

การแนะนำการใช้ยาต้านจุลชีพในสัตว์โดยใช้การเรียนรู้ของเครื่อง  
และการแก้ไขปัญหาชุดข้อมูลไม่สมดุล

**ANTIMICROBIAL RECOMMENDATION IN PETS USING  
MACHINE LEARNING AND IMBALANCED DATASET  
PROBLEM SOLVING**

รวิสรา หนูเพ็ง

วัชรินทร์ กันหา

ปริญญาอนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต  
สาขาวิชาจุลทรรศน์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์  
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง  
ปีการศึกษา 2564

ปริญญาบัตรปีการศึกษา 2564

ภาควิชาชีวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง การแนะนำการใช้ยาต้านจุลชีพในสัตว์โดยใช้การเรียนรู้ของเครื่องและการแก้ไขปัญหาชุด

ข้อมูลไม่สมดุล

ANTIMICROBIAL RECOMMENDATION IN PETS USING MACHINE LEARNING  
AND IMBALANCED DATASET PROBLEM SOLVING

ผู้จัดทำ

1. นางสาวริษรา หนูเพ็ง รหัสนักศึกษา 61010888

2. นายวชิรินทร์ กัมaha รหัสนักศึกษา 61010960

อาจารย์ที่ปรึกษา  
(ผศ. ดร. ชุดิเมฆน์ ศรีนิตา)

# การแนะนำการใช้ยาต้านจุลชีพในสัตว์โดยใช้ การเรียนรู้ของเครื่องและการแก้ไขปัญหาชุดข้อมูลไม่สมดุล

นางสาววิสรา หนูเพ็ง 61010888  
นายวชิรินทร์ กันหา 61010960  
ผศ. ดร. ชุดเมฆญ์ ศรีนิลatha อาจารย์ที่ปรึกษา  
ปีการศึกษา 2564

## บทคัดย่อ

ปัจจุบันนี้ได้มีปัญหาการดื้อยาต้านจุลชีพเพิ่มมากขึ้นเรื่อยๆ ซึ่งเป็นภัยคุกคามอย่างมากต่อวงการแพทย์ ดังนั้นการเลือกใช้ยาต้านจุลชีพให้ถูกต้อง และเหมาะสมจะช่วยลดโอกาสการดื้อยาของเชื้อจุลทรรศ์ โดยโครงการนี้ได้ใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการแนะนำยาต้านจุลชีพจากข้อมูลประจำตัวสัตว์ป่วยและผลตรวจจากห้องปฏิบัติการจุลชีววิทยา เพื่อเป็นการช่วยเหลือสัตวแพทย์ในการเลือกใช้ยาต้านจุลชีพที่เหมาะสมได้อย่างรวดเร็วมากยิ่งขึ้น อีกทั้งเพื่อผลสัมฤทธิ์ด้านการรักษา และลดโอกาสการดื้อยาของเชื้อจุลทรรศ์ โครงการนี้ใช้ชุดข้อมูลจากรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยาของห้องปฏิบัติการจุลชีววิทยา ปี พ.ศ. 2559 - 2564 และมีการศึกษาการแก้ไขปัญหาชุดข้อมูลไม่สมดุล (Imbalanced Dataset) เพื่อเพิ่มประสิทธิภาพของโมเดล รวมทั้งการนำโมเดลไปใช้งาน และมีระบบการปรับปรุงโมเดลใหม่ เพื่อให้รองรับกับข้อมูลชุดใหม่ที่เพิ่มเข้ามาในอนาคต ได้อย่างมีประสิทธิภาพ นอกจากนี้ยังได้มีการทดลองฝึกฝนโมเดลทำนายการดื้อยาต้านจุลชีพ และวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่างๆ กับการดื้อยาต้านจุลชีพแต่ละชนิด

# **ANTIMICROBIAL RECOMMENDATION IN PETS**

## **USING MACHINE LEARNING AND IMBALANCED**

### **DATASET PROBLEM SOLVING**

Ms. Rawitsara Nupeng 61010888

Mr. Watcharin Kanha 61010960

Asst. Prof. Dr. Chutimet Srinilta Advisor

Academic Year 2021

#### **ABSTRACT**

The problem of antimicrobial resistance is increasing these days, posing a threat to the medical community. Therefore, the correct and appropriate antimicrobial prescription will reduce the chances of drug resistance of microorganisms. This project uses machine learning techniques to recommend antibiotics based on patient health record and microbiology laboratory test results. This will help veterinarians choose the antimicrobial drugs appropriately for therapeutic results and reduce the chance of drug resistance of microorganisms. We use Antimicrobial Susceptibility Testing (AST) datasets 2016 – 2021 from the Microbiology Laboratory. We study how to solve imbalanced datasets to optimize model performance, including the implementation of the machine learning model and the new model improvement system to effectively support the new dataset added in the future. In addition, we build models to predict antimicrobial resistance and correlation analysis to determine the relationship between each feature and antimicrobial resistance.

## กิตติกรรมประกาศ

ปริญญา尼พนธ์ฉบับนี้สำเร็จลุล่วง ได้ด้วยคำปรีกษา และความช่วยเหลือจากอาจารย์ที่ปรีกษา พศ. ดร.ชุติเมยภู ศรีนิลatha ผู้ที่เคยให้คำแนะนำ เสนอแนะแนวทางในการพัฒนาโครงการ ตลอดจนการแก้ไขข้อบกพร่องต่าง ๆ จนปริญญา尼พนธ์ฉบับนี้เสร็จสมบูรณ์ ทางคณะผู้จัดทำจึงขอกราบขอบพระคุณเป็นอย่างสูง

ขอกราบขอบพระคุณคณะอาจารย์ และบุคลากรต่าง ๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ ที่ให้การอบรมสั่งสอน ให้ความรู้ และคำแนะนำต่าง ๆ มาโดยตลอด

สุดท้ายนี้ ขอกราบขอบพระคุณบุพการี และครอบครัวทุกท่านที่เคยให้การสนับสนุนด้านการศึกษา รวมทั้งเป็นกำลังใจที่ดีเสมอมา

รวิสรา หนูเพ็ง  
วัชรินทร์ กันหา

# สารบัญ

หน้า

บทคัดย่อภาษาไทย .....	I
บทคัดย่อภาษาอังกฤษ .....	II
สารบัญ .....	IV
สารบัญตาราง .....	IX
สารบัญรูปภาพ .....	XII
บทที่ 1 บทนำ .....	1
1.1 ความเป็นมาของปัญหา .....	1
1.2 วัตถุประสงค์ของโครงการ .....	2
1.3 เป้าหมายของโครงการ .....	2
1.4 ขอบเขตของโครงการ .....	3
1.5 แผนการดำเนินงาน .....	3
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง .....	6
2.1 ทฤษฎีที่เกี่ยวข้อง .....	6
2.1.1 ยาต้านจุลชีพ .....	6
2.1.2 การเรียนรู้ของเครื่อง (Machine Learning) .....	6
2.1.3 เค�플ดครอสเฟลิดชัน (K-Fold Cross-Validation) .....	7
2.1.4 ต้นไม้ตัดสินใจ (Decision Tree) .....	8
2.1.5 แรนดอมฟอร์สต์ (Random Forest) .....	8
2.1.6 เกรดีบันบูสติ่ง (Gradient Boosting) .....	9
2.1.7 ซัพพอร์ตเวกเตอร์แมชีน (Support Vector Machine: SVM) .....	9
2.1.8 ชุดข้อมูลไม่สมดุล (Imbalanced Dataset) .....	10

# สารบัญ (ต่อ)

หน้า

2.1.9 ประเภทของการแก้ปัญหาชุดข้อมูลไม่สมดุล .....	11
2.1.10 การวัดประสิทธิภาพของโมเดล.....	12
2.1.11 สัมประสิทธิ์ครามเมอร์ (Cramer's V Coefficient).....	13
2.2 งานวิจัยที่เกี่ยวข้อง .....	14
2.2.1 Using Machine Learning Techniques to Aid Empirical Antibiotic Therapy Decisions in the Intensive Care Unit of a General Hospital in Greece.....	14
2.2.2 Machine Learning Techniques to Identify Antimicrobial Resistance in the Intensive Care Unit .....	14
2.2.3 Machine Learning For Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy.....	15
2.2.4 Machine Learning Techniques to Identify Antibiotic Resistance in Patients Diagnosed with Various Skin and Soft Tissue Infections .....	16
2.2.5 Antimicrobial Resistance and Machine Learning: Challenges and Opportunities .....	16
2.2.6 SMOTE: Synthetic Minority Oversampling Technique.....	16
2.2.7 Borderline-SMOTE .....	17
2.2.8 ADASYN.....	17
2.2.9 Borderline Over-sampling for Imbalanced Data Classification .....	18
2.2.10 RSMOTE .....	18
2.3 เครื่องมือที่เกี่ยวข้อง .....	19
2.3.1 Python .....	19
2.3.2 Numpy .....	19
2.3.3 Pandas .....	19
2.3.4 Scikit-learn.....	19

# สารบัญ (ต่อ)

หน้า	
2.3.5 Imbalanced-learn .....	20
2.3.6 Extreme Gradient Boosting (XGBoost).....	20
2.3.7 Vue.js .....	20
2.3.8 FastAPI .....	20
2.3.9 PostgreSQL.....	20
2.3.10 Python-docx .....	20
 บทที่ 3 การออกแบบ และการพัฒนา.....	21
3.1 การออกแบบโมเดลแนะนำภาษาต้านจุลชีพ .....	21
3.2 การออกแบบขั้นตอนพัฒนาโมเดลแนะนำภาษาต้านจุลชีพ .....	22
3.2.1 การจัดเตรียมชุดข้อมูล (Dataset) .....	22
3.2.2 การทำความสะอาดชุดข้อมูล (Data Cleansing).....	22
3.2.3 การเตรียมข้อมูลให้พร้อมนำไปใช้งาน (Data Preprocessing) .....	23
3.2.4 การแบ่งชุดข้อมูล (Training Set and Test Set) .....	23
3.2.5 การแก้ไขปัญหาชุดข้อมูลไม่สมดุล .....	24
3.2.6 การฝึกฝนโมเดล .....	24
3.2.7 การวัดประสิทธิภาพของโมเดล.....	24
3.3 การออกแบบการทดลอง .....	26
3.3.1 ตอนที่ 1 การฝึกฝน โมเดลแนะนำภาษาต้านจุลชีพ .....	26
3.3.2 ตอนที่ 2 การฝึกฝน โมเดลสำหรับการคือยาต้านจุลชีพ.....	27
3.4 รายละเอียดของชุดข้อมูลที่ใช้ .....	28
3.4.1 ชุดข้อมูลแบบที่เรียกว่าบวกและชุดข้อมูลแบบที่เรียกว่าลบ โดยแบ่งตาม ชนิดของการคิดที่ใช้ตรวจสอบผลการทดสอบความไวต่อยาต้านจุลชีพด้วยเครื่อง Vitek (Vitek ID Card) .....	29
3.4.2 ชนิดของสัตว์เลี้ยง (Species) .....	30

# สารบัญ (ต่อ)

หน้า

3.4.3  สกุลแบคทีเรีย (Bacteria Genus) .....	31
3.4.4  ตัวอย่างที่ส่งตรวจ (Submitted Sample).....	32
3.4.5  ผลการทดสอบความไวต่อยาต้านจุลชีพ (Result of Antimicrobial Susceptibility Testing).....	33
3.4.6  เยต้านจุลชีพที่ได้สร้างโดยเดลเพื่อแนะนำ (Antibiotic) .....	34
3.5  ภาพรวมของระบบ .....	35
3.5.1 การแนะนำการใช้ยาต้านจุลชีพ .....	35
3.5.2 การอัปโหลดไฟล์เข้าสู่ระบบเพื่อฝึกฝนโดยเดล .....	35
3.5.3 การฝึกฝนโดยเดลใหม่เพื่ออัปเดตโดยเดล .....	35
3.6  การออกแบบและพัฒนาระบบ .....	37
3.6.1 การออกแบบ Use Case Diagram .....	37
3.6.2 การออกแบบ Sequence Diagram.....	38
3.6.3 การออกแบบ ER Diagram .....	43
3.6.4 การออกแบบส่วนติดต่อผู้ใช้ (User Interface) .....	44
บทที่ 4 การทดลอง และผลการทดลอง .....	51
4.1 การทดลองที่ 1.1 การเปรียบเทียบประสิทธิภาพของโดยเดลที่ใช้เทคนิคการจำแนกช้อมูลแบบต่าง ๆ โดยใช้ชุดช้อมูลแบคทีเรียแกรมบวก และชุดช้อมูลแบคทีเรียแกรมลบ .....	51
4.2 การทดลองที่ 1.2 การเปรียบเทียบเทคนิคการสังเคราะห์ช้อมูล (Oversampling) ในการเก็บปัญหาชุดช้อมูลไม่สมดุล โดยใช้ชุดช้อมูลแบคทีเรียแกรมบวกและแบคทีเรียแกรมลบ .....	58
4.3 การทดลองที่ 1.3 การเปรียบเทียบเทคนิคการสังเคราะห์ช้อมูล (Oversampling) ในการเก็บปัญหาชุดช้อมูลไม่สมดุล โดยใช้ชุดช้อมูลที่มาจาก UCI Machine Learning Repository.....	71

## สารบัญ (ต่อ)

หน้า

4.4 การทดลองที่ 1.4 การปรับพารามิเตอร์ของโ้มเดลกลุ่มแบบที่เรียกว่าแบบโ้มเดลกลุ่มแบบที่เรียกว่าแบบกลุ่มแบบ .....78
4.5 สรุปผลการสร้างโ้มเดลแนะนำคำอ่านจุลชีพ.....82
4.6 การทดลองที่ 2.1 การทำนายการดื้อยาต้านจุลชีพของแบบที่เรียกว่าแบบ .....84
4.7 การทดลองที่ 2.2 การทำนายการดื้อยาต้านจุลชีพของแบบที่เรียกว่าแบบ .....89
บทที่ 5 การสรุปผล และข้อเสนอแนะ .....95
5.1 สรุปผลที่ได้จากการวิเคราะห์ .....95
5.2 ปัญหาและอุปสรรค.....95
5.3 แนวทางในการพัฒนาต่อ .....96
บรรณานุกรม .....97
ภาคผนวก ก .....99

# สารบัญตาราง

ตาราง	หน้า
1.1 แผนการดำเนินงานในภาคเรียนที่ 1 .....	3
1.2 แผนการดำเนินงานในภาคเรียนที่ 2 .....	5
4.1 การเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า Accuracy, Precision, Recall และ F1-Score ของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	52
4.2 การเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า Accuracy, Precision, Recall และ F1-Score ของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	54
4.3 การจัดอันดับประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า F1-Score ของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	56
4.4 การจัดอันดับประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า F1-Score ของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	56
4.5 อัตราความไม่สมดุลของชุดข้อมูลฝึกฝนของแบนค์ที่เรียกว่า Cross-Validation) .....	59
4.6 อัตราความไม่สมดุลของชุดข้อมูลฝึกฝนของแบนค์ที่เรียกว่า Cross-Validation) .....	60
4.7 การเปรียบเทียบประสิทธิภาพด้วยค่า Precision ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	61
4.8 การเปรียบเทียบประสิทธิภาพด้วยค่า Recall ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	62
4.9 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	63
4.10 การเปรียบเทียบประสิทธิภาพด้วยค่า Precision ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	64
4.11 การเปรียบเทียบประสิทธิภาพด้วยค่า Recall ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบนค์ที่เรียกว่า Cross-Validation) .....	65

## สารบัญตาราง (ต่อ)

ตาราง	หน้า
4.12 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบบที่เรียกว่า CV (10-Fold Cross-Validation) .....	66
4.13 ประสิทธิภาพของโมเดลที่ใช้เทคนิคการสังเคราะห์ข้อมูลที่เหมาะสม โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบบที่เรียกว่า CV (10-Fold Cross-Validation) .....	70
4.14 ประสิทธิภาพของโมเดลที่ใช้เทคนิคการสังเคราะห์ข้อมูลที่เหมาะสม โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบบที่เรียกว่า CV (10-Fold Cross-Validation) .....	70
4.15 ชุดข้อมูลที่ใช้ในการทดลอง .....	72
4.16 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า Precision โดยใช้ชุดข้อมูลทั่วไป (10-Fold Cross-Validation) .....	73
4.17 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า Recall โดยใช้ชุดข้อมูลทั่วไป (10-Fold Cross-Validation) .....	74
4.18 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้ชุดข้อมูลทั่วไป (10-Fold Cross-Validation) .....	75
4.19 การเปรียบเทียบประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score ของโมเดล XGBoost ก่อน-หลังปรับจูนพารามิเตอร์ของชุดข้อมูลแบบที่เรียกว่า CV .....	79
4.20 การเปรียบเทียบประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score ของโมเดล XGBoost ก่อน-หลังปรับจูนพารามิเตอร์ของชุดข้อมูลแบบที่เรียกว่า CV .....	80
4.21 การวัดประสิทธิภาพของโมเดลกลุ่มแบบที่เรียกว่า CV ด้วยค่า Accuracy, Precision, Recall และ F1-Score .....	82
4.22 การวัดประสิทธิภาพของโมเดลกลุ่มแบบที่เรียกว่า CV ด้วยค่า Accuracy, Precision, Recall และ F1-Score .....	83
4.23 การวัดประสิทธิภาพของโมเดลแบบทดสอบรายเคส (Test By Case) ด้วยค่า Accuracy, Precision, Recall และ F1-Score .....	83
4.24 ข้อมูลการดีอิยาต้านจุลชีพของแบบที่เรียกว่า CV .....	86

## สารบัญตาราง (ต่อ)

ตาราง	หน้า
4.25 ผลการวัดประสิทธิภาพของโมเดลทำการคือยาต้านจุลชีพในกลุ่มแบคทีเรียแกรมบวกด้วยค่า Accuracy, Precision, Recall และ F1-Score .....	87
4.26 ข้อมูลการคือยาต้านจุลชีพของแบคทีเรียแกรมลบ .....	91
4.27 ผลการวัดประสิทธิภาพของโมเดลทำการคือยาต้านจุลชีพในกลุ่มแบคทีเรียแกรมลบด้วยค่า Accuracy, Precision, Recall และ F1-Score .....	92

# สารบัญรูปภาพ

รูป	หน้า
2.1 การทดสอบประสิทธิภาพของโมเดลแบบ K-Fold Cross-Validation .....	7
2.2 ต้นไม้ตัดสินใจ .....	8
2.3 แรนdomฟอร์ेसต์ .....	8
2.4 เกรเดียนบูสติ้ง .....	9
2.5 ชัฟฟอร์ตเวกเตอร์แมชชีน .....	10
2.6 ตัวอย่างชุดข้อมูลสมดุลและชุดข้อมูลไม่สมดุล .....	10
2.7 กอนฟิวชันเมทริกซ์ .....	12
2.8 การถังเคราะห์แบบขยาย (a) และการถังเคราะห์แบบแทรก (b) .....	18
3.1 การออกแบบโมเดลแนะนำตัวนจุลชีพ .....	22
3.2 วิธีการเข้ารหัสแบบวันซอต .....	23
3.3 ตารางผลลัพธ์จากการทดสอบ .....	25
3.4 ตัวอย่างชุดข้อมูลที่ใช้ .....	29
3.5 แผนภูมิแท่งแสดงจำนวนชุดข้อมูลแบคทีเรียแกรมบวกและแบคทีเรียแกรมลบ .....	29
3.6 แผนภูมิแท่งแสดงจำนวนข้อมูลชนิดของสัตว์เลี้ยง .....	30
3.7 แผนภูมิแท่งแสดงจำนวนข้อมูลสกุลแบคทีเรีย 10 อันดับแรก .....	31
3.8 แผนภูมิแท่งแสดงจำนวนข้อมูลตัวอย่างที่ส่งตรวจ 10 อันดับแรก .....	32
3.9 แผนภูมิแท่งแสดงจำนวนข้อมูลผลการทดสอบต่อイヤต้านจุลชีพแต่ละชนิด .....	33
3.10 แผนภูมิแท่งแสดงจำนวนครั้งที่มีการใช้ยาต้านจุลชีพ .....	34
3.11 ภาพรวมของระบบ .....	36
3.12 Use Case Diagram .....	37
3.13 Sequence Diagram ของการแนะนำตัวนจุลชีพ .....	38
3.14 Sequence Diagram ของการอัปโหลดไฟล์เพื่อฝึกฝนโมเดลใหม่ .....	38
3.15 Sequence Diagram ของฝึกฝนโมเดลใหม่ .....	39
3.16 Sequence Diagram ของการยกเลิกการฝึกฝนโมเดล .....	39
3.17 Sequence Diagram ของการตรวจสอบการยกเลิกการฝึกฝนโมเดล .....	40

# สารบัญรูปภาพ (ต่อ)

รูป	หน้า
3.18 Sequence Diagram ของการคุ้นไฟล์ชุดข้อมูลสำหรับฝึกฝนโมเดลที่มีอยู่ในระบบ .....	40
3.19 Sequence Diagram ของการลบไฟล์ชุดข้อมูลสำหรับฝึกฝนโมเดล .....	40
3.20 Sequence Diagram ของการคุ้นข้อมูลบันทึกการอัปโหลดไฟล์ .....	41
3.21 Sequence Diagram ของการคุ้นข้อมูลบันทึกการฝึกฝนโมเดล .....	41
3.22 Sequence Diagram ของการคุ้นเดชนอร์ด .....	41
3.23 Sequence Diagram ของการคุ้นการกำหนดค่าโมเดล .....	42
3.24 ER Diagram.....	43
3.25 หน้าแรกของเว็บแอปพลิเคชัน .....	44
3.26 หน้าแนะนำต้านจุลเชื้อ .....	44
3.27 หน้าอัปโหลดไฟล์สำหรับฝึกฝนโมเดลใหม่ .....	45
3.28 ป้อนอัปแสดงผลลัพธ์การอัปโหลดไฟล์สำเร็จ .....	46
3.29 ป้อนอัปแสดงผลลัพธ์การอัปโหลดไฟล์ไม่สำเร็จ .....	46
3.30 หน้าจัดการไฟล์และฝึกฝนโมเดล .....	46
3.31 หน้าแสดงข้อมูลบันทึกการฝึกฝนโมเดล .....	47
3.32 ป้อนอัปแสดงรายละเอียดของไฟล์ที่ใช้ในการฝึกฝนโมเดลใหม่ .....	47
3.33 หน้าเดชนอร์ดแสดงประสิทธิภาพของโมเดลในแต่ละเวอร์ชัน .....	48
3.34 หน้าเดชนอร์ดแสดงชุดข้อมูลที่ใช้ในการฝึกฝนโมเดล .....	49
3.35 หน้าเดชนอร์ดแสดงประสิทธิภาพของโมเดลยาแต่ละชนิด .....	49
3.36 หน้าแสดงการกำหนดค่าของโมเดล .....	50
4.1 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบคทีเรียแกรมบวก .....	67
4.2 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้เทคนิคการจำแนกข้อมูลด้วย SVM ของชุดข้อมูลแบคทีเรียแกรมบวก .....	67
4.3 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบคทีเรียแกรมลบ .....	68

# สารบัญรูปภาพ (ต่อ)

รูป	หน้า
4.4 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้เทคนิคการจำแนกข้อมูลด้วย SVM ของชุดข้อมูลแบนค์ที่เรียกว่า XGBoost .....	68
4.5 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost .....	76
4.6 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้เทคนิคการจำแนกข้อมูลด้วย SVM .....	76
4.7 ค่าสัมประสิทธิ์ความอิรักระหว่างฟีเจอร์ต่าง ๆ กับการดึงยาต้านจุลชีพในกลุ่มแบนค์ที่เรียกว่า บวก .....	86
4.8 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของโมเดลกลุ่มแบนค์ที่เรียกว่า บวก ที่ไม่มีสังเคราะห์ข้อมูลกับโมเดลที่มีการสังเคราะห์ข้อมูลพร้อมกับการปรับจูน พารามิเตอร์ของ XGBoost.....	87
4.9 ค่าสัมประสิทธิ์ความอิรักระหว่างฟีเจอร์ต่าง ๆ กับการดึงยาต้านจุลชีพในกลุ่มแบนค์ที่เรียกว่า ลบ .....	91
4.10 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของโมเดลกลุ่มแบนค์ที่เรียกว่า บวก ที่ไม่มีสังเคราะห์ข้อมูลกับโมเดลที่มีการสังเคราะห์ข้อมูลพร้อมกับการปรับจูน พารามิเตอร์ของ XGBoost.....	92
ก.1 แผนภาพการกระจายของชุดข้อมูลแบนค์ที่เรียกว่า บวก .....	99
ก.2 แผนภาพการกระจายของชุดข้อมูลแบนค์ที่เรียกว่า ลบ .....	100
ก.3 แผนภาพการกระจายของชุดข้อมูลที่ใช้ในการทดลองที่ 1.3 .....	101

## บทที่ 1

### บทนำ

#### 1.1 ความเป็นมาของปัญหา

ยาต้านจุลชีพเป็นยาที่ใช้ในการรักษาโรคติดเชื้อจุลทรรศ์ ซึ่งออกฤทธิ์ฆ่าหรือยับยั้งการเจริญเติบโตของจุลทรรศ์ แต่ปัจจุบันนี้ได้มีปัญหาการดื้อยาต้านจุลชีพเพิ่มมากขึ้นเรื่อยๆ ซึ่งเป็นภัยคุกคามอย่างมากต่อวงการแพทย์ คนและสัตว์ที่มีภาวะติดเชื้อดื้อยาจะมีอัตราการป่วยและเสียชีวิตในระดับที่สูงกว่าการติดเชื้อปกติ ทำให้การรักษาอาจไม่ตอบสนองต่อยาต้านจุลชีพพื้นฐานที่มีอยู่ในโรงพยาบาล สาเหตุของการดื้อยามีหลายประการ เช่น การเลือกใช้ยาต้านจุลชีพที่ไม่เหมาะสมต่อเชื้อ การใช้ยาต้านจุลชีพที่มากเกินความจำเป็น เป็นต้น ทำให้จุลทรรศ์มีวิวัฒนาการของการดื้อยาเพิ่มมากขึ้นเรื่อยๆ ดังนั้นจึงการเลือกใช้ยาต้านจุลชีพให้ถูกต้องและเหมาะสมเพื่อลดโอกาสการดื้อยาของเชื้อจุลทรรศ์ โดยทั่วไปการเลือกใช้ยาต้านจุลชีพที่เหมาะสม สัตวแพทย์จะพิจารณาจากผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยา (Result of Antimicrobial Susceptibility Test: AST) ซึ่งเป็นการตรวจหายาต้านจุลชีพที่มีความไวต่อการตอบสนองต่อเชื้อจุลทรรศ์ที่ก่อให้เกิดโรค และตรวจหายาต้านจุลชีพที่ดื้อต่อการรักษา

การช่วยเหลือสัตวแพทย์ในการเลือกใช้ยาต้านจุลชีพที่เหมาะสม ได้อย่างรวดเร็วมากขึ้นนี้ สามารถใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อแนะนำการใช้ยาต้านจุลชีพที่เหมาะสมได้ โดยใช้ชุดข้อมูลจากการรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยาของห้องปฏิบัติการจุลชีววิทยา แต่ชุดข้อมูลนี้พบปัญหาความไม่สมดุลของชุดข้อมูล (Imbalanced Dataset) ซึ่งเป็นปัญหาที่ชุดข้อมูลนี้มีจำนวนข้อมูลของกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งเป็นจำนวนมาก ทำให้การทำนายข้อมูลที่อยู่ในกลุ่มน้อยไม่สามารถทำนายได้อย่างมีประสิทธิภาพ ซึ่งปัญหานี้ไม่สมดุลของชุดข้อมูลนี้พบได้บ่อยครั้งในการนำชุดข้อมูลทางการแพทย์มาใช้งาน

จากปัญหาดังกล่าว ทางผู้จัดทำจึงได้นำเทคนิคการเรียนรู้ของเครื่องมาใช้ในการสร้างโมเดลแนะนำยาต้านจุลชีพ เพื่อเป็นการช่วยเหลือสัตวแพทย์ในการเลือกใช้ยาต้านจุลชีพที่เหมาะสม ได้อย่างรวดเร็วมากขึ้น อีกทั้งเพื่อผลสัมฤทธิ์ด้านการรักษา และลดโอกาสการดื้อยาของเชื้อจุลทรรศ์ โดยใช้ชุดข้อมูลจากการรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยาของห้องปฏิบัติการจุลชีววิทยาปี พ.ศ. 2559 - 2564 และมีการศึกษาการแก้ไขปัญหาชุดข้อมูลไม่สมดุลเพื่อเพิ่มประสิทธิภาพของโมเดล รวมทั้งการนำโมเดลไปใช้งาน และมีระบบปรับปรุงโมเดล

ใหม่เพื่อให้รองรับกับข้อมูลชุดใหม่ที่เพิ่มเข้ามาในอนาคต ได้อ้างมีประสิทธิภาพ นอกจากนี้ยังได้มีการทดลองสร้างโมเดลนำ�性การดื้อยาต้านจุลชีพ และวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิด

## 1.2 วัตถุประสงค์ของโครงการ

- 1) เพื่อสร้างโมเดลแนะนำการใช้ยาต้านจุลชีพที่เหมาะสมตามคำแนะนำของสัตวแพทย์
- 2) เพื่อเปรียบเทียบประสิทธิภาพของโมเดลที่ได้จากการตัดไม้ตัดสินใจ (Decision Tree), แรนดอมฟอร์สต์ (Random Forest), เอ็กซ์ตรีมเกรดบูสติ้ง (XGBoost) และซัพพอร์ตเวกเตอร์แมชีน (SVM)
- 3) เพื่อศึกษาวิธีการแก้ปัญหาชุดข้อมูลไม่สมดุล
- 4) เพื่อเปรียบเทียบทكنิคการสังเคราะห์ข้อมูล (Oversampling) ในการแก้ไขปัญหาชุดข้อมูลไม่สมดุล
- 5) เพื่อพัฒนาเว็บแอปพลิเคชันสำหรับทดสอบการใช้งานโมเดลแนะนำยาต้านจุลชีพ และปรับปรุงประสิทธิภาพของโมเดลให้รองรับกับข้อมูลชุดใหม่ที่เพิ่มเข้ามาในอนาคต
- 6) เพื่อสร้างโมเดลนำ�性การดื้อยาต้านจุลชีพ และวิเคราะห์ปัจจัยที่มีผลต่อการดื้อยา

## 1.3 เป้าหมายของโครงการ

- 1) สามารถสร้างโมเดลเพื่อช่วยเหลือสัตวแพทย์ในการแนะนำการใช้ยาต้านจุลชีพที่เหมาะสม และรวดเร็วมากยิ่งขึ้นได้
- 2) สามารถวิเคราะห์การแก้ไขปัญหาชุดข้อมูลไม่สมดุลได้
- 3) สามารถสร้างส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interface: API) สำหรับเรียกใช้งานโมเดลเพื่อแนะนำยาต้านจุลชีพได้
- 4) สามารถปรับปรุงประสิทธิภาพของโมเดลให้รองรับกับข้อมูลชุดใหม่ที่เพิ่มเข้ามาในอนาคตได้

#### 1.4 ขอบเขตของโครงการ

- 1) ใช้ชุดข้อมูลจากการรายงานผลการระบุชนิดจุลชีพและการทดสอบความไว้วับต่อของห้องปฏิบัติการจุลชีววิทยาปี พ.ศ. 2559 - 2564 ของโรงพยาบาลสัตว์เล็ก คณะสัตวแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เพื่อสร้างโมเดลแนวนำมายาต้านจุลชีพ
  - 2) ใช้เทคนิคการจำแนกประเภทข้อมูล (Classification) 4 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ, แรนdomฟอร์เรสต์, เอ็กซ์ตรีมเกรเดียนบูสติ้ง, ชัฟฟอร์ตเวกเตอร์แมชชีน
  - 3) ใช้เทคนิคการสังเคราะห์ข้อมูล 4 เทคนิคในการแก้ปัญหาชุดข้อมูลไม่สมดุล ได้แก่ SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN และ R-SMOTE
  - 4) ใช้เครื่องมือ Vue.js ในการพัฒนาระบบหน้าบ้าน (Frontend)
  - 5) ใช้เครื่องมือ FastAPI ในการพัฒนาระบบหลังบ้าน (Backend)

