คือ ผลลัพธ์ของการคำนวณจากชั้นประสาทก่อนหน้า ซึ่งได้มาจากผลรวมของการนำค่าของตัวแปร (X) ที่ i มาคูณกับค่าน้ำหนัก (w) ที่จับคู่โยง i,j มายัง net ที่ j บวกด้วยค่าความเอนเอียง (Bias) จากนั้นผลลัพธ์ที่ได้จะถูกนำไปเข้าสมการที่ใช้ในแสดงผลว่าเซลล์ประสาทนั้น ๆ มีลักษณะเป็นอย่างไร (Activate Function) ซึ่งในปกติแล้วเซลล์ประสาทของมนุษย์บางเซลล์จะส่งคลื่นสัญญาณที่ชัดเจนเมื่อมีความสัมพันธ์กับการกระทำใด ๆ ที่กำลังเกิดขึ้นในสมอง การแสดงลักษณะดังกล่าวสามารถเลียนแบบให้อยู่ในรูปของฟังก์ชั่นต่าง ๆ ได้ เช่น Sigmoid โดยสามารถเขียนเป็นสมการดังนี้

$$S(x) = \frac{1}{1 + e^{-x}}$$

โดยที่ x หมายถึงผลลัพธ์ที่ได้มาจากสมการก่อนหน้า ซึ่งการกระทำนี้จะทำต่อเนื่องไปเรื่อย ๆ จนถึง ชั้นสุดท้ายที่ใช้นำข้อมูลออก (Feed forward) แต่อย่างไรก็ดี ผลลัพธ์จากขั้นตอนที่กล่าวมาข้างต้นอาจจะมีความคลาดเคลื่อนอยู่ กล่าวคือโมเดลยังไม่สามารถจำแนกข้อมูลได้แม่นยำนัก ดังนั้นจึงได้มีการใช้วิธีการปรับค่าความคลาดเคลื่อน (Back propagation) ทันทีหลังจากที่ได้ผลลัพธ์ในชั้นสุดท้าย การกระทำนี้จะช่วยให้โมเดลสามารถปรับค่าน้ำหนักและค่าความเอนเอียงแบบย้อนกลับได้โดยใช้ฟังก์ชั่น Sum-Square Error ซึ่งสามารถเขียนเป็นสมการดังนี้

$$E = \frac{1}{2}\sum_{k}(t_k - a_k)^2$$

กำหนดให้ E คือค่าความคลาดเคลื่อนของการทำงานในรอบใด ๆ (Epoch) ที่คำนวณจากค่าคำตอบระหว่างความเป็นจริงกับสิ่งที่โมเดลทายออกมา จากนั้นจึงจะได้ค่าความคลาดเคลื่อนมาใช้ใน

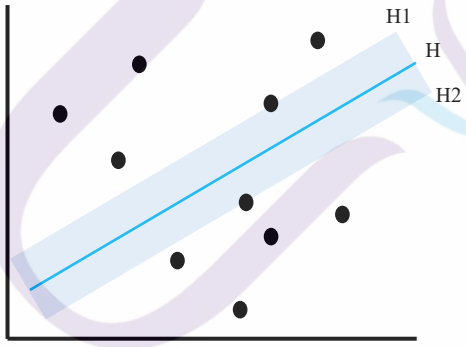
การหาว่าควรจะปรับค่าน้ำหนักที่โยงระหว่างชั้นเซลล์ประสาทก่อนหน้าเท่าใดจึงจะลดค่าความคลาดเคลื่อนได้ด้วยการหาอนุพันธ์ย้อนกลับ

$$\Delta w_{kj} \propto -\frac{\partial E}{\partial w_{kj}}$$

ข้อดีของโมเดลนี้คือมีความแม่นยำสูง สามารถเรียนรู้ข้อมูลได้อย่างละเอียด แต่อย่างไรก็ดีเนื่องจากการเรียนรู้ได้อย่างละเอียดนั้นจะทำให้มีโอกาสเกิดภาวะการเรียนรู้มากเกินไปได้ง่าย มากไปกว่านั้นโมเดลนี้มีความซับซ้อนสูงจึงส่งผลให้การแปรผลจากการทำงานทำได้ยากตามไปด้วย การออกแบบโครงสร้างเครือข่ายประสาทและการกำหนดพารามิเตอร์ต่าง ๆ จึงเป็นสิ่งที่สำคัญเพราะส่งผลต่อความสามารถของโมเดลโดยตรง

2.1.3.6 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

เป็นโมเดลที่ใช้ในการจำแนกข้อมูลที่มีการเพิ่มมิติปริภูมิให้มากขึ้นเพื่อที่จะสามารถนำสมการเชิงเส้น (Hyperplane) มาใช้ในการสร้างขอบเขตเพื่อแบ่งข้อมูลที่ป้ายกำกับต่างกันออกจากกันได้ง่ายขึ้น



ภาพที่ 2.7 การจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์

เส้นแบ่งจะประกอบไปด้วย 3 เส้นขนาน คือ H, H1, H2 โดยที่ H จะเป็นเส้นที่ขั้นกลางระหว่าง H1 และ H2 โดยระยะห่างระหว่าง H และ H1 จะมีค่าตั้งแต่ 0 ถึง 1 และระยะห่างระหว่าง H

และ H2 จะมีค่าตั้งแต่ -1 ไปถึง 0 โดยระยะห่างเหล่านี้ (Margin) จะถูกนำมาใช้พิจารณาในการแบ่งแยกข้อมูล เวกเตอร์สนับสนุน (Support Vector) จะตั้งอยู่บนเส้น H1 และ H2 และระยะห่างจาก H1 และ H2 ไปยัง H นั้นควรจะมีความยาวมากที่มากสุดเท่าที่จะเป็นไปได้ โมเดลจะสร้างเส้นแบ่งข้อมูลโดยอ้างอิงจากข้อมูลทุกจุดบนปริภูมิและทำงานวนซ้ำจนกว่าจะจนพบเส้นแบ่งที่ดีที่สุดภายใต้เงื่อนไขที่กำหนด ทั้งนี้การนำโมเดลมาใช้งานกับข้อมูลข้อมูลที่ตัวแปรที่สัมพันธ์กันแบบไม่เป็นเส้นตรงบนปริภูมิตั้งต้นสามารถทำได้ด้วยการสร้างมิติใหม่ขึ้นเรื่อย ๆ จนกว่าจะสามารถใช้สมการเชิงเส้นแบ่งแยกข้อมูลได้ (Kernel Trick) ซึ่งหนึ่งในวิธีการที่มีประสิทธิภาพความแม่นยำสูงและนิยมใช้มากที่สุดคือ Radial Basis Function (RBF)

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

$$f(\mathbf{x}) = \sum_{i}^{N} \alpha_i y_i \exp\left(-||\mathbf{x} - \mathbf{x}_i||^2/2\sigma^2\right) + b$$

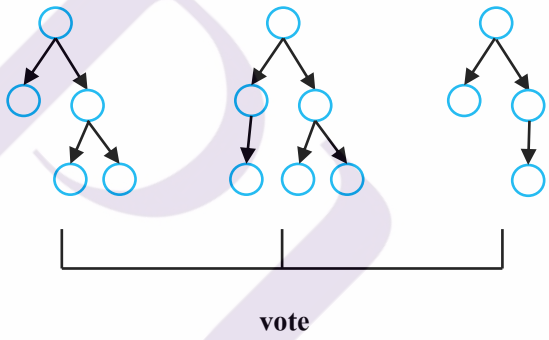
ข้อดีของโมเดลนี้คือมีความแม่นยำในการจำแนกข้อมูลสูงเพราะสามารถมองหารูปแบบที่ใช้ในการเรียนรู้ได้ในปริภูมิหลายมิติ แต่ทั้งนี้อาจจะส่งผลให้เกิดการเรียนรู้มากเกินไป (Overfitting) การฝึกโมเดลใช้เวลานานมากเนื่องจากต้องมีการคำนวณหาเส้นแบ่งข้อมูลที่ดีที่สุดโดยเทียบกับข้อมูลทุกจุดในทุกปริภูมิ รวมถึงการเลือกหาพารามิเตอร์ที่เหมาะสมก็ส่งผลทำให้ระยะเวลาที่ต้องใช้ฝึกเพิ่มขึ้นจากปกติอีกด้วย การดำเนินการสร้างปริภูมิแต่ละมิตินั้นจะถูกพักไว้บนหน่วยความจำชั่วคราว (RAM) ดังนั้นหากข้อมูลมีปริมาณมากจะทำให้โมเดลต้องการพื้นที่หน่วยความจำมากขึ้นตามลำดับ

2.1.3.7 การเรียนรู้แบบกลุ่ม (Ensemble)

การเรียนรู้แบบเป็นกลุ่มคือการฝึกโมเดลมากกว่าหนึ่งตัวแล้วนำมาใช้ในการทำงานร่วมกันเพื่อค้นหาผลลัพธ์สุดท้าย

2.1.3.7.1 ป่าไม้ตัดสินใจ (Random Forest)

คือการนำโมเดลต้นไม้ตัดสินใจมาใช้งานจำนวนหลายต้น ต้นไม้ทุกต้นจะถูกสร้างขึ้นมาจากชุดข้อมูลฝึกเดียวกัน ต้นไม้แต่ละต้นจะมีตัวแปรที่ใช้ต่างกันแต่มีจำนวนเท่ากัน โดยการเลือกตัวแปรและตัวอย่างจากชุดข้อมูลฝึกนั้นจะถูกหยิบออกมาด้วยวิธีการสุ่มแบบแทนที่ (Bagging) เพื่อนำมาสร้างเป็นต้นไม้ตัดสินใจหลายต้นแล้วจึงนำไปใช้ในการจำแนกข้อมูลใหม่ โดยที่ผลการแยะแยกของต้นไม้ทุกต้นจะถูกนำมาพิจารณาร่วมกันเพื่อตัดสินใจคำตอบที่ดีที่สุดด้วยการเลือกจากเสียงข้างมาก (Majority Vote)



ภาพที่ 2.8 ลักษณะของป่าไม้ตัดสินใจที่มีจำนวนต้นไม้ 3 ต้น

ข้อดีของป่าไม้ตันสินใจที่เด่นชัดคือโมเดลที่ได้จากการฝึกจะรู้จักข้อมูลชุดฝึกหลาย ๆ ส่วนเนื่องมาจากผลการสุ่มหยิบแบบแทนที่ ซึ่งทำให้โมเดลสามารถเป็นตัวแทนข้อมูลชุดฝึก (Representation) ได้ดี จึงทำให้มีประสิทธิภาพความแม่นยำสูง อย่างไรก็ดี โมเดลนี้สามารถเกิดภาวะเรียนรู้มากเกินไป (Overfitting) ได้ง่ายหากต้นไม้ตัดสินใจมีความลึก (Depth) ที่ไม่เหมาะสมกับชุด

ข้อมูล และเนื่องจากโมเดลนี้ได้นำต้นไม้ตัดสินใจมาใช้ร่วมกันหลายต้น ทำให้การแปรผลจากการทำงานเป็นไปได้ยาก การใช้โมเดลนี้กับข้อมูลขนาดเล็กอาจจะให้ผลลัพธ์ประสิทธิภาพไม่ดีเท่าใดนัก

2.1.3.7.2 ป่าไม้ตัดสินใจที่มีการลดค่าความคลาดเคลื่อน (Extreme Gradient Boosting Tree)

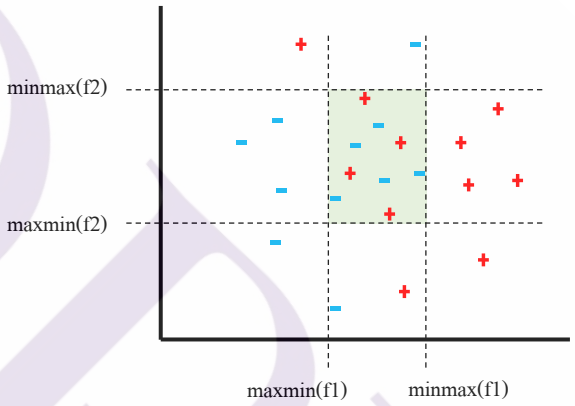
เป็นการนำต้นไม้ตัดสินใจมาเรียนรู้ข้อมูลชุดฝึก โดยในการเรียนรู้นั้นโมเดลจะสุ่มหยิบข้อมูลจำนวนหนึ่งออกมาจากชุดฝึกด้วยวิธีการสุ่มแบบแทนที่ (Bagging) เพื่อนำไปสร้างต้นไม้ที่มีความลึกไม่มากขึ้นมา (Shallow Tree) หลาย ๆ ต้น แล้วจึงจะได้เป็นโมเดลมีประสิทธิภาพต่ำ (Weak Learner) หลาย ๆ โมเดล แต่ในขณะเดียวกันโมเดลได้มีการนำฟังก์ชั่นที่ใช้ในการลดความผิดพลาดมาใช้ (Cost Function) โดยจะมีการกำหนดค่าน้ำหนัก (Weight) ให้กับข้อมูลตัวอย่างที่โมเดลจำแนกผิดพลาดให้มีน้ำหนักมากขึ้น หลังจากนั้นจะวนไปสร้างต้นไม้ตัดสินใจใหม่อีกครั้งไปเรื่อย ๆ จนกว่าค่าความผิดพลาดจะลดลงจนนิ่ง (Coverage) ซึ่งในการปรับค่าน้ำหนักนั้น โมเดลได้นำวิธีการค้นหาค่าต่ำสุดด้วยความลาดชัน (Gradient Descent) มาใช้งาน

ข้อดีที่ชัดเจนข้อโมเดลนี้คือได้มีการแก้ปัญหาความอคติ (Bias) ด้วยการปรับค่าน้ำหนัก (Boosting) และความแปรปรวน (Variance) ด้วยการสุ่มหยิบข้อมูลด้วยวิธีการสุ่มแบบแทนที่ (Bagging) ซึ่งส่งผลให้โมเดลมีประสิทธิภาพความแม่นยำสูง อย่างไรก็ดี โมเดลนี้มีความซับซ้อนสูงจึงจำเป็นต้องใช้พลังในการประมวลผลมาก ดังนั้นการฝึกโมเดลจะไม่เป็นไปอย่างรวดเร็ว และการแปรผลการทำงานทำได้ยาก ตลอดจนสามารถเกิดภาวะการเรียนรู้มากเกินไปได้ง่าย หากไม่ใช้พารามิเตอร์ที่เหมาะสม

2.1.4 ความซับซ้อนของข้อมูล (Data Complexity)

การตรวจสอบความซับซ้อนของข้อมูลมีความจำเป็นอย่างยิ่งก่อนที่จะส่งข้อมูลให้เครื่องวิเคราะห์และประมวลผลอย่างเต็มรูปแบบ เนื่องจากข้อมูลที่จะนำไปใช้อาจจะมีความผิดปกติแอบแฝงอยู่ เช่น ความทับซ้อนของข้อมูล ซึ่งปัญหานี้จะส่งผลต่อประสิทธิภาพและความแม่นยำของโมเดล วิธีการวัดระดับความซับซ้อนได้ถูกนำเสนอไว้ในงานวิจัยของ A. C. Lorena, et el. [26] ในปี 2018

2.1.4.1 ค่าของขนาดของพื้นที่ทับซ้อน (Volume of Overlapping Region – F2)



ภาพที่ 2.9 ลักษณะการพิจารณาข้อมูลทับซ้อน F2

สามารถคำนวณได้จากพื้นที่ที่ทับซ้อนของแต่ละตัวแปรภายในลาเบลทั้งหมดดังสมการข้างล่างนี้

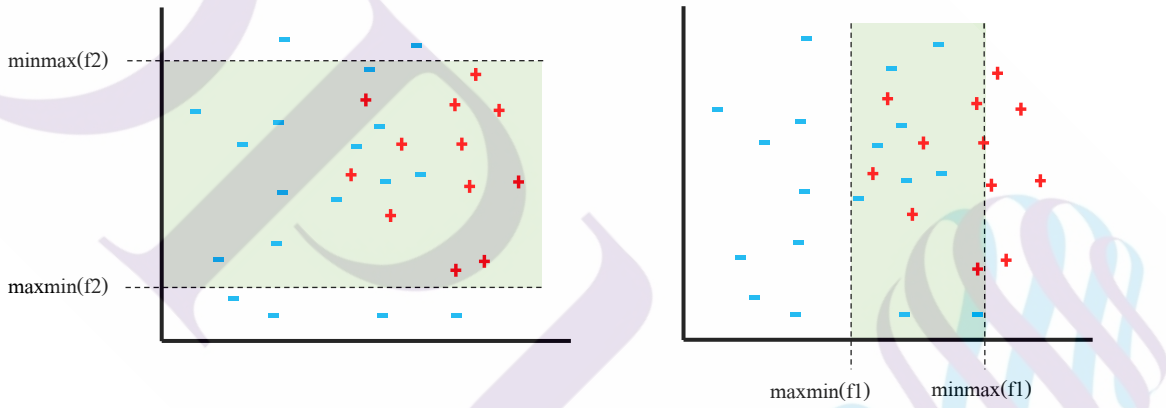
$$F2 = \prod_{i}^{m} \frac{overlap(f_i)}{range(f_i)} = \prod_{i}^{m} \frac{\max\{0, \min\max(f_i) - \max\min(f_i)\}}{\max\max(f_i) - \min\min(f_i)},$$

โดยที่ขอบเขตของพื้นที่ได้มาจากสมการดังข้างล่างนี้

$$\begin{aligned}
\min\max(f_i) &= \min(\max(f_i^{c_1}), \max(f_i^{c_2})),\\
\max\min(f_i) &= \max(\min(f_i^{c_1}), \min(f_i^{c_2})),\\
\max\max(f_i) &= \max(\max(f_i^{c_1}), \max(f_i^{c_2})),\\
\min\min(f_i) &= \min(\min(f_i^{c_1}), \min(f_i^{c_2})).
\end{aligned}$$

กำหนดให้ C คือลาเบลคำตอบ และ f_i คือ ตัวแปรใด ๆ ซึ่งหากผลลัพธ์ที่ได้มีค่าสูง หมายความว่าข้อมูลนั้น ๆ มีการทับซ้อนแฝงอยู่ในระดับที่สูง โดยผลลัพธ์จะมีค่าอยู่ในช่วง 0 ถึง 1 เท่านั้น อย่างไรก็ดีจำนวนของตัวแปรที่ใช้สามารถส่งผลต่อค่าของขนาดของพื้นที่ทับซ้อนนี้ได้

2.1.4.2 ค่าความมีประสิทธิภาพของตัวแปรแต่ละตัว (Maximum Individual Feature Efficiency – F3)



ภาพที่ 2.10 ลักษณะการพิจารณาข้อมูลทับซ้อน F3

สามารถนำมาใช้ประมาณค่าประสิทธิภาพของแต่ละตัวแปรว่าสามารถแบ่งแยกลาเบลภายในชุดของมูลได้ดีมากน้อยเพียงใด สามารถเขียนเป็นสมการ

$$F3 = \max_{i=1}^{m} \frac{n - n_o(f_i)}{n},$$

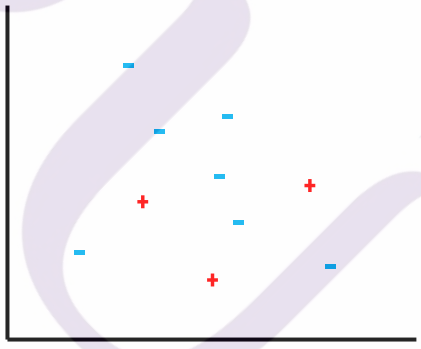
โดยกำหนดให้$n_o(f_i)$ คือ จำนวนตัวอย่างของตัวแปร fi ใด ๆ ที่อยู่ในพื้นที่ทับซ้อน ซึ่งวิธีการคำนวณหาพื้นที่ทับซ้อนทำได้ด้วยสมการ

$$n_o(f_i) = \sum_{j=1}^{n} I(x_{ji} > \max\min(f_i) \wedge x_{ji} < \min\max(f_i))$$

กำหนดให้ I หมายถึงตัวชี้วัดว่ามีการทับซ้อนหรือไม่ (0 หรือ 1) วิธีการนี้จะคำนวณประสิทธิภาพของตัวแปรด้วยการพิจารณาจากสัดส่วนระหว่างจำนวนตัวอย่างข้อมูลที่ไม่อยู่ในพื้นที่ทับซ้อนและจำนวนตัวอย่างทั้งหมด ผลลัพธ์ที่ได้จะมีค่าอยู่ในช่วงระหว่าง 0 ถึง 1 โดยที่ค่ายิ่งต่ำหมายถึงมีการทับซ้อนของข้อมูลอยู่ในระดับสูง

2.1.5 การปรับความสมดุลข้อมูล (Re-balancing)

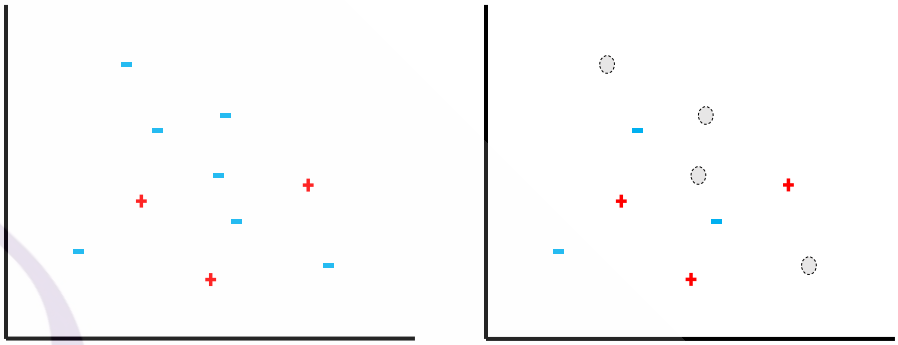
เป็นการปรับเปลี่ยนสัดส่วนของข้อมูลตัวอย่างให้มีความใกล้เคียงกัน หากตัวอย่างกลุ่มใดมีจำนวนน้อยกว่าจะถูกเพิ่มปริมาณขึ้น หากตัวอย่างกลุ่มใดมีจำนวนมากกว่าจะถูกลดปริมาณลง อย่างไรก็ดีการเพิ่มหรือลดข้อมูลขึ้นอยู่กับลักษณะของข้อมูลที่นำมาใช้ และผลลัพธ์การปรับเพิ่มและลดข้อมูลขึ้นอยู่กับโมเดลที่ใช้ในการปรับสมดุลข้อมูล



ภาพที่ 2.11 ลักษณะข้อมูลที่ไม่สมดุล

2.1.6 โมเดลที่ใช้ในการปรับสมดุลข้อมูล (Re-balancing Algorithm)

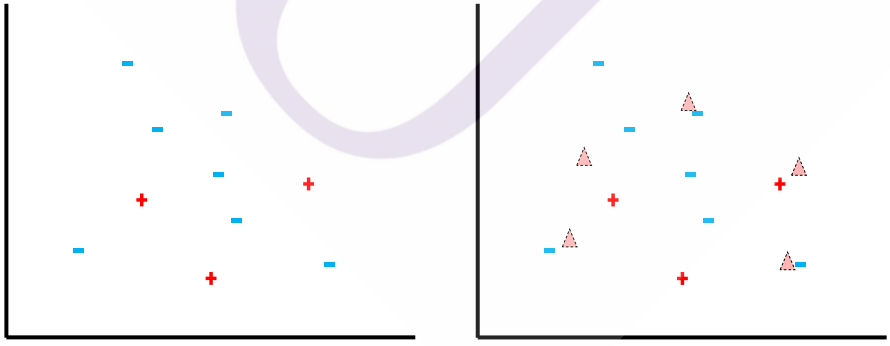
2.1.6.1 การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม (Random Under-Sampling)



ภาพที่ 2.12 ลักษณะการลดขนาดข้อมูลกลุ่มที่มีจำนวนมากกว่า

วิธีการลดตัวอย่างข้อมูลโดยการสุ่มลบออก (วงกลม) โดยจะทำการเลือกเฉพาะข้อมูล Negative เพื่อลดปริมาณตัวอย่างให้ลงมาเท่า ๆ กับปริมาณตัวอย่าง Positive ข้อดีของวิธีการนี้คือสามารถช่วยลดขนาดข้อมูล Negative ซึ่งเป็นการทำงานที่ดีหากข้อมูลมีปริมาณมากและมีลักษณะคล้ายกัน แต่ข้อเสียคือบางครั้งการลดข้อมูลด้วยการสุ่มอาจทำให้ข้อมูลที่สำคัญหายไป

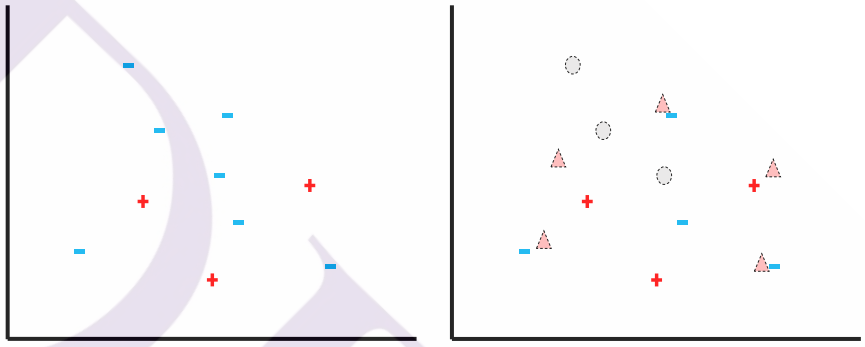
2.1.6.2 การเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่ม (Random Over-Sampling)



ภาพที่ 2.13 ลักษณะการเพิ่มขนาดข้อมูลกลุ่มที่มีจำนวนน้อยกว่า

วิธีการเพิ่มตัวอย่างข้อมูลโดยการสุ่มสร้างตัวอย่างข้อมูลขึ้นมาใหม่ (สามเหลี่ยม) โดยพิจารณาเพียงข้อมูลกลุ่มที่มีจำนวนตัวอย่างน้อยกว่า แล้วจึงสุ่มสร้างตัวอย่างของข้อมูลกลุ่มนั้น ข้อดีคือสามารถเพิ่มปริมาณข้อมูล Positive ให้มากขึ้นได้ แต่ข้อเสียคือเกิดเพิ่มข้อมูลลักษณะนี้อาจจะทำให้เกิดปัญหาข้อมูลถูกรบกวนได้ง่าย

2.1.6.3 การเพิ่มและลดจำนวนตัวอย่างข้อมูลแบบสุ่ม (Random Over and Under-Sampling)



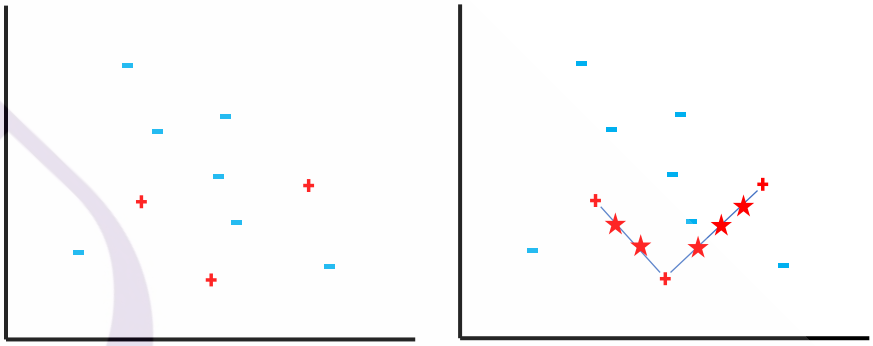
ภาพที่ 2.14 ลักษณะการเพิ่มขนาดข้อมูลกลุ่มที่มีจำนวนน้อยกว่าและลดขนาดข้อมูลกลุ่มที่มีจำนวนมากกว่า

วิธีการเพิ่มตัวอย่างข้อมูลโดยการสุ่มสร้างตัวอย่างข้อมูลขึ้นมาใหม่แล้วจึงลดข้อมูล โดยเป็นการผสมระหว่างการลดจำนวนตัวอย่างข้อมูลแบบสุ่มและการเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่ม โดยข้อดีและข้อเสียของวิธีการนี้เหมือนกับวิธีการเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่ม และวิธีการลดจำนวนตัวอย่างข้อมูลแบบสุ่ม ในขณะที่วิธีการนี้ข้อมูลจะผิดเพี้ยนน้อยกว่า

2.1.6.4 การสุ่มสร้างข้อมูลโดยอ้างอิงจากเพื่อนบ้านที่มีลักษณะคล้ายคลึง (Synthetic Minority Over-Sampling Technique)

ในปี 2002 N. V. Chawla et el. [24] ได้นำเสนอวิธีการเพิ่มตัวอย่างข้อมูลโดยการสุ่มสร้างตัวอย่างข้อมูลขึ้นมาใหม่ด้วยการนำตัวอย่างข้อมูลจากกลุ่มตัวอย่างที่มีจำนวนน้อยกว่ามาพิจารณาทีละตัวจนครบทุกตัว หลักการคือกำหนดจำนวนเพื่อนบ้านที่ใกล้ที่สุดจำนวน k ตัว แล้วทำการสุ่มสร้าง

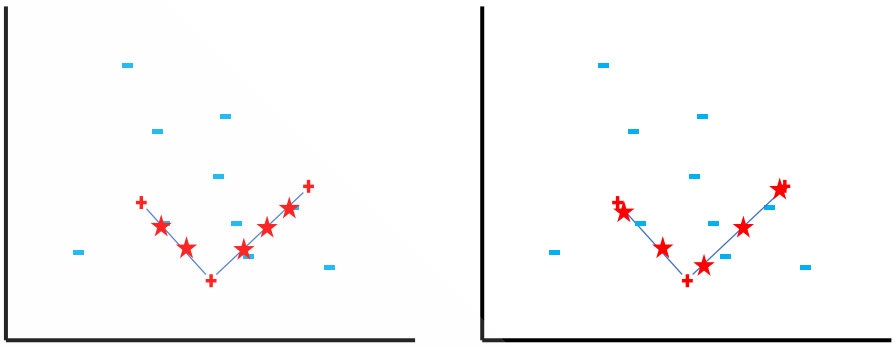
ข้อมูลขึ้นมาใหม่ในพื้นที่ใด ๆ บนทางที่เชื่อมโยงระหว่างจุดข้อมูลที่กำลังพิจารณาและจุดของข้อมูลเพื่อนบ้านที่ใกล้ที่สุด โดยปริมาณของข้อมูลสุ่มสร้างขึ้นมานั้นสามารถกำหนดได้ว่าจะสร้างขึ้นมามากน้อยเพียงใด โดยวิธีการสามารถเขียนได้ดังนี้



ภาพที่ 2.15 ลักษณะการเพิ่มขนาดข้อมูลกลุ่มที่มีจำนวนน้อยกว่าด้วยวิธีการ SMOTE

ข้อดีของวิธีการนี้คือสามารถเพิ่มจำนวนข้อมูล Positive ได้โดยไม่รบกวนพื้นที่บนปริภูมิมากเกินไป ทำให้ข้อมูลที่สร้างขึ้นมาใหม่ (สัญลักษณ์ดาว) ไม่รบกวนข้อมูลทั้งหมด อย่างไรก็ดีหลักการนี้ไม่สามารถหลบหลบหลีกการสร้างข้อมูล Positive ใหม่โดยไม่ให้ทับซ้อนกับข้อมูล Negative ได้ในบางโอกาส ซึ่งอาจส่งผลให้โมเดลเกิดความผิดพลาดได้

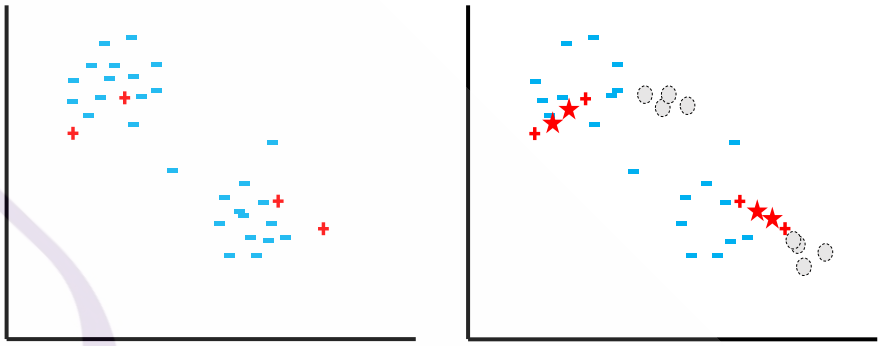
2.1.6.7 การสุ่มสร้างข้อมูลโดยอ้างอิงจากเพื่อนบ้านที่มีลักษณะคล้ายคลึงภายในพื้นที่ปลอดภัยที่สามารถย้ายที่ได้ (Relocate Safe-Level SMOTE)



ภาพที่ 2.16 ลักษณะการทำงานของวิธีการ RSLS

Relocate Safe-Level SMOTE (RSLS) ได้ถูกนำเสนอโดย W. Siriseriwan et el. [9] ในปี 2016 ซึ่งเป็นโมเดลที่พัฒนาต่อยอดมาจากวิธีการสุ่มสร้างข้อมูลโดยอ้างอิงจากเพื่อนบ้านที่มีลักษณะคล้ายคลึงภายในพื้นที่ปลอดภัย Safe-Level -SMOTE แต่โมเดลนี้จะสามารถเลื่อนการกำเนิดของของตัวอย่างที่จะสร้างขึ้นมาใหม่หากพบว่าตำแหน่งที่จะสร้างนั้นใกล้กับตัวอย่างของกลุ่มข้อมูลที่มากกว่าเพื่อลดการทับซ้อนของตำแหน่งข้อมูล และนอกเหนือจากนี้ วิธีการดังกล่าวยังพิจารณาว่าหากพบข้อมูล Positive ที่ไกลจากกลุ่มตัวเองมากเกินไปก็จะไม่สร้างข้อมูลใหม่บนระยะทางไปยังเพื่อนบ้านที่ใกล้ที่สุด เพราะจะทำให้ข้อมูลทั้งชุดอาจเกิดการรบกวนได้ ข้อดีคือวิธีการนี้ได้มีการหลีกเลี่ยงการทับซ้อนของข้อมูลระหว่างคลาส

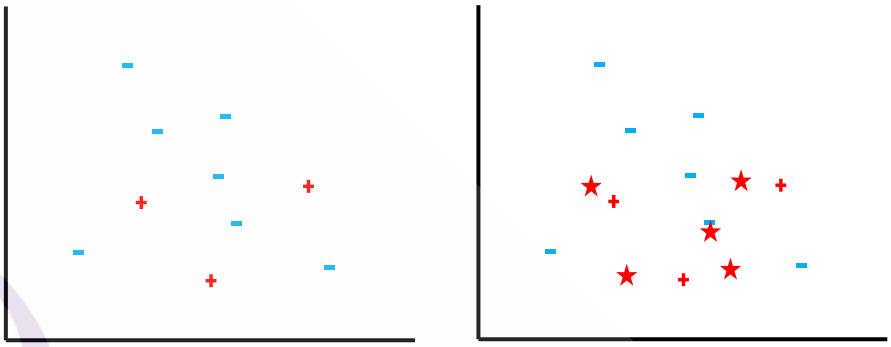
2.1.6.8 การสุ่มสร้างข้อมูลโดยอ้างอิงจากความหนาแน่นของข้อมูล (Density Based Synthetic Minority Over-Sampling Technique)



ภาพที่ 2.17 ลักษณะการทำงานของวิธีการ DBSM

ในปี 2016 Y. Sanguanmak et el. [8] ได้นำเสนอวิธีการแก้ไขข้อมูล Density Based Synthetic Minority Over-Sampling Technique (DBSM) ซึ่งวิธีการนี้ได้มีการนำวิธีการจัดกลุ่มข้อมูลที่อ้างอิงความหนาแน่น (DBSCAN) มาช่วยในการแบ่งข้อมูลออกเป็นหลายกลุ่ม เพื่อที่จะแบ่งแยกลักษณะของความหนาแน่นที่คล้ายกันไว้ด้วยกัน หลังจากที่แบ่งกลุ่มแล้ว การทำงานจะถูกแบ่งออกเป็นสองส่วน ในส่วนที่หนึ่งจะเป็นการลดปริมาณข้อมูล Negative ออก โดยการลบข้อมูลจะอ้างอิงจากข้อมูลที่อยู่ใกล้จุดศูนย์กลางของกลุ่ม (Centroid) โดยไม่สนใจข้อมูล Positive ภายใน จากนั้นจึงหยิบข้อมูล Negative ทั้งหมดจากทุกกลุ่มออกมาเป็นข้อมูลฝั่ง Negative ชุดใหม่ ในส่วนที่สองจะนำกลุ่มข้อมูลชุดเดิมที่มีข้อมูล Positive อยู่มาเพิ่มปริมาณด้วย SMOTE เมื่อดำเนินการครบทุกกลุ่มแล้ว ข้อมูล Positive ทั้งหมดของการทำงานในส่วนนี้จะถูกหยิบออกมาเป็นข้อมูล Positive ชุดใหม่ สุดท้ายข้อมูล Negative ชุดใหม่ และข้อมูล Positive ชุดใหม่จะถูกนำมารวมกันเป็นข้อมูลชุดใหม่ ข้อดีคือสามารถสร้างข้อมูล Positive และลบข้อมูล Negative โดยอ้างอิงจากความหนาแน่นภายในชุดข้อมูล แต่ทั้งนี้ปัญหาข้อมูลทับซ้อนยังคงไม่ได้ถูกแก้ไข

2.1.6.9 การสุ่มสร้างข้อมูลโดยอ้างอิงจากเพื่อนบ้านที่มีลักษณะคล้ายคลึง (Random Over-Sampling Examples)



ภาพที่ 2.18 ลักษณะการทำงานของวิธีการ ROSE

N. Lunardon et el. [17] ได้นำเสนอวิธีการ Random Over-Sampling Examples (ROSE) ในปี 2014 ซึ่งเป็นวิธีการเพิ่มตัวอย่างข้อมูลโดยการสุ่มสร้างตัวอย่างข้อมูลขึ้นมาใหม่ โดยนำตัวอย่างข้อมูลจากกลุ่มตัวอย่างที่มีจำนวนน้อยกว่ามาพิจารณาทีละตัวจนครบทุกตัว หลักการคือกำหนดจำนวนเพื่อนบ้านที่ใกล้ที่สุดจำนวน k ตัว จากนั้นทำการสุ่มสร้างข้อมูลขึ้นมาใหม่บนพื้นที่รอบ ๆ ตัวอย่างกลุ่มข้อมูลที่มีจำนวนน้อยกว่า ด้วยหลักการสุ่มหยิบแบบยอมรับการหยิบซ้ำ (Bootstrap) โดยมีการเติมเต็มข้อมูลเข้าไปในช่วงมากกว่าน้อยกว่าเพิ่มทำให้เกิดความเรียบเนียน (Smooth) ก่อนที่จะสร้างข้อมูล ดังนั้นข้อมูลที่ถูกสุ่มสร้างขึ้นมาใหม่จะมีค่าอยู่ภายในช่วงเดียวกับตัวอย่างที่กำลังพิจารณา

ภาพที่ 2.19 ลักษณะการทำงานของวิธีการ Kernel Density Bootstrapping

วิธีการดังกล่าวสามารถช่วยให้เพิ่มข้อมูลขึ้นมาได้โดยอ้างอิงไปตามการกระจายตัวของข้อมูล หากข้อมูลมีลักษณะที่เบ้และกระจุกตัวในพื้นที่ใกล้กัน อาจจะทำให้ข้อมูล Positive ที่สร้างขึ้นมาใหม่มีโอกาสทับซ้อนกับข้อมูล Negative ได้สูง กล่าวคือข้อมูลตั้งต้นจะถูกรบกวนมากเกินไป ซึ่งจะนำมาซึ่งประสิทธิภาพของโมเดลที่น้อยลง

ตารางที่ 2.2 การเปรียบเทียบความสามารถระหว่างวิธีปรับสมดุลข้อมูล

วิธีการ	เพิ่มจำนวน Positive	ลดจำนวน Negative	พิจารณาระยะทางระหว่างข้อมูล	พิจารณาระยะห่างPositive ที่ไกลจากกลุ่ม	พิจารณาความหนาแน่นของข้อมูล	พิจารณาการทับซ้อนของข้อมูล	สามารถปรับเปลี่ยนวิธีการแก้ไขข้อมูล Positive	สามารถปรับเปลี่ยนวิธีการแก้ไขข้อมูล Negative
ROS	✓							
RUS		✓						
OVUN	✓	✓						
SMOTE	✓		✓					
ROSE	✓	✓						
RSLS	✓		✓	✓		✓		
DBSM	✓	✓	✓		✓			
TOP	✓	✓	✓	✓	✓	✓	✓	✓

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวกับการเพิ่มประสิทธิภาพความถูกต้องแม่นยำของโมเดลที่ใช้ในการพยากรณ์และจำแนกข้อมูลที่ถูกสร้างมาจากข้อมูลฝึกที่มีปัญหาตัวอย่างขาดความสมดุลในแง่ปริมาณอย่างสูง ได้มีผู้นำเสนอการนำโมเดลต่าง ๆ มาประยุกต์ใช้เพื่อช่วยแก้ไขปัญหาดังกล่าว รวมไปถึงได้แนะนำแนวต่างการจัดการเบื้องต้นไว้ด้วย

2.2.1 Learning from imbalanced data: open challenges and future directions

เป็นงานวิจัยที่ถูกตีพิมพ์ไว้เมื่อปี 2009 โดย G. Hoang et el. [23] โดยได้อธิบายถึงความหมายของข้อมูลที่ขาดความสมดุล ปัญหาที่เกิดขึ้นต่อโมเดล และแนวทางต่าง ๆ ที่สามารถใช้เพื่อแก้ปัญหา

ในการสร้างโมเดลเพื่อการเรียนรู้ของเครื่องที่จะนำไปใช้ในการจำแนกข้อมูลนั้นตั้งอยู่บนสมมติฐานที่ว่าชุดข้อมูลฝึกมีปริมาณของข้อมูลตัวอย่างระหว่างกลุ่มในระดับที่เท่า ๆ กัน ซึ่งในความเป็นจริงนั้น ข้อมูลที่นำมาใช้มิได้เป็นเช่นนั้นเสมอไปเพราะข้อมูลฝึกมักมีปริมาณระหว่างกลุ่มตัวอย่างไม่เท่ากัน เมื่อนำข้อมูลฝึกลักษณะดังกล่าวมาฝึกโมเดลก็จะส่งผลในประสิทธิภาพของโมเดลในการจำแนกข้อมูลชุดใหม่ในอนาคตมีความผิดพลาดสูง เพราะโมเดลได้ถูกฝึกมาจากข้อมูลที่ไม่สมประกอบ กลุ่มข้อมูลตัวอย่างที่มีจำนวนมากกว่าจะครอบงำความสามารถของโมเดลให้เอนเอียงไปหาข้อมูลตัวอย่างนั้นมาก ในขณะที่กลุ่มข้อมูลตัวอย่างที่มีปริมาณน้อยกว่าจะถูกลดทอนความสำคัญไปโดยปริยาย ดังนั้นงานวิจัยนี้จึงได้แสดงแนวทางการแก้ปัญหาดังกล่าวไว้ดังนี้

2.2.1.1 การแก้ปัญหาในระดับข้อมูลฝึก

เป็นการปรับการกระจายตัวของข้อมูลฝึกก่อนที่จะนำไปใช้ฝึกโมเดล กล่าวคือการสุ่มเพิ่มหรือลดปริมาณข้อมูลตัวอย่างในชุดข้อมูลฝึก เช่นนำโมเดลการสุ่มสร้างข้อมูลโดยอ้างอิงจากเพื่อนบ้านที่มีลักษณะคล้ายคลึง (Synthetic Minority Over-sampling Technique) มาใช้งาน การแก้ปัญหาในระดับข้อมูลฝึกนั้นจัดอยู่ในขั้นตอนของการประมวลข้อมูลเบื้องต้น

2.2.1.2 การแก้ปัญหาในระดับโมเดลที่ใช้จำแนก

เป็นการเน้นการแก้ปัญหาความสมดุลของข้อมูลที่สมการการคำนวณที่โมเดล เพื่อเพิ่มความสามารถในการเรียนรู้กับข้อมูลฝึกประเภทดังกล่าว โดยการแก้ปัญหาจะเน้นไปที่การให้น้ำหนักของผลลัพธ์ที่โมเดลจำแนกออกมา (Cost-sensitive) ผลคำตอบที่จำแนกผิดพลาดจะถูกเพิ่มน้ำหนักความสำคัญขึ้น และลดน้ำหนักของตัวอย่างที่จำแนกได้ถูกต้องลง แล้วทำการปรับจูนโมเดลใหม่เพื่อพยายามลดจำนวนข้อมูลที่โมเดลทายผิด โดยไม่มีการแก้ไขข้อมูลที่ใช้ฝึกใด ๆ ทั้งสิ้น

2.2.1.3 การการแก้ปัญหาแบบผสม

เป็นการนำวิธีการแก้ปัญหาในระดับข้อมูลและในระดับโมเดลมาใช้งานร่วมกัน เพื่อเพิ่มความทนทานและประสิทธิภาพให้กับโมเดลที่ถูกฝึก

นอกเหนือจากนี้ผู้วิจัยยังได้กล่าวถึงข้อมูลที่ขาดความสมดุลกันอย่างสูง (Extremely Class Imbalanced) โดยได้นิยามไว้ว่าหากข้อมูลนั้นมีสัดส่วนระหว่างตัวอย่างข้อมูลต่างกันมากราว 1 ต่อ 5000 จะจัดอยู่ในหมวดของข้อมูลที่ขาดสมดุลอย่างสูง

2.2.2 A Data Mining Approach to Automate Fault Detection Model Development in the Semiconductor Manufacturing Process

ในปี 2011 K. Kerdprasop et el. [27] ได้นำเสนองานวิจัยที่ได้นำองค์ความรู้ทางด้านการทำเหมืองข้อมูลเข้ามาใช้เพื่อเพิ่มประสิทธิภาพในการตรวจจับความผิดพลาด (False Alarm) ในการผลิตสินค้าในธุรกิจโรงงานการผลิตชิ้นส่วนอิเล็กทรอนิกส์ ข้อมูลที่ใช้มีปริมาณมากแต่ขาดความสมดุลระหว่างตัวอย่าง ผู้วิจัยจึงได้แนะนำขั้นตอนการสร้างโมเดลด้วยข้อมูลดังกล่าวในงานวิจัยชิ้นนี้เนื่องจากตัวอย่างข้อมูลมีปัญหาเรื่องการขาดความสมดุลอยู่ด้วยดังนั้นขั้นตอนการเตรียมข้อมูลจึงมีขั้นตอนมากขึ้นจากการฝึกโมเดล โดยขั้นตอนในการปรับสมดุลข้อมูลมีขั้นตอนหลักดังนี้

1. นำวิธีการที่ใช้ในการเลือกคุณสมบัติที่สำคัญ (Feature Selection) จากข้อมูลฝึก เช่น การจัดกลุ่มข้อมูล (Clustering) มาใช้

2. หลังจากนั้นนำข้อมูลที่ได้จากขั้นตอนแรกมาทำการปรับสมดุลด้วยการเพิ่มจำนวนข้อมูลตัวอย่างที่มีปริมาณน้อยกว่าให้มากเท่ากับข้อมูลที่มีปริมาณมากกว่าด้วยวิธีการคัดลอก

3. หลังจากนั้นนำข้อมูลจากขั้นตอนที่ 2 ไปใช้ฝนการฝึกโมเดลและวัดประสิทธิภาพของโมเดล

จากการทดลองพบว่าโมเดลในการสุ่มเพิ่มข้อมูลตัวอย่างและการใช้วิธีการเลือกคุณสมบัติที่สำคัญจากชุดข้อมูลฝึกสามารถเพิ่มประสิทธิภาพให้แก่โมเดลที่ใช้ในการเรียนรู้ได้

บทที่ 3

ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นงานวิจัยเชิงพื้นฐาน (Basic Research) ที่สามารถนำไปประยุกต์ใช้ในงานด้านต่าง ๆ ที่มีการเรียนรู้ของเครื่องกับข้อมูลที่ขาดความสมดุล ตัวอย่างเช่น การทำนายทำงานข้อผิดพลาดของเซ็นเซอร์ในอุปกรณ์หลายประเภท การทำนายการโกงที่มีลักษณะแนบเนียน วิธีการนี้มีเป้าหมายคือเพิ่มประสิทธิภาพความแม่นยำของโมเดลที่ใช้ในการจำแนกประเภทข้อมูล

3.1 แนวทางการวิจัย

3.1.1 ศึกษาทฤษฎีการเรียนรู้ของเครื่องเพื่อใช้ในการประมวลผลข้อมูลที่ขาดความสมดุล

3.1.2 ศึกษาลักษณะของข้อมูลที่นำมาใช้ในการฝึกโมเดล เพื่อให้ผู้วิจัยเข้าใจถึงลักษณะของข้อมูลที่จะนำไปใช้ในการฝึกโมเดลเพื่อแยกแยะข้อมูล

เนื่องจากแนวทางของการฝึกโมเดลด้วยข้อมูลที่ขาดสมดุลมีหลากหลายรูปแบบ ผู้วิจัยจึงต้องค้นคว้าหาข้อมูลเพื่อที่จะทำความเข้าใจการทำงานของแนวทางเหล่านั้น แล้วจึงสามารถนำมาประยุกต์ใช้กับข้อมูลที่ใช้ในงานวิจัยนี้ได้อย่างมีประสิทธิภาพ เพราะเนื่องจากข้อมูลในงานวิจัยที่เกี่ยวข้องนั้นมีความแตกต่างกับข้อมูลที่ใช้ในงานวิจัยนี้

3.1.3 พัฒนาขั้นตอนและแนวทางการฝึกโมเดลด้วยข้อมูลที่ขาดความสมดุล

หลังจากศึกษาค้นคว้าสิ่งที่จำเป็นต่อการเพิ่มประสิทธิภาพในกับโมเดลบนชุดข้อมูลในงานวิจัยนี้แล้ว ผู้วิจัยจึงทำการพัฒนาแนวทางที่เหมาะสมสำหรับการฝึกโมเดลด้วยชุดข้อมูลที่ขาดความสมดุล

3.1.4 ดำเนินการประมวลผลข้อมูลและบันทึกผล

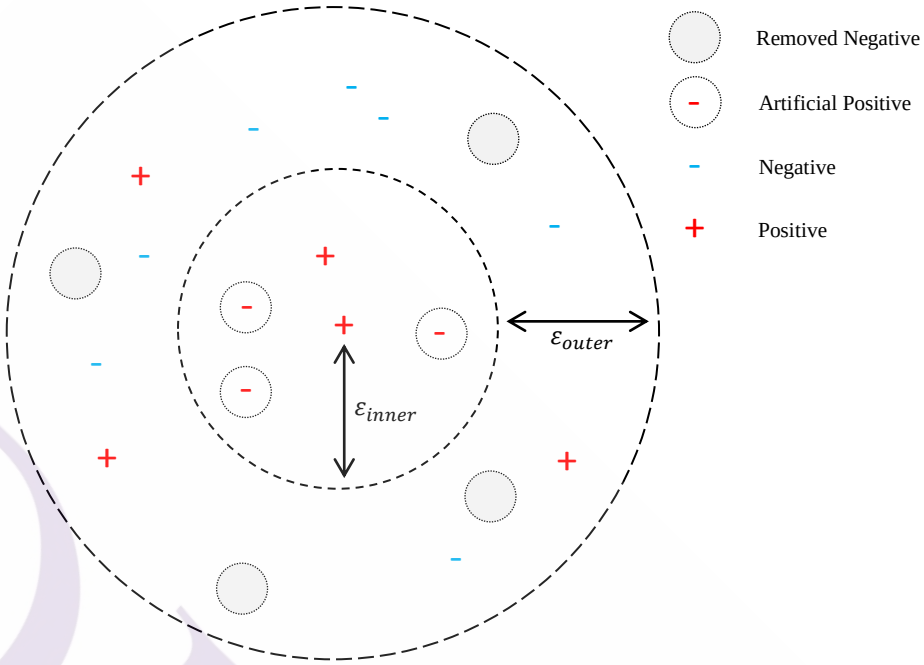
เมื่อพัฒนาแนวทางการฝึกโมเดลเสร็จสิ้น ผู้วิจัยจึงได้เตรียมเครื่องคอมพิวเตอร์ที่เหมาะสมสำหรับการประมวลผลข้อมูลที่ใช้ แล้วจึงนำข้อมูลมาประมวลและเก็บผลลัพธ์ไว้

3.1.5 วิเคราะห์และสรุปผลการทดลอง

หลังจากที่ได้ผลลัพธ์ของการทดลองทั้งหมดแล้วนั้น ผู้วิจัยจึงนำผลทั้งหมดมาวิเคราะห์ร่วมกันแล้วสรุปผลว่าวิธีการสุ่มตัวอย่างแบบที่สามารถเพิ่มประสิทธิภาพความแม่นยำได้ดีที่สุด รวมถึงโมเดลที่สามารถจำแนกข้อมูลได้อย่างแม่นยำที่สุด เพื่อพิสูจน์ว่าสามารถนำขั้นตอนดังกล่าวไปใช้งานได้จริงหรือไม่

3.2 แนวคิดและการทำงานของวิธีการ TwO-levels of Positive Resampling Framework

ความหนาแน่นโดยรอบของข้อมูลนั้นอาจส่งผลให้การสร้างขอบเขตการตัดสินใจของโมเดล การที่ข้อมูลที่มีคลาสต่างกันอยู่ใกล้กันมากจะสามารถทำให้การสร้างขอบเขตดังกล่าวมีความยากลำบาก และส่งผลให้เกิดข้อมูลพลาดในการคาดการณ์ในท้ายที่สุด โดยเฉพาะอย่างยิ่งหากข้อมูลคลาสที่สนใจ (Positive) นั้นถูกล้อมไว้ด้วยคลาสที่ไม่สนใจ (Negative) เป็นจำนวนมาก กล่าวคือมีความคล้ายคลึงกันมากเกินไปจนแยกไม่สามารถแยกแยะได้ดีเท่าใดนัก และในขณะเดียวกันถ้าหากข้อมูลชุดดังกล่าวนั้นมีความไม่สมดุลระหว่างคลาสเกิดขึ้น โดยที่คลาส Negative มากกว่า Positive ก็ยิ่งเป็นการยากที่โมเดลจะสามารถแยกแยะข้อมูลได้แม้ว่าจะการนำวิธีการเพิ่มข้อมูลแบบต่าง ๆ มาใช้เพิ่มปริมาณข้อมูลให้สมดุลกันมากขึ้นก็ตาม ผู้วิจัยเห็นว่าปัญหาดังกล่าวอาจถูกแก้ไขได้บนข้อมูลบางประเภท แต่ทั้งนี้ถ้าหากข้อมูลที่มีปัญหาของความสมดุลและปัญหาความหนาแน่นรอบข้อมูล Positive ก็อาจจะไม่สามารถถูกแก้ไขได้ดีเท่าที่ควร ดังนั้นเพื่อเป็นการแก้ปัญหาดังกล่าว งานวิจัยนี้จึงนำเสนอวิธีการที่เรียกว่าการสุ่มข้อมูลโดยพิจารณาตัวอย่างที่สนใจในพื้นที่โดยรอบข้อมูล Positive สองชั้น

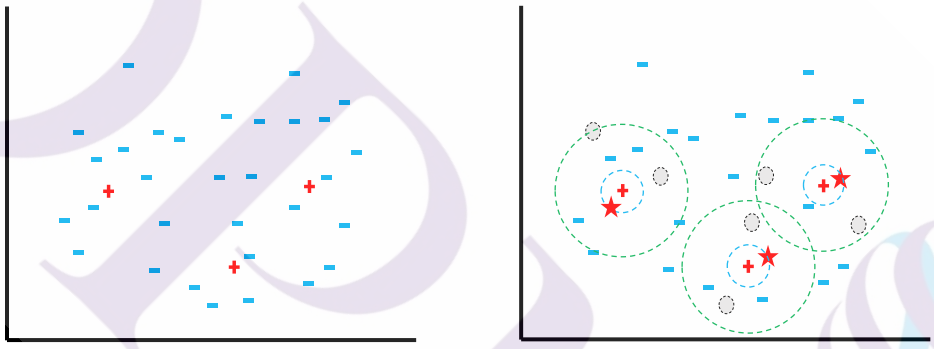


ภาพที่ 3.1 ลักษณะการทำงานของวิธีการ TwO-levels of Positive Resampling Framework

วิธีการนี้พิจารณาข้อมูล Positive ทุกตัวอย่าง โดยทำการสร้างพื้นที่รอบ ๆ ตัวอย่างเหล่านั้นตามค่าระยะที่กำหนด ซึ่งพื้นที่ดังกล่าวจะถูกแบ่งออกเป็นสองระยะเรียกว่า พื้นที่ภายในและพื้นที่ภายนอก ระยะของพื้นที่ดังกล่าวจะถูกสร้างขึ้นบนปริภูมิยุคลิด (Euclidean) โดยระยะห่างจะเริ่มวัดจากข้อมูล Positive ที่กำลังพิจารณาไปยังระยะที่กำหนด (พื้นที่ภายใน) ในพื้นที่นี้จะตรวจสอบว่าจำนวนของตัวอย่าง Positive มีปริมาณคิดเป็นร้อยละเท่าใดเมื่อเทียบเคียงกับจำนวนข้อมูล Positive ทั้งหมดของชุดข้อมูล แล้วจึงดำเนินการแก้ไขข้อมูล วิธีการที่นำเสนอให้ความสำคัญกับข้อมูลตัวอย่างที่สนใจ ดังนั้นการแก้ไขข้อมูลจะมุ่งเน้นที่พื้นที่ดังกล่าว อย่างไรก็ดี ข้อมูลแต่ละชุดมีลักษณะเฉพาะตัวที่แตกต่างกัน วิธีการที่นำเสนอสามารถทำงานร่วมกับวิธีการแก้ไขเพื่อเพิ่มปริมาณข้อมูลแบบอื่นได้ เช่น SMOTE เป็นต้น ในส่วนของพื้นภายนอกจะนับระยะห่างเริ่มตั้งแต่ขอบของพื้นที่ภายในออกไปยังระยะที่กำหนด โดยข้อมูล Negative ที่อยู่ในพื้นที่ภายนอกจะถูกสุ่มลบในปริมาณร้อยละที่กำหนด

3.2.1 วิธีการแบบ TwO-levels of Positive + Vanilla (TOP+V)

วิธีการนี้พิจารณาข้อมูล Positive ทุกตัวอย่าง โดยทำการสร้างพื้นที่รอบ ๆ ตัวอย่างเหล่านั้นตามค่าระยะที่กำหนด ซึ่งพื้นที่ดังกล่าวจะถูกแบ่งออกเป็นสองระยะเรียกว่า พื้นที่ภายใน และพื้นที่ภายนอก ระยะของพื้นที่ดังกล่าวจะถูกสร้างขึ้นบนปริภูมิยุคลิด (Euclidean) โดยระยะห่างจะเริ่มวัดจากข้อมูล Positive ที่กำลังพิจารณาไปยังระยะที่กำหนด (พื้นที่ภายใน) ในพื้นที่นี้จะตรวจสอบว่าจำนวนของตัวอย่าง Positive มีปริมาณคิดเป็นร้อยละเท่าใดเมื่อเทียบเคียงกับจำนวนข้อมูล Positive ทั้งหมดของชุดข้อมูล หากมีจำนวนน้อยกว่าปริมาณที่กำหนด ข้อมูล Negative ที่อยู่ภายในพื้นที่นี้จะถูกเปลี่ยนให้เป็นข้อมูล Positive ทั้งหมด ซึ่งวิธีการเปลี่ยนลักษณะนี้เรียกว่า Vanilla



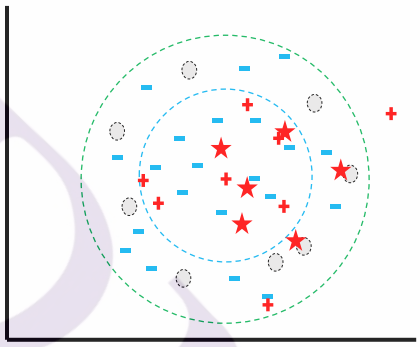
ภาพที่ 3.2 การเปรียบลักษณะของข้อมูลตั้งต้นและข้อมูลที่แก้ไขด้วยวิธีการ TOP เมื่อทำงานร่วมกับ Vanilla

จากภาพ 3.2 จะเห็นได้ว่าข้อมูล Positive ใหม่จะถูกสร้างขึ้นโดยรอบข้อมูล Positive เก่า และข้อมูล Negative จะถูกลดปริมาณลง อย่างไรก็ดีจุดสังเกตหนึ่งของการทำงานดังกล่าวคือ สัดส่วนของข้อมูลระหว่าง Positive และ Negative สำหรับข้อมูลบางชุดอาจจะไม่มีการเปลี่ยนแปลงมาอยู่ในระดับที่เท่า ๆ กันถ้าหากกำหนดพื้นที่ภายในให้มีระยะที่แคบและจำนวนการสุ่มลบข้อมูล Negative นั้นมีปริมาณน้อย

ลักษณะการทำงานดังกล่าวมุ่งเน้นที่การลดปริมาณข้อมูลของ Negative ที่อยู่ใกล้กับ Positive มาก เหมาะสำหรับข้อมูลที่มีความคลุมเครือระหว่าง Positive และ Negative สูงเมื่อเทียบกับวิธีการอื่น ๆ ที่นำมาใช้ในงานวิจัย ดังนั้นอาณาเขตของของพื้นที่ภายในควรมีลักษณะแคบ

3.2.2 วิธีการแบบ TwO-levels of Positive + ROS (TOP+ROS)

ลักษณะการทำงานร่วมกันระหว่าง TOP และ ROS จะเป็นการเพิ่มปริมาณข้อมูลในระยะภายในแบบสุ่ม

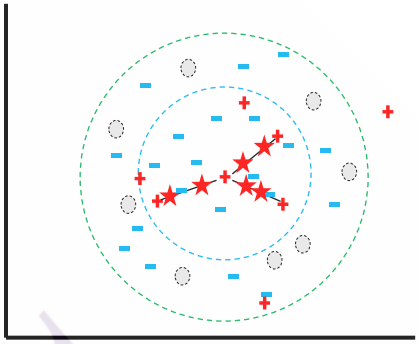


ภาพที่ 3.3 การเปรียบลักษณะของข้อมูลตั้งต้นและข้อมูลที่แก้ไขด้วยวิธีการ TOP เมื่อทำงานร่วมกับ ROS

ปริมาณของข้อมูล Positive ที่เพิ่มขึ้นดังภาพที่ 3.3 เพิ่มขึ้นแบบสุ่ม การทำงานนี้จะเกิดขึ้นในพื้นที่ภายใน ข้อมูล Negative ในพื้นที่ดังกล่าวไม่มีการเปลี่ยนแปลงใด ๆ ตรงข้ามกับพื้นที่ภายนอกที่มีการลดปริมาณข้อมูล Negative ลงอย่างสุ่มในปริมาณที่กำหนด อย่างไรก็ดี การเพิ่มปริมาณข้อมูล Positive ด้วยวิธีการนี้เหมาะกับข้อมูล Positive ที่มี Negative โดยรอบไม่มากนัก เพราะการเพิ่มปริมาณ Positive อย่างสุ่มนี้ อาจส่งผลให้เกิดการทับซ้อนข้อมูลระหว่าง Positive ที่เพิ่มขึ้นและข้อมูล Negative ที่อยู่ใกล้เคียง ดังนั้นความกว้างของพื้นที่ภายในจึงควรมีลักษณะไม่กว้างจนเกินไป

3.2.3 วิธีการแบบ TwO-levels of Positive + SMOTE (TOP+SMOTE)

ลักษณะการทำงานร่วมกันระหว่าง TOP และ SMOTE จะเป็นการเพิ่มปริมาณข้อมูลในระยะภายในแบบสุ่ม

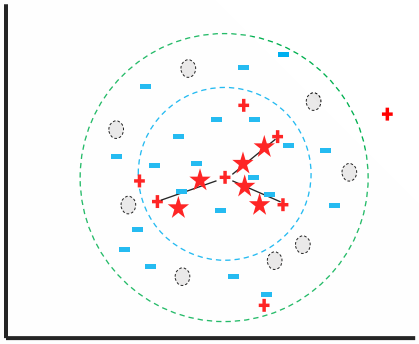


ภาพที่ 3.3 การเปรียบลักษณะของข้อมูลตั้งต้นและข้อมูลที่แก้ไขด้วยวิธีการ TOP เมื่อทำงานร่วมกับ SMOTE กำหนด k เท่ากับ 3

เนื่องจากการเพิ่มปริมาณ Positive แบบสุ่มนั้นอาจส่งผลให้เกิดการทับซ้อนของข้อมูลได้ ดังนั้นการเพิ่มปริมาณข้อมูล Positive จึงควรดำเนินการอย่างระมัดระวัง วิธีการนี้ได้นำข้อดีข้อง SMOTE มากำหนดพื้นที่ในการเพิ่มข้อมูล Positive ใหม่ ด้วยการสร้างเส้นเชื่อมโยงระหว่าง Positive ที่สนใจไปยัง Positive ที่ใกล้เคียงแล้วจึงสร้าง Positive ขึ้นมาใหม่บนเส้นนั้น ดังภาพที่ 3.3 อย่างไรก็ดี ในบางครั้งข้อมูล Positive ที่สร้างขึ้นมา อาจจะทับซ้อนกับข้อมูล Negative ได้ หากข้อมูลดังกล่าวตั้งอยู่ระหว่างข้อมูล Positive ที่สนใจและข้อมูล Positive ที่ใกล้เคียง ความกว้างของพื้นที่ภายในไม่ควรมากจนเกินไปเพราะจะเพิ่มโอกาสในการทับซ้อนของข้อมูลได้

3.2.4. วิธีการแบบ TwO-levels of Positive + RSLS (TOP+RSLS)

ลักษณะการทำงานร่วมกันระหว่าง TOP และ RSLS จะเป็นการเพิ่มปริมาณข้อมูลในระยะภายในแบบสุ่ม

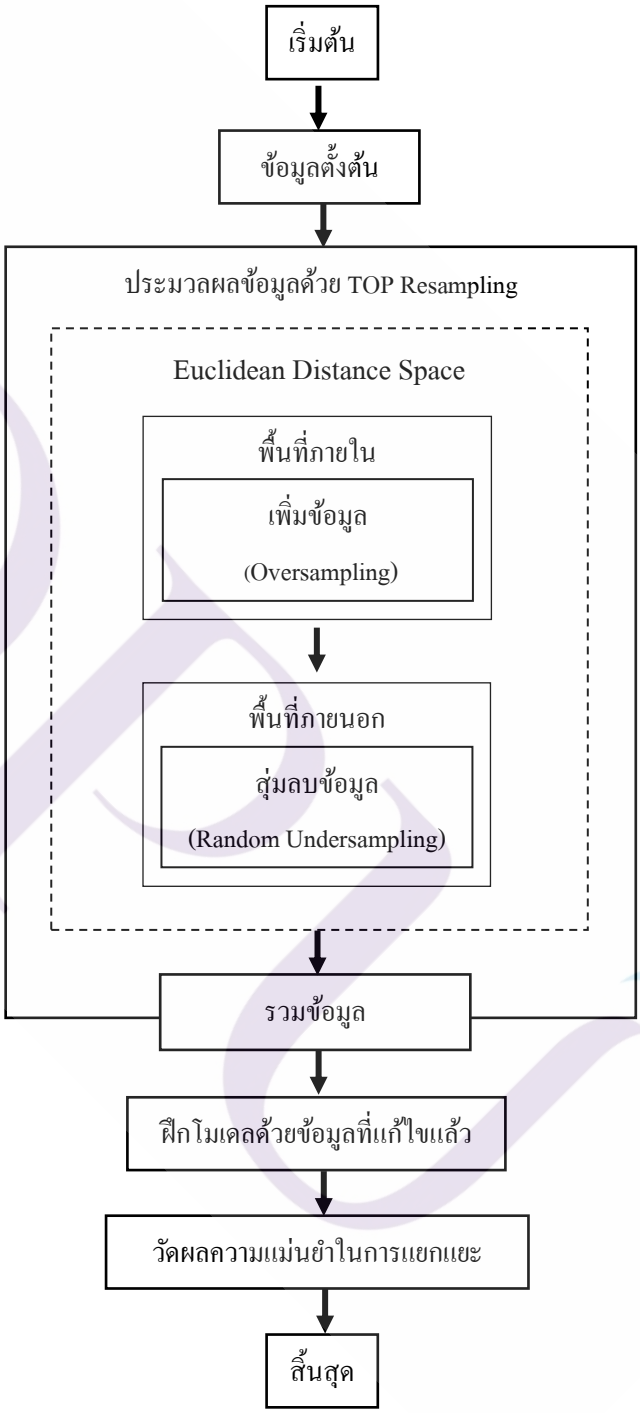


ภาพที่ 3.4 การเปรียบลักษณะของข้อมูลตั้งต้นและข้อมูลที่แก้ไขด้วยวิธีการ TOP เมื่อทำงานร่วมกับ RSLS กำหนด k เท่ากับ 3

ลักษณะการทำงานของ RSLS นั้นมีความคล้างคลึงกับ SMOTE แต่มีความแตกต่างตรงการสร้างข้อมูล Positive ที่มีความระมัดระวังมากกว่า กล่าวคือ หากพบว่ามีข้อมูล Negative ตั้งอยู่บนเส้นที่เชื่อมระหว่างข้อมูล Positive ที่กำลังพิจารณากับข้อมูล Positive ที่ใกล้เคียง วิธีการ RSLS จะไม่สร้างข้อมูล Positive ทับลงบนข้อมูล Negative เหล่านั้น แต่จะพยายามเลื่อนข้อมูล Positive ดังกล่าวออกเล็กน้อย เพื่อป้องกันการทับซ้อน ความกว้างของพื้นที่ภายในสามารถกำหนดให้กว้างกว่าวิธีการแบบ TOP+SMOTE ได้ เพราะการสร้างข้อมูล Positive ใหม่มีโอกาสที่จะทับซ้อนกับข้อมูล Negative น้อยกว่า

จากแนวคิดจากหัวข้อ 3.2.4 สามารถแสดงเป็นแนวทาง (Concept) และ Pseudocode ได้ดังแผนภาพที่ 3.4 และ 3.5 ตามลำดับ

3.3 ขั้นตอนการทำงานของวิธีการที่นำเสนอโดยสังเขป



ภาพที่ 3.4 ขั้นตอนการทำงานของวิธีการ TwO-levels of Positive Resampling Framework โดยสรุป

3.4 ขั้นตอนการทำงานของวิธีการที่นำเสนอโดยละเอียด

Algorithm 1 : TwO-level of Positive resampling

Input:
S is an imbalanced $m \times n$ dataset.

