การประยุกต์ใช้การเรียนรู้ด้วยเครื่องเพื่อตรวจจับงานเสียที่เกิดขึ้นในกระบวนการทดสอบฮาร์ดดิสก์ไดรฟ์

น.ส.อรุณี ศรีดี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

6270314021_2039304467

CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11
2039304467

APPLYING MACHINE LEARNING TECHNIQUES TO DETECT FAILURE IN HARD DISK DRIVE TEST PROCESS

Miss Arunee Sridee

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2021

Copyright of Chulalongkorn University

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

หัวข้อวิทยานิพนธ์	การประยุกต์ใช้การเรียนรู้ด้วยเครื่องเพื่อตรวจจับงานเสียที่เกิดขึ้นในกระบวนการทดสอบฮาร์ดดิสก์ไดรฟ์
โดย	น.ส.อรุณี ศรีดี
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.วิวัฒน์ วัฒนาวุฒิ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร.วรเศรษฐ สุวรรณิก)

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

อรุณี ศรีดี : การประยุกต์ใช้การเรียนรู้ด้วยเครื่องเพื่อตรวจจับงานเสียที่เกิดขึ้นในกระบวนการทดสอบฮาร์ดดิสก์ไดรฟ์. (APPLYING MACHINE LEARNING TECHNIQUES TO DETECT FAILURE IN HARD DISK DRIVE TEST PROCESS) อ.ที่ปรึกษาหลัก : ศ. ดร.ประภาส จงสถิตย์วัฒนา

งานวิจัยฉบับนี้นำเสนอเทคนิคการเรียนรู้ด้วยเครื่องในการตรวจจับงานเสียที่เกิดจากการอ่านสัญญาณเซอร์โวในกระบวนการทดสอบ ลักษณะของชุดข้อมูลมีจำนวนหลายมิติและมีความไม่สมดุลสูง มีการเลือกคุณลักษณะด้วย Filter Method และ Embedded Method เพื่อลดมิติของข้อมูล มีการประยุกต์ใช้เทคนิคการเรียนรู้ด้วยเครื่อง 2 อัลกอริทึม คือ SVM และ XGBoost ร่วมกับ 3 วิธีการในการจัดการข้อมูลที่ไม่สมดุล คือ SMOTE , Different Cost Learner และ SMOTE กับ Different Cost รวมเป็น 6 วิธีการและทำการเปรียบเทียบผลการทดลอง SVM ให้ประสิทธิภาพที่ดีในการวัดด้วย ROC AUC แต่ให้ประสิทธิภาพที่ค่อนข้างต่ำในการวัดด้วย PRC AUC ขณะที่ XGB ให้ประสิทธิภาพที่ดีทั้ง ROC AUC และ PRC AUC โดยวิธีการของ XGB SMOTE ให้ประสิทธิภาพที่ดีที่สุดที่ ROC AUC 91%, PRC AUC 73% และ Accuracy, Precision, Recall และ F1-Score ที่ 97%

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

6270314021 : MAJOR COMPUTER SCIENCE

KEYWORD: SVM, imbalance data, hard disk drive, feature selection, XGBoost

Arunee Sridee : APPLYING MACHINE LEARNING TECHNIQUES TO DETECT FAILURE IN HARD DISK DRIVE TEST PROCESS. Advisor: PRABHAS CHONGSTITVATANA

This paper presents machine learning techniques to detect servo track read back failure in hard disk drive manufacturing test process. The data is high-dimensional and highly imbalanced. Feature selection techniques with filter method and embedded method are used to reduce the dimension of data. We apply two machine learning algorithms, each algorithm applied three different imbalanced data handling methods; total six methods to compare: SMOTE, Different Cost and SMOTE with Different Cost to handle imbalance data. Several machine learning methods are compared. The SVM algorithm shows good performance on ROC AUC while low performance on PRC AUC. The XGB algorithm shows good performance on both ROC AUC and PRC AUC. The XGB SMOTE achieved the best performance with ROC AUC 91%, PRC AUC 73% and Accuracy, Precision, Recall, F1-score 97%.

Field of Study: Computer Science Student's Signature .....

Academic Year: 2021 Advisor's Signature .....

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สามารถสำเร็จลุล่วงไปได้ด้วยดีด้วยความอนุเคราะห์จาก บริษัท Western Digital ที่สนับสนุนทุนการศึกษาและข้อมูลที่ใช้ในงานวิจัย

ขอกราบขอบพระคุณ ศ.ดร. ประภาส จงสถิตย์วัฒนา อ.ที่ปรึกษา ที่สละเวลา ให้คำปรึกษา ให้ความรู้ คอยช่วยเหลือและแก้ไขข้อบกพร่องต่างๆ จนทำให้วิทยานิพนธ์เสร็จสมบูรณ์

ขอกราบขอบพระคุณ รศ.ดร. วิวัฒน์ วัฒนาวุฒิ และ รศ.ดร. วรเศรษฐ สุวรรณิก คณะกรรมการสอบวิทยานิพนธ์ ที่ให้คำแนะนำ แนวคิด และชี้ข้อบกพร่อง เพื่อนำมาปรับปรุงวิทยานิพนธ์

ขอกราบขอบพระคุณ พ่อ แม่ ครูบาอาจารย์ ที่คอยเป็นกำลังใจ เป็นที่ปรึกษา เป็นแรงผลักดันและคอยสนับสนุนช่วยเหลือ ตลอดเวลา

ขอขอบคุณหัวหน้างาน เพื่อนๆบริษัท Western Digital ที่ให้คำแนะนำและแลกเปลี่ยนความรู้

ขอบคุณเพื่อนๆ พี่ๆ น้องๆ และเจ้าหน้าที่ ปริญญาโทสาขาวิทยาศาสตร์คอมพิวเตอร์ที่คอยช่วยเหลือ ให้คำปรึกษา

ขอบคุณค่ะ

อรุณี ศรีดี

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

สารบัญ

หน้า

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

สารบัญตาราง

หน้า

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

สารบัญรูปภาพ

หน้า

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

เนื่องจากความสำคัญของข้อมูลที่มีผลต่อการดำเนินธุรกิจในปัจจุบัน จึงทำให้มีความต้องการในการเก็บข้อมูลที่สูงมาก รวมถึงการใช้งานโซเชียลมีเดียที่เพิ่มขึ้น ไฟล์ภาพและวีดีโอที่มีขนาดใหญ่จากคุณภาพที่สูงขึ้น ดังนั้นจึงมีความต้องการพื้นที่ขนาดใหญ่ในการเก็บข้อมูลเพื่อรองรับกับขนาดและจำนวนของข้อมูล ระบบคลาวด์หรือการเก็บข้อมูลแบบออนไลน์จึงได้รับความนิยมสูงมาก และมีการใช้งานอย่างแพร่หลาย ฮาร์ดดิสก์ไดรฟ์เป็นอุปกรณ์สำคัญของการเก็บข้อมูลด้วยระบบคลาวด์ เพื่อตอบสนองการใช้งานสำหรับการเก็บข้อมูลที่มีขนาดใหญ่ขึ้น จำเป็นต้องมีการพัฒนาประสิทธิภาพของตัวฮาร์ดดิสก์ไดรฟ์ให้มีความสามารถในการเก็บข้อมูลที่เพิ่มขึ้น รวมถึงการพัฒนาความสามารถของชิ้นส่วนต่างๆ เพื่อให้รองรับประสิทธิภาพที่สูงขึ้น เช่น การเพิ่มประสิทธิภาพของแผ่นดิสก์และหัวอ่านหัวเขียน เพื่อให้ฮาร์ดดิสก์ 1 ตัวมีความสามารถในการเก็บข้อมูลได้มากขึ้น ในส่วนของการผลิตก็จำเป็นต้องมีการพัฒนากระบวนการเพื่อรองรับเทคโนโลยีใหม่ๆ จึงส่งผลให้กระบวนผลิตซับซ้อนมากขึ้น ใช้เวลาในการประกอบและทดสอบเพิ่มขึ้น ทำให้ต้นทุนในการผลิตสูงขึ้นตามไปด้วย

เทคนิคการเรียนรู้ด้วยเครื่องเป็นวิธีการที่นิยมใช้อย่างแพร่ในอุตสาหกรรมทุกวันนี้ ในการผลิตฮาร์ดดิสก์ไดรฟ์ก็มีความต้องการที่จะนำเทคนิคของการเรียนรู้ด้วยเครื่องมาใช้ในการปรับปรุงการผลิต เช่น การตรวจจับความผิดปกติที่เกิดขึ้นในกระบวนการ การตรวจจับงานเสียได้เร็วขึ้น การวิเคราะห์ข้อมูล และการลดเวลาในการผลิต ด้วยเช่นกัน แต่งานวิจัยที่มีการนำเทคนิคการเรียนรู้ด้วยเครื่องไปใช้ในอุตสาหกรรมการผลิตฮาร์ดดิสก์ไดรฟ์ยังมีจำนวนน้อย

ในงานวิจัยนี้จึงได้ทำการศึกษาปัญหาที่เกิดขึ้นจริงในการผลิตฮาร์ดดิสก์ไดรฟ์กับเทคนิคการเรียนรู้ด้วยเครื่อง เนื่องจากความต้องการความจุของฮาร์ดดิสก์ไดรฟ์ที่มีมากขึ้น จึงทำให้กระบวนการผลิต โดยเฉพาะกระบวนการทดสอบใช้เวลาในการทดสอบฮาร์ดดิสก์ไดรฟ์นานขึ้น รวมถึงงานเสียที่ถูกทดสอบในระบบก็ใช้เวลานานขึ้นด้วยเช่นกัน ซึ่งงานเสียเหล่านี้เป็นความสูญเสียในการผลิตทำให้ต้นทุนในการผลิตสูงขึ้น ดังนั้นการตรวจจับงานเสียและสามารถนำงานเสียออกจากกระบวนการผลิตได้เร็วขึ้น จะทำให้สามารถลดเวลาและค่าใช้จ่ายที่ไม่จำเป็นในการผลิตได้ งานวิจัยนี้ได้ทำการศึกษาปัญหางานเสียที่เกิดจากความผิดปกติของการอ่านสัญญาณเซอร์โวในกระบวนการทดสอบ โดยใช้

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ข้อมูลของกระบวนการผลิตที่เกิดขึ้นก่อนหน้ากระบวนการทดสอบ มาใช้ร่วมกับเทคนิคของ SVM และ XGBoost อัลกอริทึม เพื่อทำนายงานเสียที่อาจจะเกิดขึ้นขณะทำการทดสอบและแสดงผลการทำนายก่อนที่ตัวงานจะเข้าสู่กระบวนการทดสอบด้วยเครื่องจักร โดยปกติแล้วการตรวจจับงานเสียที่เกิดขึ้นจริงในกระบวนการผลิต มาจากการวัดค่าจริงในกระบวนการทดสอบเปรียบเทียบกับค่าควบคุม ซึ่งใช้เวลาในการทดสอบค่อนข้างนานก่อนที่จะตรวจพบงานเสีย จากข้อมูลงานเสียที่เกิดขึ้นจากการอ่านสัญญาณเซอร์โวมีเพียง 1% โดยเฉลี่ย เมื่อเทียบกับงานดีในกระบวนการผลิต ทำให้ข้อมูลที่นำมาใช้ในงานวิจัยมีลักษณะไม่สมดุล เนื่องจากความแตกต่างของจำนวนกลุ่มงานดีและงานเสียค่อนข้างสูง ดังนั้นในการทำให้โมเดลมีประสิทธิภาพที่ดีขึ้น จึงต้องมีการศึกษาและทดลองในส่วนของการจัดการข้อมูลที่ไม่สมดุล รวมถึงการวัดประสิทธิภาพด้วยวิธีการที่เหมาะสมจำเป็นต้องพิจารณาในงานวิจัยนี้ด้วยเช่นกัน

1.2 วัตถุประสงค์ของการวิจัย

เพื่อทำระบบตรวจจับงานเสียที่เกิดขึ้นจากปัญหาการอ่านสัญญาณเซอร์โวในกระบวนการทดสอบฮาร์ดดิสก์ไดรฟ์ โดยใช้ข้อมูลที่ได้จากกระบวนการประกอบและการเขียนสัญญาณเซอร์โว เพื่อให้ตรวจพบปัญหาก่อนที่จะนำตัวงานเข้าไปทดสอบจริงด้วยเครื่องจักร เพื่อให้สามารถนำงานเสียออกจากกระบวนการผลิตให้เร็วที่สุด ลดเวลาและเครื่องจักรที่ใช้ในการทดสอบงานเสีย เป็นการลดค่าใช้จ่ายที่ไม่จำเป็นของกระบวนการผลิต

1.3 ขอบเขตของงานวิจัย

1.3.1 ข้อมูลที่นำมาใช้ในงานวิจัยนี้เป็นข้อมูลจริงจากการผลิตฮาร์ดดิสก์ไดรฟ์ ชนิดที่มีจำนวนแผ่นดิสก์ 9 แผ่น หัวอ่านหัวเขียนจำนวน 18 หัว โดยทำการเก็บข้อมูลเป็นระยะเวลา 1 เดือน

1.3.2 ข้อมูลจะแบ่งเป็น 2 กลุ่มคือกลุ่มงานดีและงานเสีย งานดีคืองานที่ผ่านตั้งแต่กระบวนการประกอบจนถึงกระบวนการทดสอบ งานเสียคืองานที่พบความผิดปกติในการอ่านสัญญาณเซอร์โวในกระบวนการทดสอบ

1.3.3 งานวิจัยนี้จะใช้อัลกอริทึมการเรียนรู้ด้วยเครื่อง 2 วิธีการ คือ SVM และ XGBoost

1.3.4 งานวิจัยนี้จะมีการจัดการกับข้อมูลที่ไม่สมดุล 3 วิธีการ คือ SMOTE, Different Cost Learner และ SMOTE with Different Cost

1.3.5 งานวิจัยนี้มีการใช้เทคนิคในการเลือกคุณลักษณะเพื่อลดขนาดของข้อมูลเรียนรู้ และเพิ่มประสิทธิภาพของโมเดล

1.3.6 เป้าหมายของโมเดลที่คาดหวังคือ ค่า ROC AUC 90% และ PRC AUC 70%

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 สามารถทำนายงานเสียที่จะเกิดขึ้นในกระบวนการทดสอบ และตรวจพบความผิดปกติที่เกิดขึ้นในการผลิตได้เร็วขึ้น

1.4.2 ลดเวลาและเครื่องจักรที่ใช้ในการทดสอบงานเสีย

1.4.3 ลดค่าใช้จ่ายที่ไม่จำเป็นในการผลิตจากการทดสอบงานเสีย

1.4.4 สามารถนำวิธีการที่ได้จากงานวิจัยนี้ไปประยุกต์ใช้กับงานเสียที่เกิดขึ้นจากปัญหาอื่นๆ และฮาร์ดดิสก์ไดรฟ์ชนิดอื่นๆได้

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 กระบวนการผลิตฮาร์ดดิสก์ไดรฟ์ (Hard Disk Drive Manufacturing Process)