## 1.5 แผนการดำเนินงาน

#### ตาราง 1.1 แผนการดำเนินงานในภาคเรียนที่ 1

6) ศึกษางานวิจัยที่เกี่ยวข้อง									
ส่วนการออกแบบ									
7) ออกแบบการจัดเตรียมข้อมูล									
8) ออกแบบการทดสอบและวัดประสิทธิภาพของโมเดล									
9) ออกแบบการเลือกอัลกอริทึมและโมเดล									
10) ออกแบบ Use Case									
11) ออกแบบ Sequence Diagram									
12) ออกแบบส่วนติดต่อผู้ใช้									
ส่วนการทดลอง									
13) ทดลองสร้างโมเดล									
14) ทดลองใช้ SMOTE									
ส่วนการ Implement									
15) แปลงชุดข้อมูลจากไฟล์ .docx เป็น .csv									
16) ทำความสะอาดข้อมูล									
17) วิเคราะห์ข้อมูลเชิงสำรวจ									
18) กระบวนการแปลงข้อมูล									
19) เลือกอัลกอริทึมและสร้างโมเดล									
20) จัดการชุดข้อมูลไม่สมดุลโดยใช้เทคนิค SMOTE แบบต่างๆ									
21) ทดสอบและวัดประสิทธิภาพของโมเดล									
ส่วนการสรุปผล									
22) สรุปผลและจัดทำรายงาน									

ตาราง 1.2 แผนการดำเนินงานในภาคเรียนที่ 2

รายการ	ภาคเรียนที่ 2															
	มกราคม				กุมภาพันธ์				มีนาคม				เมษายน			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
<b>ส่วนการออกแบบ</b>																
1) ออกแบบ ER Diagram	◀	▶														
2) ออกแบบ API	◀	▶														
<b>ส่วนการพัฒนาเว็บ</b>	◀								▶							
3) ศึกษาเครื่องมือพัฒนาเว็บ	↔															
4) พัฒนาระบบหน้าบ้าน		◀				▶										
5) พัฒนาระบบหลังบ้าน		◀			▶											
6) ทดสอบเว็บ								↔								
<b>ส่วนการปรับปรุงโมเดล</b>								◀	▶							
7) พัฒนาระบบปรับปรุง ประสิทธิภาพ โมเดล								◀	▶							
<b>ส่วนการ Deployment</b>									◀	▶						
8) ทำการ Deploy เว็บแอป- พลิกซัน								◀	▶							
9) จัดทำเอกสาร Deployment									↔							
<b>ส่วนการสรุปผล</b>										◀		▶				
10) สรุปผลและจัดทำเล่ม โครงการ									◀			▶				

## บทที่ 2

# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

### 2.1 ทฤษฎีที่เกี่ยวข้อง

#### 2.1.1 ยาต้านจุลชีพ

ยาต้านจุลชีพ เป็นยาที่ใช้ในการรักษาโรคติดเชื้อ โดยออกฤทธิ์ฆ่าหรือยับยั้งการเจริญเติบโตของเชื้อจุลินทรีย์ เช่น แบคทีเรีย เชื้อรา หรือไวรัส โดยการใช้ยาต้านจุลชีพเพื่อรักษาโรคติดเชื้อมีปัจจัยที่ต้องคำนึง ได้แก่ ความไวของเชื้อ ระดับยาในร่างกาย และอันตรายหรือฤทธิ์อันพึงประสงค์ของยาที่มีต่อร่างกาย ดังนั้นการเลือกใช้ยาต้านจุลชีพจะต้องทราบถึงความไวของเชื้อแบคทีเรียต่อยาต้านจุลชีพ

การทดสอบความไวต่อยาต้านจุลชีพ (Antimicrobial Susceptibility Testing) เป็นการทดสอบเพื่อตรวจหายาที่มีความไวในการตอบสนองต่อการรักษาโรคติดเชื้อ และตรวจหายาที่ต้องต่อการรักษา เพื่อเป็นแนวทางในการเลือกใช้ยาต้านจุลชีพให้เหมาะสม โดยผลการทดสอบความไวต่อยาต้านจุลชีพ จะมีผลการทดสอบ 3 ประเภท ดังนี้

- 1) Susceptible/Sensitive (S) หมายถึง ยาที่ใช้ขนาดหรือความเข้มข้นปกติแล้วมีประสิทธิภาพในการรักษาโรคติดเชื้อได้
- 2) Intermediate (I) หมายถึง ยาที่มีโอกาสใช้ในการรักษาโรคติดเชื้อได้ถ้าใช้ขนาดสูง และมีแนวโน้มจะมีประสิทธิภาพต่ำกว่ายาที่มีผลการทดสอบเป็น Susceptible
- 3) Resistant (R) หมายถึง ยาที่ไม่สามารถใช้รักษาโรคติดเชื้อได้

#### 2.1.2 การเรียนรู้ของเครื่อง (Machine Learning)

เป็นการเรียนรู้อัลกอริทึมของคอมพิวเตอร์ โดยมีการปรับปรุงผ่านประสบการณ์และข้อมูล ซึ่งเป็นสาขานึงของปัญญาประดิษฐ์ อัลกอริทึมการเรียนรู้ของเครื่องจะสร้างโมเดลจากข้อมูลตัวอย่างเพื่อทำการคาดการณ์หรือตัดสินใจผลลัพธ์โดยไม่ใช้การโปรแกรมแบบทั่วไป ซึ่งการเรียนรู้ของเครื่องนี้แบ่งออกเป็น 3 ประเภท ดังนี้

### 2.1.2.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning)

เป็นการสอนคอมพิวเตอร์โดยมีตัวอย่างข้อมูลมาเข้า (Input) และข้อมูลมาออก (Output) ที่ต้องการให้คอมพิวเตอร์ทำการสร้างฟังก์ชันที่เกี่ยวข้องระหว่างข้อมูลมาเข้าและข้อมูลมาออก ตัวอย่างการเรียนรู้แบบมีผู้สอน เช่น การ回帰 (Regression), ต้นไม้ตัดสินใจ (Decision Tree), ซัพพอร์ตเวกเตอร์แมชีน (Support Vector Machine) เป็นต้น

### 2.1.2.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

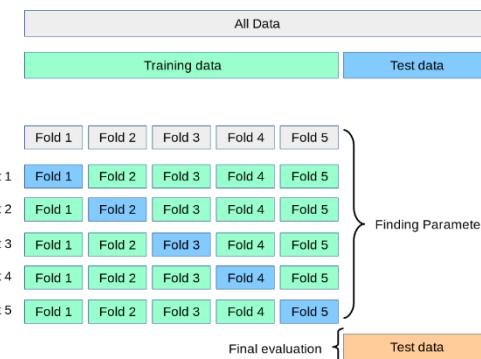
เป็นการสอนคอมพิวเตอร์โดยไม่มีการกำหนดป้ายชื่อให้กับอัลกอริทึม คอมพิวเตอร์จะทำการเรียนรู้โครงสร้างของข้อมูลมาเข้า (Input) โดยการหารูปแบบของข้อมูลและทำการจัดกลุ่มของข้อมูลด้วยตัวเอง ตัวอย่างการเรียนรู้แบบไม่มีผู้สอน เช่น การแบ่งกลุ่มข้อมูลแบบเคมิน (K-Means Clustering), ดิบีสแกน (DBSCAN) เป็นต้น

### 2.1.2.3 การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning)

เป็นการที่คอมพิวเตอร์ได้ตอบกับสภาพแวดล้อมซึ่งจะต้องทำตามเป้าหมายโดยการลองผิดลองถูก ถ้าทำถูกจะได้รางวัล (Reward) ตัวอย่างการเรียนรู้แบบนี้ เช่น กระบวนการตัดสินใจของมาร์คอฟ (Markov Decision Process) เป็นต้น

### 2.1.3 เคลฟoldคross-validation (K-Fold Cross-Validation)

เป็นวิธีการแบ่งข้อมูลรูปแบบหนึ่งเพื่อสำหรับทดสอบประสิทธิภาพของโมเดลที่สร้างขึ้น โดยจะแบ่งข้อมูลออกเป็น K ส่วนเท่า ๆ กัน และใช้ชุดข้อมูล K – 1 ชุดสำหรับฝึกฝนโมเดล ส่วนชุดข้อมูลที่เหลืออีก 1 ชุด จะใช้เป็นชุดข้อมูลทดสอบประสิทธิภาพของโมเดล จากนั้นก็ทำการสลับชุดข้อมูลที่ใช้สำหรับฝึกฝนและทดสอบจนครบ K ชุด

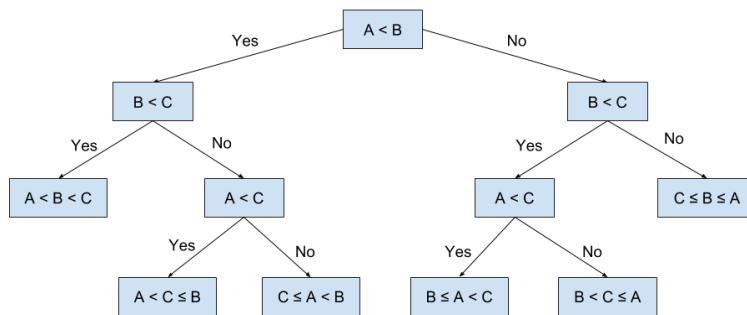


รูปที่ 2.1 การทดสอบประสิทธิภาพของโมเดลแบบ K-Fold Cross-Validation

(ที่มา: [https://scikit-learn.org/stable/\\_images/grid\\_search\\_cross\\_validation.png](https://scikit-learn.org/stable/_images/grid_search_cross_validation.png))

#### 2.1.4 ต้นไม้ตัดสินใจ (Decision Tree)

เป็นเครื่องมือที่ช่วยสนับสนุนการตัดสินใจโดยใช้โมเดลแบบต้นไม้ในการแสดงทุกความเป็นไปได้ของการตัดสินใจรวมถึงโอกาสของเหตุการณ์ และอื่น ๆ โดยเป็นการแสดงอัลกอริทึมที่มีเฉพาะคำสั่งแบบมีเงื่อนไข

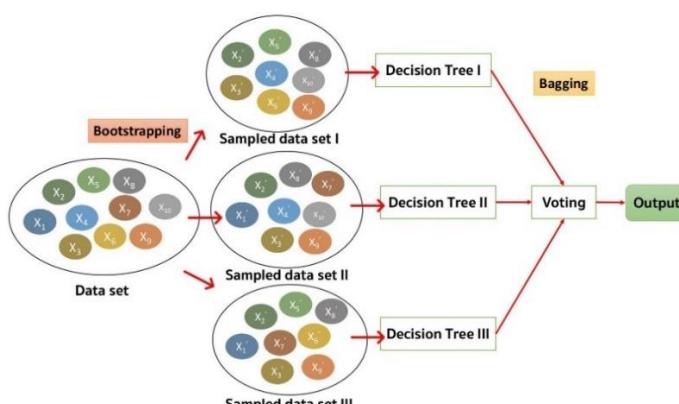


รูปที่ 2.2 ต้นไม้ตัดสินใจ

(ที่มา: [https://miro.medium.com/max/1400/1\\*aapKldn9D21nvuAekT7wKQ.png](https://miro.medium.com/max/1400/1*aapKldn9D21nvuAekT7wKQ.png))

#### 2.1.5 แรนdomฟอร์ेस्ट (Random Forest)

เป็นโมเดลที่ถูกพัฒนาขึ้นมาจากต้นไม้ตัดสินใจ หลักการคือการสร้างโมเดลจากต้นไม้ตัดสินใจหลาย ๆ โมเดล โดยแต่ละโมเดลจะเลือกฟีเจอร์และส่วนของข้อมูลสำหรับฝึกฝนที่ไม่เหมือนกัน แล้วคำนวณผลลัพธ์จากการให้คะแนนของโมเดลว่าคลาสใดถูกเลือกมากที่สุด ทำให้มีประสิทธิภาพการทำงานสูง และมีความแม่นยำมากขึ้น

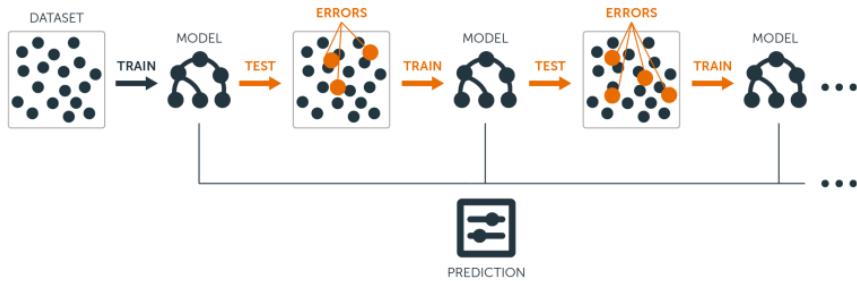


รูปที่ 2.3 แรนdomฟอร์ेस्ट

(ที่มา: [https://miro.medium.com/max/1400/1\\*iFgI9nTtiCupbcK8D9mCXg.jpeg](https://miro.medium.com/max/1400/1*iFgI9nTtiCupbcK8D9mCXg.jpeg))

### 2.1.6 เกรเดียนบูสติ้ง (Gradient Boosting)

เป็นวิธีการแบบรวมกลุ่ม (Ensemble Method) ที่พัฒนามาจากต้นไม้ตัดสินใจ โดยใช้กระบวนการบูสติ้ง (Boosting) ซึ่งเป็นกระบวนการที่ทำการสร้างต้นไม้ตัดสินใจขึ้นมา แล้วนำความผิดพลาดของต้นไม้ต้นก่อน มาเป็นต้นไม้ตัดสินใจต้นใหม่ จากนั้นทำแบบนี้ไปเรื่อยๆ จนได้เป็นโมเดลที่ดีที่สุด



รูปที่ 2.4 เกรเดียนบูสติ้ง

(ที่มา: <http://uc-r.github.io/public/images/analytics/gbm/boosted-trees-process.png>)

### 2.1.7 ชัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

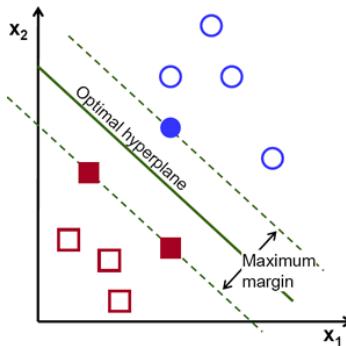
ชัพพอร์ตเวกเตอร์แมชชีน (SVM) เป็นการเรียนรู้แบบมีผู้สอนซึ่งเป็นอัลกอริทึมที่สามารถใช้ได้ทั้งการจำแนกประเภทข้อมูลและการคัดถอย แต่ส่วนใหญ่มักจะใช้กับปัญหาการจำแนกประเภทข้อมูล โดยจะทำการพลีอตข้อมูลแต่ละรายการเป็นจุดในพื้นที่  $n$  มิติ ( $n$  คือจำนวนฟีเจอร์) ซึ่งค่าของแต่ละฟีเจอร์จะเป็นค่าของพิกัดเฉพาะ จากนั้นจึงทำการจำแนกประเภทโดยการหาเส้นแบ่ง (Hyperplane) ที่จำแนกข้อมูลของทั้งสองคลาส ได้อย่างชัดเจน นอกจากนี้ยังมีเทคนิคที่เรียกว่า “เกอร์เนลทริก (Kernel Trick)” ซึ่งเป็นฟังก์ชันที่สามารถแปลงข้อมูลที่มีมิติที่ต่ำกว่าให้มีมิติที่สูงขึ้นได้ โดยการจำแนกประเภท ถ้าส่วนของระยะห่าง (Margin) มากจะจำแนกประเภทได้ดี โดยสมการการหาระยะห่างสูงสุดเพื่อให้ทำการจำแนกประเภทได้ มีดังนี้

Minimize :

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \zeta$$

Subject to :

$$\begin{aligned} y_i(w \cdot \phi(x_i) + b) &\geq 1 - \zeta_i, \forall i \\ \zeta_i &\geq 0, \forall i \end{aligned}$$

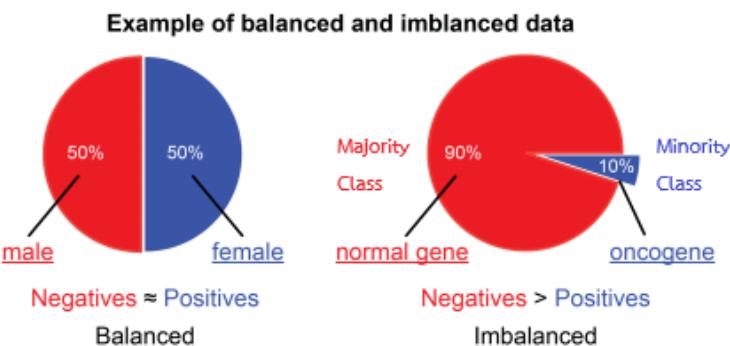


รูปที่ 2.5 ชั้พพอร์ตเวกเตอร์แมชชีน

(ที่มา: <https://www.aitrends.com/wp-content/uploads/2018/01/1-19SVM-2.jpg>)

### 2.1.8 ชุดข้อมูลไม่สมดุล (Imbalanced Dataset)

เป็นชุดข้อมูลที่มีสัดส่วนของคลาสใดคลาสหนึ่งเมี้ยไปในทางใดทางหนึ่ง ซึ่งคลาสที่มีสัดส่วนมากของชุดข้อมูลจะเรียกว่า คลาสส่วนใหญ่ (Majority Class) และคลาสที่มีสัดส่วนน้อยกว่าจะเรียกว่า คลาสส่วนน้อย (Minority Class)



รูปที่ 2.6 ตัวอย่างชุดข้อมูลสมดุลและชุดข้อมูลไม่สมดุล

(ที่มา: <https://www.aitrends.com/wp-content/uploads/2018/01/1-19SVM-2.jpg>)

### 2.1.9 ประเภทของการแก้ปัญหาชุดข้อมูลไม่สมดุล (Types of Imbalanced Methods)

การแก้ปัญหาชุดข้อมูลไม่สมดุลด้วยการเรียนรู้ของเครื่องได้ถูกนำมาใช้อ้างกว้างขวาง โดยทำการปรับเปลี่ยนชุดข้อมูลสำหรับฝึกฝน (Training set) เพื่อลดความไม่สมดุลของชุดข้อมูล หรือทำการปรับเปลี่ยนกระบวนการเรียนรู้หรือการตัดสินใจพื้นฐานของโมเดลเพื่อเพิ่มความสำคัญให้กับคลาสส่วนน้อย ดังนั้น วิธีสำหรับจัดการกับชุดข้อมูลไม่สมดุลสามารถจัดกลุ่มได้ ดังนี้

#### 2.1.9.1 วิธีการระดับข้อมูล (Data-level methods)

เป็นวิธีที่ทำการปรับเปลี่ยนกระบวนการจ่ายชุดข้อมูลสำหรับฝึกฝนเพื่อลดระดับความไม่สมดุลของชุดข้อมูล แบ่งออกเป็น 3 กลุ่ม ดังนี้

1) วิธีเพิ่มปริมาณข้อมูล (Over-sampling) เป็นวิธีการเพิ่มจำนวนข้อมูลที่อยู่ในคลาสส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับข้อมูลที่อยู่ในคลาสส่วนมาก เช่น Random Over-Sampling, Adaptive Synthetic (ADASYN), Synthetic Minority Over-Sampling Technique (SMOTE) เป็นต้น

2) วิธีลดปริมาณข้อมูล (Under-sampling) เป็นวิธีการลดจำนวนข้อมูลที่อยู่ในคลาสส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับข้อมูลที่อยู่ในคลาสส่วนน้อย เช่น Random Under-Sampling (RUS), Neat Miss, Edited Nearest Neighbors (ENN), Tomek Links (TL) เป็นต้น

3) วิธีแบบไฮบริด (Hybrid Methods) เป็นวิธารการที่ใช้การรวมกันระหว่าง Over-sampling กับ Under-sampling เช่น SMOTE-Tomek, SMOTE-ENN เป็นต้น

#### 2.1.9.2 วิธีการระดับอัลกอริทึม (Algorithm-level methods)

เป็นวิธีที่ทำการปรับเปลี่ยนกระบวนการเรียนรู้เพื่อเพิ่มความสำคัญให้กับคลาสส่วนน้อยในการฝึกชุดข้อมูลสำหรับฝึกฝน

#### 2.1.9.3 วิธีการแบบไฮบริด (Hybrid methods)

เป็นการนำวิธีการระดับข้อมูล (Data-level methods) กับวิธีการระดับอัลกอริทึม (Algorithm-level methods) มารวมกันในรูปแบบต่าง ๆ เช่น Balanced Bagging, Balanced Random Forest, Easy Ensemble, Random Under-Sampling Boost เป็นต้น

### 2.1.10 การวัดประสิทธิภาพของโมเดล

#### 1) ค่อนฟิวชันเมทริกซ์ (Confusion Matrix)

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

รูปที่ 2.7 ค่อนฟิวชันเมทริกซ์

ความหมายของ TP, FP, TN และ FN ในโครงการนี้ มีดังนี้

- True Positive (TP) หมายถึง โมเดลแนะนำว่าให้ใช้ยาต้านจุลชีพชนิดนี้ และสัตวแพทย์ก็แนะนำว่าให้ใช้ยาต้านจุลชีพชนิดนี้
- False Positive (FP) หมายถึง โมเดลแนะนำว่าให้ใช้ยาต้านจุลชีพชนิดนี้ แต่สัตวแพทย์ไม่ได้แนะนำว่าให้ใช้ยาต้านจุลชีพชนิดนี้
- True Negative (TN) หมายถึง โมเดลแนะนำว่าไม่ให้ใช้ยาต้านจุลชีพชนิดนี้ และสัตวแพทย์ไม่ได้แนะนำว่าให้ใช้ยาต้านจุลชีพชนิดนี้
- False Negative (FN) หมายถึง โมเดลแนะนำว่าไม่ให้ใช้ยาต้านจุลชีพชนิดนี้ แต่สัตวแพทย์แนะนำว่าให้ใช้ยาต้านจุลชีพชนิดนี้

#### 2) ค่าความถูกต้อง (Accuracy)

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

#### 3) ค่าความแม่นยำ (Precision)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

#### 4) ค่าความระลึก (Recall)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

#### 5) ค่าความถ่วงดูล (F1-Score)

$$\text{F-Measure} = 2 \times \left( \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

#### 2.1.11 สัมประสิทธิ์ครามอร์ (Cramer's V Coefficient)

การหาความสัมพันธ์ของข้อมูลระหว่างตัวแปรที่มีลักษณะแบบบกถุ่ม (Nominal Scale) 2 ตัวแปร ซึ่งแสดงค่าออกมาเป็นตัวเลข โดยจะมีค่าอยู่ระหว่าง 0 ถึง 1 ซึ่งบ่งบอกว่าตัวแปรทั้งสอง มีความสัมพันธ์กันมากน้อยเพียงใด หากค่าเข้าใกล้ 1 หมายถึง ตัวแปรทั้งสองมีความสัมพันธ์กัน ในระดับมาก หากค่าเข้าใกล้ 0 หมายถึง ตัวแปรทั้งสองมีความสัมพันธ์กันในระดับน้อยหรือไม่มี สัมพันธ์กันเลย และการหาค่าสัมประสิทธิ์ครามอร์ได้มีการอาศัยการวัดจากค่าสถิติ  $\chi^2$  (Chi-Square) ซึ่งสูตรในการคำนวณมีดังนี้

$$C^2 = \sqrt{\frac{\chi^2}{N(t - 1)}}$$

$$\text{เมื่อ } t = \min(r, k) \text{ และ } \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

ดังนั้นค่า  $C^2$  จะมีค่า  $0 \leq C^2 \leq 1$  โดยไม่มีค่าเป็นลบ

## การแปลความหมายทำได้ดังนี้

ค่า $C^2$	การแปลผล
0 – 0.25	สัมพันธ์น้อย
0.26 – 0.50	สัมพันธ์ปานกลาง
0.51 – 0.75	สัมพันธ์ค่อนข้างมาก
0.76 – 1.0	สัมพันธ์มาก

## 2.2 งานวิจัยที่เกี่ยวข้อง

### 2.2.1 Using Machine Learning Techniques to Aid Empirical Antibiotic Therapy

#### Decisions in the Intensive Care Unit of a General Hospital in Greece [1]

งานวิจัยฉบับนี้เสนอการใช้เทคนิคการเรียนรู้ของเครื่องในการทำการคือยาต้านจุลชีพ เพื่อช่วยแพทย์ในการพิจารณาว่าแบคทีเรียมีภาวะคือยาหรือไม่ ซึ่งจะเป็นแนวทางให้แพทย์สามารถตัดสินใจเลือกใช้ยาต้านจุลชีพที่เหมาะสมได้ โดยทราบเพียงแค่การย้อมสีแกรมของตัวอย่างบริเวณที่ติดเชื้อ และข้อมูลประชากร งานวิจัยนี้ทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่อง 8 อัลกอริทึม ได้แก่ Linear SVC, SVM, SMO, KNN, J48, Random Forest, RIPPER และ MLP โดยชุดข้อมูลที่ใช้เป็นข้อมูลของห้องปฏิบัติการจุลชีววิทยาจากห้องปฏิบัติที่มีการติดเชื้อในโรงพยาบาลของรัฐในกรีซในช่วงปี พ.ศ. 2560 – 2561 ซึ่งชุดข้อมูลประกอบไปด้วย ออกทริบิวต์ต่าง ๆ ได้แก่ เพศ, อายุ, ประเภทของตัวอย่าง, การย้อมสีแกรม, ยาต้านจุลชีพ และคลาส แบคทีเรียม โดยมีค่าตัดสินใจคือความไวต่อยาต้านจุลชีพ (S/R) ในงานวิจัยนี้ได้เลือกใช้ซอฟต์แวร์ WEKA ซึ่งเป็นหนึ่งในเครื่องมือการเรียนรู้ของเครื่องที่ได้รับความนิยมมากที่สุดอีกทั้งยังมีอัลกอริทึมที่หลากหลาย โดยมีค่าทดสอบประสิทธิภาพของโมเดลด้วย 10-Fold Cross-Validation และวัดประสิทธิภาพของโมเดลจากค่า TP Rate, FP Rate, Precision, Recall, F-Measure, MMC, ROC Area และ PRC Area

### 2.2.2 Machine Learning Techniques to Identify Antimicrobial Resistance in the Intensive Care Unit [2]

งานวิจัยฉบับนี้เสนอการใช้เทคนิคการเรียนรู้ของเครื่องเพื่อตรวจสอบว่าแบคทีเรียมีภาวะคือต่อยาต้านจุลชีพในกลุ่มต่าง ๆ หรือไม่ โดยพิจารณาจากข้อมูลทางคลินิกและข้อมูลประชากรของผู้ป่วย ตลอดจนข้อมูลการเพาะเลี้ยงเชื้อ และบันทึกการคือยาต้านจุลชีพของแบคทีเรียม

(Antibiogram) นอกจากนี้ได้แสดงให้เห็นถึงความสัมพันธ์ระหว่างแบคทีเรียชนิดต่าง ๆ และกลุ่มของยาต้านจุลชีพ โดยทำการวิเคราะห์การสมนัย (Correspondence Analysis) โดยผลลัพธ์ของการใช้เทคนิคการเรียนรู้ของเครื่องแสดงให้เห็นถึงความสัมพันธ์ที่ไม่เป็นเชิงเส้นซึ่งจะช่วยระบุการคือยาต้านจุลชีพในห้องปฏิบัติปัจจุบัน โดยประสิทธิภาพจะขึ้นอยู่กับกลุ่มของยาต้านจุลชีพ

### **2.2.3 Machine Learning For Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy [3]**

ในงานวิจัยฉบับนี้เสนอการใช้เทคนิคการเรียนรู้ของเครื่องแบบอัตโนมัติ (Automated machine learning: AutoML) ซึ่งข้อมูลสำหรับฝึกฝนโมเดลได้ใช้ข้อมูลจากห้องปฏิบัติการโดยมีจำนวน 11,496 ตัวอย่าง ซึ่งประกอบไปด้วยแอ็ตทริบิวต์ต่าง ๆ ได้แก่ เพศ, อายุ, ประเภทของตัวอย่าง, การข้อมูลสีแกรม, สารต้านจุลชีพ (antimicrobial substances) และผลการทดสอบความไวต่อยาต้านจุลชีพ ในงานวิจัยนี้ได้มีการใช้งานเทคนิคการเรียนรู้ของเครื่องแบบอัตโนมัติ (AutoML) ในการพัฒนาโมเดลซึ่งทำให้สามารถพัฒนาโมเดลการเรียนรู้ของเครื่องได้โดยง่าย สำหรับการทดลองนี้จะใช้การตั้งค่าพื้นฐาน (Default Configuration) เพื่อที่จะสามารถพัฒนาโมเดลได้โดยง่ายแม้ว่าจะไม่เขียนภาษา Python โดยใช้เทคนิค 10-Fold Cross-Validation ในการวัดผล และเนื่องจากข้อมูลนี้มีความไม่สมดุล ในงานวิจัยนี้ได้มีการใช้เทคนิค SMOTE เพื่อทำให้ข้อมูลมีความสมดุล ในงานวิจัยนี้ได้ใช้ค่าในการวัดผล 4 ค่า ได้แก่ AUCW, APSW, F1W, Accuracy

งานวิจัยนี้ได้แสดงผลการทดสอบประสิทธิภาพอัลกอริทึม 4 อันดับแรก (Top 4) ของโมเดลการเรียนรู้ของเครื่องแบบอัตโนมัติ (AutoML) โดยใช้ข้อมูลดิบและข้อมูลที่ผ่านการสังเคราะห์เพิ่มด้วยเทคนิค SMOTE ซึ่งอัลกอริทึม 4 อันดับแรกจะเป็นอัลกอริทึมที่ใช้งานแล้วให้ประสิทธิภาพสูงที่สุด สำหรับอัลกอริทึม 4 อันดับแรกของการฝึกฝนโมเดลโดยใช้ข้อมูลดิบ ได้แก่ StackEnsemble, VotingEnsemble, (MaxAbsScaler + LightGBM) และ (SparseNormalizer + XGBoostClassifier) ตามลำดับ ส่วนอัลกอริทึม 4 อันดับแรกของการฝึกฝนโมเดลโดยใช้ข้อมูลที่ผ่านการสังเคราะห์เพิ่มด้วยเทคนิค SMOTE ได้แก่ Stack Ensemble, Voting Ensemble, (SparseNormalizer + XGBoostClassifier) และ (SparseNormalizer + LightGBM) ตามลำดับ

#### **2.2.4 Machine Learning Techniques to Identify Antibiotic Resistance in Patients**

##### **Diagnosed with Various Skin and Soft Tissue Infections [4]**

ในงานวิจัยฉบับนี้เสนอการใช้เทคนิคการเรียนรู้ของเครื่องเพื่อฝึกฝนโมเดลในการทำนายการดื้อยาต้านจุลชีพของผู้ป่วยที่ตรวจพบการติดเชื้อที่ผิวนัง และเนื้อเยื่ออ่อนต่าง ๆ โดยใช้ข้อมูลประชากรผู้ป่วย และข้อมูลจากผลการทดสอบความไวต่อยาต้านจุลชีพ (Antibiotic Susceptibility Test) สำหรับอัลกอริทึมที่ใช้ฝึกฝนโมเดล ได้แก่ Random Forest (RF), Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) ซึ่งจากการฝึกฝนโมเดลได้ค่า AUC อยู่ที่ 0.68 – 0.98 ใน GPC (Gram Positive Cocci Bacteria) และ 0.56 – 0.93 ใน GNB (Gram Negative Bacilli Bacteria) โดยขึ้นอยู่กับยาต้านจุลชีพแต่ละชนิด นอกจากนี้ยังได้ทำการวิเคราะห์สหสัมพันธ์ (Correlation) เพื่อหาความสัมพันธ์เชิงเส้นระหว่างแต่ละฟีเจอร์กับตระกูลยาต้านจุลชีพที่แตกต่างกัน โดยเทคนิคการเรียนรู้ด้วยเครื่องแสดงให้เห็นว่ามีความสัมพันธ์แบบไม่เชิงเส้นที่คาดการณ์ได้ระหว่างลักษณะข้อมูลประชากรทางคลินิกของผู้ป่วยและยาต้านจุลชีพ ซึ่งความแม่นยำของการทำนายจะขึ้นอยู่กับชนิดของตระกูลยาต้านจุลชีพ