Parameter:
ε_{in} is an inner level epsilon (0 to 1).
ε_{out} is an outer level epsilon (0 to 1).
pos_{in} is a percentage of positive in inner level (0 to 1).
neg_{rm} is a percentage of negative to be removed (0 to 1).
t is an oversampling technique.
e is a number of errors allowed (1 to $+\infty$).
... depends on oversampling technique.

Output:
Modified Dataset

function TOP($S, \varepsilon_{in}, \varepsilon_{out}, pos_{in}, neg_{rm}, t, ...$)
  MAT_{dist} = EUCLIDDISTANCE(S) ▷ Produce $m \times n$ matrix of distance between data point in S
  MAT_{in} = FINDMEMBER($MAT_{dist}, 0, \varepsilon_{in}$) ▷ Produce $m \times n$ matrix of inner area
  MAT_{out} = FINDMEMBER($MAT_{dist}, \varepsilon_{in}, \varepsilon_{out}$) ▷ Produce $m \times n$ matrix of outer area
  MAT_{ooa} = FINDMEMBER($MAT_{dist}, \varepsilon_{out}, 1$) ▷ Produce $m \times n$ matrix of out of area
  P = COUNT($positive \in S$) ▷ Produce total number of positive instance in S
  $MAT_{OSresult}$ = OVERSAMPLING($t, MAT_{inner}, pos_{in}, P$) ▷ Produce $m \times n$ matrix of oversampled data
  $MAT_{USresult}$ = UNDERSAMPLING(MAT_{outer}, neg_{out}) ▷ Produce $m \times n$ matrix of undersampled data
  return $MAT_{OSresult} \cup MAT_{USresult} \cup MAT_{non}$
end function

function FINDMEMBER($MAT_d, \varepsilon_{min}, \varepsilon_{max}$)
  MAT_{ds} = RANGESCALER(MAT_d) ▷ Rescale distance value into range 0 to 1
  for all $d \in MAT_{ds}$ **do** ▷ Loop check all distance value between data point through MAT_{ds}
    if $d > \varepsilon_{min}$ **and** $d \leq \varepsilon_{max}$ **then**
      $MAT_{member} = d$ ▷ Add data point which lies in given area
    else
      if $\varepsilon_{min} \neq 0$ **and** $d > \varepsilon_{max}$ **then**
        $MAT_{nonmember} = d$ ▷ Add data point which is not located in given area
      end if
    end if
  end for
  if $MAT_{nonmember} = \varnothing$ **then**
    return MAT_{member}
  else
    return $MAT_{nonmember}$
  end if
end function

function OVERSAMPLING($MAT_{inner}, pos_{in}, P, t, ...$)
  for all $pos \in MAT_{inner}$ **do** ▷ Where pos is consideration positive instance
    p = COUNT($p \in pos$) ▷ Where p is positive in pos area
    if $\frac{p}{P} \times 100 < pos_{in}$ **then**
      if t == "$vanilla$" **then**
        MAT_{result} = VANILLA($MAT_{inner_{pos}}$)
      else
        MAT_{result} = OSWRAPPER($MAT_{inner_{pos}}, t, e, ...$)
        MAT_{result} = REDUCER(MAT_{result}) ▷ Reduce duplicate positive which generated by OSWRAPPER()
      end if
    end if
  end for
  return MAT_{result}
end function

function VANILLA(Mat_{inner}, pos_{in})
  MAT_{result} = SWAPCLASS($negative, positive$) ▷ Convert negative instance to positive instance
  return MAT_{result}
end function

Algorithm 1 : TwO-level of Positive resampling (continued)

function OSWRAPPER($MAT_{inner}, t, e, ...$)
  $i = 0$
  $MAT_{innerbk} = MAT_{inner}$ ▷ Backup original MAT_{inner}
  while $i \neq e$ **do**
    try
      MAT_{result} = FUNCTION $t(MAT_{inner}, ...)$ ▷ Where t is an algorithm such as SMOTE, RSLS etc.
      return MAT_{result}
    catch *error*
      if $i == e$ **then**
        return $MAT_{innerbk}$ ▷ Return $MAT_{innerbk}$ if error cannot be resolved
      else
        $MAT_{inner} = randomly\ duplicate\ 1\ pos$ ▷ Randomly select pos from MAT_{inner} and add to self
        $i \leftarrow i + 1$
      end if
    end try
  end while
end function

function UNDERSAMPLING(MAT_{outer}, neg_{rm})
  for all $pos \in MAT_{outer}$ **do** ▷ Where pos is consideration positive instance
    n = COUNT$(n \in pos)$ ▷ Where n is negative in pos area
    rm = ROUND$(n \times neg_{rm})$ ▷ Round value into number of negative to be removed
    MAT_{result} = RANDREMOVE(MAT_{outer}, rm) ▷ Randomly remove negative at rm size in pos area from MAT_{outer}
  end for
  return MAT_{result}
end function

ภาพที่ 3.5 ขั้นตอนการทำงานของวิธีการ TwO-levels of Positive Resampling Framework อย่างละเอียด

ตารางที่ 3.1 พารามิเตอร์และความหมาย

ชื่อพารามิเตอร์	คำอธิบาย
ε_{inner}	ระยะห่างระหว่างข้อมูล Positive ที่กำลังพิจารณาไปยังขอบเขตของพื้นที่ภายใน
ε_{outer}	ระยะห่างระหว่างขอบเขตชั้นในไปยังขอบเขตของพื้นที่ภายนอก
pos_{in}	ร้อยละของจำนวนข้อมูล Positive ที่อยู่ภายในพื้นที่ภายในเมื่อเปรียบเทียบกับจำนวน Positive ทั้งหมดในชุดข้อมูล
neg_{rm}	ร้อยละของจำนวนข้อมูล Negative ที่อยู่ภายในพื้นที่ภายนอกที่จะทำการลบ
t	วิธีการเพิ่มข้อมูล Positive แบบต่าง ๆ เช่น vanilla, SMOTE และ RSLS
e	จำนวนครั้งที่ยอมรับในการทำซ้ำในกรณีที่เกิดข้อผิดพลาดในการทำงานของฟังก์ชั่น
...	พารามิเตอร์ใด ๆ ที่อาจจำเป็นเมื่อใช้วิธีการเพิ่มข้อมูล Positive แบบอื่น ๆ นอกจาก Vanilla เช่น ค่า k ของ SMOTE เป็นต้น

คำอธิบายการทำงาน

จากแผนภาพที่ 3.5 ขั้นตอนการแก้ไขข้อมูลด้วยวิธีการ TOP เริ่มจากการนำเข้าข้อมูลตั้งต้นเพื่อนำมาหาระยะห่างระหว่างตัวอย่างทั้งหมดในชุดข้อมูลด้วยการวัดระยะแบบยูคลิด (บรรทัดที่ 2) จากนั้นเมื่อได้ระยะห่างระหว่างข้อมูลแล้วจึงนำมาใช้หาว่าข้อมูลใดอยู่ในระยะพื้นที่ที่กำหนดบ้าง (บรรทัดที่ 3 ถึง 5) การหาข้อมูลภายในระยะที่กำหนดจะถูกแบ่งออกเป็นสามส่วนคือ 1. ข้อมูลที่อยู่ภายในพื้นที่ชั้นใน (inner area) 2. ข้อมูลที่อยู่ภายในพื้นที่ชั้นนอก (outer area) 3. ข้อมูลที่ไม่อยู่ในระยะที่กำหนด (out of area) โดยในขั้นตอนแรกข้อมูลระยะห่างระหว่างข้อมูลที่หมดจะถูกนำมาปรับให้อยู่ในช่วงระหว่าง 0 ถึง 1 (บรรทัดที่ 12) แล้วจึงนำมาใช้ต่อในขั้นตอนถัดไป การหาข้อมูลที่อยู่ในพื้นที่ชั้นในจะถูกค้นหาด้วยการวัดระยะห่างระหว่างข้อมูลจาก Positive ที่กำลังพิจารณาไปยังระยะที่กำหนด โดยค่าเริ่มต้นของระยะห่างคือ 0 ไปยัง ε_{inner} หากข้อมูลที่มีระยะมากกว่า 0 และน้อยกว่าหรือเท่ากับ ε_{inner} หมายความว่าตัวอย่างข้อมูลที่อยู่ในระยะดังกล่าวคือสมาชิกในพื้นที่ชั้นในของตัวอย่าง Positive ที่กำลังพิจารณา (บรรทัดที่ 14 ถึง 15) โดยจะวนพิจารณาทุกตัวอย่าง Positive ภายในชุดข้อมูล โดยข้อมูล Positive ทุกตัวอย่างจะมีสมาชิกอยู่ในพื้นที่ชั้นในและชั้นนอกเป็นของตัวเอง ในส่วนของพื้นที่ภายนอกก็ดำเนินการเหมือนกับพื้นที่ภายในแต่ต่างกันในส่วนของเงื่อนไขการตัดสินใจ โดยถ้าหากข้อมูลตกอยู่ในระยะที่มีค่ามากกว่า ε_{inner} และน้อยกว่าหรือเท่ากับ ε_{outer} หมายความว่าข้อมูลดังกล่าวตกอยู่ในพื้นที่ชั้นใน อย่างไรก็ดีข้อมูลที่ไม่ตกอยู่ภายใต้เงื่อนไขที่กล่าวมาจะถือว่าไม่อยู่ในพื้นที่ที่กำหนด (บรรทัดที่ 17 ถึง 18) หลังจากนั้นจำนวนตัวอย่าง Positive ทั้งหมดในชุดข้อมูลจะถูกนำมานับ (บรรทัดที่ 6) แล้วจะนำไปใช้ตัดสินใจต่อไป เมื่อข้อมูลพร้อมแล้วจึงดำเนินการแก้ไขข้อมูลด้วยวิธีการเพิ่มจำนวน Positive (บรรทัดที่ 7) โดยข้อมูลที่อยู่ภายในพื้นที่ชั้นในจะถูกนำมาใช้ในขั้นตอนนี้ ซึ่งข้อมูลตัวอย่าง Positive ทุกตัวจะถูกนำมาพิจารณาจำนวนสมาชิกด้วยการนับว่าถ้าหากมี Positive เป็นสมาชิกอยู่จำนวนเท่ากับ p หารด้วย P แล้วน้อยกว่าหรือเท่ากับ pos_{in} ดังนั้นจะถูกนำไปดำเนินเพิ่มข้อมูลด้วยวิธีการแบบ t แต่ถ้าหากไม่เข้าเงื่อนไขที่กำหนด ข้อมูลในพื้นที่ชั้นในของ Positive ที่สนใจก็จะไม่ถูกนำมาเพิ่ม (บรรทัดที่ 29 ถึง 37 และ 40) โดยที่จะมีการวนดำเนินการทำวิธีการดังกล่าวกับตัวอย่าง Positive ทุกตัวจนครบ (ไม่นับ Positive ที่ถูกสร้างขึ้นมาใหม่) ในการเพิ่มข้อมูล Positive ใหม่นั้นถ้าหาก t ถูกกำหนดเป็นวิธีการ Vanilla ข้อมูลสมาชิกของทุก Positive ที่เป็น Negative จะถูก

เปลี่ยนเป็น Positive ใหม่ (บรรทัดที่ 42 ถึง 45) แต่ถ้าหาก t เป็นวิธีการอื่น เช่น SMOTE ข้อมูลก็จะถูกเพิ่มด้วยวิธีการดังกล่าว (บรรทัดที่ 35) อย่างไรก็ดีในบางครั้งการดำเนินการเพิ่มข้อมูลอาจมีข้อผิดพลาดเกิดขึ้นได้ ยกตัวอย่างเช่น ข้อมูลที่นำมาแก้ไขมีจำนวนเพื่อนบ้านน้อยกว่าค่า k ที่กำหนด เนื่องจากไม่มีข้อมูลสมาชิก Positive มากพอ จึงส่งผล SMOTE จะไม่สามารถทำงานได้ ดังนั้นจึงจำเป็นแก้ไขโดยการคัดลอกข้อมูลสมาชิกที่เป็น Positive เพิ่มขึ้นทีละหนึ่งตัว ดังนั้นการวนลูปการทำงานของการแก้ไขจะทำเป็นจำนวน e ครั้ง เพราะไม่สามารถคาดการได้ล่วงหน้าว่าข้อผิดพลาดดังกล่าวจะมีจำนวนเท่าใด (บรรทัดที่ 46 ถึง 62) เมื่อข้อมูลได้ผ่านการเพิ่มจำนวนแล้วจะได้ออกมาเป็นข้อมูลชุดใหม่ จากนั้นจึงดำเนินการนำข้อมูลที่อยู่ในพื้นที่ชั้นนอกมาสุ่มลดขนาดข้อมูลลงด้วยปริมาณร้อยละ neg_{rm} ซึ่งสมาชิกในพื้นของ Positive ทุกตัวจะถูกสุ่มลบ (บรรทัด 63 ถึง 70) หลังจากนั้นจะได้เป็นข้อมูลใหม่ สุดท้ายแล้วข้อมูลทั้งสามชุดที่ประกอบด้วย 1. ข้อมูลจากพื้นที่ภายในที่ถูกเพิ่มปริมาณ Positive 2. ข้อมูลจากพื้นที่ภายในที่ถูกลดปริมาณ Negative 3. ข้อมูลที่อยู่ภายนอกระยะ จะถูกนำมารวมกันเป็นข้อมุลชุดใหม่ที่ได้ถูกปรับปรุงแล้ว (บรรทัดที่ 9) เพื่อส่งไปเป็นข้อมูลที่ใช้สอนโมเดลต่อไป

วิธีการทดลอง

ผู้วิจัยได้ตั้งค่าการทดลองไว้หลายส่วนด้วยกัน ประกอบด้วย การจัดเตรียมข้อมูล การตั้งค่าการแก้ไขข้อมูลด้วยวิธีการต่าง ๆ ตลอดจนการเก็บและสรุปผล

3.4.1 จัดเตรียมข้อมูลที่จะนำมาใช้ในงานวิจัย

ข้อมูลที่นำมาใช้ในงานวิจัยนั้นได้ถูกนำมาจากแหล่งข้อมูลสาธารณะที่เชื่อถือได้ จำนวนข้อมูลที่นำมาใช้ทั้งหมดมีจำนวน 15 ชุด โดยแบ่งออกเป็นข้อมูลจากเว็บไซต์ UCI จำนวน 2 ชุด และข้อมูลจากเว็บไซต์ KEEL จำนวน 13 ชุด

3.4.2 ประมวลผลข้อมูลเบื้องต้น

การประมวลผลข้อมูลเบื้องต้นนั้นเป็นการเตรียมความพร้อมของชุดข้อมูลที่จะนำไปใช้ในการฝึกโมเดล เริ่มตั้งแต่การทำความเข้าใจลักษณะของข้อมูล การปรับหน่วยข้อมูล และการปรับสมดุลข้อมูล

3.4.3 การทำความเข้าใจข้อมูล

ผู้วิจัยได้ทำการประมวลผลลักษณะการขาดความสมดุลและความทับซ้อน โดยลักษณะการแจกแจงสัดส่วนระหว่างกลุ่มข้อมูลของตัวแปรป้ายกำกับ และระดับการทับซ้อนข้อมูล ดังตาราง 3.2

ตารางที่ 3.2 ลักษณะของชุดข้อมูล

ลำดับที่	ชื่อชุดข้อมูล	อัตราส่วนความสมดุล	จำนวนข้อมูลที่ไม่พอ	จำนวนข้อมูล Positive	จำนวนข้อมูล Negative	ระดับการทับซ้อน F2	ระดับการทับซ้อน F3
1	ecoli2	5.46	232	52	284	0.00	0.21
2	glass0	2.05	74	70	144	0.00	0.29
3	glass1	1.81	62	76	138	0.01	0.10
4	glass6	6.37	156	29	185	0.01	0.62
5	haberman	2.77	144	81	225	0.72	0.03
6	Liver-disorders	1.37	55	200	145	0.07	0.03
7	new-thyroid1	5.14	145	35	180	0.00	0.81
8	new-thyroid2	5.14	145	35	180	0.00	0.81
9	page-blocks-1-3 vs 4	15.8	416	444	28	0.00	0.83
10	pima	1.86	232	500	268	0.25	0.01
11	vehicle1	2.89	412	217	629	0.00	0.06
12	vehicle2	2.88	410	218	628	0.00	0.23
13	wisconsin	1.85	205	444	239	0.22	0.12
14	yeast1	2.45	626	1055	429	0.00	0.04
15	yeast3	8.1	1158	1321	163	0.00	0.54

3.4.4 การปรับสมดุลข้อมูล

หลังจากที่เตรียมข้อมูลเสร็จเรียบร้อยแล้วจึงข้อมูลมาทำการปรับความสมดุลระหว่างกลุ่มตัวอย่างด้วยวิธีการที่ใช้ในการปรับดุลต่าง ๆ ที่มีรายชื่อดังต่อไปนี้

1. Baseline (BASE) ไม่มีการสุ่มใด ๆ
2. Random Over Sampling (ROS) เพื่อเพิ่มจำนวนกลุ่มตัวอย่างข้อมูล Positive
3. Synthetic Minority Over-Sampling Technique (SMOTE) เพื่อเพิ่มจำนวนกลุ่มตัวอย่างข้อมูล Positive
4. Relocate Safe-Level-Synthetic Minority Over-Sampling Technique (RSLS) เพื่อเพิ่มจำนวนกลุ่มตัวอย่างข้อมูล Positive
5. Random Over-Sampling Examples (ROSE) เพื่อเพิ่มจำนวนกลุ่มตัวอย่างข้อมูล Positive
6. Random Over and Under-Sampling (OVUN) เพื่อเพิ่มและลดจำนวนกลุ่มตัวอย่างข้อมูล Positive และ Negative ตามลำดับ
7. TwO-levels of Positive & Vanilla (TOP+V) เพื่อเพิ่มและลดจำนวนกลุ่มตัวอย่างข้อมูล Positive และ Negative ตามลำดับ
8. TwO-levels of Positive Resampling & Random Over Sampling (TOP+ROS) เพื่อเพิ่มและลดจำนวนกลุ่มตัวอย่างข้อมูล Positive และ Negative ตามลำดับ
9. TwO-levels of Positive Resampling & Synthetic Minority Over-Sampling Technique (TOP+SMOTE) เพื่อเพิ่มและลดจำนวนกลุ่มตัวอย่างข้อมูล Positive และ Negative ตามลำดับ
10. TwO-levels of Positive Resampling & Relocate Safe-Level-Synthetic Minority Over-Sampling Technique (TOP+RSLS) เพื่อเพิ่มและลดจำนวนกลุ่มตัวอย่างข้อมูล Positive และ Negative ตามลำดับ

ทั้งหมดนี้จะถูกนำมาใช้ให้การปรับข้อมูลชุดเดียวกัน ภายใต้สภาพแวดล้อม (Environment) ที่เหมือนกัน แต่ดำเนินการอย่างอิสระต่อกัน ดังนั้นผลลัพธ์ของข้อมูลชุดใหม่จะเกิดขึ้นตามจำนวนวิธีการที่ใช้ในการปรับสมดุล ซึ่งมีทั้งหมด 10 แบบ จึงจะได้ข้อมูลใหม่ 10 ชุดต่อข้อมูลเริ่มแรก 1 ชุด

จำนวนพารามิเตอร์เพื่อนบ้านที่ใกล้ที่สุดสำหรับวิธีการสุ่มทุกแบบ คือ k = 5 ผลการทำนายของทุกโมเดลจะใช้ความน่าจะเป็นมีค่าเท่ากับ 0.5 ในการแบ่งคลาส อย่างไรก็ดีวิธีการ DBSM ไม่สามารถนำมาใช้ได้ในการทดลองเนื่องจากข้อจำกัดทางด้าน Implementation

สำหรับพารามิเตอร์ของวิธีการสุ่มข้อมูลโดยพิจารณาตัวอย่างที่สนใจในพื้นที่โดยรอบข้อมูล Positive สองชั้น จะทำค้นหาค่าที่ดีที่สุดด้วยวิธีการค้นหาแบบตาราง (Grid Search) ด้วยค่าดังตารางที่ 3.3

ตารางที่ 3.3 ค่าพารามิเตอร์ที่ใช้สำหรับวิธีการ TwO-levels of Positive Resampling Framework

ชื่อพารามิเตอร์	ค่าที่ใช้
ε_{inner}	0.01, 0.05, 0.10, 0.15, 0.20
ε_{outer}	0.01, 0.05, 0.10, 0.15, 0.20
pos_{in}	0.01, 0.05, 0.10, 0.15, 0.20
neg_{rm}	0.01, 0.05, 0.10, 0.15, 0.20

ซึ่งค่าทั้งหมดจะถูกนำมาจับคู่ได้เป็นจำนวนเท่ากับ 5 ค่า ยกกำลังด้วย 4 พารามิเตอร์ จะได้เป็นชุดพารามิเตอร์ทั้งหมดจำนวน 625 ชุด ไว้ใช้สำหรับแต่ละวิธีการแก้ไขข้อมูล ดังนั้น 625 คูณด้วยวิธีการแก้ไข 4 แบบ จะได้เป็น 2,500 ชุดต่อหนึ่งโมเดลต่อหนึ่งชุดข้อมูล

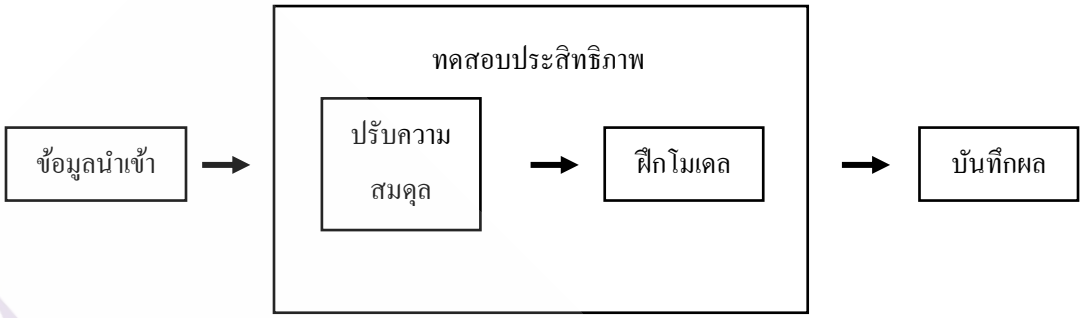
3.4.5 ฝึกโมเดลเพื่อประมวลผลข้อมูลและทดสอบประสิทธิภาพ

เมื่อได้ข้อมูลฝึกที่แก้ไขแล้ว ก็จะนำข้อมูลเหล่านั้นฝึกโมเดลที่ใช้ในการแยกแยะข้อมูล โดยที่โมเดลเดียวกันจะถูกนำมาฝึกตามจำนวนข้อมูลที่สมดุลของแต่ละวิธี กล่าวคือ หากข้อมูลที่ถูกปรับสมดุลที่ถูกสร้างมาใหม่ 15 ชุด โมเดลเดียวกันก็จะถูกฝึกออกมา 15 ตัว ด้วยข้อมูลดังกล่าว โดยโมเดลทั้งหมดที่ใช้จะประกอบด้วย

1. Decision Tree (C4.5)
2. Decision Tree (C5.0)
3. Random Forest (RF)
4. eXtreme Gradient Boosting Tree (XGB)
5. Naïve Bayes (NB)
6. Logistics Regression (LR)
7. 1-Nearest Neighbor (1-NN)
8. 3-Nearest Neighbor (3-NN)
9. 5-Nearest Neighbor (5-NN)
10. Neural Network (NN)
11. Support Vector Machine with RBF (SVM)

โมเดลทั้ง 11 แบบจะถูกฝึกร่วมกับการใช้วิธีการแก้ไขข้อมูล TOP แบบต่าง ๆ จำนวน 2,500 แบบ บนข้อมูล 15 ชุด จะได้เป็นผลลัพธ์จำนวน 412,500 โมเดลสำหรับวิธีการที่นำเสนอ ในส่วนของวิธีการแก้ไขข้อมูลแบบอื่น ๆ นั้น มีจำนวน 5 แบบ และฝึกแบบไม่แก้ไขข้อมูลจำนวน 1 แบบ บนข้อมูล 15 ชุด จะได้ผลลัพธ์เป็น 990 โมเดล ดังนั้นสามารถสรุปได้ว่าจำนวนโมเดลทั้งหมดในการทดลองของงานวิจัยชิ้นนี้ได้มีการฝึกโมเดลรวมทั้งหมดจำนวน 413,490 โมเดล

โมเดลเดียวกันจะถูกตั้งค่าพารามิเตอร์ไว้เหมือนกันทั้งหมดในสำหรับแต่ละชุดข้อมูลฝึก จากนั้นทดสอบประสิทธิภาพด้วยการทดสอบแบบไขว้ (Cross Validation) กระบวนการนี้จะต้องนำขั้นตอน 1. ปรับสมดุลข้อมูล 2. ฝึกโมเดล รวมเข้าไว้ภายใต้การทดสอบประสิทธิภาพเดียวกัน เพื่อป้องการกระจายตัวของข้อมูลผิดเพี้ยน ขั้นตอนการฝึกสามารถเขียนเป็นแผนภาพโดยสรุปได้ดังนี้



ภาพที่ 3.6 ตอนขั้นตอนการฝึกโมเดลด้วยข้อมูลที่ขาดสมดุลรายชุดข้อมูล

ผลการทดลองบันทึกไว้ในรูปแบบของตารางบันทึกผลที่มีลักษณะดังนี้

ตารางที่ 3.4 ตารางตัวอย่างที่แสดงการแจกแจงค่าเฉลี่ยผลลัทธ์ที่ได้จากขั้นตอนการทดสอบโมเดล

	Model A	**Model B**	**Model C**	**Model D**	**Model E**	**Model F**	…
Sampling A	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	…
Sampling B	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	…
Sampling C	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	…
Sampling D	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	…
Sampling E	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	ผลลัพธ์เฉลี่ย	...
…	…	…	…	…	…	…	…

ตารางที่ 3.5 ตารางตัวอย่างที่แสดงการแจกแจงผลลัพธ์ที่ได้จากขั้นตอนการทดสอบโมเดล

	Model A	**Model B**	**Model C**	**Model D**	**Model E**	**Model F**	…
Sampling A	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	…
Sampling B	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	…
Sampling C	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	…
Sampling D	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	…
Sampling E	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	ผลลัพธ์	...
…	…	…	…	…	…	…	…

3.4.6 วิเคราะห์และสรุปผลลัพธ์

หลังจากได้ผลการทดสอบของโมเดลทั้งหมด ผู้วิจัยจะตรวจสอบผลโดยใช้หน่วยวัดประสิทธิภาพ F1 ในการวัดความสามารถของโมเดลแต่ละตัว โดยหาว่าวิธีการปรับสมดุลแบบใดสามารถเพิ่มความถูกต้องแม่นยำให้กับโมเดลหลายตัวที่สุด ได้อันดับสูงมากที่สุด และหาว่าโมเดลใดที่มีประสิทธิภาพดีที่สุดโดยดูว่าโมเดลใดที่สามารถใช้วิธีการปรับสมดุลแล้วเพิ่มความถูกต้องได้หลายวิธีที่สุด ในส่วนของหน่วยวัดผลความแม่นยำอื่น ๆ ที่ใช้ในงานวิจัยจะมีไว้เพื่อสนับสนุนการประสิทธิภาพ F1 ของโมเดล

3.5 เครื่องมือที่ใช้ในงานวิจัย

3.5.1 คอมพิวเตอร์เสมือน (VM) สำหรับประมวลผลจำนวน 5 เครื่อง โดยแต่ละเครื่องมีคุณสมบัติที่สำคัญดังนี้ 1) CPU Intel Xeon E5-2680 v2 2.80 GHz 80 Cores

2) Memory 128 GB

3) Storage 15 GB

3.5.2 ซอร์ฟแวร์ที่ใช้ในงานวิจัยมีรายละเอียดดังนี้

1) R: A language and environment for statistical computing (version 3.4.4)

2) R: Package

ตารางที่ 3.6 Package ที่ใช้ในการทดลอง

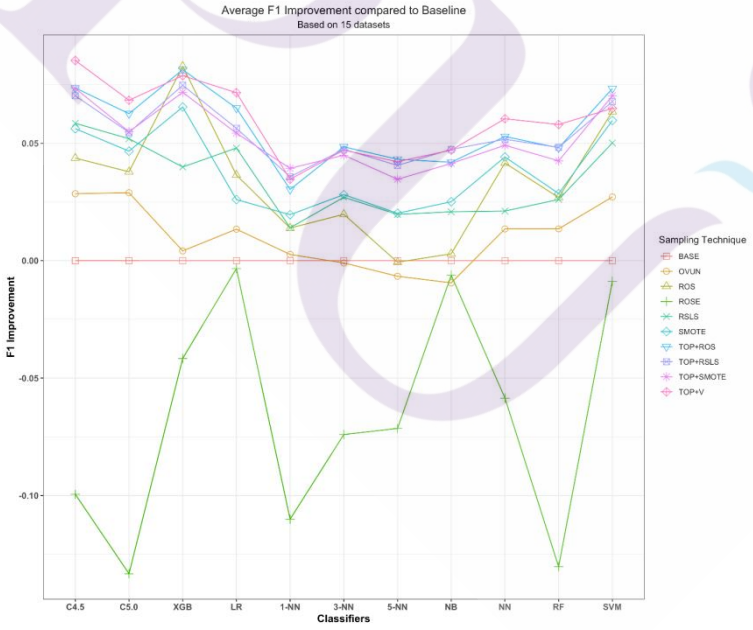
ชื่อ Package	วัตถุประสงค์
methods	เพื่อใช้งานภาษา R ผ่าน Command Line
parallel	เพื่อกระจายการประมวลผลแบบขนาน
dplyr	เพื่อจัดการข้อมูล
data.table	เพื่อจัดการข้อมูล
reshape	เพื่อจัดการข้อมูล
ggplot2	เพื่อวาดกราฟ
ECoL	เพื่อคำนวณค่าทับซ้อน F2, F3
fields	เพื่อคำนวณระยะห่างด้วย Euclidean
MLmetrics	เพื่อคำนวณค่า Confusion Matrix
precrec	เพื่อคำนวณค่า AUPRC
ModelMetrics	เพื่อคำนวณค่า AUROC
smotefamily	เพื่อแก้ไขข้อมูลด้วยวิธีการ SMOTE, RSLS
ROSE	เพื่อแก้ไขข้อมูลด้วยวิธีการ ROSE, OVUN, ROS, RUS, ROUS
RWeka	เพื่อใช้งานโมเดล C4.5
caret	เพื่อใช้งานโมเดลอื่น ๆ ตั้งค่าและประมวลผลการฝึกของโมเดล

บทที่ 4

ผลการวิจัย

ในบทนี้กล่าวถึงผลลัพธ์ของการทดลองประสิทธิภาพในแง่ความถูกต้องแม่นยำของวิธีการแก้ไขข้อมูลเมื่อนำไปใช้งานร่วมกับหลายโมเดลบนชุดข้อมูลจำนวน 15 ชุด ซึ่งผลสรุปจะประกอบด้วยผลเฉลี่ยของอัตราการพัฒนาของโมเดลเมื่อนำวิธีการแก้ไขข้อมูลแบบต่าง ๆ มาใช้ร่วมด้วย และตารางการจัดอับดับของประสิทธิภาพโมเดลโดยเฉลี่ยจากชุดข้อมูลทุกชุดโดยแบ่งออกตามวิธีการแก้ไขข้อมูลแบบต่าง ๆ ผลลัพธ์ทั้งหมดใช้มาตารวัดจำนวน 4 แบบ ประกอบด้วย F1, GM, AUROC, AUPRC

4.1 ผลการทดสอบประสิทธิภาพและการอภิปรายผล



ภาพที่ 4.1 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE)

จากข้อมูลทุกชุดด้วยหน่วยวัดแบบ F1

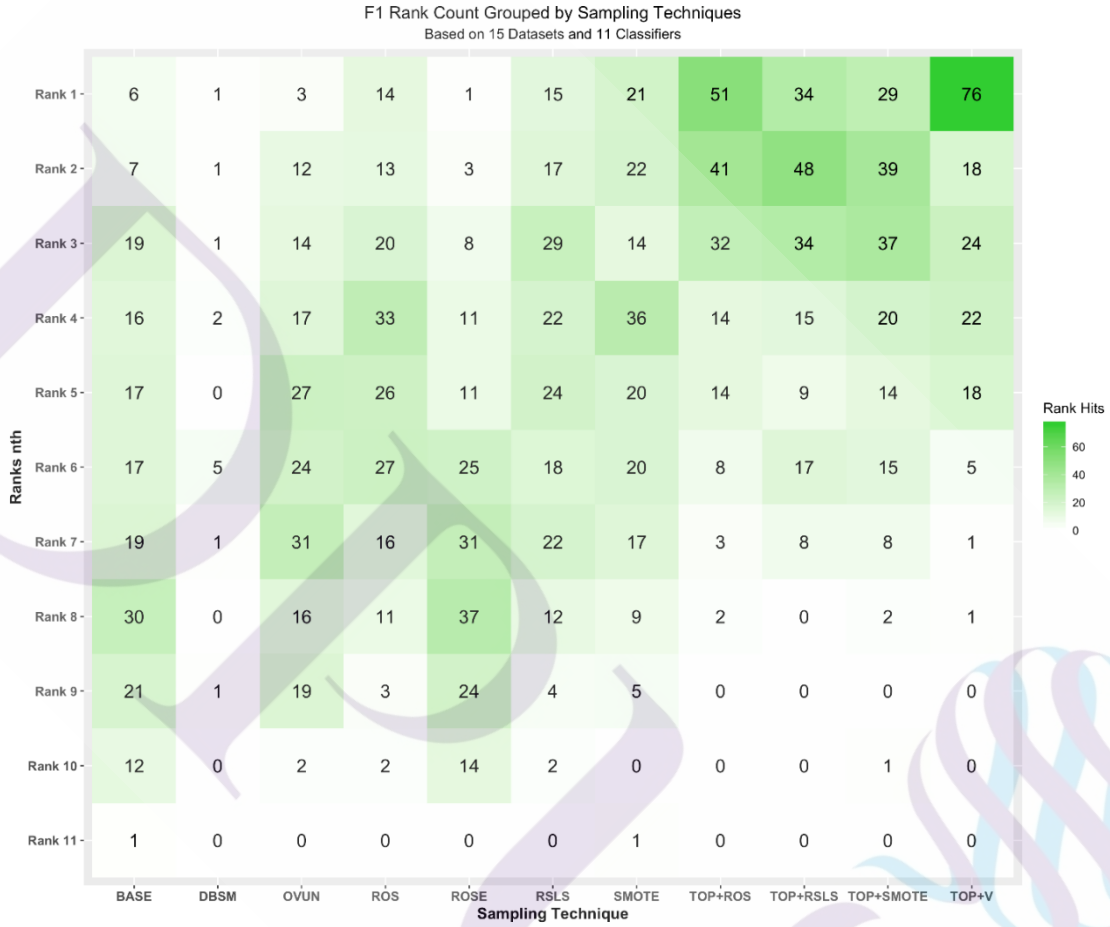
ตารางที่ 4.1 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE) จากข้อมูลทุกชุดด้วยหน่วยวัดแบบ F1

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	.0436±.09	.0378±.07	.0828±0.1	.0364±.08	.0139±.04	.0197±.06	-.0007±.08	.0029±.06	.0416±.04	.0269±.04	.0635±.08
SMOTE	.0562±.08	.0466±0.1	.0654±.07	.0260±.05	.0195±.06	.0281±.06	.0202±.07	.0251±.05	.0441±.03	.0286±.07	.0596±0.1
RSLS	.0585±0.1	.0519±.08	.0398±.05	.0479±.06	.0140±.03	.0270±.04	.0197±.03	.0208±.05	.0211±.04	.0260±.03	.0501±.06
OVUN	.0285±.22	.0289±.21	.0041±.22	.0134±.21	.0026±.23	-.0010±.22	-.0066±0.2	-.0095±.24	.0135±.22	.0135±.21	.0271±.21
ROSE	-.0994±.23	-.1333±.21	-.0416±.22	-.0034±.21	-.1101±.22	-.0740±.21	-.0714±0.2	-.0062±.24	-.0585±.23	-.1303±.21	-.0088±.21
TOP+V	**.0852±0.1***	.0683±0.1	.0787±.07	.0715±.08	.0347±.05	.0470±.08	.0421±.09	.0470±.12	.0604±.06	.0579±.07	.0648±.09
TOP+ ROS	.0735±.29	.0626±.29	.0813±.24	.0649±.21	.0303±.22	.0484±.21	.0432±.21	.0419±.25	.0528±.23	.0480±.28	.0732±.21
TOP+ SMOTE	.0733±.22	.0549±.21	.0716±0.2	.0543±.21	.0393±.22	.0449±.21	.0346±0.2	.0414±.24	.0491±.23	.0425±.21	.0702±.22
TOP+ RSLS	.0702±.23	.0546±.21	.0746±.21	.0563±.21	.0357±.22	.0474±.21	.0405±0.2	.0473±.24	.0517±.23	.0482±.21	.0677±.22

จากตารางที่ 4.1 พบว่า ประสิทธิภาพการแยกแยะข้อมูลที่ได้จากวิธีการแบบ TOP สามารถเพิ่มประสิทธิภาพสูงสุดของโมเดลได้จำนวน 9 แบบเมื่อเทียบกับวิธีแก้ไขข้อมูลแบบอื่น ซึ่งรายระเอียดของประสิทธิภาพที่แบ่งตามชุดข้อมูลนำมาจากจากตาราง 4.1.1 ถึง 4.1.15 (ภาคผนวก) โมเดลที่สามารถพัฒนาประสิทธิภาพเมื่อเทียบกับ BASE ได้มากที่สุดคือ C4.5 เมื่อนำมาใช้ร่วมกับการแก้ไขข้อมูลแบบ TOP+V ซึ่งมีค่าเท่ากับร้อยละ 8.52 ในลำดับถัดมาคือโมเดล XGB ที่ใช้ร่วมกับ ROS มีค่าเท่ากับร้อยละ 8.28 และลำดับที่สามคือ XGB ที่ใช้ร่วมกับ TOP+ROS ซึ่งมีค่าเท่ากับร้อยละ 8.13

นอกเหนือจากนี้ วิธีการแก้ไขข้อมูลแบบ TOP+V, TOP+ROS, TOP+SMOTE, TOP+RSLS สามารถเพิ่มประสิทธิภาพโมเดลเมื่อวัดผลด้วยมาตารวัด F1 ได้มากกว่าวิธีการแก้ไขข้อมูลแบบอื่นทั้งหมด ยกเว้นเพียงแค่ ROS ที่ใช้กับ XBG เท่านั้น และจากการใช้งานวิธีการดังกล่าวร่วมกับโมเดล 3-NN, 5-NN, NB, RF จะเห็นได้ว่าค่าผลลัพธ์ที่ได้นั้นได้สูงกว่าวิธีการอื่นเมื่อเทียบกับการใช้ร่วมกับโมเดลแบบอื่น ๆ ในส่วนของ ROSE นั้น ประสิทธิภาพลดลงเพราะเนื่องจากวิธีการนี้ได้มีการลด

จำนวนข้อมูล Negative ลงไประดับหนึ่งจึงอาจส่งผลให้สูญเสียข้อมูลที่สำคัญไปจึงทำให้ผลการพัฒนาความสามารถของโมเดลเมื่อเทียบกับ BASE นั้นมีค่าน้อยกว่าศูนย์

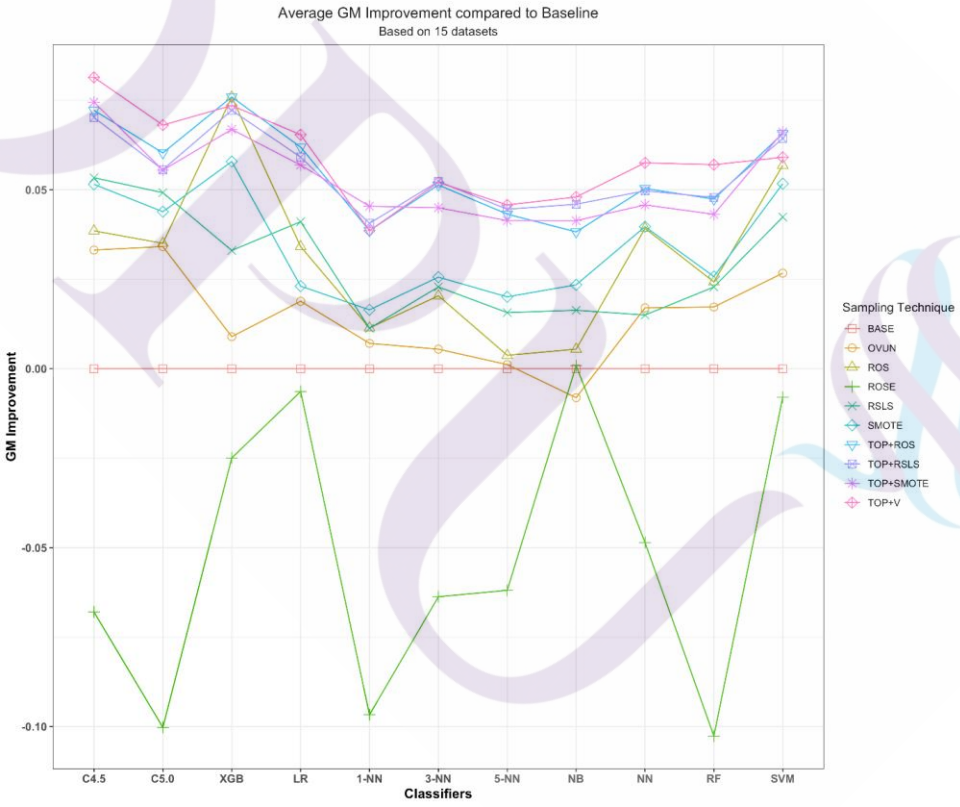


ภาพที่ 4.2 ตารางสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพจากโมเดลทุกแบบและข้อมูลทุกชุดด้วยหน่วยวัดแบบ F1

จากภาพที่ 4.2 จะเห็นได้อย่างชัดเจนว่า ประสิทธิภาพของวิธีการ TOP แบบต่าง ๆ อยู่ลำดับสูงสุดจำนวนมากที่สุด โดยที่ Rank 1 หมายถึงดีที่สุด และ Rank 11 หมายถึงแย่ที่สุด ผลการจัดลำดับ

นำมาจากประสิทธิภาพที่แบ่งตามรายโมเดลมาบวกรวมกันโดยใช้ข้อมูลจากตาราง 4.5.1 ถึง 4.5.15 (ภาคผนวก)

วิธีการที่ชนะมากที่สุดคือ TOP+V เมื่อนำมาใช้กับโมเดลทุกประเภทและเมื่อเทียบกับวิธีการแก้ไขข้อมูลแบบอื่น ๆ โดยจำนวนการชนะมีค่าสูงถึง 76 ครั้ง ในลำดับสองของ Rank 1 คือ TOP+ROS ที่สามารถชนะวิธีการอื่นได้เป็นจำนวน 51 ครั้ง และลำดับที่สามใน Rank เดียวกับคือ TOP+RSLS ที่สามารถเอาชนะได้จำนวน 34 ครั้ง เมื่อพิจารณาโดยการเปรียบเทียบวิธีการแก้ไขข้อมูลทุกแบบพบว่า วิธีการ TOP+V, TOP+ROS, TOP+SMOTE, TOP+RSLS สามารถทำคะแนนนำได้ใน Rank 1, 2, 3 ในขณะที่วิธีการอื่นทำคะแนนนำได้ใน Rank ลำดับรองลงมา



ภาพที่ 4.3 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE) จากข้อมูลทุกชุดด้วยหน่วยวัดแบบ GM

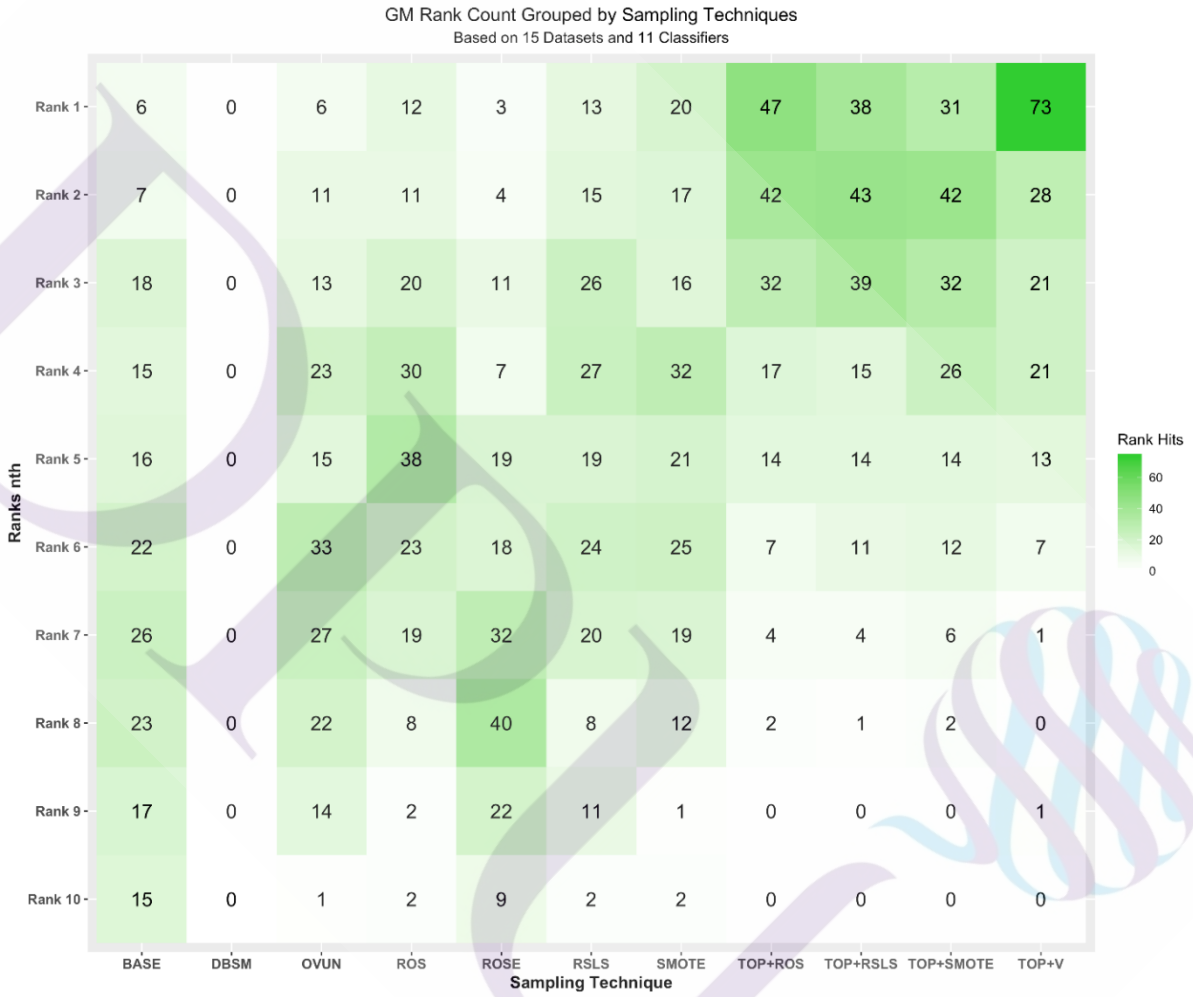
ตารางที่ 4.2 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE) จากข้อมูลทุกชุดด้วยหน่วยวัดแบบ GM

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	.0385±.09	.0350±.07	.0759±0.1	.0341±.07	.0114±.03	.0204±.06	.0037±.07	.0055±.06	.0392±.04	.0242±.04	.0568±.07
SMOTE	.0515±0.1	.0439±.08	.0579±.09	.0231±.06	.0164±.05	.0256±.06	.0201±.06	.0234±.07	.0397±.05	.0257±.03	.0517±.06
RSLS	.0533±.09	.0492±.08	.0330±.04	.0411±.05	.0113±.03	.0228±.04	.0157±.03	.0163±.04	.0150±.04	.0228±.02	.0424±.05
OVUN	.0331±.22	.0341±.21	.0089±.21	.0188±0.2	.0071±.22	.0055±.21	.0011±.19	-.0081±.22	.0170±.22	.0172±.21	.0267±.21
ROSE	-.0680±.22	-.1002±.21	-.0249±.21	-.0064±0.2	-.0967±.21	-.0637±.21	-.0619±0.2	.0008±.22	-.0486±.22	-.1026±.21	-0.008±.21
TOP+V	**.0814±0.1***	.0681±0.1	.0735±.06	.0654±.07	.0386±.05	.0520±.07	.0457±.08	.0480±.12	.0575±.05	.0570±.07	.0591±.08
TOP+ ROS	.0722±.26	.0603±.26	.0759±.22	.0619±0.2	.0386±.21	.0513±.21	.0432±0.2	.0382±.22	.0504±.22	.0474±.25	.0657±0.2
TOP+ SMOTE	.0744±.22	.0555±.21	.0668±.19	.0569±0.2	.0454±.22	.0450±0.2	.0414±0.2	.0414±.22	.0458±.22	.0431±.21	.0661±.21
TOP+ RSLS	.0702±.22	.0556±.21	.0722±0.2	.0591±0.2	.0406±.22	.0523±0.2	.0445±.19	.0460±.22	.0497±.22	.0478±.21	.0643±.21

จากตารางที่ 4.2 พบว่า ประสิทธิภาพการแยกแยะข้อมูลที่ได้จากวิธีการแบบ TOP สามรถเพิ่มประสิทธิภาพสูงสุดของโมเดลได้ทุกแบบ ซึ่งรายระเอียดของประสิทธิภาพที่แบ่งตามชุดข้อมูลได้นำมาจากตาราง 4.2.1 ถึง 4.2.15 (ภาคผนวก) โมเดลที่สามารถพัฒนาได้มากที่สุดคือ C4.5 เมื่อใช้งานร่วมกับ TOP+V โดยมีค่าที่พัฒนาจาก BASE สูงถึงร้อยละ 8.14 ในส่วนของลำดับถัดมานั้นคือวิธีการ XGB แต่ผลลัพธ์นั้นมีค่าเท่ากัน เมื่อใช้วิธีการแก้ไขข้อมูลแบบ TOP+ROS และ ROS โดยมีค่าอยู่ที่ร้อยละ 7.59 ซึ่งต่างจากการที่ใช้มาตรวัดแบบ F1 ที่เมื่อ XGB ทำงานร่วมกับ ROS จะสามารถพัฒนาประสิทธิภาพได้สูงกว่า โดยในลำดับที่สามคือ C4.5 ที่ใช้ข้อมูลจากวิธีการแก้ไขแบบ TOP+ROS โดยได้รับค่าสูงถึงร้อยละ 7.44

เช่นเดียวกันกับมาตารวัดแบบ F1 ประสิทธิภาพของทุกโมเดลได้ถูกพัฒนามากที่สุดด้วยวิธีการ TOP+V, TOP+ROS, TOP+SMOTE, TOP+RSLS เมื่อใช้มาตารวัดแบบ GM โดยมีเพียงหนึ่งกรณีที่เสมอในระดับสูงคือ XGB จับคู่กับ ROS และ XBG จับคู่กับ TOP+ROS นอกเหนือจากนี้ความแตกต่างระหว่างวิธีการ TOP ทุกแบบกับวิธีการอื่นในขณะที่ใช้โมเดล 3-NN, 5-NN, NB, RF นั้นอยู่ระดับที่มาก