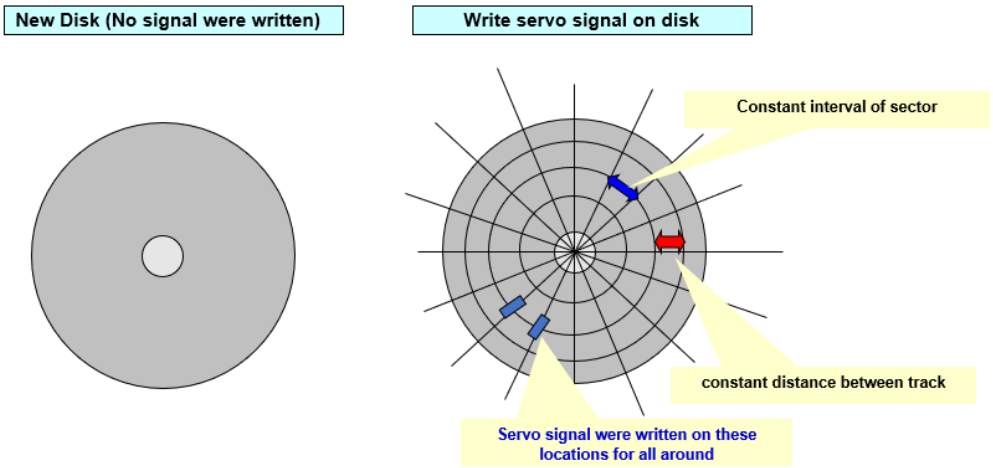
2.1.1 การประกอบชิ้นส่วน (Assembly Process) เป็นกระบวนการเริ่มต้นของการผลิตฮาร์ดดิสก์ไดรฟ์ ซึ่งเป็นกระบวนการที่ทำการประกอบชิ้นส่วนต่างๆ ของฮาร์ดดิสก์ไดรฟ์ เช่น เบส มอเตอร์ หัวอ่านหัวเขียน แผ่นดิสก์ เข้าด้วยกัน ชิ้นส่วนต่างๆ ของฮาร์ดดิสก์ไดรฟ์แสดงดังรูปที่ 1 เนื่องจากการประกอบชิ้นส่วนของฮาร์ดดิสก์ไดรฟ์ ต้องการความสะอาด ต้องมีการควบคุมอุณหภูมิและความชื้น เพื่อป้องกันสิ่งปนเปื้อนที่จะเข้าไปอยู่ในตัวฮาร์ดดิสก์ไดรฟ์ รวมถึงป้องกันการเกิดไฟฟ้าสถิตที่จะสร้างความเสียหายแก่ชิ้นส่วนภายในของฮาร์ดดิสก์ไดรฟ์ การประกอบชิ้นส่วนนี้ต้องทำในห้องสะอาด (Clean Room) หลังจากประกอบชิ้นส่วนและทำการปิดฝา (Top Cover) จึงจะนำตัวงานออกจากห้องสะอาดเพื่อส่งไปยังกระบวนการถัดไป

รูปที่ 1 *ส่วนประกอบของฮาร์ดดิสก์ไดรฟ์*

2.1.2 2.1.2 กระบวนการเขียนสัญญาเซอร์โว (Servo Track Writer Process) กระบวนการนี้เกิดขึ้นหลังจากทำการประกอบชิ้นส่วนฮาร์ดดิสก์ไดรฟ์เสร็จสมบูรณ์ และนำตัวงานออกจากห้องสะอาด โดยก่อนเริ่มทำการเขียนสัญญาณเซอร์โว (Servo Signal) จะนำตัวฮาร์ดดิสก์ไดรฟ์ไปบรรจุก๊าซฮีเลียม เพื่อช่วยในการรักษาเสถียรภาพในการหมุนของแผ่นดิสก์ และการเคลื่อนที่ของหัวอ่านหัวเขียน โดยความสำคัญของการเขียนสัญญาณเซอร์โว คือการเขียนสัญญาณเพื่อระบุตำแหน่งบนแผ่นดิสก์ เพื่อให้หัวอ่านหัวเขียน สามารถเคลื่อนที่ไปอ่านและเขียนข้อมูล ณ ตำแหน่ง ที่ต้องการ

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

บนแผ่นดิสก์ได้ โดยสัญญาณเซอร์โวที่เขียน จะมีการกำหนดคุณภาพของเส้นรอบวง (track) ให้มีระยะห่างที่เท่ากัน และมีการแบ่งจำนวนช่องหรือเซ็กเตอร์ (Sector) ให้มีระยะห่างเชิงมุมที่เท่ากัน รวมถึงมีการควบคุมขนาดของสัญญาณ ให้มีแอมพลิจูดอยู่ในค่าที่กำหนด เพื่อประสิทธิภาพในการเคลื่อนที่ของหัวอ่านไปยังตำแหน่งที่ต้องการได้ถูกต้อง และใช้เวลาน้อยที่สุด



รูปที่ 2 *การเขียนสัญญาณ Servo*

2.1.3 กระบวนการทดสอบ (Test Process) เป็นกระบวนการที่ทำต่อจากการเขียนสัญญาณเซอร์โว จะประกอบไปด้วย 4 กระบวนการหลักคือ

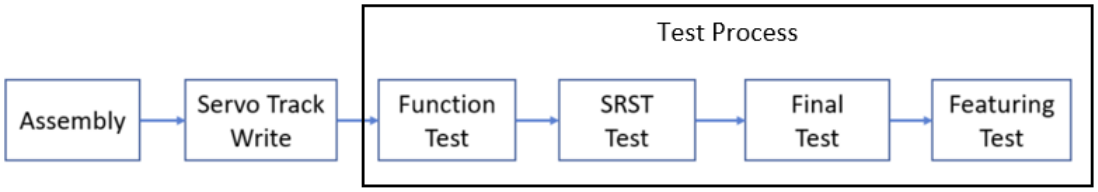
2.1.3.1 Function Test กระบวนการนี้เป็นกระบวนการแรกของการทดสอบ เริ่มจากทำการดาวน์โหลดไมโครโค้ดเพื่อให้ตัวฮาร์ดดิสก์ไดรฟ์สามารถเริ่มการทำงานได้ หลังจากนั้นจะทำการวัดและปรับค่าพารามิเตอร์ต่างๆ ที่เกี่ยวข้องกับการค้นหาสัญญาณและการเคลื่อนที่ของหัวอ่านหัวเขียน เพื่อให้ได้ค่าที่ดีที่สุดในการค้นหาสัญญาณขณะที่หัวอ่านเคลื่อนที่ไปบนแผ่นดิสก์ รวมถึงมีการปรับค่าพารามิเตอร์ที่เกี่ยวข้องกับการอ่านและเขียนสัญญาณ เพื่อให้ได้ประสิทธิภาพของการอ่านและเขียนสัญญาณที่ดีที่สุด

2.1.3.2 Self Run Stress Test (SRST) เป็นกระบวนการทดสอบต่อจาก Function Test กระบวนการทดสอบนี้จะทำการวัดและวิเคราะห์ลักษณะพื้นผิวของแผ่นดิสก์ จากนั้นทำการบันทึกค่าตำแหน่งที่ตรวจพบจุดบกพร่องของพื้นผิว กระบวนการทดสอบนี้จะทำงานที่อุณหภูมิสูง รวมถึงมีการวัดและการปรับค่าพารามิเตอร์ต่างๆ ให้ตรงกับลักษณะงานที่ลูกค้าจะนำไปใช้ เป็นกระบวนการทดสอบที่ใช้เวลาในการทดสอบนานที่สุด

2.1.3.3 Final Test สามารถเรียกได้ว่าเป็นการทดสอบประสิทธิผล จะทำการทดสอบประสิทธิภาพในการเขียนและอ่านสัญญาณบนพื้นผิวของแผ่นดิสก์ทั่วทั้งแผ่น

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

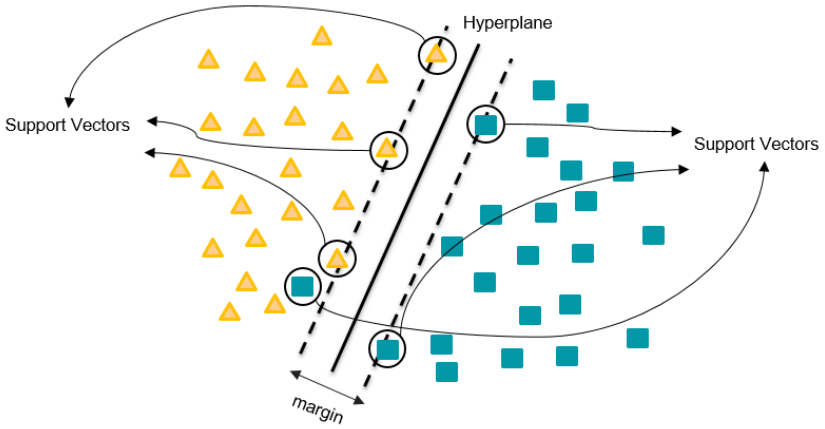
2.1.3.4 Featuring Test เป็นการทดสอบขั้นสุดท้ายก่อนส่งมอบ จะทำการปรับค่าพารามิเตอร์ต่างๆ เป็นครั้งสุดท้ายเพื่อให้ตรงตามที่ลูกค้ากำหนด



รูปที่ 3 *กระบวนการผลิตฮาร์ดดิสก์ไดรฟ์*

2.2 ทฤษฎีการเรียนรู้ด้วยเครื่อง (Machine Learning)

2.2.1 Support Vector Machine (SVM) เป็นอัลกอริทึม การเรียนรู้ด้วยเครื่องแบบมีผู้สอน (Supervised Learning) เป็นหนึ่งในอัลกอริทึมที่นิยมนำมาใช้กับปัญหาในการจำแนกกลุ่ม กับปัญหาที่มีจำนวนข้อมูลไม่เยอะมาก หลักการทำงานของ SVM จะทำการหาเส้นแบ่งหรือไฮเปอร์เพลน (Hyperplane) ที่สามารถแบ่งกลุ่มของข้อมูลออกจากกัน โดยจะทำการหาไฮเปอร์แพลนที่ใช้ในการแบ่งกลุ่มของข้อมูลที่ดีที่สุด โดยให้มีระยะห่าง (Margin) มากที่สุด ข้อมูลที่มีตำแหน่งใกล้กับไฮเปอร์เพลนมากที่สุดในแต่ละกลุ่มเรียกว่าซัพพอร์ตเวกเตอร์ และระยะห่าง (Margin) คือค่าที่วัดจากระยะของซัพพอร์ตเวกเตอร์ไปยังไฮเปอร์เพลน ดังรูปที่ 4 แสดงไฮเปอร์เพลนเส้นตรง ซึ่งเหมาะกับการจำแนกปัญหาที่มีข้อมูลเป็นเชิงเส้น

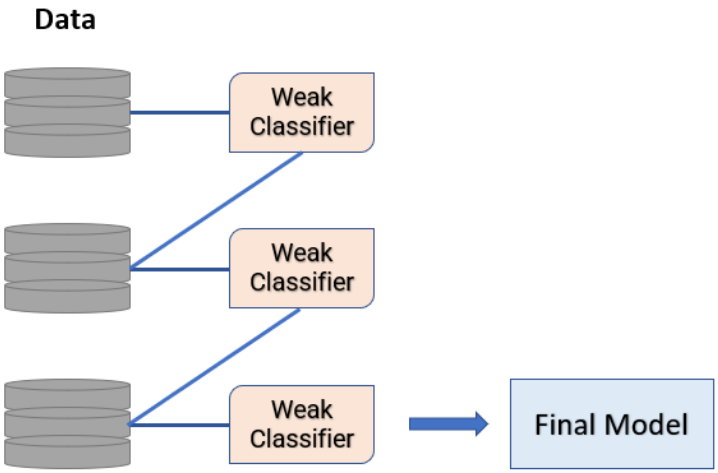


รูปที่ 4 *Support Vector Machine*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ในกรณีที่ข้อมูลมีลักษณะไม่เป็นเชิงเส้น สามารถทำการปรับค่าของ Kernel Function เพื่อให้สามารถสร้างไฮเปอร์เพลนที่มีลักษณะไม่เป็นเส้นตรงได้ โดย Kernel Function ที่นิยมใช้มีด้วยกัน 3 แบบคือ Polynomials , Radial Basis Function (RBF) และ Sigmoid

2.2.2 Extreme Gradient Boosting (XGBoost) เป็นวิธีการเรียนรู้แบบกลุ่ม (Ensemble Learning) โดยใช้เทคนิคการนำการเรียนรู้ต้นไม้ตัดสินใจหลายๆ โมเดลมาต่อกัน ทำการสร้างและเรียนรู้ โดยนำค่าผิดพลาดที่เกิดขึ้นจากโมเดลก่อนหน้ามาปรับปรุงในการสร้างโมเดลถัดไปเพื่อเพิ่มประสิทธิภาพ โมเดลจะถูกสร้างขึ้นจนกระทั่งไม่สามารถเพิ่มประสิทธิภาพของโมเดลได้อีก XGBoost เป็นอัลกอริทึม ที่พัฒนามาจาก Gradient Boosting โดยจะทำให้การเรียนรู้เร็วขึ้นและใช้ทรัพยากรของเครื่องน้อยลง XGBoost สามารถทำงานกับข้อมูลจำนวนมากกว่าสิบล้าน เป็นอัลกอริทึมที่นิยมใช้และได้ประสิทธิภาพสูงในปัจจุบัน [1]



รูปที่ 5 *Boosting*

2.3 วิธีการเลือกคุณลักษณะ (Feature Selection Method)

การเลือกคุณลักษณะเป็นหนึ่งในวิธีการที่ช่วยเพิ่มประสิทธิภาพของการเรียนรู้ด้วยเครื่อง โดยการตัดคุณลักษณะที่ไม่จำเป็นและซ้ำซ้อนออก [2] ทำการเลือกเฉพาะคุณลักษณะที่มีความสำคัญต่อการเรียนรู้ เทคนิคในการเลือกคุณลักษณะ สามารถแบ่งเป็น 3 วิธีการหลักๆ คือ

2.3.1 Filter Method เป็นการตัดคุณลักษณะที่ไม่จำเป็นออก เช่น คุณลักษณะที่เป็นค่าคงที่ คุณลักษณะที่เป็นค่าเฉพาะ รวมถึงการใช้ค่าทางสถิติมาช่วยเลือกคุณลักษณะโดยทำการเรียงลำดับความสำคัญ เช่น Chi-square, ANOVA และ ค่าสัมประสิทธิสหสัมพันธ์ (Correlation Coefficient)

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

2.3.2 Wrapper Method เป็นการเลือกกลุ่มย่อยของคุณลักษณะที่ให้ประสิทธิภาพของโมเดลการเรียนรู้ด้วยเครื่องสูงที่สุด โดยการเลือกคุณลักษณะด้วยวิธีการนี้จะทำการสร้างโมเดลการเรียนรู้ด้วยเครื่องอย่างง่ายมาช่วยในการเลือกคุณลักษณะ เช่น Forward Selection เป็นการเพิ่มคุณลักษณะเข้าไปในโมเดลที่ละตัวโดยจะเลือกคุณลักษณะที่ให้ประสิทธิภาพสูงสุดในแต่ละรอบ ทำซ้ำจนกระทั่งได้ประสิทธิภาพของโมเดลดีที่สุด, Backward Elimination เป็นการตัดคุณลักษณะออกจากโมเดลทีละตัว โดยเลือกจากคุณลักษณะที่ตัดออกแล้วทำให้ประสิทธิภาพของโมเดลเพิ่มขึ้น ทำจนกระทั่งเมื่อตัดคุณลักษณะออกแล้วประสิทธิภาพของโมเดลไม่เพิ่มขึ้น และ Recursive Feature Elimination ทำการเพิ่มและตัดคุณลักษณะไปพร้อมกัน ซึ่งวิธีการเลือกคุณลักษณะแบบ Wrapper นี้จะใช้เวลาและทรัพยากรในการคำนวณและเลือกคุณลักษณะสูง

2.3.3 Embedded Method เป็นวิธีการที่ฝังอยู่ในอัลกอริทึมการเรียนรู้ด้วยเครื่อง เป็นการรวมกันของวิธีการ Filter และ Wrapper เช่น Ridge และ Lasso เป็นวิธีการที่ฝังอยู่ในการสร้างโมเดลการเรียนรู้ด้วยเครื่อง โดยทั้งสองวิธีการนี้จะทำการตัดคุณลักษณะที่มีความสำคัญน้อยและไม่จำเป็นสำหรับการสร้างโมเดลการเรียนรู้ด้วยเครื่อง การเลือกคุณลักษณะด้วยอัลกอริทึมที่พัฒนามาจากต้นไม้ตัดสินใจเช่น Random Forest และ Light Gradient Boost จะทำการเลือกคุณลักษณะจากการเรียงลำดับความสำคัญของคุณลักษณะที่มีต่อการสร้างโมเดล โดยจะทำการเลือกคุณลักษณะจะพิจารณาจากการเรียนรู้ของต้นไม้หลายๆต้น หรือหลายๆโมเดล เพื่อลดค่าแปรปรวนของความสำคัญของคุณลักษณะ

2.4 การจัดการข้อมูลที่ไม่สมดุล (Imbalance Data Handling)