#### **2.2.5 Antimicrobial Resistance and Machine Learning: Challenges and Opportunities**

##### **[5]**

ในงานวิจัยฉบับนี้เสนอการใช้เครื่องมือการเรียนรู้ของเครื่องในการช่วยทำนายผลการดื้อยาต้านจุลชีพ เนื่องจากองค์การอนามัยโลกได้เลือกเห็นถึงปัญหาการใช้ยาต้านจุลชีพที่ไม่เหมาะสมโดยเฉพาะประเทศที่มีรายได้ต่ำถึงปานกลาง ซึ่งในช่วงไม่กี่ปีที่ผ่านมา มีการใช้เทคนิคการเรียนรู้ของเครื่องในการช่วยจัดการแนะนำยาต้านจุลชีพซึ่งได้ผลลัพธ์ที่น่าพึงพอใจ โดยงานวิจัยนี้ได้แสดงถึงการใช้เทคนิคเรียนรู้ของเครื่องที่มีอยู่มาช่วยจัดการกับการดื้อยา โดยใช้ข้อมูลทางประชากรศาสตร์และข้อมูลของผลการทดสอบที่ได้จากห้องปฏิบัติการ โดยได้อธิบายถึงความท้าทาย ข้อจำกัด และอุปสรรคในการขยายเพื่อใช้เทคนิคการเรียนรู้ของเครื่องในจัดการกับการดื้อยาร่วมถึงแนวทางต่อไปที่จะใช้ในอนาคต

#### **2.2.6 SMOTE: Synthetic Minority Oversampling Technique [6]**

ในงานวิจัยฉบับนี้เสนอเทคนิคการแก้ปัญหาชุดข้อมูลไม่สมดุลโดยใช้วิธีการเพิ่มปริมาณข้อมูล (Over-sampling) ซึ่งเป็นการนำคลาสส่วนน้อย (Minority Class) มาทำการสังเคราะห์

เพิ่มปริมาณข้อมูล โดยการสร้างตัวอย่างสังเคราะห์ (Synthetic examples) หรือเทคนิคนี้ว่า Synthetic Minority Oversampling Technique (SMOTE) ซึ่งให้ประสิทธิภาพที่ดีกว่าการใช้วิธีการเพิ่มปริมาณข้อมูลโดยการสุ่มตัวอย่างแบบไส้คืน (Over-sampling with replacement) โดยหลักการสังเคราะห์ข้อมูลจะทำการนำคลาสส่วนน้อยแต่ละตัวมาทำการสุ่มเพื่อนบ้านที่เป็นคลาสส่วนน้อยมาหนึ่งตัวจากนั้นทำการสังเคราะห์ข้อมูลด้วยการหาระยะห่างระหว่างข้อมูลสองตัวนั้น จากนั้นทำการคูณด้วยค่าสุ่มในช่วง 0 – 1 แล้วทำการนำค่าที่สังเคราะห์ได้ไปใส่ระหว่างข้อมูลทั้งสองตัวนั้น

### 2.2.7 Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets

#### Learning [7]

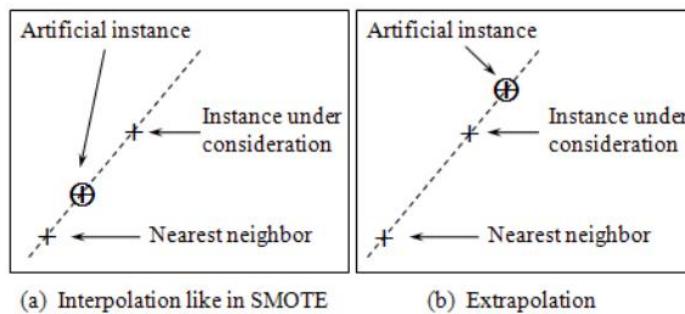
ในงานวิจัยฉบับนี้ได้คิดวิธีการแก้ปัญหาชุดข้อมูลไม่สมดุล โดยพื้นฐานมาจากเทคนิค SMOTE ปรับปรุงเป็น Borderline-SMOTE โดยมีแนวคิดว่าจะสังเคราะห์ข้อมูลตรงช่วงขอบของคลาสส่วนน้อยเท่านั้น แทนที่จะเลือกสังเคราะห์จากกลุ่มตัวอย่างทั้งหมด โดยหลักการหาขอบของข้อมูลจะใช้หลักการหาเพื่อนบ้านใกล้เคียงของคลาสส่วนน้อย โดยถ้าเพื่อนบ้านของคลาสส่วนน้อยนั้นมีเพื่อนบ้านที่เป็นคลาสส่วนมากมากกว่าเพื่อนบ้านที่คลาสส่วนน้อยจะมองว่าข้อมูลตัวนั้นเป็นขอบ หลังจากนั้นก็ทำการสังเคราะห์ข้อมูลตามหลักการของ SMOTE แต่จะมีการสังเคราะห์ข้อมูลต่อคลาสส่วนน้อยที่ถูกเลือกมากกว่า SMOTE โดย SMOTE จะทำการสุ่มเพื่อนบ้านมาหนึ่งตัวแล้วสังเคราะห์แต่ Borderline-SMOTE จะเลือกเพื่อนบ้านหลายตัว

### 2.2.8 ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning [8]

ในงานวิจัยฉบับนี้ได้คิดวิธีการแก้ปัญหาชุดข้อมูลที่ไม่สมดุลโดยจะใช้วิธี Adaptive Synthetic (ADASYN) Sampling โดย ADASYN จะมีแนวคิดในการแก้ปัญหา คือ การให้น้ำหนักการกระจายของข้อมูลคลาสน้อยในแต่ละกลุ่มต่างกัน โดยจะสังเคราะห์ข้อมูลในชุดที่ยากต่อการเรียนรู้มากกว่าในส่วนที่เรียนรู้ง่าย ซึ่ง ADASYN มีหลักการคือ ทำการหาเพื่อนบ้านของคลาสส่วนน้อย จากนั้นาอัตราส่วนของเพื่อนบ้านที่เป็นคลาสส่วนมากต่อจำนวนเพื่อนบ้านทั้งหมด เพื่อเป็นตัวบ่งบอกว่าจะทำการสังเคราะห์ข้อมูลเพิ่มจากคลาสส่วนน้อยว่าเป็นจำนวนเท่าใดต่อจำนวนข้อมูลที่ต้องสังเคราะห์ทั้งหมด หลังจากนั้นก็ทำการสังเคราะห์ข้อมูลโดยใช้คลาสส่วนน้อยที่ถูกเลือกับเพื่อนบ้านที่เป็นคลาสส่วนน้อยเหมือนกัน ซึ่งเลือกจากการสุ่มโดยทำการสังเคราะห์ข้อมูลระหว่างข้อมูลทั้งสองตัวนั้น

### 2.2.9 Borderline Over-sampling for Imbalanced Data Classification [9]

ในงานวิจัยฉบับนี้ได้คิดวิธีการแก้ปัญหาชุดข้อมูลที่ไม่สมดุลโดยจะใช้วิธี Borderline Over-sampling (BOS) หรือ SVM-SMOTE โดย BOS จะมีแนวคิดในการแก้ปัญหาความไม่สมดุลของข้อมูลโดยจะสังเคราะห์ข้อมูลบริเวณขอบเขตของเพราะบริเวณของเป็นพื้นที่ที่ใกล้กับขอบเขตตัดสินใจจะมีผลต่อการทำนายมากกว่าบริเวณที่มีคลาสส่วนน้อยอยู่มาก โดยหลักการการหาขอบเขตนำข้อมูลมาสร้างชั้พพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)! แล้วหาชั้พพอร์ตเวกเตอร์ (Support Vector) ของข้อมูลคลาสส่วนน้อย จากนั้นก็นำชั้พพอร์ตเวกเตอร์มาหาเพื่อนบ้านแล้วทำการสังเคราะห์ข้อมูล โดยถ้ามีเพื่อนบ้านที่เป็นคลาสส่วนมากมากกว่าคลาสส่วนน้อยจะทำการสังเคราะห์แบบขยาย (Extrapolation) โดยใช้ชั้พพอร์ตเวกเตอร์ที่เลือกกับเพื่อนบ้านมาหาระยะห่างแล้วทำการคูณด้วยค่าสุ่มที่อยู่ในช่วง 0 – 1 แล้วขยายต่อจากข้อมูลที่เลือก แต่ถ้ามีเพื่อนบ้านที่เป็นคลาสส่วนน้อยมากกว่าคลาสส่วนมากจะทำการสังเคราะห์แบบแทรก (Interpolation) โดยใช้ชั้พพอร์ตเวกเตอร์ที่เลือกกับเพื่อนบ้านมาหาระยะห่างแล้วทำการคูณด้วยค่าสุ่มที่อยู่ในช่วง 0 – 1 แล้วหดลงจากข้อมูลที่เลือก



รูปที่ 2.8 การสังเคราะห์แบบขยาย (a) และการสังเคราะห์แบบแทรก (b)

(ที่มา: Borderline Over-sampling for Imbalanced Data Classification, 2552)

### 2.2.10 RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise [10]

ในงานวิจัยฉบับนี้ได้คิดวิธีการแก้ปัญหาชุดข้อมูลที่ไม่สมดุลโดยจะใช้วิธี self-adaptive robust SMOTE (RSMOTE) โดยมีแนวคิดที่ว่า SMOTE นั้นอาจจะมีประสิทธิภาพไม่ดี เพราะอาจจะสังเคราะห์ข้อมูลโดยใช้ข้อมูลที่เป็นน้อยส์ (Noise) ซึ่งเป็นเหมือนการขยายตัวน้อยส์ให้มีมากขึ้น วิธี RSMOTE จึงทำการแยกประเภทของข้อมูลคลาสส่วนน้อยที่เป็นน้อยส์ออก โดยพิจารณาจากเพื่อน

ข้านว่าหากเป็นคลาสส่วนมากทั้งหมดจะนับว่าเป็นน้อยส์ หลังจากนั้นก็จะสังเคราะห์ข้อมูลโดย หากเป็นส่วนที่มีความคลุมอย่างกว้างขวางและส่วนที่มีความรุนแรงมากจะสังเคราะห์ข้อมูลน้อย โดยใช้ค่าความหนาแน่นสัมพันธ์ของข้อมูลคลาสส่วนน้อยแต่ละตัว แล้วทำการแบ่งกลุ่มข้อมูลแบบทวิมิน (2-Means Clustering) เพื่อทำการแยกว่าข้อมูลตัวไหนปลดล็อก และข้อมูลตัวไหนเป็นของ จากนั้นทำการหาหนักของคลาสส่วนน้อยแต่ละตัวว่าจะใช้ในการสังเคราะห์เท่าใด สุดท้ายก็ทำการสังเคราะห์ข้อมูลคลาสส่วนน้อยเพิ่มขึ้น โดยใช้เพื่อนบ้าน โดยจะเป็นเทคนิคเดียวกัน กับ SMOTE

## 2.3 เครื่องมือที่เกี่ยวข้อง

### 2.3.1 Python

Python เป็นภาษาโปรแกรมคอมพิวเตอร์ระดับสูงที่แพร่หลาย ใช้งานได้ง่าย นิยมมากในปัจจุบัน และยังมีไลบรารีต่าง ๆ มากมายที่ช่วยในการใช้งานเทคนิคการเรียนรู้ของเครื่อง

### 2.3.2 Numpy

Numpy เป็นไลบรารีโอเพนซอร์ซในภาษาไพทอน โดยสามารถรองรับการคำนวณแบบอาร์เรย์และเมทริกหลาย ๆ มิติ การคำนวณตัวเลขจำนวนมากๆ โดยผู้ใช้สามารถดำเนินการใช้งานอาร์เรย์และเมทริกได้ง่ายและรวดเร็ว

### 2.3.3 Pandas

Pandas เป็นไลบรารีโอเพนซอร์ซในภาษาไพทอนสำหรับจัดการและวิเคราะห์ข้อมูล โดยมีโครงสร้างข้อมูลและการดำเนินการอยู่ในรูปแบบของอนุกรม (Series) และ ตาราง (Table)

### 2.3.4 Scikit-learn

Scikit-learn เป็นไลบรารีในภาษาไพทอนสำหรับการพัฒนาโปรแกรมโดยใช้เทคนิคการเรียนรู้ของเครื่อง นอกเหนือไปนี้ยังได้ถูกออกแบบมาให้ทำงานร่วมกับไลบรารี Numpy และ Scipy ของไพทอนได้อีกด้วย

### 2.3.5 Imbalanced-learn

Imbalanced-learn เป็นไลบรารีในภาษา Python สำหรับจัดการกับชุดข้อมูลที่มีความไม่สมดุล ซึ่งเป็นไลบรารีที่รวมรวมเทคนิคต่าง ๆ ที่ใช้จัดการกับชุดข้อมูลที่ไม่สมดุล ไว้มากmany

### 2.3.6 Extreme Gradient Boosting (XGBoost)

XGBoost เป็นไลบรารีโอเพนซอร์ซที่นิยม แพร่หลาย และมีประสิทธิภาพซึ่งสามารถใช้ได้ทางด้านการจำแนกประเภทข้อมูล และด้านการ回帰 (Regression) โดยใช้อัลกอริทึมเกรเดียนบูสติ้ง (Gradient Boosting)

### 2.3.7 Vue.js

Vue.js เป็นภาษาสคริปต์เฟรมเวิร์ค โอเพนซอร์ซที่ใช้สำหรับพัฒนาในส่วนของการติดต่อผู้ใช้งาน (User Interface) แบบแอปพลิเคชันหน้าเดียว (Single-Page Application) โดย Vue นั้นเป็นสถาปัตยกรรมแบบ Model-View-ViewModel (MVVM)

### 2.3.8 FastAPI

FastAPI เป็นเว็บเฟรมเวิร์คที่ใช้สำหรับสร้างส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interface: API) โดยใช้ภาษา Python ซึ่งเป็นหนึ่งในเฟรมเวิร์คที่เร็วที่สุดของ Python และมีการสร้างเอกสาร (Document) ของ API แบบอัตโนมัติโดยเป็นไปตามรูปแบบของโอเพนอเพิร์ช (OpenAPI)

### 2.3.9 PostgreSQL

PostgreSQL เป็นโอเพนซอร์ชระบบจัดการฐานข้อมูลเชิงวัตถุ-สัมพันธ์ (Object-Relation Database Management system: RDBMS) โดยมีรูปแบบคำสั่งของภาษา SQL โดยสามารถทำงานได้บนระบบปฏิบัติการลินุกซ์ แมค และวินโดวส์

### 2.3.10 Python-docx

Python-docx เป็นไลบรารีที่เอาไว้สำหรับให้ภาษา Python สามารถอ่าน สร้าง และปรับแก้เอกสารในโครงฟอร์แมตได้

## บทที่ 3

### การออกแบบ และการพัฒนา

#### 3.1 การออกแบบโภเมเดลแนะนำนำยาต้านจุลชีพ

โครงการนี้ออกแบบโภเมเดลแนะนำนำยาต้านจุลชีพ โดยแบ่งโภเมเดลอออกเป็น 2 กลุ่ม คือ โภเมเดลแนะนำนำยาสำหรับโรคติดเชื้อแบคทีเรียแกรมบวก (GP) และ โภเมเดลแนะนำนำยาสำหรับโรคติดเชื้อแบคทีเรียแกรมลบ (GN) ซึ่งโภเมเดลแต่ละกลุ่มนี้มีโภเมเดลของยาต้านจุลชีพชนิดละ 1 โภเมเดล โดยออกแบบให้ผลลัพธ์ของแต่ละโภเมเดลมี 2 คลาส คือ แนะนำให้ใช้ยาต้านจุลชีพชนิดนี้ และไม่แนะนำให้ใช้ยาต้านจุลชีพชนิดนี้ โครงการนี้จะฝึกฝนโภเมเดลแนะนำนำยาสำหรับโรคติดเชื้อแบคทีเรียแกรมบวกและแบคทีเรียแกรมลบอย่างละ 11 ชนิด โดยเลือกจากยาต้านจุลชีพที่มีจำนวนครั้งของการแนะนำจากสัตวแพทย์มากที่สุด ดังนี้

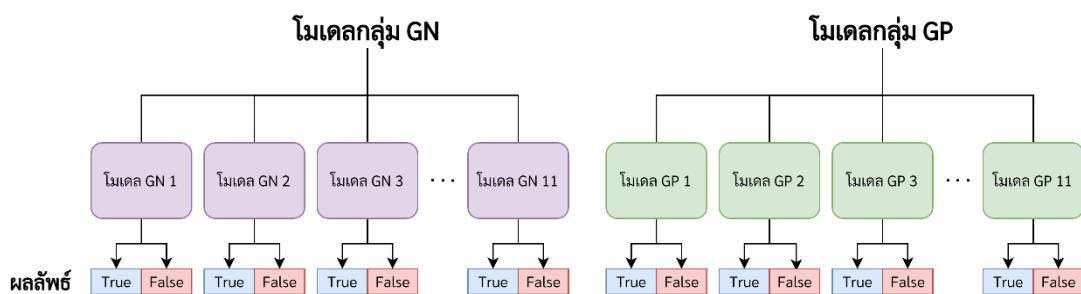
โภเมเดลแนะนำนำยาสำหรับโรคติดเชื้อแบคทีเรียแกรมบวก 11 โภเมเดล ได้แก่

- 1) โภเมเดลแนะนำนำยา Amikacin
- 2) โภเมเดลแนะนำนำยา Amoxicillin/clavulanic acid
- 3) โภเมเดลแนะนำนำยา Cefalexin
- 4) โภเมเดลแนะนำนำยา Cefovecin
- 5) โภเมเดลแนะนำนำยา Clindamycin
- 6) โภเมเดลแนะนำนำยา Doxycycline
- 7) โภเมเดลแนะนำนำยา Enrofloxacin
- 8) โภเมเดลแนะนำนำยา Marbofloxacin
- 9) โภเมเดลแนะนำนำยา Nitrofurantoin
- 10) โภเมเดลแนะนำนำยา Trimethoprim/sulfamethoxazole
- 11) โภเมเดลแนะนำนำยา Vancomycin

โภเมเดลแนะนำนำยาสำหรับโรคติดเชื้อแบคทีเรียแกรมลบ 11 โภเมเดล ได้แก่

- 1) โภเมเดลแนะนำนำยา Amikacin
- 2) โภเมเดลแนะนำนำยา Amoxicillin/clavulanic acid
- 3) โภเมเดลแนะนำนำยา Cefalexin
- 4) โภเมเดลแนะนำนำยา Cefovecin

- 5) ไมเดลແນະນໍາຍາ Doxycycline
- 6) ไมเดลແນະນໍາຍາ Enrofloxacin
- 7) ไมเดลແນະນໍາຍາ Gentamicin
- 8) ไมเดลແນະນໍາຍາ Imipenem
- 9) ไมเดลແນະນໍາຍາ Marbofloxacin
- 10) ไมเดลແນະນໍາຍາ Nitrofurantoin
- 11) ไมเดลແນະນໍາຍາ Trimethoprim/sulfamethoxazole



รูปที่ 3.1 การออกแบบไมเดลແນະນໍາຍາຕ້ານຈຸລືບີພ

## 3.2 การออกแบบขั้นตอนพัฒนาไมเดลແນະນໍາຍາຕ້ານຈຸລືບີພ

### 3.2.1 การจัดเตรียมชุดข้อมูล (Dataset)

ชุดข้อมูลที่ใช้มาจากการรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยาปี พ.ศ. 2559 - 2564 ของห้องปฏิบัติการจุลชีววิทยา โรงพยาบาลสัตว์เล็ก คณะสัตวแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ซึ่งข้อมูลชุดนี้มีการเก็บรวบรวมมาเป็นไฟล์ .docx จึงต้องทำการแปลงข้อมูลจากไฟล์ .docx เป็นไฟล์ .csv ก่อน จากนั้นจึงแบ่งชุดข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลแบบที่เรียกว่าบาก และชุดข้อมูลแบบที่เรียกว่ากลบ

### 3.2.2 การทำความสะอาดชุดข้อมูล (Data Cleansing)

ทำการตรวจสอบและแก้ไขข้อมูลที่มีความผิดพลาดให้ถูกต้อง จัดการแก้ไขข้อมูลของคอลัมน์ date\_of\_submission และ report\_issued\_date ให้มีรูปแบบเป็นวันที่ และจัดการแก้ไขข้อมูลที่มีค่าว่างของคอลัมน์ vitek\_id, species, submitted\_sample และ bacteria\_species

### 3.2.3 การเตรียมข้อมูลให้พร้อมนำไปใช้งาน (Data Preprocessing)

#### 3.2.3.1 การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)

วิเคราะห์ตรวจสอบข้อมูลเบื้องต้นโดยใช้การนำเสนอแผนภาพข้อมูล (Data Visualization) เช่น แผนภูมิแท่ง (Bar Chart), แผนภูมิวงกลม (Pie Chart) เป็นต้น เพื่อคุณวัดและกระกระจายตัวของข้อมูล

#### 3.2.3.2 กระบวนการแปลงข้อมูล (Feature Engineering)

สร้างฟีเจอร์ใหม่ขึ้นมาชื่อว่า “bacteria\_genus” โดยตัดเอาเฉพาะคำแรกจากฟีเจอร์สายพันธุ์แบคทีเรีย (bacteria\_species) จะได้เป็นสกุลของแบคทีเรีย

จัดกลุ่มหมวดหมู่เล็ก ๆ ของฟีเจอร์ตัวอย่างที่ส่งตรวจ (submitted\_sample) ที่มีจำนวนน้อยกว่า 10 รวมกันให้เป็นหมวดหมู่ใหม่ชื่อว่า “other”

ใช้วิธีการเข้ารหัสแบบวันshot (One-Hot Encoding) ในการเข้ารหัสข้อมูลประเภทนามบัญญัติ (Nominal) ซึ่งเป็นการแปลงข้อมูลให้เป็นเลข 0 และ 1 โดยเป็น 1 ก็ต่อเมื่อมีข้อมูลชนิดนั้น และเป็น 0 เมื่อมีไม่มีข้อมูลชนิดนั้น

index	species	index	species_cat	species_dog	species_other
0	cat	0	1	0	0
1	dog	1	0	1	0
2	other	2	0	0	1

รูปที่ 3.2 วิธีการเข้ารหัสแบบวันshot

### 3.2.4 การแบ่งชุดข้อมูล (Training Set and Test Set)

#### 3.2.4.1 การแบ่งชุดข้อมูลสำหรับฝึกฝนและชุดข้อมูลสำหรับทดสอบประสิทธิภาพ

##### แบบทดสอบรายเคส (Test by Case)

นำชุดข้อมูลมาแบ่งออกเป็น 2 ชุด คือ ชุดข้อมูลสำหรับฝึกฝน โฉนด และชุดข้อมูลสำหรับทดสอบรายเคส โดยการเลือกข้อมูลที่จะนำมาทดสอบจะทำการสุ่มเลือกข้อมูลมา 20% ของแต่ละเดือน เริ่มตั้งแต่เดือนกรกฎาคมปี พ.ศ. 2562 จนถึงเดือนพฤษภาคม ปี 2564 เพื่อที่จะได้ข้อมูลของปัจจุบันเป็นชุดข้อมูลสำหรับทดสอบแบบรายเคส

### **3.2.4.2 การแบ่งชุดข้อมูลสำหรับฝึกฝนและชุดข้อมูลสำหรับทดสอบประสิทธิภาพของแต่ละโมเดล**

นำชุดข้อมูลสำหรับฝึกฝนไม่เดลจากข้อที่ 3.2.5.1 มาแบ่งออกเป็น 2 ชุดเพื่อที่จะนำไปฝึกฝนและทดสอบ โดยเดลยาต้านจุลชีพแต่ละชนิด โดยแบ่งเป็นชุดข้อมูลสำหรับฝึกฝน (Training Set) 80% และชุดข้อมูลสำหรับทดสอบ (Test Set) 20% โดยจะนำข้อมูลสำหรับฝึกฝนมาทำ 10-Fold Cross-Validation เพื่อปรับจูนพารามิเตอร์ของโมเดล และทำการทดสอบประสิทธิภาพของโมเดลด้วยชุดข้อมูลสำหรับทดสอบ

### **3.2.5 การแก้ไขปัญหาชุดข้อมูลไม่สมดุล**

ใช้วิธีการแก้ไขปัญหาชุดข้อมูลไม่สมดุลด้วยการสังเคราะห์ข้อมูล (Oversampling) โดยใช้เทคนิคการสังเคราะห์ข้อมูลของคลาสส่วนน้อยให้มีความใกล้เคียงหรือเท่ากันกับคลาสส่วนมากในโครงการนี้จะทำการศึกษาเทคนิคการสังเคราะห์ข้อมูลต่าง ๆ ได้แก่ SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN และ R-SMOTE

### **3.2.6 การฝึกฝนโมเดล**

นำชุดข้อมูลสำหรับฝึกฝนจากข้อที่ 3.2.5.1 มาฝึกฝนโมเดล ในโครงการนี้จะทำการศึกษาเทคนิคการจำแนกข้อมูล (Classification) ต่าง ๆ เพื่อเปรียบเทียบประสิทธิภาพ ได้แก่ ต้นไม้ตัดสินใจ, แรนdomฟอร์สต์, เอ็กซ์ตรีมเกรเดียนบูสติ้ง และชัพพร์ตเวกเตอร์แมชชีน

### **3.2.7 การวัดประสิทธิภาพของโมเดล**

#### **3.2.7.1 การวัดประสิทธิภาพของโมเดลโดยแต่ละชนิด**

นำชุดข้อมูลสำหรับทดสอบจากข้อที่ 3.2.5.2 มาทำการทดสอบกับโมเดลยาแต่ละชนิด แล้วทำการวัดประสิทธิภาพโดยใช้ค่า Accuracy, Precision, Recall และ F1-Score

#### **3.2.7.2 การวัดประสิทธิภาพของโมเดลแบบทดสอบรายคดส (Test by Case)**

นำชุดข้อมูลสำหรับทดสอบรายคดส จากข้อที่ 3.2.5.1 มาทำการทดสอบกับทุกโมเดล แล้วทำการวัดประสิทธิภาพแบบทดสอบรายคดส โดยมีวิธีการ ดังนี้

1) นำข้อมูลสำหรับทดสอบรายเคสไปทำการทดสอบกับทุกโมเดล จะได้ออกมาเป็นตารางผลลัพธ์จากการทดสอบเป็นค่า T (True) และค่า F (False) โดยมีคอลัมน์เป็นโมเดล ยาต้านจุลชีพแต่ละชนิด ดังนี้

Case	M 1	M 2	M 3	M 4	M 5	...	M n
1	T	T	F	F	F	...	T
2	F	F	F	F	F	...	F
3	F	F	F	F	T	...	F

รูปที่ 3.3 ตารางผลลัพธ์จากการทดสอบ

2) หา TP, TN, FP, FN ของแต่ละเคส (Case) และคำนวณหาค่า Accuracy, Precision, Recall, F1-Score ของแต่ละเคส

3) สำหรับกรณีที่คำตอบจริงเป็น F ทุกโมเดล (คำตอบจริงไม่มีการแนะนำยา เช่น เป็นเคสดื้อยาทุกชนิด) จะไม่สามารถคำนวณ Precision, Recall, F1-Score ได้ ดังนั้นสำหรับในกรณีนี้จะใช้วิธีการคำนวณโดย หากเคสใดได้ผลลัพธ์จากการทำนายเป็น F ทุกโมเดล จะให้ค่า Precision, Recall, F1-Score เท่ากับ 1 แต่หากผลลัพธ์จากการทำนายไม่ใช่ F ทุกโมเดล จะให้ค่า Precision, Recall, F1-Score เท่ากับ 0

4) นำผลลัพธ์ที่ได้จากการคำนวณค่า Accuracy, Precision, Recall, F1-Score ของแต่ละเคสมาหาค่าเฉลี่ย จะได้เป็นผลการวัดประสิทธิภาพแบบทดสอบรายเคส

### 3.3 การออกแบบการทดลอง

โครงการนี้ได้ออกแบบการทดลองออกเป็น 2 ตอน ได้แก่ การฝึกฝนโมเดลแนะนำตัวนักเรียนชั้นปีที่ 6 และการฝึกฝนโมเดลทำนายการดื้อยาต้านจุลชีพ

#### 3.3.1 ตอนที่ 1 การฝึกฝนโมเดลแนะนำตัวนักเรียนชั้นปีที่ 6 แบ่งออกเป็น 4 การทดลอง ดังนี้

3.3.1.1 การทดลองที่ 1.1 การเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ โดยใช้ชุดข้อมูลแบคทีเรียแกรมบวก และชุดข้อมูลแบคทีเรียแกรมลบ

โดยมีจุดประสงค์การทดลอง ดังนี้

1) เพื่อฝึกฝนโมเดลแนะนำตัวนักเรียนชั้นปีที่ 6 โดยใช้ชุดข้อมูลจากรายงานผลการระบุชนิดจุลชีพและการทดสอบความไว้วarnต่อข้อมูลห้องปฏิบัติการจุลชีววิทยา

2) เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูล 4 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ, แรนdomฟอร์สต์, เอ็กซ์ตรีมเกรเดียนบูสติ้ง และชัพพอร์ตเเกะเตอร์แมชชีน ด้วยค่า Accuracy, Precision, Recall และ F1-Score

3) เพื่อเลือกเทคนิคการจำแนกข้อมูลที่เหมาะสมกับชุดข้อมูลแบคทีเรียแกรมบวกและชุดข้อมูลแบคทีเรียแกรมลบ

4) เพื่อเลือกการจำแนกข้อมูล 2 เทคนิคที่ดีที่สุดสำหรับนำไปทำการทดลองที่ 1.2 (การแก้ไขปัญหาชุดข้อมูลไม่สมดุล)

3.3.1.2 การทดลองที่ 1.2 การเปรียบเทียบทekenิคการสังเคราะห์ข้อมูลสำหรับการแก้ปัญหาชุดข้อมูลไม่สมดุล โดยใช้ชุดข้อมูลแบคทีเรียแกรมบวก และชุดข้อมูลแบคทีเรียแกรมลบ

โดยมีจุดประสงค์การทดลอง ดังนี้

1) เพื่อศึกษาเทคนิคการสังเคราะห์ข้อมูลในการแก้ไขปัญหาชุดข้อมูลไม่สมดุล ได้แก่ SMOTE, Borderline-SMOTE, ADASYN, SVM-SMOTE และ R-SMOTE โดยใช้ชุดข้อมูลแบคทีเรียแกรมบวก และชุดข้อมูลแบคทีเรียแกรมลบ ซึ่งเป็นข้อมูลเชิงคุณภาพ (Qualitative Data)

2) เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูล ด้วยค่า Precision, Recall และ F1-Score

3) เพื่อเลือกเทคนิคการจำแนกข้อมูลที่เหมาะสมกับชุดข้อมูล

4) เพื่อเลือกเทคนิคการสังเคราะห์ข้อมูลที่เหมาะสมกับชุดข้อมูลแบคทีเรียแกรมบวก และชุดข้อมูลแบคทีเรียแกรมลบ

**3.3.1.3 การทดลองที่ 1.3 การเปรียบเทียบทักษิณิกการสังเคราะห์ข้อมูลสำหรับการแก้ปัญหาชุดข้อมูลไม่สมดุล โดยใช้ชุดข้อมูลทั่วไปจาก UCI Machine Learning Repository**

โดยมีจุดประสงค์การทดลอง ดังนี้

- 1) เพื่อศึกษาทักษิณิกการสังเคราะห์ข้อมูลต่าง ๆ ในการแก้ไขปัญหาชุดข้อมูลไม่สมดุล ได้แก่ SMOTE, Borderline-SMOTE, ADASYN, SVM-SMOTE และ R-SMOTE โดยใช้ชุดข้อมูลทั่วไปที่มีฟีเจอร์เป็นข้อมูลเชิงปริมาณ (Quantitative Data)
- 2) เพื่อเปรียบเทียบประสิทธิภาพของทักษิณิกการสังเคราะห์ข้อมูลด้วยค่า Precision, Recall และ F1-Score

- 3) เพื่อเปรียบเทียบผลลัพธ์ที่ได้กับการทดลองที่ 1.2

**3.3.1.4 การทดลองที่ 1.4 การปรับพารามิเตอร์ของโมเดลกลุ่มแบบที่เรียกว่า “โมเดลกลุ่มแบบที่เรียกว่า”**

โดยมีจุดประสงค์การทดลอง ดังนี้

- 1) เพื่อนำผลจากการทดลองที่ 1.2 มาทำการฝึกฝนโมเดล แล้วพารามิเตอร์ของโมเดลที่ดีที่สุด โดยใช้การวัดผลด้วยค่า F1-Score
- 2) เพื่อเปรียบเทียบประสิทธิภาพก่อนและหลังการปรับพารามิเตอร์ด้วยค่า Accuracy, Precision, Recall และ F1-Score