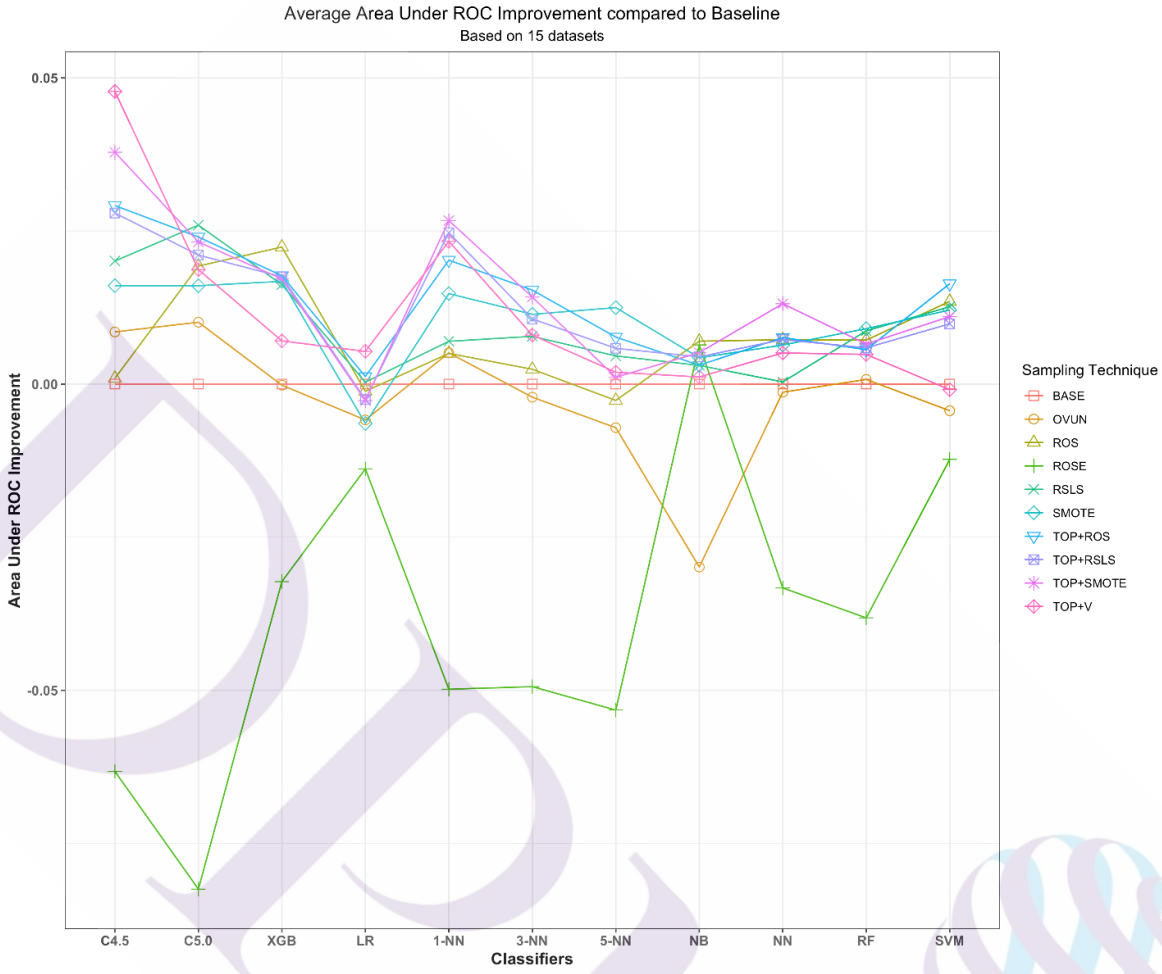
ขึ้นเมื่อเทียบกับ F1 และจะสังเกตได้ว่าเมื่อใช้มาตารวัดแบบ GM จะทำให้ผลของ 1-NN, LR นั้นมีค่าความแตกต่างจากวิธีการอื่นขึ้นชัดเจนมากขึ้น



ภาพที่ 4.4 ตารางสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพจากโมเดลแบบและข้อมูลทุกชุดด้วยหน่วยวัดแบบ GM

จากภาพที่ 4.4 ประสิทธิภาพของวิธีการ TOP แบบต่าง ๆ อยู่ลำดับสูงสุดจำนวนมากที่สุด Rank 1 หมายถึงดีที่สุด และ Rank 11 หมายถึงแย่ที่สุด ซึ่งรายระเอียดของการจัดลำดับประสิทธิภาพได้นำผลจากตาราง 4.6.1 ถึง 4.6.15 (ภาคผนวก) มาบวกรวมกันโดยแบ่งตาม Rank

วิธีการที่ดีที่สุดคือ TOP+V ที่สามารถทำคะแนนอับดับหนึ่งได้มากที่สุดซึ่งมีค่าเท่ากับ 73 ครั้งโดยรวมจากผลการจัดอันดับจากโมเดลทุกแบบเมื่อเทียบกับวิธีการแก้ไขข้อมูลแบบอื่น วิธีการที่สามารถทำคะแนนได้ในอันดับเดียวกันกับแบบแรกคือ TOP+ROS และ TOP+RSLS ที่มีค่าเท่ากับ 47 ครั้ง และ 38 ครั้งตามลำดับ ในส่วนของ Rank 2 วิธีการ TOP ทุกแบบสามารถชนะวิธีการอื่นได้ทั้งหมด อย่างไรก็ดี ใน Rank 3 วิธีการ RSLS สามารถเอาชนะ TOP+V ที่ซึ่งอยู่ในลำดับที่ 4 ของ Rank ดังกล่าว



ภาพที่ 4.5 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE) จากข้อมูลทุกชุดด้วยหน่วยวัด AUROC

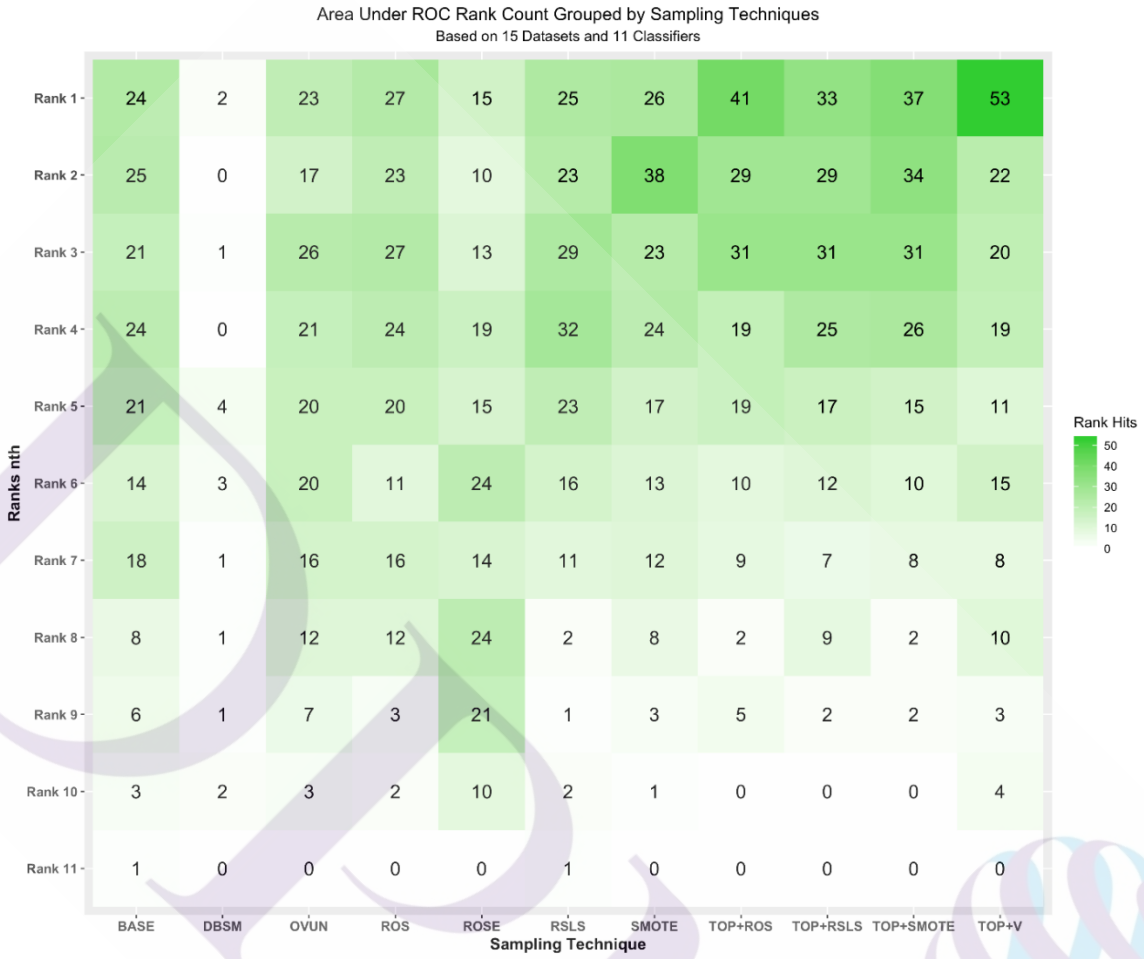
ตารางที่ 4.3 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE) จากข้อมูลทุกชุด

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	.0009±.05	.0193±.04	.0224±.02	-.0012±.02	.0050±.03	.0024±.03	-.0027±.03	.0070±.03	.0073±.03	.0072±.02	.0135±.06
SMOTE	.0161±.04	.0161±.04	.0168±.04	-.0064±.02	.0148±.03	.0114±.04	.0125±.03	.0043±.03	.0064±.03	.0091±.02	.0121±.05
RSLS	.0201±.06	.0260±.04	.0162±.04	.0003±.02	.0070±.02	.0078±.03	.0046±.04	.0031±.03	.0004±.04	.0087±.03	.0126±.02
OVUN	.0085±.15	.0101±.15	-.0002±.14	-.0059±0.2	.0051±.19	-.0021±.19	-.0071±.19	-.0299±.18	-.0013±.15	.0008±.15	-.0043±.18
ROSE	-.0632±.19	-.0825±.15	-.0322±.15	-.0138±.14	-.0498±0.2	-.0494±.19	-.0532±.18	.0064±.18	-.0333±.19	-.0382±.14	-.0123±.15
TOP+V	**.0478±.06***	.0187±.06	.0071±.05	.0054±.03	.0233±.03	.0080±.05	.0019±.06	.0012±.09	.0051±.04	.0048±.05	-.0009±.04
TOP+ ROS	.0292±.23	.0240±.23	.0177±.18	.0012±.15	.0202±0.2	.0154±.21	.0077±.21	.0031±.19	.0076±.21	.0056±.17	.0164±.17
TOP+ SMOTE	.0379±.18	.0232±.15	.0172±.15	-.0027±.14	.0267±.21	.0142±0.2	.0011±.18	.0051±.19	.0131±.19	.0065±.14	.0110±.16
TOP+ RSLS	.0279±0.2	.0211±.15	.0175±.15	-.0025±.14	.0247±0.2	.0106±.19	.0058±.18	.0045±.18	.0073±.19	.0059±.14	.0098±.15

จากตารางที่ 4.3 พบว่า ประสิทธิภาพการแยกแยะข้อมูลที่ได้จากวิธีการแบบ TOP สามรถเพิ่มประสิทธิภาพสูงสุดของโมเดลได้จำนวน 6 แบบ ซึ่งรายระเอียดของประสิทธิภาพที่แบ่งตามชุดข้อมูลนำมาจากตาราง 4.3.1 ถึง 4.3.15 (ภาคผนวก)

โมเดลที่ C4.5 สามารถเพิ่มประสิทธิภาพได้สูงที่สุดเมื่อทำงานร่วมกับ TOP+V โดยมีค่าที่พัฒนาจาก BASE เท่ากับร้อยละ 4.78 ในส่วนของ และในส่วนของลำดับที่สองและสามนั้นอัตราการเพิ่มประสิทธิภาพยังคงเพิ่มมากที่สุดกับโมเดล C4.5 โดยเมื่อจับคู่กับ TOP+SMOTE และ TOP+ROS ที่มีค่าเท่ากับร้อยละ 3.79 และ 2.92 ตามลำดับ

จากภาพ 4.3 สังเกตได้ว่าการเพิ่มขึ้นของประสิทธิภาพของโมเดลด้วยการใช้วิธีการ TOP แบบต่าง ๆ นั้นเพิ่มมากขึ้นเมื่อเทียบกับ BASE โดยที่มีเพียงโมเดล SVM และ LR ที่เมื่อทำงานร่วมกับ TOP บางแบบส่งผลให้ประสิทธิภาพการพัฒนาลดลงเมื่อเทียบกับ BASE

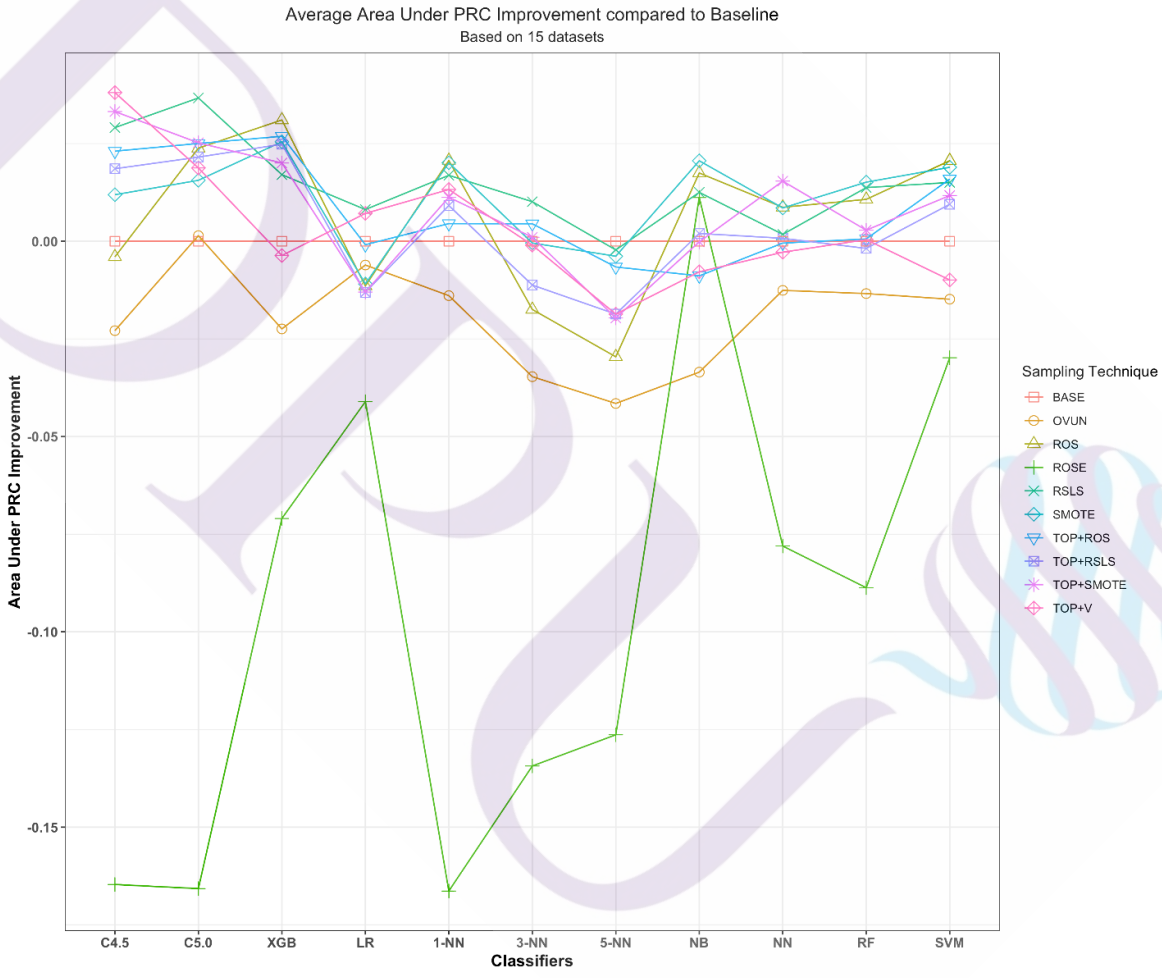


ภาพที่ 4.6 ตารางสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพจากโมเดลทุกแบบและข้อมูลทุกชุดด้วยหน่วยวัด AUROC

จากภาพที่ 4.6 พบว่าประสิทธิภาพของวิธีการ TOP แบบต่าง ๆ อยู่ลำดับสูงสุดจำนวนหลายครั้ง Rank 1 หมายถึงดีที่สุด และ Rank 11 หมายถึงแย่ที่สุด ซึ่งรายระเอียดของประสิทธิภาพที่แบ่งรายโมเดลตามตารางที่ 4.7.1 ถึง 4.7.15 (ภาคผนวก)

วิธีการ TOP+V สามารถทำคะแนนใน Rank 1 ได้เป็นจำนวนครั้งมากที่สุด ซึ่งมีค่าเท่ากับ 53 ครั้ง ลำดับถัดมาคือ TOP+ROS ที่มีจำนวน 41 ครั้ง และลำดับที่สามคือ TOP+SMOTE ที่ทำได้ 37

ครั้ง โดยวิธีการ TOP ทุกแบบสามาทำคะแนนได้มากกว่าวิธีการอื่นทั้งหมดใน Rank 1 ในขณะที่ Rank 2 นั้นวิธีการ TOP แบบต่าง ๆ ทำคะแนนได้น้อยกว่าวิธีการ SMOTE และในขณะเดียวกัน TOP+V นั้นทำคะแนนได้น้อยกว่า BASE, ROS, RSLS อยู่ 3 ครั้ง 1 ครั้ง และ 1 ครั้ง ตามลำดับ ในส่วนของ Rank ที่ 3 วิธีการ TOP แบบต่าง ๆ สามารถทำคะแนนได้มากที่สุด ยกเว้น TOP+V ที่มีจำนวน 20 ครั้ง ซึ่งน้อยกว่าวิธีการ BASE, OVUN, SMOTE, RSLS ที่มีค่าเท่ากับ 21, 26, 29, 23 ตามลำดับ



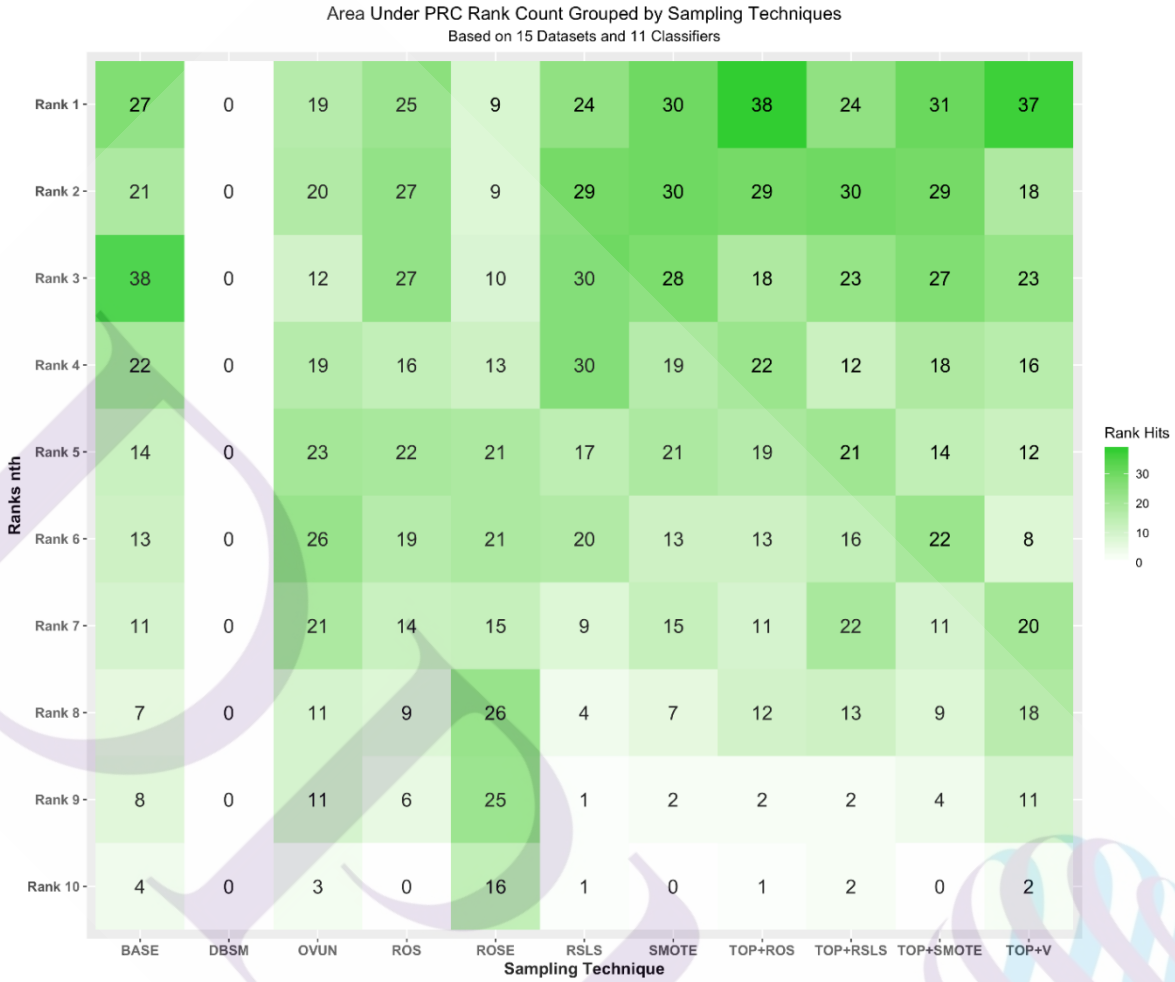
ภาพที่ 4.7 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE) จากข้อมูลทุกชุดด้วยหน่วยวัด AUROC

ตารางที่ 4.4 ค่าเฉลี่ยของประสิทธิภาพที่เปลี่ยนแปลงจากประสิทธิภาพของโมเดลต้นฉบับ (BASE) จากข้อมูลทุกชุดด้วยหน่วยวัด AUROC

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	-0.004±.12	.0238±.05	.0310±.05	-.0114±.06	.0207±.06	-.0175±.06	-.0296±.06	.0175±.04	.0087±.04	.0107±.05	.0206±.04
SMOTE	.0119±.11	.0156±.05	.0255±.06	-.0109±.07	.0201±.08	-.0005±.07	-.0038±.07	.0206±.04	.0085±.05	.0151±.05	.0189±.05
RSLS	.0291±0.1	.0367±.05	.0170±.04	.0081±.03	.0169±.06	.0101±.06	-.0022±.06	.0125±.05	.0018±.04	.0137±.06	.0151±.03
OVUN	-.0229±.27	.0014±.21	-.0224±.21	-.0061±.19	-.0139±.28	-.0347±.27	-.0415±.26	-.0335±.21	-.0126±.24	-.0134±.22	-.0148±.19
ROSE	-.1647±.28	-.1657±.21	-.0710±.22	-.0411±.19	-.1663±.27	-.1343±.28	-.1264±.26	-.0112±.21	-.0780±.25	-.0887±.22	-.0298±.19
TOP+V	**.0381±.12***	.0188±0.1	-.0037±.08	.0071±.05	.0132±.09	-.0010±0.1	-.0187±.11	-.0078±0.1	-.0028±.07	.0004±.09	-.0099±.08
TOP+ ROS	.0231±.31	.0250±.33	.0269±.23	-.0009±.19	.0045±.26	.0045±.26	-.0066±.26	-.0089±.22	-.0005±.26	.0006±.24	.0159±0.2
TOP+ SMOTE	.0332±.25	.0252±.22	.0200±0.2	-.0130±0.2	.0113±.28	.0010±.28	-.0196±.26	-.0002±.22	.0154±.26	.0028±.22	.0117±.21
TOP+ RSLS	.0186±.27	.0215±.21	.0248±.21	-.0131±.19	.0091±.28	-.0112±.27	-.0187±.26	.0020±.22	-.0008±.25	.0018±.22	.0095±0.2

จากตารางที่ 4.4 พบว่าประสิทธิภาพการแยกแยะข้อมูลที่ได้จากวิธีการแบบ TOP สามารถเพิ่มประสิทธิภาพสูงสุดของโมเดลได้จำนวน 1 แบบ ซึ่งรายระเอียดของประสิทธิภาพที่แบ่งตามชุดข้อมูลสามารถดูได้จากตาราง 4.4.1 ถึง 4.4.15 (ภาคผนวก)

จากผลการวัดการเพิ่มขึ้นของประสิทธิภาพของโมเดลด้วยมาตารวัดแบบ AUPRC แสดงให้เห็นว่า C4.5 คือโมเดลที่สามารถเพิ่มประสิทธิภาพได้มากที่สุด โดยเมื่อทำงานร่วมกับ TOP+V ซึ่งมีประสิทธิภาพที่เพิ่มขึ้นร้อยละ 3.81 ลำดับที่สองคือโมเดล C5.0 ที่มีประสิทธิภาพเพิ่มขึ้นร้อยละ 3.67 เมื่อทำงานร่วมกับวิธีการแก้ไขข้อมูลแบบ RSLS และลำดับที่สามคือ C4.5 ที่ทำงานร่วมกับ TOP+SMOTE โดยประสิทธิภาพของโมเดลนี้เพิ่มขึ้นร้อยละ 3.32 เมื่อเทียบกับ BASE



ภาพที่ 4.8 ตารางสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพจากโมเดลทุกแบบและข้อมูลทุกชุดด้วยหน่วยวัด AUPRC

จากภาพที่ 4.8 พบว่าประสิทธิภาพของวิธีการ TOP แบบต่าง ๆ อยู่ลำดับสูงสุดมากกว่าวิธีการอื่น ๆพอสังเขป Rank 1 หมายถึงดีที่สุด และ Rank 11 หมายถึงแย่ที่สุด ซึ่งรายระเอียดของประสิทธิภาพที่นำมาใช้จัดอันดับได้นำมาจากตาราง 4.5.1 ถึง 4.5.15 (ภาคผนวก)

ภาพนี้แสดงให้เห็นถึงผลการจัดอันดับด้วยมาตารวัดแบบ AUPRC โดยสังเกตได้ว่าคะแนนของวิธีการ TOP ในแบบต่าง ๆ สามารถเอาชนะวิธีการอื่นใน Rank 1 ได้ทั้งหมด ยกเว้นเพียงแต่

TOP+RSLS ที่ทำคะแนนได้น้อยกว่า BASE และ ROS อยู่ที่จำนวน 3 และ 1 ครั้ง ในส่วนของ Rank 2 นั้น TOP+V และ SMOTE ทำคะแนนได้สูงที่สุด โดยอยู่ที่ 30 ครั้ง และลำดับถังมาใน Rank เดียวกัน RSLS, TOP+ROS, TOP+SMOTE ทำคะแนนได้เท่ากันคือ 29 ครั้ง ในขณะที่ TOP+V ทำคะแนนได้เพียง 18 ครั้ง

จากผลการทดสอบประสิทธิภาพด้วยมาตรวัด 4 ประเภท พบว่าวิธีการแก้ไขข้อมูล TOP ในแบบต่าง ๆ สามารถเพิ่มประสิทธิภาพให้กับโมเดลได้หลายประเภทในอัตราที่มากกว่าวิธีการแก้ไขข้อมูลแบบอื่นบนข้อมูลหลายชุด นอกจากนี้วิธีการแก้ไขข้อมูลดังกล่าวส่งผลให้โมเดลหลายประเภทมีประสิทธิภาพที่สูงขึ้นทั้งหมด โดยผลลัพธ์การทำงานส่วนมากนั้นอยู่ใน Rank สามอับดับแรกเมื่อตัดสินด้วยมาตรวัดทุกแบบ วิธีการนี้ยังสามารถช่วยให้โมเดลแยกแยะข้อมูลได้ดีขึ้นเมื่อนำข้อมูลที่ขาดความสมดุลที่มีปัญหาความทับซ้อนแฝงอยู่มาใช้ฝึก ดังนั้นจึงสามารถสรุปได้ว่า วิธีการปรับสมดุลข้อมูลแบบ TOP สามารถนำมาใช้เป็นทางเลือกหนึ่งในประสิทธิภาพการทำงานของโมเดลเมื่อฝึกด้วยข้อมูลที่ขาดความสมดุลและทับซ้อนในระดับหนึ่ง

บทที่ 5

สรุปผลและข้อเสนอแนะ

ในบทนี้เป็นบทสรุปที่ได้รับจากการทดสอบงานวิจัย ตลอดจนถึงการอธิบายข้อจำกัดของวิธีการที่พบจากการทดสอบ และข้อเสนอแนะสำหรับแนวทางในการในพัฒนางานวิจัยต่อนี้ต่อไป เพื่อปรับปรุงและแก้ไขข้อบกพร่องของระบบในมีประสิทธิภาพมากยิ่งขึ้น

5.1 สรุปผลการวิจัย

งานวิจัยชิ้นนี้ได้นำเสนอแนวทางการเพิ่มประสิทธิภาพการแยกแยะของตัวแบบ โดยนำหลักการกำหนดขนาดเส้นรอบวงที่คล้ายคลึงกับหลักการจัดกลุ่มโดยอ้างอิงจากความหนาแน่นของข้อมูลมาใช้งานเพื่อจำกัดพื้นที่ของการสุ่ม ผนวกกับแนวทางการสุ่มเพิ่มและลดข้อมูลโดยการเปลี่ยนข้อมูล negative ไปเป็น positive มากไปกว่านั้นงานวิจัยนี้ได้นำวิธีการสุ่มเพิ่มและลดข้อมูลแบบต่าง ๆ มาใช้ในการเพิ่มประสิทธิภาพตัวแบบอีกด้วย ซึ่งจากผลการทดลองพบว่า วิธีการที่นำเสนอนั้นสามารถเพิ่มประสิทธิภาพการคัดแยกข้อมูลให้กับตัวแบบได้จริง โดยที่สามารถเพื่อได้มากที่สุดถึงร้อยละ 8.52 และนอกจากนี้ วิธีการที่นำเสนอยังสามารถช่วยให้โมเดลหลายประเภทมีประสิทธิภาพในการคัดแยกข้อมูลที่ไม่สมดุลและทับซ้อนได้ดียิ่งขึ้น ในขณะที่ตัวอย่างข้อมูล positive ไม่ได้ถูกเพิ่มขึ้นมามากเท่ากับวิธีการปรับสมดุลอื่น

5.2 ข้อจำกัดและแนวทางการแก้ไขของงานวิจัย

ข้อกำจัดและแนวทางแก้ไขของงานวิจัยสามารถแบ่งออกได้เป็นสองข้อดังต่อไปนี้

5.2.1 วิธีการที่นำเสนอในงานวิจัยชิ้นนี้มีสมมติฐานว่า ข้อมูลโดยรอบของข้อมูลที่มีจำนวนน้อยทุกตัวมีความหนาแน่นในลักษณะเดียวกัน ความกว้างของพื้นที่ทั้งสองระดับถูกนำไปใช้กับข้อมูลตัวอย่าง Positive ทั้งหมด ดังนั้นหากข้อมูลโดยรอบมีความหนาแน่นในลักษณะที่แตกต่างกัน อาจจะส่งผลให้วิธีการนี้ทำงานได้อย่างไม่เต็มประสิทธิภาพหรือผิดพลาดได้ เพื่อพัฒนาประสิทธิภาพของวิธีการนี้ให้ดียิ่งขึ้น ระยะความกว้างของชั้นทั้งสองระดับควรยืดหยุ่นได้โดยอ้างอิงจากลักษณะความหนาแน่นโดยรอบของข้อมูลที่มีจำนวนน้อย

5.2.2 เนื่องจากแนวทางที่นำเสนอในงานวิจัยนี้ได้มีการใช้พารามิเตอร์จำนวน 4 ประเภท (ในบางกรณีอาจจะมากกว่าซึ่งขึ้นอยู่กับวิธีการปรับสมดุลที่นำมาใช้ในพื้นที่ภายใน) การค้นหาชุดพารามิเตอร์ที่เหมาะสมที่สุดกับข้อมูลแต่ละชุดนั้นจึงเป็นไปอย่างล่าช้า ส่งผลให้ระยะเวลาที่ใช้ในการการฝึกตัวแบบโดยรวมช้าลงอย่างเห็นได้ชัด ดังนั้นการลดระยะเวลาการค้นหาพารามิเตอร์ที่เหมาะสมจึงจำเป็นอย่างยิ่ง การใช้วิธีการค้นหาด้วยวิธีการวัฒนาการของสิ่งมีชีวิตหรือการเรียนรู้แบบเสริมกำลังอาจจะเป็นวิธีการที่ดีที่จะนำมาใช้เพื่อลดข้อจำกัดนี้

5.2.3 การปรับค่าพารามิเตอร์ของวิธีการที่นำเสนอสามารถใช้ระยะเวลานานในการค้นหา อย่างไรก็ดีค่าที่แนะนำสำหรับการเริ่มต้นนั้นไม่ควรกำหนดสูงจนเกินไป ค่าระยะห่างของพื้นที่ภายในและภายนอกสามารถเริ่มจาก 0.10 แล้วจึงเพิ่มระดับขึ้นภายหลัง เพราะข้อมูลอาจจะถูกแก้ไขมากเกินไป ค่าของการสุ่มลบข้อมูล dnegative สามารถเริ่มที่ 0.10 จากน้อยไปหามากเช่นกัน เนื่องจากจะทำให้ข้อมูลถูกลบมากเกินความจำเป็น ส่งผลให้สูญเสียข้อมูลที่สำคัญ

บรรณานุกรม

บรรณานุกรม

A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. Souto, and T. K. Ho, "How Complex is your classification problem? A survey on measuring classification complexity," arXiv preprint arXiv:1808.03591, 2018.

A. C. Schierz, "Virtual screening of bioassay data," *J. Cheminform.*, vol. 1, no. 1, pp. 231–235, 2009.

C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Folleco, "An Empirical Study of the Classification Performance of Learners on Imbalanced and Noisy Software Quality Data," in *2007 IEEE International Conference on Information Reuse and Integration*, 2007, pp. 651–658.

G. Hoang, A. Bouzerdoum, and S. Lam, "Learning Pattern Classification Tasks with Imbalanced Data Sets," Pattern Recognit., pp. 193–208, 2009.

I. H. Witten, E. Frank, and M. A. Hall, *Data mining*. 2011.

J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 233–240.

J. M. Lobo, A. Jiménez-valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, no. 2. pp. 145–151, 2008.

J. Keilwagen, I. Grosse, and J. Grau, "Area under precision-recall curves for weighted and unweighted data," *PLoS One*, vol. 9, no. 3, 2014.

J. R. Quinlan, *C4.5: Programs for Machine Learning*. 1992.

J. S. Akosa, "Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data," *SAS Glob. Forum*, 2017.

K. H. Brodersen, C. S. Ong, K. E. Stephany, and J. M. Buhmann, “The binormal assumption on precision-recall curves,” in *Proceedings - International Conference on Pattern Recognition*, 2010, pp. 4263–4266.

K. Kerdprasop and N. Kerdprasop, “A data mining approach to automate fault detection model development in the semiconductor manufacturing process,” Int. J. Mech., 2011.

M. N. Wright and A. Ziegler, “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R,” *J. Stat. Softw.*, 2017.

M. Vuk, “ROC Curve , Lift Chart and Calibration Plot,” *Metod. Zv.*, vol. 3, no. 1, pp. 89–108, 2006.

N. Lunardon, G. Menardi, and N. Torelli, “ROSE : A Package for Binary Imbalanced Learning,” *R J.*, vol. 6, no. June, pp. 79–89, 2014.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

R. Pandya and J. Pandya, “C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning,” *Int. J. Comput. Appl.*, 2015.

S. Cateni, V. Colla, and M. Vannucci, “A method for resampling imbalanced datasets in binary classification tasks for real-world problems,” Neurocomputing, vol. 135, pp. 32–41, 2014.

S. Maldonado, R. Weber, and F. Famili, “Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines,” *Inf. Sci. (Ny).*, vol. 286, pp. 228–246, 2014.

S. Visa and A. Ralescu, “Issues in mining imbalanced data sets-a review paper,” in Proceedings of the sixteen midwest artificial intelligence and cognitive science conference, 2005, no. August 2016, pp. 67–73.

T. Bäck and H.-P. Schwefel, “An Overview of Evolutionary Algorithms for Parameter Optimization,” *Evol. Comput.*, 1993.

T. Chen and C. Guestrin, “XGBoost : Reliable Large-scale Tree Boosting System,” *arXiv*, 2016.

T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLoS One, vol. 10, no. 3, 2015.

V. García, J. S. Sánchez, and R. A. Mollineda, “On the effectiveness of preprocessing methods when dealing with different levels of class imbalance,” *Knowledge-Based Systems*, vol. 25, no. 1. pp. 13–21, 2012.

W. Siriseriwan and K. Sinapiromsaran, “The effective redistribution for imbalance dataset: Relocating safe-eevel SMOTE with minority outcast handling,” *Chiang Mai J. Sci.*, vol. 43, no. 1, pp. 234–246, 2016.

X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, “On the class imbalance problem,” in *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, 2008, vol. 4, pp. 192–201.

Y. Sanguanmak and A. Hanskunatai, “DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification,” in *2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016*, 2016.

ภาคผนวก

ภาคผนวก ก

7.1 ผลการทดสอบประสิทธิภาพด้วยหน่วยวัดแบบ F1

ตารางที่ 7.1.1 ผลการทดสอบด้วยข้อมูลชุดที่ 1

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.786	0.825	0.735	0.666	0.832	0.859	0.894	0.469	0.690	0.815	0.859
ROS	0.788	0.809	0.778	0.720	0.832	0.730	0.713	0.390	0.782	0.824	0.846
SMOTE	0.763	0.800	0.755	0.747	0.773	0.763	0.774	0.514	0.768	0.836	0.886
RSLS	0.758	0.820	0.755	0.741	0.832	0.858	0.873	0.569	0.738	0.821	0.892
OVUN	0.812	0.801	0.723	0.732	0.836	0.778	0.865	0.434	0.725	0.852	0.886
ROSE	0.275	0.275	0.428	0.737	0.633	0.677	0.673	0.342	0.728	0.311	0.769
DBSM	0.767	-	-	-	-	-	-	-	-	-	-
TOP+V	0.811	0.877	0.794	0.728	0.832	0.870	0.894	0.469	0.759	0.875	0.905
TOP+ROSE	0.832	0.877	0.785	0.738	0.846	0.870	0.894	0.469	0.746	0.850	0.905
TOP+SMOTE	0.811	0.877	0.787	0.730	0.846	0.870	0.894	0.522	0.748	0.857	0.894
TOP+RSLS	0.811	0.877	0.816	0.743	0.846	0.870	0.894	0.617	0.769	0.868	0.894

ตารางที่ 7.1.2 ผลการทดสอบด้วยข้อมูลชุดที่ 2

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.704	0.747	0.670	0.595	0.711	0.663	0.692	0.622	0.634	0.787	0.688
ROS	0.721	0.786	0.763	0.638	0.711	0.727	0.708	0.635	0.692	0.780	0.743
SMOTE	0.675	0.768	0.691	0.622	0.718	0.709	0.717	0.622	0.650	0.807	0.724
RSLS	0.708	0.781	0.711	0.614	0.714	0.685	0.691	0.620	0.659	0.814	0.738
OVUN	0.716	0.763	0.689	0.606	0.700	0.687	0.686	0.625	0.643	0.775	0.684
ROSE	0.644	0.639	0.631	0.640	0.607	0.631	0.623	0.636	0.637	0.636	0.641
DBSM	0.711	-	-	-	-	-	-	-	-	-	-
TOP+V	0.759	0.806	0.747	0.713	0.745	0.748	0.743	0.661	0.714	0.831	0.743
TOP+ROSE	0.743	0.763	0.741	0.652	0.711	0.716	0.732	0.658	0.692	0.819	0.704
TOP+SMOTE	0.724	0.747	0.730	0.659	0.711	0.716	0.732	0.658	0.695	0.799	0.712
TOP+RSLS	0.724	0.747	0.741	0.659	0.711	0.716	0.732	0.658	0.692	0.806	0.710

ตารางที่ 7.1.3 ผลการทดสอบด้วยข้อมูลชุดที่ 3

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.608	0.679	0.590	0.435	0.783	0.709	0.716	0.612	0.663	0.731	0.602
ROS	0.621	0.720	0.690	0.575	0.778	0.751	0.732	0.606	0.672	0.771	0.612
SMOTE	0.571	0.679	0.601	0.435	0.783	0.709	0.716	0.612	0.635	0.760	0.596
RSLS	0.673	0.782	0.654	0.576	0.783	0.755	0.760	0.617	0.649	0.778	0.634
OVUN	0.642	0.760	0.629	0.524	0.787	0.731	0.746	0.597	0.669	0.745	0.611
ROSE	0.578	0.567	0.628	0.442	0.625	0.646	0.614	0.594	0.647	0.670	0.588
DBSM	0.710	-	-	-	-	-	-	-	-	-	-
TOP+V	0.748	0.798	0.693	0.621	0.819	0.757	0.771	0.628	0.726	0.805	0.638
TOP+ROSE	0.678	0.769	0.662	0.612	0.792	0.776	0.752	0.620	0.683	0.768	0.658
TOP+SMOTE	0.698	0.757	0.650	0.600	0.792	0.772	0.744	0.620	0.703	0.772	0.675
TOP+RSLS	0.662	0.750	0.646	0.605	0.797	0.772	0.770	0.618	0.703	0.771	0.648

ตารางที่ 7.1.4 ผลการทดสอบด้วยข้อมูลชุดที่ 4

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.766	0.868	0.858	0.850	0.864	0.805	0.810	0.844	0.825	0.886	0.800
ROS	0.763	0.883	0.869	0.773	0.864	0.782	0.740	0.808	0.867	0.886	0.867
SMOTE	0.789	0.912	0.850	0.797	0.886	0.859	0.812	0.818	0.883	0.891	0.886
RSLS	0.824	0.896	0.883	0.830	0.864	0.805	0.782	0.775	0.810	0.891	0.800
OVUN	0.797	0.864	0.844	0.850	0.864	0.800	0.782	0.796	0.859	0.872	0.800
ROSE	0.828	0.861	0.872	0.859	0.791	0.849	0.847	0.790	0.839	0.848	0.775
DBSM	0.829	-	-	-	-	-	-	-	-	-	-
TOP+V	0.830	0.868	0.911	0.850	0.864	0.805	0.810	0.844	0.892	0.926	0.830
TOP+ROSE	0.870	0.873	0.911	0.898	0.864	0.820	0.810	0.844	0.892	0.906	0.866
TOP+SMOTE	0.839	0.898	0.891	0.850	0.897	0.850	0.810	0.844	0.892	0.926	0.860
TOP+RSLS	0.829	0.898	0.901	0.850	0.864	0.820	0.810	0.844	0.892	0.926	0.860

ตารางที่ 7.1.5 ผลการทดสอบด้วยข้อมูลชุดที่ 5

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.170	0.190	0.341	0.242	0.410	0.344	0.308	0.289	0.337	0.287	0.223
ROS	0.458	0.466	0.512	0.447	0.438	0.459	0.452	0.385	0.425	0.355	0.496
SMOTE	0.498	0.471	0.495	0.432	0.408	0.411	0.400	0.361	0.469	0.355	0.438
RSLS	0.466	0.456	0.460	0.424	0.399	0.399	0.345	0.351	0.389	0.337	0.446
OVUN	0.497	0.500	0.507	0.457	0.453	0.466	0.473	0.431	0.502	0.483	0.488
ROSE	0.475	0.486	0.478	0.417	0.366	0.391	0.395	0.368	0.433	0.430	0.396
DBSM	0.467	-	-	-	-	-	-	-	-	-	-
TOP+V	0.484	0.471	0.497	0.492	0.456	0.457	0.469	0.455	0.478	0.455	0.473
TOP+ROSE	0.509	0.502	0.485	0.462	0.465	0.481	0.435	0.437	0.480	0.474	0.481
TOP+SMOTE	0.452	0.453	0.465	0.427	0.469	0.457	0.441	0.423	0.464	0.465	0.443
TOP+RSLS	0.493	0.453	0.474	0.427	0.469	0.480	0.494	0.423	0.463	0.471	0.445

ตารางที่ 7.1.6 ผลการทดสอบด้วยข้อมูลชุดที่ 6

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.700	0.746	0.739	0.696	0.664	0.678	0.703	0.352	0.686	0.763	0.657
ROS	0.682	0.740	0.785	0.692	0.708	0.680	0.686	0.412	0.746	0.774	0.720
SMOTE	0.746	0.746	0.774	0.733	0.681	0.692	0.735	0.501	0.729	0.759	0.721
RSLS	0.736	0.746	0.772	0.733	0.681	0.692	0.735	0.501	0.672	0.784	0.725
OVUN	0.581	0.627	0.641	0.614	0.568	0.575	0.596	0.615	0.627	0.621	0.584
ROSE	0.564	0.552	0.567	0.556	0.545	0.538	0.526	0.584	0.549	0.543	0.570
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.645	0.662	0.693	0.653	0.614	0.624	0.628	0.621	0.644	0.686	0.630
TOP+ROSE	0.606	0.663	0.679	0.631	0.621	0.618	0.616	0.608	0.628	0.678	0.628
TOP+SMOTE	0.629	0.649	0.660	0.628	0.623	0.608	0.608	0.618	0.643	0.656	0.620
TOP+RSLS	0.609	0.649	0.660	0.639	0.623	0.618	0.608	0.619	0.641	0.678	0.623

ตารางที่ 7.1.7 ผลการทดสอบด้วยข้อมูลชุดที่ 7

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.870	0.920	0.868	0.940	0.930	0.871	0.849	0.898	0.989	0.946	0.879
ROS	0.909	0.920	0.943	0.940	0.930	0.890	0.818	0.868	0.989	0.923	0.926
SMOTE	0.927	0.941	0.929	0.960	0.950	0.905	0.859	0.888	0.975	0.955	0.922
RSLS	0.835	0.888	0.896	0.940	0.930	0.857	0.877	0.887	0.989	0.951	0.872
OVUN	0.862	0.901	0.871	0.940	0.930	0.905	0.869	0.907	0.989	0.966	0.873
ROSE	0.870	0.873	0.923	0.920	0.698	0.737	0.757	0.813	0.860	0.923	0.896
DBSM	0.913	-	-	-	-	-	-	-	-	-	-
TOP+V	0.932	0.960	0.966	0.975	0.950	0.908	0.890	0.898	0.989	0.986	0.916
TOP+ROSE	0.916	0.949	0.946	0.975	0.939	0.919	0.906	0.903	0.989	0.969	0.955
TOP+SMOTE	0.963	0.963	0.927	0.949	0.930	0.919	0.854	0.923	0.989	0.955	0.921
TOP+RSLS	0.940	0.960	0.927	0.949	0.939	0.919	0.854	0.923	0.989	0.966	0.912

ตารางที่ 7.1.8 ผลการทดสอบด้วยข้อมูลชุดที่ 8

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.848	0.847	0.844	0.909	0.910	0.876	0.820	0.927	0.969	0.916	0.913
ROS	0.852	0.849	0.866	0.909	0.910	0.914	0.883	0.875	0.975	0.916	0.891
SMOTE	0.863	0.863	0.905	0.909	0.925	0.896	0.891	0.878	0.975	0.925	0.911
RSLS	0.840	0.882	0.902	0.909	0.910	0.882	0.822	0.913	0.989	0.930	0.907
OVUN	0.836	0.843	0.867	0.909	0.910	0.895	0.887	0.902	0.975	0.919	0.867
ROSE	0.838	0.801	0.897	0.900	0.696	0.743	0.771	0.847	0.850	0.866	0.876
DBSM	0.951	-	-	-	-	-	-	-	-	-	-
TOP+V	0.902	0.941	0.920	0.943	0.910	0.910	0.865	0.927	0.989	0.989	0.944
TOP+ROSE	0.871	0.897	0.907	0.929	0.925	0.921	0.910	0.927	0.989	0.939	0.941
TOP+SMOTE	0.886	0.897	0.913	0.958	0.963	0.901	0.867	0.927	0.989	0.939	0.933
TOP+RSLS	0.886	0.897	0.913	0.958	0.960	0.926	0.867	0.927	0.989	0.939	0.933

ตารางที่ 7.1.9 ผลการทดสอบด้วยข้อมูลชุดที่ 9

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.769	0.811	0.602	0.594	0.667	0.618	0.611	0.468	0.802	0.860	0.739
ROS	0.967	0.910	0.967	0.627	0.798	0.701	0.612	0.575	0.867	0.950	0.907
SMOTE	0.967	0.932	0.967	0.549	0.867	0.778	0.686	0.675	0.932	0.917	0.907
RSLS	0.967	0.967	0.553	0.645	0.798	0.754	0.666	0.496	0.799	0.933	0.797
OVUN	0.908	0.967	0.540	0.583	0.798	0.764	0.666	0.496	0.865	0.919	0.917
ROSE	0.318	0.328	0.401	0.544	0.528	0.418	0.441	0.573	0.423	0.438	0.690
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.967	0.986	0.834	0.745	0.837	0.786	0.725	0.518	0.932	1.000	0.921
TOP+ROSE	0.967	0.967	0.967	0.713	0.812	0.786	0.730	0.559	0.907	1.000	0.936
TOP+SMOTE	0.967	0.967	0.967	0.713	0.856	0.786	0.728	0.496	0.881	0.967	0.936
TOP+RSLS	0.967	0.967	0.967	0.713	0.819	0.786	0.728	0.496	0.907	0.980	0.936

ตารางที่ 7.1.10 ผลการทดสอบด้วยข้อมูลชุดที่ 10

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.578	0.568	0.559	0.636	0.515	0.513	0.519	0.606	0.503	0.569	0.619
ROS	0.634	0.588	0.680	0.661	0.543	0.557	0.573	0.652	0.560	0.645	0.654
SMOTE	0.599	0.643	0.633	0.637	0.545	0.530	0.546	0.623	0.536	0.641	0.634
RSLS	0.600	0.643	0.621	0.637	0.545	0.530	0.546	0.623	0.476	0.632	0.631
OVUN	0.620	0.634	0.660	0.665	0.559	0.563	0.576	0.626	0.533	0.641	0.655
ROSE	0.611	0.627	0.642	0.640	0.559	0.564	0.569	0.627	0.476	0.642	0.629
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.672	0.670	0.670	0.674	0.578	0.584	0.601	0.660	0.583	0.670	0.664
TOP+ROSE	0.656	0.669	0.686	0.672	0.593	0.594	0.621	0.656	0.579	0.661	0.665
TOP+SMOTE	0.658	0.653	0.684	0.668	0.592	0.593	0.619	0.660	0.579	0.670	0.677
TOP+RSLS	0.658	0.655	0.686	0.670	0.593	0.594	0.621	0.656	0.579	0.666	0.666

ตารางที่ 7.1.11 ผลการทดสอบด้วยข้อมูลชุดที่ 11

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.405	0.522	0.440	0.587	0.423	0.438	0.454	0.513	0.528	0.501	0.464
ROS	0.492	0.580	0.619	0.664	0.415	0.501	0.514	0.517	0.597	0.595	0.627
SMOTE	0.561	0.609	0.586	0.644	0.437	0.505	0.522	0.520	0.601	0.561	0.605
RSLS	0.605	0.596	0.589	0.660	0.436	0.498	0.510	0.514	0.622	0.547	0.596
OVUN	0.452	0.576	0.501	0.592	0.428	0.462	0.484	0.510	0.605	0.551	0.513
ROSE	0.492	0.565	0.589	0.529	0.426	0.468	0.500	0.509	0.518	0.573	0.559
DBSM	0.581	-	-	-	-	-	-	-	-	-	-
TOP+V	0.599	0.619	0.582	0.639	0.551	0.554	0.540	0.513	0.624	0.624	0.580
TOP+ROSE	0.560	0.640	0.589	0.653	0.527	0.542	0.536	0.532	0.642	0.642	0.631
TOP+SMOTE	0.567	0.619	0.553	0.637	0.535	0.534	0.536	0.517	0.640	0.612	0.651
TOP+RSLS	0.576	0.619	0.540	0.637	0.540	0.538	0.536	0.517	0.613	0.620	0.659

ตารางที่ 7.1.12 ผลการทดสอบด้วยข้อมูลชุดที่ 12

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.922	0.963	0.936	0.941	0.854	0.839	0.815	0.593	0.964	0.977	0.936
ROS	0.940	0.973	0.888	0.935	0.850	0.815	0.790	0.651	0.908	0.981	0.949
SMOTE	0.922	0.975	0.894	0.938	0.860	0.845	0.825	0.632	0.952	0.968	0.950
RSLS	0.941	0.966	0.892	0.942	0.857	0.845	0.828	0.621	0.937	0.966	0.956
OVUN	0.923	0.960	0.868	0.924	0.768	0.695	0.640	0.262	0.892	0.964	0.949
ROSE	0.746	0.784	0.847	0.799	0.567	0.577	0.546	0.634	0.656	0.847	0.836
DBSM	0.919	-	-	-	-	-	-	-	-	-	-
TOP+V	0.938	0.979	0.940	0.948	0.854	0.841	0.815	0.699	0.964	0.981	0.944
TOP+ROSE	0.935	0.977	0.936	0.941	0.854	0.852	0.838	0.595	0.964	0.977	0.942
TOP+SMOTE	0.939	0.977	0.937	0.943	0.854	0.842	0.815	0.593	0.943	0.979	0.955
TOP+RSLS	0.932	0.977	0.937	0.943	0.857	0.845	0.815	0.593	0.964	0.977	0.952

ตารางที่ 7.1.13 ผลการทดสอบด้วยข้อมูลชุดที่ 13

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.945	0.954	0.952	0.954	0.942	0.956	0.960	0.945	0.947	0.963	0.942
ROS	0.943	0.954	0.946	0.956	0.938	0.957	0.958	0.948	0.947	0.953	0.946
SMOTE	0.945	0.954	0.948	0.954	0.942	0.956	0.960	0.945	0.948	0.959	0.942
RSLS	0.931	0.950	0.946	0.961	0.947	0.963	0.963	0.948	0.954	0.961	0.950
OVUN	0.932	0.961	0.953	0.959	0.943	0.955	0.957	0.944	0.942	0.953	0.944
ROSE	0.936	0.943	0.957	0.954	0.929	0.949	0.945	0.942	0.950	0.951	0.945
DBSM	0.931	-	-	-	-	-	-	-	-	-	-
TOP+V	0.951	0.957	0.960	0.961	0.945	0.958	0.962	0.952	0.954	0.965	0.945
TOP+ROSE	0.945	0.963	0.959	0.959	0.962	0.964	0.964	0.950	0.963	0.965	0.948
TOP+SMOTE	0.945	0.956	0.953	0.956	0.957	0.962	0.962	0.958	0.953	0.963	0.951
TOP+RSLS	0.945	0.958	0.953	0.956	0.958	0.966	0.962	0.958	0.955	0.965	0.953

ตารางที่ 7.1.14 ผลการทดสอบด้วยข้อมูลชุดที่ 14

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.504	0.551	0.533	0.446	0.492	0.535	0.499	0.496	0.521	0.553	0.471
ROS	0.528	0.563	0.597	0.582	0.490	0.569	0.562	0.486	0.597	0.563	0.584
SMOTE	0.595	0.596	0.608	0.584	0.505	0.589	0.573	0.490	0.612	0.615	0.580
RSLS	0.571	0.581	0.612	0.579	0.509	0.562	0.546	0.493	0.603	0.589	0.584
OVUN	0.501	0.495	0.475	0.476	0.500	0.492	0.487	0.457	0.482	0.508	0.511
ROSE	0.452	0.452	0.452	0.583	0.524	0.555	0.568	0.464	0.584	0.453	0.575
DBSM	0.573	-	-	-	-	-	-	-	-	-	-
TOP+V	0.585	0.595	0.601	0.591	0.539	0.564	0.560	0.496	0.600	0.592	0.586
TOP+ROSE	0.573	0.605	0.606	0.592	0.520	0.566	0.554	0.503	0.605	0.592	0.591
TOP+SMOTE	0.579	0.586	0.593	0.559	0.539	0.561	0.559	0.496	0.597	0.592	0.583
TOP+RSLS	0.579	0.587	0.593	0.559	0.536	0.561	0.567	0.496	0.598	0.598	0.583

ตารางที่ 7.1.15 ผลการทดสอบด้วยข้อมูลชุดที่ 15

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.768	0.759	0.765	0.738	0.649	0.696	0.749	0.429	0.697	0.746	0.749
ROS	0.700	0.776	0.773	0.655	0.649	0.664	0.647	0.301	0.752	0.784	0.727
SMOTE	0.766	0.759	0.781	0.676	0.659	0.674	0.686	0.361	0.753	0.779	0.737
RSLS	0.765	0.776	0.784	0.756	0.651	0.722	0.752	0.447	0.785	0.755	0.769
OVUN	0.691	0.730	0.728	0.599	0.642	0.618	0.584	0.320	0.649	0.732	0.668
ROSE	0.225	0.199	0.497	0.658	0.499	0.546	0.553	0.250	0.728	0.212	0.668
DBSM	0.705	-	-	-	-	-	-	-	-	-	-
TOP+V	0.798	0.784	0.808	0.769	0.671	0.739	0.756	0.429	0.813	0.783	0.796
TOP+ROSE	0.786	0.777	0.796	0.775	0.671	0.702	0.749	0.431	0.788	0.780	0.790
TOP+SMOTE	0.786	0.776	0.799	0.764	0.671	0.702	0.749	0.429	0.775	0.785	0.786
TOP+RSLS	0.786	0.776	0.800	0.764	0.671	0.702	0.749	0.429	0.775	0.790	0.786

7.2 ผลการทดสอบประสิทธิภาพด้วยหน่วยวัดแบบ GM

ตารางที่ 7.2.1 ผลการทดสอบด้วยข้อมูลชุดที่ 1

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.797	0.831	0.748	0.684	0.836	0.865	0.898	0.543	0.713	0.826	0.865
ROS	0.794	0.813	0.788	0.739	0.836	0.745	0.733	0.481	0.794	0.831	0.850
SMOTE	0.773	0.804	0.766	0.762	0.782	0.774	0.785	0.578	0.781	0.840	0.889
RSLS	0.767	0.823	0.766	0.748	0.836	0.861	0.876	0.621	0.746	0.830	0.895
OVUN	0.819	0.807	0.739	0.742	0.842	0.788	0.868	0.515	0.737	0.858	0.889
ROSE	0.399	0.399	0.515	0.755	0.666	0.704	0.701	0.447	0.744	0.425	0.781
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.821	0.881	0.805	0.733	0.836	0.875	0.898	0.543	0.765	0.880	0.908
TOP+ROSE	0.843	0.881	0.796	0.745	0.851	0.875	0.898	0.543	0.755	0.858	0.908
TOP+SMOTE	0.821	0.881	0.795	0.738	0.851	0.875	0.898	0.584	0.765	0.862	0.897
TOP+RSLS	0.821	0.881	0.823	0.760	0.851	0.875	0.898	0.659	0.782	0.873	0.897

ตารางที่ 7.2.2 ผลการทดสอบด้วยข้อมูลชุดที่ 2

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.714	0.755	0.678	0.604	0.717	0.672	0.694	0.654	0.639	0.794	0.696
ROS	0.732	0.793	0.774	0.648	0.717	0.739	0.722	0.669	0.701	0.789	0.754
SMOTE	0.684	0.772	0.706	0.630	0.725	0.716	0.729	0.654	0.663	0.812	0.733
RSLS	0.717	0.786	0.720	0.622	0.721	0.690	0.700	0.653	0.671	0.819	0.745
OVUN	0.725	0.774	0.699	0.615	0.705	0.692	0.689	0.656	0.649	0.781	0.690
ROSE	0.682	0.680	0.676	0.658	0.648	0.679	0.673	0.678	0.679	0.678	0.680
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.773	0.815	0.757	0.727	0.753	0.757	0.751	0.700	0.734	0.836	0.751
TOP+ROSE	0.754	0.772	0.752	0.667	0.717	0.739	0.748	0.698	0.708	0.822	0.713
TOP+SMOTE	0.736	0.755	0.741	0.675	0.717	0.739	0.748	0.698	0.711	0.806	0.722
TOP+RSLS	0.736	0.755	0.754	0.675	0.717	0.739	0.748	0.698	0.708	0.812	0.721

ตารางที่ 7.2.3 ผลการทดสอบด้วยข้อมูลชุดที่ 3

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.629	0.693	0.611	0.460	0.789	0.718	0.724	0.645	0.678	0.742	0.611
ROS	0.628	0.728	0.699	0.592	0.787	0.756	0.740	0.646	0.690	0.782	0.619
SMOTE	0.578	0.693	0.617	0.460	0.789	0.718	0.724	0.645	0.651	0.776	0.609
RSLS	0.676	0.785	0.661	0.590	0.787	0.759	0.765	0.654	0.657	0.787	0.639
OVUN	0.657	0.767	0.641	0.541	0.793	0.741	0.754	0.636	0.688	0.751	0.615
ROSE	0.604	0.582	0.648	0.447	0.642	0.666	0.635	0.638	0.665	0.682	0.606
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.751	0.804	0.707	0.640	0.824	0.762	0.775	0.666	0.738	0.810	0.654
TOP+ROSE	0.682	0.776	0.668	0.644	0.798	0.781	0.759	0.659	0.703	0.775	0.669
TOP+SMOTE	0.711	0.766	0.657	0.639	0.802	0.777	0.755	0.659	0.714	0.774	0.683
TOP+RSLS	0.674	0.756	0.682	0.643	0.798	0.777	0.779	0.657	0.711	0.773	0.654

ตารางที่ 7.2.4 ผลการทดสอบด้วยข้อมูลชุดที่ 4

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.781	0.879	0.871	0.862	0.875	0.819	0.822	0.853	0.837	0.895	0.824
ROS	0.781	0.891	0.878	0.786	0.875	0.789	0.753	0.821	0.878	0.895	0.879
SMOTE	0.805	0.917	0.858	0.810	0.896	0.867	0.825	0.830	0.891	0.900	0.895
RSLS	0.835	0.908	0.891	0.842	0.875	0.819	0.795	0.788	0.824	0.900	0.824
OVUN	0.805	0.871	0.857	0.862	0.875	0.806	0.795	0.807	0.869	0.884	0.824
ROSE	0.844	0.873	0.882	0.867	0.802	0.858	0.860	0.804	0.849	0.858	0.792
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.838	0.879	0.918	0.862	0.875	0.819	0.822	0.853	0.898	0.932	0.848
TOP+ROSE	0.876	0.885	0.918	0.903	0.875	0.826	0.822	0.853	0.898	0.913	0.877
TOP+SMOTE	0.849	0.903	0.900	0.862	0.905	0.858	0.822	0.853	0.898	0.932	0.872
TOP+RSLS	0.841	0.903	0.909	0.862	0.875	0.826	0.822	0.853	0.898	0.932	0.872