ข้อมูลที่ไม่สมดุล คือชุดข้อมูลที่มีจำนวนของข้อมูลในแต่ละกลุ่ม (Class) แตกต่างกันมาก โดยธรรมชาติของชุดข้อมูล กลุ่มของข้อมูลที่เราให้ความสนใจจะมีจำนวนน้อยกว่ากลุ่มของข้อมูลที่ไม่ได้สนใจ กลุ่มของข้อมูลที่มีจำนวนน้อยกว่าจะเรียกว่า คลาสรอง (Minority Class) และกลุ่มของข้อมูลที่มีจำนวนเยอะกว่าเรียกว่า คลาสหลัก (Majority Class) ตัวอย่างของชุดข้อมูลที่ไม่สมดุลได้แก่ ข้อมูลของผู้ป่วยโรคมะเร็ง ข้อมูลของความผิดปกติของบัตรเครดิต รวมถึงข้อมูลของงานเสียที่เกิดขึ้นในการผลิตฮาร์ดดิสก์ไดรฟ์

วิธีการจัดการกับข้อมูลที่ไม่สมดุล จะแบ่งออกเป็น 2 ส่วนใหญ่ๆ คือ

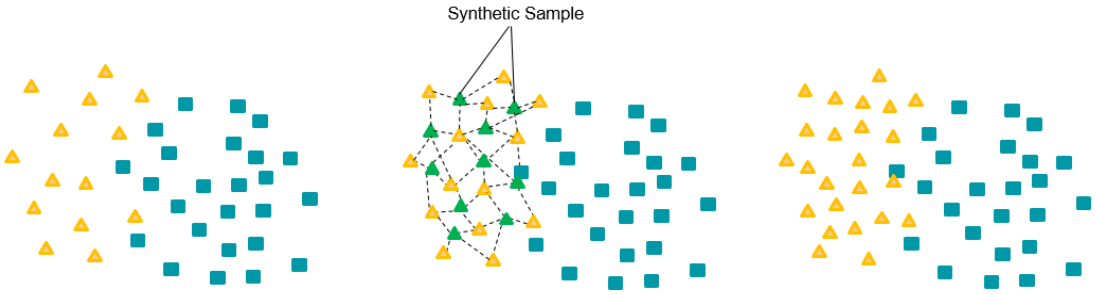
2.4.1 การจัดการในส่วนของข้อมูล (Data-Level Approach)

2.4.1.1 การปรับลดจำนวนข้อมูลคลาสหลักด้วยการสุ่ม (Random Under-sampling) และการปรับเพิ่มจำนวนคลาสรองด้วยวิธีการสุ่ม (Random Over-sampling)

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

เป็นการปรับสมดุลของข้อมูลด้วยการสุ่มลดจำนวนของคลาสหลัก และเพิ่มจำนวนของคลาสรองเพื่อให้จำนวนของข้อมูลทั้งสองกลุ่มมีจำนวนใกล้เคียงกัน

2.4.1.2 SMOTE (Synthetic Minority Over-Sampling Technique) เป็นหนึ่งในวิธีการจัดการกับข้อมูลที่ไม่สมดุล ด้วยการเพิ่มจำนวนคลาสรอง โดยใช้ข้อมูลจริงในการสร้างข้อมูลสังเคราะห์ของคลาสรองขึ้นมา โดยการสุ่มเลือกข้อมูลที่อยู่ในคลาสรองขึ้นมา 1 ค่า และพิจารณาข้อมูลที่อยู่ใกล้เคียงจำนวน K ตัว (K-Nearest Neighbor) แล้วคำนวณหา Euclidean Distance ระหว่างข้อมูลที่สุ่มเลือกกับข้อมูลใกล้เคียงจำนวน K ตัว แล้วทำการสร้างข้อมูลสังเคราะห์ขึ้นมาให้อยู่ในระยะของ Euclidean Distance ของข้อมูลที่สุ่มเลือกและข้อมูลใกล้เคียงจำนวน K ตัว ดังรูปที่ 6



รูปที่ 6 *Synthetic Minority Over-sampling Technique (SMOTE)*

2.4.2 การจัดการในส่วนของอัลกอริทึม

ในงานวิจัยนี้ จะใช้ วิธีการของ Different Cost Learner เป็นการทำงานในส่วนของอัลกอริทึมการเรียนรู้ โดยจะกำหนดค่าใช้จ่ายของความผิดพลาด (Cost) ที่เกิดขึ้นในการเรียนรู้ของแต่ละคลาสแตกต่างกัน โดยค่าใช้จ่ายที่เกิดจากความผิดพลาดในการทำนายคลาสรองจะมีค่าสูงกว่าความผิดพลาดที่เกิดขึ้นจากการทำนายของคลาสหลัก ซึ่งวิธีการนี้จะให้ความสำคัญในการทำนายของกลุ่มคลาสรองมากกว่าคลาสหลัก เนื่องจากในการเรียนรู้ของเครื่องต้องทำการสร้างโมเดลที่มีค่าใช้จ่ายของความผิดพลาดในการทำนายที่น้อยที่สุด

2.5 การวัดประสิทธิภาพ (Performance Measurement)

ในการวัดประสิทธิภาพของโมเดลสำหรับการแบ่งกลุ่มของข้อมูลจำนวน 2 กลุ่ม โดยเป็นข้อมูลที่เป็นกลุ่มบวกและกลุ่มลบ มีการวัดค่าประสิทธิภาพของโมเดลได้ดังนี้

Confusion Matrix คือตารางประเมินผลลัพธ์ของการทำนายของโมเดล

True Positive (TP) ทำนายว่าเป็นบวก และ ค่าจริงเป็นบวก

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

True Negative (TN) ทำนายว่าเป็นลบ และ ค่าจริงเป็นลบ

False Positive (FP) ทำนายว่าเป็นบวก แต่ค่าจริงเป็นลบ

False Negative (FN) ทำนายว่าเป็นลบ แต่ค่าจริงเป็นบวก

Accuracy คือค่าที่บอกความถูกต้องในการทำนายของโมเดล หาได้จากอัตราส่วนของค่าที่โมเดลทายถูกต่อค่าทั้งหมด

$$\frac{(TP + TN)}{(TP + TN + FP + FN)}$$

True Positive Rate (TPR) หรือ Recall คืออัตราส่วนของกลุ่มของงานบวกที่โมเดลทายถูกต่อกลุ่มของงานบวกทั้งหมด

$$\frac{TP}{(TP + FN)}$$

False Negative Rate (FNR) คืออัตราส่วนการทำนายผิดของกลุ่มของงานลบ ต่อกลุ่มของงานลบทั้งหมด

$$\frac{FP}{(TN + FP)}$$

Precision คือ ค่าที่บอกความถูกต้องของการทำนายค่าบวกของโมเดล

$$\frac{TP}{(TP + FP)}$$

F1-score คือ ค่าเฉลี่ย Harmonic ของ Precision และ Recall

$$\frac{2\ x\ (Precision\ x\ Recall)}{(Precision + Recall)}$$

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

Receiver Operating Characteristic (ROC) Curves คือ กราฟความสัมพันธ์ระหว่าง True Positive Rate และ False Positive Rate ที่เกิดจากการปรับค่าเกณฑ์ของความน่าจะเป็นในการทำนายที่ต่างกัน

ROC AUC คือ การวัดพื้นที่ใต้กราฟ ROC โดยค่า ROC AUC ที่เข้าใกล้ 1 แสดงว่าโมเดลมีประสิทธิภาพดี

Precision-Recall (PRC) Curves คือกราฟความสัมพันธ์ระหว่าง Precision และ Recall ที่ปรับค่าเกณฑ์ของความน่าจะเป็นในการทำนายที่ต่างกัน

PRC AUC คือ การวัดพื้นที่ใต้กราฟ PRC โดยค่า PRC AUC ที่เข้าใกล้ 1 แสดงว่าโมเดลมีประสิทธิภาพดี

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

บทที่ 3
งานวิจัยที่เกี่ยวข้อง

มีงานวิจัยหลายๆตัวที่ใช้เทคนิคการเรียนรู้ด้วยเครื่องในการตรวจจับความเสียหายที่เกิดขึ้นกับตัวฮาร์ดดิสก์ไดรฟ์ในการใช้งานทั่วไป แต่งานวิจัยของเทคนิคการเรียนรู้ด้วยเครื่องที่ศึกษาเกี่ยวกับกระบวนการผลิตฮาร์ดดิสก์ไดรฟ์ยังคงไม่แพร่หลาย เมื่อเทียบกับงานวิจัยที่ใช้เทคนิคการเรียนรู้ด้วยเครื่องกับปัญหาการตรวจจับตัวฮาร์ดดิสก์ไดรฟ์ที่เสียจากการใช้งาน

3.1 งานวิจัยที่ใช้เทคนิคการเรียนรู้ด้วยเครื่องในกระบวนการผลิตฮาร์ดดิสก์ไดรฟ์

งานวิจัย [3] นำเสนอวิธีการใหม่ในการปรับปรุงทำนายผลผลิต (Yield) ของการผลิตฮาร์ดดิสก์ไดรฟ์ ด้วยการวิเคราะห์ทางสถิติ ร่วมกับ MLR, ANN และ CART อัลกอริทึม ในการทำนายมีการนำเสนอการจัดกลุ่มจำนวนข้อมูลด้วยวิธีการใหม่ ซึ่งวิธีการจัดกลุ่มข้อมูลใหม่ร่วมกับ MLR และ CART ให้ประสิทธิภาพดีที่สุด MAE = 0.01

SVM เป็นอัลกอริทึมการเรียนรู้ด้วยเครื่องที่นิยมนำมาใช้ในการจำแนกปัญหาสำหรับชุดข้อมูลที่มี 2 กลุ่ม ในงานวิจัย [4] ใช้ SVM ในการตรวจจับงานเสียที่เกิดจากการประกอบฮาร์ดดิสก์ไดรฟ์ โดยใช้กระแสของมอเตอร์ขดลวด (Voice Coil Motor) ในการเรียนรู้ของโมเดล ชุดข้อมูลมีลักษณะไม่สมดุล งานเสียมีเพียงแค่ 3% ซึ่งการใช้ SVM อัลกอริทึมสามารถจำแนกงานดีงานเสียได้ความถูกต้อง 100%

เปรียบเทียบกับในงานวิจัยฉบับนี้ ที่ศึกษาปัญหาของงานเสียที่เกิดขึ้นเนื่องจากความผิดพลาดที่เกิดขึ้นจากการอ่านสัญญาณเซอร์โว ชุดของข้อมูลที่ใช้คือค่าพารามิเตอร์ ในกระบวนการประกอบชิ้นส่วนและการเขียนสัญญาณเซอร์โว สามารถเปรียบเทียบได้ดังตารางที่ 1

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ตารางที่ 1 เปรียบเทียบงานวิจัย

หัวข้อ	งานวิจัย [4]	งานวิจัยฉบับนี้
ปัญหา	ปัญหาของการประกอบชิ้นส่วน เช่น ชิ้นส่วนไม่ครบ หรือตำแหน่งในการประกอบมีความคลาดเคลื่อน	ปัญหาที่เกี่ยวข้องกับการอ่านสัญญาณเซอร์โวที่เกิดขึ้นในกระบวนการทดสอบ ซึ่งปัญหาอาจจะเกิดจาก หัวอ่านหัวเขียน สิ่งปนเปื้อนในตัวฮาร์ดดิสก์ไดรฟ์ และความผิดปกติที่เกิดขึ้นขณะเขียนสัญญาณเซอร์โว
ชุดข้อมูลที่ใช้ในการวิเคราะห์	กระแสของมอเตอร์ขดลวด (VCM) ขณะที่หัวอ่านเคลื่อนที่ลงไปยังแผ่นดิสก์	ค่าพารามิเตอร์ที่วัดตอนประกอบชิ้นส่วนและเขียนสัญญาณเซอร์โว เช่น ระยะห่างระหว่างหัวอ่านกับแผ่นดิสก์, ค่าความต้านทานของหัวอ่าน, ตำแหน่งของสัญญาณที่ผิดพลาด, กระแสของมอเตอร์, พารามิเตอร์ที่บอกคุณภาพของสัญญาณ
อัลกอริทึม	SVM	SVM, XGBoost
การจัดการข้อมูลที่ไม่สมดุล	ไม่มี	การสุ่มลดจำนวน SMOTE Different Cost Learner

3.2 งานวิจัยการตรวจจับฮาร์ดดิสก์ไดรฟ์เสียในการใช้งาน

มีงานวิจัยหลายฉบับที่เกี่ยวข้องการทำนายความผิดปกติของการใช้งานตัวฮาร์ดดิสก์ไดรฟ์ งานวิจัย [5] ใช้ข้อมูลสาธารณะของ Operation จาก Blackblaze ซึ่งชุดข้อมูลมีความไม่สมดุลสูงมากในอัตราส่วน 5000:1 มีการศึกษาในอัลกอริทึมของ SVM, RF และ GBT กับ SMOTE ในการทำนายของฮาร์ดดิสก์ไดรฟ์เสียด้วย SMART Parameter โดย RF และ GBT อัลกอริทึมให้ประสิทธิภาพที่ดีที่สุด งานวิจัย [6] ทำการศึกษาและใช้ 13 SMART Parameter ที่มีนัยยะสำคัญสูงสุดของของโมเดล

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

การตรวจจับความผิดปกติเพื่อทำนายฮาร์ดดิสก์ไดรฟ์เสีย โดยสามารถตรวจจับฮาร์ดดิสก์ไดรฟ์เสียได้ Accuracy 96.11% งานวิจัย [7, 8] เป็นงานวิจัยที่ศึกษาการทำนายฮาร์ดดิสก์ไดรฟ์เสียด้วย SMART Parameter เช่นกัน ซึ่งงานวิจัยเหล่านี้เป็นงานวิจัยที่ศึกษาปัญหาของฮาร์ดดิสก์ไดรฟ์เสียจากการใช้งาน ส่วนในงานวิจัยฉบับนี้ทำการศึกษาปัญหาของฮาร์ดดิสก์ไดรฟ์ในกระบวนการผลิต

3.3 งานวิจัยของ SVM กับข้อมูลที่ไม่สมดุล

งานวิจัยที่มีการใช้ SVM อัลกอริทึมสำหรับข้อมูลที่มีความไม่สมดุลสูง ได้แก่ งานวิจัย [9-11] มีการศึกษา SVM กับ SMOTE เทคนิคในการจัดการงานไม่สมดุล ใน [9] มีการประยุกต์ใช้ SVM กับเทคนิคของ SMOTE ร่วมกับ Different Cost เรียกว่า SDC (SMOTE with Different Cost) โดยวิธีการ SDC ให้ประสิทธิภาพที่ดีที่สุด เปรียบเทียบกับ SVM และ SVM SMOTE ซึ่งในงานวิจัยฉบับนี้มีการศึกษาเรื่องของ SVM SMOTE และ SVM SDC ด้วยเช่นกัน

งานวิจัย [10] ใช้ SVM ร่วมกับวิธีการต่างๆในการจัดการกับข้อมูลที่ไม่สมดุล มีการศึกษา SVM-Weight ซึ่งเป็นการเรียนรู้ด้วยค่าใช้จ่ายที่ต่างกัน SVM-SMOTE ในการเพิ่มจำนวนของกลุ่มข้อมูลรอง SVM-RANDU การสุ่มลดจำนวน และ GSVM-RU ซึ่งเป็นการสุ่มลดจำนวนเฉพาะข้อมูลที่ไม่ได้เป็นซัพพอร์ตเวกเตอร์ ทำการเปรียบเทียบประสิทธิภาพ โดย GSVM-RU ให้ประสิทธิภาพที่ดีที่สุด งานวิจัย [12] ใช้ SVM ในชุดข้อมูลที่ไม่สมดุล โดยการประยุกต์ใช้ Active Learning เพื่อลดขนาดของชุดข้อมูลเรียนรู้ และได้ประสิทธิภาพดีที่สุด งานวิจัย [11] มีการศึกษาอัลกอริทึม C50,KNN,NN,RF และ SVM กับการจัดการข้อมูลที่ไม่สมดุลด้วย การสุ่มเพิ่มจำนวน (ROSE) และ SMOTE ในการเพิ่มจำนวนข้อมูลกลุ่มรอง ด้วยอัตราส่วนที่แตกต่างกัน โดยวิธีการ SMOTE ด้วยอัตราส่วนในการเพิ่มจำนวนที่ 1:3 ให้ประสิทธิภาพดีที่สุดสำหรับทุกอัลกอริทึม งานวิจัย [13] นำเสนอวิธีการ fuzzy SVM ในการเรียนรู้สำหรับชุดข้อมูลที่ไม่สมดุล และยังมีงานวิจัยอีกหลายๆ ตัวที่ใช้ SVM กับปัญหาของชุดข้อมูลที่ไม่สมดุล เช่น งานวิจัย [14, 15]