**3.3.2 ตอนที่ 2 การฝึกฝนโมเดลทำนายการดื้อยาต้านจุลชีพ แบ่งออกเป็น 2 การทดลอง ดังนี้**

**3.3.2.1 การทดลองที่ 2.1 การทำนายการดื้อยาต้านจุลชีพของแบบที่เรียกว่า**

โดยมีจุดประสงค์การทดลอง ดังนี้

- 1) เพื่อนำผลการทำนายการดื้อยาต้านจุลชีพของแบบที่เรียกว่า โดยใช้ชุดข้อมูลรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยา และวัดประสิทธิภาพของโมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score
- 2) วิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ โดยใช้ค่าสัมประสิทธิ์ครามอร์ (Cramer's V Coefficient) เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิดในกลุ่มแบบที่เรียกว่า

### 3.3.2.2 การทดลองที่ 2.2 การทำนายการดื้อยาต้านจุลชีพของเบคทีเรียแกรมลบ

โดยมีจุดประสงค์การทดลอง ดังนี้

1) เพื่อฝนฟันโนเมเดลทำนายการดื้อยาต้านจุลชีพของเบคทีเรียแกรมลบ โดยใช้ชุดข้อมูลรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยา และวัดประสิทธิภาพของโนเมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score

2) วิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ โดยใช้ค่าสัมประสิทธิ์ครามอร์ (Cramer's V Coefficient) เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิดในกลุ่มเบคทีเรียแกรมลบ

## 3.4 รายละเอียดของชุดข้อมูลที่ใช้

ชุดข้อมูลที่ใช้เป็นชุดข้อมูลจากรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยาปี พ.ศ. 2559 - 2564 ของห้องปฏิบัติการจุลชีววิทยา โรงพยาบาลสัตว์เล็ก คณะสัตวแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย โดยนำชุดข้อมูลมาแบ่งออกเป็น 2 ชุด คือ ชุดข้อมูลเบคทีเรียแกรมลบ และชุดข้อมูลเบคทีเรียแกรมบวก ซึ่งแบ่งตามชนิดของการคัดที่ใช้ตรวจผลการทดสอบความไวต่อยาต้านจุลชีพด้วยเครื่อง Vitek

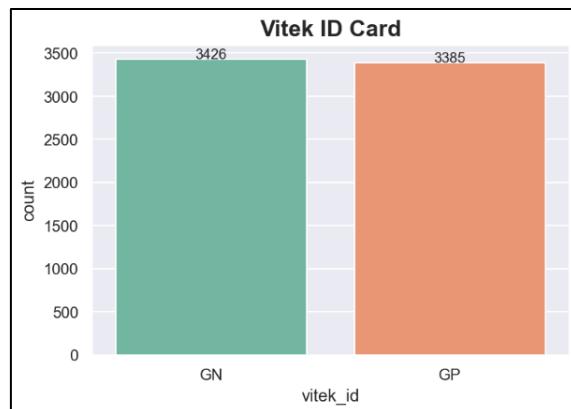
ชุดข้อมูลที่ใช้ประกอบไปด้วยคอลัมน์ ดังนี้

- hn คือ รหัสของสัตว์ป่วย
- date\_of\_submission คือ วันที่ส่งตรวจ
- report\_issued\_date คือ วันที่เขียนรายงาน
- vitek\_id คือ ชนิดของการคัดที่ใช้ตรวจผลการทดสอบความไวต่อยาด้วยเครื่อง Vitek
- species คือ ชนิดของสัตว์เลี้ยง
- submitted\_sample คือ ตัวอย่างที่ส่งตรวจ
- bacteria\_species คือ สายพันธุ์เบคทีเรีย
- bacteria\_genus คือ สกุลเบคทีเรีย (เป็นฟีเจอร์ที่สร้างขึ้นมาใหม่จาก bacteria\_species)
- S/I/R\_[ชื่อยาต้านจุลชีพ] คือ ผลตรวจความไวต่อยาต้านจุลชีพแต่ละชนิด
- ans\_[ชื่อยาต้านจุลชีพ] คือ ยาต้านจุลชีพที่สัตวแพทย์แนะนำ

A	B	C	D	E	F	G	H	I	J	K	L	M	N
hn	date_of_submission report_issued_date	vitek_id	species	submitted_sample	bacteria_species	bacteria_genus	S/I/R_amika	S/I/R_amox	S/I/R_marbo	ans_amika	ans_amox	ans_cefal	
1	6303	19/3/2020	26/3/2020 GN	cat	urine	klebsiella pneum klebsiella	R	R	R	FALSE	FALSE	FALSE	
3	6210	20/2/2020	25/2/2020 GN	cat	nasal cavity	pasteurella canis pasteurella	S	S	S	FALSE	TRUE	TRUE	
4	6201	25/2/2019	28/2/2019 GN	dog	urine	escherichia coli	escherichia	S	S	R	FALSE	FALSE	TRUE
5	6210	20/1/2020	23/1/2020 GN	cat	opened wound	pseudomonas ae pseudomonas	S		R	TRUE	FALSE	FALSE	
6	6014	30/10/2017	6/11/2017 GN	dog	abdominal fluid	escherichia coli	escherichia	S	R	R	TRUE	FALSE	FALSE
7	5801	11/1/2018	17/1/2018 GP	dog	ub mucosa	staphylococcus a staphylococcus		S	S	FALSE	TRUE	TRUE	
8	6004	26/5/2017	1/6/2017 GP	dog	surgical site	staphylococcus l staphylococcus		S	R	FALSE	FALSE	FALSE	
9	6019	28/12/2017	4/1/2018 GP	cat	wound	gemella morbillic gemella				FALSE	TRUE	TRUE	
10	6019	29/12/2017	8/1/2018 GP	cat	wound	enterococcus hir enterococcus		S	R	FALSE	TRUE	FALSE	
11	5907	9/7/2019	12/7/2019 GP	dog	wound	staphylococcus a staphylococcus		S	R	FALSE	FALSE	FALSE	

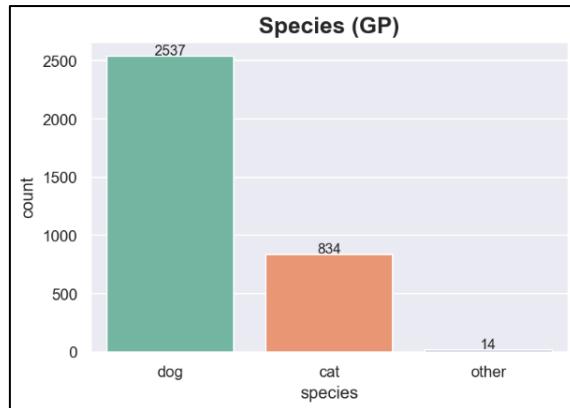
รูปที่ 3.4 ตัวอย่างชุดข้อมูลที่ใช้

3.4.1 ชุดข้อมูลแบบที่เรียกว่าแบบนวนและชุดข้อมูลแบบที่เรียกว่าแบบลบ โดยแบ่งตามชนิดของการ์ดที่ใช้ตรวจสอบผลการทดสอบความไวต่อยาต้านจุลชีพด้วยเครื่อง Vitek (Vitek ID Card)

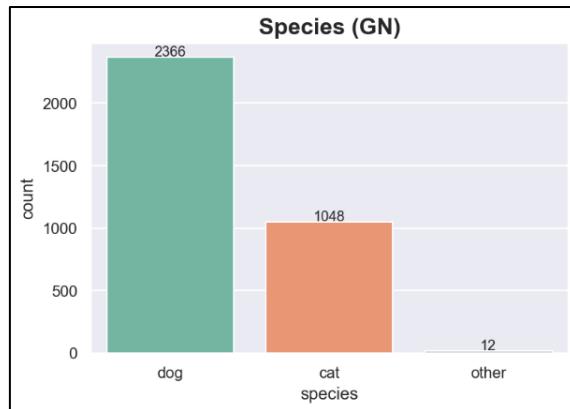


รูปที่ 3.5 แผนภูมิแท่งแสดงจำนวนชุดข้อมูลแบบที่เรียกว่าแบบนวนและแบบที่เรียกว่าแบบลบ โดยแบ่งตามชนิดของการ์ดที่ใช้ตรวจสอบ

### 3.4.2 ชนิดของสัตว์เลี้ยง (Species)



(ก)



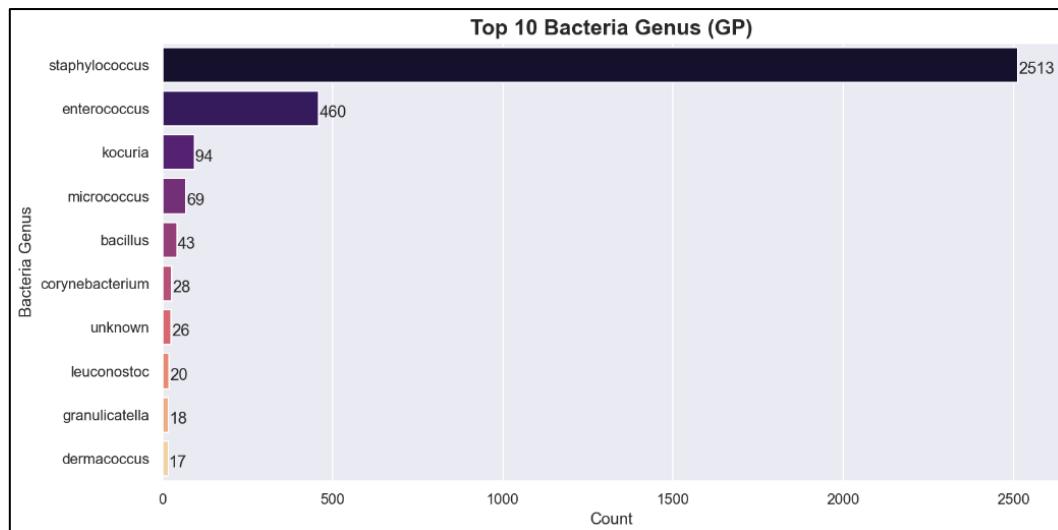
(ข)

รูปที่ 3.6 แผนภูมิแท่งแสดงจำนวนข้อมูลชนิดของสัตว์เลี้ยง

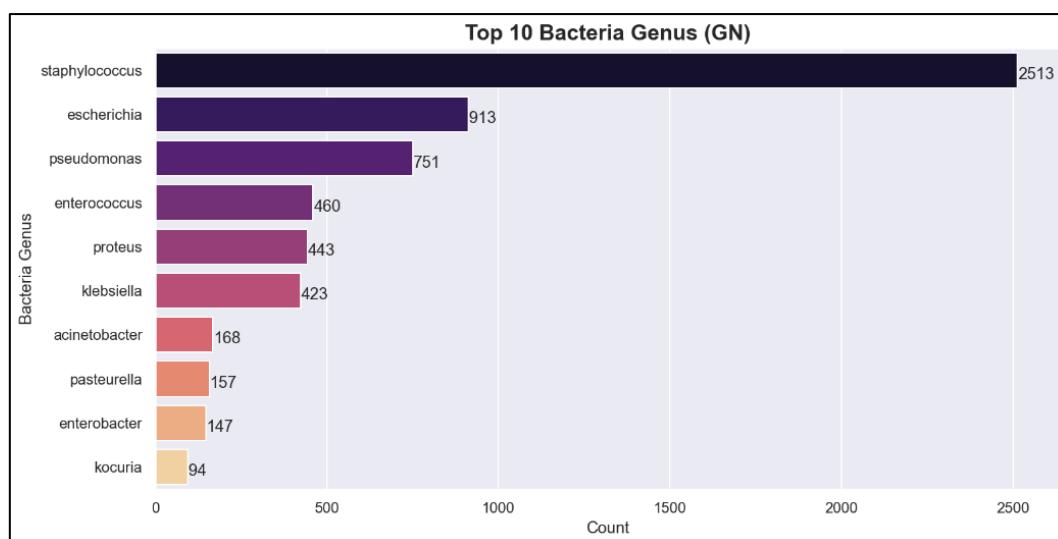
ก) ชุดข้อมูลแบนค์ที่เรียกว่ารวมบวก

ข) ชุดข้อมูลแบนค์ที่เรียกว่ารวมลบ

### 3.4.3 ສຸກລແບບທີ່ເຮືອຍ (Bacteria Genus)



(ก)



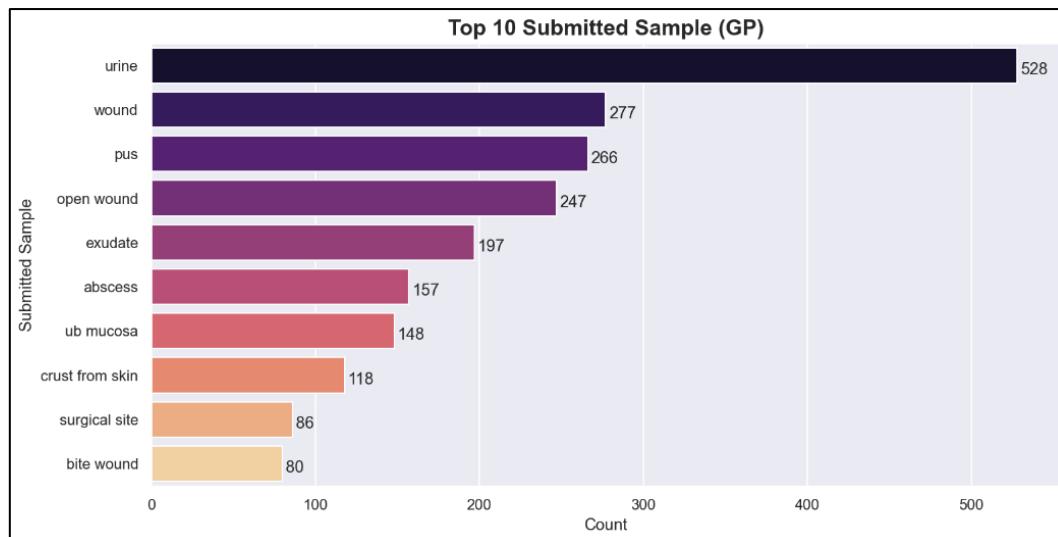
(ข)

ຮັບທີ່ 3.7 ແຜນກົມິແທ່ງແສດງຈຳນວນຂໍອມູລສຸກລແບບທີ່ເຮືອຍ 10 ອັນດັບແຮກ

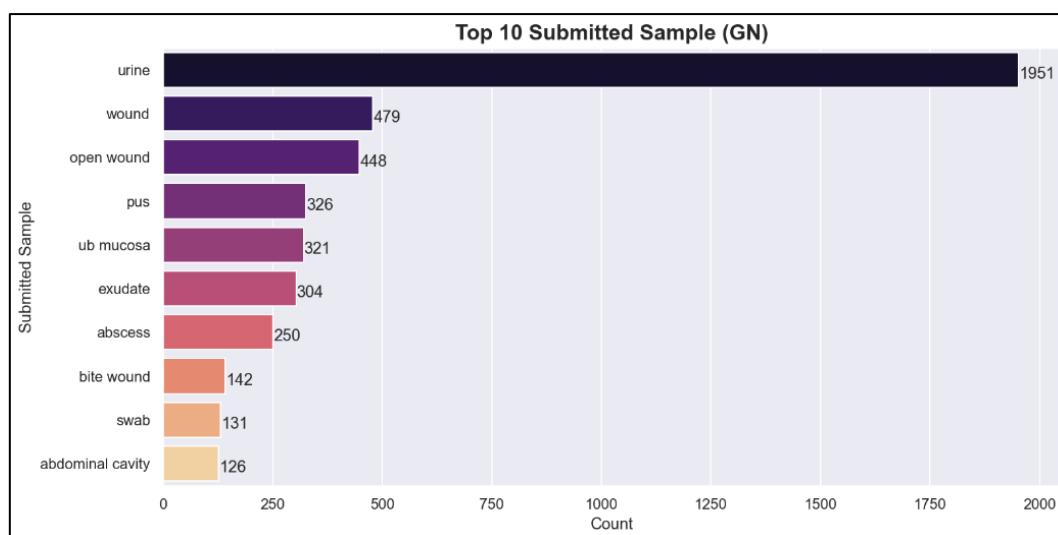
ກ) ຊຸດຂໍ້ອມູລແບບທີ່ເຮືອຍແກຣມບວກ

ຂ) ຊຸດຂໍ້ອມູລແບບທີ່ເຮືອຍແກຣມລົບ

### 3.4.4 ตัวอย่างที่ส่งตรวจ (Submitted Sample)



(ก)



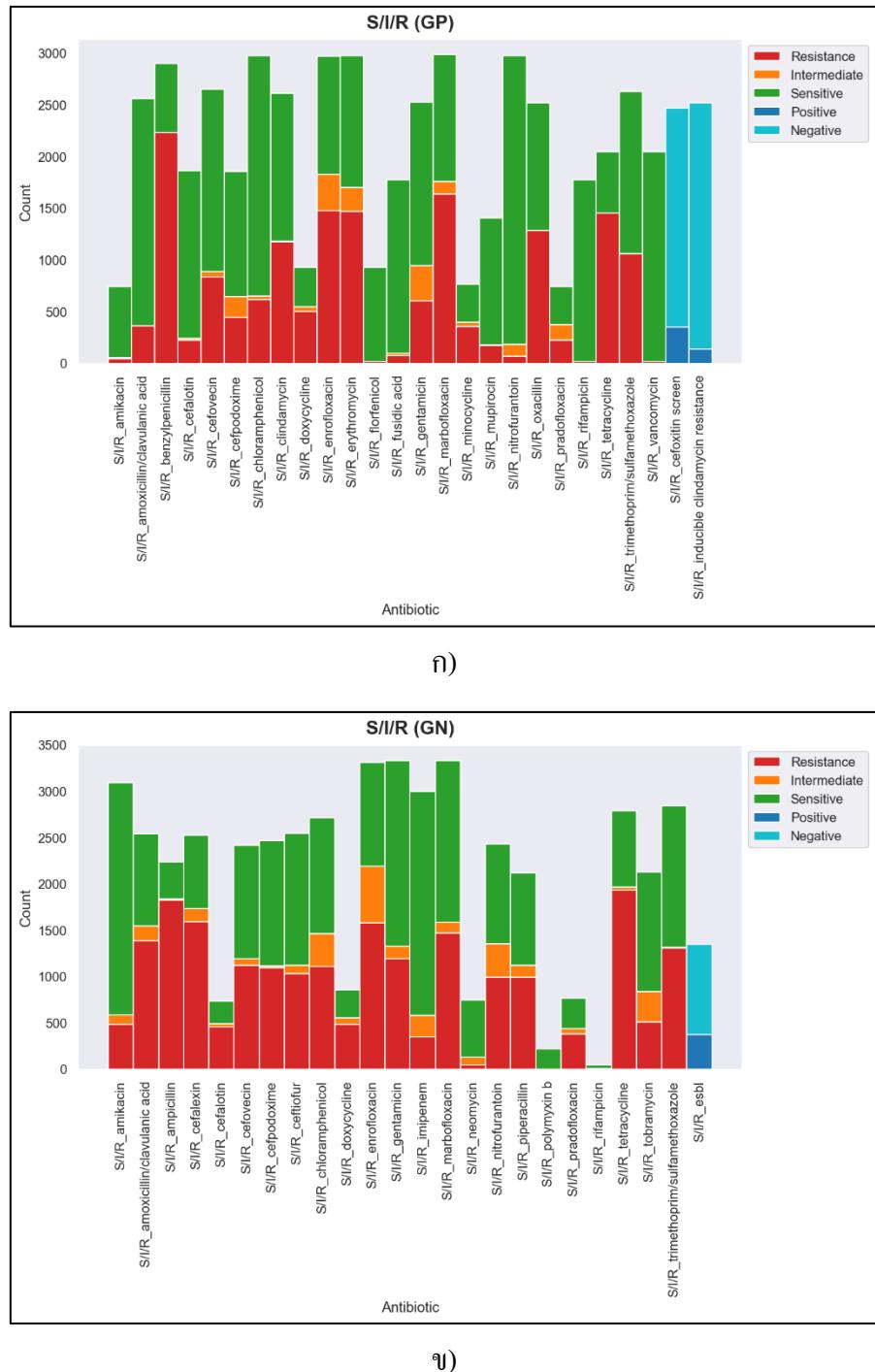
(ข)

รูปที่ 3.8 แผนภูมิแท่งแสดงจำนวนข้อมูลตัวอย่างที่ส่งตรวจ 10 อันดับแรก

ก) ชุดข้อมูลแบบที่เรียกว่าบวก

ข) ชุดข้อมูลแบบที่เรียกว่าลบ

### 3.4.5 ผลการทดสอบความไวต่อยาต้านจุลชีพ (Result of Antimicrobial Susceptibility Testing)

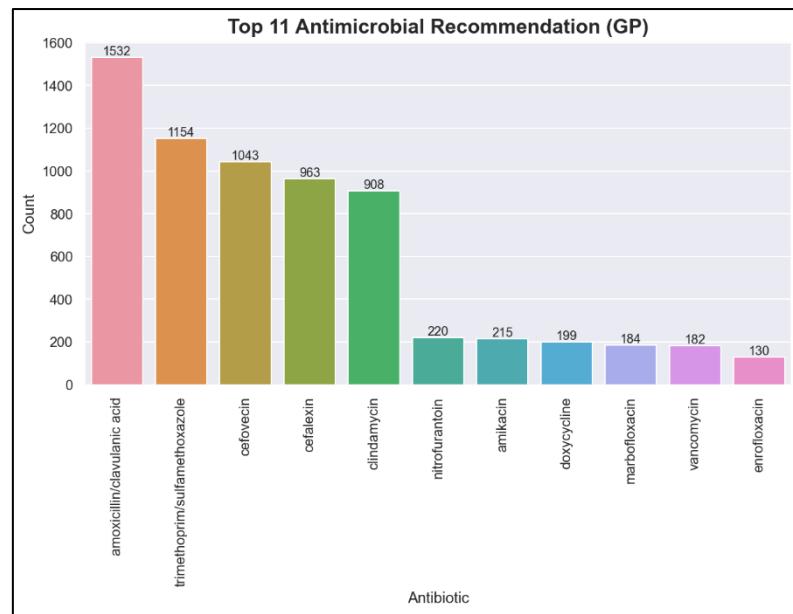


รูปที่ 3.9 แผนภูมิแท่งแสดงจำนวนข้อมูลผลการทดสอบต่อยาต้านจุลชีพแต่ละชนิด

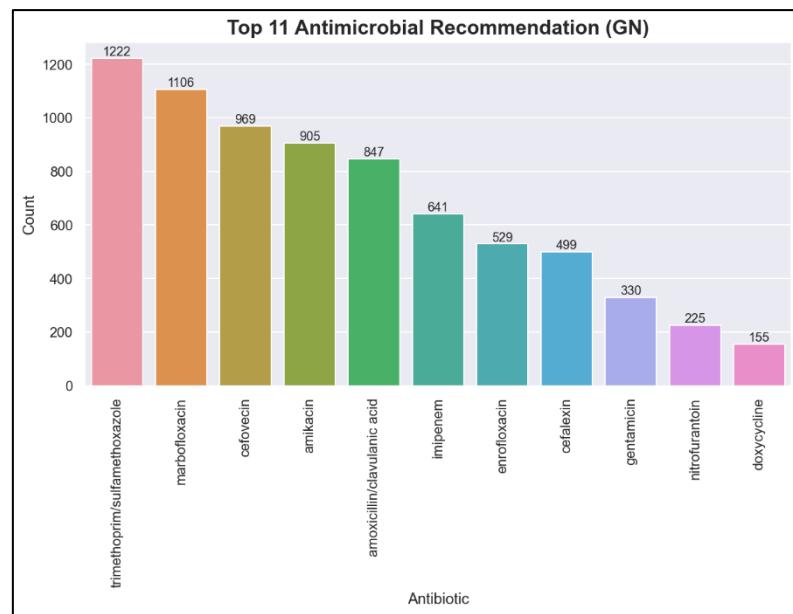
ก) ชุดข้อมูลแบบที่เรียกว่ารูปแบบบาก

ข) ชุดข้อมูลแบบที่เรียกว่ารูปแบบคลบ

### 3.4.6 ยาต้านจุลชีพที่ได้สร้างໂມದລເພື່ອແນະນຳ (Antibiotic)



(ii)



(iii)

ຮັບກີ່ 3.10 ແຜນກຸມແທ່ງແສດງຈຳນວນຄວັງທີ່ມີການໃຊ້ຢາຕ້ານຈຸລື້ພ

ກ) ຜູດຂໍ້ອມຸລແບບທີ່ເຮີຍແກຣນບາກ

ຂ) ຜູດຂໍ້ອມຸລແບບທີ່ເຮີຍແກຣນຄົມ

### 3.5 ภาพรวมของระบบ

โครงการนี้ได้ทำการออกแบบเว็บแอปพลิเคชันสำหรับทดสอบการเรียกใช้งานโมเดลแนะนำข่ายต้านจุลชีพ และการอัปโหลดข้อมูลเพื่อฝึกฝนโมเดล ซึ่งจะแบ่งออกเป็น 3 ส่วน ดังนี้

#### 3.5.1 การแนะนำการใช้ยาต้านจุลชีพ

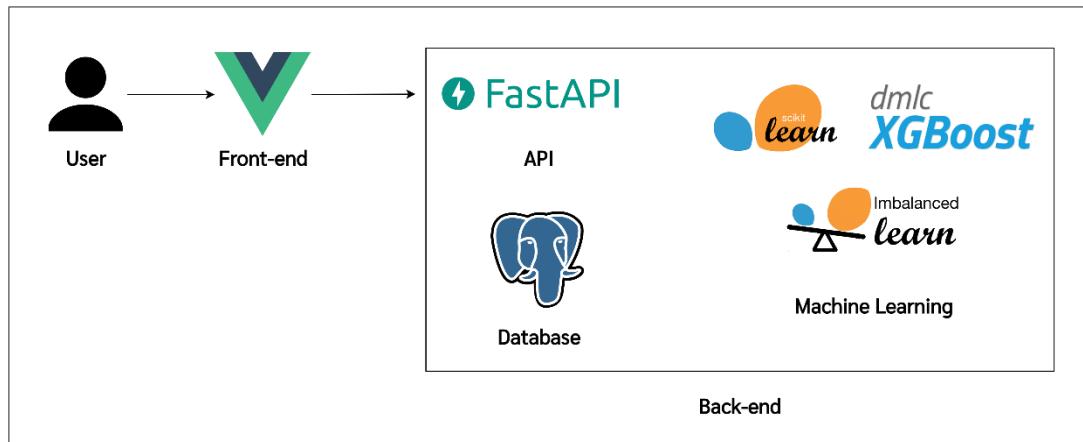
ผู้ใช้งานจะทำการกรอกข้อมูลผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยาได้แก่ ชนิดของสัตว์ (Species), ชนิดของแบคทีเรีย (Bacteria Genus), ตัวอย่างที่ส่งตรวจ (Submitted Sample), ชนิดของการคัดที่ใช้ตรวจด้วยเครื่อง Vitek (Vitek ID Card) และผลตรวจความไวต่อยาแต่ละชนิด (S/I/R) แล้วกดปุ่มส่งข้อมูลเพื่อที่จะให้โมเดลแนะนำยา โดยทางระบบจะทำการนำข้อมูลเข้าไปประมวลผลโดยจัดรูปแบบของข้อมูลที่รับมาให้เป็นข้อมูลขาเข้า (Input) สำหรับโมเดลจากนั้นส่งข้อมูลเข้าไปให้แต่ละโมเดล จะได้เป็นผลลัพธ์ออกมาเป็นชื่อยาต้านจุลชีพที่แนะนำให้ใช้และความมั่นใจของโมเดลเป็นเปอร์เซ็นต์ แล้วระบบจะส่งผลลัพธ์กลับไปให้ผู้ใช้

#### 3.5.2 การอัปโหลดไฟล์เข้าสู่ระบบเพื่อฝึกฝนโมเดล

ผู้ใช้งานจะทำการเลือกไฟล์แล้วอัปโหลดไฟล์เข้าสู่ระบบ จากนั้นทางระบบจะทำการตรวจสอบความถูกต้องของไฟล์ที่ผู้ใช้อัปโหลด ในการนี้ที่ไฟล์ต้องตามเงื่อนไขก็จะอัปโหลดสำเร็จ แล้วทำการบันทึกข้อมูลไว้ในระบบ แต่ถ้าไม่ตรงตามเงื่อนไขก็จะอัปโหลดไม่สำเร็จ แล้วทำการส่งข้อผิดพลาดในการอัปโหลดไฟล์กลับไปให้ผู้ใช้ทราบ หากผู้ใช้ต้องการลบไฟล์ที่ได้อัปโหลดไปสามารถทำได้โดยจะต้องเข้าไปที่หน้าจัดการไฟล์และฝึกฝนโมเดล จากนั้นก็ทำการเลือกไฟล์ แล้วคลิกปุ่มลบไฟล์

#### 3.5.3 การฝึกฝนโมเดลใหม่เพื่ออัปเดตโมเดล

เมื่อต้องการฝึกฝนโมเดลใหม่ ผู้ใช้งานจะต้องคลิกปุ่มทำ “Retraining Now” ในหน้าจัดการไฟล์และฝึกฝนโมเดล เมื่อคลิกแล้วระบบจะทำการตรวจสอบว่ามีข้อมูลใหม่สำหรับฝึกฝนโมเดลใหม่หรือไม่ หากมีก็จะนำข้อมูลใหม่ไปรวมกับข้อมูลเดิมแล้วทำการฝึกฝนโมเดล รวมทั้งวัดประสิทธิภาพของโมเดลที่ได้ หากโมเดลใหม่มีประสิทธิภาพดีกว่าหรือเท่ากับโมเดลปัจจุบันก็จะทำการเปลี่ยนเป็นโมเดลใหม่ แต่หากประสิทธิภาพต่ำกว่าโมเดลเดิมจะยังคงใช้โมเดลเดิมต่อไป (ประเมินผลด้วยค่า F1-Score)