ตารางที่ 7.2.5 ผลการทดสอบด้วยข้อมูลชุดที่ 5

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.172	0.193	0.360	0.295	0.415	0.357	0.325	0.336	0.358	0.289	0.265
ROS	0.464	0.470	0.516	0.452	0.438	0.470	0.462	0.403	0.431	0.358	0.498
SMOTE	0.500	0.474	0.498	0.446	0.410	0.416	0.407	0.387	0.472	0.357	0.444
RSLS	0.472	0.466	0.463	0.438	0.401	0.401	0.347	0.379	0.390	0.337	0.452
OVUN	0.519	0.519	0.535	0.507	0.463	0.485	0.497	0.479	0.520	0.497	0.520
ROSE	0.479	0.490	0.479	0.420	0.376	0.401	0.402	0.389	0.438	0.440	0.399
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.512	0.497	0.536	0.498	0.464	0.525	0.530	0.524	0.509	0.503	0.508
TOP+ROSE	0.518	0.512	0.518	0.503	0.532	0.494	0.457	0.447	0.505	0.510	0.484
TOP+SMOTE	0.499	0.503	0.522	0.504	0.532	0.460	0.524	0.499	0.499	0.499	0.514
TOP+RSLS	0.517	0.503	0.531	0.504	0.532	0.526	0.545	0.499	0.496	0.520	0.518

ตารางที่ 7.2.6 ผลการทดสอบด้วยข้อมูลชุดที่ 6

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.703	0.749	0.741	0.701	0.668	0.680	0.704	0.399	0.688	0.766	0.659
ROS	0.685	0.742	0.786	0.696	0.711	0.683	0.688	0.448	0.748	0.778	0.722
SMOTE	0.750	0.748	0.780	0.735	0.684	0.694	0.738	0.525	0.731	0.762	0.724
RSLS	0.740	0.748	0.777	0.735	0.684	0.694	0.738	0.525	0.674	0.785	0.728
OVUN	0.591	0.642	0.664	0.632	0.575	0.587	0.610	0.647	0.634	0.631	0.595
ROSE	0.586	0.573	0.573	0.560	0.553	0.546	0.534	0.606	0.556	0.545	0.574
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.651	0.675	0.701	0.664	0.650	0.654	0.636	0.652	0.648	0.691	0.641
TOP+ROSE	0.644	0.667	0.689	0.642	0.646	0.645	0.639	0.622	0.640	0.687	0.631
TOP+SMOTE	0.644	0.654	0.664	0.647	0.647	0.637	0.644	0.645	0.646	0.676	0.624
TOP+RSLS	0.636	0.654	0.664	0.657	0.647	0.645	0.644	0.646	0.654	0.687	0.633

ตารางที่ 7.2.7 ผลการทดสอบด้วยข้อมูลชุดที่ 7

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.880	0.926	0.883	0.944	0.939	0.886	0.867	0.906	0.989	0.950	0.888
ROS	0.915	0.926	0.947	0.944	0.939	0.902	0.835	0.879	0.989	0.929	0.930
SMOTE	0.933	0.947	0.934	0.963	0.958	0.915	0.870	0.897	0.976	0.958	0.927
RSLS	0.851	0.899	0.905	0.944	0.939	0.873	0.891	0.896	0.989	0.955	0.885
OVUN	0.872	0.913	0.886	0.944	0.939	0.915	0.886	0.914	0.989	0.968	0.882
ROSE	0.877	0.881	0.929	0.926	0.736	0.768	0.784	0.831	0.871	0.929	0.900
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.937	0.963	0.968	0.976	0.958	0.918	0.902	0.906	0.989	0.987	0.922
TOP+ROSE	0.923	0.952	0.950	0.976	0.947	0.929	0.918	0.910	0.989	0.971	0.958
TOP+SMOTE	0.965	0.965	0.934	0.952	0.939	0.929	0.869	0.929	0.989	0.958	0.925
TOP+RSLS	0.944	0.963	0.934	0.952	0.947	0.929	0.870	0.929	0.989	0.968	0.915

ตารางที่ 7.2.8 ผลการทดสอบด้วยข้อมูลชุดที่ 8

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.854	0.852	0.853	0.912	0.920	0.889	0.837	0.933	0.971	0.926	0.917
ROS	0.866	0.863	0.872	0.912	0.920	0.922	0.893	0.888	0.976	0.926	0.901
SMOTE	0.877	0.877	0.907	0.912	0.934	0.900	0.898	0.891	0.976	0.934	0.919
RSLS	0.845	0.897	0.904	0.912	0.920	0.887	0.833	0.920	0.989	0.939	0.911
OVUN	0.853	0.858	0.870	0.912	0.920	0.899	0.899	0.911	0.976	0.929	0.874
ROSE	0.845	0.823	0.899	0.902	0.722	0.767	0.794	0.863	0.865	0.873	0.882
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.906	0.944	0.922	0.947	0.920	0.920	0.878	0.933	0.989	0.989	0.946
TOP+ROSE	0.886	0.910	0.907	0.930	0.934	0.924	0.917	0.933	0.989	0.947	0.943
TOP+SMOTE	0.900	0.910	0.915	0.962	0.965	0.907	0.875	0.933	0.989	0.947	0.935
TOP+RSLS	0.900	0.910	0.915	0.962	0.963	0.931	0.875	0.933	0.989	0.947	0.935

ตารางที่ 7.2.9 ผลการทดสอบด้วยข้อมูลชุดที่ 9

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.784	0.822	0.619	0.620	0.688	0.642	0.636	0.485	0.823	0.870	0.763
ROS	0.971	0.915	0.971	0.677	0.811	0.733	0.659	0.599	0.880	0.958	0.916
SMOTE	0.971	0.939	0.971	0.596	0.873	0.801	0.725	0.701	0.939	0.928	0.916
RSLS	0.971	0.971	0.583	0.666	0.811	0.770	0.672	0.514	0.806	0.941	0.806
OVUN	0.919	0.971	0.569	0.616	0.811	0.783	0.672	0.514	0.878	0.928	0.917
ROSE	0.428	0.426	0.490	0.584	0.544	0.429	0.455	0.609	0.516	0.516	0.725
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.971	0.987	0.845	0.751	0.850	0.801	0.752	0.542	0.935	1.000	0.923
TOP+ROSE	0.971	0.971	0.971	0.727	0.830	0.801	0.742	0.582	0.916	1.000	0.937
TOP+SMOTE	0.971	0.971	0.971	0.727	0.864	0.801	0.739	0.514	0.887	0.971	0.937
TOP+RSLS	0.971	0.971	0.971	0.727	0.826	0.801	0.739	0.514	0.916	0.982	0.937

ตารางที่ 7.2.10 ผลการทดสอบด้วยข้อมูลชุดที่ 10

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.601	0.600	0.593	0.651	0.545	0.543	0.547	0.621	0.530	0.600	0.644
ROS	0.637	0.593	0.685	0.663	0.544	0.563	0.579	0.654	0.564	0.648	0.658
SMOTE	0.607	0.647	0.637	0.644	0.546	0.531	0.550	0.627	0.538	0.643	0.637
RSLS	0.608	0.647	0.626	0.644	0.546	0.531	0.550	0.627	0.481	0.636	0.634
OVUN	0.622	0.638	0.661	0.666	0.560	0.564	0.578	0.629	0.535	0.644	0.657
ROSE	0.619	0.630	0.644	0.642	0.561	0.566	0.570	0.629	0.482	0.645	0.632
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.675	0.674	0.672	0.677	0.587	0.589	0.605	0.665	0.586	0.674	0.668
TOP+ROSE	0.658	0.670	0.695	0.686	0.597	0.603	0.631	0.658	0.590	0.669	0.674
TOP+SMOTE	0.658	0.662	0.694	0.677	0.597	0.603	0.630	0.663	0.584	0.675	0.688
TOP+RSLS	0.658	0.664	0.696	0.674	0.597	0.603	0.631	0.659	0.598	0.669	0.677

ตารางที่ 7.2.11 ผลการทดสอบด้วยข้อมูลชุดที่ 11

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.441	0.530	0.471	0.591	0.424	0.441	0.457	0.523	0.533	0.507	0.487
ROS	0.494	0.584	0.632	0.676	0.417	0.515	0.534	0.537	0.615	0.596	0.646
SMOTE	0.565	0.613	0.589	0.648	0.439	0.513	0.533	0.536	0.608	0.564	0.610
RSLS	0.610	0.599	0.591	0.664	0.438	0.503	0.518	0.528	0.625	0.549	0.599
OVUN	0.478	0.582	0.519	0.595	0.429	0.466	0.487	0.522	0.608	0.560	0.525
ROSE	0.534	0.596	0.605	0.544	0.448	0.486	0.519	0.536	0.535	0.592	0.575
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.605	0.633	0.587	0.653	0.567	0.582	0.556	0.523	0.634	0.634	0.587
TOP+ROSE	0.589	0.649	0.606	0.664	0.558	0.596	0.549	0.551	0.648	0.652	0.665
TOP+SMOTE	0.617	0.629	0.555	0.646	0.585	0.542	0.549	0.530	0.645	0.630	0.664
TOP+RSLS	0.614	0.629	0.549	0.646	0.582	0.581	0.549	0.530	0.620	0.628	0.675

ตารางที่ 7.2.12 ผลการทดสอบด้วยข้อมูลชุดที่ 12

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.923	0.963	0.937	0.941	0.856	0.840	0.819	0.598	0.964	0.977	0.938
ROS	0.941	0.973	0.891	0.936	0.852	0.822	0.799	0.658	0.909	0.982	0.949
SMOTE	0.923	0.975	0.895	0.939	0.862	0.849	0.829	0.636	0.953	0.968	0.950
RSLS	0.942	0.966	0.894	0.942	0.859	0.847	0.830	0.626	0.938	0.966	0.957
OVUN	0.925	0.961	0.873	0.925	0.782	0.724	0.679	0.266	0.894	0.964	0.950
ROSE	0.762	0.795	0.851	0.807	0.586	0.595	0.556	0.645	0.677	0.853	0.844
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.939	0.980	0.941	0.948	0.856	0.843	0.819	0.700	0.964	0.982	0.945
TOP+ROSE	0.935	0.977	0.937	0.942	0.856	0.853	0.840	0.601	0.964	0.977	0.942
TOP+SMOTE	0.939	0.977	0.938	0.944	0.856	0.843	0.819	0.598	0.944	0.979	0.956
TOP+RSLS	0.933	0.977	0.938	0.944	0.860	0.848	0.819	0.598	0.964	0.977	0.954

ตารางที่ 7.2.13 ผลการทดสอบด้วยข้อมูลชุดที่ 13

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.946	0.955	0.953	0.954	0.943	0.956	0.960	0.946	0.948	0.963	0.943
ROS	0.944	0.955	0.947	0.956	0.938	0.957	0.959	0.948	0.947	0.953	0.947
SMOTE	0.945	0.955	0.948	0.954	0.943	0.956	0.960	0.946	0.949	0.959	0.943
RSLS	0.932	0.950	0.947	0.961	0.947	0.963	0.963	0.948	0.954	0.961	0.951
OVUN	0.933	0.962	0.954	0.959	0.943	0.955	0.958	0.944	0.943	0.954	0.945
ROSE	0.936	0.943	0.958	0.954	0.930	0.950	0.945	0.942	0.950	0.952	0.946
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.951	0.957	0.960	0.961	0.945	0.958	0.962	0.953	0.954	0.965	0.946
TOP+ROSE	0.946	0.964	0.959	0.960	0.963	0.964	0.964	0.951	0.964	0.965	0.949
TOP+SMOTE	0.946	0.956	0.953	0.956	0.957	0.963	0.962	0.958	0.954	0.963	0.952
TOP+RSLS	0.946	0.958	0.953	0.956	0.959	0.966	0.962	0.958	0.956	0.965	0.954

ตารางที่ 7.2.14 ผลการทดสอบด้วยข้อมูลชุดที่ 14

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.510	0.556	0.547	0.470	0.493	0.539	0.505	0.569	0.536	0.559	0.495
ROS	0.530	0.567	0.605	0.591	0.492	0.578	0.571	0.562	0.610	0.564	0.592
SMOTE	0.598	0.602	0.610	0.587	0.505	0.595	0.578	0.565	0.616	0.616	0.583
RSLS	0.574	0.584	0.614	0.581	0.510	0.566	0.549	0.567	0.605	0.590	0.585
OVUN	0.574	0.569	0.554	0.555	0.570	0.568	0.565	0.540	0.560	0.577	0.577
ROSE	0.540	0.540	0.540	0.592	0.546	0.577	0.590	0.547	0.612	0.541	0.597
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.598	0.610	0.605	0.603	0.579	0.581	0.588	0.569	0.616	0.598	0.589
TOP+ROSE	0.575	0.606	0.609	0.595	0.560	0.578	0.558	0.574	0.608	0.593	0.594
TOP+SMOTE	0.580	0.588	0.594	0.591	0.549	0.581	0.582	0.569	0.598	0.598	0.585
TOP+RSLS	0.580	0.599	0.594	0.591	0.540	0.577	0.583	0.569	0.599	0.601	0.585

ตารางที่ 7.2.15 ผลการทดสอบด้วยข้อมูลชุดที่ 15

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.772	0.763	0.770	0.742	0.656	0.700	0.753	0.518	0.705	0.750	0.755
ROS	0.703	0.780	0.784	0.676	0.656	0.678	0.676	0.419	0.767	0.787	0.739
SMOTE	0.772	0.764	0.789	0.693	0.664	0.683	0.698	0.464	0.764	0.783	0.745
RSLS	0.770	0.778	0.787	0.758	0.658	0.724	0.755	0.530	0.787	0.759	0.773
OVUN	0.713	0.745	0.750	0.643	0.661	0.653	0.628	0.428	0.687	0.745	0.689
ROSE	0.354	0.331	0.571	0.679	0.553	0.597	0.602	0.377	0.745	0.343	0.696
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.801	0.791	0.813	0.771	0.677	0.741	0.758	0.518	0.815	0.788	0.799
TOP+ROSE	0.790	0.781	0.799	0.776	0.677	0.705	0.753	0.521	0.789	0.783	0.793
TOP+SMOTE	0.790	0.779	0.804	0.765	0.677	0.705	0.753	0.518	0.776	0.789	0.787
TOP+RSLS	0.790	0.779	0.805	0.765	0.677	0.705	0.753	0.518	0.776	0.796	0.787

7.3 ผลการทดสอบประสิทธิภาพด้วยหน่วยวัดแบบ AUROC

ตารางที่ 7.3.1 ผลการทดสอบด้วยข้อมูลชุดที่ 1

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.849	0.935	0.930	0.938	0.912	0.949	0.952	0.939	0.937	0.967	0.962
ROS	0.882	0.964	0.954	0.937	0.912	0.935	0.938	0.936	0.941	0.960	0.962
SMOTE	0.858	0.957	0.945	0.936	0.903	0.936	0.942	0.944	0.936	0.969	0.963
RSLS	0.862	0.962	0.941	0.937	0.912	0.948	0.951	0.938	0.939	0.960	0.956
OVUN	0.876	0.950	0.935	0.935	0.921	0.940	0.948	0.939	0.937	0.965	0.960
ROSE	0.518	0.518	0.937	0.934	0.863	0.928	0.944	0.946	0.934	0.945	0.952
DBSM	0.894	-	-	-	-	-	-	-	-	-	-
TOP+V	0.869	0.957	0.912	0.931	0.912	0.950	0.952	0.939	0.911	0.954	0.969
TOP+ROSE	0.864	0.957	0.925	0.939	0.930	0.950	0.951	0.939	0.933	0.973	0.976
TOP+SMOTE	0.869	0.957	0.951	0.935	0.930	0.950	0.951	0.945	0.952	0.970	0.968
TOP+RSLS	0.869	0.957	0.949	0.939	0.930	0.950	0.951	0.950	0.937	0.978	0.968

ตารางที่ 7.3.2 ผลการทดสอบด้วยข้อมูลชุดที่ 2

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.772	0.904	0.871	0.818	0.785	0.847	0.880	0.762	0.828	0.937	0.874
ROS	0.792	0.898	0.903	0.812	0.785	0.843	0.856	0.760	0.825	0.929	0.886
SMOTE	0.724	0.900	0.896	0.816	0.792	0.853	0.866	0.766	0.818	0.931	0.879
RSLS	0.764	0.923	0.897	0.809	0.789	0.852	0.866	0.764	0.825	0.934	0.878
OVUN	0.780	0.892	0.871	0.816	0.778	0.851	0.877	0.788	0.830	0.934	0.878
ROSE	0.729	0.823	0.818	0.792	0.688	0.715	0.729	0.751	0.800	0.789	0.814
DBSM	0.788	-	-	-	-	-	-	-	-	-	-
TOP+V	0.826	0.918	0.867	0.817	0.817	0.876	0.889	0.762	0.815	0.938	0.854
TOP+ROSE	0.819	0.911	0.913	0.816	0.785	0.858	0.862	0.756	0.822	0.939	0.869
TOP+SMOTE	0.815	0.904	0.901	0.816	0.785	0.856	0.863	0.756	0.828	0.940	0.884
TOP+RSLS	0.815	0.904	0.902	0.816	0.785	0.856	0.863	0.756	0.830	0.937	0.879

ตารางที่ 7.3.3 ผลการทดสอบด้วยข้อมูลชุดที่ 3

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.735	0.860	0.828	0.703	0.832	0.879	0.895	0.714	0.852	0.915	0.789
ROS	0.673	0.889	0.848	0.715	0.827	0.883	0.875	0.701	0.842	0.914	0.812
SMOTE	0.692	0.860	0.827	0.703	0.832	0.879	0.895	0.714	0.859	0.904	0.793
RSLS	0.777	0.896	0.839	0.706	0.832	0.887	0.897	0.688	0.842	0.907	0.796
OVUN	0.748	0.906	0.843	0.735	0.836	0.878	0.899	0.646	0.857	0.919	0.798
ROSE	0.643	0.699	0.734	0.647	0.695	0.736	0.720	0.724	0.747	0.744	0.741
DBSM	0.777	-	-	-	-	-	-	-	-	-	-
TOP+V	0.829	0.899	0.802	0.719	0.864	0.895	0.903	0.696	0.864	0.921	0.781
TOP+ROSE	0.746	0.899	0.856	0.727	0.842	0.892	0.904	0.716	0.848	0.913	0.821
TOP+SMOTE	0.784	0.903	0.834	0.724	0.847	0.885	0.912	0.706	0.851	0.908	0.802
TOP+RSLS	0.723	0.902	0.839	0.726	0.845	0.888	0.910	0.683	0.849	0.892	0.802

ตารางที่ 7.3.4 ผลการทดสอบด้วยข้อมูลชุดที่ 4

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.907	0.933	0.961	0.947	0.917	0.920	0.912	0.930	0.975	0.979	0.995
ROS	0.852	0.958	0.967	0.945	0.917	0.913	0.911	0.922	0.971	0.970	0.995
SMOTE	0.882	0.953	0.964	0.948	0.939	0.939	0.954	0.885	0.986	0.970	0.991
RSLS	0.939	0.965	0.964	0.949	0.917	0.915	0.911	0.906	0.975	0.982	0.995
OVUN	0.900	0.921	0.965	0.942	0.917	0.914	0.912	0.949	0.972	0.993	0.995
ROSE	0.927	0.974	0.964	0.948	0.920	0.958	0.963	0.946	0.949	0.972	0.970
DBSM	0.903	-	-	-	-	-	-	-	-	-	-
TOP+V	0.934	0.933	0.964	0.947	0.917	0.920	0.912	0.930	0.975	0.974	0.995
TOP+ROSE	0.937	0.945	0.949	0.935	0.917	0.915	0.912	0.930	0.982	0.953	0.993
TOP+SMOTE	0.902	0.953	0.961	0.951	0.942	0.939	0.908	0.919	0.988	0.989	0.995
TOP+RSLS	0.900	0.953	0.971	0.951	0.917	0.915	0.908	0.919	0.973	0.969	0.995

ตารางที่ 7.3.5 ผลการทดสอบด้วยข้อมูลชุดที่ 5

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.516	0.529	0.670	0.676	0.609	0.630	0.619	0.631	0.679	0.652	0.671
ROS	0.627	0.698	0.703	0.667	0.613	0.632	0.627	0.622	0.646	0.642	0.694
SMOTE	0.660	0.627	0.702	0.678	0.603	0.617	0.632	0.615	0.663	0.665	0.715
RSLS	0.625	0.650	0.683	0.675	0.600	0.619	0.607	0.611	0.608	0.648	0.717
OVUN	0.668	0.674	0.702	0.623	0.614	0.664	0.661	0.584	0.675	0.677	0.671
ROSE	0.640	0.645	0.648	0.653	0.542	0.568	0.592	0.611	0.648	0.640	0.645
DBSM	0.622	-	-	-	-	-	-	-	-	-	-
TOP+V	0.609	0.607	0.662	0.652	0.640	0.555	0.548	0.563	0.658	0.604	0.618
TOP+ROSE	0.635	0.643	0.696	0.672	0.594	0.653	0.621	0.616	0.671	0.636	0.714
TOP+SMOTE	0.640	0.677	0.685	0.633	0.600	0.651	0.555	0.593	0.653	0.639	0.632
TOP+RSLS	0.652	0.677	0.680	0.633	0.600	0.638	0.641	0.593	0.665	0.638	0.630

ตารางที่ 7.3.6 ผลการทดสอบด้วยข้อมูลชุดที่ 6

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.617	0.747	0.744	0.708	0.620	0.640	0.679	0.634	0.700	0.765	0.659
ROS	0.631	0.732	0.783	0.710	0.623	0.650	0.662	0.647	0.732	0.754	0.696
SMOTE	0.642	0.735	0.753	0.707	0.617	0.633	0.683	0.644	0.686	0.753	0.687
RSLS	0.642	0.735	0.770	0.707	0.617	0.633	0.683	0.644	0.664	0.759	0.691
OVUN	0.594	0.700	0.704	0.696	0.602	0.620	0.633	0.624	0.699	0.733	0.653
ROSE	0.541	0.547	0.673	0.699	0.565	0.569	0.550	0.620	0.620	0.617	0.656
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.676	0.723	0.772	0.738	0.575	0.625	0.689	0.636	0.719	0.766	0.700
TOP+ROSE	0.577	0.762	0.765	0.714	0.601	0.669	0.678	0.636	0.704	0.771	0.729
TOP+SMOTE	0.657	0.735	0.789	0.700	0.604	0.665	0.661	0.643	0.741	0.739	0.715
TOP+RSLS	0.593	0.735	0.789	0.707	0.604	0.679	0.661	0.650	0.703	0.771	0.709

ตารางที่ 7.3.7 ผลการทดสอบด้วยข้อมูลชุดที่ 7

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.944	0.996	1.000	0.985	0.950	0.965	0.963	0.996	1.000	1.000	0.994
ROS	0.936	0.965	0.999	0.968	0.950	0.967	0.967	0.996	1.000	1.000	0.997
SMOTE	0.957	0.972	0.999	0.988	0.967	0.965	0.964	0.996	1.000	1.000	0.997
RSLS	0.872	0.999	1.000	0.985	0.950	0.964	0.963	0.996	1.000	1.000	0.994
OVUN	0.914	0.999	1.000	0.968	0.950	0.965	0.965	0.996	1.000	1.000	0.994
ROSE	0.955	0.997	1.000	1.000	0.903	0.960	0.973	0.995	1.000	0.999	0.993
DBSM	0.933	-	-	-	-	-	-	-	-	-	-
TOP+V	0.973	0.996	1.000	1.000	0.967	0.964	0.962	0.996	1.000	1.000	0.992
TOP+ROSE	0.962	0.983	0.999	0.988	0.964	0.967	0.964	0.996	1.000	1.000	1.000
TOP+SMOTE	0.982	0.983	1.000	0.997	0.950	0.967	0.982	0.996	1.000	0.998	0.997
TOP+RSLS	0.956	0.974	1.000	0.997	0.964	0.967	0.958	0.996	1.000	1.000	0.996

ตารางที่ 7.3.8 ผลการทดสอบด้วยข้อมูลชุดที่ 8

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.929	0.968	0.993	0.964	0.949	0.995	0.994	0.996	1.000	0.993	0.990
ROS	0.914	0.945	0.990	0.964	0.949	0.992	0.994	0.996	1.000	0.996	0.994
SMOTE	0.920	0.942	0.984	0.964	0.961	0.990	0.991	0.996	1.000	0.997	0.988
RSLS	0.875	0.995	0.988	0.964	0.949	0.992	0.992	0.996	1.000	0.996	0.992
OVUN	0.903	0.948	0.988	0.981	0.949	0.990	0.994	0.996	1.000	0.990	0.988
ROSE	0.937	0.986	0.989	0.993	0.894	0.974	0.981	0.994	1.000	0.990	0.993
DBSM	0.974	-	-	-	-	-	-	-	-	-	-
TOP+V	0.964	0.999	0.991	1.000	0.949	0.995	0.997	0.996	1.000	1.000	0.990
TOP+ROSE	0.927	0.991	0.991	0.981	0.961	0.993	0.997	0.997	1.000	0.996	0.993
TOP+SMOTE	0.930	0.991	0.991	0.996	0.992	0.992	0.986	0.996	1.000	0.998	0.990
TOP+RSLS	0.930	0.991	0.991	0.996	0.992	0.992	0.986	0.996	1.000	0.996	0.990

ตารางที่ 7.3.9 ผลการทดสอบด้วยข้อมูลชุดที่ 9

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.955	0.989	0.991	0.976	0.905	0.960	0.975	0.931	0.983	0.996	0.987
ROS	0.998	0.996	1.000	0.920	0.903	0.957	0.957	0.948	0.997	1.000	0.997
SMOTE	0.998	0.998	1.000	0.821	0.953	0.993	0.992	0.948	0.992	1.000	1.000
RSLS	0.998	0.998	0.988	0.930	0.903	0.957	0.947	0.947	0.996	1.000	0.996
OVUN	0.990	0.998	0.992	0.947	0.903	0.958	0.945	0.943	0.994	0.999	0.997
ROSE	0.847	0.925	0.943	0.960	0.753	0.746	0.742	0.956	0.954	0.977	0.984
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.998	0.999	0.993	0.977	0.957	0.960	0.965	0.949	0.997	1.000	0.998
TOP+ROSE	0.998	0.998	1.000	0.915	0.941	0.959	0.955	0.947	0.997	1.000	0.998
TOP+SMOTE	0.998	0.998	1.000	0.915	0.952	0.959	0.954	0.945	0.997	1.000	0.998
TOP+RSLS	0.998	0.998	1.000	0.915	0.918	0.959	0.954	0.945	0.997	0.999	0.998

ตารางที่ 7.3.10 ผลการทดสอบด้วยข้อมูลชุดที่ 10

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.627	0.685	0.687	0.775	0.566	0.592	0.596	0.713	0.581	0.692	0.755
ROS	0.712	0.801	0.828	0.834	0.652	0.684	0.707	0.814	0.702	0.820	0.827
SMOTE	0.715	0.817	0.832	0.833	0.651	0.704	0.734	0.816	0.695	0.819	0.830
RSLS	0.725	0.817	0.826	0.833	0.651	0.704	0.734	0.816	0.688	0.825	0.831
OVUN	0.732	0.814	0.818	0.829	0.656	0.705	0.733	0.801	0.680	0.812	0.826
ROSE	0.744	0.792	0.812	0.829	0.655	0.705	0.727	0.802	0.605	0.807	0.815
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.753	0.821	0.833	0.817	0.656	0.704	0.728	0.778	0.692	0.821	0.822
TOP+ROSE	0.747	0.816	0.837	0.832	0.678	0.712	0.747	0.816	0.694	0.800	0.829
TOP+SMOTE	0.741	0.787	0.828	0.833	0.675	0.710	0.743	0.815	0.703	0.815	0.830
TOP+RSLS	0.741	0.788	0.830	0.823	0.676	0.712	0.746	0.815	0.687	0.810	0.829

ตารางที่ 7.3.11 ผลการทดสอบด้วยข้อมูลชุดที่ 11

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.647	0.854	0.822	0.875	0.613	0.704	0.757	0.709	0.837	0.850	0.845
ROS	0.653	0.857	0.847	0.875	0.609	0.692	0.721	0.709	0.829	0.865	0.864
SMOTE	0.708	0.876	0.847	0.874	0.619	0.719	0.749	0.709	0.827	0.861	0.856
RSLS	0.762	0.863	0.844	0.875	0.620	0.722	0.747	0.705	0.842	0.860	0.857
OVUN	0.654	0.863	0.826	0.876	0.616	0.716	0.761	0.710	0.851	0.863	0.850
ROSE	0.721	0.784	0.799	0.757	0.589	0.651	0.680	0.707	0.725	0.798	0.786
DBSM	0.730	-	-	-	-	-	-	-	-	-	-
TOP+V	0.760	0.841	0.808	0.856	0.706	0.740	0.761	0.709	0.838	0.844	0.816
TOP+ROSE	0.717	0.854	0.815	0.864	0.685	0.739	0.756	0.729	0.834	0.845	0.844
TOP+SMOTE	0.749	0.856	0.804	0.866	0.696	0.734	0.755	0.710	0.858	0.842	0.877
TOP+RSLS	0.752	0.856	0.799	0.866	0.700	0.690	0.755	0.710	0.841	0.847	0.871

ตารางที่ 7.3.12 ผลการทดสอบด้วยข้อมูลชุดที่ 12

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.946	0.998	0.991	0.992	0.907	0.952	0.958	0.854	0.991	0.998	0.993
ROS	0.964	0.999	0.988	0.991	0.903	0.943	0.949	0.867	0.981	0.999	0.995
SMOTE	0.942	0.999	0.984	0.992	0.918	0.956	0.960	0.850	0.994	0.999	0.995
RSLS	0.962	0.998	0.984	0.992	0.913	0.958	0.960	0.851	0.982	0.999	0.995
OVUN	0.967	0.997	0.979	0.989	0.880	0.906	0.900	0.541	0.966	0.998	0.995
ROSE	0.891	0.962	0.978	0.937	0.719	0.781	0.790	0.853	0.890	0.980	0.981
DBSM	0.948	-	-	-	-	-	-	-	-	-	-
TOP+V	0.970	0.998	0.991	0.992	0.907	0.955	0.958	0.870	0.991	0.999	0.992
TOP+ROSE	0.962	0.999	0.991	0.992	0.907	0.955	0.962	0.795	0.991	0.999	0.994
TOP+SMOTE	0.963	0.999	0.991	0.992	0.907	0.953	0.958	0.854	0.988	0.999	0.995
TOP+RSLS	0.953	0.998	0.991	0.992	0.915	0.955	0.958	0.854	0.991	0.999	0.995

ตารางที่ 7.3.13 ผลการทดสอบด้วยข้อมูลชุดที่ 13

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.959	0.995	0.991	0.994	0.955	0.986	0.992	0.985	0.992	0.994	0.987
ROS	0.952	0.994	0.989	0.995	0.952	0.983	0.987	0.985	0.993	0.991	0.987
SMOTE	0.962	0.995	0.990	0.994	0.955	0.986	0.992	0.985	0.991	0.990	0.988
RSLS	0.953	0.993	0.989	0.995	0.960	0.986	0.987	0.985	0.992	0.992	0.988
OVUN	0.950	0.993	0.989	0.995	0.955	0.982	0.988	0.987	0.991	0.992	0.987
ROSE	0.952	0.990	0.991	0.995	0.943	0.978	0.985	0.988	0.992	0.989	0.989
DBSM	0.952	-	-	-	-	-	-	-	-	-	-
TOP+V	0.967	0.992	0.988	0.995	0.957	0.986	0.992	0.988	0.993	0.993	0.988
TOP+ROSE	0.959	0.993	0.991	0.994	0.979	0.980	0.986	0.985	0.993	0.994	0.983
TOP+SMOTE	0.959	0.993	0.990	0.995	0.970	0.980	0.990	0.992	0.992	0.993	0.977
TOP+RSLS	0.959	0.989	0.990	0.995	0.975	0.981	0.990	0.992	0.991	0.994	0.981

ตารางที่ 7.3.14 ผลการทดสอบด้วยข้อมูลชุดที่ 14

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.734	0.788	0.783	0.791	0.643	0.737	0.757	0.762	0.804	0.799	0.781
ROS	0.645	0.775	0.794	0.790	0.642	0.726	0.751	0.758	0.805	0.798	0.785
SMOTE	0.734	0.791	0.789	0.791	0.650	0.738	0.754	0.760	0.806	0.807	0.789
RSLS	0.716	0.780	0.798	0.790	0.653	0.731	0.751	0.760	0.807	0.799	0.791
OVUN	0.614	0.683	0.646	0.724	0.597	0.603	0.606	0.645	0.690	0.665	0.646
ROSE	0.507	0.507	0.507	0.786	0.651	0.716	0.743	0.759	0.794	0.715	0.783
DBSM	0.698	-	-	-	-	-	-	-	-	-	-
TOP+V	0.741	0.778	0.782	0.781	0.659	0.731	0.702	0.762	0.789	0.784	0.763
TOP+ROSE	0.726	0.798	0.799	0.792	0.638	0.731	0.753	0.760	0.805	0.792	0.788
TOP+SMOTE	0.718	0.794	0.794	0.748	0.669	0.717	0.729	0.762	0.805	0.794	0.789
TOP+RSLS	0.718	0.777	0.794	0.748	0.668	0.721	0.737	0.762	0.805	0.793	0.789

ตารางที่ 7.3.15 ผลการทดสอบด้วยข้อมูลชุดที่ 15

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.934	0.971	0.968	0.962	0.801	0.887	0.929	0.953	0.972	0.961	0.973
ROS	0.854	0.970	0.974	0.962	0.801	0.880	0.919	0.952	0.976	0.968	0.969
SMOTE	0.918	0.970	0.968	0.963	0.826	0.906	0.938	0.945	0.975	0.969	0.969
RSLS	0.899	0.968	0.963	0.963	0.803	0.892	0.931	0.947	0.976	0.967	0.973
OVUN	0.909	0.967	0.968	0.958	0.868	0.920	0.928	0.911	0.972	0.969	0.956
ROSE	0.570	0.767	0.954	0.966	0.837	0.918	0.939	0.950	0.976	0.964	0.971
DBSM	0.875	-	-	-	-	-	-	-	-	-	-
TOP+V	0.918	0.971	0.970	0.961	0.832	0.910	0.930	0.953	0.968	0.973	0.967
TOP+ROSE	0.932	0.964	0.969	0.962	0.845	0.901	0.929	0.937	0.973	0.971	0.974
TOP+SMOTE	0.932	0.969	0.968	0.961	0.845	0.901	0.929	0.953	0.973	0.974	0.973
TOP+RSLS	0.932	0.969	0.968	0.961	0.845	0.901	0.929	0.953	0.973	0.963	0.973

7.4 ผลการทดสอบประสิทธิภาพด้วยหน่วยวัดแบบ AUPRC

ตารางที่ 7.4.1 ผลการทดสอบด้วยข้อมูลชุดที่ 1

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.730	0.875	0.825	0.849	0.745	0.885	0.911	0.816	0.845	0.911	0.905
ROS	0.627	0.900	0.888	0.837	0.745	0.765	0.783	0.815	0.840	0.913	0.921
SMOTE	0.556	0.875	0.881	0.843	0.654	0.780	0.843	0.845	0.846	0.927	0.917
RSLS	0.679	0.902	0.840	0.845	0.745	0.873	0.900	0.828	0.850	0.918	0.897
OVUN	0.719	0.876	0.832	0.838	0.749	0.813	0.872	0.831	0.839	0.900	0.908
ROSE	0.160	0.160	0.848	0.801	0.472	0.729	0.784	0.866	0.797	0.854	0.889
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.752	0.900	0.819	0.838	0.745	0.890	0.911	0.816	0.818	0.892	0.905
TOP+ROSE	0.764	0.900	0.855	0.850	0.756	0.890	0.896	0.816	0.841	0.922	0.913
TOP+SMOTE	0.752	0.900	0.886	0.837	0.756	0.890	0.896	0.834	0.859	0.911	0.914
TOP+RSLS	0.752	0.900	0.899	0.852	0.756	0.890	0.896	0.876	0.837	0.926	0.914

ตารางที่ 7.4.2 ผลการทดสอบด้วยข้อมูลชุดที่ 2

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.629	0.863	0.754	0.668	0.625	0.709	0.773	0.573	0.696	0.901	0.776
ROS	0.655	0.848	0.824	0.659	0.625	0.678	0.714	0.588	0.640	0.870	0.800
SMOTE	0.550	0.836	0.800	0.667	0.620	0.701	0.715	0.580	0.628	0.879	0.785
RSLS	0.579	0.868	0.801	0.661	0.615	0.698	0.730	0.577	0.662	0.881	0.781
OVUN	0.609	0.822	0.758	0.666	0.606	0.717	0.767	0.606	0.689	0.895	0.782
ROSE	0.484	0.635	0.685	0.612	0.450	0.472	0.497	0.574	0.624	0.596	0.665
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.638	0.865	0.750	0.656	0.648	0.730	0.783	0.560	0.650	0.888	0.740
TOP+ROSE	0.633	0.838	0.849	0.673	0.625	0.696	0.715	0.548	0.664	0.874	0.784
TOP+SMOTE	0.611	0.863	0.822	0.671	0.625	0.694	0.716	0.548	0.677	0.885	0.805
TOP+RSLS	0.611	0.863	0.823	0.671	0.625	0.694	0.716	0.548	0.680	0.875	0.800

ตารางที่ 7.4.3 ผลการทดสอบด้วยข้อมูลชุดที่ 3

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.646	0.754	0.746	0.564	0.762	0.817	0.814	0.569	0.755	0.866	0.699
ROS	0.584	0.841	0.786	0.578	0.771	0.824	0.811	0.564	0.761	0.866	0.739
SMOTE	0.569	0.754	0.745	0.564	0.762	0.817	0.814	0.569	0.786	0.858	0.721
RSLS	0.673	0.833	0.761	0.569	0.750	0.826	0.832	0.517	0.748	0.843	0.695
OVUN	0.646	0.865	0.753	0.603	0.761	0.817	0.821	0.514	0.788	0.861	0.719
ROSE	0.478	0.535	0.610	0.479	0.505	0.557	0.549	0.597	0.609	0.607	0.632
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.745	0.834	0.648	0.552	0.784	0.826	0.822	0.541	0.813	0.875	0.676
TOP+ROSE	0.611	0.829	0.793	0.597	0.753	0.830	0.811	0.570	0.716	0.870	0.733
TOP+SMOTE	0.650	0.854	0.767	0.599	0.691	0.814	0.839	0.566	0.743	0.855	0.703
TOP+RSLS	0.556	0.845	0.777	0.600	0.740	0.820	0.826	0.554	0.774	0.834	0.709

ตารางที่ 7.4.4 ผลการทดสอบด้วยข้อมูลชุดที่ 4

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.805	0.896	0.895	0.832	0.813	0.878	0.854	0.783	0.930	0.962	0.976
ROS	0.624	0.918	0.933	0.786	0.813	0.808	0.807	0.792	0.911	0.922	0.976
SMOTE	0.721	0.928	0.930	0.805	0.838	0.850	0.872	0.811	0.945	0.942	0.963
RSLS	0.822	0.927	0.896	0.842	0.813	0.828	0.851	0.771	0.930	0.970	0.976
OVUN	0.737	0.854	0.905	0.826	0.813	0.823	0.858	0.794	0.894	0.959	0.976
ROSE	0.761	0.955	0.937	0.831	0.681	0.806	0.821	0.774	0.944	0.954	0.957
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.775	0.896	0.939	0.832	0.813	0.878	0.854	0.783	0.942	0.946	0.976
TOP+ROSE	0.805	0.915	0.898	0.883	0.813	0.828	0.854	0.783	0.951	0.917	0.972
TOP+SMOTE	0.766	0.904	0.902	0.847	0.853	0.850	0.842	0.788	0.967	0.939	0.976
TOP+RSLS	0.726	0.904	0.951	0.847	0.813	0.828	0.842	0.788	0.933	0.937	0.976

ตารางที่ 7.4.5 ผลการทดสอบด้วยข้อมูลชุดที่ 5

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.290	0.314	0.470	0.493	0.393	0.433	0.420	0.470	0.453	0.404	0.464
ROS	0.407	0.438	0.474	0.488	0.401	0.414	0.407	0.468	0.484	0.392	0.486
SMOTE	0.429	0.370	0.505	0.495	0.365	0.414	0.433	0.464	0.448	0.420	0.506
RSLS	0.393	0.400	0.491	0.494	0.375	0.400	0.384	0.460	0.408	0.376	0.506
OVUN	0.413	0.454	0.471	0.447	0.364	0.436	0.439	0.415	0.489	0.411	0.455
ROSE	0.427	0.424	0.439	0.472	0.296	0.340	0.377	0.456	0.454	0.406	0.456
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.350	0.347	0.435	0.449	0.375	0.304	0.296	0.338	0.441	0.324	0.377
TOP+ROSE	0.361	0.385	0.518	0.496	0.321	0.446	0.391	0.465	0.456	0.366	0.508
TOP+SMOTE	0.369	0.434	0.450	0.433	0.326	0.455	0.298	0.421	0.460	0.376	0.464
TOP+RSLS	0.402	0.434	0.470	0.433	0.326	0.350	0.354	0.421	0.437	0.397	0.459

ตารางที่ 7.4.6 ผลการทดสอบด้วยข้อมูลชุดที่ 6

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.653	0.754	0.771	0.757	0.668	0.677	0.717	0.696	0.759	0.777	0.715
ROS	0.673	0.765	0.814	0.765	0.661	0.688	0.701	0.714	0.753	0.792	0.753
SMOTE	0.679	0.763	0.785	0.759	0.662	0.672	0.708	0.706	0.715	0.778	0.745
RSLS	0.686	0.763	0.801	0.759	0.662	0.672	0.708	0.706	0.715	0.778	0.748
OVUN	0.490	0.624	0.651	0.644	0.493	0.532	0.542	0.536	0.638	0.671	0.611
ROSE	0.457	0.444	0.613	0.631	0.470	0.475	0.482	0.555	0.546	0.569	0.592
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.576	0.613	0.716	0.700	0.464	0.505	0.625	0.546	0.635	0.742	0.665
TOP+ROSE	0.472	0.727	0.727	0.666	0.485	0.573	0.598	0.547	0.648	0.740	0.718
TOP+SMOTE	0.536	0.699	0.760	0.642	0.487	0.563	0.563	0.551	0.717	0.700	0.708
TOP+RSLS	0.482	0.699	0.760	0.650	0.487	0.573	0.563	0.552	0.625	0.740	0.680

ตารางที่ 7.4.7 ผลการทดสอบด้วยข้อมูลชุดที่ 7

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.827	0.968	1.000	0.959	0.925	0.943	0.935	0.977	1.000	1.000	0.974
ROS	0.880	0.937	0.994	0.935	0.925	0.949	0.949	0.977	1.000	1.000	0.989
SMOTE	0.905	0.965	0.994	0.984	0.949	0.943	0.936	0.977	1.000	1.000	0.989
RSLS	0.780	0.994	1.000	0.959	0.925	0.938	0.935	0.977	1.000	1.000	0.978
OVUN	0.771	0.994	1.000	0.935	0.925	0.939	0.943	0.977	1.000	1.000	0.974
ROSE	0.756	0.989	1.000	1.000	0.557	0.768	0.862	0.965	1.000	0.994	0.972
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.900	0.968	1.000	1.000	0.949	0.929	0.924	0.977	1.000	1.000	0.947
TOP+ROSE	0.878	0.962	0.994	0.984	0.929	0.949	0.938	0.977	1.000	1.000	1.000
TOP+SMOTE	0.944	0.962	1.000	0.975	0.925	0.949	0.970	0.977	1.000	0.993	0.987
TOP+RSLS	0.924	0.966	1.000	0.975	0.929	0.949	0.918	0.977	1.000	1.000	0.982

ตารางที่ 7.4.8 ผลการทดสอบด้วยข้อมูลชุดที่ 8

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.847	0.889	0.966	0.938	0.888	0.975	0.977	0.971	1.000	0.966	0.949
ROS	0.814	0.863	0.962	0.938	0.888	0.946	0.952	0.971	1.000	0.977	0.968
SMOTE	0.836	0.854	0.925	0.938	0.904	0.930	0.946	0.971	1.000	0.988	0.945
RSLS	0.777	0.978	0.931	0.938	0.888	0.954	0.958	0.971	1.000	0.977	0.960
OVUN	0.782	0.857	0.931	0.962	0.888	0.944	0.973	0.971	1.000	0.961	0.953
ROSE	0.745	0.943	0.933	0.966	0.548	0.863	0.893	0.974	1.000	0.955	0.966
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.840	0.997	0.943	1.000	0.888	0.975	0.986	0.971	1.000	1.000	0.949
TOP+ROSE	0.833	0.961	0.956	0.962	0.904	0.945	0.983	0.987	1.000	0.977	0.966
TOP+SMOTE	0.858	0.961	0.956	0.977	0.935	0.935	0.952	0.971	1.000	0.990	0.949
TOP+RSLS	0.858	0.961	0.956	0.977	0.930	0.930	0.952	0.971	1.000	0.977	0.949

ตารางที่ 7.4.9 ผลการทดสอบด้วยข้อมูลชุดที่ 9

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.659	0.928	0.864	0.715	0.514	0.681	0.776	0.598	0.835	0.949	0.858
ROS	0.950	0.935	1.000	0.526	0.709	0.774	0.773	0.658	0.947	1.000	0.979
SMOTE	0.950	0.950	1.000	0.459	0.793	0.837	0.820	0.689	0.951	1.000	1.000
RSLS	0.950	0.950	0.901	0.732	0.709	0.831	0.755	0.664	0.928	1.000	0.931
OVUN	0.875	0.950	0.916	0.785	0.709	0.825	0.748	0.643	0.907	0.990	0.951
ROSE	0.192	0.526	0.670	0.640	0.405	0.456	0.455	0.699	0.555	0.731	0.832
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.950	0.975	0.914	0.850	0.754	0.853	0.771	0.628	0.958	1.000	0.970
TOP+ROSE	0.950	0.950	1.000	0.602	0.729	0.852	0.820	0.711	0.938	1.000	0.970
TOP+SMOTE	0.950	0.950	1.000	0.602	0.778	0.852	0.801	0.598	0.967	1.000	0.970
TOP+RSLS	0.950	0.950	1.000	0.602	0.728	0.852	0.801	0.598	0.938	0.990	0.970

ตารางที่ 7.4.10 ผลการทดสอบด้วยข้อมูลชุดที่ 10

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.434	0.515	0.570	0.635	0.388	0.412	0.418	0.524	0.439	0.515	0.629
ROS	0.549	0.659	0.694	0.729	0.494	0.523	0.550	0.676	0.533	0.695	0.704
SMOTE	0.567	0.691	0.734	0.731	0.487	0.563	0.624	0.677	0.545	0.691	0.723
RSLS	0.595	0.691	0.723	0.731	0.487	0.563	0.624	0.677	0.511	0.721	0.724
OVUN	0.571	0.704	0.706	0.722	0.486	0.547	0.594	0.661	0.496	0.698	0.711
ROSE	0.582	0.674	0.697	0.715	0.486	0.545	0.578	0.663	0.453	0.685	0.707
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.583	0.697	0.714	0.691	0.470	0.520	0.542	0.612	0.491	0.694	0.718
TOP+ROSE	0.604	0.702	0.728	0.724	0.498	0.536	0.573	0.682	0.508	0.653	0.710
TOP+SMOTE	0.585	0.634	0.708	0.730	0.497	0.534	0.569	0.681	0.537	0.694	0.709
TOP+RSLS	0.585	0.639	0.713	0.706	0.497	0.536	0.572	0.682	0.508	0.672	0.713

ตารางที่ 7.4.11 ผลการทดสอบด้วยข้อมูลชุดที่ 11

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.451	0.651	0.627	0.695	0.358	0.427	0.498	0.532	0.622	0.645	0.652
ROS	0.379	0.661	0.625	0.688	0.356	0.399	0.416	0.529	0.606	0.644	0.658
SMOTE	0.454	0.674	0.631	0.692	0.357	0.421	0.450	0.532	0.581	0.654	0.639
RSLS	0.527	0.657	0.634	0.694	0.360	0.423	0.438	0.532	0.617	0.650	0.649
OVUN	0.433	0.664	0.612	0.701	0.360	0.447	0.511	0.537	0.659	0.656	0.640
ROSE	0.428	0.538	0.575	0.530	0.311	0.373	0.405	0.512	0.450	0.561	0.548
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.502	0.610	0.589	0.653	0.415	0.424	0.452	0.532	0.570	0.586	0.576
TOP+ROSE	0.404	0.602	0.565	0.653	0.380	0.422	0.447	0.535	0.606	0.592	0.606
TOP+SMOTE	0.437	0.611	0.560	0.667	0.374	0.432	0.447	0.544	0.655	0.604	0.645
TOP+RSLS	0.428	0.611	0.550	0.667	0.384	0.370	0.447	0.544	0.620	0.592	0.634

ตารางที่ 7.4.12 ผลการทดสอบด้วยข้อมูลชุดที่ 12

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.878	0.994	0.980	0.982	0.789	0.892	0.911	0.673	0.980	0.996	0.983
ROS	0.893	0.996	0.966	0.979	0.786	0.833	0.859	0.699	0.946	0.997	0.987
SMOTE	0.872	0.997	0.958	0.981	0.786	0.891	0.894	0.683	0.984	0.997	0.986
RSLS	0.903	0.996	0.952	0.981	0.785	0.894	0.896	0.696	0.955	0.998	0.985
OVUN	0.895	0.993	0.933	0.974	0.639	0.691	0.690	0.390	0.922	0.994	0.987
ROSE	0.702	0.908	0.945	0.801	0.427	0.574	0.599	0.697	0.708	0.951	0.954
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.898	0.995	0.981	0.981	0.789	0.895	0.911	0.753	0.980	0.996	0.981
TOP+ROSE	0.900	0.997	0.983	0.981	0.789	0.898	0.920	0.627	0.980	0.997	0.983
TOP+SMOTE	0.923	0.997	0.980	0.981	0.789	0.893	0.911	0.673	0.966	0.997	0.988
TOP+RSLS	0.888	0.996	0.980	0.981	0.786	0.873	0.911	0.673	0.980	0.997	0.988

ตารางที่ 7.4.13 ผลการทดสอบด้วยข้อมูลชุดที่ 13

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.949	0.990	0.979	0.989	0.930	0.962	0.979	0.950	0.983	0.987	0.964
ROS	0.933	0.990	0.974	0.991	0.930	0.958	0.960	0.950	0.984	0.981	0.956
SMOTE	0.951	0.990	0.977	0.989	0.930	0.962	0.979	0.950	0.980	0.980	0.964
RSLS	0.907	0.983	0.971	0.989	0.932	0.957	0.961	0.950	0.983	0.980	0.960
OVUN	0.911	0.988	0.970	0.989	0.925	0.950	0.967	0.960	0.981	0.980	0.955
ROSE	0.913	0.974	0.977	0.989	0.913	0.952	0.968	0.964	0.985	0.969	0.961
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.916	0.988	0.967	0.990	0.932	0.962	0.980	0.960	0.983	0.986	0.966
TOP+ROSE	0.949	0.982	0.978	0.988	0.933	0.943	0.957	0.951	0.982	0.988	0.936
TOP+SMOTE	0.949	0.983	0.978	0.989	0.930	0.937	0.971	0.979	0.982	0.985	0.913
TOP+RSLS	0.949	0.962	0.978	0.989	0.922	0.940	0.971	0.979	0.977	0.988	0.924

ตารางที่ 7.4.14 ผลการทดสอบด้วยข้อมูลชุดที่ 14

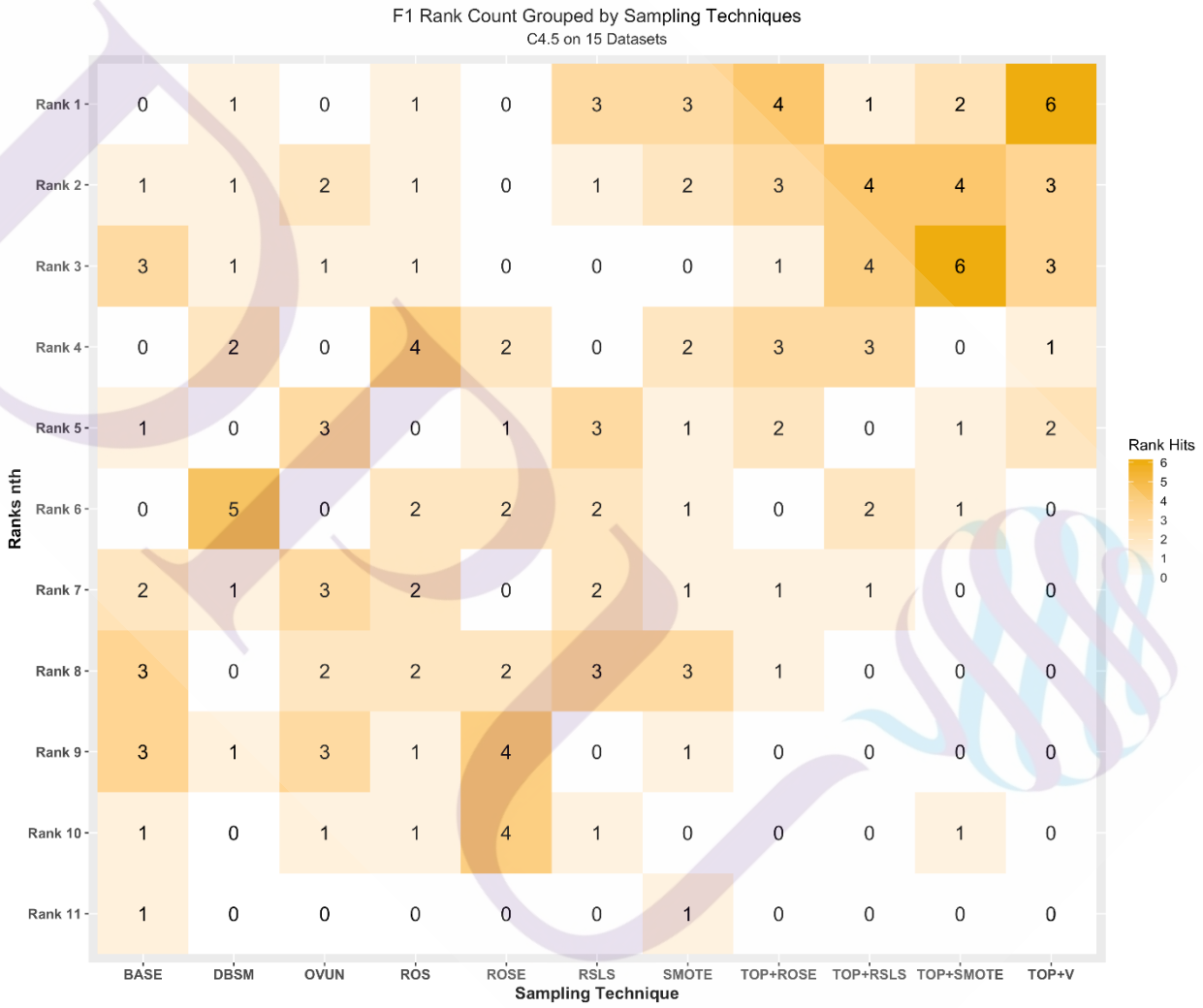
	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.497	0.595	0.615	0.595	0.426	0.530	0.556	0.533	0.620	0.629	0.612
ROS	0.433	0.602	0.616	0.596	0.429	0.491	0.529	0.528	0.623	0.628	0.579
SMOTE	0.502	0.586	0.605	0.598	0.423	0.516	0.527	0.532	0.623	0.636	0.594
RSLS	0.481	0.585	0.623	0.596	0.432	0.501	0.533	0.534	0.619	0.620	0.600
OVUN	0.337	0.406	0.370	0.496	0.337	0.340	0.342	0.381	0.424	0.386	0.376
ROSE	0.292	0.292	0.292	0.589	0.391	0.453	0.492	0.535	0.610	0.439	0.594
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.478	0.568	0.610	0.585	0.386	0.497	0.418	0.533	0.605	0.577	0.584
TOP+ROSE	0.500	0.613	0.619	0.598	0.372	0.503	0.536	0.507	0.624	0.617	0.598
TOP+SMOTE	0.487	0.610	0.603	0.522	0.420	0.464	0.469	0.533	0.618	0.603	0.603
TOP+RSLS	0.487	0.577	0.603	0.522	0.432	0.472	0.490	0.533	0.618	0.608	0.603

ตารางที่ 7.4.15 ผลการทดสอบด้วยข้อมูลชุดที่ 15

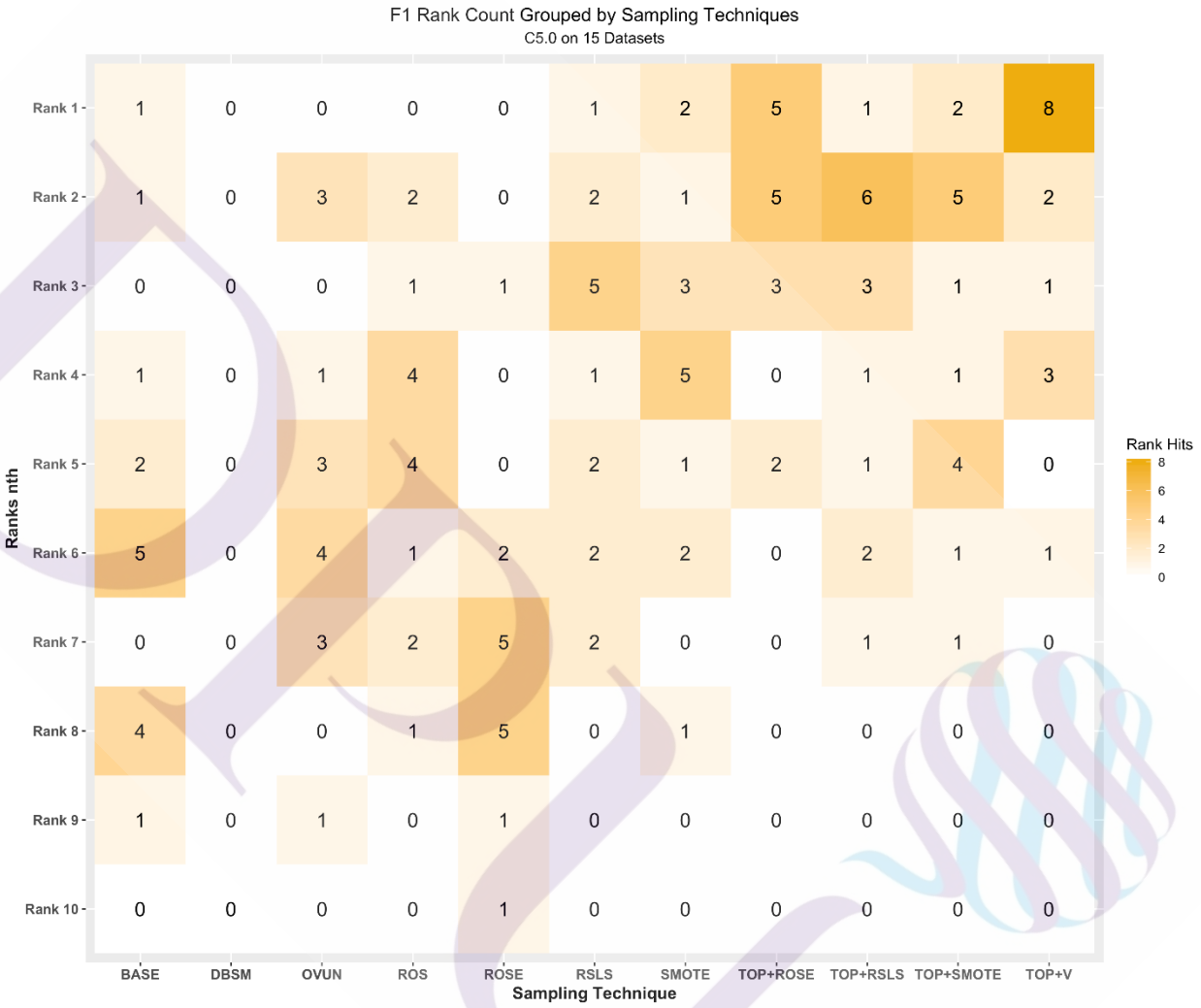
	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
BASE	0.704	0.804	0.848	0.794	0.515	0.689	0.757	0.671	0.819	0.826	0.840
ROS	0.539	0.797	0.828	0.796	0.515	0.597	0.639	0.670	0.838	0.819	0.809
SMOTE	0.641	0.794	0.825	0.795	0.510	0.608	0.676	0.659	0.831	0.811	0.802
RSLS	0.683	0.816	0.842	0.796	0.514	0.705	0.756	0.665	0.835	0.829	0.829
OVUN	0.468	0.763	0.766	0.786	0.475	0.568	0.604	0.619	0.821	0.769	0.774
ROSE	0.153	0.311	0.626	0.793	0.332	0.534	0.638	0.673	0.831	0.731	0.822
DBSM	-	-	-	-	-	-	-	-	-	-	-
TOP+V	0.668	0.821	0.832	0.796	0.525	0.706	0.741	0.671	0.807	0.833	0.817
TOP+ROSE	0.681	0.804	0.851	0.797	0.520	0.665	0.757	0.499	0.815	0.830	0.835
TOP+SMOTE	0.681	0.807	0.839	0.798	0.520	0.665	0.757	0.671	0.820	0.844	0.836
TOP+RSLS	0.681	0.807	0.824	0.798	0.520	0.665	0.757	0.671	0.819	0.773	0.836

7.5 การสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพของโมเดลด้วยหน่วยวัด F1