3.4 งานวิจัยที่ศึกษาเทคนิคการจัดการความไม่สมดุล และการเลือกคุณลักษณะ

งานวิจัย [16] นำเสนอวิธีการของ SMOTE ซึ่งเป็นวิธีการที่นิยมใช้ในการสุ่มเพิ่มจำนวนคลาสรอง งานวิจัย [17] นำเสนอวิธีการ Borderline SMOTE ซึ่งเป็นวิธีการที่พัฒนามาจาก SMOTE โดยจะทำการสุ่มเพิ่มจำนวนคลาสรองเฉพาะตรงขอบของข้อมูลคลาสรองเท่านั้น งานวิจัย [18] ศึกษาการเพิ่มจำนวนข้อมูลคลาสรองด้วยการพิจารณาคลาสหลัก (SWIM) และสร้างข้อมูลคลาสรองที่อยู่บนระยะทาง Mahalanobis เดียวกันกับคลาสหลัก งานวิจัย [19] ศึกษาคุณลักษณะที่ป้อนเข้าไปยังโมเดล ซึ่งคุณลักษณะที่ป้อนเข้าไปยังโมเดลส่งผลกับความสำคัญและประสิทธิภาพของโมเดล งานวิจัย [20]

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ใช้ SSVM-FS ในการเลือกคุณลักษณะ ซึ่งวิธีการนี้มีการให้ความสนใจที่ความไม่สมดุลของคลาส และค่าน้ำหนักของ SVM เพื่อระบุความสำคัญของคุณลักษณะ

3.5 งานวิจัยที่มีการศึกษาอัลกอริทึม XGBoost

งานวิจัย [21] ศึกษา XGBoost อัลกอริทึม กับการทำนายค่า Transient Stability ของ Power System ซึ่งได้ค่า Accuracy สูงมาก และใช้เวลาในการคำนวณค่อนข้างน้อย งานวิจัย [22] ศึกษา XGBoost อัลกอริทึมกับการจัดกลุ่ม Network Traffic เทียบกับอัลกอริทึม Naïve Bay, KNN และ Tree-Base โดย XGBoost อัลกอริทึมให้ประสิทธิภาพที่ดีที่สุดที่ 99.5% Accuracy ซึ่ง XGBoost เป็นอัลกอริทึมที่ให้ประสิทธิภาพสูงในการทดลองกับหลายชุดข้อมูล อย่างไรก็ตามยังไม่มีงานวิจัยที่นำ XGBoost อัลกอริทึมมาใช้กับปัญหาของฮาร์ดดิสก์ไดรฟ์ ดังนั้นในงานวิจัยฉบับนี้จึงนำ XGBoost มาประยุกต์ใช้กับชุดข้อมูลของการผลิตฮาร์ดดิสก์ไดรฟ์

3.6 งานวิจัยที่ศึกษาการวัดประสิทธิภาพของอัลกอริทึมกับชุดข้อมูลที่ไม่สมดุล

การใช้ Accuracy ในการวัดประสิทธิภาพของอัลกอริทึมการเรียนรู้ด้วยเครื่องกับชุดข้อมูลที่ไม่สมดุลนั้นไม่เหมาะสม เนื่องจากประสิทธิภาพของโมเดลยังให้ค่าที่ดีแม้ว่าจะไม่สามารถทำนายข้อมูลออกมาเป็นกลุ่มคลาสรองได้ ก็ยังให้ค่าของ Accuracy ที่ยังคงสูงอยู่ จาก [23] การวัดประสิทธิภาพด้วย AUC ไม่มีผลกระทบจากอัตราส่วนความไม่สมดุลของคลาส (Skew) ในขณะที่ Accuracy, F1-score มีผลกระทบจาก Skew ของชุดข้อมูล

การวัดประสิทธิภาพของโมเดลด้วย ROC สามารถบอกได้แค่ภาพรวมของประสิทธิภาพเท่านั้น ในขณะที่การวัดประสิทธิภาพด้วย PRC กราฟ สามารถบอกความละเอียดของประสิทธิภาพในการทำนายได้ดีกว่า [24] ดังนั้นในงานวิจัยฉบับนี้จึงมีการใช้ค่า ROC AUC และ PRC AUC ในการวัดประสิทธิภาพของโมเดลด้วยเช่นกัน

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

‘

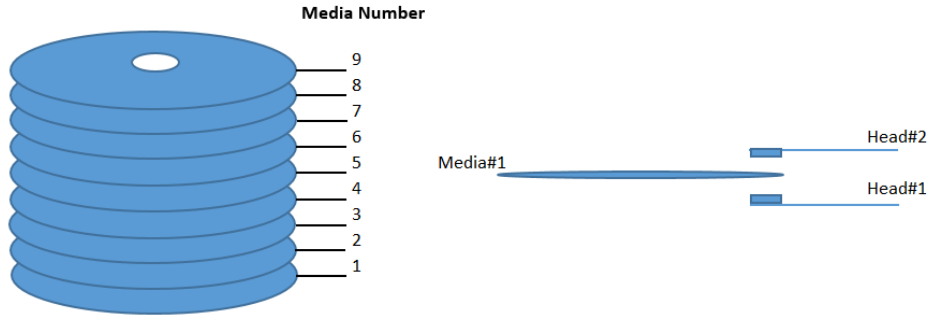
บทที่ 4
วิธีดำเนินการวิจัย

4.1 ศึกษาและเก็บรวบรวมข้อมูล

ทำการศึกษาและรวบรวมข้อมูลของปัญหาที่เกิดขึ้นในกระบวนการผลิตฮาร์ดดิสก์ไดรฟ์ นำปัญหาความผิดปกติในการการอ่านสัญญาณเซอร์โวในกระบวนการทดสอบมาใช้ในงานวิจัย เพื่อศึกษาและหาวิธีการในการตรวจจับปัญหาให้เร็วขึ้น

4.2 เก็บข้อมูลสำหรับงานวิจัย (Data Collection)

ในงานวิจัยนี้ มีการเก็บข้อมูลจากฮาร์ดดิสก์ไดรฟ์ชนิดที่มี 9 แผ่นดิสก์ 18 หัว โดยแผ่นดิสก์ 1 แผ่นจะถูกประกอบเข้ากับหัวอ่านหัวเขียนจำนวน 2 หัว คือ 1 หัวอยู่ด้านบน และอีก 1 หัว ที่ด้านล่างของแผ่นดิสก์ ดังรูปที่ 7 ดังนั้นในการเก็บค่าพารามิเตอร์ของฮาร์ดดิสก์ 1 ตัว จะมีการเก็บค่าของหัวอ่านทุกหัว ดังนั้นข้อมูลของฮาร์ดดิสก์ไดรฟ์แต่ละตัวจะประกอบด้วย ข้อมูลจำนวน 18 แถว



รูปที่ 7 *การประกอบหัวอ่านหัวเขียนกับแผ่นดิสก์*

ซึ่งชุดข้อมูลที่นำมาใช้ในงานวิจัยเก็บมาจากข้อมูลจริงของการผลิตฮาร์ดดิสก์ไดรฟ์เป็นระยะเวลา 1 เดือน โดยการเก็บข้อมูลจะแบ่งเป็น 2 ส่วน ส่วนแรกทำการเก็บจากพารามิเตอร์ที่วัดในการประกอบชิ้นส่วน และกระบวนการเขียนสัญญาณเซอร์โว เช่น

4.2.1 กระแสมอเตอร์ คือค่ากระแสที่วัดจากมอเตอร์ขณะหมุนด้วยภาระต่างกัน เช่น กระแสของมอเตอร์ที่หมุนแบบไม่มีภาระ กระแสมอเตอร์ขณะที่ทำการเคลื่อนที่ของหัวอ่านลงบนแผ่นดิสก์ไปยังตำแหน่งต่างๆ กระแสของมอเตอร์ขณะที่ทำการอ่านและเขียนสัญญาณบนแผ่นดิสก์

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

4.2.2 ความต้านทานของหัวอ่านหัวเขียน จะเป็นค่าความต้านทานที่วัดจากหัวอ่านหัวเขียนก่อนตอนเริ่มต้นกระบวนการเขียนสัญญาณ และหลังจากการเขียนสัญญาณเสร็จสิ้น

4.2.3 คุณภาพของสัญญาณเซอร์โว (Servo Signal Quality) ได้แก่ ค่าแอมพลิจูดของสัญญาณ ค่าเฟสของสัญญาณ ค่าการเรียงตัวของสัญญาณ เป็นต้น

4.2.4 ค่าของตำแหน่งที่ผิดปกติ (Positioning Error Signal) ค่าความถี่ของตำแหน่งที่ผิดปกติแบบเกิดซ้ำ (Repeatable Run Out: RRO) และความถี่ของตำแหน่งที่ผิดปกติแบบเกิดไม่ซ้ำ (Non-Repeatable Run Out: NRO)

4.2.5 ระยะห่างระหว่างหัวอ่านและแผ่นดิสก์ (Fly Height) เป็นค่าที่ทำการวัดและปรับให้ได้ระยะห่างในค่าที่กำหนด เพื่อให้แอมพลิจูดของสัญญาณคงที่ขณะทำการอ่านและเขียน

ข้อมูลทั้งหมดจะมีการวัดและเก็บค่าในแต่ละหัวที่ประกอบเข้ากับแผ่นดิสก์ โดยข้อมูลของหัวอ่าน 1 หัวจะทำการวัดและเก็บข้อมูลจากหลายตำแหน่งที่แตกต่างกันบนแผ่นดิสก์ โดยค่าพารามิเตอร์ที่เก็บมาได้ทั้งหมดจะมี 359 พารามิเตอร์ ดังตารางที่ 2

ส่วนที่ 2 จะเป็นการเก็บข้อมูลจากกระบวนการทดสอบซึ่งจะเป็นค่าที่ใช้ในการบอกว่าเป็นงานดี (Passer) หรืองานเสีย (Failure) จากการอ่านสัญญาณซอร์โว ซึ่งอัตราส่วนของงานเสีย มีเพียง 1% ของงานดี ซึ่งในงาน 1 ตัว มี 18 หัวอ่านหัวเขียน ในงานแต่ละตัวมี 1 หัวที่เสีย อัตราส่วนของหัวอ่านหัวเขียนที่เสีย จะเป็น 0.06% เท่านั้น ดังนั้นอัตราส่วนของงานเสียต่องานดีคือ 1:1,667 ซึ่งจะเห็นได้ว่าความไม่สมดุลของข้อมูลอยู่ในอัตราที่สูงมาก

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ตารางที่ 2 ข้อมูลที่ใช้ในงานวิจัย

Input parameter		Target output	
Motor current	p1 - p3	Failure	Class 1
Spindle motor parameter	p4 - p10	Passer	Class 0
Head resistance	p11 - p15		
Servo signal quality	p16 - p44, p154 - p182		
Fly height	p45 - p106		
Positioning error signal	p107 - p119		
RRO/NRO	p120 - p127.		
Process time	p128 - p130		
Flag count	p131 - p133		
Temperature	p134 - p139		
Other Configuration parameters	p140 – p153, p183 – p221		
ID, Unique parameters, Date/time, etc.	p222 – p359		

ทำการเก็บข้อมูลโดยใช้ข้อมูลของงานเสียที่เกิดขึ้นทุกๆ หัวอ่านหัวเขียนเป็นระยะเวลา 1 เดือนและเก็บข้อมูลของงานดีด้วยการสุ่มข้อมูลของงานดีจำนวน 1% จากข้อมูลของงานดีทั้งหมดด้วยวิธีการ Bernoulli เพื่อให้การกระจายของข้อมูลงานดีสามารถเป็นตัวแทนของข้อมูลงานดีทั้งหมดได้

4.3 การเตรียมชุดข้อมูล และจัดการชุดข้อมูลเบื้องต้น (Data Preprocessing)

4.3.1 ชุดข้อมูลที่นำมาใช้ในงานวิจัยประกอบไปด้วยข้อมูลของฮาร์ดดิสก์ทั้งหมด 84,221 ตัว แบ่งเป็นงานดี 79,344 ตัว และงานเสีย 4,877 ตัว ซึ่งงานเสียเป็นกลุ่มของงานที่ต้องการตรวจจับและมีจำนวนน้อย ดังนั้นจึงทำการกำหนดคลาสของงานเสียเป็น 1 และงานดีเป็น 0

4.3.2 ทำการแบ่งชุดข้อมูลเพื่อใช้ในการเรียนรู้และการทดสอบ โดยแบ่งเป็น 70% ของข้อมูลใช้ในการเรียนรู้ของโมเดล และ 30% เพื่อใช้ในการทดสอบประสิทธิภาพของโมเดล โดยจะได้จำนวนข้อมูลตามตารางที่ 3

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ตารางที่ 3 จำนวนข้อมูลที่ใช้ในงานวิจัย

	Input Data	Train Data	Test Data	Unit
Total Data	84,221	58,954	25,267	HDD
Passer	79,344	55,556	23,788	HDD
Failure	4,877	3,398	1,479	HDD

4.3.3 จัดการกับข้อมูลที่หายไปด้วยการตัดค่าของพารามิเตอร์ที่มีอัตราการหายไปของข้อมูลเกินกว่า 40% ออก และแทนค่าของข้อมูลที่หายไปในพารามิเตอร์แต่ละตัวที่เหลือด้วยค่าเฉลี่ยของพารามิเตอร์แต่ละตัว

4.3.4 ทำให้ค่าของพารามิเตอร์แต่ละตัวให้เป็นมาตรฐาน (Standardize) ด้วยวิธีคะแนนมาตรฐาน (Z-score) เพื่อปรับค่าเฉลี่ยของพารามิเตอร์แต่ละตัวให้อยู่ในช่วงเดียวกัน โดยมีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนให้อยู่ในช่วงค่า 1

4.4 เลือกคุณลักษณะ (Feature Selection)

ชุดข้อมูลที่ทำการเก็บมาจากข้อ 4.2 ประกอบไปด้วยพารามิเตอร์หรือคุณลักษณะทั้งหมด 359 พารามิเตอร์ ซึ่งเป็นขนาดของข้อมูลที่มีจำนวนมิติขนาดใหญ่ ในการใช้ข้อมูลที่มีมิติของข้อมูลขนาดใหญ่ในการเรียนรู้ของเครื่องทำให้เสียเวลาและสิ้นเปลืองทรัพยากรที่ใช้ในการคำนวณ รวมถึงประสิทธิภาพของโมเดลที่ได้จะไม่ดีเท่าที่ควร ดังนั้นในการเลือกคุณลักษณะจึงมีความจำเป็นเพื่อลดเวลาและลดทรัพยากรที่ใช้ในการคำนวณและเพิ่มประสิทธิภาพของโมเดล การเลือกคุณลักษณะที่มีการศึกษาในงานวิจัยนี้จะใช้ 2 วิธีการคือ