รูปที่ 3.11 ภาพรวมของระบบ

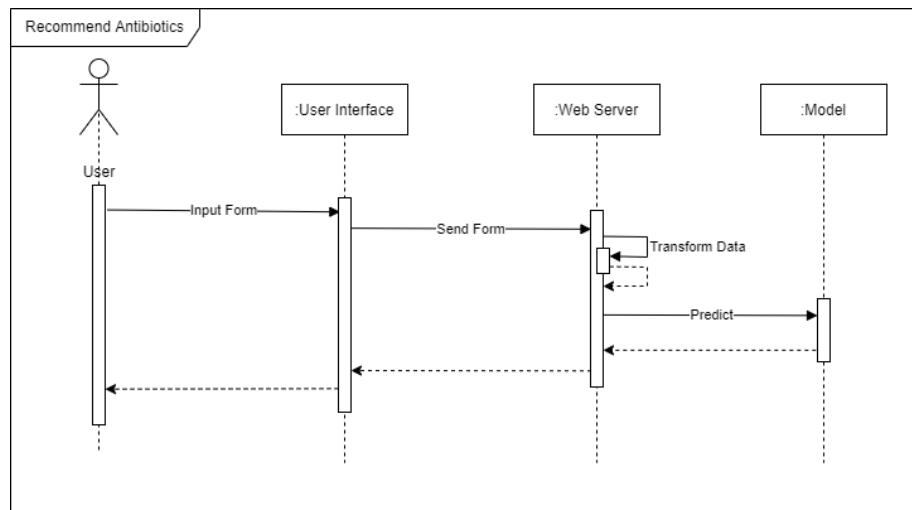
### 3.6 การออกแบบและพัฒนาระบบ

#### 3.6.1 การออกแบบ Use Case Diagram

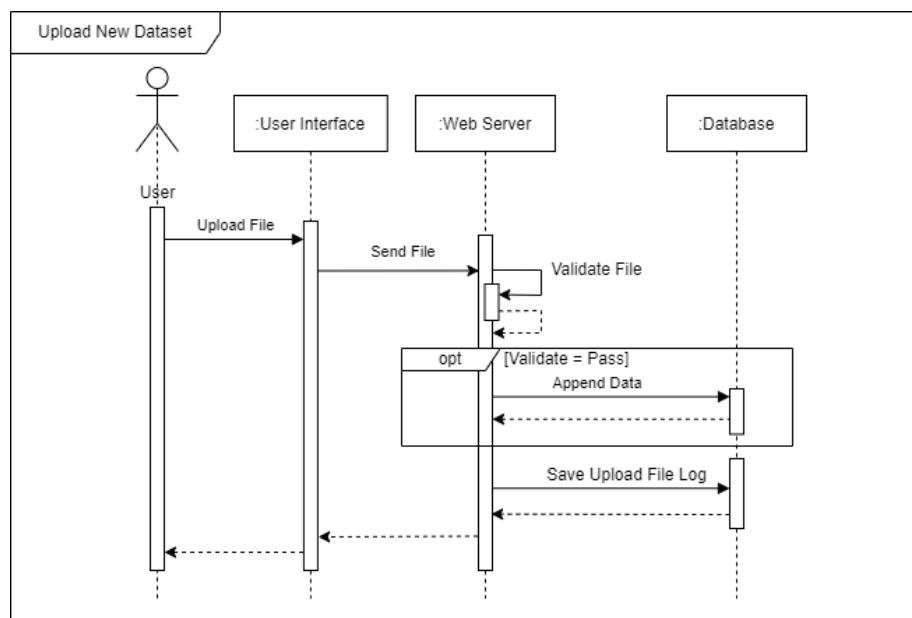


รูปที่ 3.12 Use Case Diagram

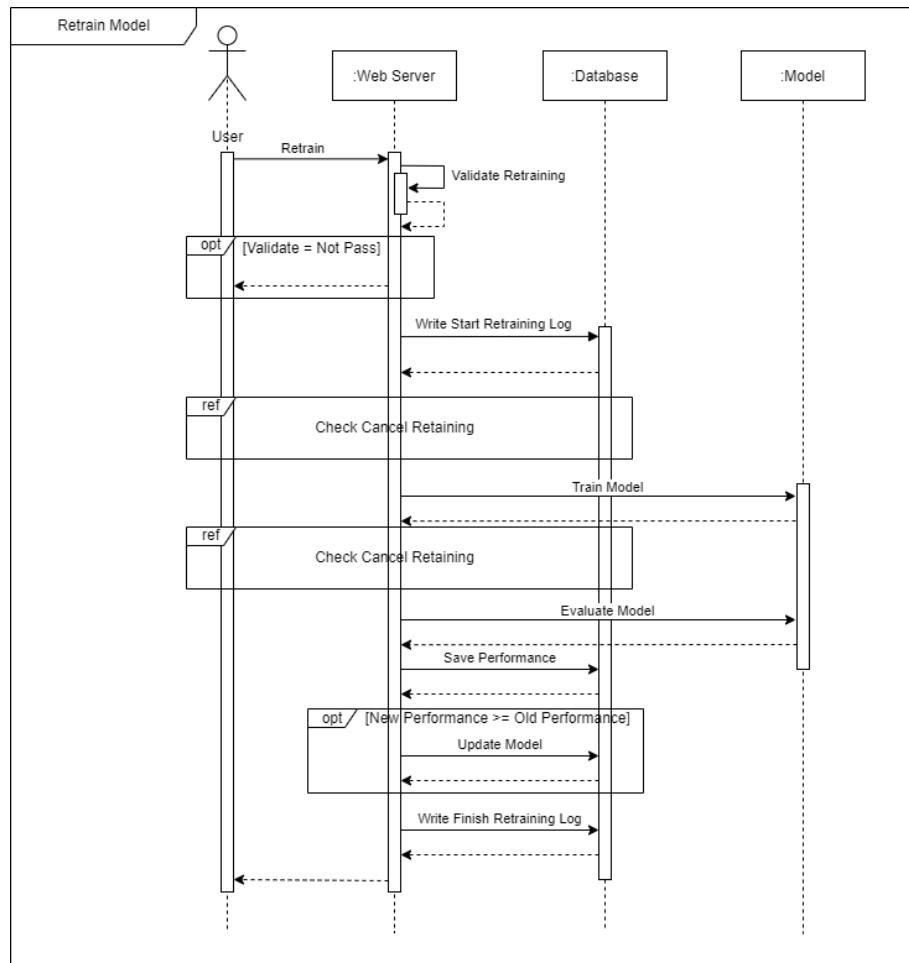
### 3.6.2 การออกแบบ Sequence Diagram



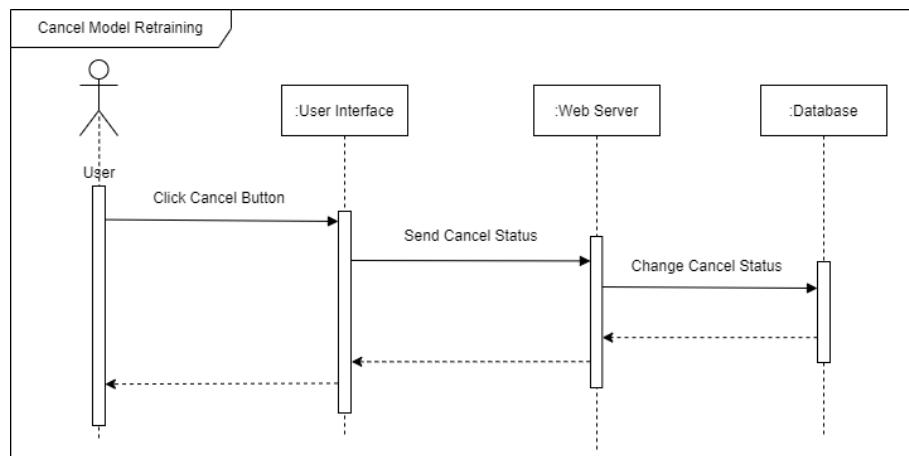
รูปที่ 3.13 Sequence Diagram ของการแนะนำยาต้านจุลชีพ



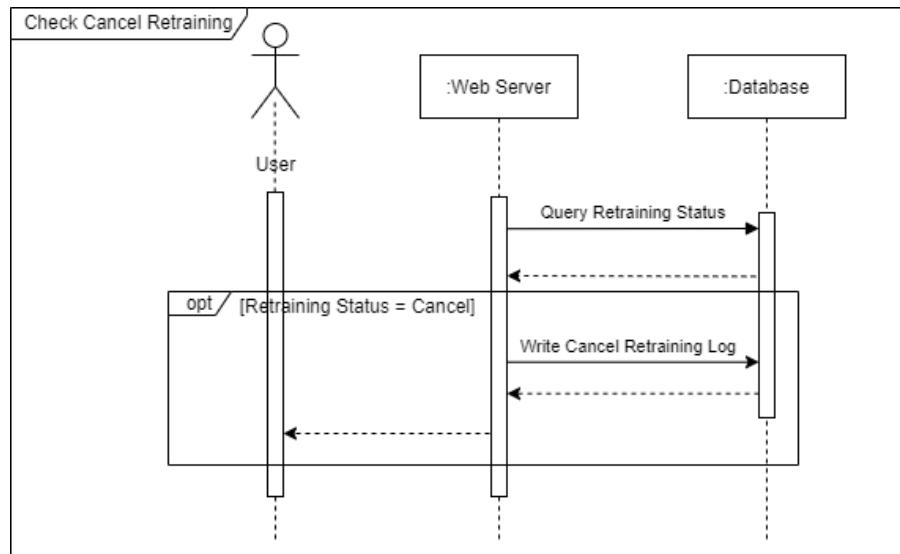
รูปที่ 3.14 Sequence Diagram ของการอัปโหลดไฟล์เพื่อฝึกสอนโมเดลใหม่



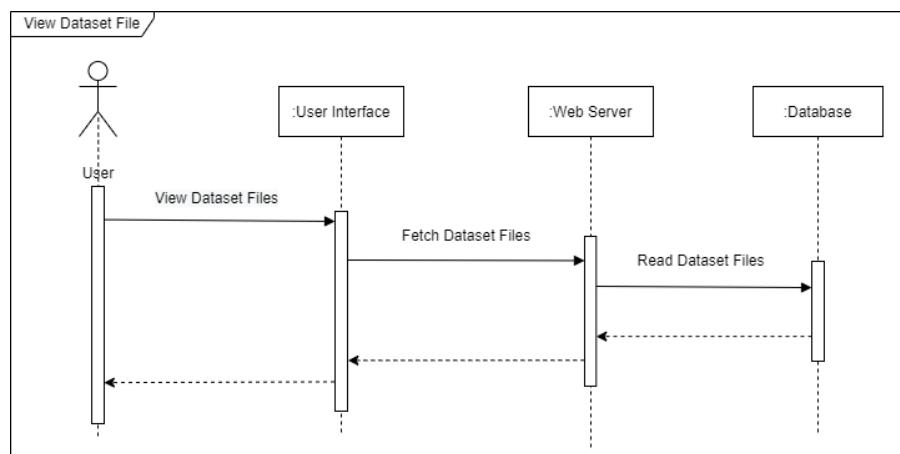
รูปที่ 3.15 Sequence Diagram ของฝึกฝนโมเดลใหม่



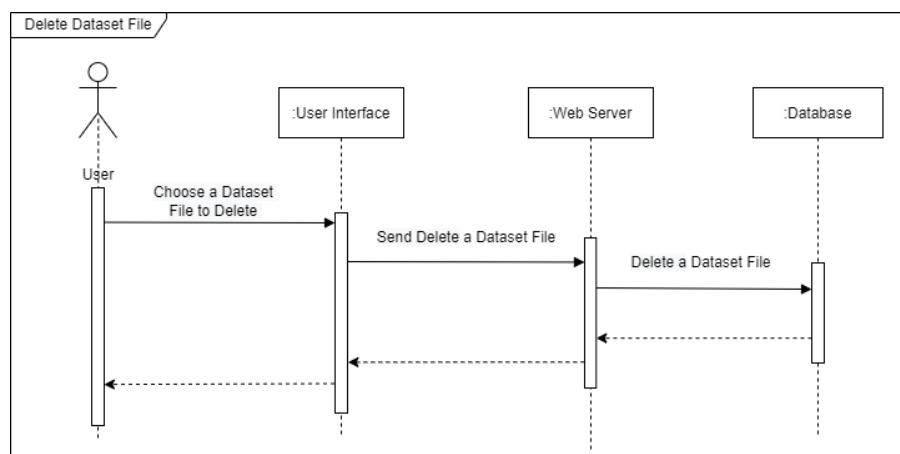
รูปที่ 3.16 Sequence Diagram ของการยกเลิกการฝึกฝนโมเดล



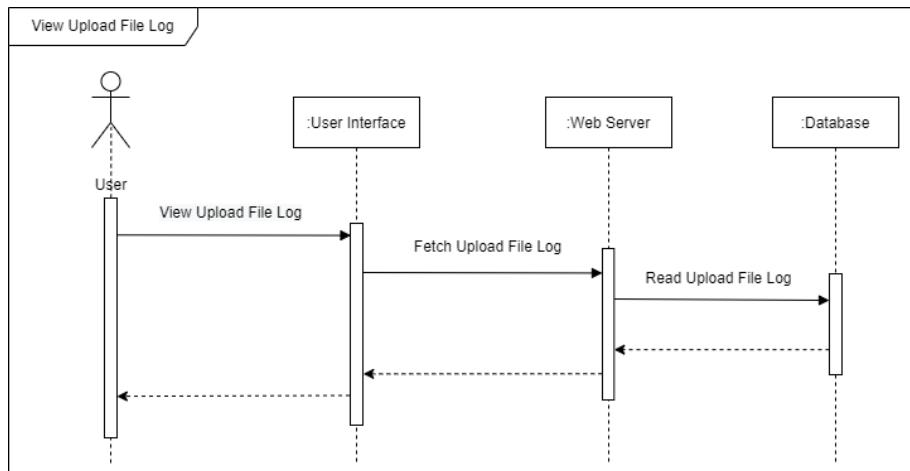
รูปที่ 3.17 Sequence Diagram ของการตรวจสอบการยกเลิกการฝึกฝนโมเดล



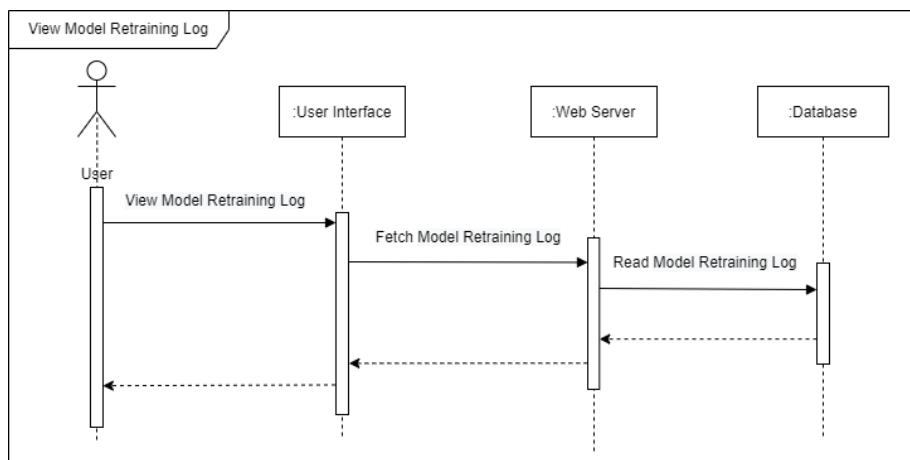
รูปที่ 3.18 Sequence Diagram ของการดูไฟล์ชุดข้อมูลสำหรับฝึกฝนโมเดลที่มีอยู่ในระบบ



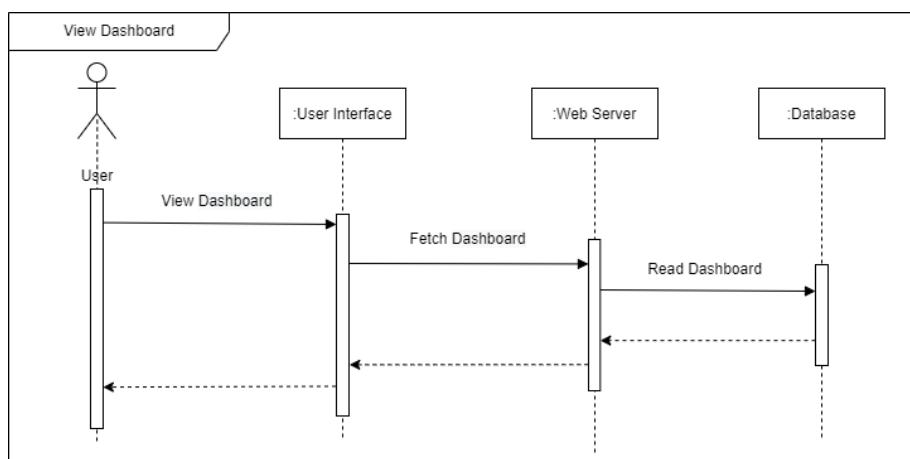
รูปที่ 3.19 Sequence Diagram ของการลบไฟล์ชุดข้อมูลสำหรับฝึกฝนโมเดล



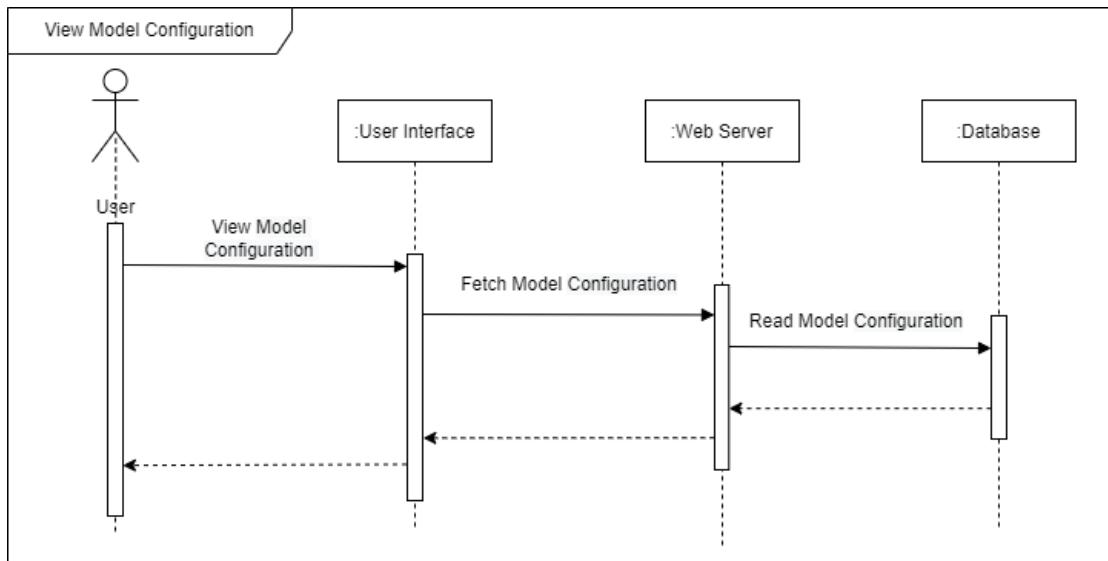
รูปที่ 3.20 Sequence Diagram ของการดูข้อมูลบันทึกการอัปโหลดไฟล์



รูปที่ 3.21 Sequence Diagram ของการดูข้อมูลบันทึกการฝึกฝนโมเดล

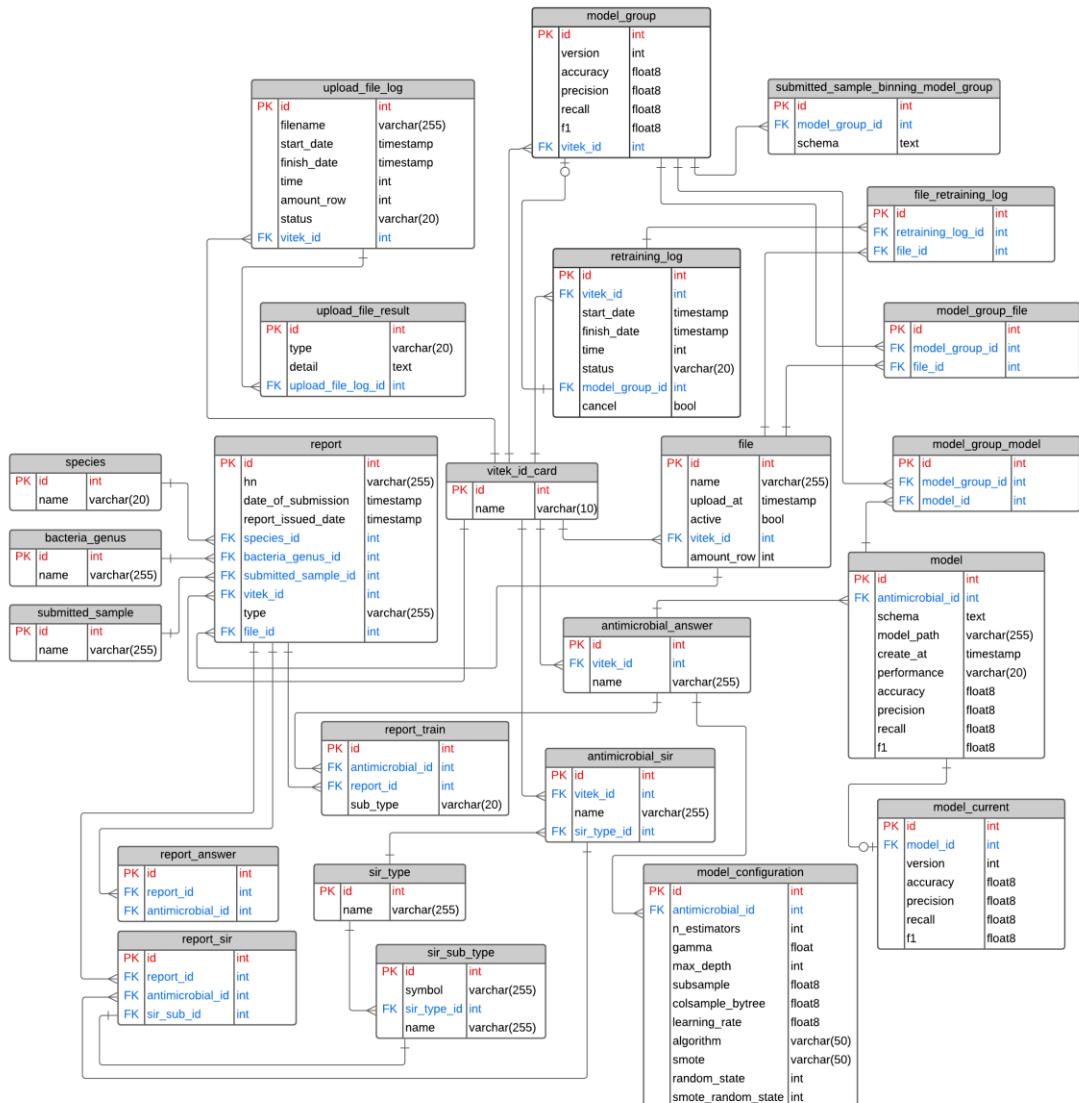


รูปที่ 3.22 Sequence Diagram ของการดูแดชบอร์ด



รูปที่ 3.23 Sequence Diagram ของการดูการกำหนดค่าของโมเดล

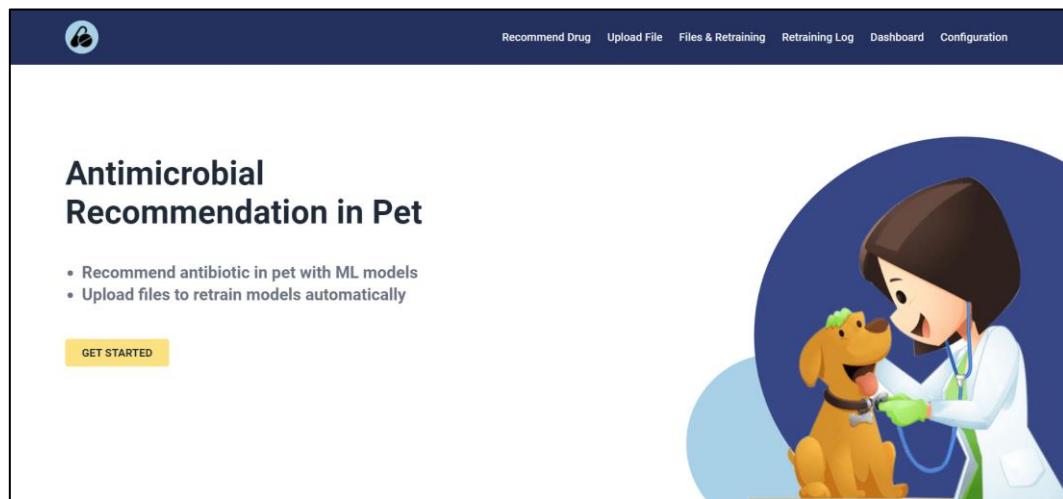
### 3.6.3 ດາວອອກແບບ ER Diagram



ສັບຖິ່ນ 3.24 ER Diagram

### 3.6.4 การออกแบบส่วนติดต่อผู้ใช้ (User Interface)

1) หน้าแรกของเว็บแอปพลิเคชัน เมื่อผู้ใช้เข้าสู่เว็บแอปพลิเคชันจะแสดงหน้าแรกขึ้นมาโดยมีปุ่ม “GET START” สำหรับเริ่มต้นใช้งาน เมื่อคลิกจะเข้าสู่หน้าแนะนำต้านจุลชีพ



รูปที่ 3.25 หน้าแรกของเว็บแอปพลิเคชัน

2) หน้าแนะนำต้านจุลชีพ โดยผู้ใช้จะต้องกรอกข้อมูลผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยา จากนั้นกดปุ่ม “Recommend” เพื่อส่งข้อมูลไปให้โมเดลทำการแนะนำต้านจุลชีพ จากนั้นระบบจะส่งผลลัพธ์กลับมาให้ผู้ใช้เป็นชื่อยาต้านจุลชีพพร้อมกับความมั่นใจในการแนะนำของโมเดลเป็นเปอร์เซ็นต์

Antimicrobial	S/I/R	Antimicrobial	S/I/R
Amikacin	S	Gentamicin	R
Aztreonam	R	Imipenem	R
Ampicillin	S	Marbofloxacin	S
Cefalexin	R	Neomycin	I
Cefalotin	R	Nitrofurantoin	R
Cefovicin	R	Piperacillin	
Cefpodoxime	R	Polymyxin b	
Cetisotur	I	Pradofloxacin	I
Chloramphenicol	R	Rifampicin	
Doxycycline	R	Tetracycline	R
Enrofloxacin	S	Tobramycin	
ESBL		Trimethoprim/sulfamethoxazole	R

รูปที่ 3.26 หน้าแนะนำต้านจุลชีพ

3) หน้าอัปโหลดไฟล์สำหรับฝึกฝนโมเดลใหม่ โดยผู้ใช้จะต้องอัปโหลดไฟล์ .csv ที่มีเงื่อนไขตามที่ระบบได้กำหนดไว้ดังนี้

- ไฟล์จะต้องมีจำนวนแถวอย่างน้อย 300 แถว
- ไฟล์จะต้องมีคอลัมน์ ดังนี้ hn (รหัสสัตว์ป่วย), date\_of\_submission (วันที่ส่งตรวจ), report\_issued\_date (วันที่เขียนรายงาน), species (ชนิดของสัตว์), bacteria\_genus (สกุลแบปค์ทีเรีย), submitted\_sample (ตัวอย่างที่ส่งตรวจ), vitek\_id (ชนิดของการรักษาที่ใช้ตรวจด้วยเครื่อง Vitek), S/I/R\_[ชื่อยาต้านจุลชีพ] (ผลตรวจความไวต่อยาต้านจุลชีพแต่ละชนิด), ans\_[ชื่อยาต้านจุลชีพ] (ยาต้านจุลชีพที่แนะนำ)
- คอลัมน์ date\_of\_submission และ report\_issued\_date จะต้องใช้รูปแบบวันที่เป็น yyyy-mm-dd
  - คอลัมน์ ans\_[ชื่อยาต้านจุลชีพ] จะต้องมีค่าเป็น True หรือ False
  - ไฟล์จะต้องไม่มีแถวที่ซ้ำกันในไฟล์และในฐานข้อมูล เมื่อผู้ใช้คลิกปุ่ม “Upload” ระบบจะทำการตรวจสอบไฟล์ว่ามีความถูกต้องหรือไม่ หากถูกต้องจะทำการอัปโหลดไฟล์เข้าสู่ระบบ หากไม่ถูกต้องระบบจะส่งข้อผิดพลาดกลับไปให้ผู้ใช้ โดยหน้าเว็บจะแสดงตารางข้อมูลบันทึกการอัปโหลดไฟล์ (Upload Files Log) สำหรับให้ผู้ใช้ดูสถานะการอัปโหลดไฟล์ เมื่อคลิกปุ่ม “Result” จะแสดงป็อบอัพ (Popup) ผลลัพธ์ของการอัปโหลดไฟล์

The screenshot shows a web application interface for file upload. At the top, there's a navigation bar with links: Recommend Drug, Upload File, Files & Retraining, Retraining Log, Dashboard, and Configuration. Below the navigation is a section titled "Upload File" with a large dashed box for dragging files. A "Choose a File" button is below the box. There are two radio buttons: "GN" and "GP". A blue "Upload" button is at the bottom of the file area. Below this is a section titled "File Upload Log" with a table. The table has columns: #, Filename, Vitek ID, Start Date, Finish Date, Time, Total Rows, and Status. It lists three entries:

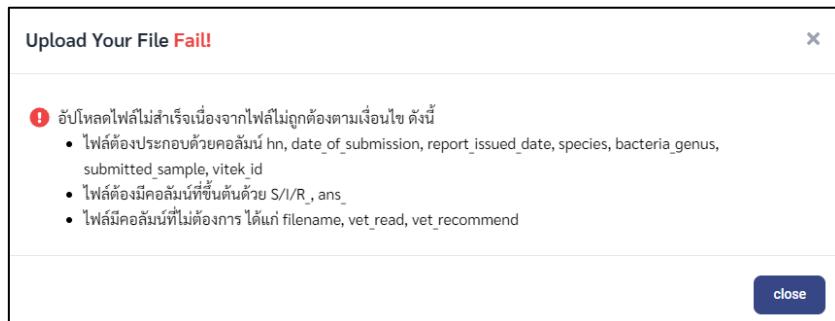
#	Filename	Vitek ID	Start Date	Finish Date	Time	Total Rows	Status
1	gn_2021.csv	GN	03-Apr-2022 19:34:30	-	-	-	<span style="color: blue;">● Uploading</span>
2	gp_2021.csv	GP	03-Apr-2022 19:34:19	03-Apr-2022 19:34:22	00:00:03	500	<span style="color: green;">● Success</span>
3	gp_2021.csv	GN	03-Apr-2022 19:33:59	03-Apr-2022 19:33:59	00:00:00	0	<span style="color: red;">● Fail</span>

Below the table, it says "Showing 1 to 3 of 3 entries." On the right side of the log table, there are three "Result" buttons, each corresponding to one of the log entries.