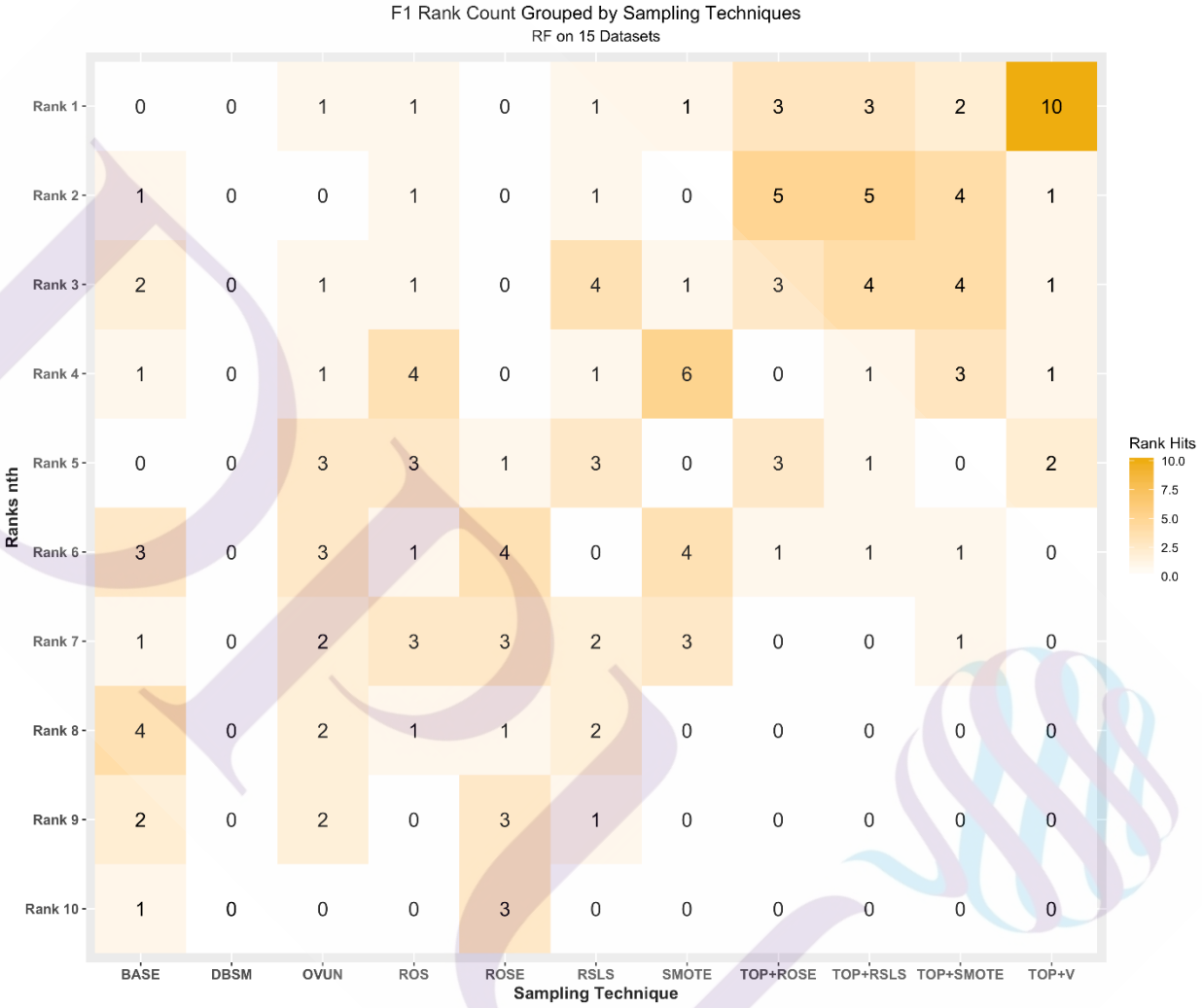
ภาพที่ 7.5.1 โมเดล C4.5 บนข้อมูลทุกชุด



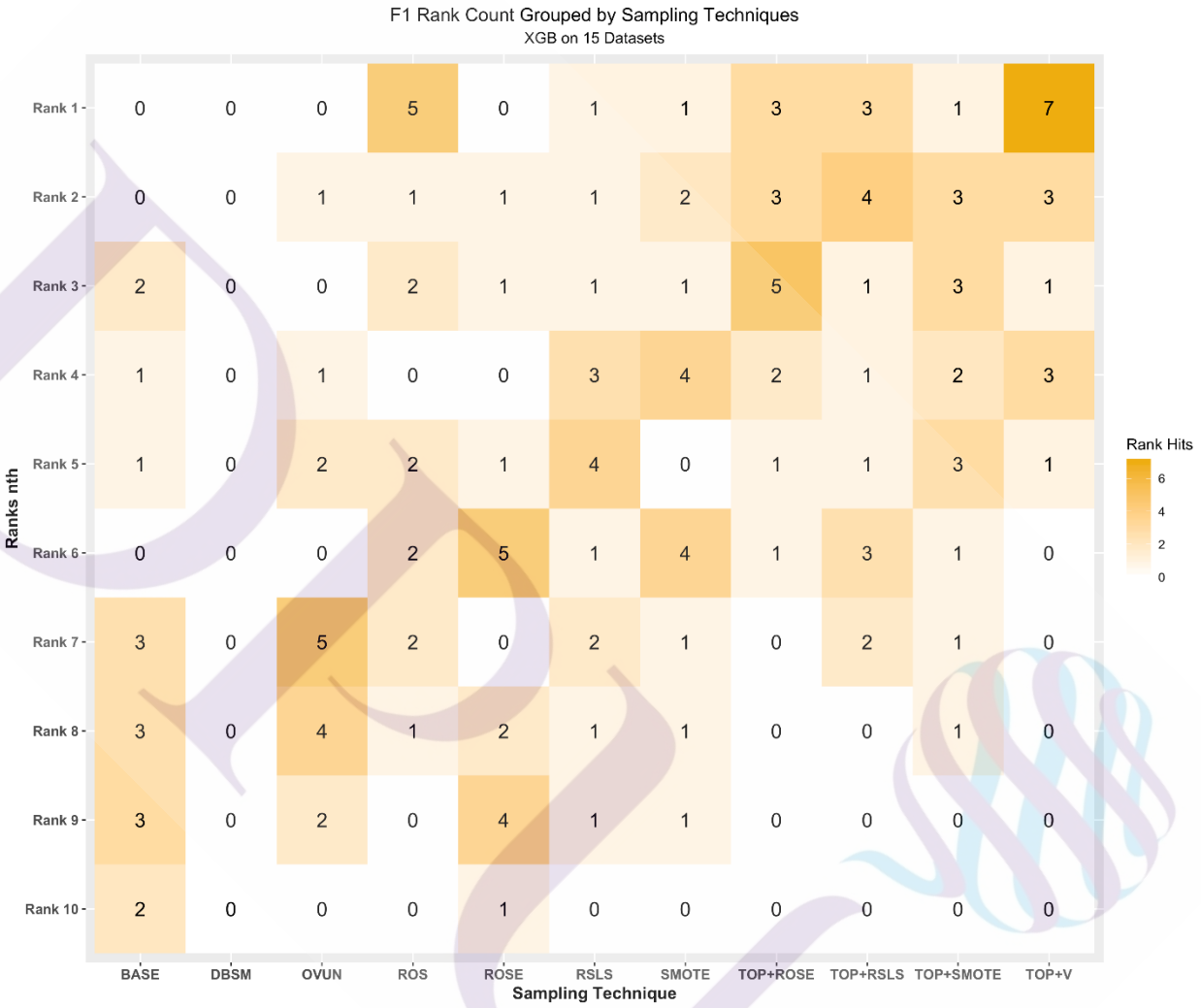
ภาพที่ 7.5.2 โมเดล C5.0 บนข้อมูลทุกชุด



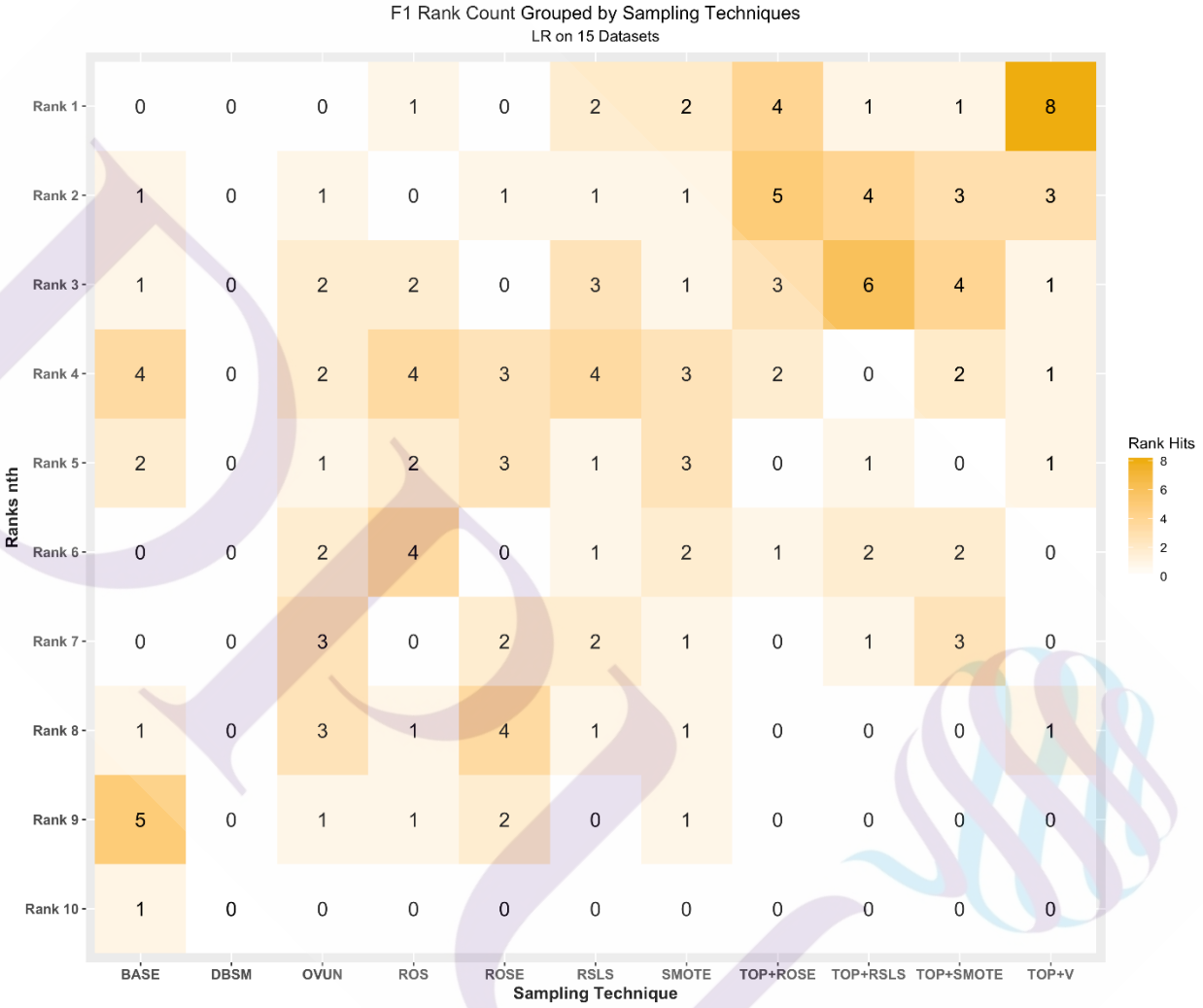
ภาพที่ 7.5.3 โมเดล RF บนข้อมูลทุกชุด



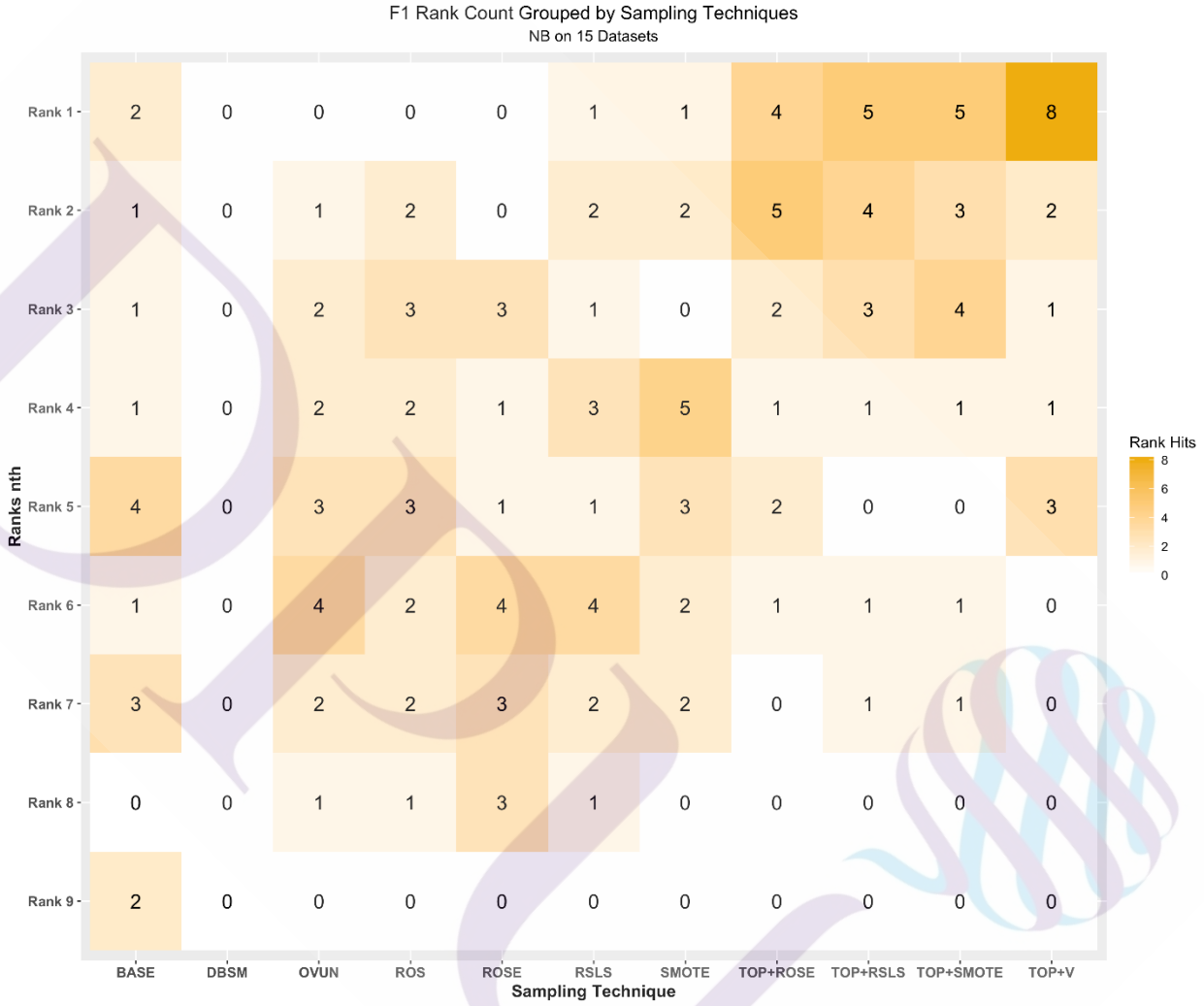
ภาพที่ 7.5.4 โมเดล XGB บนข้อมูลทุกชุด



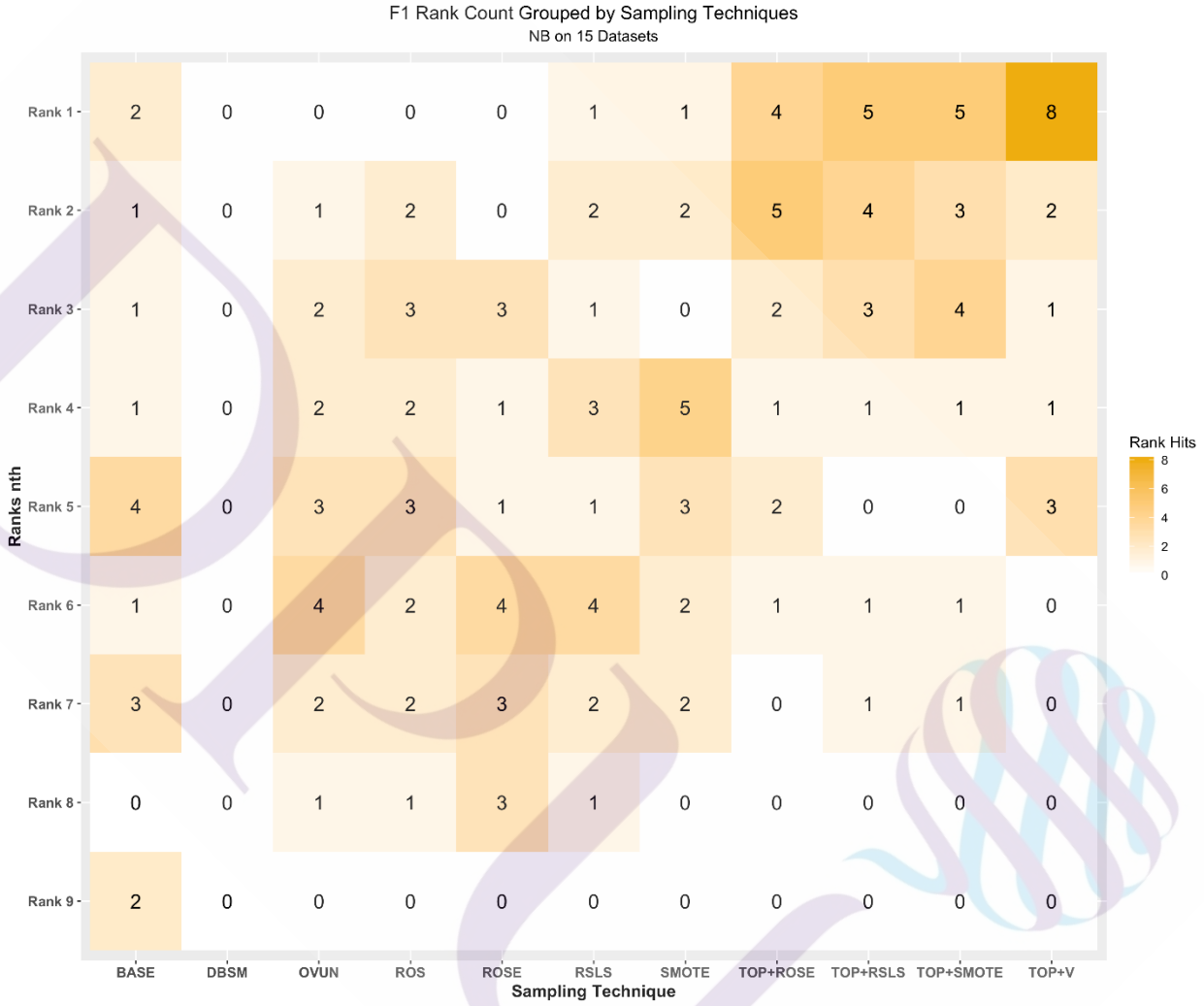
ภาพที่ 7.5.5 โมเดล LR บนข้อมูลทุกชุด



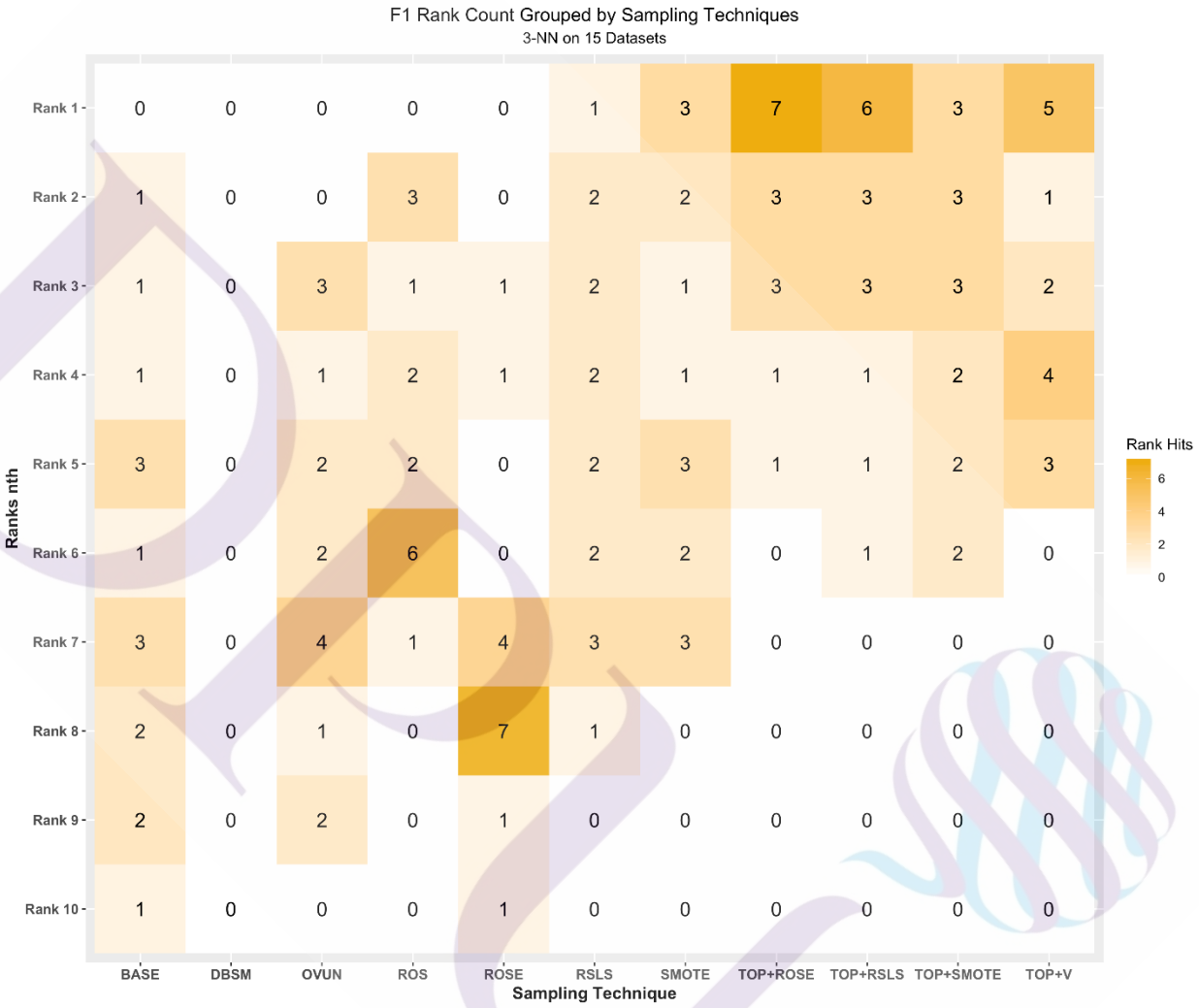
ภาพที่ 7.5.6 โมเดล NB บนข้อมูลทุกชุด



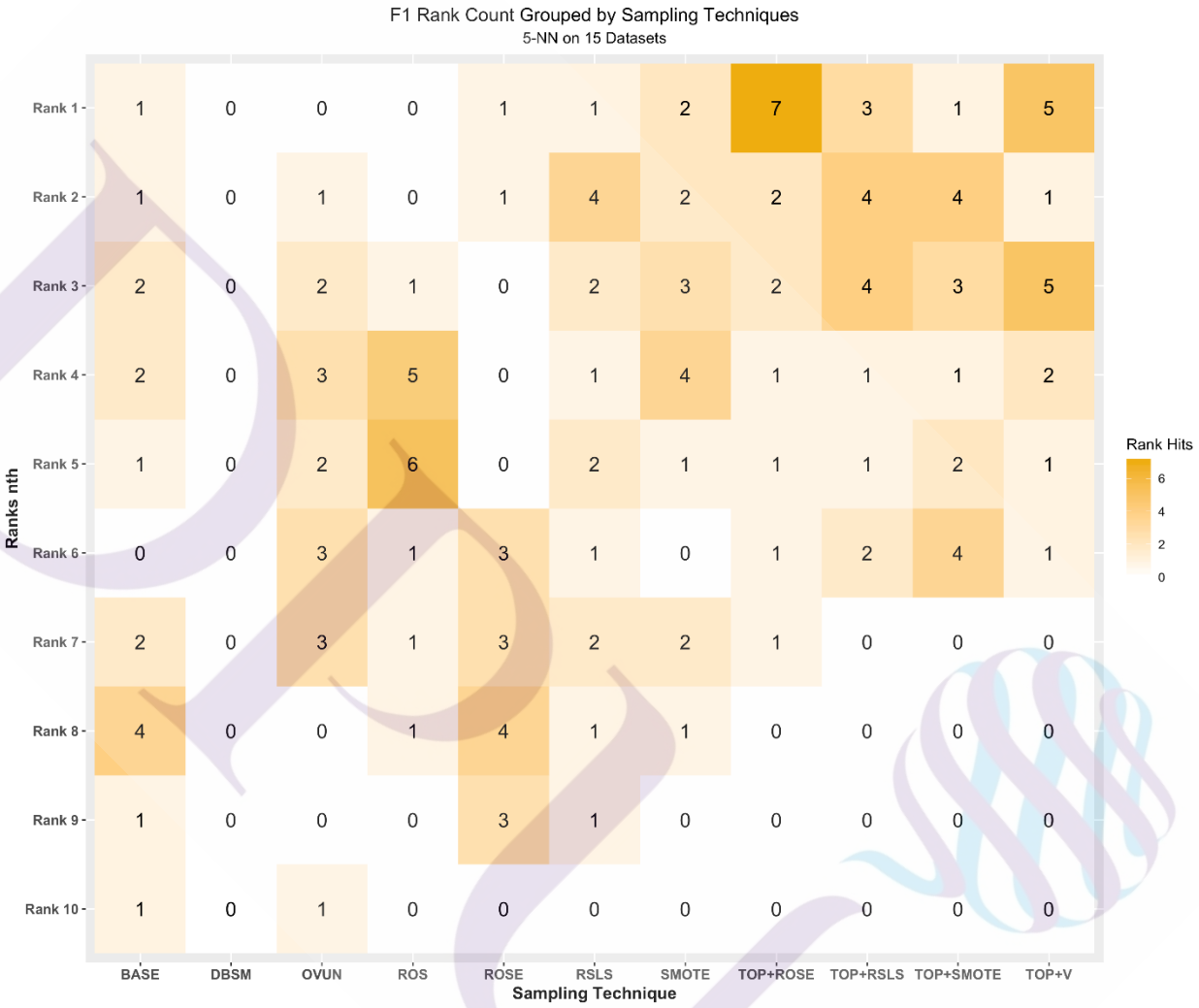
ภาพที่ 7.5.7 โมเดล 1-NN บนข้อมูลทุกชุด



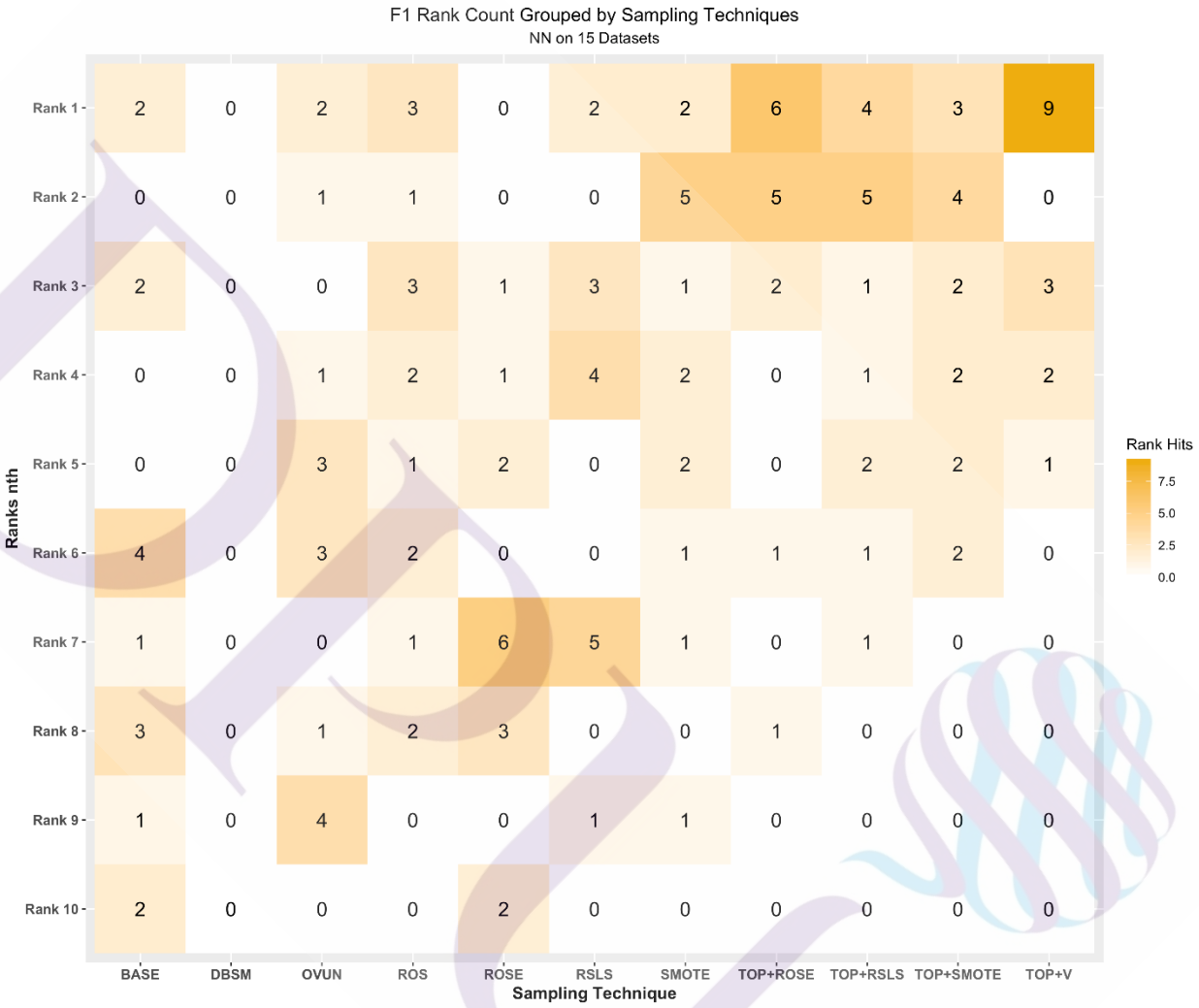
ภาพที่ 7.5.8 โมเดล 3-NN บนข้อมูลทุกชุด



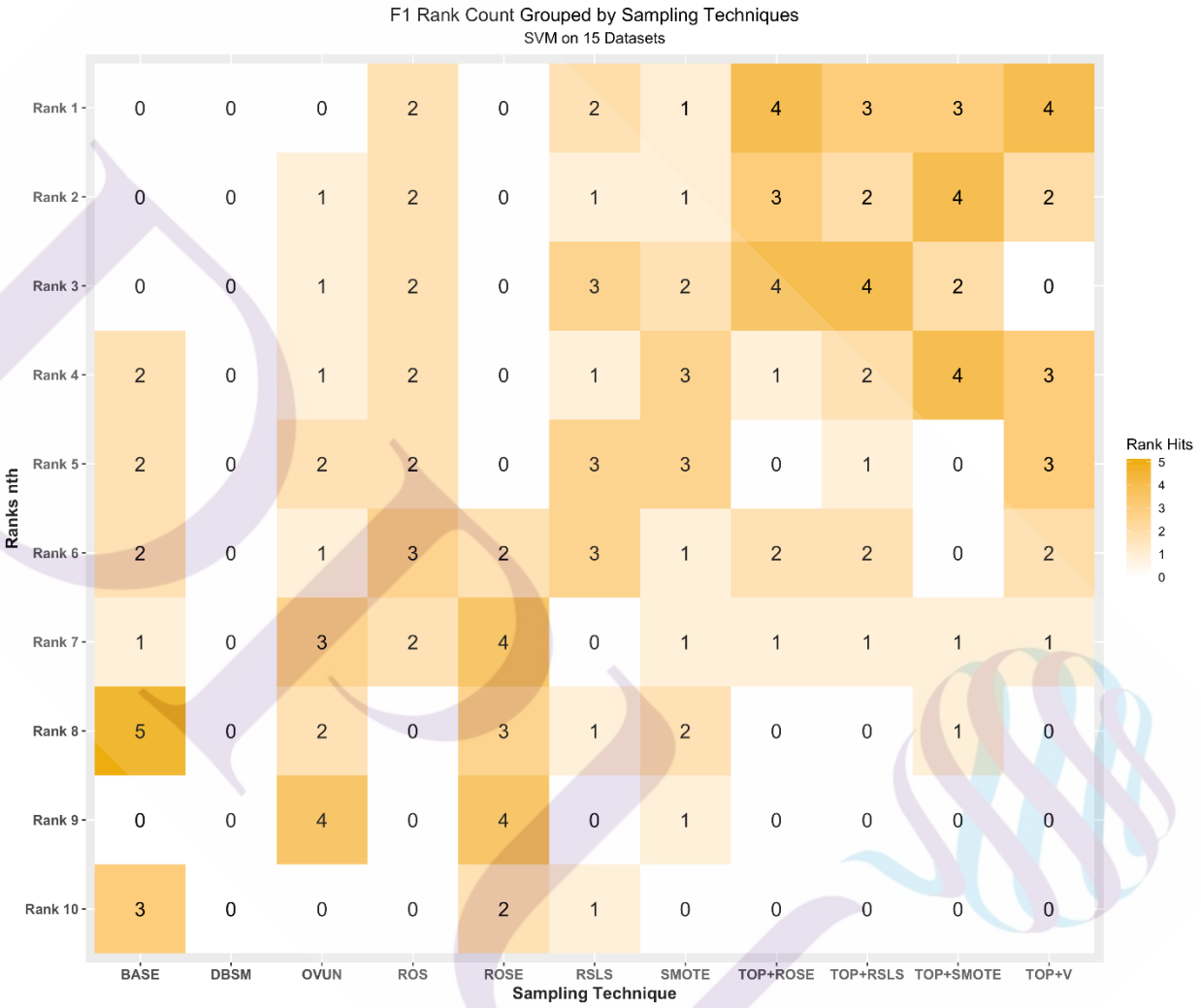
ภาพที่ 7.5.9 โมเดล 5-NN บนข้อมูลทุกชุด



ภาพที่ 7.5.10 โมเดล NN บนข้อมูลทุกชุด

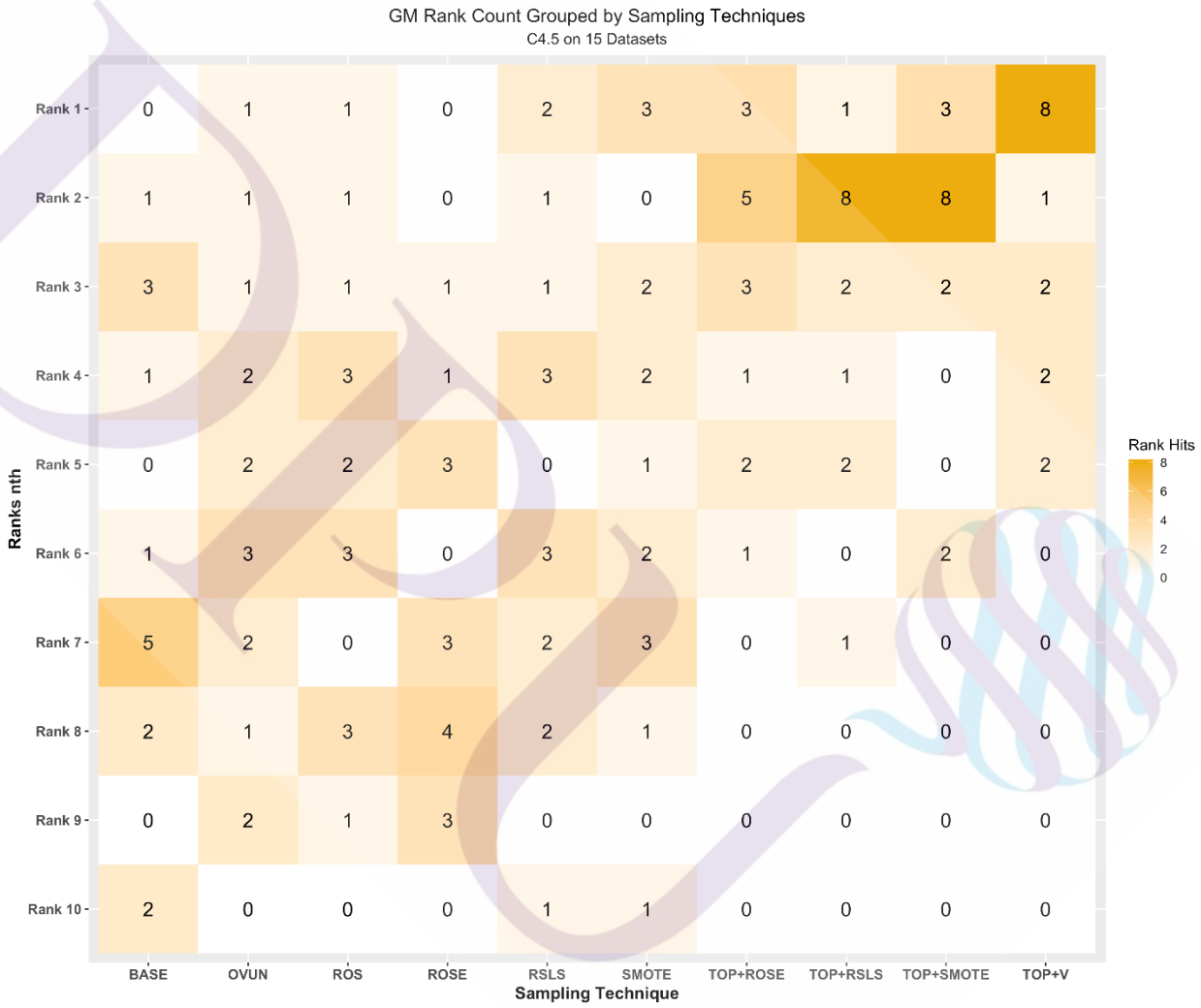


ภาพที่ 7.5.11 โมเดล SVM บนข้อมูลทุกชุด

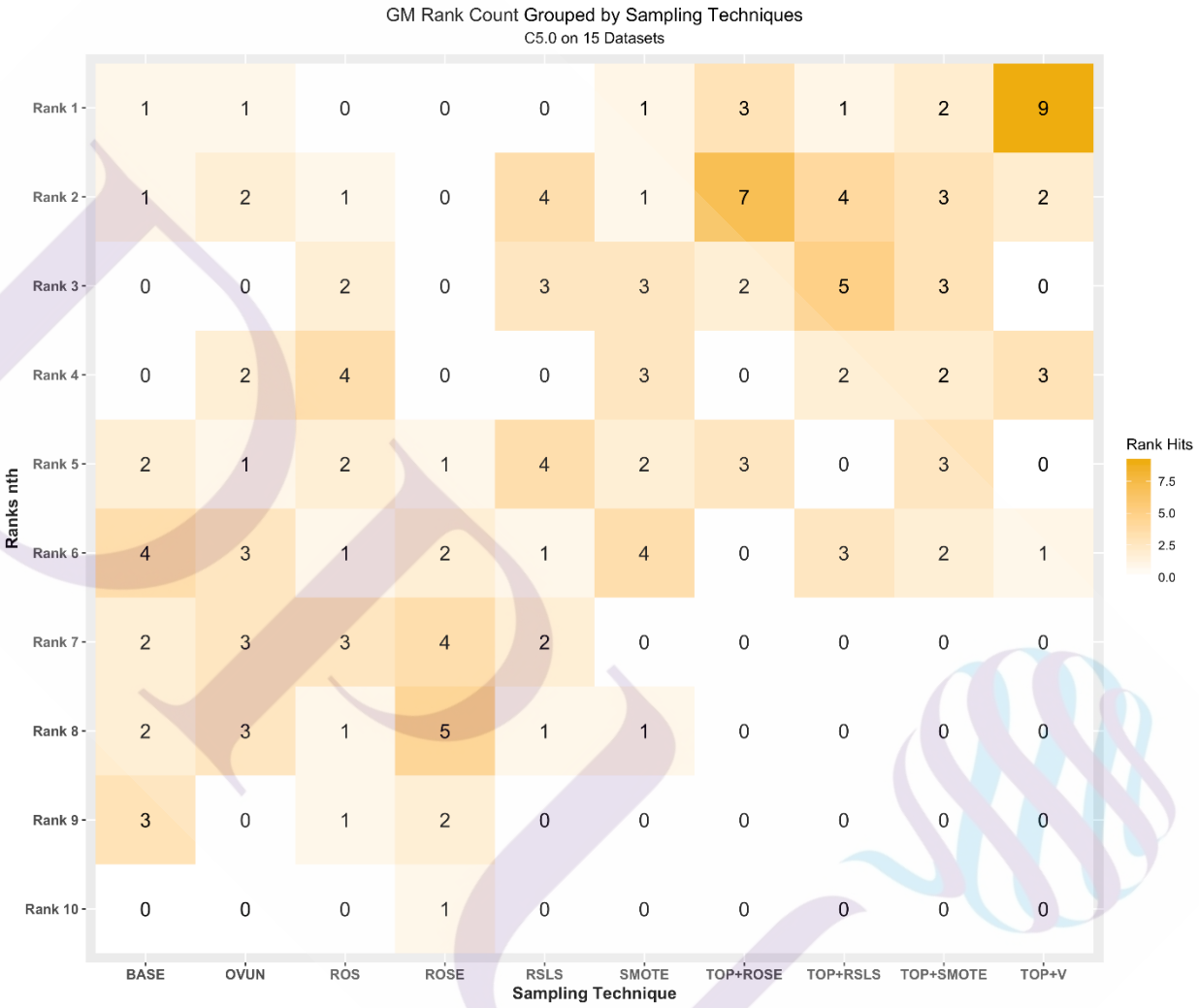


7.6 การสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพของโมเดลด้วยหน่วยวัด GM