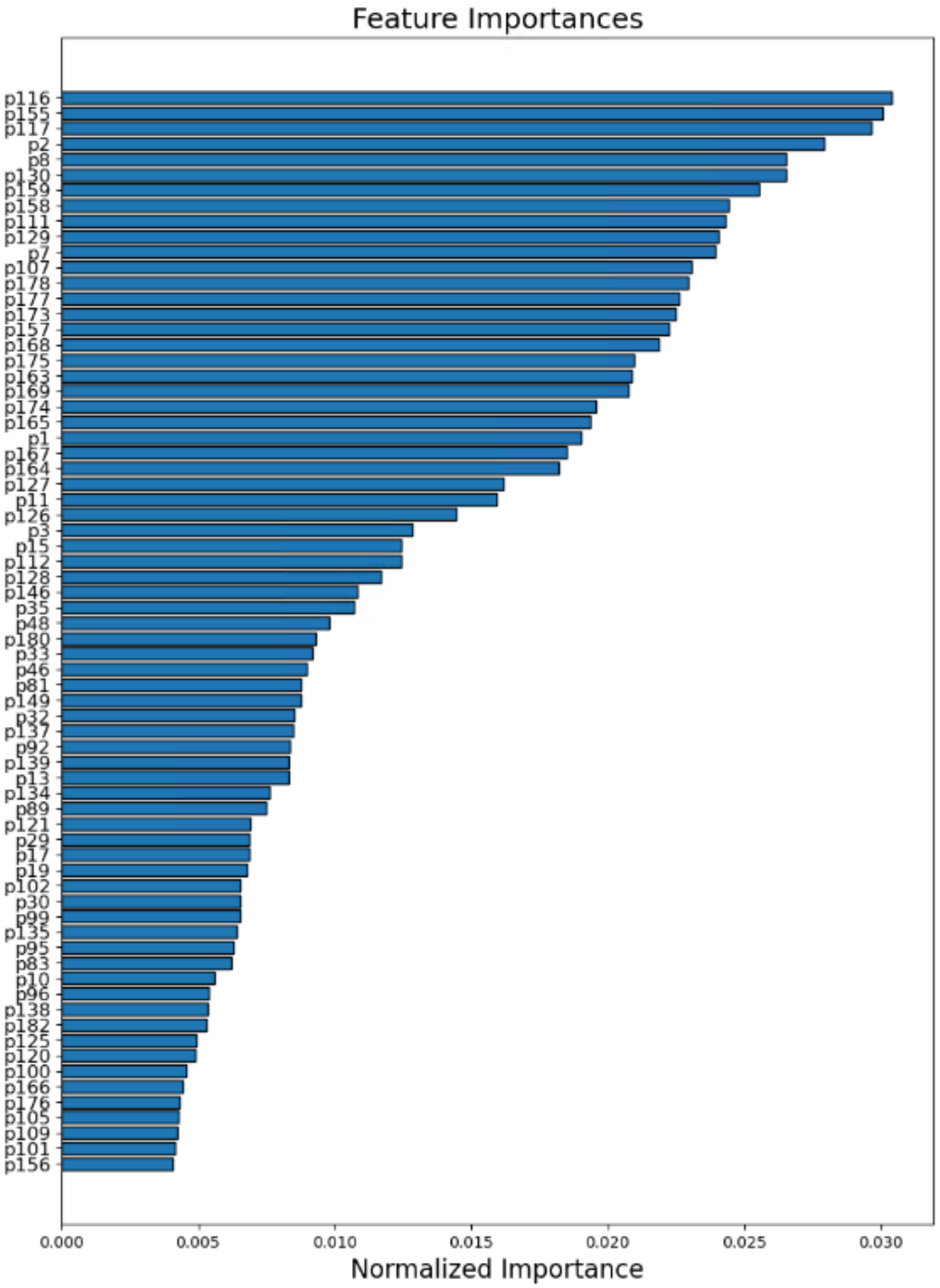
4.4.1 Filter Method เพื่อทำการตัดคุณลักษณะที่ไม่มีความสำคัญในการเรียนรู้ของโมเดล โดยทำการตัดคุณลักษณะที่เป็นค่าคงที่ ซึ่งเป็นพารามิเตอร์ที่ค่าของข้อมูลทุกตัวเป็นค่าเดียวกัน คุณลักษณะเฉพาะ เช่น หมายเลขของตัวฮาร์ดดิสก์ไดรฟ์ ตัดพารามิเตอร์ที่ซ้ำกันออก สำหรับพารามิเตอร์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์มากกว่า 95% จะทำการเลือกเพียง 1 ตัว และตัดพารามิเตอร์ที่เหลือออก โดยจะทำการเลือกจากพารามิเตอร์ที่ให้ ค่า ROC AUC สูงที่สุดจากการนำกลุ่มของพารามิเตอร์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์กันสูง ไปสร้าง Random Forest โมเดลอย่างง่าย

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

4.4.2 Embedded Method โดยใช้ Light Gradient Boost อัลกอริทึม โดยจะเลือกพารามิเตอร์ที่มีความสำคัญมากที่สุดในการเรียนรู้ของโมเดล โดยทำการเรียงลำดับความสำคัญของพารามิเตอร์จากการเรียนรู้ของโมเดลจำนวน 10 รอบ หาค่าเฉลี่ยความสำคัญของแต่ละพารามิเตอร์จากการเรียนรู้ทั้ง 10 รอบ แล้วทำการเรียงลำดับความสำคัญของพารามิเตอร์แต่ละตัว เลือกพารามิเตอร์โดยการรวมค่าความสำคัญ ให้ได้ค่าความสำคัญรวมที่ 95% และตัดพารามิเตอร์ที่เหลือออกเนื่องจากไม่มีความสำคัญในการเรียนรู้ของโมเดล

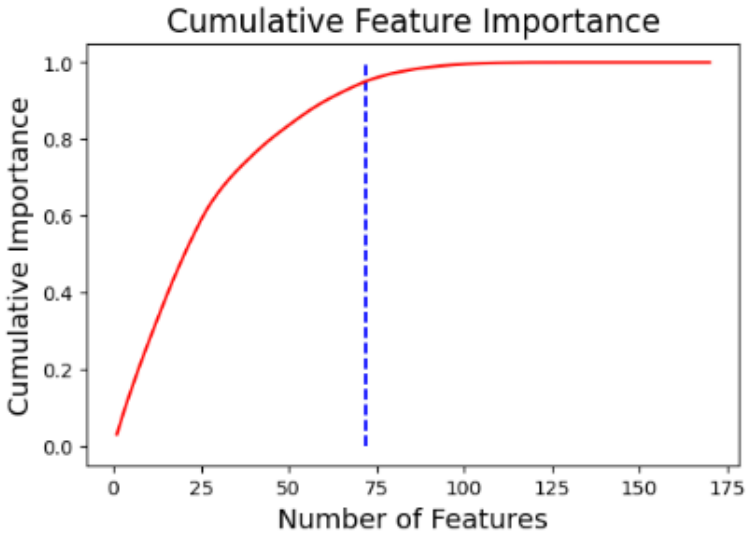
ในขั้นตอนการเลือกคุณลักษณะนี้ สามารถตัดพารามิเตอร์ที่ไม่จำเป็นต่อการเรียนรู้ของเครื่องออกไปจำนวน 289 พารามิเตอร์ และมี 70 พารามิเตอร์ที่ถูกเลือกไปใช้สำหรับการเรียนรู้ของเครื่อง การเรียงลำดับค่าความสำคัญของพารามิเตอร์แบบ Normalized แสดงได้ดังรูปที่ 8 และ กราฟแสดงการรวมค่าความสำคัญแบบ Normalized เทียบกับจำนวนพารามิเตอร์แสดงดังรูปที่ 9

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11



รูปที่ 8 *การเรียงลำดับความสำคัญของคุณลักษณะ แบบ Normalized*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11



รูปที่ 9 *การรวมค่าสำคัญของลักษณะแบบ Normalized เทียบกับจำนวนคุณลักษณะ*

จากรูปที่ 8 จะเห็นได้ว่าพารามิเตอร์ที่มีความสำคัญแบบ Normalized มากที่สุด 5 อันดับแรก คือ p116, p155, p117, p2 และ p8 ซึ่งเป็นพารามิเตอร์ที่เกี่ยวข้องกับค่าตำแหน่งที่ผิดปกติ คุณภาพของสัญญาณเซอร์โว,กระแสไฟฟ้าของมอเตอร์ และค่าพารามิเตอร์ของมอเตอร์

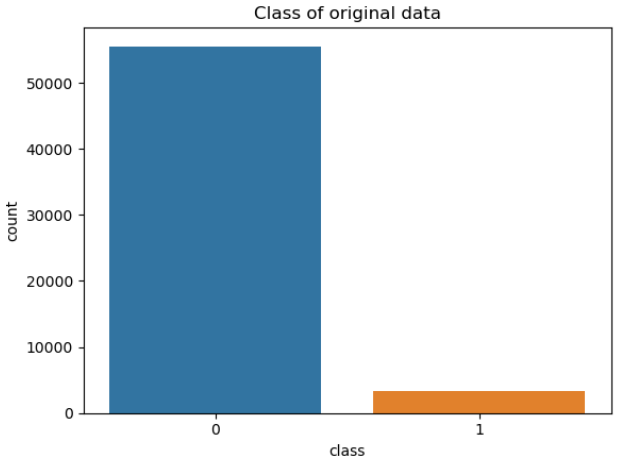
4.5 การจัดการกับข้อมูลที่ไม่สมดุล (Imbalance Data Handling)

เนื่องจากจำนวนของงานดีและงานเสียที่มีความแตกต่างกันมาก ในการสร้างโมเดลให้มีประสิทธิภาพที่ดี จึงมีการนำเทคนิคในการจัดการข้อมูลที่ไม่สมดุลหลายๆ วิธีการมาประยุกต์ใช้ในงานวิจัย

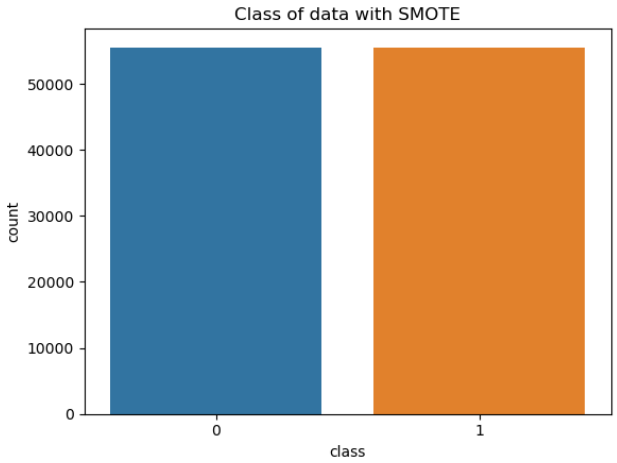
4.5.1 การลดจำนวนข้อมูลของกลุ่มงานดี ด้วยวิธี Bernoulli ซึ่งเทคนิคนี้ใช้ตั้งแต่การเก็บข้อมูลในขั้นตอน 4.1 โดยทำการเก็บข้อมูลของงานดี 1% จากงานดีทั้งหมด

4.5.2 ใช้เทคนิค SMOTE ในการเพิ่มจำนวนข้อมูลของกลุ่มงานเสีย โดยทำการสร้างชุดข้อมูลสังเคราะห์ของกลุ่มงานเสียจากชุดข้อมูลของงานเสียที่เก็บมาได้ โดยชุดข้อมูลเรียนรู้มีอัตราส่วนของงานดีต่องานเสียที่ 17:1 หลังจากเพิ่มจำนวนงานเสียด้วยวิธี SMOTE ทำให้อัตราส่วนของงานดีต่องานเสียอยู่ที่ 1:1 จำนวนข้อมูลงานดีเทียบกับงานเสียก่อนการทำ SMOTE แสดงดังรูปที่ 10 และหลังจากการทำ SMOTE ดังรูปที่ 11

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11



รูปที่ 10 *ข้อมูลของกลุ่มงานดีเทียบกับงานเสียก่อนการทำ SMOTE*



รูปที่ 11 *ข้อมูลของกลุ่มงานดีเทียบกับงานเสียหลังการทำ SMOTE*

4.4.3 ใช้อัลกอริทึมการเรียนรู้ด้วยเครื่อง ร่วมกับ Different Cost Learner เพื่อกำหนดค่าใช้จ่ายที่เกิดจากความผิดพลาดในการทำนายทำนายงานเสียว่าเป็นงานดี ให้สูงกว่าความผิดพลาดจากการทำนายงานดีว่าเป็นงานเสียด้วยค่าอัตราส่วนของงานดีต่องานเสีย ซึ่งวิธีการนี้จะทำการกำหนดค่าในไฮเปอร์พารามิเตอร์ของแต่ละอัลกอริทึมการเรียนรู้ของเครื่อง สำหรับ SVM ใช้ไฮเปอร์พารามิเตอร์ชื่อ class_weight และ XGBoost ใช้ ไฮเปอร์พารามิเตอร์ชื่อ scale_pos_weight

4.4.4 SMOTE with Different Cost เป็นการใช้เทคนิคของ SMOTE ร่วมกับ Different Cost Learner เพื่อเพิ่มข้อมูลของกลุ่มงานเสีย และกำหนดค่าใช้จ่ายของความผิดพลาดจากการทำนายงานเสียว่าเป็นงานดี ให้สูงกว่าความผิดพลาดจากการทำนายงานดีว่าเป็นงานเสีย

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

4.6 สร้างโมเดลการเรียนรู้ (Training Model)

การเรียนรู้ของโมเดลจะประกอบไปด้วย 6 วิธีที่แตกต่างกัน มาจาก 2 อัลกอริทึมการเรียนรู้ของเครื่อง และ 3 เทคนิคที่ใช้ในการจัดการข้อมูลที่ไม่สมดุล

4.6.1 SVM with SMOTE (SVM SMOTE) เป็นการนำ SVM อัลกอริทึมมาประยุกต์ใช้กับการเพิ่มจำนวนงานเสียด้วยวิธีการ SMOTE

4.6.2 SVM with Different Cost Learner (SVM DC) เป็นการใช้ SVM อัลกอริทึมโดยปรับค่าไฮเปอร์พารามิเตอร์ให้ค่าผิดพลาดที่เกิดจากการทำนายงานเสียมีความสำคัญมากกว่าความผิดพลาดในการทำนายงานดีนดี

4.6.3 SVM with SMOTE and Different Cost (SVM SDC) เป็นการนำวิธีการ SVM SMOTE และ SVM DC มาใช้ร่วมกัน ซึ่งค่าไฮเปอร์พารามิเตอร์ของ SVM ที่ใช้ เป็นดังตารางที่ 4 ส่วนค่าอื่นๆเป็นค่าเริ่มต้นของ Sklearn

ตารางที่ 4 ค่าไฮเปอร์พารามิเตอร์ของ SVM อัลกอริทึม

Hyperparameter	SVM SMOTE	SVM DC	SVM SDC
Kernel	rbf	rbf	rbf
Degree	3	3	3
Gamma	scale	scale	scale
Max_iter	1	1	1
Class_weight	default	pos_ratio	pos_ratio

4.6.4 XGBoost with SMOTE (XGB SMOTE) เป็นการนำ XGBoost อัลกอริทึมมาประยุกต์ใช้กับการเพิ่มจำนวนงานเสียด้วยวิธีการ SMOTE

4.6.5 XGBoost with Different Cost Learner (XGB DC) เป็นการใช้ XGBoost อัลกอริทึมโดยปรับค่าไฮเปอร์พารามิเตอร์ให้ค่าผิดพลาดที่เกิดจากการทำนายงานเสียมีความสำคัญมากกว่าความผิดพลาดในการทำนายงานดี

4.6.6 XGBoost with SMOTE and Different Cost (XGB SDC) เป็นการนำวิธีการ XGB SMOTE และ XGB DC มาใช้ร่วมกัน โดยมีการเพิ่มจำนวนงานเสียด้วย SMOTE และมีการกำหนดค่าผิดพลาดจากการทำนายงานเสียมีค่าสูงกว่างานดี ซึ่งค่าไฮเปอร์พารามิเตอร์ของ XGBoost ที่ใช้ เป็นดังตารางที่ 5 ส่วนค่าอื่นๆ เป็นค่าเริ่มต้นของ XGBoost Library

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ตารางที่ 5 ค่าไฮเปอร์พารามิเตอร์ของ XGBoost อัลกอริทึม

Hyperparameter	XGB SMOTE	XGB DC	XGB SDC
Booster	gbtree	gbtree	gbtree
Eta	0.3	0.3	0.3
Gamma	0	0	0
Max_depth	6	6	6
Scale_pos_weight	default	pos_ratio	pos_ratio

4.7 การวัดประสิทธิภาพของโมเดล (Performance Measurement)