รูปที่ 3.27 หน้าอัปโหลดไฟล์สำหรับฝึกฝนโมเดลใหม่



รูปที่ 3.28 ป้อนอัพแสดงผลลัพธ์การอัปโหลดไฟล์สำเร็จ



รูปที่ 3.29 ป้อนอัพแสดงผลลัพธ์การอัปโหลดไฟล์ไม่สำเร็จ

4) หน้าจัดการไฟล์และฝึกฝนโมเดล (File & Retraining) เป็นหน้าที่แสดงไฟล์ต่าง ๆ ในระบบ และมีปุ่มสำหรับฝึกฝนโมเดลใหม่ โดยจะแสดงรายละเอียดของไฟล์ที่ผู้ใช้ได้ทำการอัปโหลดเข้าสู่ระบบ ซึ่งจะมีปุ่มให้ผู้ใช้สามารถลบไฟล์ได้ และมีปุ่มสำหรับให้ผู้ใช้คลิกเมื่อต้องการจะฝึกฝนโมเดลใหม่ เมื่อคลิกที่ปุ่ม “Retraining Now” ระบบจะทำการตรวจสอบว่ามีข้อมูลใหม่สำหรับฝึกฝนโมเดลหรือไม่ หากมีข้อมูลใหม่ก็จะทำการเริ่มฝึกฝนโมเดล

#	Filename	Vitek ID	Upload Date	Total Rows	Delete
1	gn_2021.csv	GN	03-Apr-2022 19:34:30	500	
2	gp_2021.csv	GP	03-Apr-2022 19:34:19	500	
3	origin	GN	31-Dec-2021 00:00:00	3426	
4	origin	GP	31-Dec-2021 00:00:00	3385	

รูปที่ 3.30 หน้าจัดการไฟล์และฝึกฝนโมเดล

5) หน้าแสดงข้อมูลบันทึกการฝึกฝนโมเดล โดยจะแสดงเป็นตารางซึ่งจะระบุรายละเอียดต่าง ๆ ได้แก่ กลุ่มของโมเดล (GN/GP), วันเวลาที่เริ่มและสิ้นสุดการฝึกฝนโมเดล, จำนวนเวลาที่ใช้ และสถานะของการฝึกฝนโมเดล (Training, Success, Cancel) เมื่อกlikปุ่ม “View” จะแสดงป้อนอัพรายละเอียดของไฟล์ที่ใช้ในการฝึกฝนโมเดล และเมื่อเสร็จสิ้นการฝึกฝนโมเดล จะมีปุ่ม “Dashboard” ให้กlikเพื่อเข้าสู่หน้าแดชบอร์ดแสดงประสิทธิภาพของโมเดลที่ได้ฝึกฝนใหม่ และรายละเอียดของชุดข้อมูลที่ใช้ในเวอร์ชันนี้

#	Vitek ID	Start Date	Finish Date	Time	Status	File
1	GN	03-Apr-2022 19:44:57	-	-	Training	<a href="#">View</a> <a href="#">Cancel</a>
2	GP	03-Apr-2022 19:41:25	03-Apr-2022 19:42:58	00:01:32	Success	<a href="#">View</a> <a href="#">Dashboard</a>
3	GN	03-Apr-2022 19:40:53	03-Apr-2022 19:41:25	00:00:32	Cancel	<a href="#">View</a> <a href="#">Dashboard</a>

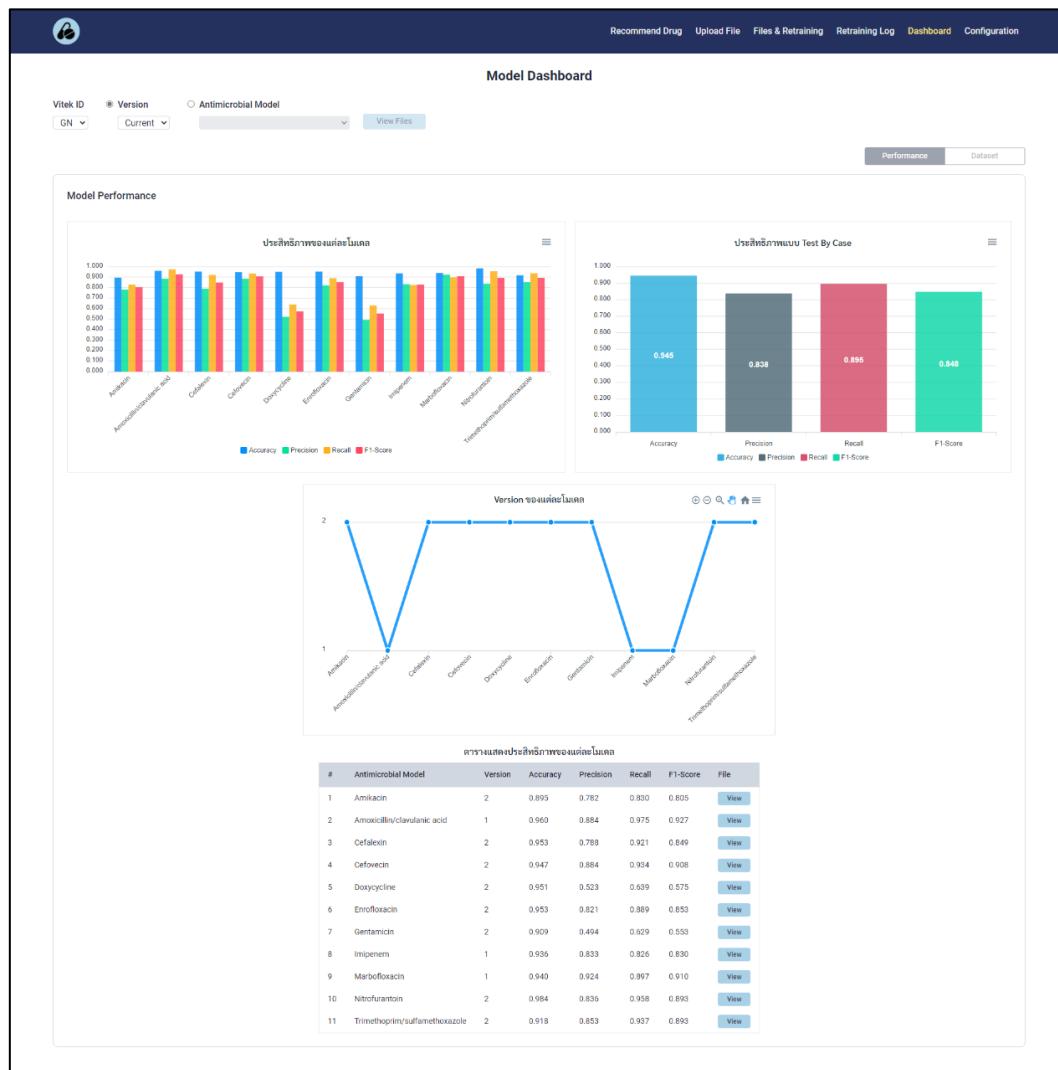
รูปที่ 3.31 หน้าแสดงข้อมูลบันทึกการฝึกฝนโมเดล



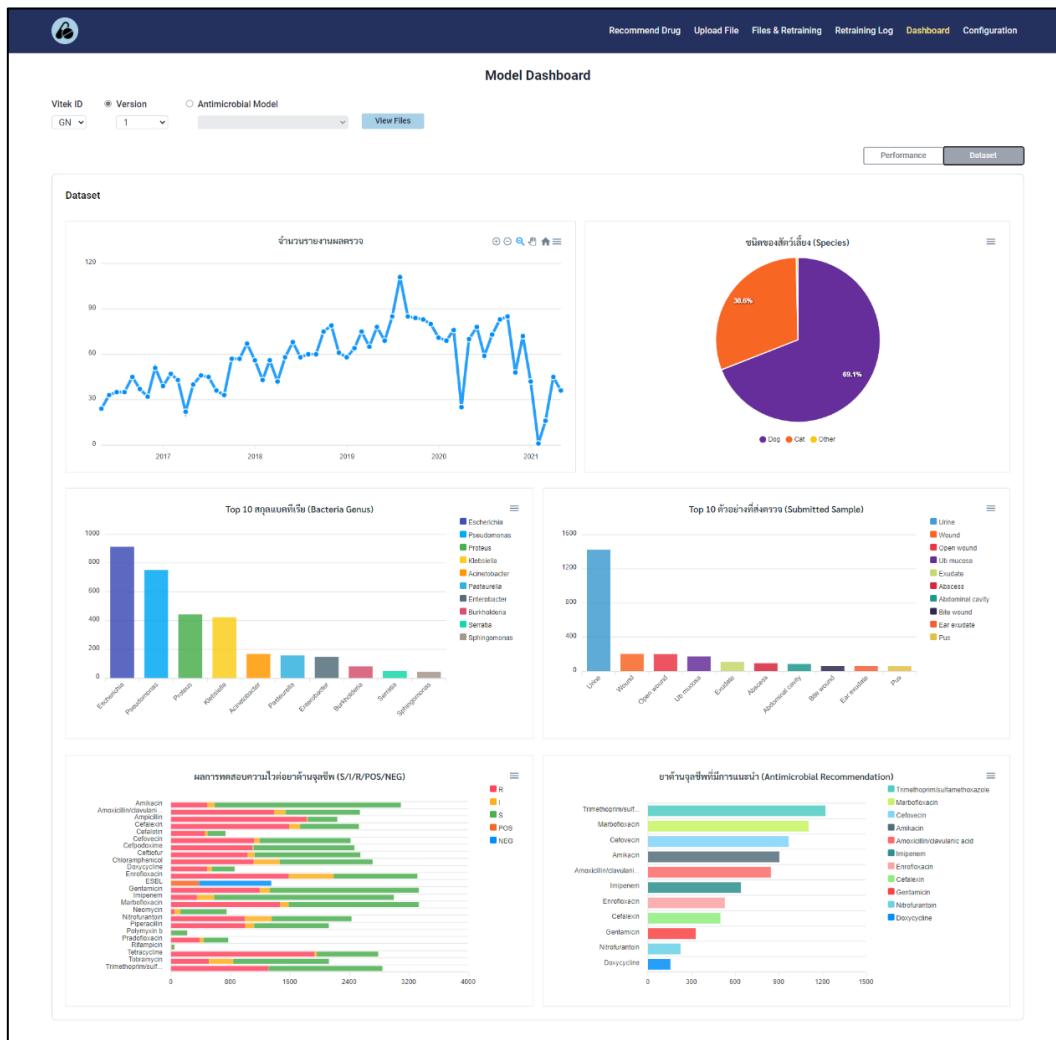
รูปที่ 3.32 ป้อนอัพแสดงรายละเอียดของไฟล์ที่ใช้ในการฝึกฝนโมเดลใหม่

6) หน้าแดชบอร์ด เป็นหน้าที่แสดงแผนภาพข้อมูลต่าง ๆ (Data Visualization) ของโมเดล โดยผู้ใช้จะสามารถเลือกดูแดชบอร์ดของโมเดลแต่ละเวอร์ชัน และแดชบอร์ดของโมเดลยาแต่ละชนิดได้ โดยหากเลือกดูแดชบอร์ดของโมเดลแต่ละเวอร์ชัน จะแสดงค่าประสิทธิภาพต่าง ๆ (Accuracy, Precision, Recall และ F1-Score) ของแต่ละโมเดล, ประสิทธิภาพแบบทดสอบรายเคส (Test by Case), เวอร์ชันของแต่ละโมเดล, รายละเอียดไฟล์ที่ใช้ในการฝึกฝนโมเดล และแสดงชุด

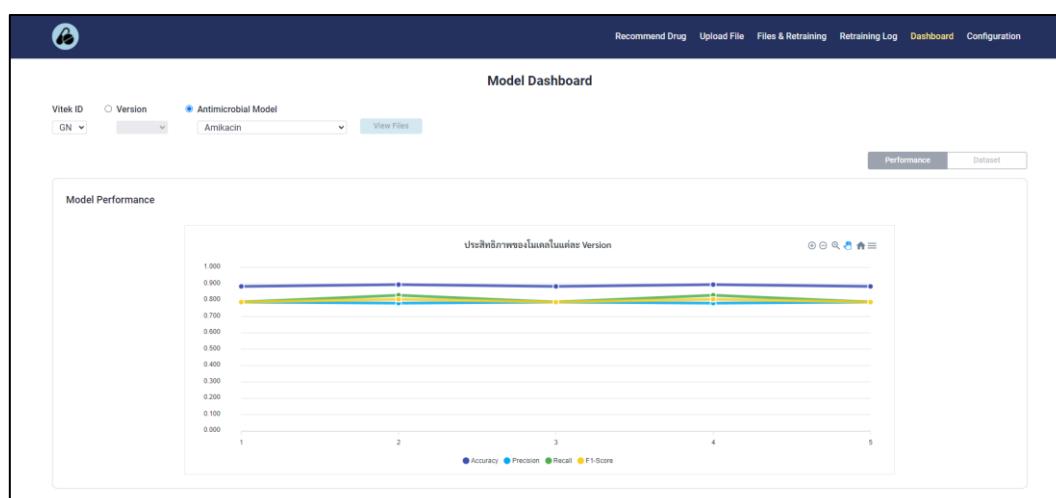
ข้อมูลที่สำหรับฝึกฝนโมเดลในแต่ละเวอร์ชัน ได้แก่ จำนวนรายงานผลตรวจ, ชนิดของสัตว์เลี้ยง, กลุ่มแบคทีเรีย 10 อันดับแรก, ตัวอย่างที่ส่งตรวจ 10 อันดับแรก, ผลการทดสอบความไวต่อยาต้านจุลชีพ และยาต้านจุลชีพที่มีการแนะนำ หากผู้ใช้เลือกคัดช่วงของข้อมูลยาแต่ละชนิด จะแสดงค่าประสิทธิภาพของโมเดลชนิดยานั้นในทุกเวอร์ชัน (Accuracy, Precision, Recall และ F1-Score)



រូបថត 3.33 หน้าแดមបែនប្រព័ន្ធដែលត្រូវនៅក្នុងកែវ



รูปที่ 3.34 หน้าแดชบอร์ดแสดงชุดข้อมูลที่ใช้ในการฝึกฝนโมเดล



รูปที่ 3.35 หน้าแดชบอร์ดแสดงประสิทธิภาพของโมเดลยาแต่ละชนิด

7) หน้าแสดงการกำหนดค่าของโมเดล เป็นหน้าที่แสดงค่าต่าง ๆ ที่ใช้ในการฝึกฝน โมเดล ได้แก่ อัตราส่วนของข้อมูลสำหรับฝึกฝนและทดสอบ (Train : Test), ปริมาณข้อมูลสำหรับทดสอบรายเคส (Test By Case), ฟีเจอร์ต่าง ๆ ที่ใช้, เทคนิคการจำแนกข้อมูลที่ใช้ (Classification Algorithm), พารามิเตอร์ต่าง ๆ และการใช้เทคนิคการสังเคราะห์ข้อมูลของโมเดลยาแต่ละชนิด

#	Antimicrobial Model (GN)	Parameters
1	Amikacin	n_estimators = 1000, gamma = 0.5, max_depth = 5, subsample = 1.0, colsample_bytree = 1.0, learning_rate = 0.3, random_state = 0
2	Amoxicillin/clavulanic acid	n_estimators = 1000, gamma = 2.0, max_depth = 6, subsample = 1.0, colsample_bytree = 0.7, learning_rate = 0.1, random_state = 0
3	Cefalexin	n_estimators = 1000, gamma = 2.0, max_depth = 6, subsample = 0.5, colsample_bytree = 0.7, learning_rate = 0.3, random_state = 0
4	Cefovecin	n_estimators = 1000, gamma = 0.5, max_depth = 5, subsample = 1.0, colsample_bytree = 0.5, learning_rate = 0.3, random_state = 0
5	Doxycycline	n_estimators = 1000, gamma = 2.0, max_depth = 5, subsample = 0.7, colsample_bytree = 0.5, learning_rate = 0.3, random_state = 0
6	Erofloxacin	n_estimators = 1000, gamma = 0.5, max_depth = 7, subsample = 1.0, colsample_bytree = 0.5, learning_rate = 0.1, random_state = 0
7	Gentamicin	n_estimators = 1000, gamma = 1.0, max_depth = 7, subsample = 0.5, colsample_bytree = 0.5, learning_rate = 0.3, random_state = 0
8	Imipenem	n_estimators = 1000, gamma = 2.0, max_depth = 6, subsample = 0.7, colsample_bytree = 0.7, learning_rate = 0.1, random_state = 0
9	Marbofloxacin	n_estimators = 1000, gamma = 2.0, max_depth = 6, subsample = 0.7, colsample_bytree = 1.0, learning_rate = 0.3, random_state = 0
10	Nitrofurantoin	n_estimators = 1000, gamma = 1.0, max_depth = 6, subsample = 0.7, colsample_bytree = 1.0, learning_rate = 0.3, random_state = 0
11	Trimethoprim/sulfamethoxazole	n_estimators = 1000, gamma = 0.5, max_depth = 6, subsample = 1.0, colsample_bytree = 1.0, learning_rate = 0.3, random_state = 0

#	Antimicrobial Model (GP)	SMOTE Algorithm
1	Amikacin	R-SMOTE
2	Amoxicillin/clavulanic acid	Borderline-SMOTE
3	Cefalexin	SMOTE
4	Cefovecin	SMOTE
5	Doxycycline	Borderline-SMOTE
6	Erofloxacin	R-SMOTE
7	Gentamicin	SVM-SMOTE
8	Imipenem	R-SMOTE
9	Marbofloxacin	SVM-SMOTE
10	Nitrofurantoin	R-SMOTE
11	Trimethoprim/sulfamethoxazole	R-SMOTE

รูปที่ 3.36 หน้าแสดงการกำหนดค่าของโมเดล

## บทที่ 4

### การทดลอง และผลการทดลอง

#### ตอนที่ 1 การฝึกฝนโมเดลแนวนำ้ยาต้านจุลชีพ

##### 4.1 การทดลองที่ 1.1 การเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ โดยใช้ชุดข้อมูลแบบที่เรียกว่า “Random” และชุดข้อมูลแบบที่เรียกว่า “Test”

###### 4.1.1 จุดประสงค์การทดลอง

- 1) เพื่อฝึกฝนโมเดลแนวนำ้ยาต้านจุลชีพโดยใช้ชุดข้อมูลจากการรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยาของห้องปฏิบัติการจุลชีววิทยา
- 2) เพื่อเปรียบเทียบประสิทธิภาพของการจำแนกข้อมูล 4 เทคนิค ได้แก่ Decision Tree, Random Forest, XGBoost และ SVM ด้วยค่า Accuracy, Precision, Recall และ F1-Score
- 3) เพื่อเลือกเทคนิคการจำแนกข้อมูลที่เหมาะสมกับชุดข้อมูลแบบที่เรียกว่า “Random” และชุดข้อมูลแบบที่เรียกว่า “Test”
- 4) เพื่อการจำแนกข้อมูล 2 เทคนิคที่ดีที่สุดสำหรับนำไปทำการทดลองที่ 1.2 (การแก้ไขปัญหาชุดข้อมูลไม่สมดุล)

###### 4.1.2 วิธีการทดลอง

- 1) นำชุดข้อมูลที่ทำความสะอาดแล้วมาทำการจัดเตรียมข้อมูล ดังนี้
  - จัดกลุ่มหมวดหมู่เด็ก ๆ ของฟีเจอร์ตั้งอย่างที่ส่งตรวจ (submitted\_sample) ที่มีจำนวนน้อยกว่า 10 รวมกันให้เป็นหมวดหมู่ใหม่ชื่อว่า “other”
  - นำชุดข้อมูลแบบที่เรียกว่า “Random” และแบบที่เรียกว่า “Test” มาทำการแบ่งข้อมูลสำหรับฝึกฝนโมเดลอย่างละ 11 โมเดล โดยชุดข้อมูลที่จะนำมาฝึกฝนแต่ละโมเดลนั้นจะแบ่งออกเป็น 2 ชุด (80:20) คือ ชุดข้อมูลสำหรับฝึกฝน 80% และชุดข้อมูลสำหรับทดสอบ 20% โดยแบ่งให้ชุดข้อมูลสำหรับฝึกฝน และชุดข้อมูลสำหรับทดสอบ มีอัตราส่วนของ 2 คลาส (True/False) เท่ากัน
    - นำชุดข้อมูลมาทำการเปลี่ยนให้อยู่ในรูปแบบตัวเลขด้วยวิธีการเข้ารหัสแบบวันออด

- 2) นำชุดข้อมูลสำหรับฝึกฝนไปทำการฝึกฝนโมเดลด้วย Decision Tree, Random Forest, XGBoost และ SVM โดยฟีเจอร์ที่ใช้ ได้แก่ ชนิดของสัตว์เลี้ยง (species), ศักดิ์แบบที่เรีย (bacteria\_genus), ตัวอย่างที่ส่งตรวจ (submitted\_sample) และผลตรวจความไวต่อยาต้านจุลชีพแต่ละชนิด (S/I/R) โดยใช้ 10-Fold Cross-Validation ในการทดสอบประสิทธิภาพ
- 3) วัดประสิทธิภาพของโมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score
  - 4) เปรียบเทียบประสิทธิภาพของแต่ละโมเดล
  - 5) สรุปผลการทดลอง

#### 4.1.3 ผลการทดลอง

ตาราง 4.1 การเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า Accuracy, Precision, Recall และ F1-Score ของชุดข้อมูลแบบที่เรียแกรมบวก (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพของโมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amikacin	DT	0.940	0.512	0.513	0.509
		RF	0.952	0.692	0.382	0.487
		XGB	0.953	0.669	0.487	<b>0.556</b>
		SVM	0.948	0.672	0.295	0.399
2	Amoxicillin/clavulanic acid	DT	0.919	0.904	0.919	0.911
		RF	0.935	0.923	0.934	0.928
		XGB	0.936	0.923	0.937	<b>0.930</b>
		SVM	0.937	0.938	0.922	<b>0.930</b>
3	Cefalexin	DT	0.926	0.866	0.884	0.874
		RF	0.941	0.886	0.913	<b>0.899</b>
		XGB	0.937	0.876	0.911	0.893
		SVM	0.940	0.894	0.900	0.897
4	Cefovecin	DT	0.927	0.863	0.906	0.884
		RF	0.940	0.866	0.952	<b>0.907</b>

		XGB	0.938	0.875	0.932	0.903
		SVM	0.937	0.845	0.978	0.906
5	Clindamycin	DT	0.881	0.758	0.809	0.781
		RF	0.883	0.768	0.809	0.784
		XGB	0.895	0.796	0.815	0.803
		SVM	0.897	0.802	0.815	<b>0.806</b>
6	Doxycycline	DT	0.950	0.561	0.582	<b>0.565</b>
		RF	0.951	0.663	0.326	0.427
		XGB	0.955	0.670	0.447	0.517
		SVM	0.954	0.698	0.369	0.468
7	Enrofloxacin	DT	0.965	0.509	0.480	0.490
		RF	0.968	0.699	0.319	0.418
		XGB	0.970	0.639	0.493	<b>0.540</b>
		SVM	0.968	0.702	0.329	0.432
8	Marbofloxacin	DT	0.953	0.589	0.546	0.546
		RF	0.963	0.746	0.446	0.542
		XGB	0.966	0.748	0.546	<b>0.611</b>
		SVM	0.964	0.732	0.485	0.563
9	Nitrofurantoin	DT	0.962	0.709	0.733	0.717
		RF	0.958	0.783	0.497	0.603
		XGB	0.972	0.798	0.770	0.780
		SVM	0.976	0.811	0.814	<b>0.812</b>
10	Trimethoprim/ sulfamethoxazole	DT	0.865	0.792	0.804	0.797
		RF	0.880	0.824	0.813	0.817
		XGB	0.883	0.832	0.810	0.821
		SVM	0.889	0.858	0.798	<b>0.826</b>
11	Vancomycin	DT	0.950	0.537	0.592	0.558
		RF	0.967	0.758	0.546	0.629
		XGB	0.963	0.699	0.546	0.608
		SVM	0.971	0.866	0.538	<b>0.650</b>

ตาราง 4.2 การเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า Accuracy, Precision, Recall และ F1-Score ของชุดข้อมูลแบนค์เรียแกรมlob (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพของโมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amikacin	DT	0.869	0.744	0.768	0.869
		RF	0.868	0.774	0.711	0.868
		XGB	0.883	0.781	0.777	0.883
		SVM	0.884	0.807	0.739	<b>0.884</b>
2	Amoxicillin/clavulanic acid	DT	0.938	0.882	0.871	0.876
		RF	0.948	0.883	0.914	0.897
		XGB	0.944	0.876	0.904	0.889
		SVM	0.948	0.862	0.947	<b>0.902</b>
3	Cefalexin	DT	0.926	0.723	0.781	0.748
		RF	0.940	0.791	0.793	0.790
		XGB	0.937	0.777	0.776	0.775
		SVM	0.942	0.808	0.781	<b>0.793</b>
4	Cefovecin	DT	0.904	0.830	0.828	0.828
		RF	0.917	0.821	0.905	0.860
		XGB	0.923	0.840	0.897	0.867
		SVM	0.925	0.833	0.916	<b>0.872</b>
5	Doxycycline	DT	0.948	0.440	0.447	0.438
		RF	0.955	0.562	0.147	0.226
		XGB	0.957	0.553	0.357	<b>0.430</b>
		SVM	0.960	0.625	0.286	0.383
6	Enrofloxacin	DT	0.928	0.772	0.766	0.768
		RF	0.945	0.845	0.800	0.818
		XGB	0.942	0.828	0.798	0.810
		SVM	0.948	0.862	0.790	<b>0.823</b>
7	Gentamicin	DT	0.901	0.487	0.533	0.507

		RF	0.925	0.657	0.454	0.534
		XGB	0.922	0.606	0.530	<b>0.565</b>
		SVM	0.930	0.751	0.404	0.514
8	Imipenem	DT	0.921	0.777	0.798	0.787
		RF	0.928	0.842	0.748	0.791
		XGB	0.933	0.829	0.800	<b>0.813</b>
		SVM	0.930	0.844	0.761	0.799
9	Marbofloxacin	DT	0.919	0.876	0.878	0.876
		RF	0.937	0.914	0.888	0.901
		XGB	0.932	0.898	0.892	0.895
		SVM	0.939	0.919	0.892	<b>0.905</b>
10	Nitrofurantoin	DT	0.968	0.747	0.803	0.769
		RF	0.969	0.837	0.669	0.739
		XGB	0.975	0.811	0.810	<b>0.808</b>
		SVM	0.971	0.845	0.699	0.759
11	Trimethoprim/ sulfamethoxazole	DT	0.866	0.806	0.822	0.813
		RF	0.887	0.813	0.886	0.847
		XGB	0.899	0.836	0.891	0.862
		SVM	0.900	0.822	0.919	<b>0.867</b>

ตาราง 4.3 การจัดอันดับประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า F1-Score ของชุดข้อมูลแบ่งที่เรียกวรมนวก (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิค			
		DT	RF	XGB	SVM
1	Amikacin	2	3	<b>1</b>	4
2	Amoxicillin/clavulanic acid	4	3	2	<b>1</b>
3	Cefalexin	4	<b>1</b>	3	2
4	Cefovecin	4	<b>1</b>	3	2
5	Clindamycin	4	3	2	<b>1</b>
6	Doxycycline	<b>1</b>	4	2	3
7	Enrofloxacin	2	4	<b>1</b>	3
8	Marbofloxacin	3	4	<b>1</b>	2
9	Nitrofurantoin	3	4	2	<b>1</b>
10	Trimethoprim/sulfamethoxazole	4	3	2	<b>1</b>
11	Vancomycin	4	2	3	<b>1</b>
	การจัดอันดับ	4	3	2	<b>1</b>

ตาราง 4.4 การจัดอันดับประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูลแบบต่าง ๆ ด้วยค่า F1-Score ของชุดข้อมูลแบ่งที่เรียกวรมฉบับ (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิค			
		DT	RF	XGB	SVM
1	Amikacin	3	4	<b>1</b>	2
2	Amoxicillin/clavulanic acid	4	2	3	<b>1</b>
3	Cefalexin	4	2	3	<b>1</b>
4	Cefovecin	4	3	2	<b>1</b>
5	Doxycycline	<b>1</b>	4	2	3
6	Enrofloxacin	4	2	3	<b>1</b>
7	Gentamicin	4	2	<b>1</b>	3
8	Imipenem	4	3	<b>1</b>	2

9	Marbofloxacin	4	2	3	<b>1</b>
10	Nitrofurantoin	2	4	<b>1</b>	3
11	Trimethoprim/sulfamethoxazole	4	3	2	<b>1</b>
	<b>การจัดอันดับ</b>	<b>3</b>	<b>4</b>	<b>2</b>	<b>1</b>

โดย DT หมายถึง ต้นไม้ตัดสินใจ (Decision Tree)

RF หมายถึง แรนดอมฟอร์ेस्ट (Random Forest)

XGB หมายถึง เอ็กซ์ट्रีมเกรเดียนบูสต์ (XGBoost)

SVM หมายถึง ชัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

#### 4.1.4 วิเคราะห์ผลการทดลอง

จากผลการทดลองของการฝึกฝน โมเดลกลุ่มแบคทีเรียแกรมบวกในตารางที่ 4.1 และตารางที่ 4.3 พบว่า เทคนิคการจำแนกข้อมูลที่ให้ค่า F1-Score สูงที่สุด คือ SVM, XGBoost, Random Forest และ Decision Tree ตามลำดับ โดยจะเห็นได้ว่าในหลาย ๆ โมเดลที่ใช้ SVM และ XGBoost จะให้ประสิทธิภาพที่ดีกว่า โมเดลที่ใช้ Random Forest และ Decision Tree ซึ่งที่เห็นได้ค่อนข้างชัดเจน เช่น โมเดลแนะนำยา Nitrofurantoin ที่ใช้ SVM มีค่า F1-Score อยู่ที่ 0.812 และ โมเดลที่ใช้ XGBoost มีค่า F1-Score อยู่ที่ 0.780 ในขณะที่เทคนิคอื่น ๆ นั้นมีค่า F1-Score อยู่ที่ 0.603 – 0.717

จากผลการทดลองของการฝึกฝน โมเดลกลุ่มแบคทีเรียแกรมลบในตารางที่ 4.2 และตารางที่ 4.4 พบว่า เทคนิคการจำแนกข้อมูลที่ให้ค่า F1-Score สูงที่สุด คือ SVM, XGBoost, Decision Tree และ Random Forest ตามลำดับ โดยจะเห็นได้ว่าในหลาย ๆ โมเดลที่ใช้ SVM และ XGBoost จะให้ประสิทธิภาพที่ดีกว่า โมเดลที่ใช้ Random Forest และ Decision Tree ซึ่งที่เห็นได้ค่อนข้างชัดเจน เช่น โมเดลแนะนำยา Trimethoprim/sulfamethoxazole ที่ใช้ SVM มีค่า F1-Score อยู่ที่ 0.867 และ โมเดลที่ใช้ XGBoost มีค่า F1-Score อยู่ที่ 0.862 ในขณะที่เทคนิคอื่น ๆ นั้นมีค่า F1-Score อยู่ที่ 0.813 – 0.817

#### 4.1.5 สรุปผลการทดลอง

จากการทดลองสรุปผลได้ว่า เทคนิคการจำแนกข้อมูลที่เหมาะสมกับชุดข้อมูลแบคทีเรีย แกรมบวก และชุดข้อมูลแบคทีเรียแกรมลบ 2 อันดับแรก โดยวัดผลจากค่า F1-Score คือ XGBoost

และ SVM โดยส่วนใหญ่ไม้modelที่ใช้ XGBoost และ SVM จะให้ค่า Accuracy, Precision, Recall และ F1-Score ที่ดีกว่าโมเดลที่ใช้ Random Forest และ Decision Tree เนื่องจาก XGBoost จะมีการปรับน้ำหนักของคลาสส่วนน้อยให้เพิ่มมากขึ้นในการสร้างต้นไม้ตัดสินใจในแต่ละครั้ง ทำให้เหมาะสมกับชุดข้อมูลที่ไม่สมดุล ส่วน SVM มีเครื่องเรียนที่ช่วยเพิ่มมิติของข้อมูล ทำให้จำแนกข้อมูลได้ดีกับข้อมูลที่มีฟีเจอร์มาก เนื่องด้วยในการทดลองนี้ใช้วิธีการเข้ารหัสแบบวันออดทำให้มีฟีเจอร์มากซึ่งจะเหมาะสมกับการใช้ SVM ดังนั้นผู้จัดทำจึงจะนำเทคนิค XGBoost และ SVM ไปใช้ในการทดลองที่ 1.2 ต่อไป

## 4.2 การทดลองที่ 1.2 การเปรียบเทียบทεcnikการสั่งเคราะห์ข้อมูล (Oversampling) ใน การแก้ปัญหาชุดข้อมูลไม่สมดุล โดยใช้ชุดข้อมูลแบบที่เรียแกรมบวกและแบบที่เรียแกรมลบ

### 4.2.1 จุดประสงค์การทดลอง

- 1) เพื่อศึกษาเทคนิคการสั่งเคราะห์ข้อมูลในการแก้ไขปัญหาชุดข้อมูลไม่สมดุล ได้แก่ SMOTE, Borderline-SMOTE, ADASYN, SVM-SMOTE และ R-SMOTE โดยใช้ชุดข้อมูลแบบที่เรียแกรมบวก และชุดข้อมูลแบบที่เรียแกรมลบ ซึ่งเป็นข้อมูลเชิงคุณภาพ (Qualitative Data)
- 2) เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการสั่งเคราะห์ข้อมูลด้วยค่า Precision, Recall และ F1-Score

- 3) เพื่อเลือกเทคนิคการจำแนกข้อมูลที่เหมาะสมกับชุดข้อมูล
- 4) เพื่อเลือกเทคนิคการสั่งเคราะห์ข้อมูลที่เหมาะสมกับชุดข้อมูลแบบที่เรียแกรมบวก และชุดข้อมูลแบบที่เรียแกรมลบ

### 4.2.2 วิธีการทดลอง

- 1) นำชุดข้อมูลที่ทำการจำแนกมาทำการเรียนรู้ข้อมูลใหม่อนการทดลองที่ 1.1
- 2) นำชุดข้อมูลสำหรับฝึกฝนมาทำการสั่งเคราะห์ข้อมูลของคลาสส่วนน้อยให้มีปริมาณเท่ากับคลาสส่วนมาก โดยใช้เทคนิค SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN และ R-SMOTE

3) นำชุดข้อมูลสำหรับฝึกฝนไปทำการฝึกฝนโมเดลด้วยเทคนิคการจำแนกข้อมูลที่ได้จากผลการทดลองที่ 1.1 ซึ่งก็คือ XGBoost และ SVM โดยใช้ 10-Fold Cross-Validation ในการทดสอบประสิทธิภาพ

- 4) วัดประสิทธิภาพของโมเดลด้วยค่า Precision, Recall และ F1-Score
- 5) เปรียบเทียบประสิทธิภาพของแต่ละโมเดล
- 6) สรุปผลการทดลอง

#### 4.2.3 ผลการทดลอง

ตาราง 4.5 อัตราความไม่สมดุลของชุดข้อมูลฝึกฝนของแบบที่เรียกวรมนวก

ลำดับ	ชุดข้อมูล	จำนวนข้อมูลของคลาสส่วนมาก	จำนวนข้อมูลของคลาสส่วนน้อย	อัตราความไม่สมดุล
1	Amikacin	2345	152	15.43
2	Amoxicillin/clavulanic acid	1368	1129	1.21
3	Cefalexin	1774	723	2.45
4	Cefovecin	1727	770	2.24
5	Clindamycin	1836	661	2.78
6	Doxycycline	2356	141	16.71
7	Enrofloxacin	2405	92	26.14
8	Marbofloxacin	2367	130	18.21
9	Nitrofurantoin	2336	161	14.51
10	Trimethoprim/sulfamethoxazole	1675	822	2.04
11	Vancomycin	2367	130	18.21