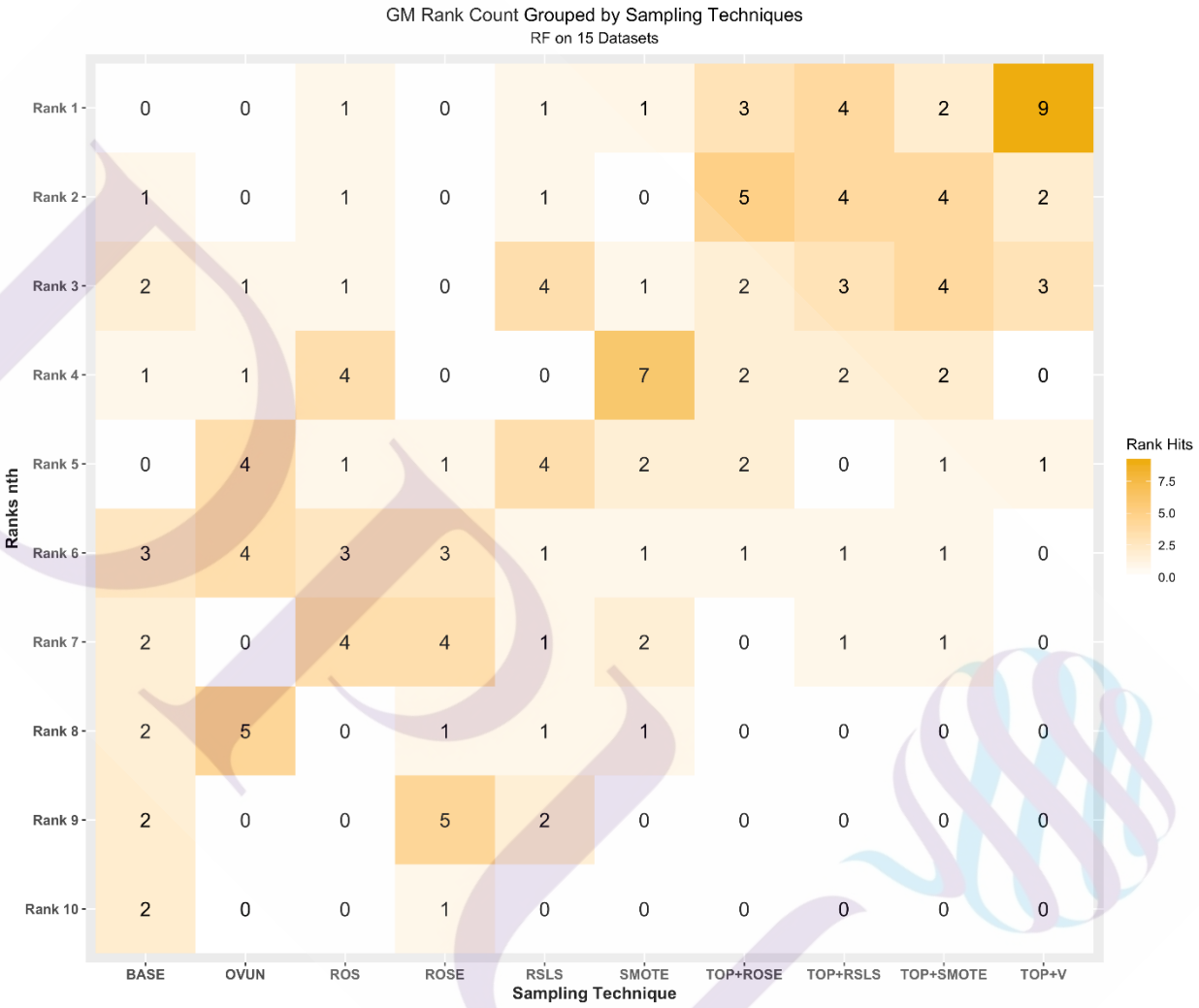
ภาพที่ 7.6.1 โมเดล C4.5 บนข้อมูลทุกชุด



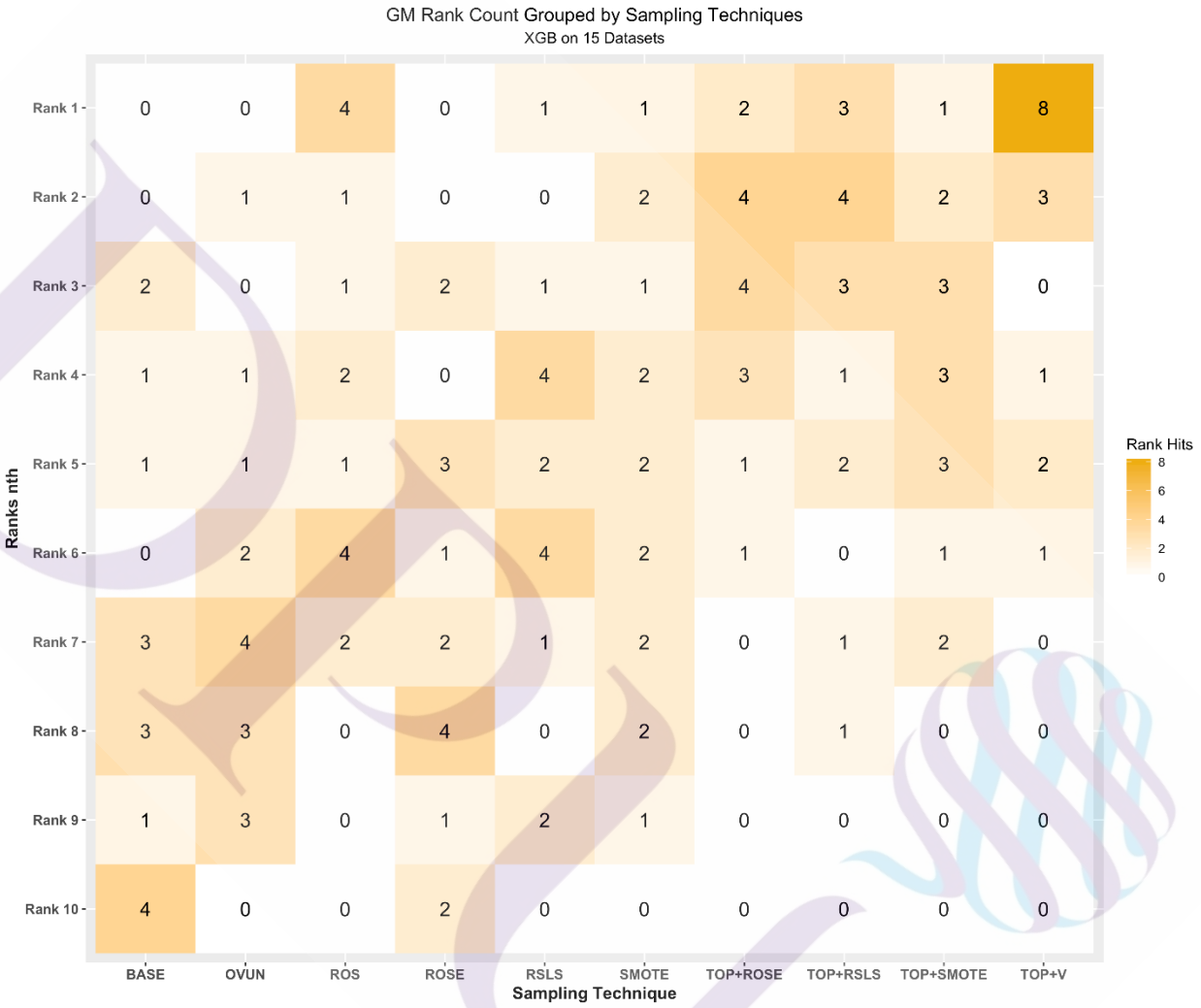
ภาพที่ 7.6.2 โมเดล C5.0 บนข้อมูลทุกชุด



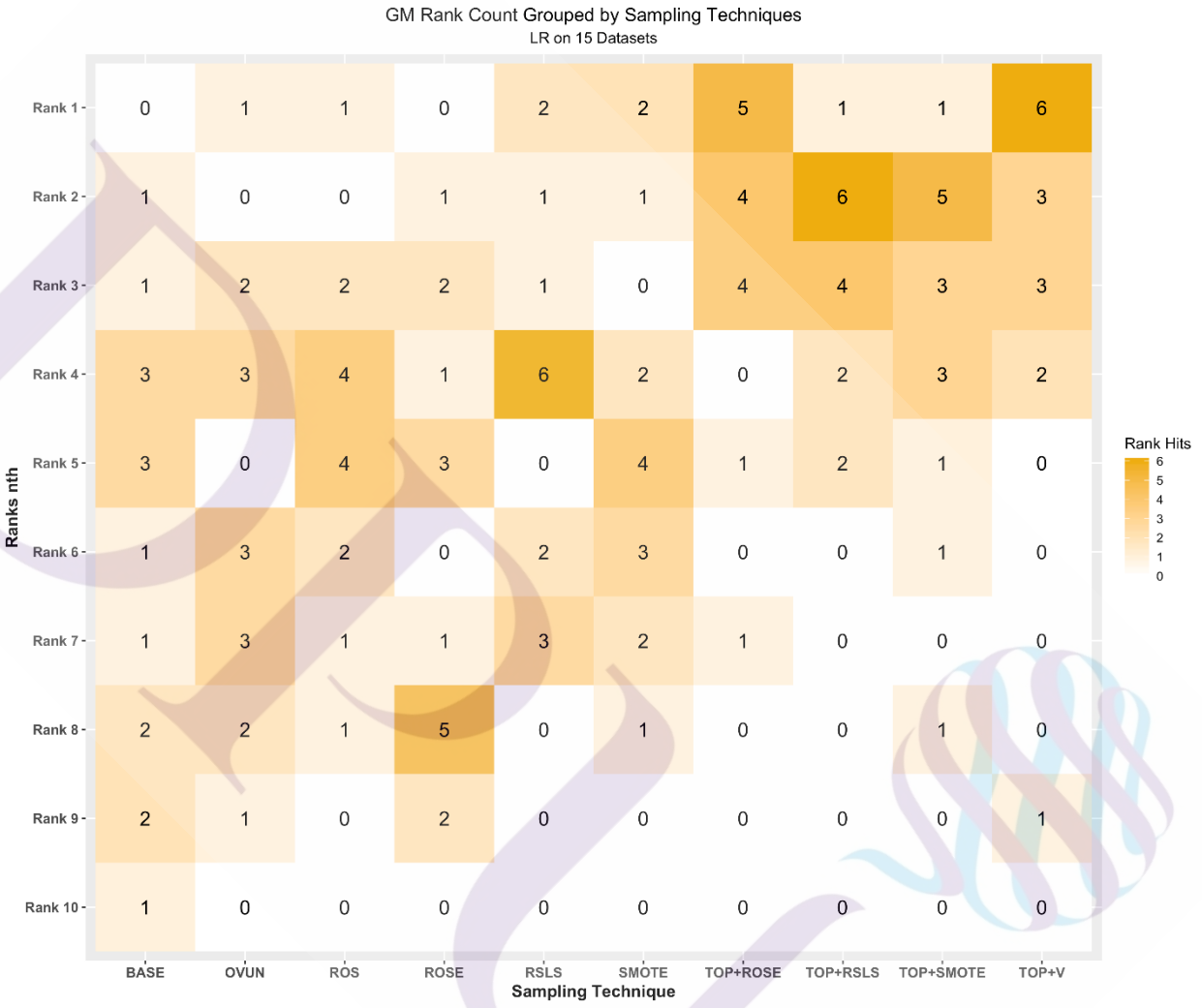
ภาพที่ 7.6.3 โมเดล RF บนข้อมูลทุกชุด



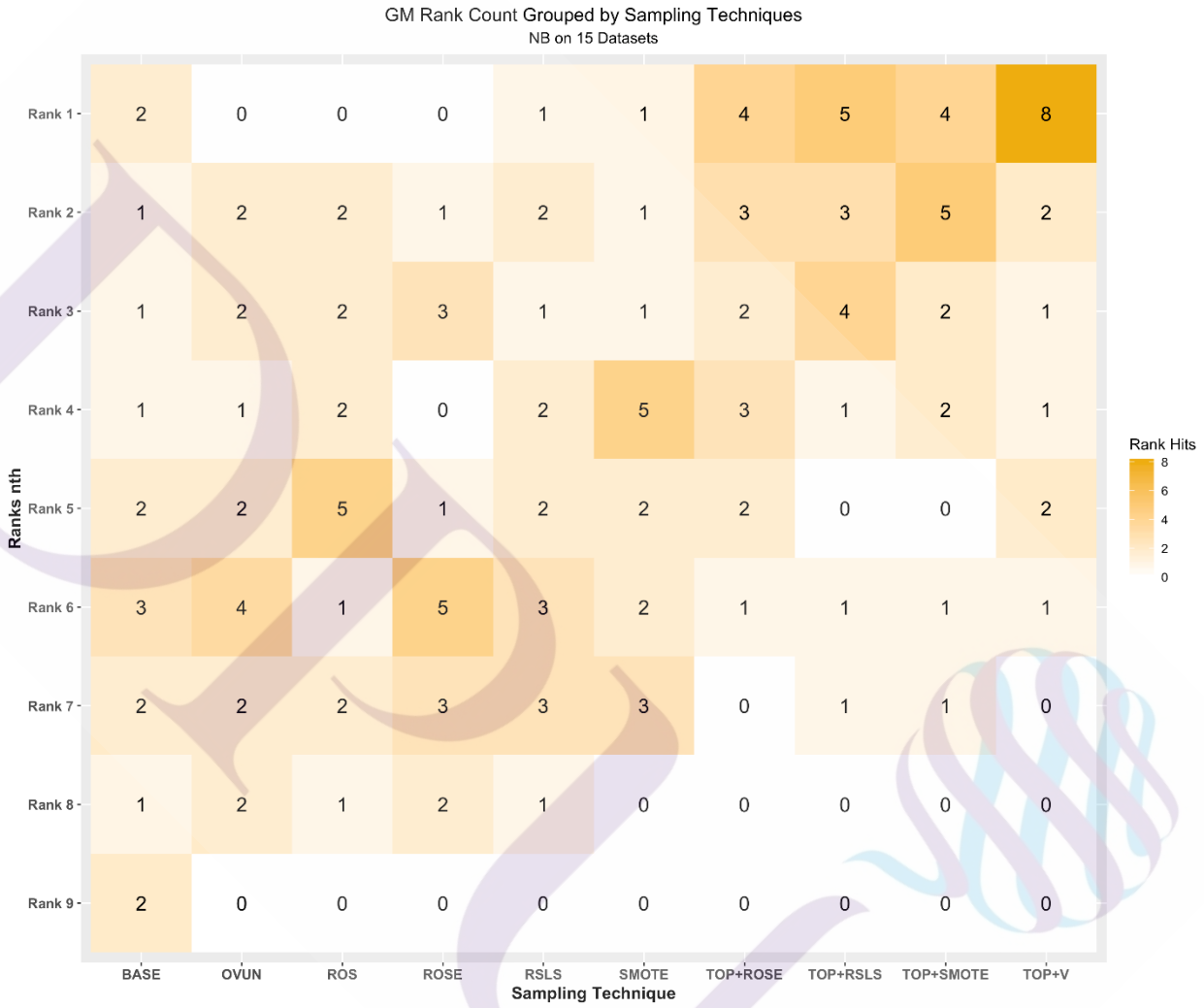
ภาพที่ 7.6.4 โมเดล XGB บนข้อมูลทุกชุด



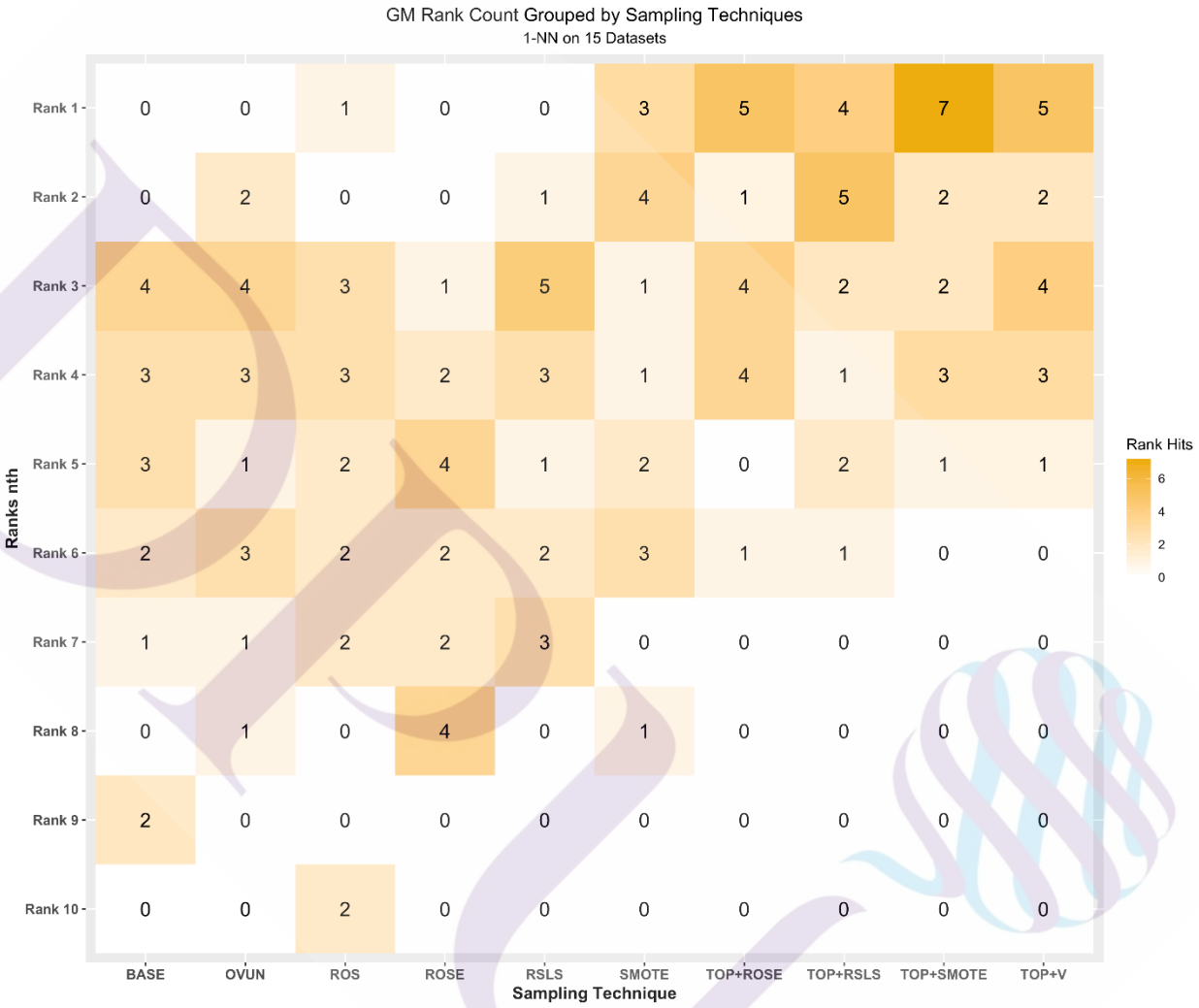
ภาพที่ 7.6.5 โมเดล LR บนข้อมูลทุกชุด



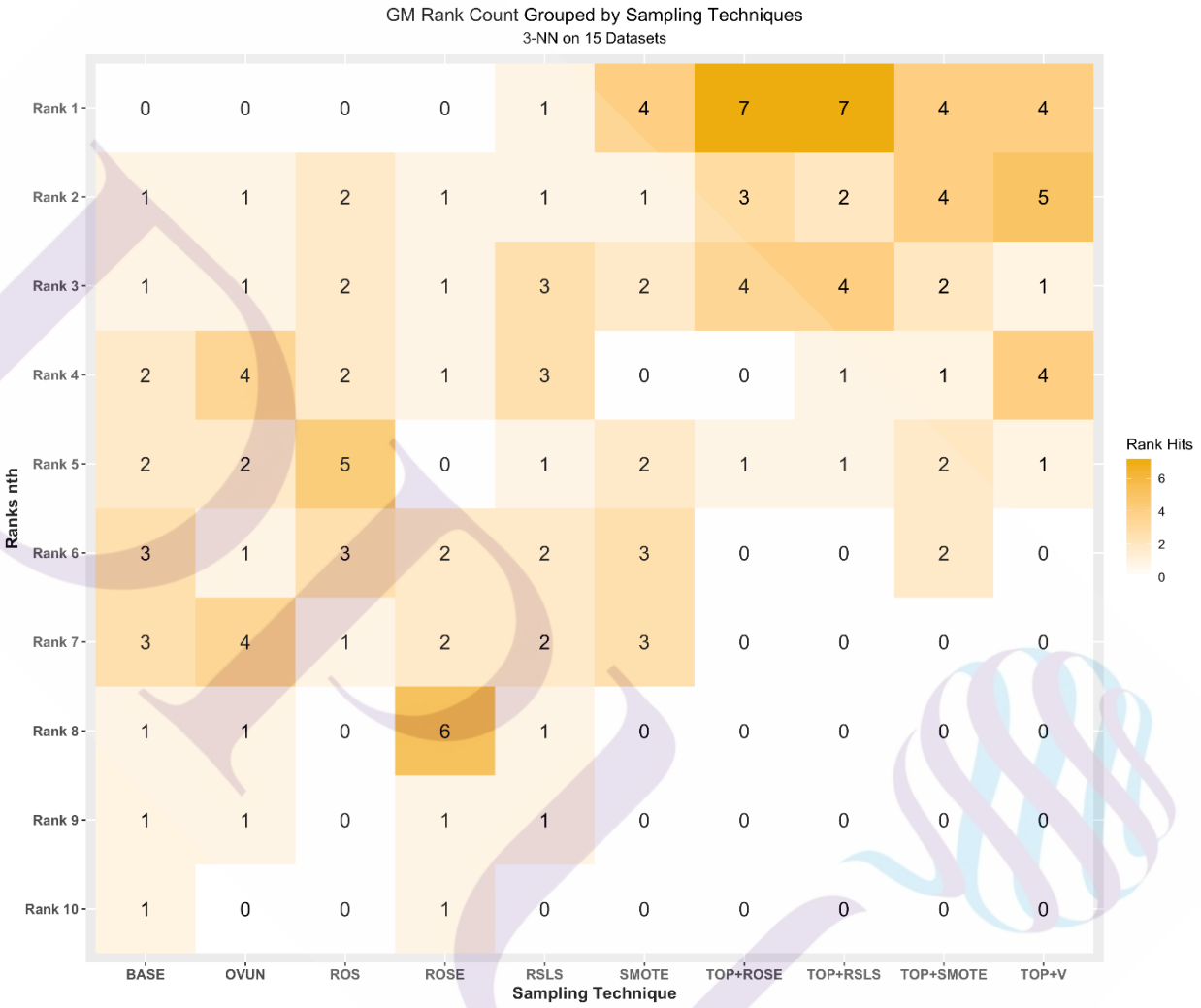
ภาพที่ 7.6.6 โมเดล NB บนข้อมูลทุกชุด



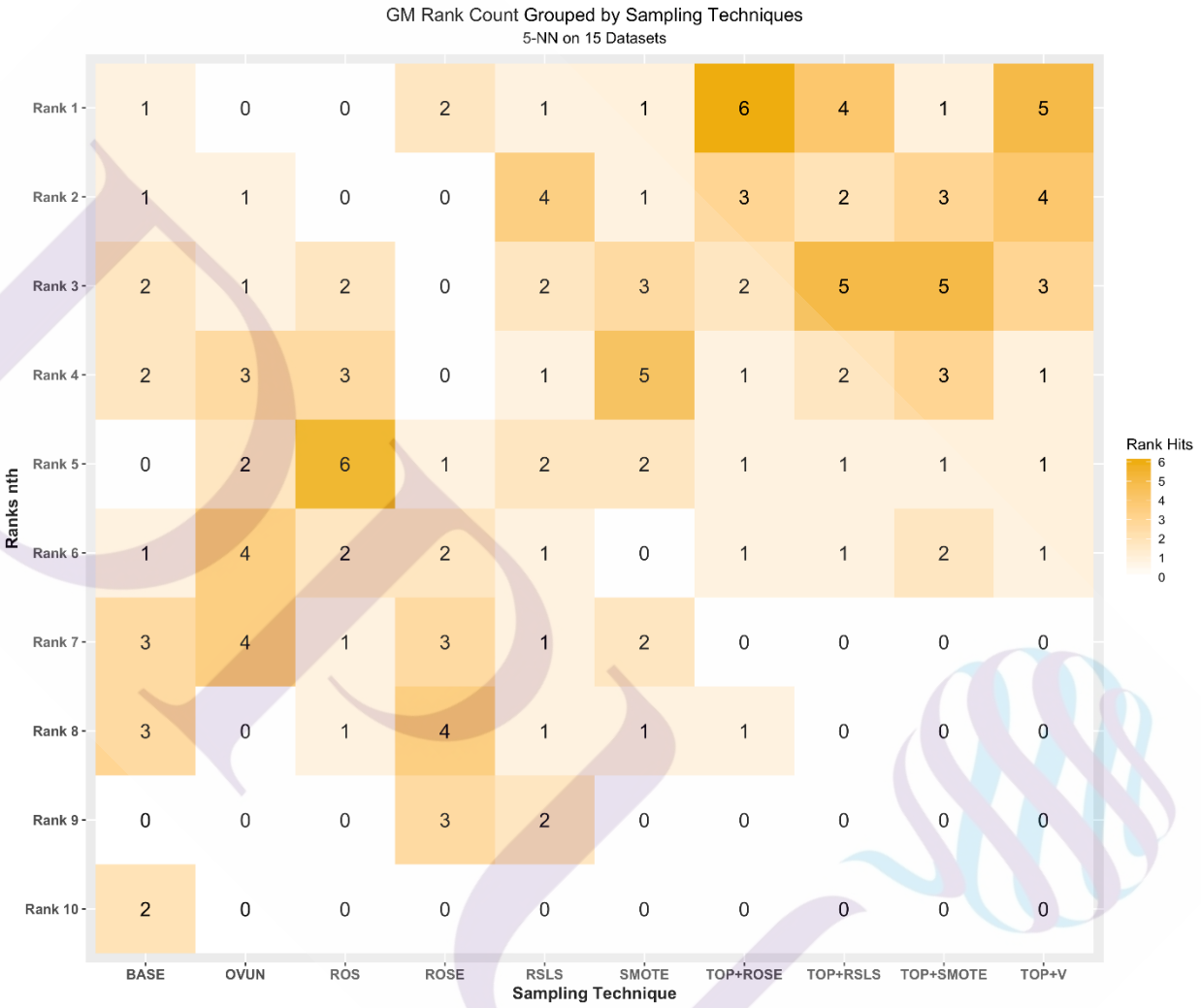
ภาพที่ 7.6.7 โมเดล 1-NN บนข้อมูลทุกชุด



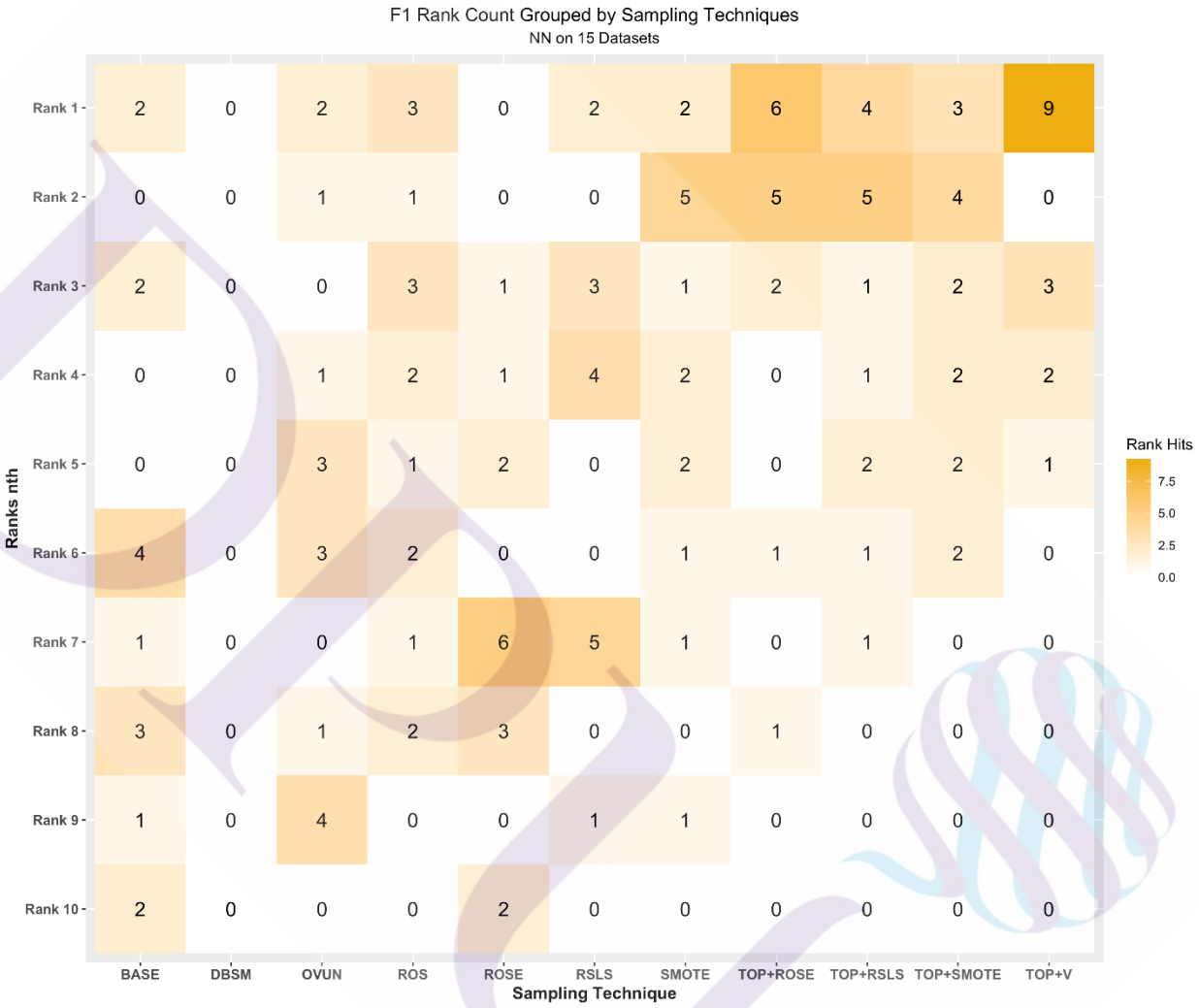
ภาพที่ 7.6.8 โมเดล 3-NN บนข้อมูลทุกชุด



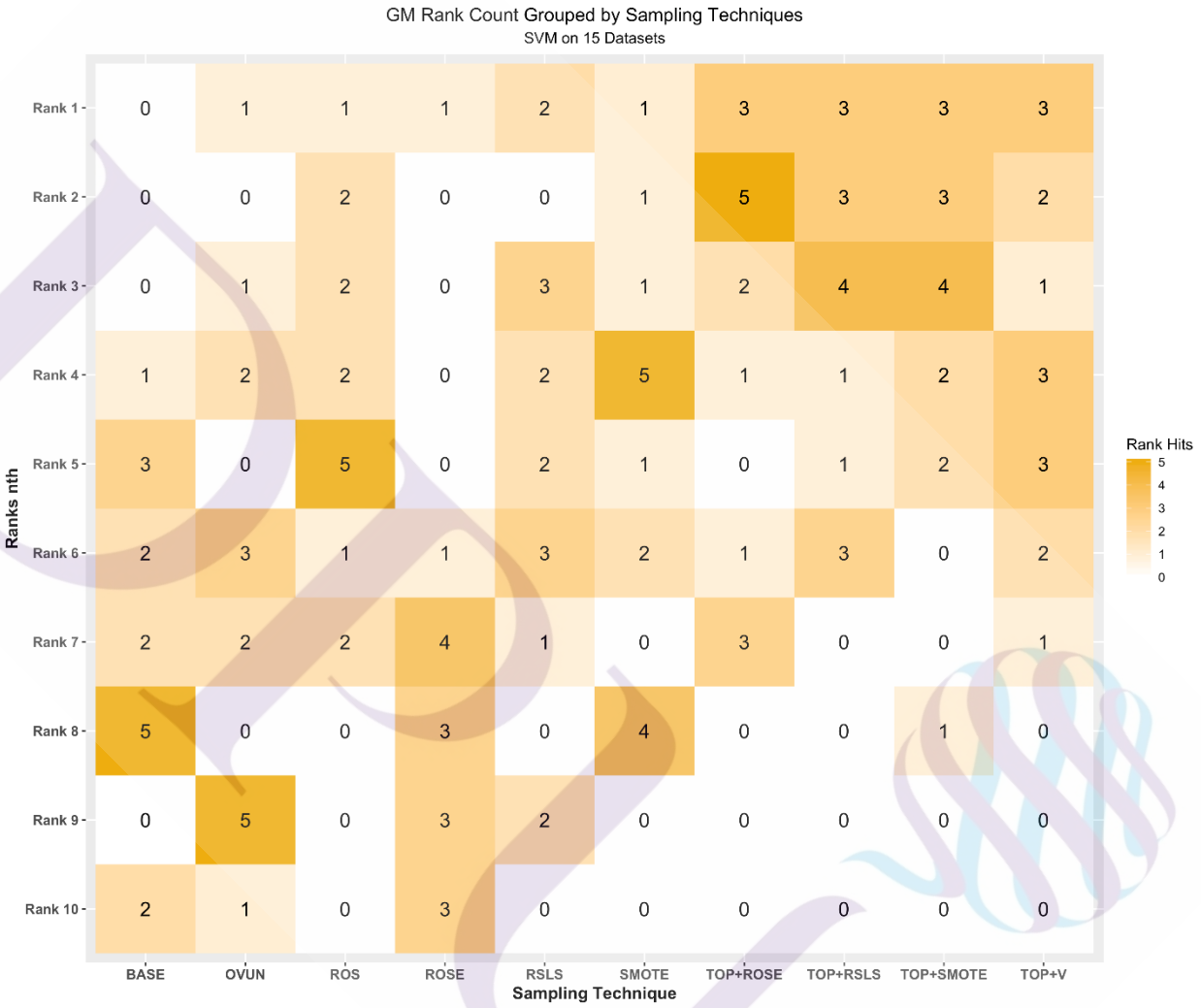
ภาพที่ 7.6.9 โมเดล 5-NN บนข้อมูลทุกชุด



ภาพที่ 7.6.10 โมเดล NN บนข้อมูลทุกชุด

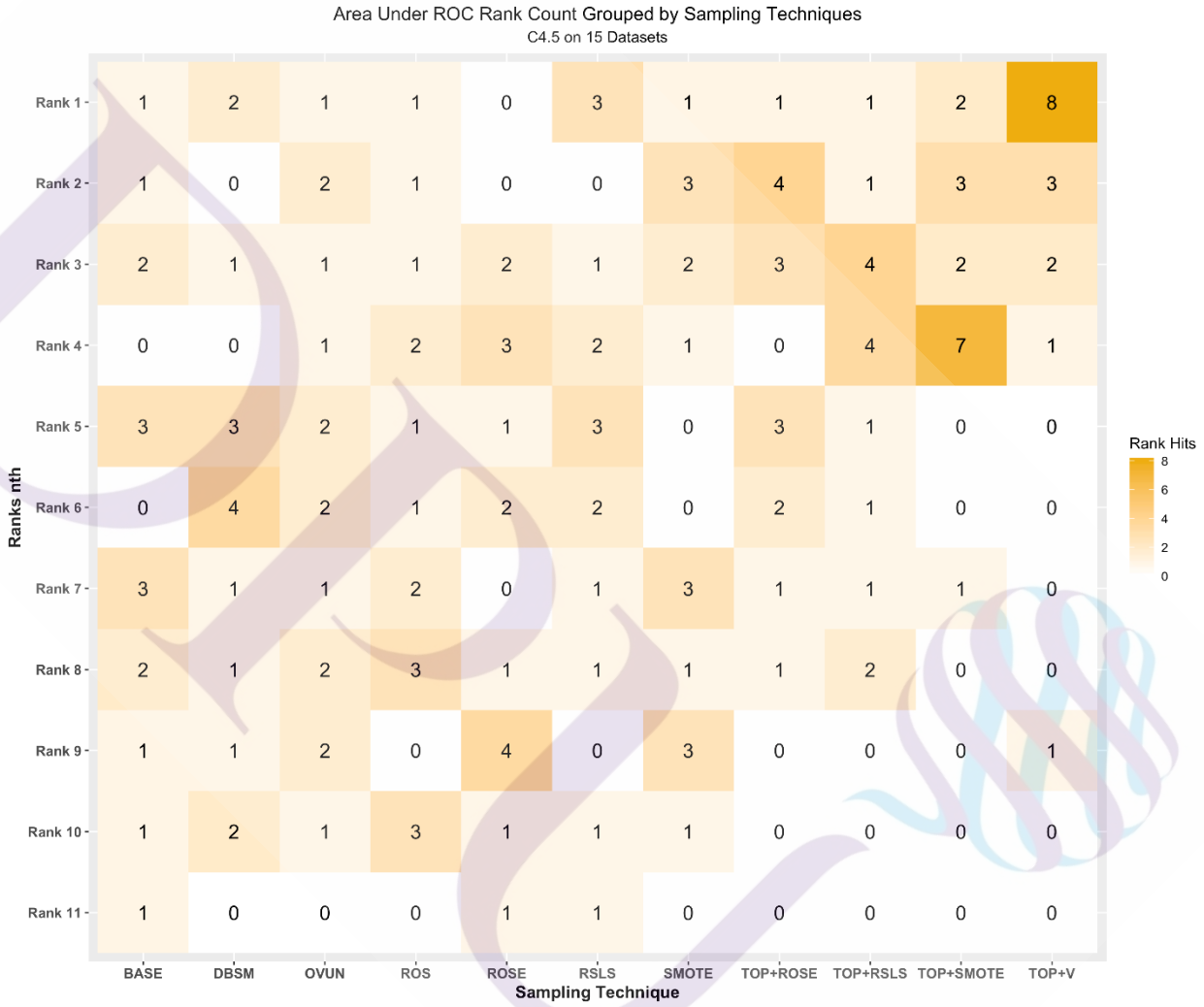


ภาพที่ 7.6.11 โมเดล SVM บนข้อมูลทุกชุด

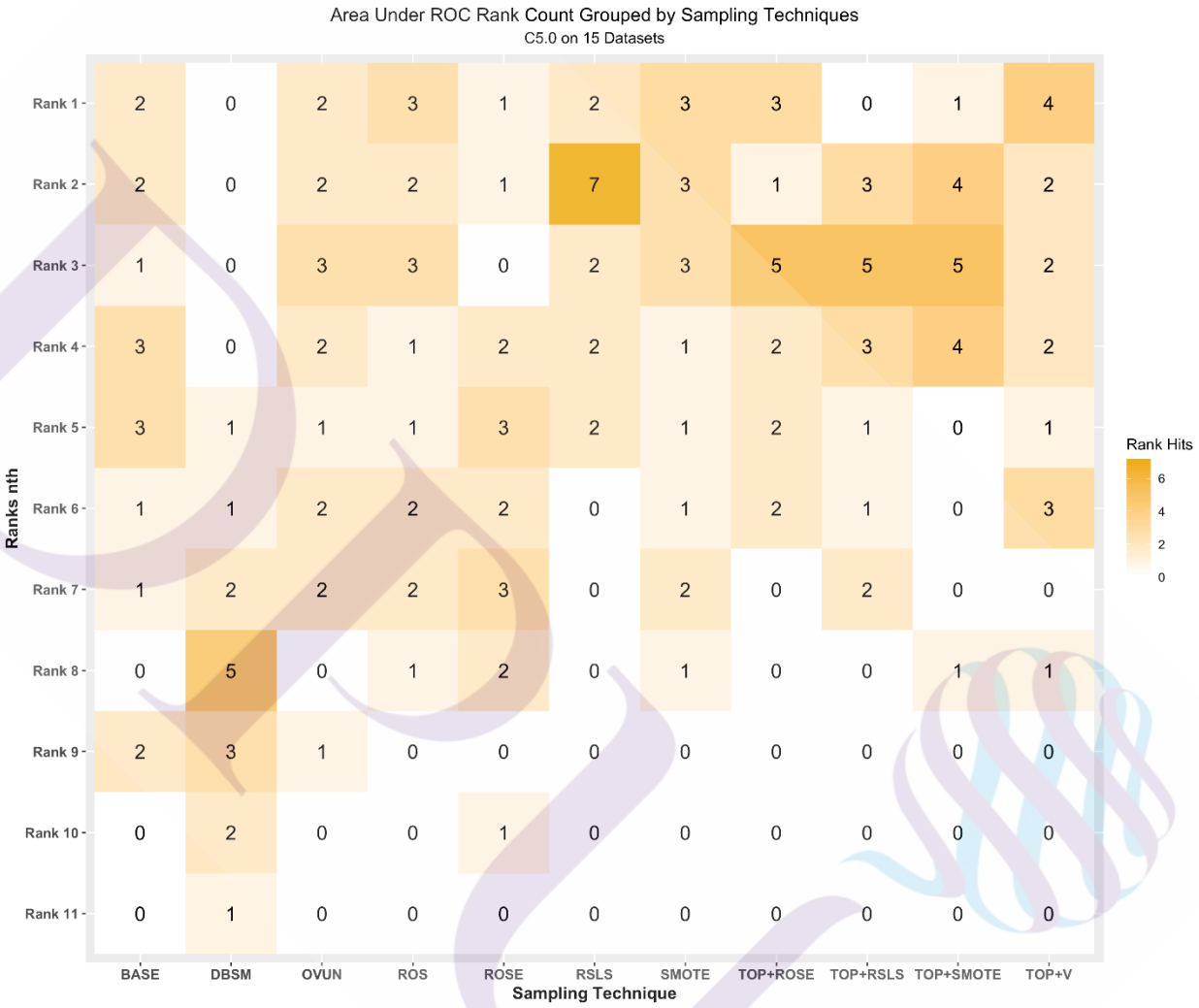


7.7 การสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพของโมเดลด้วยหน่วยวัด AUROC

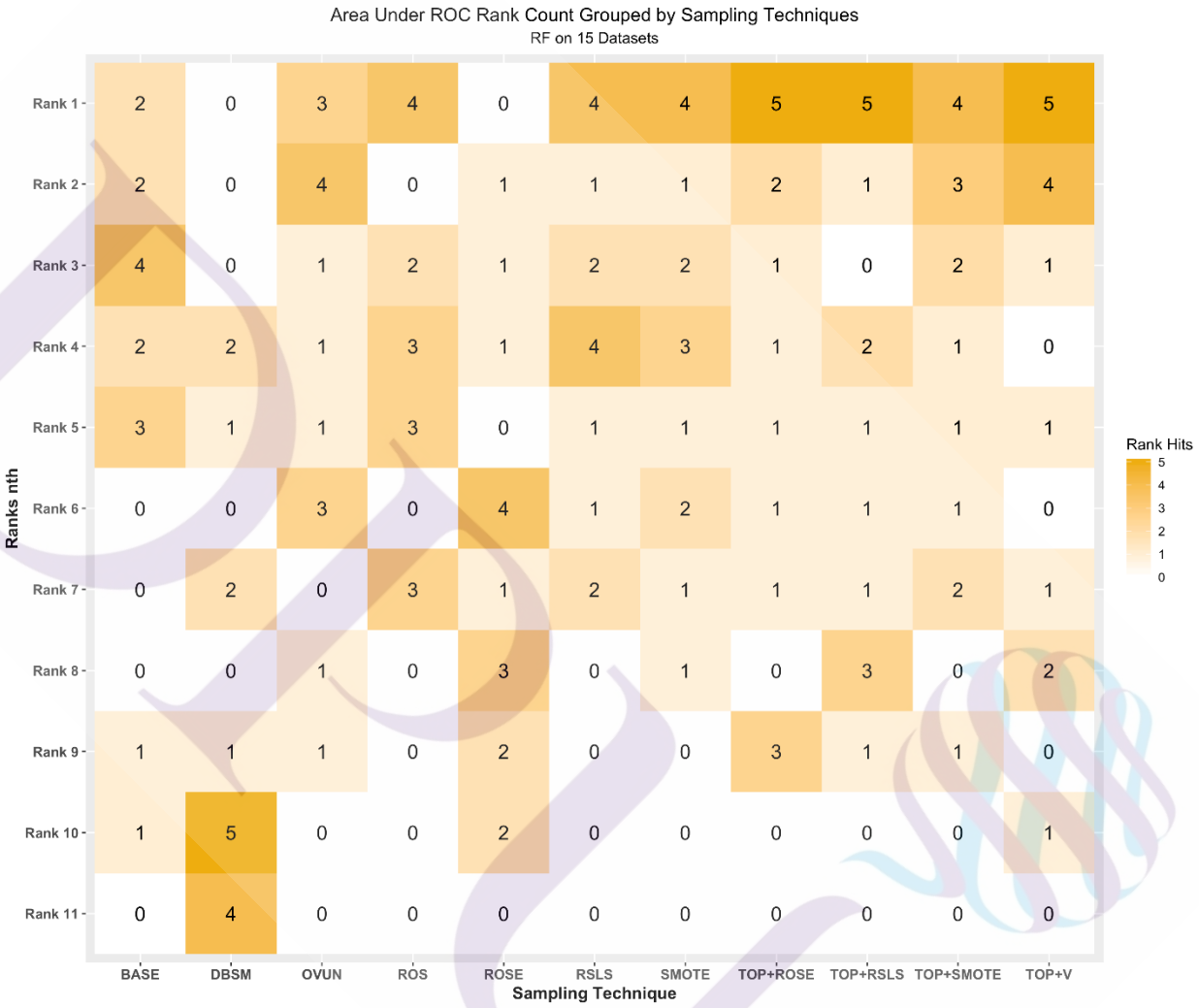
ภาพที่ 7.7.1 โมเดล C4.5 บนข้อมูลทุกชุด



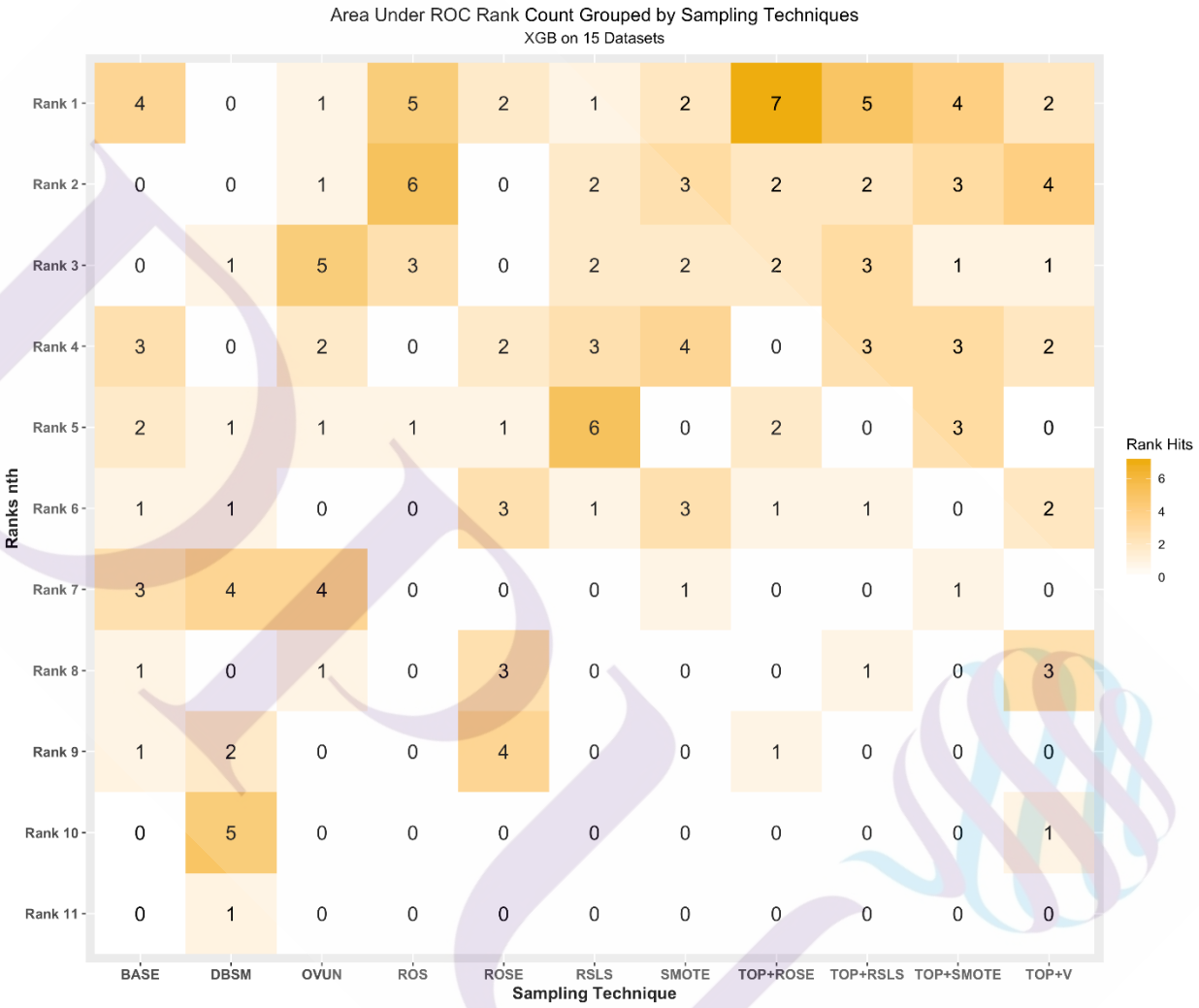
ภาพที่ 7.7.2 โมเดล C5.0 บนข้อมูลทุกชุด



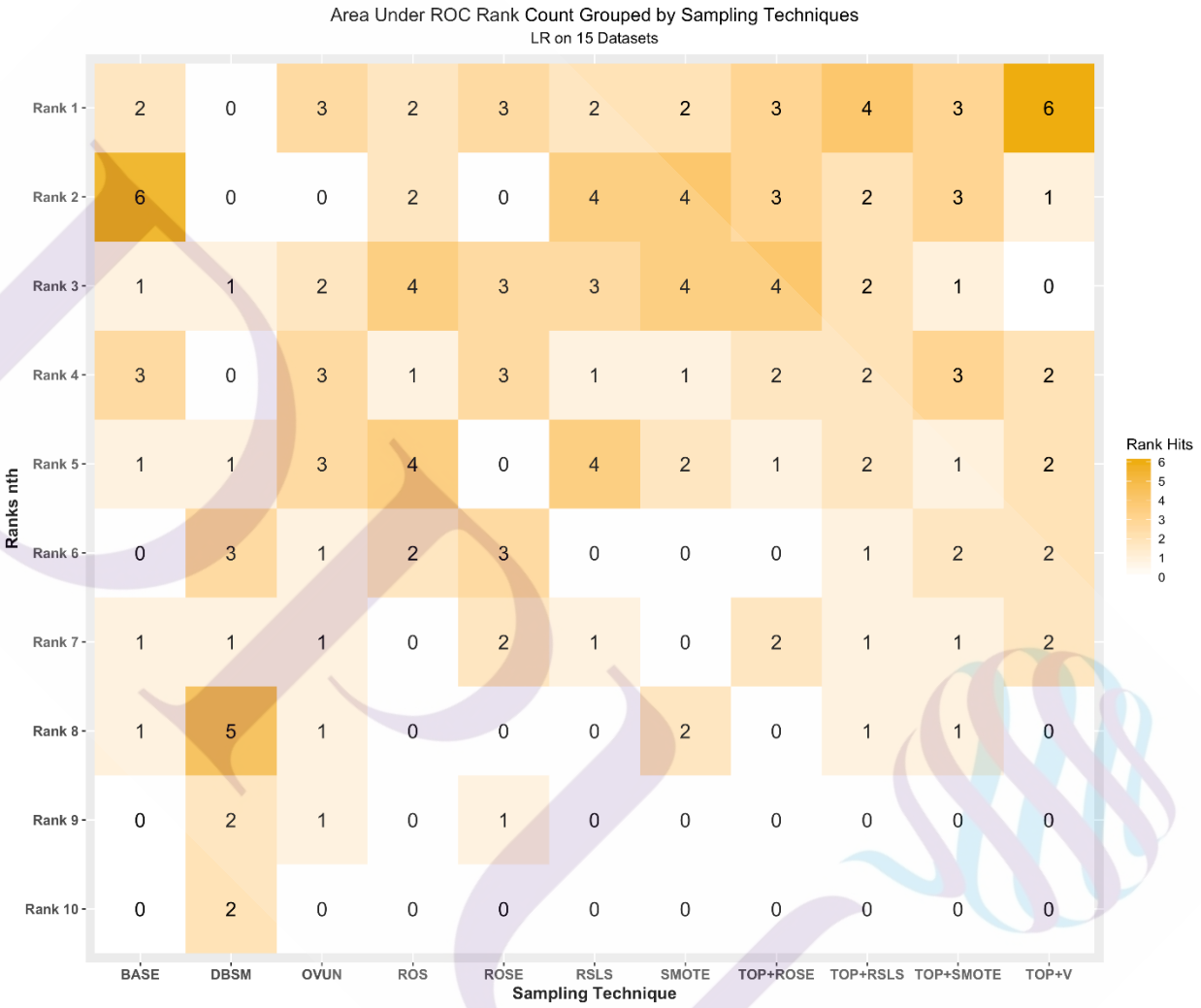
ภาพที่ 7.7.3 โมเดล RF บนข้อมูลทุกชุด



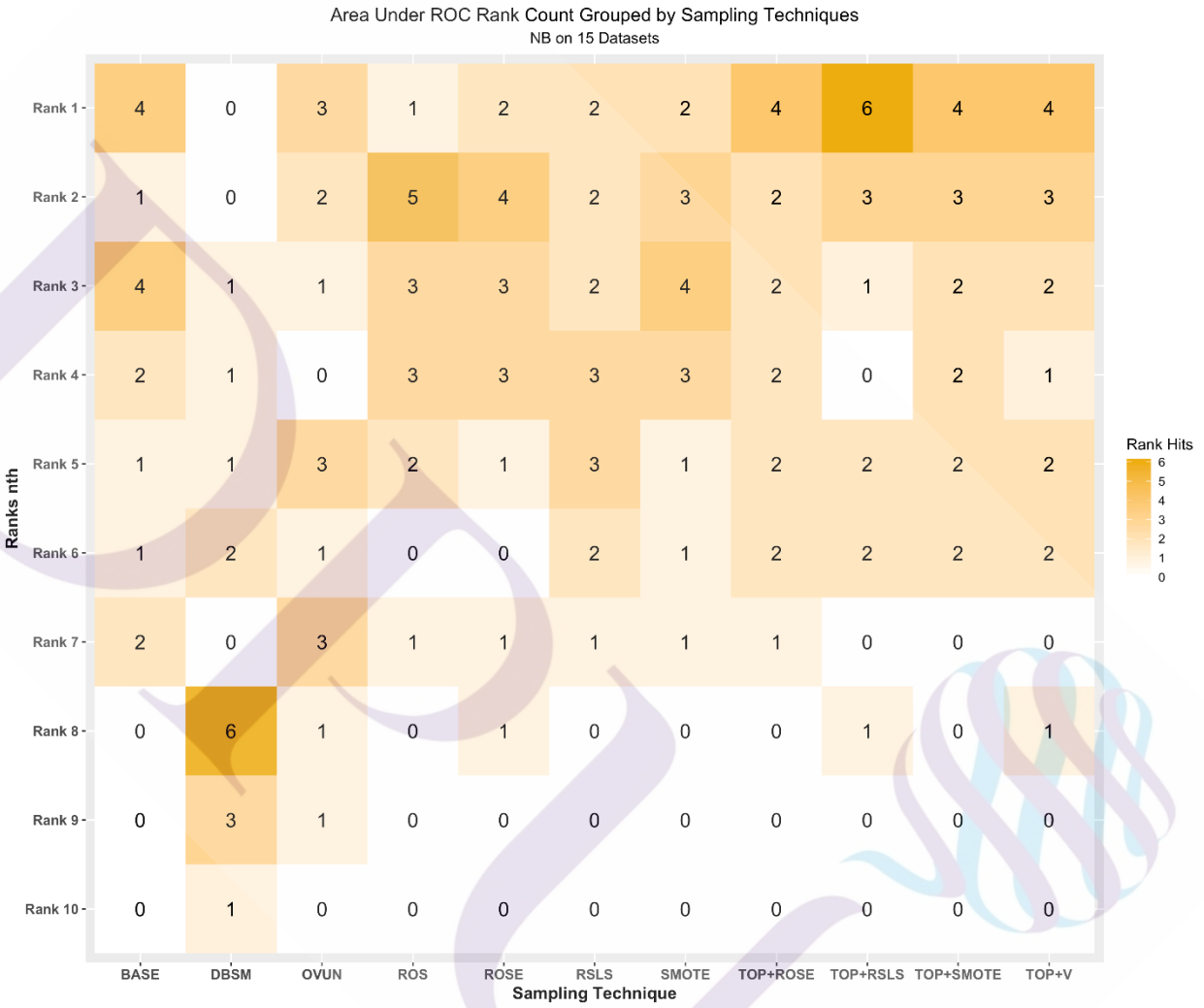
ภาพที่ 7.7.4 โมเดล XGB บนข้อมูลทุกชุด



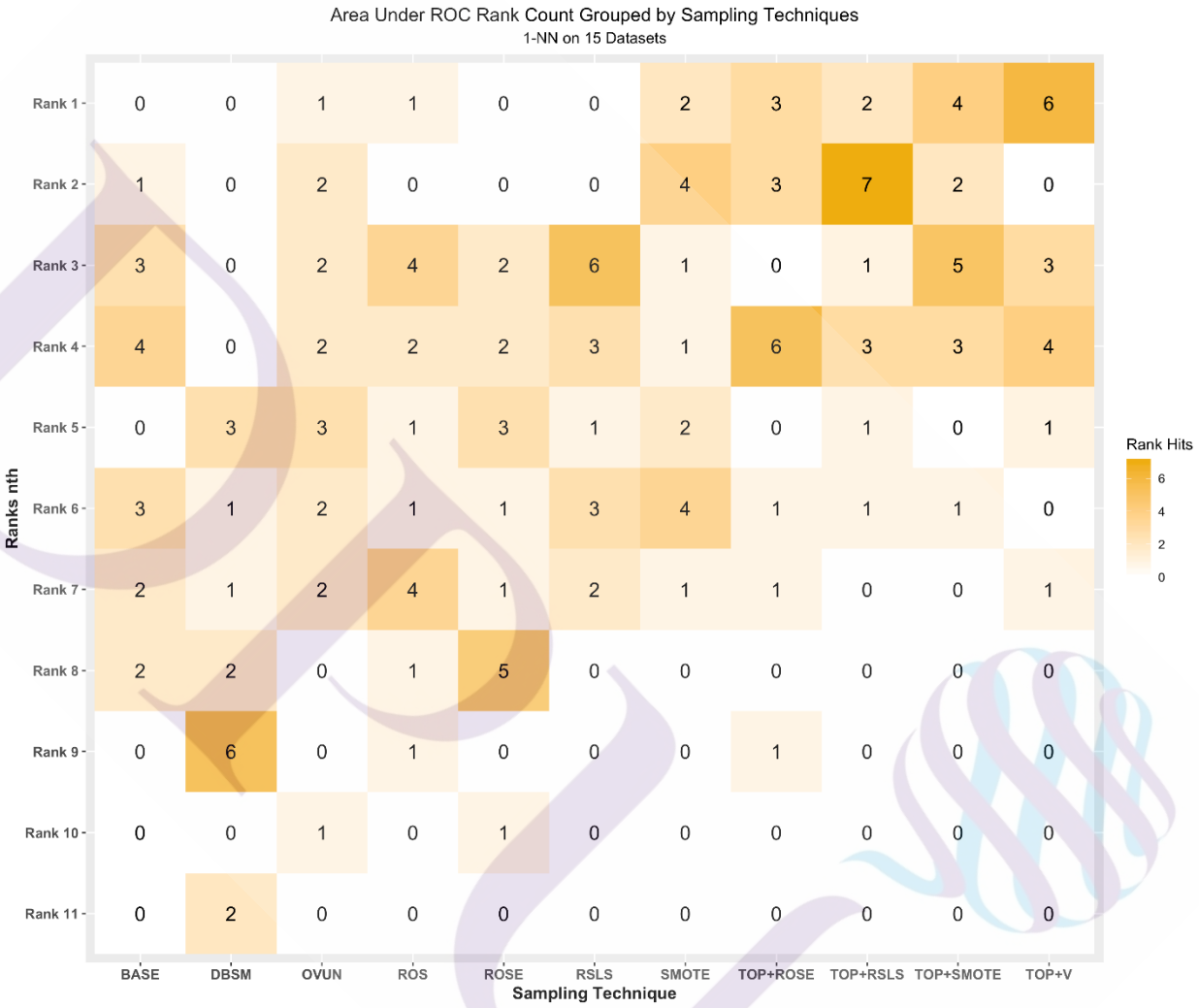
ภาพที่ 7.7.5 โมเดล LR บนข้อมูลทุกชุด



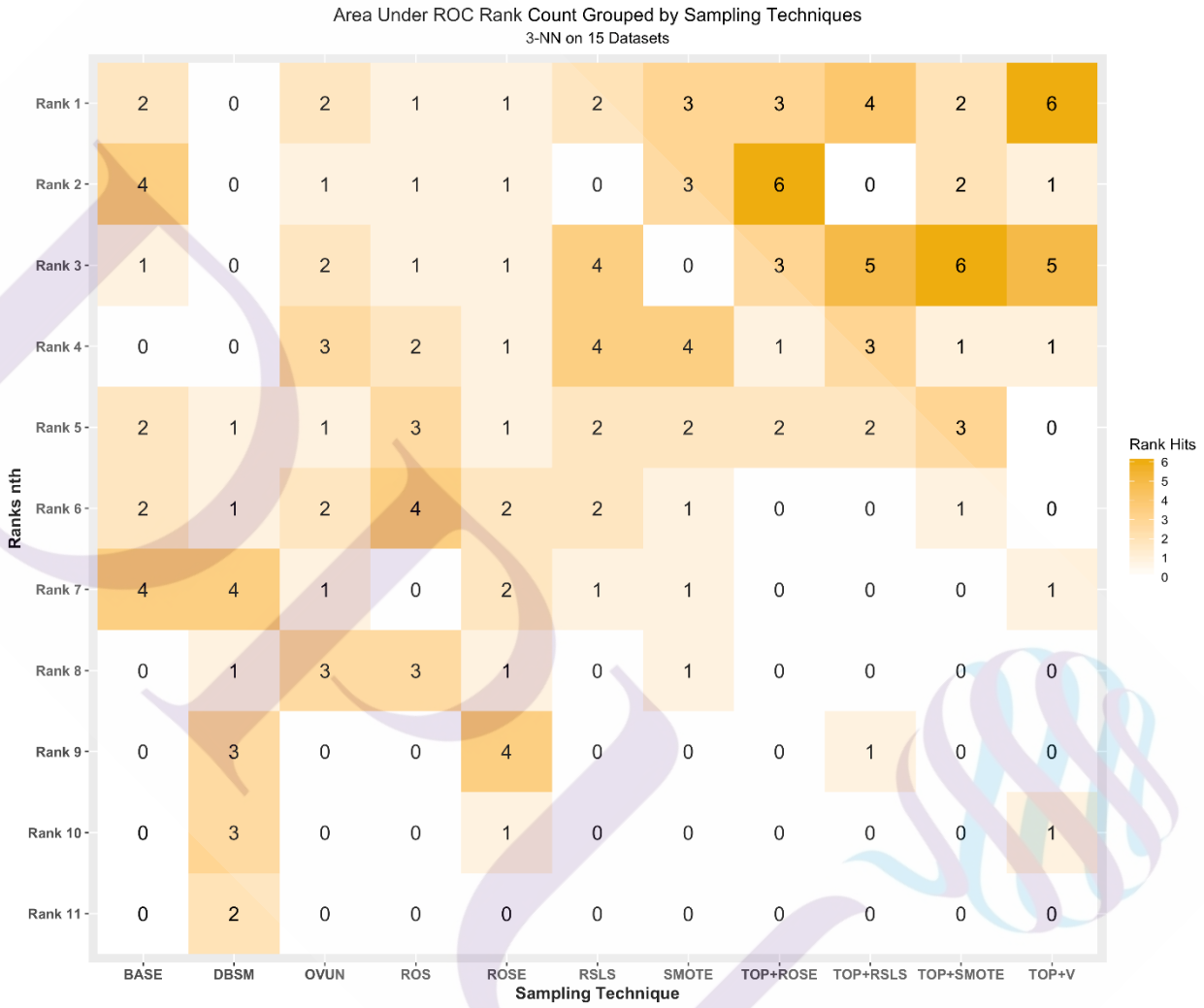
ภาพที่ 7.7.6 โมเดล NB บนข้อมูลทุกชุด



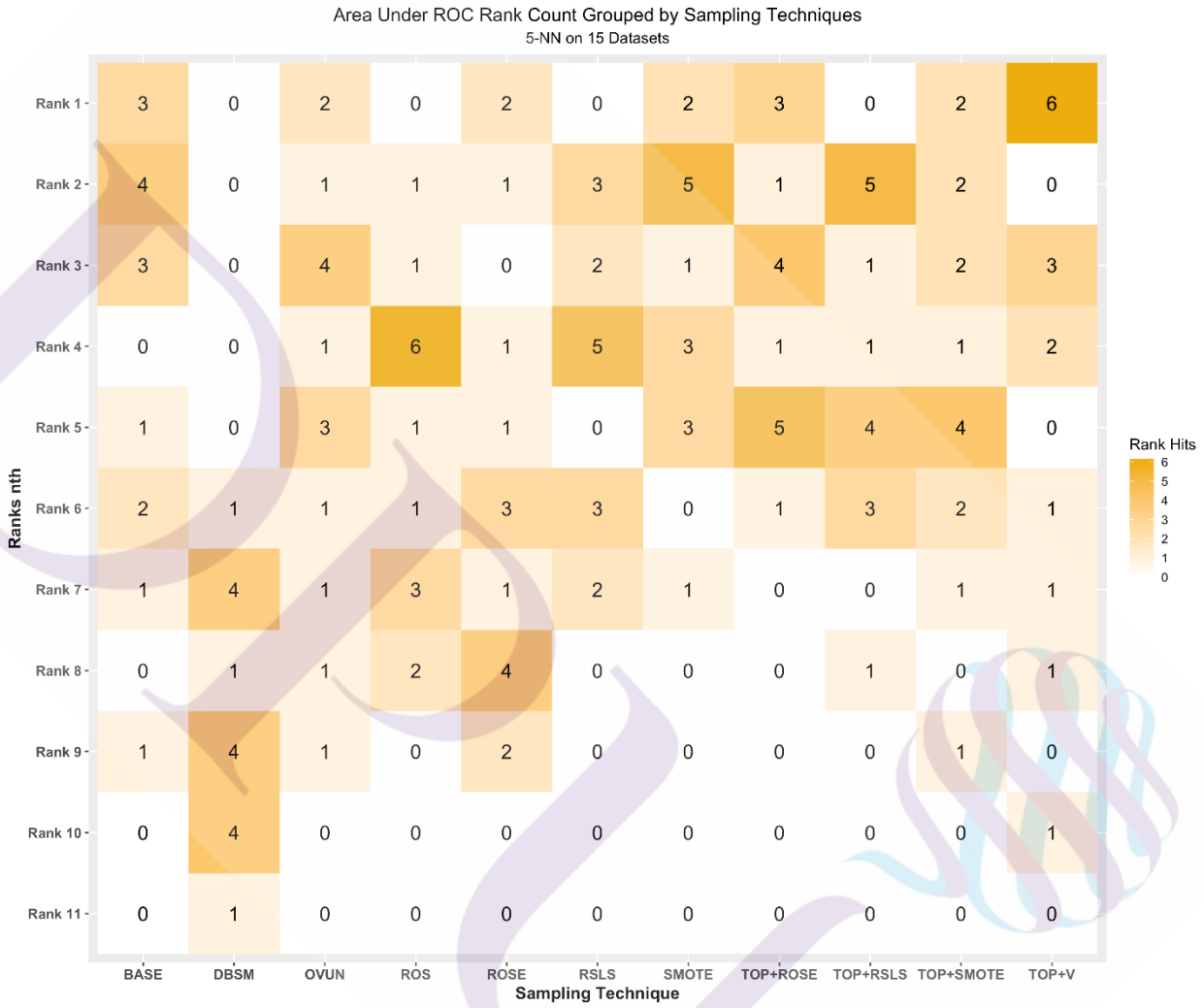
ภาพที่ 7.7.7 โมเดล 1-NN บนข้อมูลทุกชุด



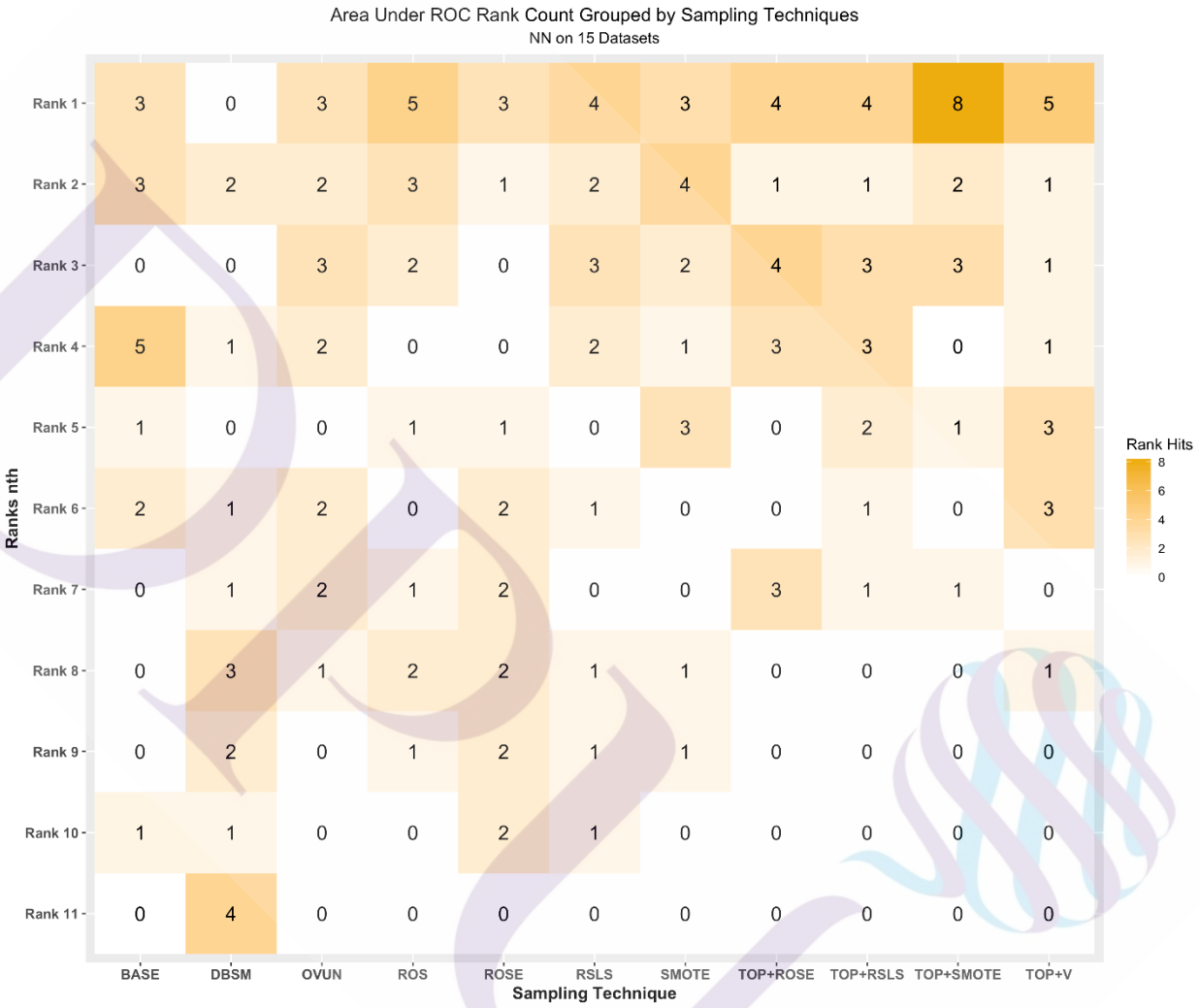
ภาพที่ 7.7.8 โมเดล 3-NN บนข้อมูลทุกชุด



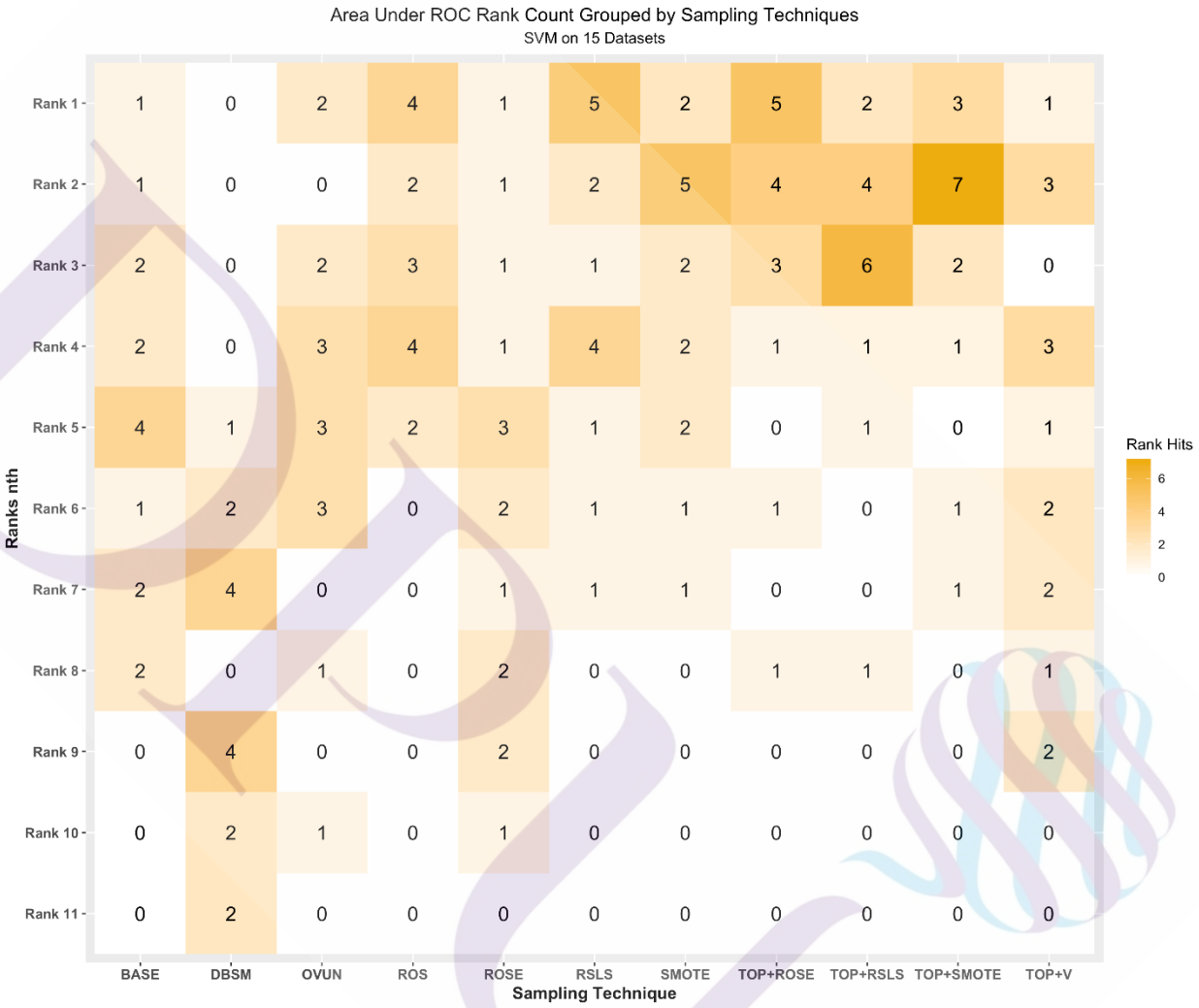
ภาพที่ 7.7.9 โมเดล 5-NN บนข้อมูลทุกชุด



ภาพที่ 7.7.10 โมเดล NN บนข้อมูลทุกชุด

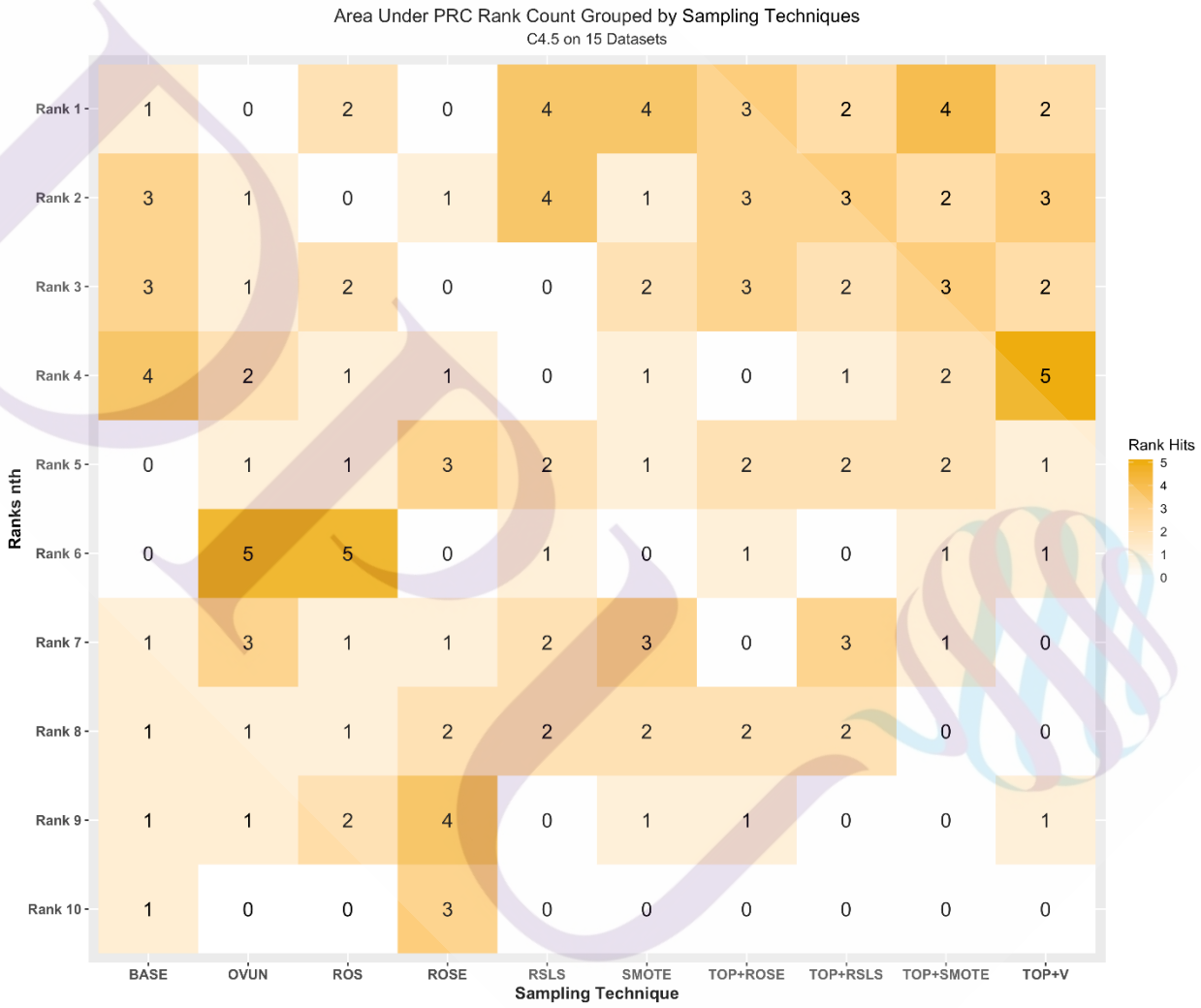


ภาพที่ 7.7.11 โมเดล SVM บนข้อมูลทุกชุด

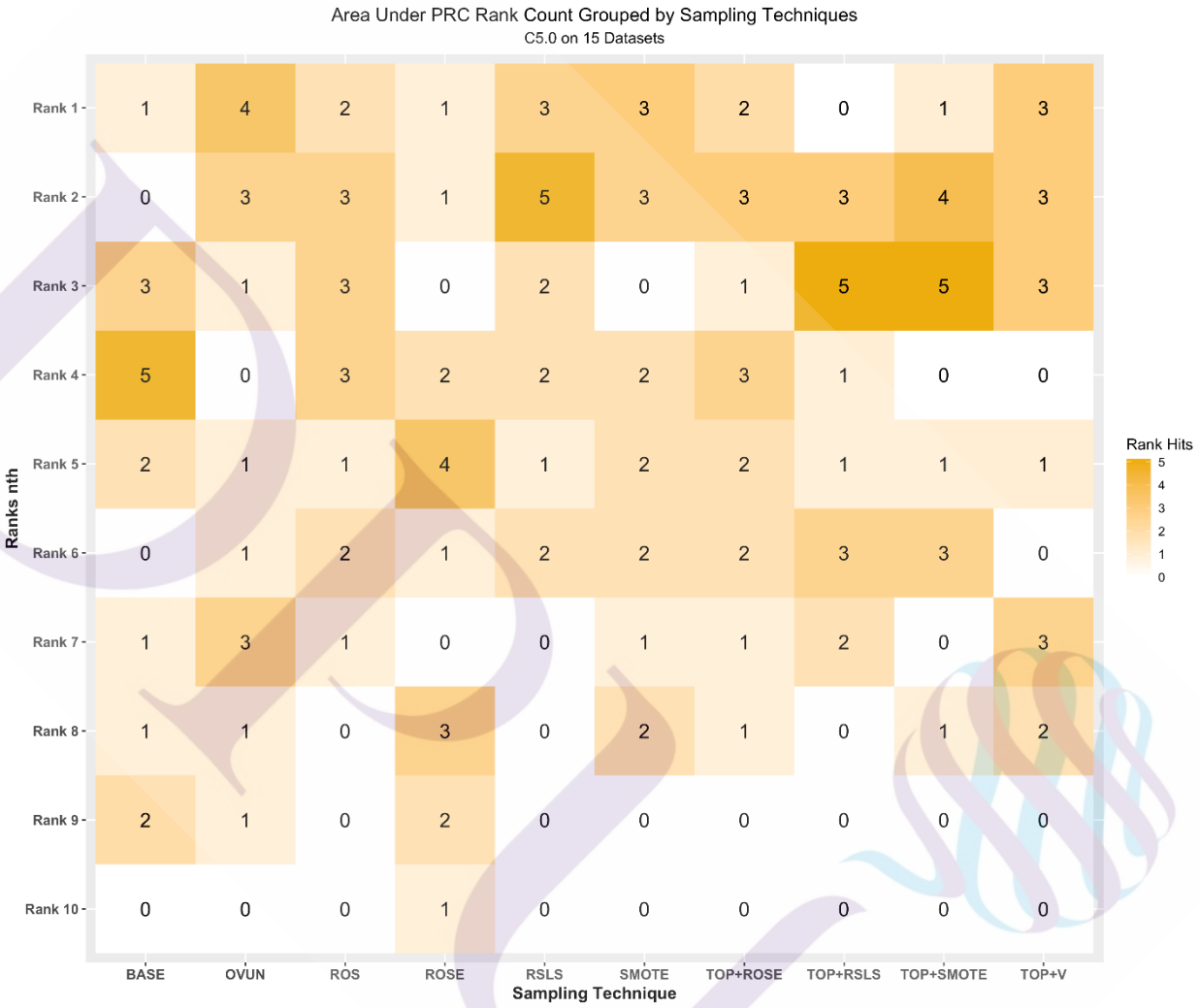


7.8 การสรุปจำนวนครั้งของการจัดอันดับโดยคิดจากค่าเฉลี่ยของประสิทธิภาพของโมเดลด้วยหน่วยวัด AUPRC