หลังจากสร้างโมเดลการเรียนรู้ของเครื่องด้วยข้อมูลเรียนรู้แล้ว ทำการทดสอบโมเดลที่สร้างขึ้นทั้ง 6 โมเดล ด้วยข้อมูลทดสอบ และทำการวัดประสิทธิภาพของแต่ละโมเดลจากผลการทำนายจากข้อมูลทดสอบ โดยในการวัดประสิทธิภาพของโมเดลจะพิจารณาและเปรียบเทียบค่าของพื้นที่ใต้กราฟของ Receiver Operating Characteristics (ROC AUC) และพื้นที่ใต้กราฟของ Precision-Racall (PRC AUC) เนื่องจากชุดข้อมูลที่ใช้ในงานวิจัยนี้มีลักษณะไม่สมดุล การวัดประสิทธิภาพของโมเดลด้วยค่า Accuracy ไม่เหมาะสมสำหรับข้อมูลที่มีความไม่สมดุล เนื่องจากในกรณีที่โมเดลทำนายค่าทั้งหมดเป็นงานดี และไม่สามารถทำนายค่าของงานเสียได้เลย ค่า Accuracy ของโมเดลยังคงให้ประสิทธิภาพสูงอยู่ การใช้ ROC AUC และ PRC AUC สามารถแสดงถึงประสิทธิภาพที่แท้จริงของแต่ละโมเดลได้ดีกว่า นอกจากการวัดค่าพื้นที่ใต้กราฟของ ROC และ PRC แล้วก็จะมีการเปรียบเทียบประสิทธิภาพของโมเดลจาก Confusion Matrix, Precision, Recall และ F1-Score ด้วย

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

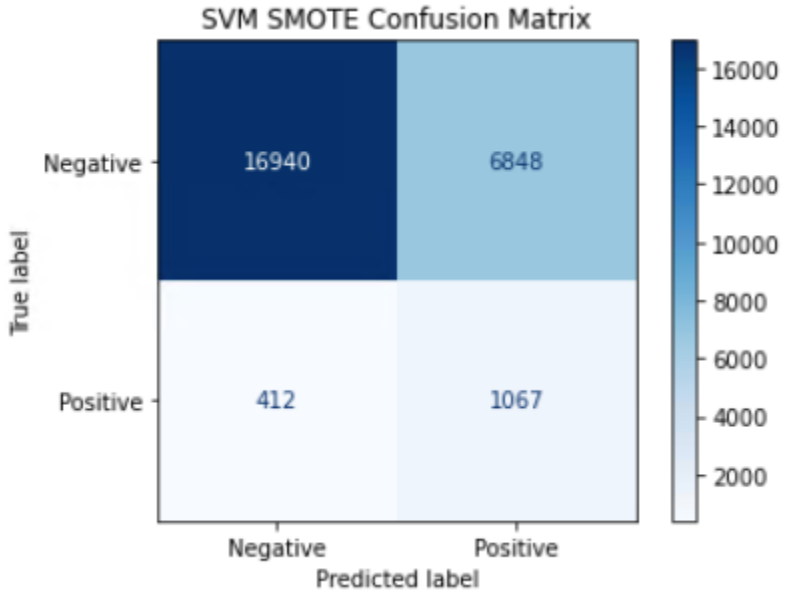
บทที่ 5
ผลการทดลอง

ผลการทดลองในการทำนายงานดีและงานเสีย จากชุดข้อมูลทดสอบของฮาร์ดดิสก์ไดรฟ์จำนวน 25,267 ตัว กับโมเดลการเรียนรู้ของเครื่องทั้ง 6 วิธีการ ได้ผลดังต่อไปนี้

5.1 การวัดประสิทธิภาพด้วย Confusion Matrix

โดยกำหนดให้งานเสียคือ Positive และงานดีเป็น Negative การวัดผลของโมเดลด้วย Confusion Matrix สามารถแสดงได้ดังนี้

5.1.1 SVM SMOTE: Confusion Matrix แสดงดังรูปที่ 12 จะเห็นได้ว่า SVM SMOTE สามารถตรวจจับงานเสียได้จำนวน 1,067 ตัว คิดเป็น 72.14% ของงานเสียทั้งหมด ในขณะที่งานดีที่ทำนายว่าเป็นงานเสีย มีสูงถึง 6,848 ตัว คิดเป็น 28.79% ของงานดีทั้งหมด ซึ่งมีความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียค่อนข้างสูง



รูปที่ 12 *Confusion Matrix ของโมเดล SVM SMOTE*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

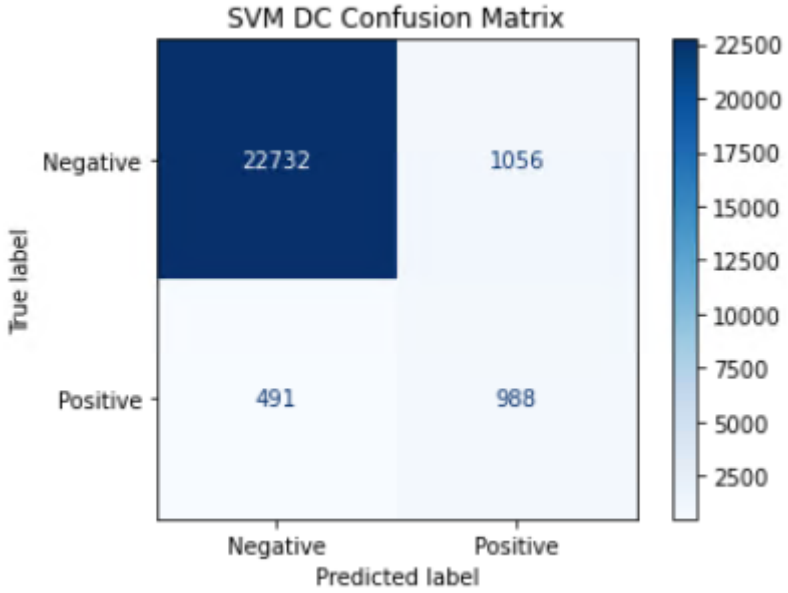
5.1.2 ผล Classification Report ของ SVM SMOTE แสดงค่า Accuracy, Precision, Recall และ F1-score ได้ดังรูปที่ 13 เมื่อพิจารณาจากงานดี (คลาส 0) จะได้ค่า Precision 98%, Recall 71% และ F1-score 82% ซึ่งให้ค่าประสิทธิภาพของโมเดลที่ค่อนข้างดี แต่เมื่อพิจารณาจากงานเสีย (คลาส 1) จะได้ค่า Precision เพียง 13%, และ F1-score เพียง 23% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ค่อนข้างต่ำมาก เนื่องจากความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียจากข้อ 5.1.1 มีอัตราสูง และเมื่อพิจารณาค่าเฉลี่ยน้ำหนักของงานดีและงานเสีย จะได้ค่า Precision 93%, Recall 71% และ F1-score 79% ซึ่งให้ประสิทธิภาพพอใช้

```
### SVM SMOTE classification report ###
              precision    recall  f1-score   support

           0       0.98      0.71      0.82     23788
           1       0.13      0.72      0.23      1479

    accuracy                           0.71     25267
   macro avg       0.56      0.72      0.53     25267
weighted avg       0.93      0.71      0.79     25267
```

รูปที่ 13 *ผล Accuracy, Precision, Recall, F1-score ของโมเดล SVM SMOTE*

5.1.3 5.1.3 SVM DC: Confusion Matrix แสดงดังรูปที่ 14 จะเห็นได้ว่า SVM DC สามารถตรวจจับงานเสียได้จำนวน 988 ตัว คิดเป็น 66.8% ของงานเสียทั้งหมด ในขณะที่งานดีที่ทำนายว่าเป็นงานเสีย มีจำนวน 1,056 ตัว คิดเป็น 4.44% ของงานดีทั้งหมด ซึ่งโมเดลนี้มีความสามารถในการทำนายงานเสียได้พอใช้และมีความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียค่อนข้างต่ำ

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11



รูปที่ 14 *Confusion Matrix ของโมเดล SVM DC*

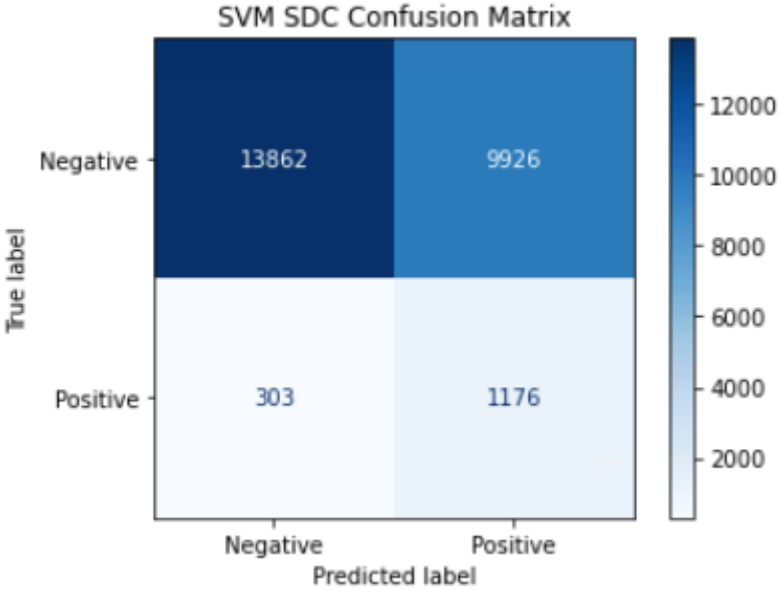
5.1.4 ผล Classification Report ของ SVM DC แสดงค่า Accuracy, Precision, Recall และ F1-score ได้ดังรูปที่ 15 เมื่อพิจารณางานดี (คลาส 0) จะได้ค่า Precision 98%, Recall 96% และ F1-score 97% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ค่อนข้างสูงมาก และเมื่อพิจารณาจากงานเสีย (คลาส 1) จะได้ค่า Precision 48%, Recall 67% และ F1-score 56% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่พอใช้ และเมื่อพิจารณาค่าเฉลี่ยน้ำหนักของงานดีและงานเสีย จะได้ค่า Precision 95%, Recall 94% และ F1-score 94% ซึ่งให้ประสิทธิภาพของโมเดลที่ค่อนข้างดี

```
### SVM DC classification report ###
              precision    recall  f1-score   support

           0       0.98      0.96      0.97     23788
           1       0.48      0.67      0.56      1479

    accuracy                           0.94     25267
   macro avg       0.73      0.81      0.76     25267
weighted avg       0.95      0.94      0.94     25267
```

รูปที่ 15 *ผล Accuracy, Precision, Recall, F1-score ของโมเดล SVM DC*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

5.1.5 SVM SDC: Confusion Matrix แสดงดังรูปที่ 16 จะเห็นได้ว่า โมเดล SVM SDC สามารถตรวจจับงานเสียได้จำนวน 1,176 ตัว คิดเป็น 79.51% ของงานเสียทั้งหมด ในขณะที่งานดีที่ทำนายผิดพลาดว่าเป็นงานเสีย มีจำนวนสูงถึง 9,926 ตัว คิดเป็น 41.73% ของงานดีทั้งหมด เป็นค่าการผิดพลาดในการทำนายที่ค่อนข้างสูง โมเดลนี้มีความสามารถในการทำนายงานเสียได้ดี แต่มีความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียค่อนข้างสูงมาก



รูปที่ 16 *Confusion Matrix ของโมเดล SVM SDC*

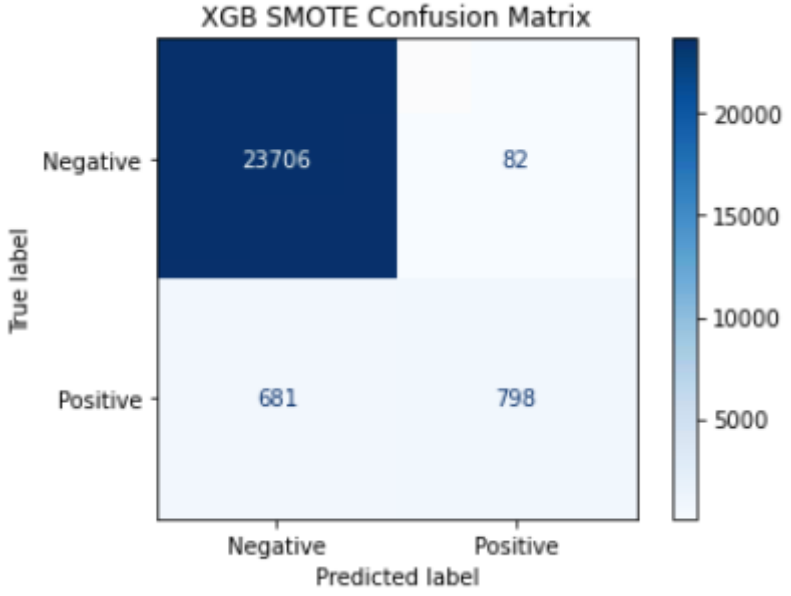
5.1.6 ผล Classification Report ของ SVM SDC แสดงค่า Accuracy, Precision, Recall และ F1-score ได้ดังรูปที่ 17 เมื่อพิจารณาการทำนายว่าเป็นงานดี (คลาส 0) จะได้ค่า Precision 98%, Recall 58% และ F1-score 73% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ค่อนข้างต่ำและเมื่อพิจารณาจากการทำนายว่าเป็นงานเสีย (คลาส 1) จะได้ค่า Precision 11%, Recall 80% และ F1-score 19% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ค่อนข้างต่ำ เนื่องจากความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียจากข้อ 5.1.5 มีอัตราสูงถึง 41.73% และเมื่อพิจารณาค่าเฉลี่ยน้ำหนักของงานดีและงานเสีย จะได้ค่า Precision 93%, Recall 60% และ F1-score 70% ซึ่งจะได้ว่าประสิทธิภาพของโมเดลค่อนข้างต่ำ

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

```
### SVM SDC classification report ###
              precision    recall  f1-score   support

           0       0.98      0.58      0.73     23788
           1       0.11      0.80      0.19      1479

    accuracy                           0.60     25267
   macro avg       0.54      0.69      0.46     25267
weighted avg       0.93      0.60      0.70     25267
```

รูปที่ 17 *ผล Accuracy, Precision, Recall, F1-score ของโมเดล SVM SDC*

5.1.7 XGB SMOTE: Confusion Matrix แสดงดังรูปที่ 18 จะเห็นได้ว่า โมเดล XGB SMOTE สามารถตรวจจับงานเสียได้จำนวน 798 ตัว คิดเป็น 53.96% ของงานเสียทั้งหมด ในขณะที่งานดีที่ทำนายผิดพลาดว่าเป็นงานเสีย มีเพียง 82 ตัว คิดเป็น 0.34% ของงานดีทั้งหมด โมเดลนี้มีความสามารถในการทำนายงานเสียได้พอใช้ แต่มีความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียต่ำมาก ซึ่งเป็นโมเดลที่เหมาะสำหรับการนำไปใช้งาน



รูปที่ 18 *Confusion Matrix ของโมเดล XGB SMOTE*

5.1.8 ผล Classification Report ของ XGB SMOTE แสดงค่า Accuracy, Precision, Recall และ F1-score ได้ดังรูปที่ 19 เมื่อพิจารณาการทำนายว่าเป็นงานดี (คลาส 0) จะได้ค่า Precision 97%, Recall 100% และ F1-score 98% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ดีมาก และเมื่อพิจารณาจากการทำนายว่าเป็นงานเสีย (คลาส 1) จะได้ค่า Precision 91%, Recall 54% และ

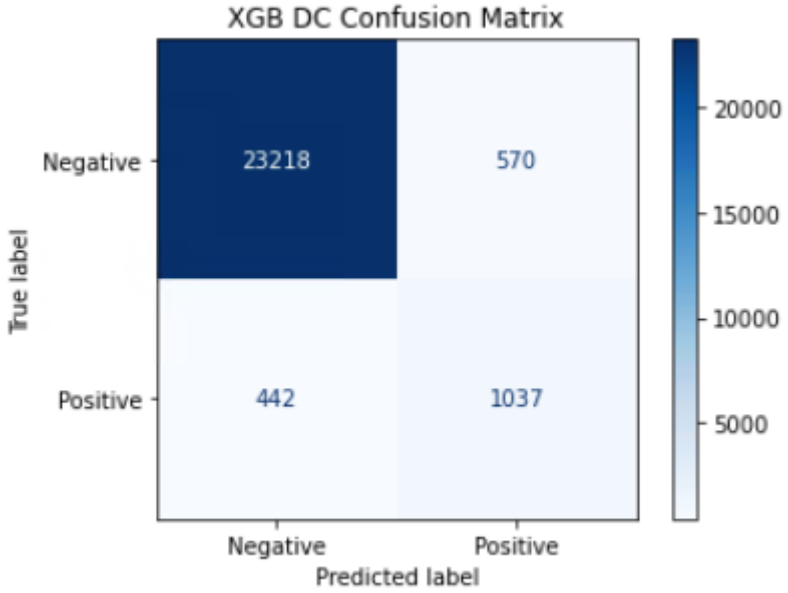
2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