ตาราง 4.6 อัตราความไม่สมดุลของชุดข้อมูลฝึกฝนของแบบที่เรียกว่ากลุ่ม

ลำดับ	ชุดข้อมูล	จำนวนข้อมูลของคลาสส่วนมาก	จำนวนข้อมูลของคลาสส่วนน้อย	อัตราความไม่สมดุล
1	Amikacin	1844	660	2.79
2	Amoxicillin/clavulanic acid	1878	626	3.00
3	Cefalexin	2152	352	6.11
4	Cefovecin	1802	702	2.57
5	Doxycycline	2390	114	20.96
6	Enrofloxacin	2118	386	5.49
7	Gentamicin	2266	238	9.52
8	Imipenem	2044	460	4.44
9	Marbofloxacin	1690	814	2.08
10	Nitrofurantoin	2341	163	14.36
11	Trimethoprim/sulfamethoxazole	1618	886	1.83

โดย อัตราความไม่สมดุลของชุดข้อมูล (Imbalanced Ratio) =  $\frac{\text{จำนวนข้อมูลของคลาสส่วนมาก}}{\text{จำนวนข้อมูลของคลาสส่วนน้อย}}$

None หมายถึง ไม่มีการสังเคราะห์ข้อมูล

SM หมายถึง SMOTE

BL-S หมายถึง Borderline-SMOTE

SVM-S หมายถึง SVM-SMOTE

ADY หมายถึง ADASYN

R-S หมายถึง R-SMOTE

ตาราง 4.7 การเปรียบเทียบประสิทธิภาพด้วยค่า Precision ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบ่งที่เรียกวิกรมวาก (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า Precision						
			None	SM	BL-S	SVM-S	ADY	R-S	
1	Amikacin	XGB	<b>0.669</b>	0.524	0.527	0.526	0.502	0.587	
		SVM	<b>0.672</b>	0.507	0.479	0.491	0.561	0.563	
2	Amoxicillin/clavulanic acid	XGB	<b>0.923</b>	0.919	0.910	0.907	0.911	0.919	
		SVM	<b>0.938</b>	0.930	0.900	0.909	0.913	0.936	
3	Cefalexin	XGB	<b>0.876</b>	0.864	0.852	0.848	0.846	0.869	
		SVM	<b>0.894</b>	0.857	0.824	0.821	0.822	0.862	
4	Cefovecin	XGB	<b>0.875</b>	0.863	0.854	0.860	0.857	0.865	
		SVM	<b>0.845</b>	0.831	0.828	0.829	0.826	0.834	
5	Clindamycin	XGB	<b>0.796</b>	0.747	0.726	0.738	0.730	0.749	
		SVM	<b>0.802</b>	0.715	0.708	0.710	0.705	0.727	
6	Doxycycline	XGB	<b>0.670</b>	0.488	0.498	0.515	0.489	0.582	
		SVM	<b>0.698</b>	0.443	0.438	0.454	0.450	0.605	
7	Enrofloxacin	XGB	<b>0.639</b>	0.431	0.453	0.462	0.431	0.518	
		SVM	<b>0.702</b>	0.422	0.392	0.443	0.428	0.538	
8	Marbofloxacin	XGB	<b>0.748</b>	0.549	0.610	0.560	0.564	0.622	
		SVM	<b>0.732</b>	0.498	0.539	0.540	0.500	0.648	
9	Nitrofurantoin	XGB	<b>0.798</b>	0.694	0.707	0.711	0.705	0.745	
		SVM	<b>0.811</b>	0.659	0.680	0.681	0.666	0.745	
10	Trimethoprim/sulfamethoxazole	XGB	<b>0.832</b>	0.775	0.756	0.773	0.754	0.782	
		SVM	<b>0.858</b>	0.764	0.719	0.735	0.723	0.768	
11	Vancomycin	XGB	<b>0.699</b>	0.584	0.551	0.562	0.552	0.648	
		SVM	<b>0.866</b>	0.557	0.537	0.551	0.555	0.666	
จำนวนครั้งที่ชนะ (Winning Times)		XGB	<b>10</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	
		SVM	<b>10</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	

ตาราง 4.8 การเปรียบเทียบประสิทธิภาพด้วยค่า Recall ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบ่งที่เรียกวิเคราะห์ (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า Recall						
			None	SM	BL-S	SVM-S	ADY	R-S	
1	Amikacin	XGB	0.487	0.624	<b>0.625</b>	0.593	0.619	0.513	
		SVM	0.295	<b>0.652</b>	0.625	0.645	0.599	0.565	
2	Amoxicillin/clavulanic acid	XGB	0.937	0.947	<b>0.957</b>	0.957	0.956	0.942	
		SVM	0.922	0.934	<b>0.966</b>	0.964	0.956	0.926	
3	Cefalexin	XGB	0.911	0.938	0.945	<b>0.950</b>	0.943	0.923	
		SVM	0.900	0.952	0.964	0.967	<b>0.968</b>	0.947	
4	Cefovecin	XGB	0.932	0.953	0.955	<b>0.964</b>	0.957	0.945	
		SVM	0.978	0.981	0.982	0.983	<b>0.984</b>	0.979	
5	Clindamycin	XGB	0.815	0.890	0.909	<b>0.920</b>	0.899	0.856	
		SVM	0.815	0.946	0.950	<b>0.955</b>	0.949	0.932	
6	Doxycycline	XGB	0.447	0.631	0.617	0.610	<b>0.638</b>	0.553	
		SVM	0.369	0.688	0.695	<b>0.688</b>	0.666	0.596	
7	Enrofloxacin	XGB	0.493	0.568	<b>0.569</b>	0.558	0.568	0.528	
		SVM	0.329	0.580	0.581	<b>0.613</b>	0.548	0.547	
8	Marbofloxacin	XGB	0.546	0.638	<b>0.654</b>	0.585	<b>0.654</b>	0.569	
		SVM	0.485	<b>0.685</b>	0.677	0.677	0.623	0.623	
9	Nitrofurantoin	XGB	0.770	0.815	0.808	0.814	<b>0.827</b>	0.777	
		SVM	0.814	0.895	0.895	<b>0.901</b>	0.858	0.826	
10	Trimethoprim/sulfamethoxazole	XGB	0.810	0.867	0.899	<b>0.905</b>	0.895	0.870	
		SVM	0.798	0.897	<b>0.931</b>	0.928	0.931	0.893	
11	Vancomycin	XGB	0.546	<b>0.700</b>	0.692	<b>0.700</b>	<b>0.700</b>	0.592	
		SVM	0.538	0.715	0.731	<b>0.754</b>	0.615	0.662	
จำนวนครั้งที่ชนะ (Winning Times)		XGB	0	1	4	5	4	0	
		SVM	0	2	2	5	2	0	

ตาราง 4.9 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบ่งที่เรียกวิเคราะห์ (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า F1-Score						
			None	SM	BL-S	SVM-S	ADY	R-S	
1	Amikacin	XGB	0.556	0.566	<b>0.568</b>	0.555	0.552	0.541	
		SVM	0.399	0.568	0.541	0.555	<b>0.575</b>	0.559	
2	Amoxicillin/clavulanic acid	XGB	0.930	0.932	<b>0.933</b>	0.931	0.933	0.930	
		SVM	0.930	0.932	0.932	<b>0.936</b>	0.934	0.931	
3	Cefalexin	XGB	0.893	<b>0.899</b>	0.896	0.896	0.892	0.894	
		SVM	0.897	0.902	<b>0.888</b>	0.888	0.889	0.902	
4	Cefovecin	XGB	0.903	0.906	0.901	<b>0.908</b>	0.904	0.903	
		SVM	<b>0.906</b>	0.899	0.898	0.899	0.898	0.900	
5	Clindamycin	XGB	0.803	0.810	0.806	<b>0.818</b>	0.804	0.796	
		SVM	0.806	0.814	0.811	0.814	0.808	<b>0.816</b>	
6	Doxycycline	XGB	0.517	0.546	0.545	0.555	0.547	<b>0.556</b>	
		SVM	0.468	0.533	0.532	0.544	0.530	<b>0.592</b>	
7	Enrofloxacin	XGB	<b>0.540</b>	0.484	0.497	0.499	0.483	0.518	
		SVM	0.432	0.472	0.457	0.506	0.468	<b>0.529</b>	
8	Marbofloxacin	XGB	0.611	<b>0.579</b>	<b>0.619</b>	0.561	0.593	0.572	
		SVM	0.563	0.555	0.580	0.577	0.537	<b>0.618</b>	
9	Nitrofurantoin	XGB	<b>0.780</b>	0.746	0.750	0.756	0.758	0.758	
		SVM	<b>0.812</b>	0.757	0.770	0.773	0.747	0.781	
10	Trimethoprim/sulfamethoxazole	XGB	0.821	0.818	0.821	<b>0.833</b>	0.818	0.823	
		SVM	<b>0.826</b>	0.824	0.810	0.820	0.813	0.825	
11	Vancomycin	XGB	0.608	<b>0.632</b>	0.608	0.618	0.611	0.612	
		SVM	0.650	0.619	0.615	0.630	0.575	<b>0.658</b>	
ค่าเฉลี่ย		XGB	<b>0.724</b>	<b>0.720</b>	<b>0.722</b>	0.721	<b>0.718</b>	0.719	
จำนวนครั้งที่ชนะ (Winning Times)		SVM	0.699	0.716	0.712	<b>0.722</b>	0.707	<b>0.737</b>	
		XGB	2	2	3	3	0	1	
		SVM	3	0	1	1	1	5	

**ตาราง 4.10 การเปรียบเทียบประสิทธิภาพด้วยค่า Precision ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบบที่เรียกวิกรมลอง (10-Fold Cross-Validation)**

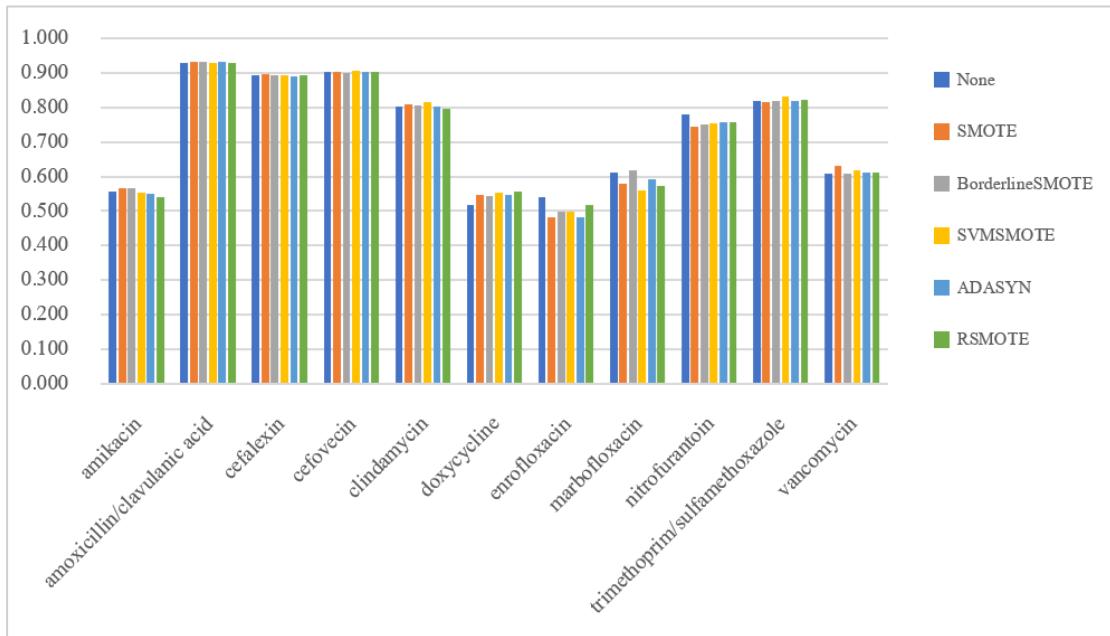
ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า Precision						
			None	SM	BL-S	SVM-S	ADY	R-S	
1	Amikacin	XGB	<b>0.781</b>	0.752	0.727	0.740	0.734	0.769	
		SVM	<b>0.807</b>	0.738	0.708	0.710	0.712	0.747	
2	Amoxicillin/clavulanic acid	XGB	<b>0.876</b>	0.850	0.853	0.855	0.853	0.868	
		SVM	<b>0.862</b>	0.836	0.834	0.830	0.831	0.842	
3	Cefalexin	XGB	<b>0.777</b>	0.733	0.725	0.718	0.717	0.758	
		SVM	<b>0.808</b>	0.693	0.689	0.684	0.684	0.732	
4	Cefovecin	XGB	<b>0.840</b>	0.825	0.820	0.818	0.818	0.837	
		SVM	<b>0.833</b>	0.796	0.807	0.801	0.798	0.805	
5	Doxycycline	XGB	<b>0.553</b>	0.432	0.449	0.440	0.417	0.481	
		SVM	<b>0.625</b>	0.491	0.411	0.408	0.485	0.439	
6	Enrofloxacin	XGB	<b>0.828</b>	0.784	0.772	0.777	0.766	0.823	
		SVM	<b>0.862</b>	0.762	0.755	0.753	0.771	0.824	
7	Gentamicin	XGB	<b>0.606</b>	0.501	0.487	0.501	0.498	0.557	
		SVM	<b>0.751</b>	0.460	0.468	0.474	0.453	0.528	
8	Imipenem	XGB	<b>0.829</b>	0.778	0.766	0.781	0.736	0.787	
		SVM	<b>0.844</b>	0.733	0.727	0.720	0.708	0.777	
9	Marbofloxacin	XGB	<b>0.898</b>	0.892	0.871	0.873	0.870	0.893	
		SVM	<b>0.919</b>	0.895	0.854	0.856	0.856	0.891	
10	Nitrofurantoin	XGB	<b>0.811</b>	0.746	0.730	0.743	0.744	0.776	
		SVM	<b>0.845</b>	0.687	0.689	0.700	0.709	0.765	
11	Trimethoprim/sulfamethoxazole	XGB	<b>0.836</b>	0.818	0.812	0.818	0.812	0.822	
		SVM	<b>0.822</b>	0.803	0.797	0.802	0.796	0.808	
จำนวนครั้งที่ชนะ (Winning Times)		XGB	<b>11</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	
		SVM	<b>11</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	

ตาราง 4.11 การเปรียบเทียบประสิทธิภาพด้วยค่า Recall ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบ่งที่เรียกวิกรมลับ (10-Fold Cross-Validation)

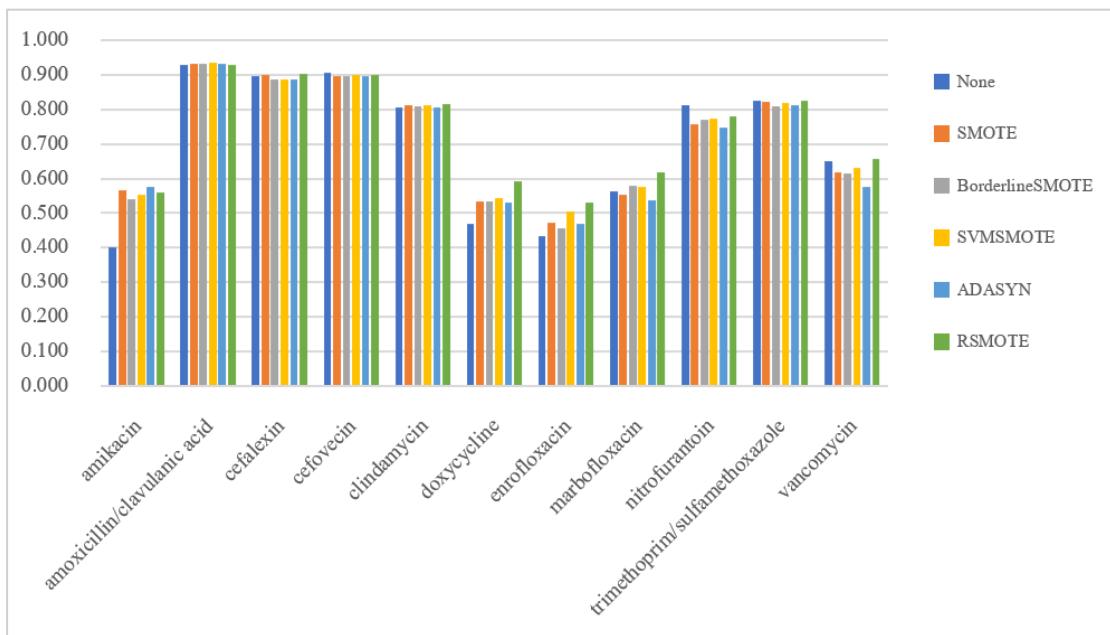
ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า Recall						
			None	SM	BL-S	SVM-S	ADY	R-S	
1	Amikacin	XGB	0.777	0.844	0.855	<b>0.858</b>	0.855	0.827	
		SVM	0.739	0.855	0.864	0.865	<b>0.883</b>	0.800	
2	Amoxicillin/clavulanic acid	XGB	0.904	0.933	0.936	<b>0.949</b>	0.941	0.912	
		SVM	0.947	<b>0.978</b>	0.971	0.976	0.976	0.968	
3	Cefalexin	XGB	0.776	0.844	0.838	<b>0.858</b>	0.852	0.827	
		SVM	0.781	0.895	0.878	0.897	<b>0.903</b>	0.867	
4	Cefovecin	XGB	0.897	0.912	0.923	0.929	<b>0.930</b>	0.905	
		SVM	0.916	<b>0.964</b>	0.953	0.957	0.957	0.954	
5	Doxycycline	XGB	0.357	0.527	0.527	0.508	<b>0.553</b>	0.456	
		SVM	0.286	0.287	0.455	<b>0.473</b>	0.261	0.436	
6	Enrofloxacin	XGB	0.798	<b>0.868</b>	0.865	0.868	0.855	0.832	
		SVM	0.790	<b>0.886</b>	0.870	0.881	0.870	0.837	
7	Gentamicin	XGB	0.530	0.694	0.698	0.686	<b>0.706</b>	0.639	
		SVM	0.404	0.702	<b>0.710</b>	0.702	0.685	0.622	
8	Imipenem	XGB	0.800	0.826	0.839	<b>0.848</b>	<b>0.848</b>	0.813	
		SVM	0.761	0.872	0.885	0.880	<b>0.887</b>	0.824	
9	Marbofloxacin	XGB	0.892	0.904	0.921	<b>0.918</b>	0.912	0.905	
		SVM	0.892	0.916	0.925	0.929	<b>0.932</b>	0.905	
10	Nitrofurantoin	XGB	0.810	0.853	0.841	<b>0.860</b>	<b>0.860</b>	0.799	
		SVM	0.699	<b>0.859</b>	0.847	<b>0.859</b>	0.810	0.786	
11	Trimethoprim/sulfamethoxazole	XGB	0.891	0.914	0.923	<b>0.928</b>	0.924	0.904	
		SVM	0.919	0.950	0.958	0.956	<b>0.958</b>	0.944	
จำนวนครั้งที่ชนะ (Winning Times)			<b>XGB</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>7</b>	<b>5</b>	
			<b>SVM</b>	<b>0</b>	<b>4</b>	<b>1</b>	<b>2</b>	<b>5</b>	
								<b>0</b>	

ตาราง 4.12 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของเทคนิคการสังเคราะห์ข้อมูลของชุดข้อมูลแบบที่เรียกวิกรมลบ (10-Fold Cross-Validation)

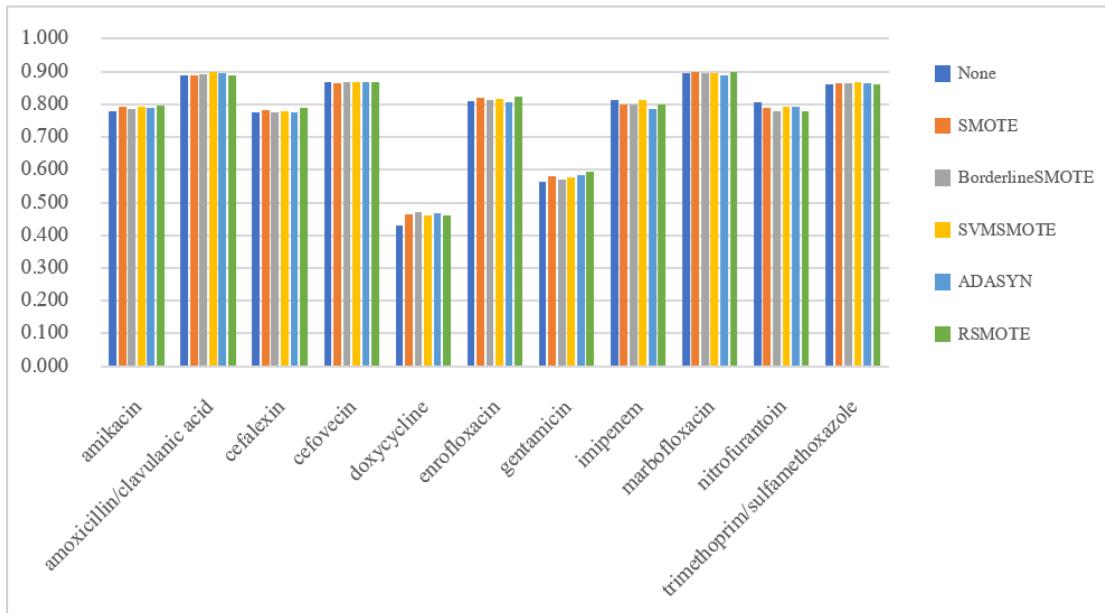
ลำดับ	โมเดล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า F1-Score						
			None	SM	BL-S	SVM-S	ADY	R-S	
1	Amikacin	XGB	0.778	0.794	0.784	0.793	0.789	<b>0.796</b>	
		SVM	0.769	<b>0.790</b>	0.777	0.779	0.787	0.771	
2	Amoxicillin/clavulanic acid	XGB	0.889	0.889	0.892	<b>0.899</b>	0.894	0.889	
		SVM	<b>0.902</b>	0.901	0.897	0.897	0.897	0.900	
3	Cefalexin	XGB	0.775	0.782	0.776	0.780	0.777	<b>0.790</b>	
		SVM	<b>0.793</b>	0.779	0.769	0.774	0.776	0.792	
4	Cefovecin	XGB	0.867	0.866	0.868	0.869	<b>0.870</b>	0.869	
		SVM	0.872	0.871	0.873	0.872	0.870	<b>0.873</b>	
5	Doxycycline	XGB	0.430	0.463	<b>0.470</b>	0.460	0.468	0.460	
		SVM	0.383	0.336	0.420	0.424	0.317	<b>0.427</b>	
6	Enrofloxacin	XGB	0.810	0.821	0.813	0.817	0.806	<b>0.825</b>	
		SVM	0.823	0.816	0.804	0.807	0.814	<b>0.828</b>	
7	Gentamicin	XGB	0.565	0.580	0.572	0.578	0.583	<b>0.593</b>	
		SVM	0.514	0.553	0.562	0.563	0.543	<b>0.567</b>	
8	Imipenem	XGB	<b>0.813</b>	0.800	0.800	0.812	0.787	0.799	
		SVM	<b>0.799</b>	0.795	0.797	0.791	0.787	0.799	
9	Marbofloxacin	XGB	0.895	0.898	0.895	0.894	0.890	<b>0.899</b>	
		SVM	0.905	<b>0.905</b>	0.888	0.891	0.892	0.898	
10	Nitrofurantoin	XGB	<b>0.808</b>	0.789	0.778	0.793	0.792	0.780	
		SVM	0.759	0.758	0.754	0.766	0.747	<b>0.771</b>	
11	Trimethoprim/sulfamethoxazole	XGB	0.862	0.863	0.864	<b>0.869</b>	0.864	0.861	
		SVM	0.867	0.870	0.870	<b>0.872</b>	0.869	0.870	
ค่าเฉลี่ย			<b>XGB</b>	<b>0.772</b>	<b>0.777</b>	<b>0.774</b>	<b>0.779</b>	<b>0.774</b>	
จำนวนครั้งที่ชนะ (Winning Times)			<b>SVM</b>	0.762	0.761	0.765	0.767	0.755	
			<b>XGB</b>	<b>2</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>1</b>	
			<b>SVM</b>	<b>3</b>	<b>2</b>	<b>0</b>	<b>1</b>	<b>0</b>	
								<b>5</b>	



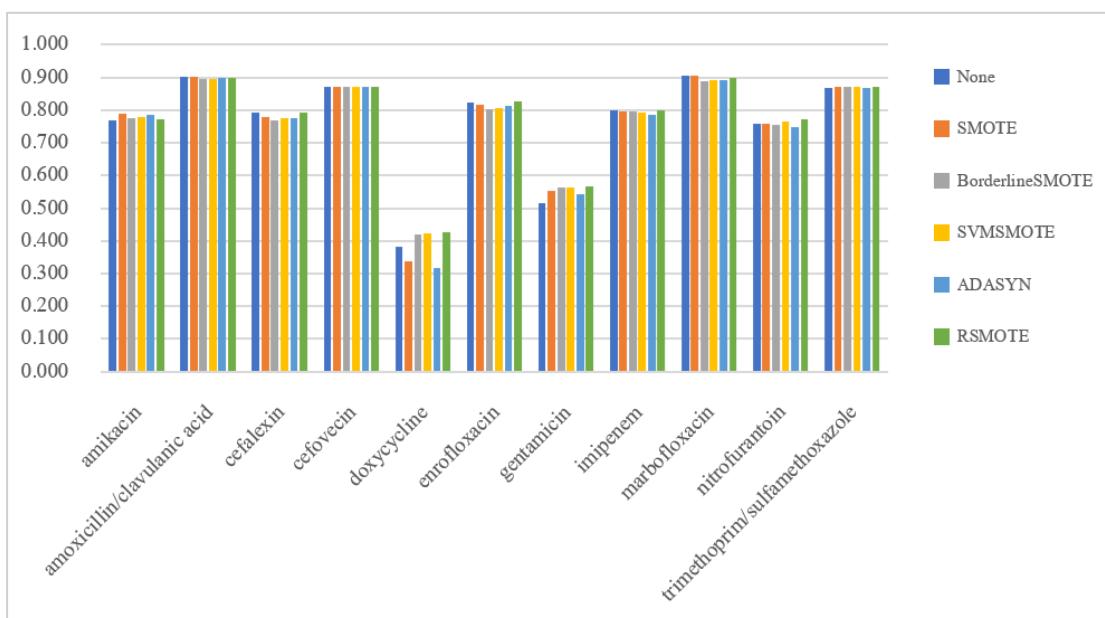
รูปที่ 4.1 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้  
เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบคทีเรียแกรมบวก



รูปที่ 4.2 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้  
เทคนิคการจำแนกข้อมูลด้วย SVM ของชุดข้อมูลแบคทีเรียแกรมบวก



รูปที่ 4.3 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้  
เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบคทีเรียแกรมลบ



รูปที่ 4.4 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้  
เทคนิคการจำแนกข้อมูลด้วย SVM ของชุดข้อมูลแบคทีเรียแกรมลบ

#### 4.2.4 วิเคราะห์ผลการทดลอง

จากการการเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของชุดข้อมูลแบคที่เรียกว่า น้ำกและชุดข้อมูลแบคที่เรียกว่า แกรม พบว่า เทคนิคการสังเคราะห์ข้อมูลที่ให้ค่า F1-Score สูงที่สุด โดยพิจารณาจากจำนวนครั้งที่ชนะ (Winning Times) ของการใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost และ SVM รวมกัน คือ R-SMOTE (ชนะ 16 ครั้ง), SVM-SMOTE (ชนะ 7 ครั้ง), Borderline-SMOTE (ชนะ 5 ครั้ง) SMOTE (ชนะ 4 ครั้ง), และ ADASYN (ชนะ 2 ครั้ง) ตามลำดับ เนื่องจาก R-SMOTE จะทำการสังเคราะห์ข้อมูลมากตรงส่วนที่มีความปลอดกัน และส่วนที่มีความวุ่นวายมากจะสังเคราะห์ข้อมูลน้อย โดยพิจารณาจากค่าความหนาแน่นสัมพันธ์ของข้อมูลคลาส ส่วนน้อย ดังนั้น การสังเคราะห์ข้อมูลด้วย R-SMOTE จะให้ค่า Precision สูงกว่าการสังเคราะห์ข้อมูลด้วยเทคนิคอื่น ๆ แม้ว่าค่า Recall จะต่ำกว่า แต่ก็ยังทำให้ค่า F1-Score สูงกว่าเมื่อนำไปใช้กับในหลาย ๆ ชุดข้อมูล

#### 4.2.5 สรุปผลการทดลอง

จากการทดลองสรุปผลได้ว่า การสังเคราะห์ข้อมูลของคลาสส่วนน้อยให้มีความสมดุล กับคลาสส่วนมากสามารถเพิ่มประสิทธิภาพของโมเดลได้เพียงเล็กน้อย โดยส่วนใหญ่การสังเคราะห์ข้อมูลจะลดค่า Precision แต่เพิ่มค่า Recall ได้มาก ทำให้ค่า F1-Score สูงมากขึ้น ซึ่งแต่ละเทคนิคจะเหมาะสมกับชุดข้อมูลที่แตกต่างกันไป การใช้การสังเคราะห์ข้อมูลกับบางชุดข้อมูลทำให้ประสิทธิภาพเพิ่มขึ้น แต่บางชุดข้อมูลทำให้ประสิทธิภาพต่ำลง

จากการใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost และ SVM ให้ค่า F1-Score ที่มีความใกล้เคียงกัน แต่ XGBoost จะให้ค่าเฉลี่ย F1-Score ของแต่ละเทคนิคการสังเคราะห์ข้อมูลที่มากกว่า และ SVM ดังนั้นจากการทดลองนี้จึงเลือก XGBoost เป็นเทคนิคที่มีเหมาะสมกับชุดข้อมูล แบคที่เรียกว่า น้ำก และแบคที่เรียกว่า แกรม โดยเทคนิคการสังเคราะห์ข้อมูลที่ให้ค่า F1-Score สูงที่สุดกับแต่ละชุดข้อมูล โดยใช้ XGBoost และดูผลตามตาราง ดังนี้

ตาราง 4.13 ประสิทธิภาพของโมเดลที่ใช้เทคนิคการสังเคราะห์ข้อมูลที่เหมาะสม โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบคทีเรียแกรมบวก (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิคการสังเคราะห์ข้อมูล	Precision	Recall	F1-Score
1	Amikacin	BL-S	0.527	0.625	0.568
2	Amoxicillin/clavulanic acid	BL-S	0.910	0.957	0.933
3	Cefalexin	SM	0.864	0.938	0.899
4	Cefovecin	SVM-S	0.860	0.964	0.908
5	Clindamycin	SVM-S	0.738	0.920	0.818
6	Doxycycline	R-S	0.582	0.553	0.556
7	Enrofloxacin	R-S	0.518	0.528	0.518
8	Marbofloxacin	BL-S	0.610	0.654	0.619
9	Nitrofurantoin	R-S	0.745	0.777	0.758
10	Trimethoprim/sulfamethoxazole	SVM-S	0.773	0.905	0.833
11	Vancomycin	SM	0.584	0.700	0.632

ตาราง 4.14 ประสิทธิภาพของโมเดลที่ใช้เทคนิคการสังเคราะห์ข้อมูลที่เหมาะสม โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost ของชุดข้อมูลแบคทีเรียแกรมลบ (10-Fold Cross-Validation)

ลำดับ	โมเดล	เทคนิคการสังเคราะห์ข้อมูล	Precision	Recall	F1-Score
1	Amikacin	R-S	0.769	0.827	0.796
2	Amoxicillin/clavulanic acid	SVM-S	0.855	0.949	0.899
3	Cefalexin	R-S	0.758	0.827	0.790
4	Cefovecin	ADY	0.818	0.930	0.870
5	Doxycycline	BL-S	0.449	0.527	0.470
6	Enrofloxacin	R-S	0.823	0.832	0.825
7	Gentamicin	R-S	0.557	0.639	0.593
8	Imipenem	SM	0.778	0.826	0.800