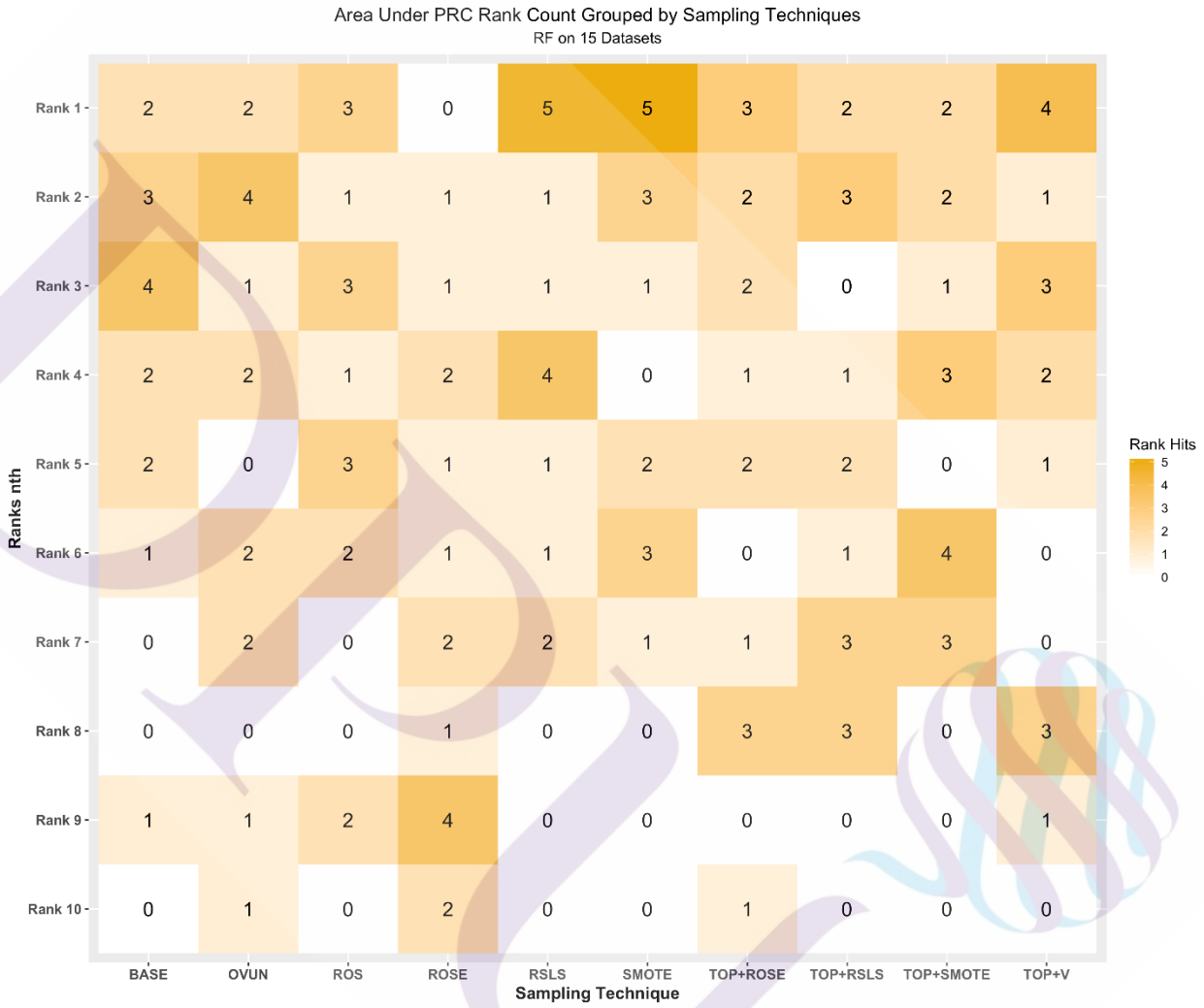
ภาพที่ 7.8.1 โมเดล C4.5 บนข้อมูลทุกชุด



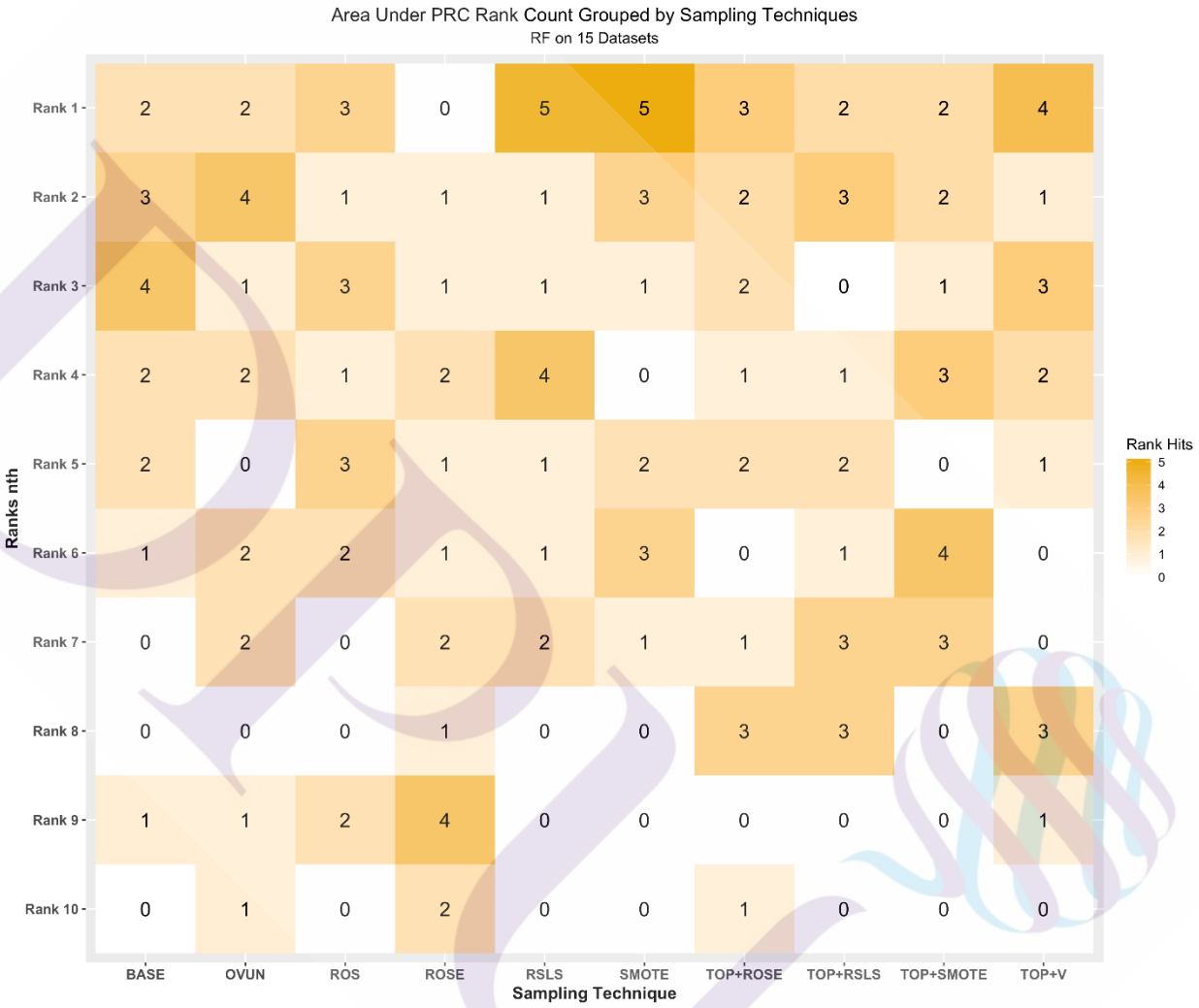
ภาพที่ 7.8.2 โมเดล C5.0 บนข้อมูลทุกชุด



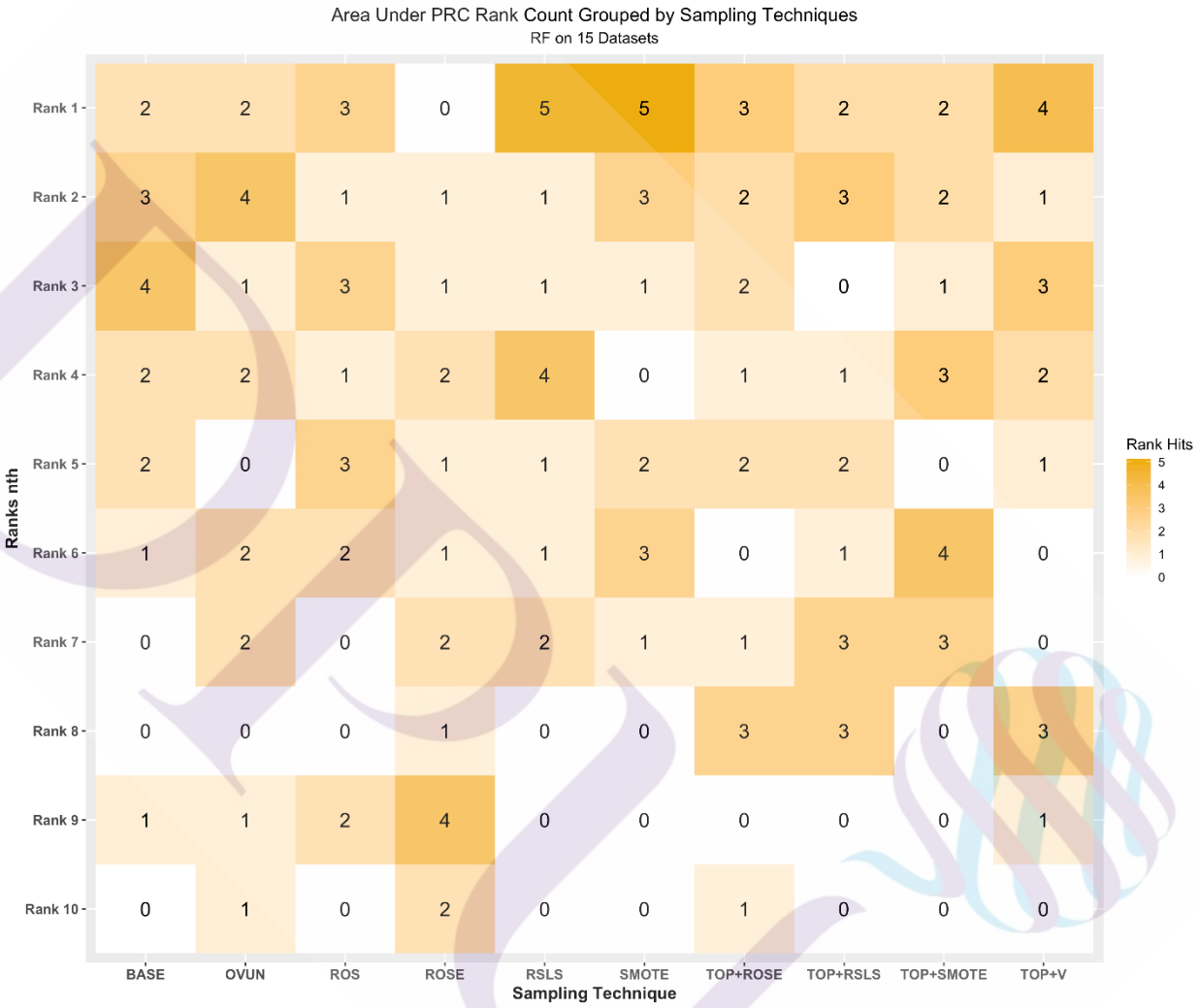
ภาพที่ 7.8.3 โมเดล RF บนข้อมูลทุกชุด



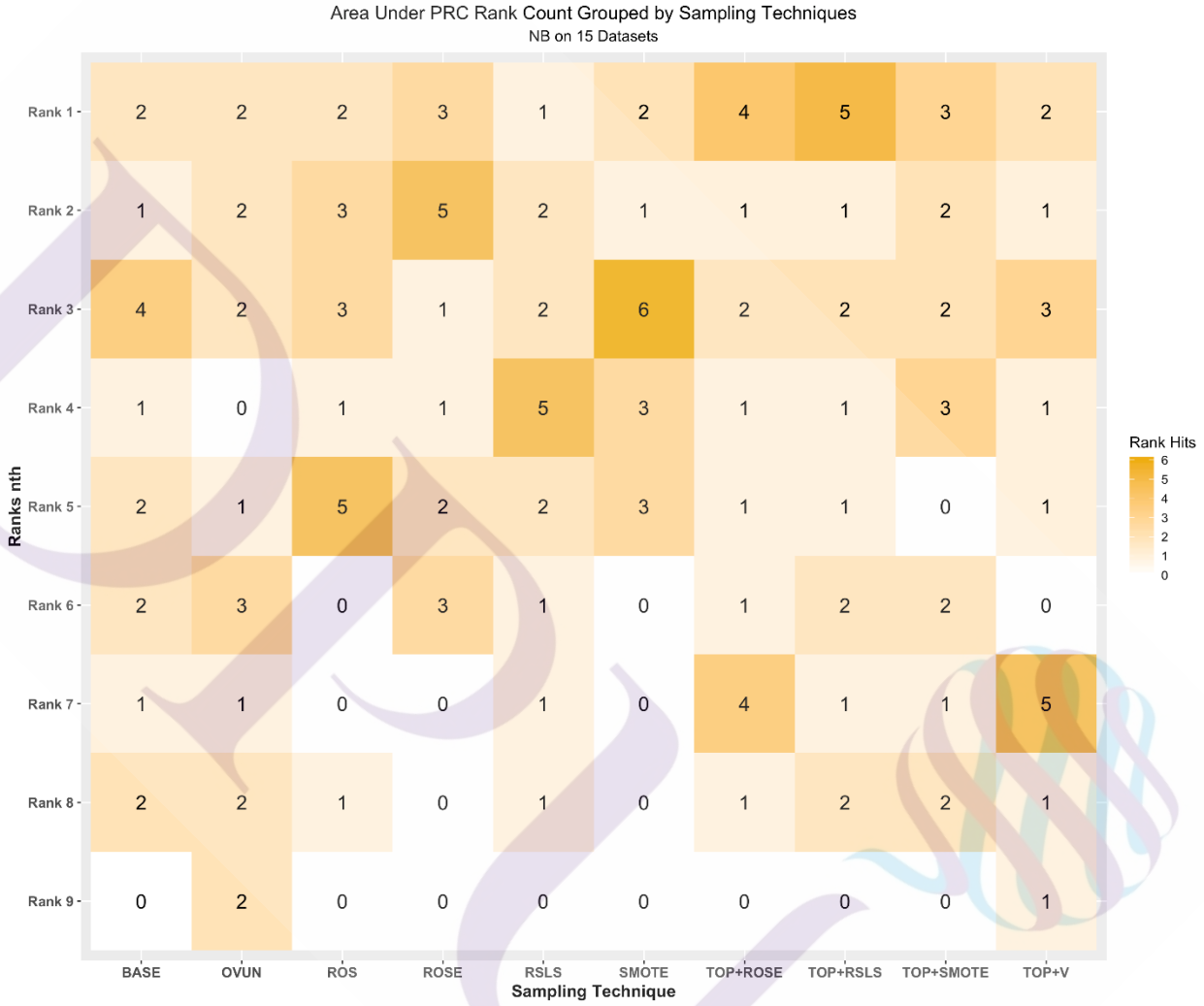
ภาพที่ 7.8.4 โมเดล XGB บนข้อมูลทุกชุด



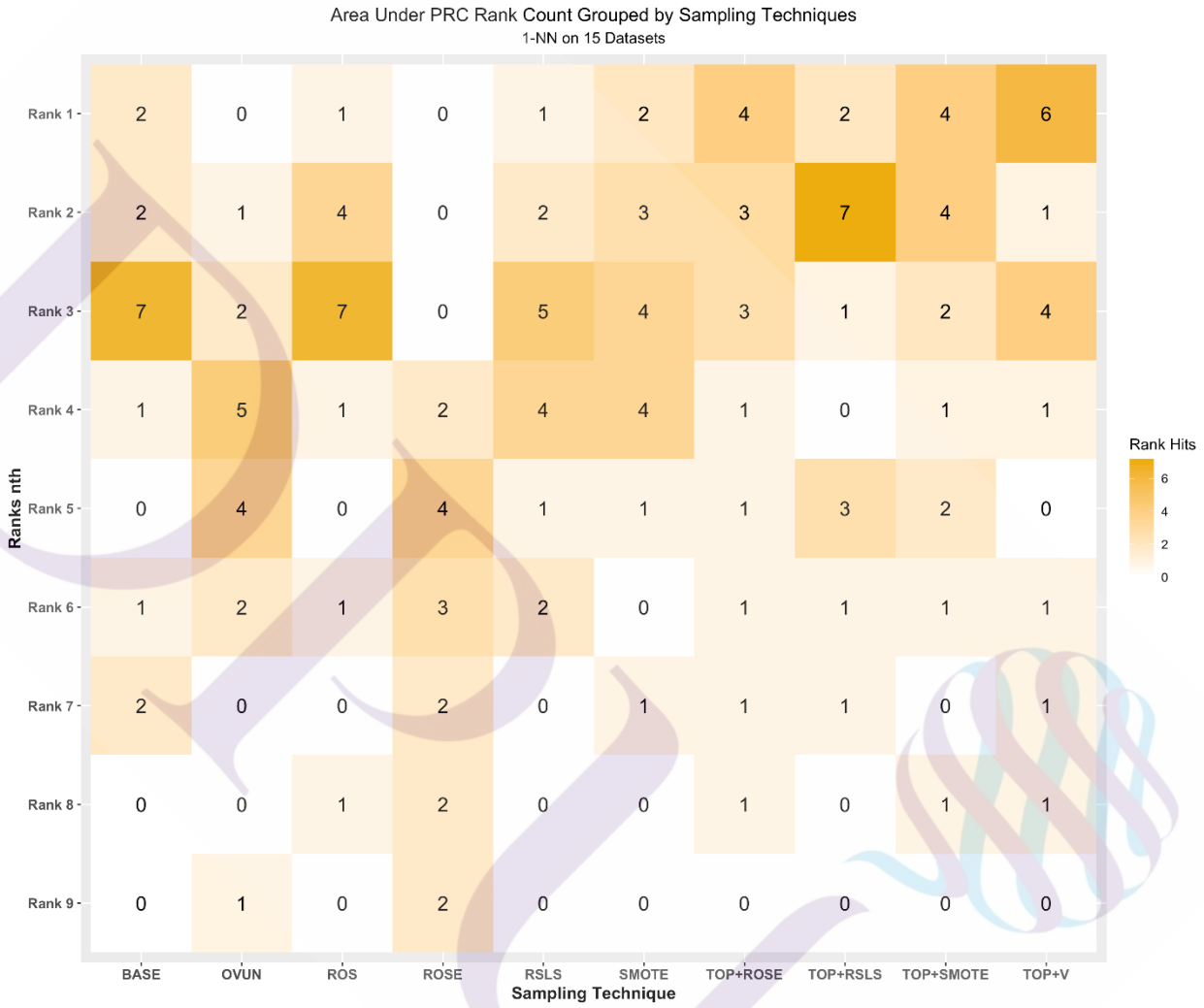
ภาพที่ 7.8.5 โมเดล LR บนข้อมูลทุกชุด



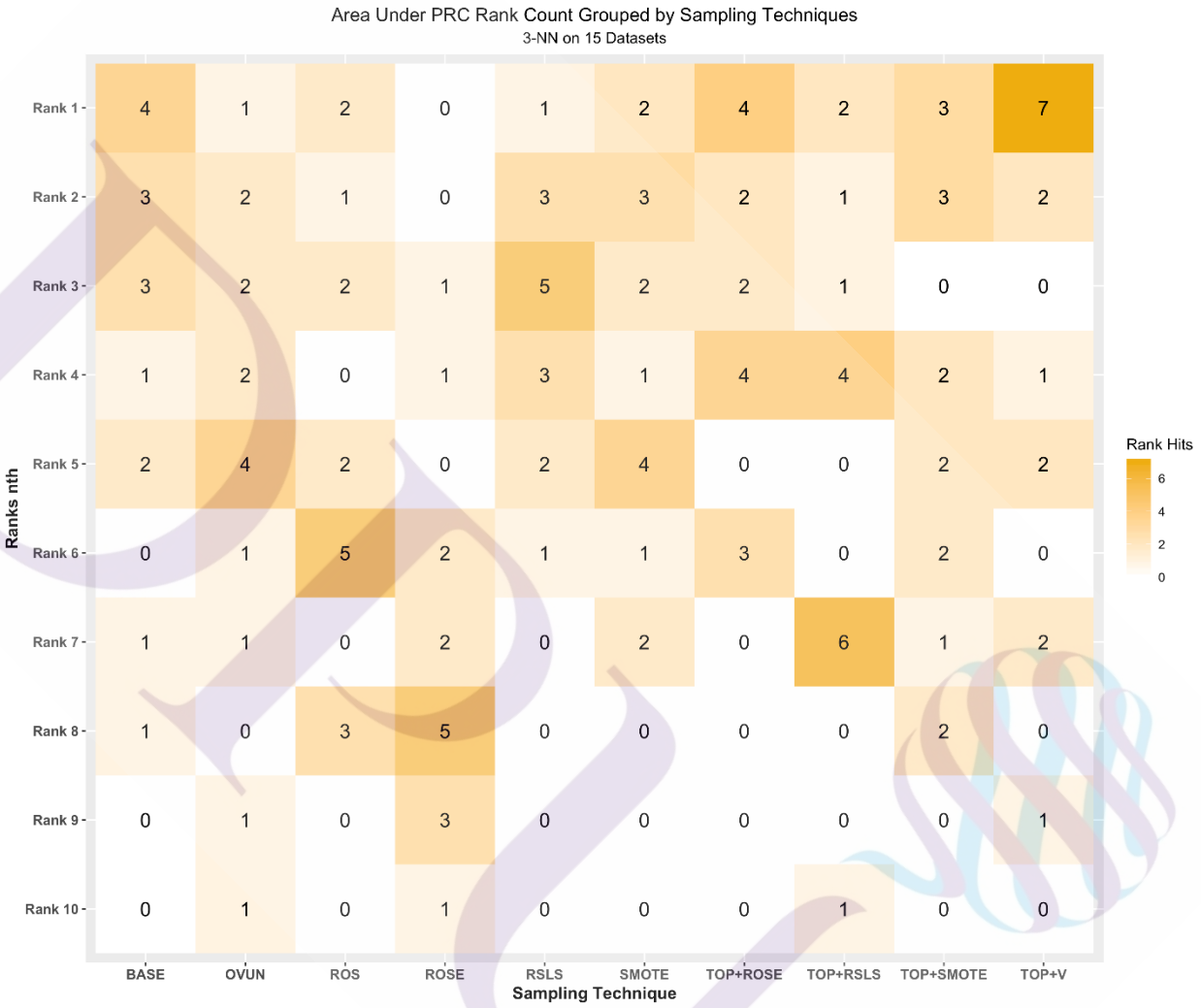
ภาพที่ 7.8.6 โมเดล NB บนข้อมูลทุกชุด



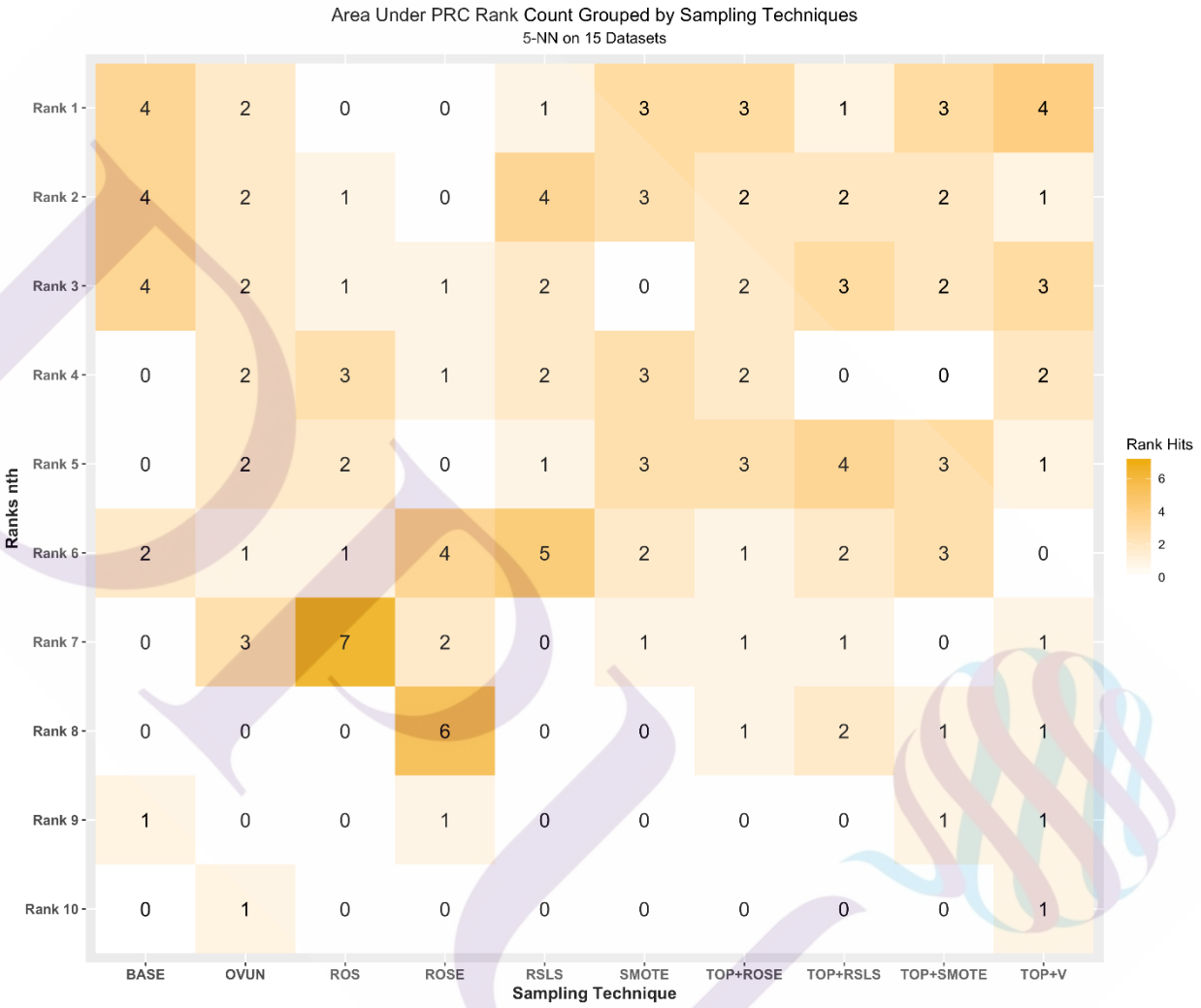
ภาพที่ 7.8.7 โมเดล 1-NN บนข้อมูลทุกชุด



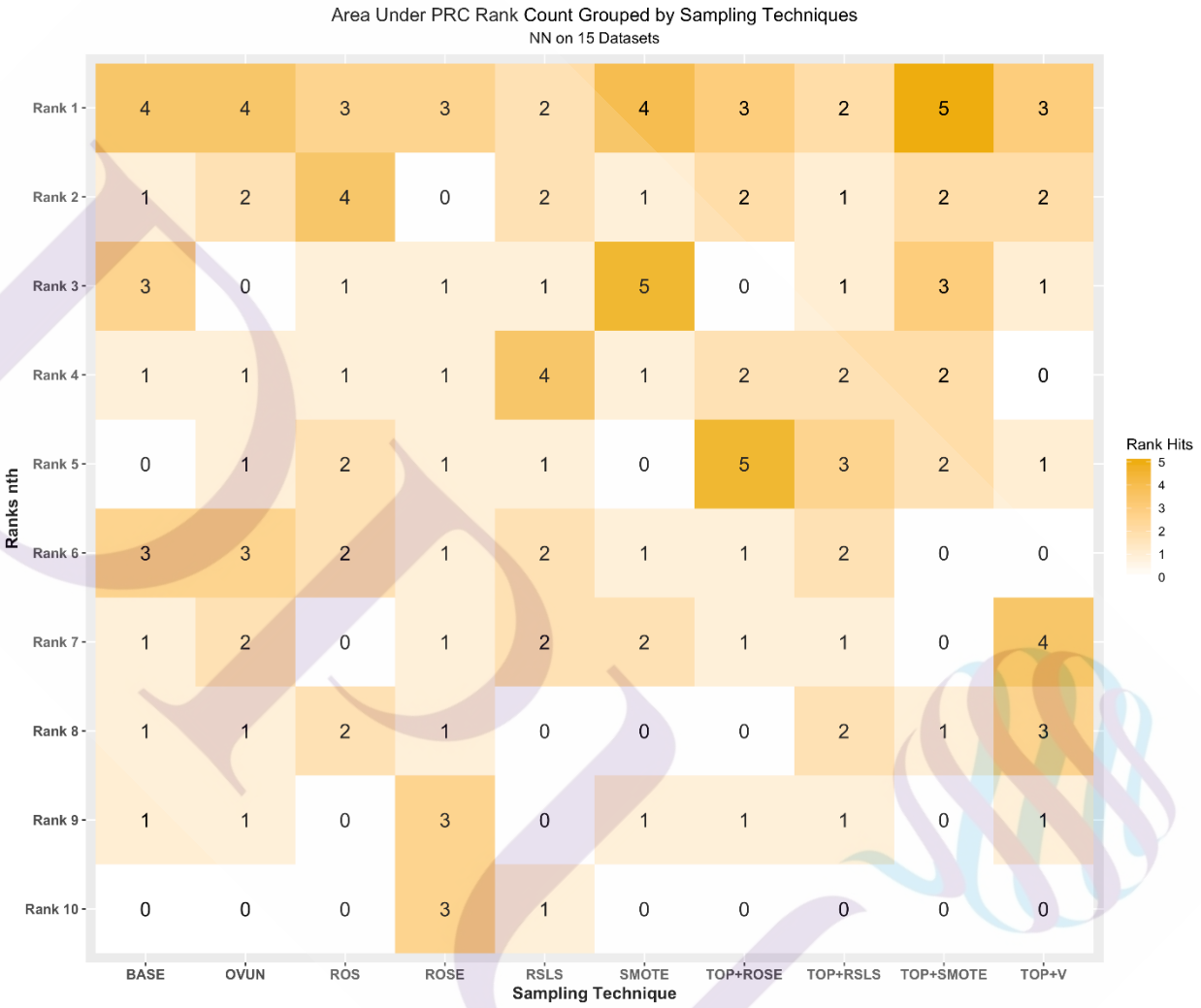
ภาพที่ 7.8.8 โมเดล 3-NN บนข้อมูลทุกชุด



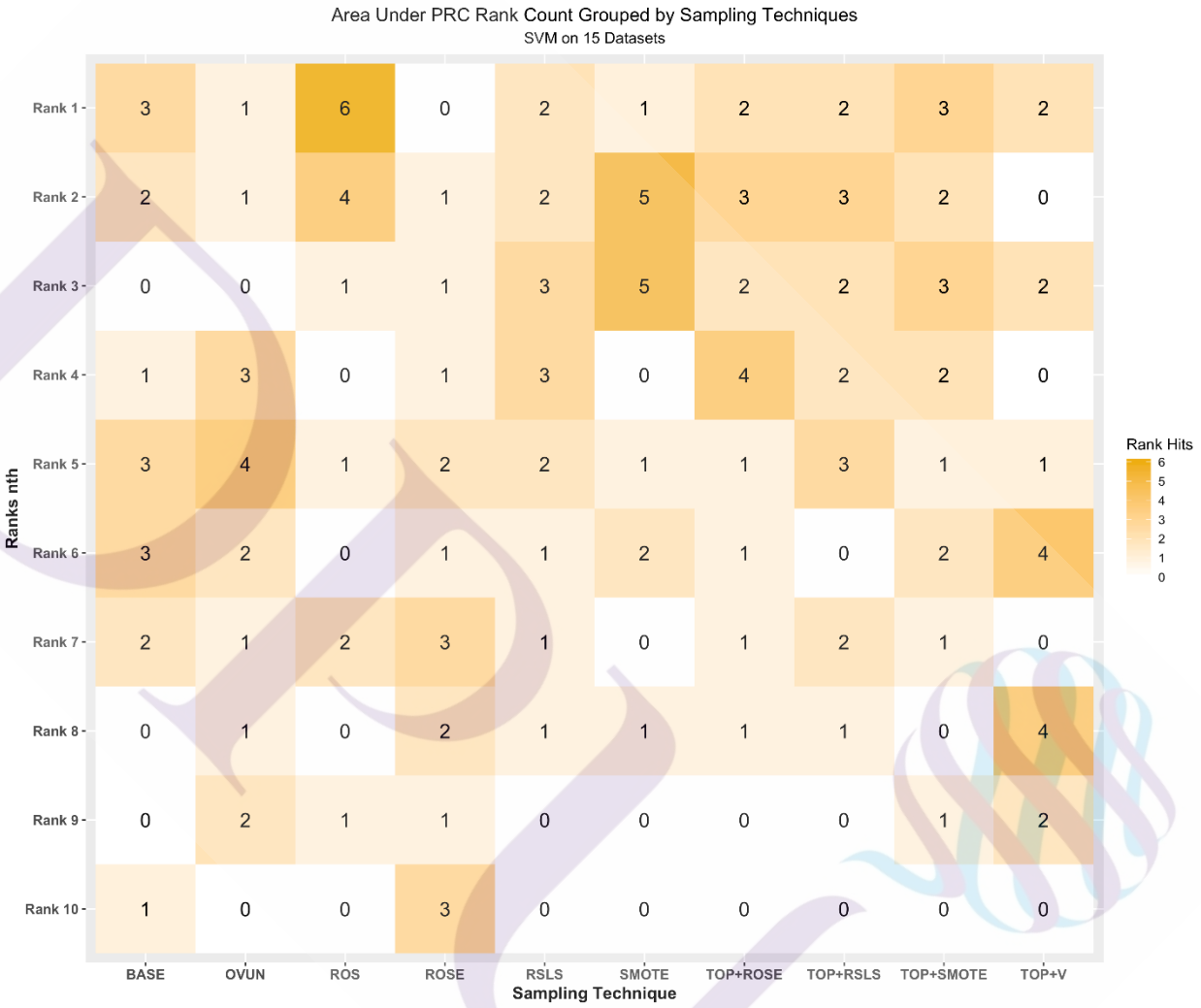
ภาพที่ 7.8.9 โมเดล 5-NN บนข้อมูลทุกชุด



ภาพที่ 7.8.10 โมเดล NN บนข้อมูลทุกชุด



ภาพที่ 7.8.11 โมเดล SVM บนข้อมูลทุกชุด



ภาคผนวก ข

Preliminary Call For Papers

3rd IEEE/ACIS International Conference on Big Data, Cloud Computing, and Data Science Engineering (BCD 2018)

July 10 – 12, 2018, Yonago, Japan

Sponsored by IEEE Computer Society and The International Association for Computer and Information Science (ACIS)

BCD 2018 Proceedings will be published by IEEE Conference Publishing Services and will be submitted to be indexed by EI, INSPEC, and DBLP.

Conference officers will select outstanding papers for publication in the following journals (www.acisinternational.org):

- *International Journal of Networked and Distributed Computing (IJNDC), Paris, France*
- *International Journal of Software Innovation (IJSI), IGI Globe, U.S.A.*
- *Studies in Computational Intelligence (SCI), Springer, Germany*

Conference officers are arranging two special issues of SCI-Indexed Journals to publish outstanding papers.

- *Special Issue: Machine Learning for Intelligent Systems Journal of Robotics, Networking and Artificial Life (JRNAL)*
- *Special Issue: Knowledge-Based Intelligent Systems for Smart Computing International Journal of Computational Intelligence Systems (IJCIS)*

The 3rd International Conference on Big Data, Cloud Computing, and Data Science (BCD 2018) brings together researchers, scientists, engineers, industry practitioners, and students to discuss, encourage and exchange new ideas, research results, and experiences on all aspects of Big Data, Cloud Computing, and Data Science. BCD 2018 aims to facilitate cross-fertilizations among, and is soliciting papers in the key technology enabling areas.

The topics of interests include but not limited to:

Big data Infrastructure	Cloud Solution Design Patterns	Hybrid cloud
Big data management	Data analysis applications - bioinformatics	Mobile processing of big data
Big data privacy and security	Data analysis applications - business intelligence	NoSQL data stores
Big data theory and algorithms	Data analysis applications - social informatics	Outlier detection in big data
Big data visualization	Data cleaning methodologies	Processing of text documents
Cloud Application Scalability and Availability	Data extraction, transformation and load	Public and private clouds

	(ETL)	
Cloud Applications Performance and Monitoring	Data mining and knowledge discovery and big data	QoS for Applications on Clouds
Cloud Computing Architecture	Data modeling	Social network analysis
Cloud Delivery Models	Distributed data structures, MapReduce	Web services
Cloud Optimization and Automation	Education in data science & engineering	Web-based data analysis

Best Paper Sward and Best Student Paper Awards will be conferred at the conference (in order to qualify for the award, the paper must be presented at the conference.)

The format of the manuscript should be in a two-column format and 6 pages in length. Up to an extra 2 pages (total of 8) can be purchased at registration time.

Papers must be submitted electronically through EasyChair (http://www.easychair.org/conferences/?conf=bcd2018)

Conference Organizers

General Chair: Roger Lee, Central Michigan University, USA
Conference Chair: Masateru Tsunoda, Kindai University, Japan
Program Chair: Akinori Ihara, NAIST, Japan
Finance Chair: Shinsuke Matsumoto, Osaka University, Japan
Registration Chair: Tetsuya Kanda, Osaka University, Japan
Publicity Chair/s: Eunjong Choi, NAIST, Japan, Shizue Izumi, Shiga University, Japan
Local Arrangement Chair: Sousuke Amasaki, Okayama Prefectural University, Japan

Important Dates

Workshop/Special session proposal:	**January 20, 2018**
Workshop/Special Session acceptance notification:	**February 16, 2018**
Full Paper Submission:	**April 18, 2018**
Acceptance Notification:	**May 2, 2018**
Camera-Ready Papers & Registration:	**May 28, 2018**
Conference Dates:	**July 10-12, 2018**

TOP: An Efficient TwO-levels of Positive Resampling Framework for Class Imbalanced Data

Nathaniel Netirungroj, Eakasit Pacharawongsakda

Big Data Engineering program, College of Innovative Technology and Engineering, Dhurakij Pundit University
Bangkok, Thailand
{585162020017, eakasit.pac}@dpu.ac.th

***Abstract*—In real-world applications such as fraud detection, target class values have an unequal size. This problem is called class-imbalanced data. Many strategies have been proposed to deal with this situation. Most of them focused on changing the data characteristics. For example, adjusting class distribution is one of the most popular approaches for this matter. In this work, we proposed TOP (TwO-levels of Positive resampling framework), an alternative framework to resolve such a problem. Our technique exploits DBSCAN mechanism and other resampling algorithms in order to maximize classification performance. It is able to dynamically draw two boundaries that represent similarity level between consideration positive and other instances. Many possible resampling techniques such as under-sampling or over-sampling are allowed to perform inside those areas. We benchmarked TOP with three types of resampling techniques including over-sampling, down-sampling, and hybrid-sampling by training eleven machine learning algorithms on fifteen datasets. As a result, our technique outperformed other techniques in several evaluation metrics.**

***Index Terms*—Class Imbalanced Data, Hybrid-sampling, Over-sampling, Under-sampling, Binary Classification**

I. Introduction

An information may be partially captured depending on a nature of its source. Many real-world applications have this in common because of uncontrollable factors, for example, one critical problem in banking area is fraud. Fraudulent are usually occur in a small amount in comparison with non-fraud. In fact, this kind of issue tends to be reduced by the bank regarding business reason. Thus, fraudulent example is naturally rare. Another example is about internet of thing logging system. Malfunction can be occurred unexpectedly while the whole system is operating. As a consequence, some pieces of information may loss from logging process. It is challenging to extract an insight from such a data because the lack of example availability.

One of the common obstacle that typically exists in various topics is called class imbalanced. It is a situation when class distribution appears with an asymmetrical magnitude. This problem consists of two example types, the majority and the minority. A majority is an example class that has more occurrence while a minority has less. Those two are also known as a negative class and a positive class or a undesirable class and a desirable class respectively.

Constructing machine learning model from a different fraction of class label would not significantly increase classification error if the gap between them is inconsiderable [1] [2]. At a certain point, proportion between class label does effect classifier performances. As machine learning model exclusively consume particular class label, it could also be overwhelmed by the amount of corresponding example at the same time. Thus, the probability of misclassification rate could raise up.

An individual data has its own determinant. Some can be noisy, some can be sparse, or some can be dense. Furthermore, some may also have a class overlapping occurred as well. Many resampling techniques such as over-sampling, under-sampling, and hybrid-sampling can be applied to increase classifier accuracies. However, with various data characteristics, only single technique may become exhausted easily. Thus, the idea of applying multiple techniques could be a potential approach in order to tackle class imbalanced issue.

The goal of this research is to propose an alternative technique called TwO-levels of Positive Resampling Framework (TOP). It aims to extend a minority class region while be able to reduce the density of majority class which is located nearby minority area. This framework employs a notion of density-based clustering algorithm (DBSCAN) and exploits the capability of existing resampling techniques.

II. Related works

In recent years, many researchers have proposed various strategies which are intend to overcome class imbalanced situation. Those can be categorized into three main areas [3] as follows. First, resolve at algorithm level by using technique called cost-sensitive [4]. This approach attempts to minimize error by assigning weight to misclassification example based

on confusion matrix results. Second, resolve at feature level by extracting or selecting variables those strengthen class separation [5]. Third, resolve at data level by performing resampling technique [6]. This approach re-calibrate data distribution by creating and/or eliminating example from training data. This work is mainly focus on resampling technique since many studies proved that it is an effective and robust way to solve the problem [7].

In this work, we focused on two resampling methods. One is single-sampling and another is hybrid-sampling. The different between them is single-sampling technique performs resampling data in one way either increase or decrease quantity of example, whereas hybrid-sampling technique attempts to resampling in both way at once.

A. Single-sampling algorithms

1) Random Over Sampling (ROS): It is a resampling algorithm that randomly generate new minority class example until its size become as equal as majority class. Drawback of this technique is that it adds more bias or irrelevant information into training data.

2) Random Under Sampling (RUS): This technique works as the same way as Random Over Sampling does. Instead of creating new minority example, it randomly wipe out the majority class. Thus, precious information may have loss.

3) Synthetic Minority Over-Sampling TEchnique (SMOTE) [8]: A state-of-art resampling technique that artificially create minority example by engaging k-nearest neighbors technique. The algorithm links all minority instance to their neighbors with a straight line, then randomly create synthetic minority example on that path. However, new example may be created extremely near or over the majority ones. In fact, this behavior increases class separation difficulty for some machine learning model.

4) Relocating Safe-level Synthetic Minority Over-Sampling TEchnique (RSLS) [9]: It is a over-sampling algorithm that based on SMOTE. The idea of this technique is to carefully increase minority instance. It draws areas for each positive class, then create artificial instance inside. In the creation process, it attempts to avoid class overlapping problem by shifting synthetic instance further away from the negative class.

B. Hybrid-sampling algorithms

1) Randomly Over Sampling Examples (OVUN) [10]: Unlike ROS and RUS, this technique reduces majority instances and increases minority instances simultaneously. Two good points of this algorithm are that the majority class is not be excessively removed as RUS technique and it also generates less synthetic minority example than ROS technique. More original information is reserved in this way. This technique is still adopt the weakness of ROS and RUS but lesser.

2) Random Over-Under sampling (ROSE) [10]: A hybrid-sampling algorithm to redistribute class imbalanced. It employs smoothed bootstrap to synthetically draw minority instance around the original ones. It also randomly erase majority example concurrently. A concern point of ROSE is that it tends to introduce more noise than SMOTE which potentially causes more class overlapping.

3) Density Based Synthetic Minority Over-Sampling Technique (DBSM) [11]: This technique takes advantage of DBSCAN algorithm to separate data example by density, then reduce the majority that lies close to centroid or minority instance of each cluster. For minority class, it applies SMOTE to artificially create new instance. Drawback of this technique is that it could discard valuable information from cluster that has only majority example.

III. Proposed Framework

In this research, we proposed technique namely, TwO-levels of Positive Resampling Framework (TOP). The motivation of this framework is to exploit an ability of proposed resampling algorithms to redistribute class imbalanced and strengthen minority instance region. It limits resampling area by drawing two boundaries; inner area and outer area, around the consideration minority instance. Synthetic instance will be created in an inner area because we supposed that similar instance should lies in its region. And the majority instance that exists in outer area will be randomly removed.

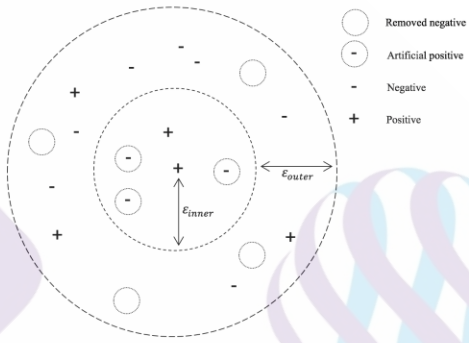


Fig. 1: TwO-levels of Positive Resampling Framework illustration

The process of this framework is divided into two parts which displayed in Algorithm 1. First part is to over-sampling minority class (line 2 to 9). It calculates euclidean distance from considering minority to all majority that lies inside inner level (line 3). If given percentage of positive membership in this area is less than positive population in dataset, then perform over-sampling (line 4 to 8). Second part focuses on outer level (line 10 to 13). Majority instances which located between the edge of inner area and outer area are randomly removed (line 11 to 12). This technique considers all positive instances in input dataset. Instances in two-levels area both inner and outer were obtained by a distance function. This function takes three parameters including consideration positive instance, minimum distance, and maximum distance. For inner level, minimum distance is zero since it starts from consideration minority. Maximum distance for this task is an

epsilon of inner level. For outer level, beginning point from this area is an epsilon value of inner level. And the maximum distance is an epsilon value of outer level. We have also created build-in over-sampling technique which is called vanilla. It checks if there is any class overlapping (zero distance between positive and negative instances) occurred in inner level or not. If yes, negative instance which located exactly the same position as positive will be converted into positive. Likewise, for non-overlapping instance. This technique can be replace at line 6. Additionally, other over-sampling algorithms such as SMOTE and RSLS can be applied instead of vanilla. Class overlapping conversion method in vanilla can be disabled or used with other resampling algorithms. For outer level, other under-sampling algorithms can be applied instead of randomly remove majority.

Algorithm 1 Two-levels Resampling

Input: S is an imbalanced dataset
P is an amount of positive in dataset
p is positive membership in two-levels area
n is negative membership in two-levels area
pos is consideration positive instance
pos_{in} percentage of positive in inner level
neg_{rm} percentage of negative to be removed
ε_{in} is an inner level epsilon
ε_{out} is an outer level epsilon
Output: Modified dataset

1: **for all** $pos \in S$ **do**
2: $member_{inner} - find_{member}(pos, 0, \varepsilon_{in})$
3: **for all** $p \subset member_{inner}$ **do**
4: **if** $\frac{count(p)}{P} \times 100 < pos_{in}$ **then**
5: **for all** $n \in member_{inner}$ **do**
6: $convert\ n\ to\ p$
7: **end for**
8: **end if**
9: **end for**
10: $member_{outer} = find_{member}(pos, \varepsilon_{in}, \varepsilon_{out})$
11: **for all** $n \in member_{outer}$ **do**
12: $randomly\ remove\ n\ at\ neg_{rm}$
13: **end for**
14: **end for**
15: **return**

IV. Experiment settings

In this work we performed 10-folds cross validation on given data sets to obtain average model performances. We implemented several pre-process techniques and machine learning classifiers sequentially in each split to avoid over-fitting situation.

A. Data descriptions

The class imbalanced data sets in this experiment are gathered from two public sources. Fifteen data sets including ecoli2, glass0, glass1, glass6, haberman, new-thyroid1, new-thyroid2, page-blocks-1-3_vs_4, pima, vehicle1, vehicle2, wisconsin, yeast1, yeast3, and Liver-disorders were from KEEL and UCI repositories. We identified class imbalanced by IR (1) and lack of minority information by LI (2) [1] . The details of those data sets are described in Table I.

$$IR = \frac{Number\ of\ Negative}{Number\ of\ Positive} \tag{1}$$

$$LI = Number\ of\ Negative - Number\ of\ Positive \tag{2}$$

TABLE I: Data characteristics

Dataset	IR	LI	Positive	Negative
ecoli2	5.46	232	52	284
glass0	2.05	74	70	144
glass1	1.81	62	76	138
glass6	6.37	156	29	185
haberman	2.77	144	81	225
Liver-disorders	1.37	55	200	145
new-thyroid1	5.14	145	35	180
new-thyroid2	5.14	145	35	180
page blocks 1 3_vs_1	15.8	416	444	28
pima	1.86	232	500	268
vehicle1	2.89	412	217	629
vehicle2	2.88	410	218	628
wisconsin	1.85	205	444	239
yeast1	2.45	626	1055	429
yeast3	8.10	1158	1321	163

B. Evaluation metrics

The experiments have been assessed by various metrics that appropriate for this topic. Instead of overall accuracy, we have applied indicators which are capable to reflect classifier performance in many aspects [12]. All those metrics are based on Confusion Matrix theorem. Their equation are described as follows:

$$Recall - \frac{TruePositive}{TruePositive + FalseNegative} \tag{3}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4}$$

$$F\text{-}score - 2 \times \frac{Precision \times Recall}{Precision \cdot Recall} \tag{5}$$

$$Gscore = \sqrt{Precision \times Recall} \tag{6}$$

Moreover, we employed Area Under the Receiver Operating Characteristic Curve (AUROC) [13] which is commonly adopted by many researchers to compare classifiers performance. However, certain studies shown that this indicator may compromised in class skewed situation [14] [15]. Hence, we have selected Area Under the Precision Recall Curve (AUPRC) [16] [17] as another measurement to reveal more information [18]. Both metrics are respectively illustrated below.

$$AU_{ROC} = \int_0^1 \frac{TP}{P}\, d\frac{FP}{N} \tag{7}$$

$$AU_{PRC} = \int_0^{\cdot} p(r)\, dr, \tag{8}$$

Table II – Experimental results

	Metrics	C4.5	XGB	RF	C4.5	XGB	RF	C4.5	XGB	RF	C4.5	XGB	RF	C4.5	XGB	RF	C4.5	XGB	RF
		ecoli2			glass0			glass1			glass6			haberman			Liver-disorders		
BASELINE	F1 / GM	706 / 797	735 / 743	815 / 806	704 / 714	670 / 673	737 / 794	608 / 629	550 / 611	731 / 742	765 / 731	853 / 871	806 / 895	170 / 172	341 / 360	237 / 239	700 / 703	739 / 741	763 / 766
	AUROC / AUPRC	849 / 730	930 / 825	967 / 911	772 / 629	871 / 754	937 / 901	735 / 646	828 / 746	915 / 866	907 / 805	961 / 895	979 / 962	516 / 290	670 / 470	652 / 404	617 / 652	744 / 771	765 / 777
ROS	F1 / GM	788 / 794	773 / 783	824 / 851	721 / 732	**.763** / 774	790 / 789	621 / 628	650 / 659	771 / 782	763 / 781	869 / 879	836 / 895	458 / 464	**.512** / 516	350 / 358	662 / 635	**.785** / 786	774 / 778
	AUROC / AUPRC	882 / 627	954 / 803	960 / 913	790 / 655	903 / 824	929 / 870	673 / 534	848 / 786	914 / 866	852 / 624	967 / 933	970 / 922	627 / 407	703 / 474	642 / 392	631 / 673	783 / 814	754 / 792
SMOTE	F1 / GM	763 / 773	755 / 766	836 / 840	675 / 684	691 / 708	807 / 812	571 / 578	601 / 617	760 / 776	789 / 805	850 / 853	891 / 900	498 / 500	495 / 493	355 / 357	**.746** / 750	774 / 780	759 / 762
	AUROC / AUPRC	858 / 656	945 / 881	969 / 927	724 / 550	896 / 800	931 / 879	692 / 569	827 / 745	904 / 858	882 / 721	964 / 930	970 / 942	660 / 429	702 / 505	665 / 420	642 / 679	753 / 785	755 / 778
BSLS	F1 / GM	758 / 767	755 / 766	821 / 830	708 / 717	711 / 720	814 / 819	673 / 676	654 / 661	778 / 787	824 / 835	883 / 891	891 / 900	466 / 472	460 / 463	337 / 337	736 / 740	772 / 777	**.784** / 785
	AUROC / AUPRC	862 / 679	941 / 840	960 / 910	764 / 579	897 / 801	934 / 881	777 / 673	839 / 761	907 / 843	939 / 822	964 / 896	932 / 970	625 / 393	683 / 491	648 / 376	642 / 636	770 / 801	759 / 770
TOP+V	F1 / GM	811 / 821	794 / 805	**.875** / 880	**.759** / 773	747 / 757	**.831** / 836	**.748** / 751	**.693** / 707	**.805** / 810	830 / 838	**.911** / 918	**.926** / 932	484 / 512	497 / 536	455 / 503	645 / 651	693 / 701	686 / 691
	AUROC / AUPRC	869 / 752	912 / 819	954 / 892	826 / 630	867 / 750	938 / 883	825 / 745	802 / 648	921 / 875	934 / 775	964 / 939	974 / 946	609 / 350	662 / 435	604 / 324	676 / 576	772 / 716	766 / 742
TOP+ROS	F1 / GM	**.832** / 843	785 / 796	850 / 858	743 / 754	741 / 752	819 / 822	678 / 682	662 / 668	768 / 775	**.870** / 876	**.911** / 913	906 / 913	**.509** / 518	485 / 513	474 / 510	606 / 644	679 / 689	678 / 687
	AUROC / AUPRC	864 / 764	925 / 855	973 / 902	819 / 603	913 / 849	939 / 874	746 / 611	856 / 783	917 / 870	937 / 805	949 / 893	952 / 917	635 / 361	696 / 513	636 / 366	577 / 472	765 / 727	771 / 740
TOP + SMOTE	F1 / GM	811 / 821	787 / 795	857 / 862	724 / 736	730 / 741	799 / 806	698 / 711	650 / 657	771 / 774	839 / 849	891 / 900	**.926** / 932	450 / 499	465 / 522	465 / 499	625 / 644	660 / 664	656 / 676
	AUROC / AUPRC	869 / 752	951 / 806	970 / 911	815 / 611	901 / 822	940 / 885	784 / 650	834 / 767	908 / 855	902 / 766	961 / 902	939 / 939	640 / 369	685 / 450	639 / 376	657 / 536	789 / 760	739 / 700
TOP + BSLS	F1 / GM	811 / 821	**.816** / 823	868 / 873	724 / 736	741 / 754	806 / 812	662 / 674	646 / 682	771 / 773	829 / 841	901 / 909	**.926** / 932	493 / 517	474 / 531	471 / 520	609 / 636	660 / 664	678 / 687
	AUROC / AUPRC	869 / 752	949 / 899	978 / 926	815 / 611	902 / 823	937 / 875	723 / 556	839 / 777	891 / 834	900 / 726	971 / 951	969 / 937	652 / 402	680 / 470	638 / 397	593 / 432	789 / 760	771 / 740
OVUN	F1 / GM	812 / 819	723 / 739	852 / 858	716 / 725	689 / 699	775 / 781	642 / 657	629 / 641	745 / 751	797 / 805	844 / 857	872 / 884	497 / 519	507 / 535	**.483** / 497	581 / 591	641 / 664	621 / 631
	AUROC / AUPRC	876 / 719	925 / 822	965 / 900	700 / 609	871 / 758	934 / 895	748 / 646	843 / 753	919 / 861	900 / 737	965 / 905	993 / 959	668 / 413	702 / 471	677 / 411	594 / 490	704 / 651	733 / 671
ROSE	F1 / GM	275 / 399	403 / 515	311 / 405	644 / 632	631 / 676	636 / 673	578 / 604	608 / 648	670 / 682	323 / 644	872 / 882	848 / 858	475 / 479	478 / 479	430 / 440	564 / 536	567 / 573	543 / 545
	AUROC / AUPRC	518 / 160	937 / 843	945 / 354	729 / 434	818 / 585	739 / 596	643 / 478	734 / 610	744 / 607	927 / 761	964 / 937	972 / 954	640 / 427	648 / 439	640 / 406	541 / 457	673 / 613	617 / 569
DESM*	F1 / GM	767 / -----	----- / -----	----- / -----	711 / -----	----- / -----	----- / -----	710 / -----	----- / -----	----- / -----	829 / -----	----- / -----	----- / -----	467 / -----	----- / -----	----- / -----	----- / -----	----- / -----	----- / -----
	AUROC / AUPRC	894 / -----	----- / -----	----- / -----	738 / -----	----- / -----	----- / -----	777 / -----	----- / -----	----- / -----	903 / -----	----- / -----	----- / -----	622 / -----	----- / -----	----- / -----	----- / -----	----- / -----	----- / -----
		new-thyroid1			new-thyroid2			page-blocks-1-3_vs_4			pima			vehicle1			vehicle2		
BASELINE	F1 / GM	870 / 880	965 / 965	946 / 950	848 / 854	844 / 855	916 / 925	769 / 794	602 / 619	860 / 870	579 / 601	510 / 555	589 / 600	405 / 441	440 / 471	501 / 507	922 / 925	956 / 957	977 / 977
	AUROC / AUPRC	944 / 827	1.00 / 1.00	1.00 / 1.00	929 / 847	993 / 966	992 / 966	955 / 659	991 / 964	996 / 949	827 / 434	687 / 570	690 / 515	647 / 451	822 / 627	850 / 645	946 / 878	991 / 960	998 / 996
ROS	F1 / GM	909 / 915	943 / 947	923 / 929	852 / 856	866 / 872	916 / 925	**.967** / 971	**.967** / 971	950 / 958	631 / 637	651 / 655	645 / 648	492 / 494	619 / 632	595 / 596	940 / 941	888 / 891	**.981** / 980
	AUROC / AUPRC	936 / 850	960 / 964	1.00 / 1.00	914 / 814	990 / 962	996 / 977	992 / 950	1.00 / 1.00	1.00 / 1.00	712 / 149	813 / 694	830 / 635	655 / 379	847 / 625	861 / 644	964 / 872	928 / 965	999 / 997
SMOTE	F1 / GM	907 / 939	929 / 934	956 / 978	865 / 877	905 / 907	921 / 924	**.967** / 971	**.967** / 971	917 / 928	599 / 607	643 / 657	641 / 643	561 / 565	566 / 589	561 / 564	922 / 925	894 / 895	966 / 968
	AUROC / AUPRC	957 / 906	959 / 964	1.00 / 1.00	920 / 836	984 / 925	997 / 983	998 / 950	1.00 / 1.00	1.00 / 1.00	715 / 567	832 / 734	819 / 691	708 / 454	847 / 631	861 / 654	940 / 872	994 / 993	999 / 997
BSLS	F1 / GM	831 / 851	956 / 965	951 / 955	840 / 845	902 / 904	950 / 959	**.967** / 971	923 / 925	932 / 941	603 / 608	611 / 616	652 / 656	**.605** / 610	**.589** / 591	547 / 549	**.941** / 942	892 / 894	966 / 966
	AUROC / AUPRC	872 / 730	1.00 / 1.00	1.00 / 1.00	875 / 777	980 / 931	996 / 977	998 / 950	980 / 901	1.00 / 1.00	725 / 595	826 / 703	825 / 721	762 / 527	844 / 634	860 / 650	962 / 902	984 / 952	999 / 990
TOP+V	F1 / GM	932 / 937	**.966** / 968	**.986** / 987	902 / 906	**.920** / 952	**.989** / 989	**.967** / 971	847 / 845	**1.00** / 1.00	**.672** / 677	671 / 672	**.670** / 674	599 / 605	560 / 587	634 / 646	945 / 955	**.940** / 941	**.981** / 980
	AUROC / AUPRC	925 / 900	1.00 / 1.00	1.00 / 1.00	964 / 840	991 / 943	1.00 / 1.00	992 / 950	993 / 914	1.00 / 1.00	753 / 555	833 / 714	821 / 694	760 / 502	808 / 589	844 / 556	970 / 872	951 / 961	999 / 996
TOP+ROS	F1 / GM	916 / 925	945 / 950	969 / 971	871 / 886	907 / 907	939 / 947	**.967** / 971	**.967** / 971	**1.00** / 1.00	656 / 652	**.686** / 695	651 / 659	560 / 589	**.589** / 605	**.642** / 650	951 / 957	956 / 957	977 / 977
	AUROC / AUPRC	962 / 878	959 / 964	1.00 / 1.00	927 / 833	991 / 956	996 / 977	998 / 950	1.00 / 1.00	1.00 / 1.00	747 / 604	837 / 703	830 / 653	717 / 404	815 / 565	845 / 592	962 / 900	991 / 963	999 / 997
TOP + SMOTE	F1 / GM	**.963** / 962	927 / 934	955 / 958	886 / 900	913 / 915	939 / 947	**.967** / 971	**.967** / 971	967 / 971	655 / 652	654 / 654	**.670** / 675	567 / 617	553 / 555	611 / 630	959 / 959	957 / 958	979 / 979
	AUROC / AUPRC	902 / 944	1.00 / 1.00	990 / 993	930 / 850	991 / 956	998 / 990	998 / 950	1.00 / 1.00	1.00 / 1.00	741 / 535	823 / 703	815 / 694	749 / 437	804 / 560	842 / 604	967 / 922	991 / 960	999 / 997
TOP + BSLS	F1 / GM	940 / 944	927 / 934	966 / 968	886 / 900	913 / 915	939 / 947	**.967** / 971	**.967** / 971	980 / 980	655 / 658	**.686** / 695	656 / 659	576 / 614	540 / 569	630 / 628	952 / 954	957 / 958	977 / 977
	AUROC / AUPRC	956 / 924	1.00 / 1.00	1.00 / 1.00	930 / 858	991 / 956	996 / 977	998 / 950	1.00 / 1.00	999 / 990	741 / 535	833 / 713	810 / 672	752 / 428	799 / 550	847 / 592	953 / 932	991 / 960	999 / 997
OVUN	F1 / GM	862 / 872	971 / 985	966 / 968	836 / 853	867 / 870	919 / 929	902 / 915	540 / 569	919 / 928	620 / 621	660 / 661	641 / 644	452 / 478	501 / 519	551 / 560	925 / 921	868 / 873	964 / 964
	AUROC / AUPRC	914 / 771	1.00 / 1.00	1.00 / 1.00	903 / 782	988 / 931	990 / 961	990 / 875	992 / 916	999 / 990	732 / 571	813 / 706	810 / 698	654 / 433	826 / 612	863 / 656	967 / 895	979 / 993	998 / 994
ROSE	F1 / GM	870 / 877	923 / 929	923 / 929	838 / 845	897 / 899	866 / 873	318 / 422	401 / 490	438 / 516	610 / 615	642 / 644	640 / 645	492 / 534	**.589** / 605	579 / 592	746 / 760	847 / 851	847 / 853
	AUROC / AUPRC	955 / 756	1.00 / 1.00	999 / 994	937 / 745	989 / 933	990 / 955	347 / 192	943 / 670	977 / 731	744 / 530	812 / 697	807 / 635	721 / 420	799 / 575	798 / 581	891 / 700	978 / 945	980 / 951
DESM*	F1 / GM	914 /	/	/	**.951** /	/	/	/	/	/	/	/	/	581 /	/	/	919 /	/	/
	AUROC / AUPRC	933 / -----	----- / -----	----- / -----	974 / -----	----- / -----	----- / -----	----- / -----	----- / -----	----- / -----	----- / -----	----- / -----	----- / -----	730 / -----	----- / -----	----- / -----	949 / -----	----- / -----	----- / -----
		wisconsin			yeast1			yeast3											
BASELINE	F1 / GM	945 / 946	952 / 953	963 / 963	504 / 510	533 / 547	557 / 559	768 / 772	765 / 770	746 / 750									
	AUROC / AUPRC	959 / 949	981 / 979	994 / 987	731 / 497	753 / 515	795 / 522	934 / 704	968 / 848	961 / 826									
ROS	F1 / GM	947 / 944	946 / 947	953 / 953	528 / 530	597 / 605	557 / 561	700 / 717	773 / 784	784 / 787									
	AUROC / AUPRC	952 / 937	989 / 974	991 / 961	645 / 433	794 / 515	798 / 523	854 / 539	974 / 828	968 / 819									
SMOTE	F1 / GM	941 / 941	948 / 948	959 / 959	**.595** / 598	608 / 610	**.615** / 618	768 / 772	751 / 769	778 / 783									
	AUROC / AUPRC	962 / 951	981 / 977	990 / 950	731 / 500	799 / 505	807 / 575	913 / 611	968 / 825	969 / 811									
BSLS	F1 / GM	931 / 932	946 / 947	961 / 961	571 / 576	**.612** / 614	589 / 591	765 / 770	784 / 787	755 / 759									
	AUROC / AUPRC	954 / 937	980 / 971	990 / 980	716 / 451	798 / 503	795 / 500	895 / 660	983 / 842	967 / 829									
TOP+V	F1 / GM	**.951** / 951	**.960** / 960	**.965** / 965	585 / 598	601 / 605	592 / 594	798 / 801	808 / 815	783 / 788									
	AUROC / AUPRC	967 / 917	983 / 967	993 / 980	741 / 478	780 / 510	781 / 577	913 / 663	970 / 870	973 / 833									
TOP+ROS	F1 / GM	941 / 940	949 / 949	**.965** / 965	575 / 575	606 / 609	590 / 596	**.786** / 790	786 / 789	780 / 788									
	AUROC / AUPRC	959 / 949	981 / 974	994 / 968	786 / 500	799 / 519	782 / 517	952 / 681	969 / 871	971 / 850									
TOP + SMOTE	F1 / GM	945 / 946	953 / 953	963 / 963	579 / 580	593 / 594	592 / 598	**.786** / 790	799 / 801	785 / 789									
	AUROC / AUPRC	959 / 949	981 / 978	993 / 985	718 / 447	794 / 505	794 / 503	932 / 681	968 / 819	974 / 813									
TOP + BSLS	F1 / GM	945 / 946	953 / 953	**.965** / 965	579 / 580	593 / 594	595 / 601	**.786** / 790	**.800** / 805	**.790** / 796									
	AUROC / AUPRC	959 / 949	981 / 978	994 / 968	718 / 427	794 / 503	797 / 502	932 / 681	968 / 821	963 / 773									
OVUN	F1 / GM	950 / 955	913 / 914	913 / 914	502 / 504	475 / 504	502 / 577	591 / 705	788 / 710	785 / 745									
	AUROC / AUPRC	960 / 931	989 / 970	990 / 960	614 / 367	646 / 570	655 / 555	908 / 466	968 / 786	965 / 769									
ROSE	F1 / GM	930 / 930	957 / 958	951 / 952	453 / 510	452 / 540	457 / 540	225 / 351	497 / 571	212 / 313									
	AUROC / AUPRC	952 / 911	981 / 977	989 / 989	507 / 290	507 / 292	707 / 452	570 / 154	914 / 868	964 / 751									
DESM*	F1 / GM	931 /	/	/	574 /	/	/	705 /	/	/									
	AUROC / AUPRC	952 /	/	/	693 /	/	/	875 /	/	/									

*DESM does not implemented in our experiment.