F1-score 68% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ค่อนข้างดีเมื่อเทียบกับโมเดลอื่นๆ และเมื่อพิจารณาค่าเฉลี่ยน้ำหนักของงานดีและงานเสีย จะได้ค่า Precision 97%, Recall 97% และ F1-score 97% จะเห็นได้ว่าโมเดลให้ประสิทธิภาพที่ดีมาก

```
### XGB SMOTE classification report ###
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     23788
           1       0.91      0.54      0.68      1479

    accuracy                           0.97     25267
   macro avg       0.94      0.77      0.83     25267
weighted avg       0.97      0.97      0.97     25267
```

รูปที่ 19 *ผล Accuracy, Precision, Recall, F1-score ของโมเดล XGB SMOTE*

5.1.9 XGB DC: Confusion Matrix แสดงดังรูปที่ 20 จะเห็นได้ว่า โมเดล XGB DC สามารถตรวจจับงานเสียได้จำนวน 1,037 ตัว คิดเป็น 70.11% ของงานเสียทั้งหมด ในขณะที่งานดีที่ทำนายผิดพลาดว่าเป็นงานเสีย มีจำนวน 570 ตัว คิดเป็น 2.4% ของงานดีทั้งหมด เป็นค่าการผิดพลาดในการทำนายที่ค่อนข้างต่ำ โมเดลนี้มีความสามารถในการทำนายงานเสียได้ค่อนข้างดี และมีความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียค่อนข้างต่ำ



รูปที่ 20 *Confusion Matrix ของโมเดล XGB DC*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

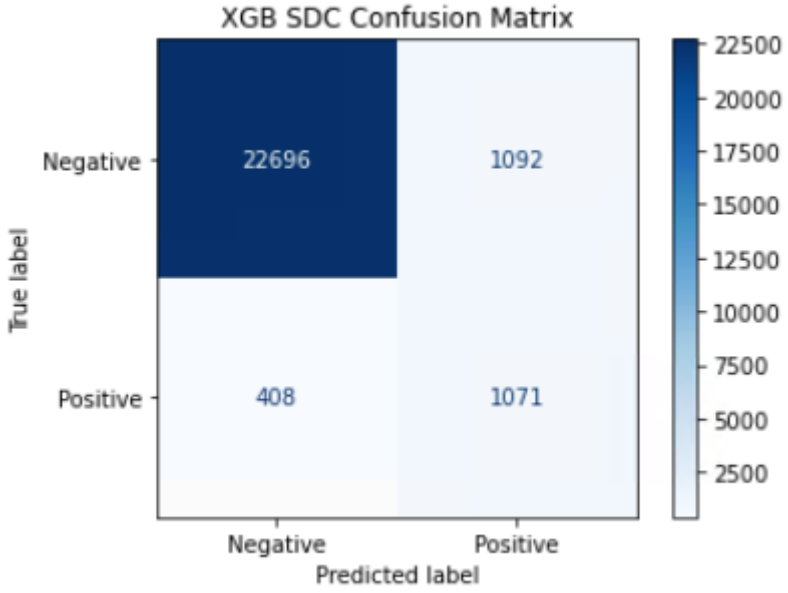
5.1.10 ผล Classification Report ของ XGB DC แสดงค่า Accuracy, Precision, Recall และ F1-score ได้ดังรูปที่ 21 เมื่อพิจารณาการทำนายว่าเป็นงานดี (คลาส 0) จะได้ค่า Precision 98%, Recall 98% และ F1-score 98% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ดีมาก และเมื่อพิจารณาจากการทำนายว่าเป็นงานเสีย (คลาส 1) จะได้ค่า Precision 65%, Recall 70% และ F1-score 67% ซึ่งเป็นค่าประสิทธิภาพของโมเดลค่อนข้างดี และเมื่อพิจารณาค่าเฉลี่ยน้ำหนักของงานดีและงานเสีย จะได้ค่า Precision 96%, Recall 96% และ F1-score 96% ซึ่งให้ประสิทธิภาพของโมเดลที่ดีมาก

```
### XGB DC classification report ###
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     23788
           1       0.65      0.70      0.67      1479

    accuracy                           0.96     25267
   macro avg       0.81      0.84      0.83     25267
weighted avg       0.96      0.96      0.96     25267
```

รูปที่ 21 *ผล Accuracy, Precision, Recall, F1-score ของโมเดล XGB DC*

5.1.11 XGB SDC: Confusion Matrix แสดงดังรูปที่ 22 จะเห็นได้ว่า โมเดล XGB SDC สามารถตรวจจับงานเสียได้จำนวน 1,071 ตัว คิดเป็น 72.41% ของงานเสียทั้งหมด ในขณะที่งานดีที่ทำนายผิดพลาดว่าเป็นงานเสีย มีจำนวน 1,092 ตัว คิดเป็น 4.6% ของงานดีทั้งหมด เป็นค่าการผิดพลาดในการทำนายที่ค่อนข้างต่ำ โมเดลนี้มีความสามารถในการทำนายงานเสียได้ค่อนข้างดี และมีความผิดพลาดในการทำนายงานดีว่าเป็นงานเสียค่อนข้างต่ำ

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11



รูปที่ 22 *Confusion Matrix ของโมเดล XGB SDC*

5.1.12 ผล Classification Report ของ XGB SDC แสดงค่า Accuracy, Precision, Recall และ F1-score ได้ดังรูปที่ 23 เมื่อพิจารณาการทำนายว่าเป็นงานดี (คลาส 0) จะได้ค่า Precision 98%, Recall 95% และ F1-score 97% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ดีมาก และเมื่อพิจารณาจากการทำนายว่าเป็นงานเสีย (คลาส 1) จะได้ค่า Precision 50%, Recall 72% และ F1-score 59% ซึ่งเป็นค่าประสิทธิภาพของโมเดลที่ระดับปานกลาง และเมื่อพิจารณาค่าเฉลี่ยน้ำหนักของงานดีและงานเสีย จะได้ค่า Precision 95%, Recall 94% และ F1-score 95% ซึ่งให้ประสิทธิภาพของโมเดลที่ดี

```
### XGB SDC classification report ###
              precision    recall  f1-score   support

           0       0.98      0.95      0.97     23788
           1       0.50      0.72      0.59      1479

    accuracy                           0.94     25267
   macro avg       0.74      0.84      0.78     25267
weighted avg       0.95      0.94      0.95     25267
```

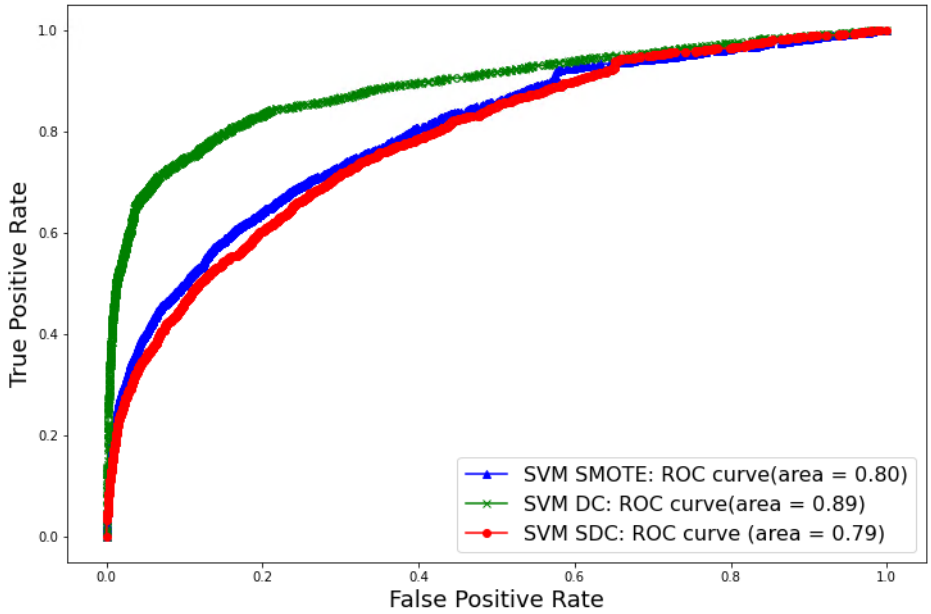
รูปที่ 23 *ผล Accuracy, Precision, Recall, F1-score ของโมเดล XGB SDC*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

5.2 การวัดประสิทธิภาพด้วย ROC AUC

กราฟของ ROC คือวัดค่าประสิทธิภาพของ True Positive Rate และ False Positive Rate ด้วยการปรับค่าเกณฑ์ความน่าจะเป็นของการทำนายที่ต่างกัน โดยค่า ROC AUC เป็นการวัดพื้นที่ใต้กราฟของ ROC ซึ่งค่าของโมเดลที่มีประสิทธิภาพดีจะให้ค่า ROC AUC เข้าใกล้ 1

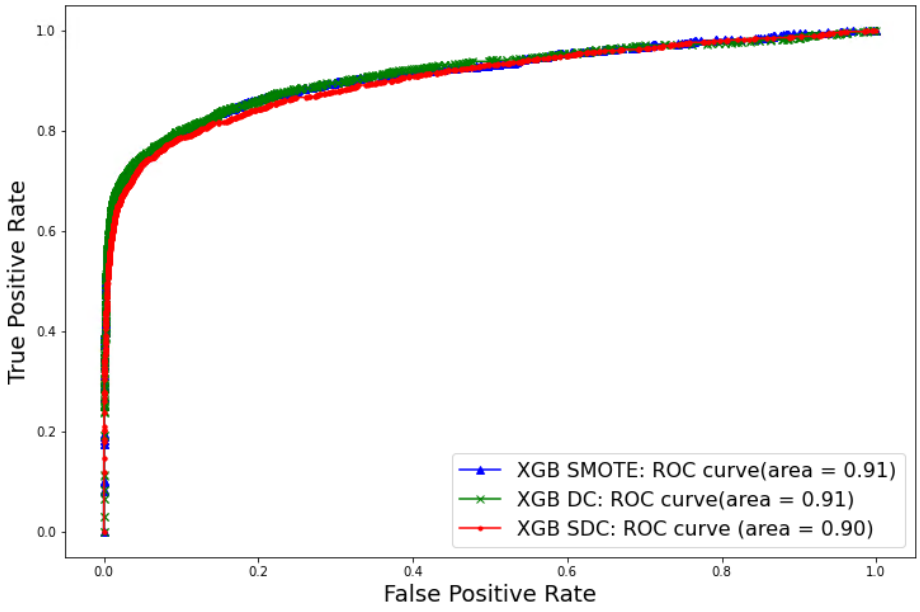
5.2.1 SVM อัลกอริทึม: ROC กราฟ และ ค่า ROC AUC สามารถ แสดงได้ดังรูปที่ 24 SVM อัลกอริทึมที่ให้ค่า ROC AUC สูงที่สุด คือ SVM DC ที่ ROC AUC 89% แสดงได้ดังเส้นสีเขียว ขณะที่ SVM SMOTE มีค่า ROC AUC ที่ 80% ดังสีน้ำเงิน และ SVM SDC ให้ประสิทธิภาพต่ำที่สุดในการวัดค่า ROC AUC ที่ 79% ดังเส้นสีแดง



รูปที่ 24 *ROC AUC ของ SVM อัลกอริทึม*

5.2.2 XGBoost อัลกอริทึม: ROC กราฟ และ ค่า ROC AUC สามารถ แสดงได้ดังรูปที่ 25 XGBoost อัลกอริทึมที่ให้ค่า ROC AUC สูงที่สุด คือ XGB SMOTE แสดงได้ดังเส้นสีน้ำเงิน และ XGB DC แสดงได้ดังเส้นสีเขียว ที่ค่า ROC AUC ที่ 91% และ XGB SDC ให้ประสิทธิภาพต่ำที่สุดในการวัดค่า ROC AUC ที่ 90% ดังเส้นสีแดง

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11



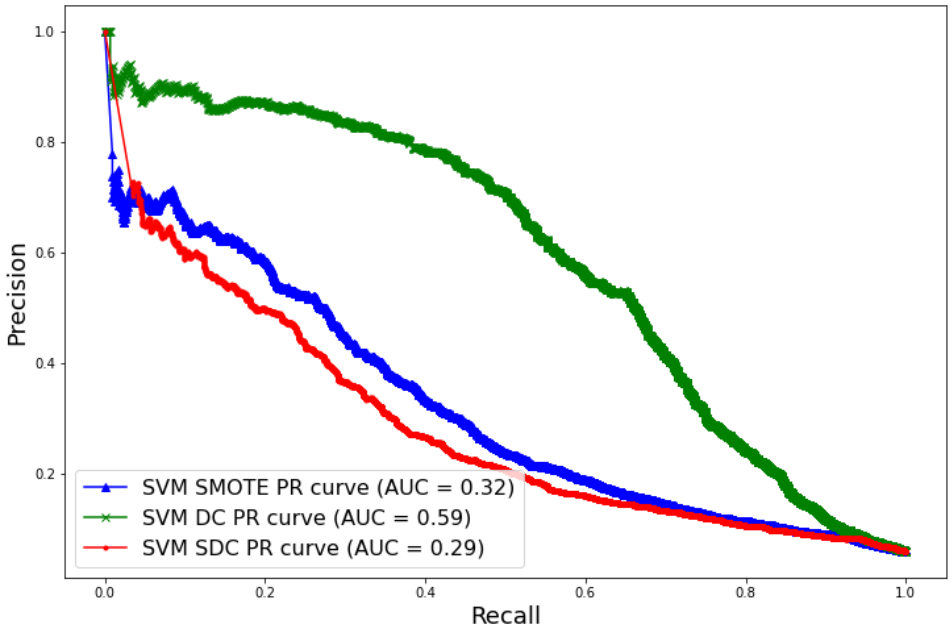
รูปที่ 25 ROC AUC ของ XGBoost อัลกอริทึม

5.3 การวัดประสิทธิภาพด้วย PRC AUC

กราฟ PRC คือกราฟที่วัดค่าประสิทธิภาพของ Precision และ Recall ด้วยการปรับค่าเกณฑ์ความน่าจะเป็นของการทำนายที่ต่างกัน โดยค่า PRC AUC เป็นการวัดพื้นที่ใต้กราฟของ PRC ซึ่งค่าของโมเดลที่มีประสิทธิภาพดีจะให้ค่า PRC AUC เข้าใกล้ 1 ซึ่งในการวัดประสิทธิภาพด้วย PRC จะให้ความสนใจกับงานเสีย และการทำนายว่าเป็นงานเสียจากโมเดล

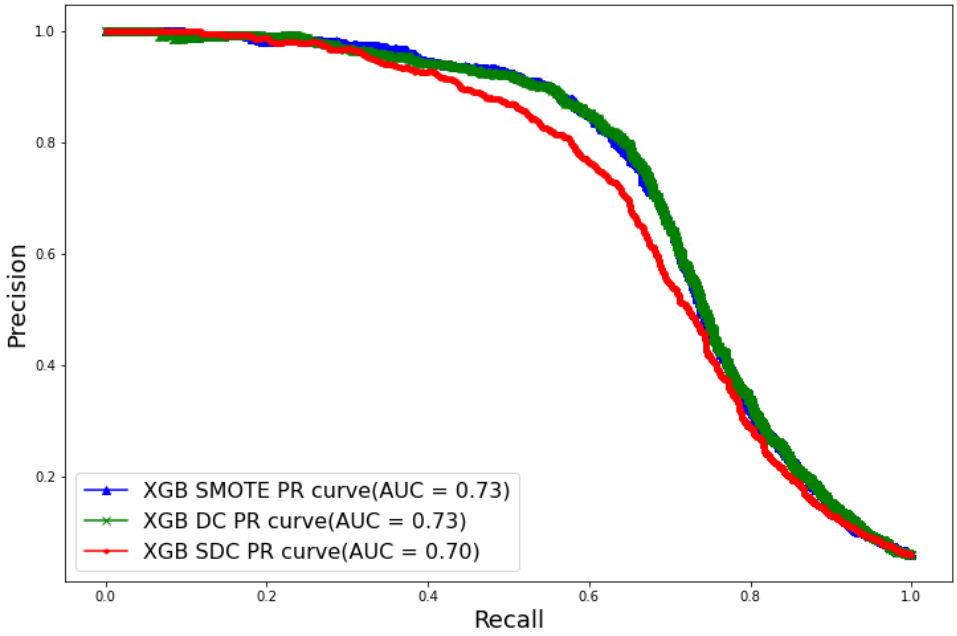
5.3.1 SVM อัลกอริทึม: กราฟ PRC และค่า PRC AUC ของ SVM แสดงได้ดังรูปที่ 26 โมเดลที่ให้ประสิทธิภาพดีที่สุดในการวัดค่าด้วย PRC AUC คือ SVM DC ที่ 59% แสดงได้ดังเส้นสีเขียว SVM SMOTE มีค่า PRC AUC ที่ 32% แสดงได้ดังเส้นสีน้ำเงิน และ SVM SDC ให้ประสิทธิผลที่ต่ำที่สุดที่ PRC AUC 29% แสดงได้ดังเส้นสีแดง เนื่องจากทั้งสองโมเดลมีงานดีที่ทำนายผิดพลาดว่าเป็นงานเสียค่อนข้างสูงมาก จึงทำให้การวัดค่าประสิทธิภาพด้วย PRC AUC มีค่าต่ำ

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11



รูปที่ 26 *PRC AUC ของ SVM อัลกอริทึม*

5.3.2 XGBoost อัลกอริทึม: กราฟ PRC และค่า PRC AUC ของ XGBoost แสดงได้ดังรูปที่ 27 โมเดลที่ให้ประสิทธิภาพดีที่สุดในการวัดค่าด้วย PRC AUC คือ XGB SMOTE และ XGB DC ที่ 73% แสดงได้ดังเส้นสีน้ำเงิน และสีเขียวตามลำดับ ส่วน XGB SDC ให้ค่า PRC AUC ต่ำสุดที่ 70% เส้นสีแดง ซึ่ง XGBoost อัลกอริทึมให้ประสิทธิภาพที่ค่อนข้างดีในการวัดค่าด้วย PRC AUC



รูปที่ 27 *PRC AUC ของ XGBoost อัลกอริทึม*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

5.4 ผลการเปรียบเทียบประสิทธิภาพของโมเดลทั้ง 6 วิธี

เปรียบเทียบประสิทธิภาพของโมเดลทั้ง 6 วิธี ด้วยค่า ROC AUC, PRC AUC, Accuracy, Precision, Recall และ F1-score วิธีการ XGB SMOTE ให้ประสิทธิภาพของโมเดลดีที่สุด ที่ ROC AUC 91%, PRC AUC 73% และ Precision, Recall, F1-score ที่ 97% ในขณะที่ SVM SDC ให้ประสิทธิภาพของโมเดลต่ำที่สุด วิธีการของ XGBoost ทั้ง 3 วิธีการให้ประสิทธิภาพที่ดีกว่าวิธีการของ SVM โดยตารางเปรียบเทียบค่าประสิทธิภาพของทั้ง 6 โมเดล ด้วยค่าเฉลี่ยน้ำหนัก ของ Accuracy, Precision, Recall และ F1-score แสดงได้ดังตารางที่ 6

ตารางที่ 6 เปรียบผลการวัดประสิทธิภาพ

Method	ROC AUC	PRC AUC	Accuracy	Precision	Recall	F1-score
SVM SMOTE	80%	32%	71%	93%	71%	79%
SVM DC	89%	59%	94%	95%	94%	94%
SVM SDC	78%	29%	60%	93%	60%	70%
XGB SMOTE	91%	73%	97%	97%	97%	97%
XGB DC	91%	73%	96%	96%	96%	96%
XGB SDC	90%	70%	94%	95%	94%	95%

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

ในงานวิจัยฉบับนี้ทำการศึกษาวิธีการในการตรวจจับงานเสียที่เกิดขึ้นจากปัญหาการอ่านสัญญาณเซอร์โว ในกระบวนการทดสอบ ด้วย SVM และ XGBoost อัลกอริทึม มีการนำเสนอวิธีการในการเลือกคุณลักษณะ การจัดการกับข้อมูลที่ไม่สมดุล และทำการทดลองด้วยชุดข้อมูล ที่เก็บจากงานจริงในการผลิตฮาร์ดดิสก์ไดรฟ์

ในการเลือกคุณลักษณะใช้วิธีการ Filter และ Embedded ทำให้สามารถลดพารามิเตอร์จาก 359 พารามิเตอร์เหลือเพียง 70 พารามิเตอร์ สำหรับใช้ในการเรียนรู้ของอัลกอริทึม และใช้วิธีการจัดการข้อมูลที่ไม่สมดุล 3 แบบคือ SMOTE, Different Cost Learner และ SMOTE with Different Cost ร่วมกับอัลกอริทึม SVM และ XGBoost รวมเป็น 6 วิธีการ และทำการทดสอบด้วยชุดข้อมูลทดสอบจำนวน 25,267 ข้อมูล

จากผลการทดลองแสดงให้เห็นว่า XGBoost อัลกอริทึมให้ประสิทธิภาพที่ดีกว่า SVM อัลกอริทึมในทุกวิธีการ การจัดการข้อมูลไม่สมดุลด้วยวิธีการ Different Cost Learner ให้ประสิทธิภาพที่ดีที่สุดสำหรับ SVM และยังให้ประสิทธิภาพที่ดีใน XGBoost อัลกอริทึมด้วยเช่นกัน การจัดการข้อมูลไม่สมดุลด้วยวิธีการ SMOTE ให้ประสิทธิภาพที่ดีที่สุดสำหรับ XGBoost อัลกอริทึม

การวัดประสิทธิภาพของโมเดลด้วย ROC AUC แสดงให้เห็นว่าประสิทธิภาพของโมเดล XGBoost มีประสิทธิภาพที่ดีมาก มีค่า ROC AUC มากกว่า 90% ในทุกวิธีการจัดการข้อมูลไม่สมดุล ขณะเดียวกัน SVM อัลกอริทึม ยังคงให้ประสิทธิภาพที่ค่อนข้างสูงในทุกวิธีการจัดการข้อมูลที่ไม่สมดุลเช่นกัน ส่วนการวัดประสิทธิภาพของโมเดลด้วย PRC AUC แสดงให้เห็นว่าประสิทธิภาพของโมเดล XGBoost ยังคงให้ประสิทธิภาพที่ดี โดย PRC AUC มีค่ามากกว่า 70% ในขณะประสิทธิภาพของ SVM SMOTE และ SMOTE SDC มีค่าต่ำ เนื่องจากการทำนายงานดีที่ผิดพลาดว่าเป็นงานเสียมีอัตราที่สูง และเมื่อเปรียบเทียบประสิทธิภาพของทุกโมเดลจากการทดลองนี้ จะเห็นได้ว่า XGB SMOTE ให้ผลที่ดีที่สุด ที่ ROC AUC 91%, PRC AUC 73%, ความถูกต้อง 97% Precision 97% Recall 97% และ F1-score 97%

ในส่วนของการนำวิธีการนี้ไปใช้งานในการตรวจจับงานเสียที่เกิดขึ้น นอกจากจำนวนงานเสียที่สามารถตรวจจับได้จากการทำนายแล้ว จำเป็นต้องมีการพิจารณาในส่วนของงานดีที่โมเดลทำนาย

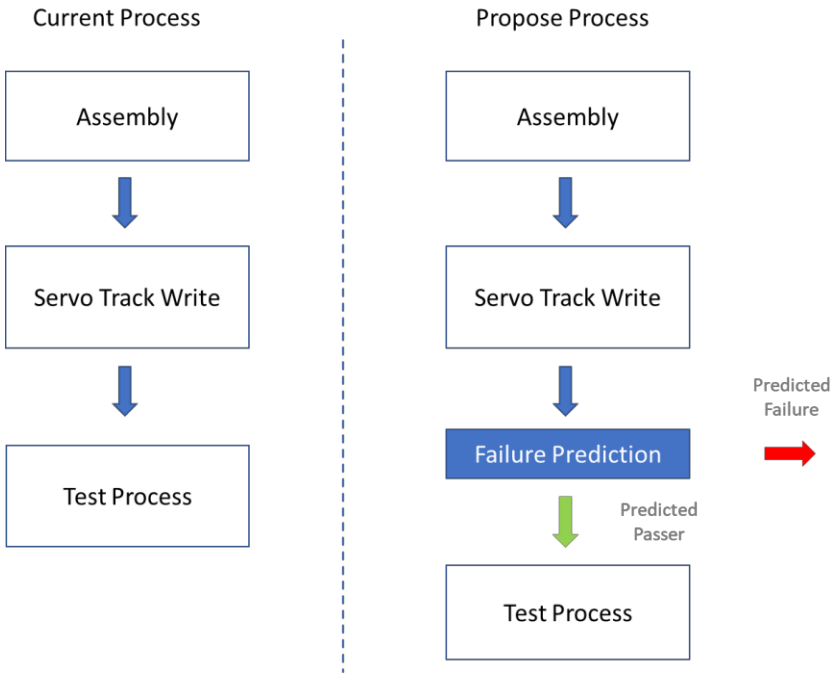
2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

ผิดพลาดว่าเป็นงานเสียด้วยเช่นกัน จากผลการทดลองถ้านำโมเดลที่ให้ประสิทธิภาพดีที่สุด คือ XGB SMOTE ไปใช้งานจะเห็นได้ว่าสามารถตรวจจับงานเสียได้ 53.96% โดยมีค่างานดีเพียง 0.34% ที่ถูกทำนายว่าเป็นงานเสีย จากวิธีการนี้จะสามารถลดค่าใช้จ่ายในการทดสอบงานเสียที่เกิดขึ้นจากปัญหาการอ่านสัญญาณเซอร์โวได้ 53.96% แต่ในขณะเดียวกันก็จะมีค่าใช้จ่ายที่เพิ่มขึ้นเนื่องจากการทำนายงานดีผิดพลาด ซึ่งสามารถนำกราฟของ PRC มาช่วยในการปรับค่าเกณฑ์ของความน่าจะเป็นในการทำนาย ให้ได้ค่าผิดพลาดในการทำนายงานดีว่าเป็นงานเสียมีค่าเข้าใกล้ 0 โดยทำการปรับให้ค่า Recall มีค่าสูงๆ และได้ค่า Precision ที่ดีที่สุด

6.2 ข้อเสนอแนะ

วิธีการที่ศึกษาในงานวิจัยฉบับนี้ สามารถนำไปประยุกต์ใช้ได้โดย

6.2.1 ทำการปรับปรุงโมเดลให้มีประสิทธิภาพในการตรวจจับงานเสียได้ดีขึ้น และลดการทำนายงานดีผิดพลาดว่าเป็นงานเสีย และนำไปใช้ในการทำนายงานเสียจากการอ่านสัญญาณเซอร์โวที่จะเกิดขึ้นในกระบวนการทดสอบ โดยระบบตรวจจับงานเสียจะอยู่ต่อจากกระบวนการเขียนสัญญาณเซอร์โว เพื่อทำนายและนำงานเสียที่ทำนายได้ออกจากกระบวนการผลิต ก่อนที่จะนำงานดีที่ทำนายได้จากโมเดลเข้าสู่กระบวนการทดสอบด้วยเครื่องจักร ดังรูปที่ 28



รูปที่ 28 *การนำระบบตรวจจับงานเสียไปใช้งาน*

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

6.2.2 การนำค่าของพารามิเตอร์ที่มีความสำคัญที่ได้จากการเลือกคุณลักษณะไปวิเคราะห์เพื่อหาสาเหตุของปัญหา รวมถึงการนำพารามิเตอร์ไปใช้ในการทำระบบเฝ้าระวัง

6.2.3 พัฒนาวิธีการเพื่อให้ตรวจจับปัญหาลักษณะเดียวกันบนฮาร์ดดิสก์ไดรฟ์ชนิดอื่น หรือตรวจจับปัญหาอื่นที่เกิดขึ้น โดยทำการเก็บข้อมูลใหม่และเรียนรู้จากชุดข้อมูลใหม่

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

บรรณานุกรม

[1] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," 2016, pp. 785-794.

[2] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014.

[3] A. Hirunyawanakul, N. Kerdprasop, and K. Kerdprasop, "Efficient Machine Learning Methods for Hard Disk Drive Yield Prediction Improvement," *International Journal of Machine Learning and Computing,* vol. 10, no. 2, 2020.

[4] M. Simongyi and P. Chongstitvatana, "Abnormality Detection in Hard Disk Drive Assembly Process Using Support Vector Machine," 2018: IEEE, pp. 612-615.

[5] N. Aussel, S. Jaulin, G. Gandon, Y. Petetin, E. Fazli, and S. Chabridon, "Predictive models of hard drive failures based on operational data," 2017: IEEE, pp. 619-625.

[6] S. M. Djurasevic, U. M. Pesovic, and B. S. Djordjevic, "Anomaly Detection Model for Predicting Hard Disk Drive Failures," *Applied Artificial Intelligence,* vol. 35, no. 8, pp. 549-566, 2021.

[7] F. Gargiulo, D. Duellmann, P. Arpaia, and R. Schiano Lo Moriello, "Predicting Hard Disk Failure by Means of Automatized Labeling and Machine Learning Approach," *Applied Sciences,* vol. 11, no. 18, p. 8293, 2021.

[8] J. Shen, J. Wan, S.-J. Lim, and L. Yu, "Random-forest-based failure prediction for hard disk drives," *International Journal of Distributed Sensor Networks,* vol. 14, no. 11, 2018.

[9] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," 2004: Springer, pp. 39-50.

[10] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 1, pp. 281-288, 2008.

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

[11] A. D. Chakravarthy, S. Bonthu, Z. Chen, and Q. Zhu, "Predictive models with resampling: A comparative study of machine learning algorithms and their performances on handling imbalanced datasets," 2019: IEEE, pp. 1492-1495.

[12] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research,* vol. 2, no. Nov, pp. 45-66, 2001.

[13] H. Ma, L. Wang, and B. Shen, "A new fuzzy support vector machines for class imbalance learning," 2011: IEEE, pp. 3781-3784.

[14] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," 2008, vol. 4: IEEE, pp. 192-201.

[15] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowledge and information systems,* vol. 25, no. 1, pp. 1-20, 2010.

[16] 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.

[17] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," 2005: Springer, pp. 878-887.

[18] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane, and N. Japkowicz, "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance," 2018: IEEE, pp. 447-456.

[19] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences,* 2019.

[20] C. Zhang *et al.*, "Feature selection for high dimensional imbalanced class data based on F-measure optimization," 2017: IEEE, pp. 278-283.

[21] M. Chen, Q. Liu, S. Chen, Y. Liu, C.-H. Zhang, and R. Liu, "XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system," *IEEE Access,* vol. 7, pp. 13149-13158, 2019.

[22] I. L. Cherif and A. Kortebi, "On using extreme gradient boosting (XGBoost) machine learning algorithm for home network traffic classification," 2019: IEEE, pp. 1-6.

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

[23] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," 2013: IEEE, pp. 245-251.

[24] 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.

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11

CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11
2039304467

ประวัติผู้เขียน

ชื่อ-สกุล	Arunee Sridee
วัน เดือน ปี เกิด	13 May 1984
สถานที่เกิด	Bangkok
วุฒิการศึกษา	Bachelor of Engineering, Electrical Engineering, King Mongkut's University of Technology North Bangkok 2006
ที่อยู่ปัจจุบัน	92/24 Moo2 NongTumlung Panthong Chonburi 20160

2039304467 CU iThesis 6270314021 thesis / recv: 18082565 22:48:00 / seq: 11