Table III - Average F1 Improvement from Baseline

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	4.36%	3.78%	**8.28%**	3.64%	1.39%	1.97%	-0.07%	0.29%	4.16%	2.69%	6.35%
SMOTE	5.62%	4.66%	6.54%	2.60%	1.95%	2.81%	2.02%	2.51%	4.41%	2.86%	5.96%
RSLS	5.85%	5.19%	3.98%	4.79%	1.40%	2.70%	1.97%	2.08%	2.11%	2.60%	5.01%
TOP+V	**8.52%**	**6.83%**	7.87%	**7.15%**	3.47%	4.70%	4.21%	4.70%	**6.04%**	**5.79%**	6.48%
TOP+ROS	7.35%	6.26%	8.13%	6.49%	3.03%	**4.84%**	**4.32%**	4.19%	5.28%	4.80%	7.32%
TOP+SMOTE	7.33%	5.49%	7.16%	5.43%	**3.93%**	4.49%	3.46%	4.14%	4.91%	4.25%	7.02%
TOP+RSLS	7.02%	5.46%	7.46%	5.63%	3.57%	4.74%	4.05%	**4.73%**	5.17%	4.82%	6.77%
OVUN	2.85%	2.89%	0.41%	1.34%	0.26%	-0.10%	-0.66%	-0.95%	1.35%	1.35%	2.71%
ROSE	-9.94%	-13.33%	-4.16%	-0.34%	-11.01%	-7.40%	-7.14%	-0.62%	-5.85%	-13.03%	-0.88%

Table IV - Average GM Improvement from Baseline

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	3.85%	3.50%	7.59%	3.41%	1.14%	2.04%	0.37%	0.55%	3.92%	2.42%	5.68%
SMOTE	5.15%	4.39%	5.79%	2.31%	1.64%	2.56%	2.01%	2.34%	3.97%	2.57%	5.17%
RSLS	5.33%	4.92%	3.30%	4.11%	1.13%	2.28%	1.57%	1.63%	1.50%	2.28%	4.24%
TOP+V	**8.14%**	**6.81%**	7.35%	**6.54%**	3.86%	5.20%	**4.57%**	**4.80%**	**5.75%**	**5.70%**	5.91%
TOP+ROS	7.22%	6.03%	7.59%	6.19%	3.86%	5.13%	4.32%	3.82%	5.04%	4.74%	6.57%
TOP+SMOTE	7.44%	5.55%	6.68%	5.69%	**4.54%**	4.50%	4.14%	4.14%	4.58%	4.31%	**6.61%**
TOP+RSLS	7.02%	5.56%	7.22%	5.91%	4.06%	**5.23%**	4.45%	4.60%	4.97%	4.78%	6.43%
OVUN	3.31%	3.41%	0.89%	1.88%	0.71%	0.55%	0.11%	-0.81%	1.70%	1.72%	2.67%
ROSE	-6.80%	-10.02%	-2.49%	-0.61%	-9.67%	-6.37%	-6.19%	0.08%	-4.86%	-10.26%	-0.80%

Table V - Average AUROC Improvement from Baseline

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	0.09%	1.93%	**2.24%**	-0.12%	0.50%	0.24%	-0.27%	**0.70%**	0.73%	0.72%	1.35%
SMOTE	1.61%	1.61%	1.68%	-0.64%	1.48%	1.14%	**1.25%**	0.43%	0.64%	**0.91%**	1.21%
RSLS	2.01%	**2.60%**	1.62%	0.03%	0.70%	0.78%	0.46%	0.31%	0.04%	0.87%	1.26%
TOP+V	**4.78%**	1.87%	0.71%	**0.54%**	2.33%	0.80%	0.19%	0.12%	0.51%	0.48%	-0.09%
TOP+ROS	2.92%	2.40%	1.77%	0.12%	2.02%	**1.54%**	0.77%	0.31%	0.76%	0.56%	**1.64%**
TOP+SMOTE	3.79%	**2.32%**	1.72%	-0.27%	**2.67%**	1.42%	0.11%	0.51%	**1.31%**	0.65%	1.16%
TOP+RSLS	2.79%	2.11%	1.75%	-0.25%	2.47%	1.06%	0.58%	0.45%	0.73%	0.59%	0.98%
OVUN	0.85%	1.01%	-0.02%	-0.59%	0.51%	-0.21%	-0.71%	-2.99%	-0.13%	0.08%	-0.43%
ROSE	-6.32%	-8.25%	-3.22%	-1.38%	-1.98%	-4.91%	-5.32%	0.64%	-3.33%	-3.82%	-1.23%

Table VI - Average AUPRC Improvement from Baseline

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM
ROS	-0.40%	2.38%	**3.10%**	-1.14%	**2.07%**	-1.75%	-2.96%	1.75%	0.87%	1.07%	**2.06%**
SMOTE	1.19%	1.56%	2.55%	-1.09%	2.01%	-0.05%	-0.38%	**2.06%**	0.85%	**1.51%**	1.89%
RSLS	2.91%	**3.67%**	1.70%	**0.81%**	1.69%	**1.01%**	-0.22%	1.25%	0.18%	1.37%	1.51%
TOP+V	**3.81%**	1.88%	-0.37%	0.71%	1.32%	-0.10%	-1.87%	-0.78%	-0.28%	0.04%	-0.99%
TOP+ROS	2.31%	2.50%	2.69%	-0.09%	0.45%	0.45%	-0.66%	-0.89%	-0.05%	0.06%	1.59%
TOP+SMOTE	3.32%	2.52%	2.00%	-1.30%	1.13%	0.10%	-1.96%	-0.02%	**1.54%**	0.28%	1.17%
TOP+RSLS	1.86%	2.15%	2.48%	-1.31%	0.91%	-1.12%	-1.87%	0.20%	0.08%	-0.18%	0.95%
OVUN	-2.29%	0.14%	-2.24%	-0.61%	-1.39%	-3.47%	-4.15%	-3.35%	-1.26%	-1.34%	-1.48%
ROSE	-16.47%	-16.57%	-7.10%	-4.11%	-16.63%	-13.43%	-12.64%	1.12%	-7.80%	-8.87%	-2.98%

Table VII - Average Ranking

	C4.5	C5.0	XGB	LR	1-NN	3-NN	5-NN	NB	NN	RF	SVM	Avg. by Sampling
BASE	7.00	4.20	6.40	6.60	6.13	**6.53**	**6.27**	7.67	5.47	2.87	6.87	**6.00**
ROS	7.13	4.47	**3.80**	6.13	6.93	6.60	7.93	9.20	5.07	3.67	4.40	**5.94**
SMOTE	6.33	**3.27**	4.93	6.47	6.67	7.27	6.87	9.33	4.73	2.87	5.40	**5.83**
RSLS	6.00	3.53	5.33	5.73	6.53	7.33	7.33	9.07	5.60	**3.60**	5.27	**5.94**
TOP+V	6.80	3.60	6.20	6.13	6.27	7.07	7.13	8.67	5.33	2.80	5.40	**5.95**
TOP+ROS	7.20	6.47	3.93	**3.67**	8.60	7.20	7.33	6.73	**5.07**	4.87	**4.00**	**5.92**
TOP+SMOTE	**5.53**	3.73	4.40	5.27	7.73	7.60	7.73	9.60	**4.60**	2.87	6.07	**5.92**
TOP+RSLS	6.60	3.80	4.33	5.87	7.33	6.80	7.60	9.60	4.73	3.20	5.00	**5.90**
OVUN	**6.00**	4.07	4.73	6.73	**5.93**	6.87	7.33	8.93	4.93	2.60	5.20	**5.76**
ROSE	6.40	4.40	4.80	6.67	6.33	6.73	7.13	**9.13**	5.33	**2.33**	5.60	**5.90**
Avg. by Model	**6.50**	**4.15**	**4.89**	**5.93**	**6.85**	**7.00**	**7.27**	**8.79**	**5.09**	**3.17**	**5.32**	**5.90**

V. Results and discussions

Regarding our settings designed in section IV, we applied data pre-process techniques including ROS, SMOTE, RSLS, TOP+V, TOP+SMOTE, TOP+RSLS, TOP+ROS, OVUN, and ROSE to adjust class distribution. We then applied processed data with eleven machine learning algorithms including C4.5, C5.0, eXtreme Gradient Boosting (XGB), Random Forest (RF), Support Vector Machine with Radial Basis Kernel function (SVM), Naive Bayes (NB), Logistics Regression (LR), Neural Network (NN) and k-Nearest Neighbors with $k-1, 3, 5$ (kNN). A number of k for resampling algorithms was set to five. Replacement of vanilla algorithm were used with class overlapping process disabled. Only three tree-based classifier; C4.5, RF, and XGB are represent in Table II. Result from other models can be found at an external source[1].

According to table II, TOP has increased F1 score, especially with vanilla. It outperformed other techniques on most datasets. For DBSM resampling technique, we were not implemented it to our experiment. The results were taken from original work [11].

The average of improvement of F1, GM, AUROC, and AUPRC are illustrated in Table III, IV, V, and VI correspondingly. Our technique achieved F1 score up to 8.52% of an average improvement rate across fifteen datasets. It also improved F1 score more than other resampling techniques in most cases. Except for certain AUPRC score in Table VI, only two cases have been improved, 3.81% with C4.5 and 1.54% with Neural Network. However, 3.81% is the maximum value in this particular table. The experimental output in Table VII shows that Random Forest is the most accurate classifier as it incorporates with Random over-sampling technique. It accomplished highest average rank in most combination (less is best). On the other hand, Naive Bayes has lowest average rank with all resampling techniques. TOP helps certain classifiers including C4.5, Logistics Regression, Neural Network, and Support Vector Machine achieved its best average ranking.

However, TOP may compromised when dealing with small or sparse imbalanced dataset since two-levels area is based on minority class. Moreover, a harsh hyper-parameter combination could result in unsatisfactory performances. This framework requires an intensive hyper-parameters tuning in order to achieve acceptable output. It can be very time consuming to perform this task. Thus, optimization technique such as genetic algorithm could be a potential solution to shorten tuning time.

VI. Conclusion

We propose an alternative strategies to handle class imbalanced data. It focuses on improving classifiers performance using resampling methods. Eleven machine learning classifiers were constructed on one original data and ten pre-processed data from ten resampling approaches including ROS, SMOTE, RSLS, TOP+V, TOP+ROS, TOP+SMOTE, TOP+RSLS, OVUN and ROSE. Experiment results show that TwO-levels of Positive Resampling Framework is able to accomplish highest classification performance. The classifier that works best with our technique is Random Forest.

[1] https://github.com/nathanielre/top_expresult

References

[1] S. Visa and A. Ralescu, "Issues in mining imbalanced data sets-a review paper," in *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*, no. August 2016, 2005, pp. 67–73.

[2] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," pp. 13–21, 2012.

[3] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, vol. 4, 2008, pp. 192–201.

[4] A. C. Schierz, "Virtual screening of bioassay data," *Journal of Cheminformatics*, vol. 1, no. 1, pp. 231–235, 2009.

[5] S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Information Sciences*, vol. 286, pp. 228–246, 2014.

[6] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32–41, 2014.

[7] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Folleco, "An Empirical Study of the Classification Performance of Learners on Imbalanced and Noisy Software Quality Data," in *2007 IEEE International Conference on Information Reuse and Integration*, 2007, pp. 651–658.

[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] W. Siriseriwan and K. Sinapiromsaran, "The effective redistribution for imbalance dataset: Relocating safe-level SMOTE with minority outcast handling," *Chiang Mai Journal of Science*, vol. 43, no. 1, pp. 234–246, 2016.

[10] N. Lunardon, G. Menardi, and N. Torelli, "ROSE : A Package for Binary Imbalanced Learning," *The R Journal*, vol. 6, no. June, pp. 79–89, 2014.

[11] Y. Sanguanmak and A. Hanskunatai, "DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification," in *2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016*, 2016.

[12] G. Hoang, A. Bouzerdoum, and S. Lam, "Learning Pattern Classification Tasks with Imbalanced Data Sets," *Pattern Recognition*, pp. 193–208, 2009.

[13] M. Vuk, "ROC Curve , Lift Chart and Calibration Plot," *Metodoloski zvezki*, vol. 3, no. 1, pp. 89–108, 2006.

[14] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 233–240.

[15] J. M. Lobo, A. Jiménez-valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," pp. 145–151, 2008.

[16] J. Keilwagen, I. Grosse, and J. Grau, "Area under precision-recall curves for weighted and unweighted data," *PLoS ONE*, vol. 9, no. 3, 2014.

[17] K. H. Brodersen, C. S. Ong, K. E. Stephany, and J. M. Buhmann, "The binormal assumption on precision recall curves," in *Proceedings International Conference on Pattern Recognition*, 2010, pp. 4263–4266.

[18] T. Saito and M. Rehmsmeier, "The precision recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, 2015.

2017 KICSS Nagoya

Once-in-a-lifetime Conference for Creativity

On behalf of the Organizing Committee, it gives us great pleasure to invite you to the The 12th International Conference on Knowledge, Information and Creativity Support Systems(KICSS2017), which will be held at the Nagoya, Japan, on November 9-11, 2017.

Important Dates:

- Paper Submission(Full/short papers): June 15, 2017
- Paper Submission(Poster papers): June 30, 2017
- Notification: July 30, 2017
- Camera Ready Submission and Author Registration: September 1, 2017
- Conference: November 9-11, 2017

Scope & Topics: Following the tradition of previous conferences on Knowledge, Information and Creativity Support Systems, KICSS 2017 will cover all aspects of knowledge management, knowledge engineering, intelligent information systems, and creativity in an information technology context, including computational creativity and its cognitive and collaborative aspects. Papers reporting original unpublished research results on theoretical foundations, IT implementations of decision support and expert systems, as well as case studies of successful applications of the above-mentioned ideas in various fields are equally solicited.

Paper Submission and Publications: We solicit research and experience papers as well as research-in-progress and practitioner reports in any of the technical areas listed under Scope & Topics. The author's name and address MUST NOT appear in the submitted paper for review. This is to facilitate a blind review. The format of the final manuscript should be in a two-column format and 6 pages for full paper, 4 pages for short papers, and 2 pages for poster papers in length.KICSS 2017 conference proceeding will be published.

Organizations

General Chair: Takayuki Ito (Nagoya Institute of Technology, Japan)
Program Chair/Financial Chair: Tokuro Matsuo (Advanced Institute of Industrial Technology, Japan)
Publication Chair: Katsuhide Fujita (Tokyo University of Agriculture and Technology, Japan)
Local Arrangements Chairs: Takanobu Otsuka and Tomomichi Hayakawa (Nagoya Institute of Technology, Japan)
Special Session Chairs: Naoki Fukuta (Shizuoka University, Japan), Tessai Hayama (Nagaoka University of Technology, Japan), Shun Shiramatsu (Nagoya Institute of Technology, Japan), Ryo Kanaomori (Nagoya University, Japan)
Publicity Chair: Takaya Yuizono (JAIST, Japan), Kiyota Hashimoto (PSU, Thailand) (Registration Desk Chair)

The Effects of Sampling Approach on Hard Disk Drive Failure Prediction with Highly Imbalanced Data

Nathaniel Netirungroj[1], Aimamorn Suvichakorn[2], Kittiphan Pomoung[2], and Eakasit Pacharawongsakda[1]

[1]Big Data Engineering program, College of Innovative Technology and Engineering, Dhurakij Pundit University
Bangkok, Thailand
{585162020017, eakasit.pac}@dpu.ac.th
[2]Western Digital (Thailand) Co., Ltd.
Phra Nakhon Si Ayutthaya, Thailand
{aimamorn.suvichakorn, kitiphan.pomoung}@wdc.com

***Abstract*—Hard Disk Drive (HDD) failures are very rare events. Predicting failure is an essential task that manufacturer applied in order to eliminate the defect HDD and improve product quality. In this paper, several sampling methods, which are able to tackle highly class imbalanced data sets from HDD production, were examined. This real world dataset has 356 attributes and 506,239 observations and were collected from several weeks. In this work, seven sampling approaches were applied to re-balanced the datasets and then used to build a predictive model. In the model training phase, seven machine learning algorithms were employed, for example, rule-based, distance-based, probabilistic-based, neural-based, and regression-based. From the experimental results with 343 combinations, we found that Random Forest is the best model as it performed better on most sampling techniques. This work can be used as a guideline in order to deal with Highly Class Imbalanced scenario.**

***Index Terms*—Highly Class Imbalanced, Hard Disk Drive Failure, Over-Sampling, Under-Sampling**

I. Introduction

In electronic hardware industry, magnetic HDD is one of the important part in all kinds of computer machines. This hardware tends to be highly robust in order to keep the machine operate properly. The causes of HDD failure can be roughly defined as two categories which are caused by usage and/or production process. This work discusses about the failure that caused by production process. Normally, a HDD failure is rarely occurred. The rare event information can be found from two main sources, from internal HDD log such as SMART (self-monitoring and report technology). A recent work that successfully improved HDD failure prediction accuracy used SMART information from the individual drive [1], [2], [3] where the data is considered as class imbalanced. However, besides information from SMART, data source that obtain by manufacturer equipments can also be useful as used in [4], [5], [6] to classify HDD failure, because the information comes from origin stage that defines how HDD is created.

In the real-world situation, a class imbalanced is one of the well-known issue in machine learning field. This problem occurs when the dataset has unequally size between class label. For clarity, the class with more observation is labeled as the majority class (MA) and the class with less observation is labeled as the minority class (MI). In training stage, learner attempt to learn from dataset in order to find the optimal parameters which fitting to label class. The dataset that contains skewed distribution of class label usually compromise classifier performance and result in MI misclassification in prediction task [7], [8], because it is influenced by the MA ones. Traditionally, a different between class label in size ratio are 70:30 can be considered as a class imbalanced. For highly class imbalanced, the MA size is usually beyond 90 whereas MI is less than 10. The data set we used in this work has 99.9 percents of MA.

There are three primarily approaches [9], [10] that can be applied to tackle class imbalanced problem including specialize algorithms, features selection, and re-balancing data. This paper concentrates on re-balancing class distribution since it has an impressive result in many cases [11].

II. Background

A class imbalanced problem requires different method from general classifier construction procedure. Additional processes can be added to the task. Exiting process also may need to be customized as well, such as evaluation metric, which describes in section IV. This section, we introduce five main data preprocessing techniques that can be applied to the problem.

A. Re-balancing Algorithms

There are several data sampling techniques, which able to increase MA samples and/or MI samples. This study employs

seven techniques that has a capability to re-sampling our data sets. The short description of each technique is describe here.

1) Random Sampling: This method is to randomly select MA and/or MI samples. It has three approaches including Random Over-Sampling (ROS), Random Under-Sampling (RUS), and Random Over and Under-Sampling (ROUS). ROS randomly duplicate the MI samples in order to increase its size to become as equal as MA samples. The drawback of this method is that it may increase undesirable noise to the data set that leads to over-fitting. On the other hand, RUS randomly dropped the MA samples. The main advantage of this method is that it does not tend to over-fit training dataset, however good samples from the MA may be discarded. ROUS randomly reduce MA samples while randomly increased MI at the same time.

2) SMOTE: Synthetic Minority Over-Sampling Technique [12] is an algorithm that simulates MI samples. It aims to increase the number of MI sample artificially. Unlike the ROS method, SMOTE doesn't duplicate MI samples, but it generates new MI samples that located nearby original MI samples. The algorithm employs k-nearest neighbors technique to find the connecting line between MI data point in feature space then generates new MI sample on that line. One good point of this method is that it reserves all of MA samples. The concern of this method is that it might generate artificial MI samples within the overlapping regions.

3) ROSE: Random Over-Sampling Examples [13] is an algorithm that employs k-nearest neighbors technique and smoothed bootstrap approach to re-balance the dataset. Comparing to SMOTE, ROSE generates artificial MI samples in the feature space neighborhood around the original MI ones by using kernel density estimation instead of knn's . The main benefit of this method is that synthetic samples size and shape can be adjusted.

4) Safe-level SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique (S-SMOTE) [14] is an algorithm that based on SMOTE. It assigns safe level for each original MI then simulates new synthetic MI that lies close to the safe level. Main advantage of this technique is that it generates less noise. The concerns are that it ignores the relation between MA and original MI and may cause artificial MI excessively generated close to MA.

5) Relocating Safe-level SMOTE: Relocate Safe-Level-Synthetic Minority Over-Sampling Technique (RS-SMOTE) [15] is an enhanced version of Safe-level SMOTE. This algorithm is able to calculate the distance between new MI and MA then moves artificial MI away from MA to the nearest original MI instance. Moreover, the outcast original MI is also considered in an attempt to reserve all of MI sample.

B. Machine Learning Algorithms

There are seven algorithms that we employ including Decision Tree, Logistics Regression, Naive Bayes, Neural Network, Support Vector Machine, Random Forest, and Extreme Gradient Boosting. Those algorithms can be categorized into single classifier and multiple classifiers, namely ensemble classifier. The details are describing in the following subsection below.

1) Decision Tree (DT): It is widely used classification algorithm due to transparency of model. It employs concept of information gain to select good variable and build tree structure model. The node in top level has capability to separate data into different class.

2) Logistics Regression (LR): It is a simple mathematical algorithm that measures relationship between the categorical dependent variable and several independent variables by estimating probability using logistic function.

3) Naive Bayes (NB): One of probabilistic algorithm that computes prior probability of each class and conditional probability of variables and classes from the training data. To predict an unseen example, these probabilities are multiplied as score for each class. The model will select the class with highest score.

4) Neural Network (NN): It is a mathematical algorithm which is inspired by structure of biological neural network. It composes of several nodes called neural and interaction between them. These neural are aligned into three main layers: (1) input layer, (2) hidden layer and (3) output layer. The algorithm adopts backpropagation concept to update weight between neural in adjacent layers.

5) Support Vector Machine (SVM): A distance-based algorithm that extensively uses in binary classification task. The main idea of this algorithm is that it projects optimal hyperplanes to draw boundaries between class across multi-dimensional spaces, we adopt Radial Basis Kernel function.

6) Random Forest (RF): It has been getting more attention recently since it widely applies in several domains and performs better performance than others while still reserving the explanation ability. The basic idea of this algorithm is similar to Decision Tree except that it grows many trees (forest) by randomly select training data in order to create different committee.

7) Extreme Gradient Boosting (XGB): This algorithm is an implementation of Boosted Decision Trees and follow the principle of Gradient Boosting. It constructs weak learners (shallow) sequentially, unlike Random Forest where trees (fully grown) are randomly created in parallel. One good point is that it has a capability of re-adjusting subsequence example weight to minimize bias by using error from previous trees to construct the new ones.

III. Methods

We propose a framework which handles highly class imbalanced by re-balancing class proportion and evaluating learner performance properly. The objective of this study is to find the most suitable approach which is able to capture MI sample as much as possible while deliver minimum misclassification result.

A. Proposed Framework

This study employs seven techniques that capable to alter our data sets. The short description of each technique is already described in section II.

Fig. 1: A propose framework conceptual

We first eliminated missing value by replacing with median value of its attribute. Then, we transformed each data set with standardization technique to adjust data scaling and applied seven sampling techniques to re-balance the data sets. After that, we evaluated classification performance with suitable metrics.

IV. Experimental Settings

The average performance of classifier is obtained by using 10-fold cross validation with same random seed. The hold-out data is set to twenty percent from total of each dataset for validation purpose.

A. Data Description

The comprehensive variables in this study is collected from HDD assembling and manufacturing process during several functional test sequences and sensors at different test stress, duration and conditions where each data point is flagged into two class labels, i.e., failures and passers. Table 1 presents the proportion between each class where the failures are considered MI and passers, MA, respectively

TABLE I: Characteristic of each dataset

Dataset	Proportion in % (MA:MI)	MA sample	MI sample
1	99.96392 : 0.03608	30478	11
2	99.97970 : 0.02030	34479	7
3	99.99311 : 0.00689	72544	5
4	99.99079 : 0.00921	119463	11
5	99.99252 : 0.00748	106954	8
6	99.99237 : 0.00763	91740	7
7	99.98814 : 0.01186	50581	6

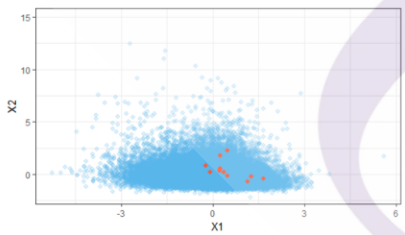


Fig. 2: A scatter plot of dataset no. 1 with two random attributes. The blue and red dots represent MA and MI class correspondingly.

B. Evaluation Metrics

For comparison, our evaluation metrics are defined based on the confusion matrix that break-downs the result from classifier versus that of the actual ones, as illustrated in Table II

f[!ht]

TABLE II: A confusion matrix

Predicted/Actual	MI	MA
MI	True Positive	False Positive
MA	False Negative	True Negative

where TP, TN, FP, FN signify stand for true positive, true negative, false positive, and false negative respectively and six metrics used in this paper are as follow:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Accuracy is a common evaluation metric that explains model overall performance. It is calculated by summarizing correctly predicted class label and divided by total number of all classes. Since the MI is our desirable class, applying accuracy metric may compromise model performance, because it appreciates the proportion of correct class to total class. A recent work shows that Precision, Recall, Specificity, F1-score, and G-mean are more suitable metrics than overall accuracy regarding this situation [16]. Additionally, geometric mean metric is broadly adopted in many class imbalanced studies. It denotes trade-off between classifier performance on both MI and MA. Recall (sensitivity) and specificity were used to calculate the metric. Our study uses this indicator to evaluate classification performance. These metrics are defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F_1 score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$GM = \sqrt{\frac{TP}{TP + FN} \times \frac{TP}{TP + FP}} \tag{6}$$

V. Experimental Results

According to our setup in section IV, we applied seven sampling techniques including RUS, ROS, ROUS, SMOTE, ROSE, S-SMOTE, RS-SMOTE to adjust original data sets class distribution. Then, we used seven classification algorithms including DT, LR, NB, NN, SVM, RF, and XGB to learn new data sets from previous stage with its default settings. All mentioned metrics were used to indicate model performances. Experimental result is presents in the table III and IV.

TABLE III
Experimental results of propose framework

RB	Metrics	Dataset 1							Dataset 2							Dataset 3							Dataset 4						
		DT	GLM	NB	NN	SVM	RF	XGB	DT	GLM	NB	NN	SVM	RF	XGB	DT	GLM	NB	NN	SVM	RF	XGB	DT	GLM	NB	NN	SVM	RF	XGB
BASELINE	Accuracy	.9996	.9990	.9995	.9995	.9996	.9996	.9996	.9998	.9996	.3227	.9998	.9998	.9998	.9998	.9999	.9996	.4199	.9999	.9999	.9999	.9999	.9999	.7208	.0354	.9999	.9999	.9999	.9999
	Recall	0	0	0	0	0	0	0	0	0	.7143	0	0	0	0	0	0	1	0	0	0	0	0	.2727	.9091	0	0	0	0
	Precision	0	0	0	0	0	0	0	0	0	.0002	0	0	0	0	0	0	.0001	0	0	0	0	0	.0001	.0001	0	0	0	0
	Specificity	1	.9994	.9998	.9998	1	1	1	1	.9998	.3227	1	1	1	1	1	.9997	.4198	1	1	1	1	1	.7208	.0353	1	1	1	1
	FP Rate	0	.0006	.0002	.0002	0	0	0	0	.0002	.6773	0	0	0	0	0	.0003	.5802	0	0	0	0	0	.2792	.9647	0	0	0	0
	F1	0	0	0	0	0	0	0	0	0	.0004	0	0	0	0	0	0	.0002	0	0	0	0	0	.0002	.0002	0	0	0	0
	GM	0	0	0	0	0	0	0	0	0	**.0124**	0	0	0	0	0	0	**.0109**	0	0	0	0	0	.0050	**.0089**	0	0	0	0
RUS	Accuracy	.9348	.7826	.9130	.8261	.8478	.9348	.9565	.6970	.7273	.9091	.8788	.9091	.9697	.9394	.4762	.7143	.619	.9048	.6667	1	.8571	.4783	.8261	.8913	.7609	.6739	.9348	.8478
	Recall	.9091	1	1	1	.9091	1	1	.5714	1	1	1	1	1	1	.1	1	.2000	1	.3000	1	1	.0909	1	.9091	1	.4545	.9091	1
	Precision	.9524	.6875	.8462	.7333	.8000	.8800	.9167	.6667	.6087	.8235	.7778	.8235	.9333	.875	.3333	.6250	1	.8333	1	1	.7692	.3333	.7333	.8696	.6667	.7692	.9524	.7586
	Specificity	.9583	.5833	.8333	.6667	.7917	.8750	.9167	.7895	.5263	.8421	.7895	.8421	.9474	.8947	.8182	.4545	1	.8182	1	1	.7273	.8333	.6667	.8750	.5417	.875	.9583	.7083
	FP Rate	.0417	.4167	.1667	.3333	.2083	.1250	.0833	.2105	.4737	.1579	.2105	.1579	.0526	.1053	.1818	.5455	0	.1818	0	0	.2727	.1667	.3333	.1250	.4583	.125	.0417	.2917
	F1	.9302	.8148	.9167	.8462	.8511	.9362	.9565	.6154	.7568	.9032	.8750	.9032	.9655	.9333	.1538	.7692	.3333	.9091	.4615	1	.8696	.1429	.8462	.8889	.8000	.5714	.9302	.8627
	GM	.9305	.8292	.9199	.8563	.8528	.9381	**.9574**	.6172	.7802	.9075	.8819	.9075	**.9661**	.9354	.1826	.7906	.4472	.9129	.5477	**1**	.8771	.1741	.8563	.8891	.8165	.5913	**.9305**	.8710
ROS	Accuracy	.9025	.9996	1	.9992	1	1	.9957	.9265	.9999	1	.9993	1	1	.9990	.9512	.9981	.9991	.9996	1	1	.9997	.8527	.8707	.5003	.9996	1	1	1
	Recall	.9425	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.8448	1	1	1	1	1
	Precision	.8719	.9991	.9999	.9983	1	1	.9914	.8713	.9997	1	.9986	1	1	.9981	.9109	.9961	.9983	.9993	1	1	.9995	.7724	.8909	.5001	.9992	1	1	.9999
	Specificity	.8629	.9991	.9999	.9983	1	1	.9914	.8537	.9997	1	.9986	1	1	.9981	.9025	.9961	.9983	.9993	1	1	.9995	.7055	.8966	.0009	.9992	1	1	.9999
	FP Rate	.1371	.0009	.0001	.0017	0	0	.0086	.1463	.0003	0	.0014	0	0	.0019	.0975	.0039	.0017	.0007	0	0	.0005	.2945	.1034	.9991	.0008	0	0	.0001
	F1	.9058	.9996	1	.9992	1	1	.9957	.9312	.9999	1	.9993	1	1	.9990	.9534	.9981	.9991	.9996	1	1	.9997	.8716	.8672	.6667	.9996	1	1	1
	GM	.9065	.9996	**1**	.9992	**1**	**1**	.9957	.9334	.9999	**1**	.9993	**1**	**1**	.9990	.9544	.9981	.9991	.9996	**1**	**1**	.9997	.8789	.8675	.7072	.9996	**1**	**1**	**1**
ROUS	Accuracy	.9080	.9993	.9998	.9985	1	1	.9963	.9235	.9998	.9999	.9989	1	1	.9990	.9455	.9997	.9983	.9996	1	1	.9998	.8507	.5930	.4978	.9993	1	1	.9924
	Recall	.9537	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.5915	1	1	1	1	1
	Precision	.8721	.9987	.9996	.9970	1	1	.9925	.8655	.9995	.9999	.9977	1	1	.9979	.9008	.9993	.9966	.9992	1	1	.9996	.7693	.5911	.4978	.9986	1	1	.9849
	Specificity	.8634	.9987	.9996	.9970	1	1	.9926	.8494	.9995	.9999	.9978	1	1	.9979	.8921	.9993	.9967	.9992	1	1	.9996	.7027	.5944	.0001	.9987	1	1	.9848
	FP Rate	.1366	.0013	.0004	.0030	0	0	.0074	.1506	.0005	.0001	.0022	0	0	.0021	.1079	.0007	.0033	.0008	0	0	.0004	.2973	.4056	.9999	.0013	0	0	.0152
	F1	.9111	.9993	.9998	.9985	1	1	.9962	.9279	.9998	.9999	.9989	1	1	.9989	.9478	.9997	.9983	.9996	1	1	.9998	.8696	.5913	.6647	.9993	1	1	.9924
	GM	.9120	.9993	.9998	.9985	**1**	**1**	.9962	.9303	.9998	.9999	.9989	**1**	**1**	.9989	.9491	.9997	.9983	.9996	**1**	**1**	.9998	.8771	.5913	.7055	.9993	**1**	**1**	.9924
SMOTE	Accuracy	.9236	.9997	.9962	.9996	1	1	.9986	.9573	.9998	.9987	.9998	1	1	.9990	.9488	.9994	.7929	.9998	1	1	.9997	.9007	.8381	.5723	.9998	1	1	.9999
	Recall	.9437	1	1	1	1	1	1	.9518	1	1	1	1	1	1	.9869	1	1	1	1	1	1	.9363	.8247	1	1	1	1	1
	Precision	.9072	.9994	.9925	.9992	1	1	.9971	.9624	.9997	.9974	.9997	1	1	.9979	.9170	.9988	.7071	.9997	1	1	.9994	.8740	.8475	.5390	.9995	1	1	.9998
	Specificity	.9035	.9994	.9925	.9992	1	1	.9971	.9628	.9997	.9974	.9997	1	1	.9979	.9107	.9988	.5858	.9997	1	1	.9994	.8650	.8516	.1447	.9995	1	1	.9998
	FP Rate	.0965	.0006	.0075	.0008	0	0	.0029	.0372	.0003	.0026	.0003	0	0	.0021	.0893	.0012	.4142	.0003	0	0	.0006	.1350	.1484	.8553	.0005	0	0	.0002
	F1	.9251	.9997	.9963	.9996	1	1	.9986	.9571	.9998	.9987	.9998	1	1	.9990	.9507	.9994	.8284	.9998	1	1	.9997	.9041	.8359	.7004	.9998	1	1	.9999
	GM	.9253	.9997	.9963	.9996	**1**	**1**	.9986	.9571	.9998	.9987	.9998	**1**	**1**	.9990	.9513	.9994	.8409	.9998	**1**	**1**	.9997	.9046	.8360	.7342	.9998	**1**	**1**	.9999
ROSE	Accuracy	.9999	.9731	1	.9747	.9999	1	1	1	.9890	1	.9876	1	1	1	1	.9921	.9999	.9923	1	1	1	1	.9412	.9999	.9736	1	1	1
	Recall	1	.9782	1	.9822	1	1	1	1	.9914	1	.9917	1	1	1	1	.9952	1	.9961	1	1	1	1	.9540	1	.9862	1	1	1
	Precision	.9999	.9678	1	.9671	.9999	1	1	.9999	.9863	1	.9832	1	1	1	1	.9890	.9998	.9885	1	1	1	1	.9297	.9999	.9617	1	1	1
	Specificity	.9999	.9682	1	.9674	.9999	1	1	.9999	.9867	1	.9836	1	1	1	1	.9891	.9998	.9886	1	1	1	1	.9285	.9999	.9611	1	1	1
	FP Rate	.0001	.0318	0	.0326	.0001	0	0	.0001	.0133	0	.0164	0	0	0	0	.0109	.0002	.0114	0	0	0	0	.0715	.0001	.0389	0	0	0
	F1	.9999	.9729	1	.9746	.9999	1	1	1	.9889	1	.9875	1	1	1	1	.9921	.9999	.9923	1	1	1	1	.9417	.9999	.9738	1	1	1
	GM	.9999	.9730	**1**	.9746	.9999	**1**	**1**	**1**	.9889	**1**	.9875	**1**	**1**	**1**	**1**	.9921	.9999	.9923	**1**	**1**	**1**	**1**	.9417	.9999	.9739	**1**	**1**	**1**
S-SMOTE	Accuracy	.9228	.9997	.9968	.9995	1	1	.9987	.9589	.9998	.9987	.9995	1	1	.9989	.9488	.9993	.7928	.9997	1	1	.9997	.9011	.6942	.5726	.9995	1	1	.9999
	Recall	.9422	1	1	1	1	1	1	.9554	1	1	1	1	1	1	.9870	1	1	1	1	1	1	.9398	.7355	1	1	1	1	1
	Precision	.9071	.9994	.9936	.9989	1	1	.9973	.9621	.9997	.9975	.9990	1	1	.9978	.9170	.9986	.7070	.9994	1	1	.9995	.8724	.6555	.5391	.9991	1	1	.9998
	Specificity	.9035	.9994	.9935	.9989	1	1	.9973	.9624	.9997	.9974	.9990	1	1	.9978	.9107	.9986	.5856	.9994	1	1	.9995	.8625	.6575	.1452	.9991	1	1	.9998
	FP Rate	.0965	.0006	.0065	.0011	0	0	.0027	.0376	.0003	.0026	.0010	0	0	.0022	.0893	.0014	.4144	.0006	0	0	.0005	.1375	.3425	.8548	.0009	0	0	.0002
	F1	.9243	.9997	.9968	.9995	1	1	.9987	.9588	.9998	.9987	.9995	1	1	.9989	.9507	.9993	.8284	.9997	1	1	.9997	.9048	.6932	.7006	.9995	**1**	1	.9999
	GM	.9245	.9997	.9968	.9995	**1**	**1**	.9987	.9588	.9998	.9987	.9995	**1**	**1**	.9989	.9513	.9993	.8408	.9997	**1**	**1**	.9997	.9054	.6943	.7343	.9995	**1**	**1**	.9999
RS-SMOTE	Accuracy	.9234	.9997	.9966	.9993	1	1	.9986	.9592	.9998	.9987	.9997	1	1	.9990	.9491	.9993	.7925	.9997	1	1	.9997	.9006	.6866	.5713	.9996	1	1	.9999
	Recall	.9433	1	1	1	1	1	1	.9556	1	1	1	1	1	1	.9875	1	1	1	1	1	1	.9363	.6935	1	1	1	1	1
	Precision	.9071	.9994	.9932	.9987	1	1	.9973	.9626	.9997	.9974	.9994	1	1	.9981	.9171	.9987	.7068	.9995	1	1	.9994	.8740	.6841	.5384	.9993	1	1	.9998
	Specificity	.9034	.9994	.9932	.9987	1	1	.9972	.9628	.9997	.9974	.9994	1	1	.9981	.9107	.9987	.5851	.9995	1	1	.9994	.8650	.6797	.1425	.9993	1	1	.9998
	FP Rate	.0966	.0006	.0068	.0013	0	0	.0028	.0372	.0003	.0026	.0006	0	0	.0019	.0893	.0013	.4149	.0005	0	0	.0006	.1350	.3203	.8575	.0007	0	0	.0002
	F1	.9249	.9997	.9966	.9993	1	1	.9986	.9591	.9998	.9987	.9997	1	1	.9990	.9510	.9993	.8282	.9997	1	1	.9997	.9040	.6887	.6999	.9996	1	1	.9999
	GM	.9250	.9997	.9966	.9993	**1**	**1**	.9986	.9591	.9998	.9987	.9997	**1**	**1**	.9990	.9516	.9993	.8407	.9997	**1**	**1**	.9997	.9046	.6888	.7337	.9996	**1**	**1**	.9999

TABLE IV
Experimental results of propose framework

RB	Metrics	Dataset 5							Dataset 6							Dataset 7						
		DT	GLM	NB	NN	SVM	RF	XGB	DT	GLM	NB	NN	SVM	RF	XGB	DT	GLM	NB	NN	SVM	RF	XGB
BASELINE	Accuracy	.9999	.9995	.9999	.9999	.9999	.9999	.9999	.9999	.9996	.1638	.9999	.9999	.9999	.9999	.9999	.9993	.9998	.9999	.9999	.9999	.9999
	Recall	0	0	0	0	0	.1250	0	0	0	.7143	0	0	0	0	0	0	0	0	0	0	0
	Precision	0	0	0	0	0	.5000	0	0	0	.0001	0	0	0	0	0	0	0	0	0	0	0
	Specificity	1	.9996	1	1	1	1	1	1	.9997	.1637	.9999	1	1	1	1	.9995	1	1	1	1	1
	FP Rate	0	.0004	0	0	0	0	0	0	.0003	.8363	.0001	0	0	0	0	.0005	0	0	0	0	0
	F1	0	0	0	0	0	.2000	0	0	0	.0001	0	0	0	0	0	0	0	0	0	0	0
	GM	0	0	0	0	0	**.2500**	0	0	0	**.0068**	0	0	0	0	0	0	0	0	0	0	0
RUS	Accuracy	.6757	.8108	.8649	.7568	.8108	.9459	.8108	.5455	.7576	.9091	.6970	.8182	.9697	.8485	.6667	.5926	.9259	.7407	.6667	.8148	.7778
	Recall	.7500	1	.8750	.8750	.8750	.8750	.8125	.7143	1	1	1	.7143	.9286	1	.8333	1	.8333	.8333	.3333	.8333	.8333
	Precision	.6000	.6957	.8235	.6667	.7368	1	.7647	.4762	.6364	.8235	.5833	.8333	1	.7368	.5882	.5217	1	.6667	.8000	.7692	.7143
	Specificity	.6190	.6667	.8571	.6667	.7619	1	.8095	.4211	.5789	.8421	.4737	.8947	1	.7368	.5333	.2667	1	.6667	.9333	.8000	.7333
	FP Rate	.3810	.3333	.1429	.3333	.2381	0	.1905	.5789	.4211	.1579	.5263	.1053	0	.2632	.4667	.7333	0	.3333	.0667	.2000	.2667
	F1	.6667	.8205	.8485	.7568	.8000	.9333	.7879	.5714	.7778	.9032	.7368	.7692	.9630	.8485	.6897	.6857	.9091	.7407	.4706	.8000	.7692
	GM	.6708	.8341	.8489	.7638	.8030	**.9354**	.7882	.5832	.7977	.9075	.7638	.7715	**.9636**	.8584	.7001	.7223	**.9129**	.7454	.5164	.8006	.7715
ROS	Accuracy	.8823	.9936	1	.9997	1	1	.9983	.8771	.9981	1	.9996	1	1	.9989	.9217	.9982	1	.9993	1	1	.9984
	Recall	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Precision	.8093	.9874	1	.9994	1	1	.9965	.8025	.9963	1	.9991	1	1	.9979	.8644	.9963	1	.9986	1	1	.9967
	Specificity	.7647	.9872	1	.9994	1	1	.9965	.7546	.9963	1	.9991	1	1	.9979	.8439	.9963	1	.9986	1	1	.9967
	FP Rate	.2353	.0128	0	.0006	0	0	.0035	.2454	.0037	0	.0009	0	0	.0021	.1561	.0037	0	.0014	0	0	.0033
	F1	.8946	.9936	1	.9997	1	1	.9983	.8905	.9982	1	.9996	1	1	.9989	.9273	.9982	1	.9993	1	1	.9984
	GM	.8996	.9937	**1**	.9997	**1**	**1**	.9983	.8958	.9982	**1**	.9996	**1**	**1**	.9989	.9297	.9982	**1**	.9993	**1**	**1**	.9984
ROUS	Accuracy	.8881	.9955	1	.9996	1	1	.9985	.8896	.9987	1	.9996	1	1	.9989	.9104	.9997	1	.9994	1	1	.9982
	Recall	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Precision	.8164	.9910	.9999	.9991	1	1	.9970	.8182	.9974	1	.9992	1	1	.9977	.8466	.9993	1	.9989	1	1	.9964
	Specificity	.7772	.9910	.9999	.9991	1	1	.9970	.7804	.9974	1	.9992	1	1	.9977	.8229	.9993	1	.9989	1	1	.9965
	FP Rate	.2228	.0090	.0001	.0009	0	0	.0030	.2196	.0026	0	.0008	0	0	.0023	.1771	.0007	0	.0011	0	0	.0035
	F1	.8989	.9955	1	.9996	1	1	.9985	.9000	.9987	1	.9996	1	1	.9989	.9169	.9997	1	.9994	1	1	.9982
	GM	.9036	.9955	**1**	.9996	**1**	**1**	.9985	.9046	.9987	1	.9996	**1**	**1**	.9989	.9201	.9997	1	.9994	**1**	**1**	.9982
SMOTE	Accuracy	.8646	.9985	.9910	.9998	1	1	.9987	.8882	.9990	.9980	.9998	1	1	.9994	.9266	.999	.9989	.9996	1	1	.9992
	Recall	.8683	1	1	1	1	1	1	.9712	1	1	1	1	1	1	1	1	1	1	1	1	1
	Precision	.8619	.9970	.9824	.9997	1	1	.9975	.8330	.9980	.9960	.9996	1	1	.9987	.8720	.9979	.9977	.9992	1	1	.9984
	Specificity	.8609	.9970	.9821	.9997	1	1	.9975	.8054	.9980	.9960	.9996	1	1	.9987	.8532	.9979	.9977	.9992	1	1	.9984
	FP Rate	.1391	.0030	.0179	.0003	0	0	.0025	.1946	.0020	.0040	.0004	0	0	.0013	.1468	.0021	.0023	.0008	0	0	.0016
	F1	.8651	.9985	.9911	.9998	1	1	.9987	.8968	.9990	.9980	.9998	1	1	.9994	.9316	.9990	.9989	.9996	1	1	.9992
	GM	.8651	.9985	.9912	.9998	**1**	**1**	.9987	.8994	.9990	.9980	.9998	**1**	**1**	.9994	.9338	.9990	.9989	.9996	**1**	**1**	.9992
ROSE	Accuracy	1	.9713	1	.9891	1	1	1	1	.9767	.9999	.9891	1	1	1	.9823	.9656	1	.9823	1	1	1
	Recall	1	.9805	1	.9914	1	1	1	1	.9814	1	.9909	1	1	1	.998	.9733	1	.9863	1	1	1
	Precision	1	.9625	1	.9868	1	1	1	1	.9720	.9999	.9873	1	1	1	.9672	.9578	1	.9782	1	1	.9999
	Specificity	1	.9621	1	.9868	1	1	1	1	.9721	.9999	.9874	1	1	1	.9669	.9581	1	.9785	1	1	.9999
	FP Rate	0	.0379	0	.0132	0	0	0	0	.0279	.0001	.0126	0	0	0	.0331	.0419	0	.0215	0	0	.0001
	F1	1	.9714	1	.9891	1	1	1	1	.9767	.9999	.9891	1	1	1	.9824	.9655	1	.9822	1	1	1
	GM	**1**	.9715	**1**	.9891	**1**	**1**	**1**	**1**	.9767	.9999	.9891	**1**	**1**	**1**	.9825	.9655	**1**	.9822	**1**	**1**	**1**
S-SMOTE	Accuracy	.8727	.9955	.9909	.9998	1	1	.9987	.8885	.9989	.9980	.9998	1	1	.9987	.9266	.9987	.9989	.9998	1	1	.9991
	Recall	.9024	1	1	1	1	1	1	.9716	1	1	1	1	1	1	1	1	1	1	1	1	1
	Precision	.8519	.9910	.9821	.9996	1	1	.9974	.8331	.9978	.9961	.9996	1	1	.9974	.8720	.9974	.9978	.9995	1	1	.9983
	Specificity	.8431	.9910	.9818	.9996	1	1	.9974	.8054	.9978	.9961	.9996	1	1	.9974	.8532	.9974	.9977	.9995	1	1	.9983
	FP Rate	.1569	.0090	.0182	.0004	0	0	.0026	.1946	.0022	.0039	.0004	0	0	.0026	.1468	.0026	.0023	.0005	0	0	.0017
	F1	.8764	.9955	.9910	.9998	1	1	.9987	.8970	.9989	.9981	.9998	1	1	.9987	.9316	.9987	.9989	.9998	1	1	.9992
	GM	.8768	.9955	.9910	.9998	**1**	**1**	.9987	.8997	.9989	.9981	.9998	**1**	**1**	.9987	.9338	.9987	.9989	.9998	**1**	**1**	.9992
RS-SMOTE	Accuracy	.8518	.9983	.9890	.9997	1	1	.9988	.8881	.9989	.9983	.9997	1	1	.9994	.9266	.9989	.9989	.9998	1	1	.9990
	Recall	.8217	1	1	1	1	1	1	.9709	1	1	1	1	1	1	1	1	1	1	1	1	1
	Precision	.8743	.9966	.9784	.9994	1	1	.9975	.8330	.9978	.9966	.9995	1	1	.9989	.872	.9978	.9978	.9996	1	1	.9979
	Specificity	.8819	.9966	.9780	.9994	1	1	.9975	.8054	.9978	.9966	.9995	1	1	.9989	.8532	.9978	.9978	.9996	1	1	.9979
	FP Rate	.1181	.0034	.0220	.0006	0	0	.0025	.1946	.0022	.0034	.0005	0	0	.0011	.1468	.0022	.0022	.0004	0	0	.0021
	F1	.8472	.9983	.9891	.9997	1	1	.9988	.8967	.9989	.9983	.9997	1	1	.9994	.9316	.9989	.9989	.9998	1	1	.9990
	GM	.8476	.9983	.9892	.9997	**1**	**1**	.9988	.8993	.9989	.9983	.9997	**1**	**1**	.9994	.9338	.9989	.9989	.9998	**1**	**1**	.9990

From our result, Random Forest and Support Vector Machine with Radial Basis Kernel function are able to classify MI better than others while Decision Tree delivered unimpressive score. Random Forest performed best with every sampling techniques except Random Under-Sampling on dataset 1 and 7. The concern point is that both SVM and RF could possibly excessively be fitted to training dataset. As such, Logistics Regression, Neural Network, and Extreme Gradient Boosting may be a potential optional model to consider regarding several sampling methods we used.

VI. Conclusion

Highly class imbalanced data is a great obstacle in machine learning and data mining field. Re-balancing data is an essential process that helps increase model performances. Indeed, evaluation measurement plays important part in this scenario as it indicates the actual model performance. We introduced data re-balancing framework that significantly improved HDD failure prediction. It allows all models to capture rare event better. We conducted experiments with various sampling strategies including Random Sampling, and Synthetic Sampling. Seven machine learning algorithms both single and ensemble classifier were used on balanced data sets. It seems to us that Random Forest is the champion model which dominated others on most of the re-balanced training data. However, our data sets contain huge amount of MA in comparison with MI ones. We have found that there are certain results that models cannot classify the data correctly. The similarity between majority and minority class is quite close to each other. There is a class overlapping occurring in the data sets. It is still an open challenge for us to overcome in future work. We need to explore further through the data in order to understand the association between class.

References

[1] G. Hamerly, and C. Elkan, "Bayesian approaches to failure prediction for disk drives", ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 202-209.

[2] M. Botezatu, I. Giurgiu, J. Bogojeska, and D. Wiesmann, "Predicting Disk Replacement towards Reliable Data Centers", KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 39-48.

[3] J. Li, "Hard Disk Drive Failure Prediction Using Classification and Regression Trees", 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Atlanta, GA, 2014, pp. 383-394.

[4] F. Sun, "Predicting Customer Integration Fallout Based on Producer's Final Quality Audit." Reliability and Maintainability Symposium, 2007. RAMS'07. Annual. IEEE, 2007.

[5] K. Kerdprasop and N. Kerdprasop, "A Data Mining Approach to Automate Fault Detection Model Development in the Semiconductor Manufacturing Process", International Journal of Mechanics, Issue 4, Volume 5, 2011, pp. 336-344.

[6] C. Chien, W. Wang, and J. Cheng, "Data mining for yield enhancement in semiconductor manufacturing and an empirical study", International Journal of Mechanics, Issue 4, Volume 5, 2011, pp. 336-344.

[7] P. Branco, L. Torgo, and R. Ribeiro, "A Survey of Predictive Modelling under Imbalanced Distributions", ACM Computing Surveys (CSUR) Volume 49 Issue 2, November 2016, Article No. 31.

[8] R. Blagus, and L. Lusa, "Class Prediction for High Dimensional Class Imbalanced Data" BMC Bioinformatics 11 (2010): 523. PMC. Web. 9 Aug. 2017.

[9] R. Longadge, S. Dongre, and L. Malik, "Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network, Volume: 2, Issue 1, 2013, pp. 83 87.

[10] X. Guo, Y. Yin, C. Dong, and G. Yang and G. Zhou, "On the Class Imbalance Problem", 2008 Fourth International Conference on Natural Computation, Jinan, 2008, pp. 192 201.

[11] C. Seiffert, T. Khoshgoftaar, J. Van Hulse and A. Folleco, "An Empirical Study of the Classification Performance of Learners on Imbalanced and Noisy Software Quality Data", 2007 IEEE International Conference on Information Reuse and Integration, Las Vegas, IL, 2007, pp. 651 658.

[12] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over sampling Technique", Journal Of Artificial Intelligence Research, 2002, Volume 16, pp. 321 357.

[13] N. Lunardon, G. Menardi, and N. Torelli, "ROSE: A Package for Binary Imbalanced Learning", The R Journal Vol. 6/1, June 2014.

[14] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe Level SMOTE:Safe Level Synthetic Minority Over Sampling TEchnique for Handling the Class Imbalanced Problem" PAKDD '09 Proceedings of the 13th Pacific Asia Conference on Advances in Knowledge Discovery and Data Mining, 2009, pp. 475 482.

[15] W. Siriseriwan and K. Sinapiromsaran, "The Effective Redistribution for Imbalance Dataset: Relocating Safe eevel SMOTE with Minority Outcast Handling", Chiang Mai J. Sci. 2016;43(1) : 234 246, http://epg.science.cmu.ac.th/ejournal/Contributed.

[16] G. Nguyen, A. Bouzerdoum, and S. Phung, "Learning pattern classification tasks with imbalanced data sets", In P. Yin (Eds.), Pattern recognition, 2009, pp. 193-208.

ประวัติผู้เขียน

ชื่อ นามสกุล	ณัฐณัย เนติรุ่งโรจน์
ประวัติการศึกษา	พ.ศ. 2556 ปริญญาตรี สาขาบริหารธุรกิจระหว่างประเทศ คณะบริหารธุรกิจ มหาวิทยาลัยหอการค้าไทย
ตำแหน่งและสถานที่ทำงานปัจจุบัน	นักวิทยาศาสตร์ข้อมูล (ระบบการเรียนรู้ของเครื่องอัตโนมัติและการสร้างตัวแปร) บริษัท ทรู ดิจิตอล แอนด์ มีเดีย แพลตฟอร์ม จำกัด