9	Marbofloxacin	R-S	0.893	0.905	0.899
10	Nitrofurantoin	SVM-S	0.743	0.860	0.793
11	Trimethoprim/sulfamethoxazole	SVM-S	0.818	0.928	0.869

### 4.3 การทดลองที่ 1.3 การเปรียบเทียบทεκνικการสังเคราะห์ข้อมูล (Oversampling) ใน การแก้ปัญหาชุดข้อมูลไม่สมดุล โดยใช้ชุดข้อมูลทั่วไปจาก UCI Machine Learning Repository

#### 4.3.1 จุดประสงค์การทดลอง

- 1) เพื่อศึกษาเทคนิคการสังเคราะห์ข้อมูลต่าง ๆ ในการแก้ไขปัญหาชุดข้อมูลไม่สมดุล ได้แก่ SMOTE, Borderline-SMOTE, ADASYN, SVM-SMOTE และ R-SMOTE โดยใช้ชุดข้อมูลทั่วไปที่มีไฟล์เป็นข้อมูลเชิงปริมาณ (Quantitative Data)
- 2) เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า Precision, Recall และ F1-Score
- 3) เพื่อเปรียบเทียบผลลัพธ์ที่ได้กับการทดลองที่ 1.2

#### 4.3.2 วิธีการทดลอง

- 1) ค้นหาชุดข้อมูลทั่วไปจาก UCI Machine Learning Repository
- 2) นำแต่ละชุดข้อมูลมาทำการสังเคราะห์เพิ่มปริมาณข้อมูลของคลาสส่วนน้อยให้มีปริมาณเท่ากับคลาสส่วนมาก โดยใช้เทคนิค SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN และ R-SMOTE
- 3) นำแต่ละชุดข้อมูลไปทำการฝึกฝนโมเดลด้วยเทคนิคการจำแนกข้อมูลเหมือนกับการทดลองที่ 1.2 ซึ่งก็คือ XGBoost และ SVM โดยใช้ 10-Fold Cross-Validation ในการทดสอบประสิทธิภาพ
- 4) วัดประสิทธิภาพของโมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score
- 5) เปรียบเทียบประสิทธิภาพของแต่ละโมเดล
- 6) เพื่อเปรียบเทียบผลลัพธ์ที่ได้กับการทดลองที่ 1.2
- 7) สรุปผลการทดลอง

### 4.3.3 ผลการทดลอง

ตาราง 4.15 ชุดข้อมูลที่ใช้ในการทดลองที่ 1.3

ชุดข้อมูล	จำนวน ฟีเจอร์	ประเภท ข้อมูล	จำนวนข้อมูลของ คลาสส่วนมาก	จำนวนข้อมูลของ คลาสส่วนน้อย	อัตราความ ไม่สมดุล
pima	8	float, int	500	268	1.86
wisconsin	9	int	444	239	1.86
glass4	9	float	201	13	2.48
haberman	3	float	225	81	2.78
vehicle0	18	int	647	199	3.25
ecoli2	7	float	284	52	5.46
page-blocks0	10	float, int	4713	559	8.43
yeast-1_vs_7	7	float	429	30	14.30
ecoli4	7	float	316	20	15.80
yeast6	8	float	1449	35	41.40

ตาราง 4.16 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า Precision โดยใช้ชุดข้อมูลทั่วไป (10-Fold Cross-Validation)

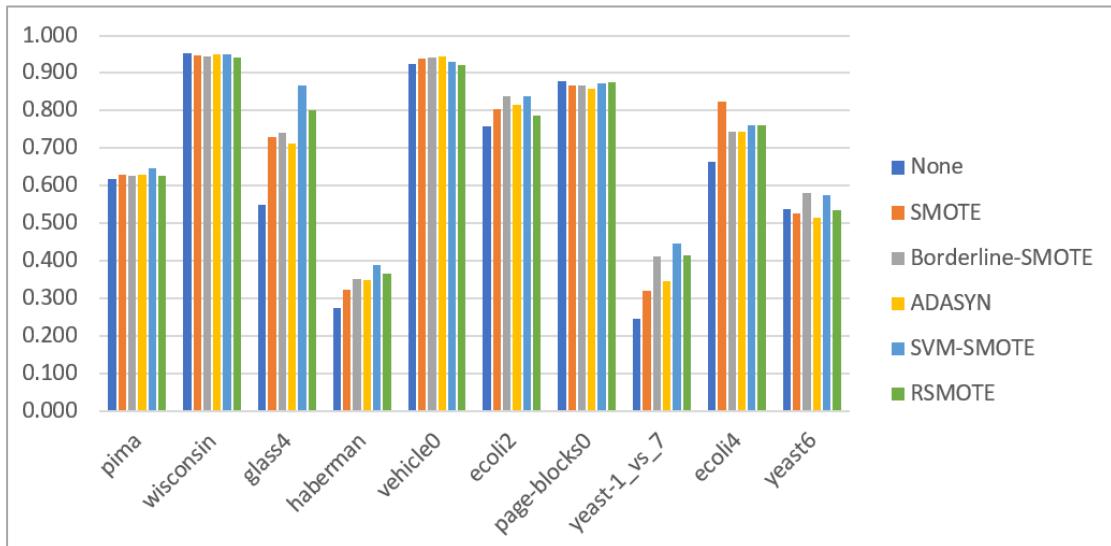
ชุดข้อมูล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า Precision					
		None	SM	BL-S	SVM-S	ADY	R-S
pima	XGB	<b>0.628</b>	0.601	0.601	0.614	0.598	0.604
	SVM	<b>0.721</b>	0.594	0.562	0.587	0.585	0.599
wisconsin	XGB	<b>0.956</b>	0.933	0.922	0.933	0.930	0.936
	SVM	<b>0.948</b>	0.928	0.921	0.917	0.906	0.940
glass4	XGB	0.583	0.750	0.725	<b>0.850</b>	0.717	0.800
	SVM	0.200	0.262	0.263	0.369	0.103	<b>0.575</b>
haberman	XGB	0.286	0.296	0.324	<b>0.362</b>	0.311	0.354
	SVM	0.040	<b>0.515</b>	0.488	0.514	0.463	0.513
vehicle0	XGB	<b>0.931</b>	0.924	0.929	0.903	0.920	0.920
	SVM	0.680	0.718	0.684	0.698	0.666	<b>0.718</b>
ecoli2	XGB	<b>0.876</b>	0.826	0.875	0.867	0.813	0.863
	SVM	<b>0.921</b>	0.848	0.870	0.851	0.715	0.856
page-blocks0	XGB	<b>0.884</b>	0.826	0.819	0.831	0.803	0.860
	SVM	<b>0.822</b>	0.234	0.287	0.240	0.251	0.255
yeast-1_vs_7	XGB	0.367	0.363	0.453	<b>0.573</b>	0.350	0.553
	SVM	<b>0.600</b>	0.227	0.277	0.300	0.217	0.363
ecoli4	XGB	0.717	<b>0.850</b>	0.750	0.750	0.750	0.800
	SVM	<b>1.000</b>	0.833	0.767	0.817	0.767	0.817
yeast6	XGB	<b>0.687</b>	0.508	0.585	0.580	0.453	0.563
	SVM	<b>0.617</b>	0.229	0.311	0.313	0.155	0.347
จำนวนครั้งที่ ชนะ	XGB	<b>6</b>	1	0	3	0	0
	SVM	<b>7</b>	1	0	0	0	2

ตาราง 4.17 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า Recall โดยใช้ชุดข้อมูลทั่วไป (10-Fold Cross-Validation)

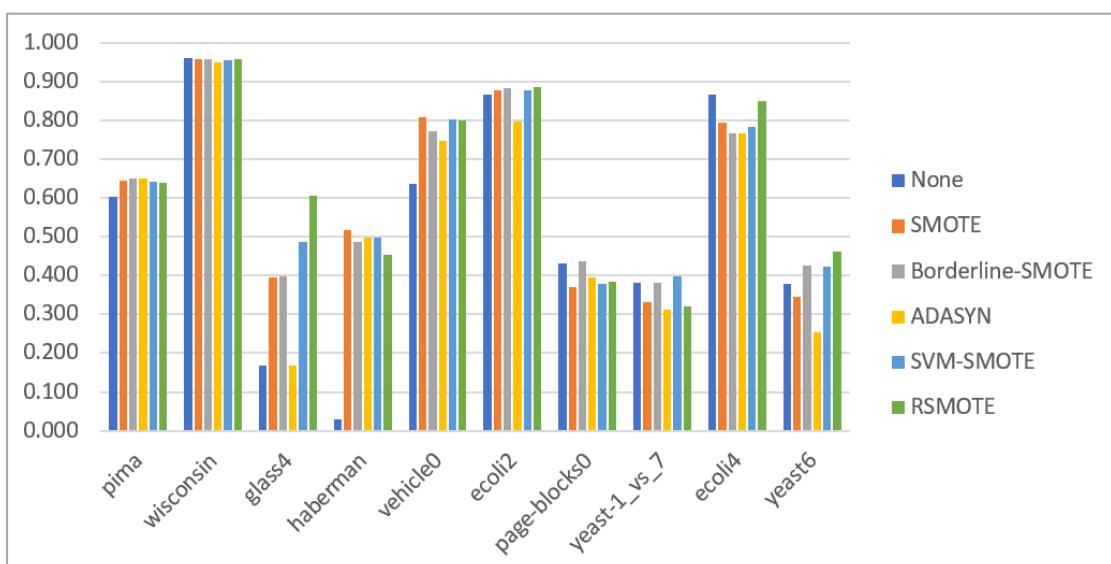
ชุดข้อมูล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า Recall					
		None	SM	BL-S	SVM-S	ADY	R-S
pima	XGB	0.615	0.664	0.656	<b>0.686</b>	0.668	0.652
	SVM	0.523	0.709	<b>0.772</b>	0.713	0.739	0.690
wisconsin	XGB	0.954	0.963	0.971	0.971	<b>0.975</b>	0.950
	SVM	0.971	0.988	<b>0.996</b>	<b>0.996</b>	<b>0.996</b>	0.975
glass4	XGB	0.600	0.800	0.850	<b>0.950</b>	0.800	0.850
	SVM	0.150	<b>0.950</b>	<b>0.950</b>	<b>0.950</b>	0.650	0.750
haberman	XGB	0.274	0.361	0.399	<b>0.435</b>	0.422	0.399
	SVM	0.025	0.532	0.494	0.496	<b>0.557</b>	0.424
vehicle0	XGB	0.920	0.955	0.955	0.960	<b>0.970</b>	0.925
	SVM	0.608	0.929	0.889	<b>0.949</b>	0.858	0.909
ecoli2	XGB	0.723	0.840	0.833	0.853	<b>0.857</b>	0.780
	SVM	0.860	<b>0.930</b>	0.910	<b>0.930</b>	<b>0.930</b>	<b>0.930</b>
page-blocks0	XGB	0.875	0.916	<b>0.925</b>	0.925	<b>0.925</b>	0.894
	SVM	0.297	0.886	0.912	0.896	<b>0.916</b>	0.793
yeast-1_vs_7	XGB	0.200	0.333	0.400	<b>0.433</b>	0.367	0.367
	SVM	0.300	<b>0.667</b>	0.633	0.633	0.633	0.333
ecoli4	XGB	0.650	<b>0.850</b>	0.800	<b>0.850</b>	0.800	0.800
	SVM	0.800	0.800	0.800	0.800	0.800	<b>0.900</b>
yeast6	XGB	0.492	0.600	0.600	0.608	<b>0.633</b>	0.542
	SVM	0.292	<b>0.808</b>	0.708	0.708	0.775	0.742
จำนวนครั้งที่ ชนะ	XGB	<b>0</b>	<b>1</b>	<b>1</b>	<b>5</b>	<b>5</b>	<b>0</b>
	SVM	<b>0</b>	<b>4</b>	<b>3</b>	<b>4</b>	<b>4</b>	<b>2</b>

ตาราง 4.18 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้ชุดข้อมูลทั่วไป (10-Fold Cross-Validation)

ชุดข้อมูล	เทคนิค	ผลการวัดประสิทธิภาพด้วยค่า F1-Score					
		None	SM	BL-S	SVM-S	ADY	R-S
pima	XGB	0.617	0.628	0.627	<b>0.646</b>	0.629	0.625
	SVM	0.604	0.644	0.648	0.642	<b>0.650</b>	0.639
wisconsin	XGB	<b>0.954</b>	0.947	0.945	0.951	0.951	0.942
	SVM	<b>0.959</b>	0.956	0.956	0.954	0.948	0.957
glass4	XGB	0.550	0.730	0.740	<b>0.867</b>	0.713	0.800
	SVM	0.167	0.396	0.397	0.488	0.167	<b>0.607</b>
haberman	XGB	0.276	0.323	0.353	<b>0.387</b>	0.350	0.365
	SVM	0.031	<b>0.517</b>	0.485	0.498	0.496	0.452
vehicle0	XGB	0.923	0.938	0.940	0.929	<b>0.944</b>	0.921
	SVM	0.636	<b>0.809</b>	0.770	0.802	0.748	0.800
ecoli2	XGB	0.758	0.805	0.838	<b>0.839</b>	0.816	0.788
	SVM	0.866	0.875	0.883	0.877	0.798	<b>0.885</b>
page-blocks0	XGB	<b>0.878</b>	0.867	0.867	0.874	0.859	0.876
	SVM	0.431	0.370	<b>0.437</b>	0.378	0.394	0.385
yeast-1_vs_7	XGB	0.247	0.320	0.413	<b>0.446</b>	0.345	0.415
	SVM	0.380	0.332	0.380	<b>0.397</b>	0.312	0.319
ecoli4	XGB	0.663	<b>0.823</b>	0.743	0.760	0.743	0.760
	SVM	<b>0.867</b>	0.793	0.767	0.783	0.767	0.850
yeast6	XGB	0.537	0.527	<b>0.582</b>	0.575	0.515	0.534
	SVM	0.377	0.345	0.426	0.422	0.254	<b>0.461</b>
จำนวนครั้งที่ ชนะ	XGB	<b>2</b>	<b>1</b>	<b>1</b>	<b>5</b>	<b>1</b>	<b>0</b>
	SVM	<b>2</b>	<b>2</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>3</b>



รูปที่ 4.5 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้  
เทคนิคการจำแนกข้อมูลด้วย XGBoost



รูปที่ 4.6 การเปรียบเทียบประสิทธิภาพของเทคนิคการสังเคราะห์ข้อมูลด้วยค่า F1-Score โดยใช้  
เทคนิคการจำแนกข้อมูลด้วย SVM

#### 4.3.4 วิเคราะห์ผลการทดลอง

จากการการเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score พบว่า เทคนิคการสังเคราะห์ข้อมูลที่ให้ค่า F1-Score สูงที่สุด โดยพิจารณาตามจำนวนครั้งที่ชนะ (Winning Times) ของ XGBoost และ SVM รวมกัน คือ SVM-SMOTE (ชนะ 6 ครั้ง), R-SMOTE (ชนะ 3 ครั้ง), SMOTE (ชนะ 3 ครั้ง), Borderline-SMOTE (ชนะ 2 ครั้ง) และ ADASYN (ชนะ 2 ครั้ง) ตามลำดับ ซึ่งลำดับจะแตกต่างจากการทดลองที่ 1.2 จึงวิเคราะห์ผลได้ว่า การใช้เทคนิคการสังเคราะห์ข้อมูลจะให้ผลลัพธ์ที่แตกต่างกันตามลักษณะของชุดข้อมูล โดย SMOTE เป็นอัลกอริทึมพื้นฐาน ซึ่งจะทำการสังเคราะห์ข้อมูลในทุก ๆ จุดของคลาสส่วนน้อย ส่วน SVM-SMOTE จะเน้นการสังเคราะห์ข้อมูลเพื่อบรยายพื้นที่ของคลาสส่วนน้อยให้ไปยังพื้นที่ของคลาสส่วนมาก ส่วน R-SMOTE จะเน้นการสังเคราะห์ข้อมูลในส่วนที่มีความปลดดกย จึงหมายความว่าชุดข้อมูลที่มี Noise มาก ส่วน Borderline-SMOTE จะเน้นการสังเคราะห์ข้อมูลตรงช่วงขอบของคลาสส่วนน้อย และ ADASYN จะเน้นการสังเคราะห์ข้อมูลตรงส่วนที่คลาสส่วนน้อยมีความหนาแน่นน้อย

จากการการเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score พบว่า โดยส่วนมากข้อมูลที่ทำการสังเคราะห์เพิ่มปริมาณข้อมูลจะมีประสิทธิภาพ F1-Score ที่สูงกว่าข้อมูลที่ไม่ได้ทำการสังเคราะห์เพิ่มปริมาณข้อมูล ซึ่งแสดงให้เห็นว่าการสังเคราะห์เพิ่มปริมาณข้อมูลสามารถช่วยเพิ่มประสิทธิภาพให้แก่การสร้างโมเดลด้วยชุดข้อมูลที่ไม่สมดุลได้

#### 4.3.5 สรุปผลการทดลอง

จากการทดลองสรุปผลได้ว่า มีความสอดคล้องกันกับการทดลองที่ 1.2 คือ การสังเคราะห์เพิ่มปริมาณข้อมูลสามารถเพิ่มประสิทธิภาพของโมเดลได้ โดยส่วนใหญ่จะลดค่า Precision แต่เพิ่มค่า Recall ได้มาก ทำให้ค่า F1-Score สูงมากขึ้น เนื่องจากการสังเคราะห์ข้อมูลช่วยเพิ่มข้อมูลของคลาสส่วนน้อยให้มากขึ้น ทำให้การทำนายคลาสส่วนน้อยมีประสิทธิภาพมากขึ้น ซึ่งในบางชุดข้อมูลจะสามารถเพิ่มประสิทธิภาพได้สูงมากโดยวัดจากค่า F1-Score เช่น ชุดข้อมูล glass4 (XGB, SVM), haberman (SVM), vehicle0 (SVM) และ yeast-1\_vs\_7 (XGB) และบางชุดข้อมูลมีประสิทธิภาพลดลง เช่น wisconsin (XGB) และ ecoli4 (SVM)

## 4.4 การทดลองที่ 1.4 การปรับพารามิเตอร์ของโมเดลกลุ่มแบบที่เรียกว่าแบบ และ โมเดลกลุ่มแบบที่เรียกว่าแบบ

### 4.4.1 จุดประสงค์การทดลอง

- 1) เพื่อนำผลจากการทดลองที่ 1.2 มาทำการฝึกฝนโมเดล และวิเคราะห์ของโมเดล ที่ดีที่สุด โดยใช้การวัดผลด้วยค่า F1-Score
- 2) เพื่อเปรียบเทียบประสิทธิภาพก่อนและหลังการปรับพารามิเตอร์ด้วยค่า Accuracy, Precision, Recall และ F1-Score

### 4.4.2 วิธีการทดลอง

- 1) นำชุดข้อมูลที่ทำความสะอาดแล้วมาทำการเตรียมข้อมูลเมื่ອนการทดลองที่ 1.1
- 2) นำชุดข้อมูลสำหรับฝึกฝนมาทำการสังเคราะห์เพิ่มปริมาณข้อมูลของคลาสส่วนน้อย ให้มีปริมาณเท่ากับคลาสส่วนมาก โดยใช้เทคนิคการสังเคราะห์ข้อมูลที่ได้จากผลการทดลองที่ 1.2
- 3) นำชุดข้อมูลสำหรับฝึกฝนไปทำการฝึกฝนโมเดลด้วยอัลกอริทึมที่ได้จากผลการทดลองที่ 1.2 ซึ่งก็คือ XGBoost !!แล้วทำการปรับจูนพารามิเตอร์ของโมเดลโดยใช้การค้นหาแบบ กริด (Grid Search) ดังนี้

n_estimators	[100, 1000]
gamma	[0.5, 1, 2]
max_depth	[5, 6, 7]
subsample	[0.5, 0.7, 1]
colsample_bytree	[0.5, 0.7, 1]
learning_rate,	[0.3, 0.1]

- 4) นำชุดข้อมูลสำหรับฝึกฝนมาสร้างโมเดลโดยใช้พารามิเตอร์ที่ปรับจูนแล้ว
- 5) วัดประสิทธิภาพของโมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score โดยใช้ชุดข้อมูลสำหรับทดสอบ
- 6) เปรียบเทียบประสิทธิภาพก่อนและหลังปรับจูนพารามิเตอร์ของแต่ละโมเดล
- 7) สรุปผลการทดลอง

#### 4.4.3 ผลการทดสอบ

ตาราง 4.19 การเปรียบเทียบประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score ของโมเดล XGBoost ก่อน-หลังปรับจูนพารามิเตอร์ของชุดข้อมูลแบ่งที่เรียกกรรมบวก

ลำดับ	โมเดล	ก่อน-หลัง การปรับจูน พารามิเตอร์	ผลการวัดประสิทธิภาพของโมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amikacin	ก่อน	<b>0.960</b>	<b>0.644</b>	<b>0.763</b>	<b>0.699</b>
		หลัง	0.946	0.537	<b>0.763</b>	0.630
2	Amoxicillin/clavulanic acid	ก่อน	<b>0.931</b>	<b>0.902</b>	0.950	<b>0.926</b>
		หลัง	0.928	0.874	<b>0.982</b>	0.925
3	Cefalexin	ก่อน	0.920	<b>0.829</b>	0.912	0.868
		หลัง	<b>0.925</b>	0.825	<b>0.939</b>	<b>0.879</b>
4	Cefovecin	ก่อน	<b>0.936</b>	<b>0.863</b>	0.943	<b>0.901</b>
		หลัง	0.934	0.852	<b>0.953</b>	0.900
5	Clindamycin	ก่อน	<b>0.918</b>	<b>0.785</b>	0.952	<b>0.860</b>
		หลัง	0.910	0.758	<b>0.970</b>	0.851
6	Doxycycline	ก่อน	0.958	0.615	0.686	0.649
		หลัง	<b>0.962</b>	<b>0.622</b>	<b>0.800</b>	<b>0.700</b>
7	Enrofloxacin	ก่อน	<b>0.971</b>	<b>0.632</b>	0.522	0.571
		หลัง	0.970	0.591	<b>0.565</b>	<b>0.578</b>
8	Marbofloxacin	ก่อน	<b>0.957</b>	<b>0.561</b>	<b>0.719</b>	<b>0.630</b>
		หลัง	<b>0.957</b>	<b>0.561</b>	<b>0.719</b>	<b>0.630</b>
9	Nitrofurantoin	ก่อน	0.968	0.727	0.800	0.762
		หลัง	<b>0.971</b>	<b>0.739</b>	<b>0.850</b>	<b>0.791</b>
10	Trimethoprim/sulfamethoxazole	ก่อน	0.845	0.717	<b>0.874</b>	<b>0.788</b>
		หลัง	<b>0.846</b>	<b>0.727</b>	0.854	0.786
11	Vancomycin	ก่อน	<b>0.962</b>	<b>0.583</b>	0.875	0.700
		หลัง	<b>0.962</b>	0.580	<b>0.906</b>	<b>0.707</b>

ตาราง 4.20 การเปรียบเทียบประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score ของโมเดล XGBoost ก่อน-หลังปรับจูนพารามิเตอร์ของชุดข้อมูลแบคทีเรียแกรมลบ

ลำดับ	โมเดล	ก่อน-หลัง การปรับจูน พารามิเตอร์	ผลการวัดประสิทธิภาพของโมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amikacin	ก่อน	0.856	0.709	<b>0.770</b>	0.738
		หลัง	<b>0.861</b>	<b>0.722</b>	<b>0.770</b>	<b>0.745</b>
2	Amoxicillin/clavulanic acid	ก่อน	0.949	<b>0.857</b>	0.955	0.904
		หลัง	<b>0.950</b>	0.842	<b>0.987</b>	<b>0.909</b>
3	Cefalexin	ก่อน	0.936	0.767	0.784	0.775
		หลัง	<b>0.950</b>	<b>0.800</b>	<b>0.864</b>	<b>0.831</b>
4	Cefovecin	ก่อน	0.911	<b>0.816</b>	0.881	0.847
		หลัง	<b>0.923</b>	0.814	<b>0.943</b>	<b>0.874</b>
5	Doxycycline	ก่อน	<b>0.947</b>	<b>0.429</b>	<b>0.536</b>	<b>0.476</b>
		หลัง	0.942	0.395	<b>0.536</b>	0.455
6	Enrofloxacin	ก่อน	0.947	<b>0.825</b>	0.833	0.829
		หลัง	<b>0.952</b>	0.817	<b>0.885</b>	<b>0.850</b>
7	Gentamicin	ก่อน	<b>0.920</b>	<b>0.565</b>	0.661	0.609
		หลัง	0.917	0.544	<b>0.729</b>	<b>0.623</b>
8	Imipenem	ก่อน	<b>0.922</b>	<b>0.758</b>	0.843	<b>0.798</b>
		หลัง	0.917	0.730	<b>0.870</b>	0.794
9	Marbofloxacin	ก่อน	0.952	<b>0.914</b>	0.941	0.927
		หลัง	<b>0.955</b>	0.907	<b>0.961</b>	<b>0.933</b>
10	Nitrofurantoin	ก่อน	<b>0.976</b>	<b>0.750</b>	0.951	<b>0.839</b>
		หลัง	0.974	0.727	<b>0.976</b>	0.833
11	Trimethoprim/ sulfamethoxazole	ก่อน	0.893	<b>0.801</b>	0.928	0.860
		หลัง	<b>0.898</b>	0.792	<b>0.964</b>	<b>0.869</b>

#### 4.4.4 วิเคราะห์ผลการทดลอง

จากผลการทดลองในตารางที่ 4.19 การเปรียบเทียบประสิทธิภาพก่อนและหลังปรับจูนพารามิเตอร์ของโมเดลกลุ่มแบคทีเรียแกรมบวก พบว่า ส่วนใหญ่หลังปรับจูนพารามิเตอร์จะมีประสิทธิภาพที่ดีขึ้น เช่น ยา Doxycycline มีค่า F1-Score เพิ่มขึ้นจาก 0.649 เป็น 0.700 และ Recall เพิ่มขึ้นจาก 0.686 เป็น 0.800 แต่ยังคงมีบางโมเดลที่ก่อนปรับจูนพารามิเตอร์ให้ประสิทธิภาพมากกว่าเพียงเล็กน้อย เช่น โมเดลยา Amoxicillin/clavulanic acid มีค่า F1-Score ลดลงจาก 0.926 เป็น 0.925

จากผลการทดลองในตารางที่ 4.20 การเปรียบเทียบประสิทธิภาพก่อน-หลังปรับจูนพารามิเตอร์ของโมเดลกลุ่มแบคทีเรียแกรมลบ พบว่า ส่วนใหญ่หลังปรับจูนพารามิเตอร์จะมีประสิทธิภาพที่ดีขึ้น เช่น ยา Cefalexin มีค่า F1-Score เพิ่มขึ้นจาก 0.775 เป็น 0.831 และ Recall เพิ่มขึ้นจาก 0.784 เป็น 0.864 แต่ยังคงมีบางโมเดลที่ก่อนปรับจูนพารามิเตอร์ให้ประสิทธิภาพมากกว่าเพียงเล็กน้อย เช่น โมเดลยา Doxycycline มีค่า F1-Score ลดลงจาก 0.476 เป็น 0.455

#### 4.4.5 สรุปผลการทดลอง

จากการทดลองสรุปผลได้ว่า การปรับจูนพารามิเตอร์ของโมเดล XGBoost สามารถช่วยเพิ่มประสิทธิภาพของโมเดลได้บางโมเดล โดยส่วนใหญ่ทั้งโมเดลกลุ่มแบคทีเรียแกรมบวก และโมเดลกลุ่มแบคทีเรียแกรมลบหลังการปรับจูนพารามิเตอร์จะมีประสิทธิภาพที่ดีขึ้น โดยโมเดลยา Cefalexin ของกลุ่มแบคทีเรียแกรมลบ สามารถเพิ่มค่า F1-Score ได้มากจาก 0.775 เป็น 0.831 และโมเดลยา Doxycycline ของกลุ่มแบคทีเรียแกรมบวก สามารถเพิ่มค่า F1-Score ได้มากจาก 0.649 เป็น 0.700

#### 4.5 สรุปผลการสร้างโมเดลแนะนำยาต้านจุลชีพ

จากการสรุปผลการทดลองในตอนที่ 1 ผู้จัดทำได้เลือกเทคนิคการจำแนกข้อมูลเป็น XGBoost ในการฝึกฝน โมเดลทั้งหมด และได้แก้ไขปัญหาชุดข้อมูลไม่สมดุล โดยใช้เทคนิคการสังเคราะห์ข้อมูลที่แตกต่างกัน ไปตามแต่ละชุดข้อมูล รวมทั้งทำการปรับจูนพารามิเตอร์ XGBoost โดยผลจากการวัดประสิทธิภาพของ โมเดลด้วยชุดข้อมูลสำหรับทดสอบแสดงตามตาราง ดังนี้

ตาราง 4.21 การวัดประสิทธิภาพของ โมเดลกู้มแบคทีเรียแกรมบวกด้วยค่า Accuracy, Precision, Recall และ F1-Score

ลำดับ	โมเดล	เทคนิคการสังเคราะห์ข้อมูล	ผลการวัดประสิทธิภาพของ โมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amikacin	BL-S	0.946	0.537	0.763	0.630
2	Amoxicillin/clavulanic acid	BL-S	0.928	0.874	0.982	0.925
3	Cefalexin	SM	0.925	0.825	0.939	0.879
4	Cefovecin	SVM-S	0.934	0.852	0.953	0.900
5	Clindamycin	SVM-S	0.910	0.758	0.970	0.851
6	Doxycycline	R-S	0.962	0.622	0.800	0.700
7	Enrofloxacin	R-S	0.970	0.591	0.565	0.578
8	Marbofloxacin	BL-S	0.957	0.561	0.719	0.630
9	Nitrofurantoin	R-S	0.971	0.739	0.850	0.791
10	Trimethoprim/sulfamethoxazole	SVM-S	0.846	0.727	0.854	0.786
11	Vancomycin	SM	0.962	0.580	0.906	0.707

ตาราง 4.22 การวัดประสิทธิภาพของโมเดลกู้่มแบคทีเรียแกรมลบด้วยค่า Accuracy, Precision, Recall และ F1-Score

ลำดับ	โมเดล	เทคนิคการ สังเคราะห์ ข้อมูล	ผลการวัดประสิทธิภาพของโมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amikacin	R-S	0.861	0.722	0.770	0.745
2	Amoxicillin/clavulanic acid	SVM-S	0.950	0.842	0.987	0.909
3	Cefalexin	R-S	0.950	0.800	0.864	0.831
4	Cefovecin	ADY	0.923	0.814	0.943	0.874
5	Doxycycline	BL-S	0.942	0.395	0.536	0.455
6	Enrofloxacin	R-S	0.952	0.817	0.885	0.850
7	Gentamicin	R-S	0.917	0.544	0.729	0.623
8	Imipenem	SM	0.917	0.730	0.870	0.794
9	Marbofloxacin	R-S	0.955	0.907	0.961	0.933
10	Nitrofurantoin	SVM-S	0.974	0.727	0.976	0.833
11	Trimethoprim/ sulfamethoxazole	SVM-S	0.898	0.792	0.964	0.869

ตาราง 4.23 การวัดประสิทธิภาพของโมเดลแบบทดสอบรายเคส (Test By Case) ด้วยค่า Accuracy, Precision, Recall และ F1-Score

โมเดล	จำนวนชุดข้อมูล สำหรับทดสอบ	ผลการวัดประสิทธิภาพ			
		Accuracy	Precision	Recall	F1-Score
กลุ่มแบคทีเรียแกรมบวก	263	0.933	0.771	0.811	0.774
กลุ่มแบคทีเรียแกรมลบ	296	0.941	0.829	0.890	0.841

การวัดประสิทธิภาพแบบทดสอบรายเคสของ โมเดลกู้่มแบคทีเรียแกรมบวก ใช้ชุดข้อมูล สำหรับทดสอบ 263 เคส โดยมีเคสที่ โมเดลแนะนำถูกต้องทุกnya 149 เคส ส่วน โมเดลกู้่มแบคทีเรีย แกรมลบ ใช้ชุดข้อมูลสำหรับทดสอบ 296 เคส โดยมีเคสที่ โมเดลแนะนำถูกต้องทุกnya 179 เคส

## ตอนที่ 2 การฝึกฝนโมเดลทำนายการดื้อยาต้านจุลชีพ

### 4.6 การทดลองที่ 2.1 การทำนายการดื้อยาต้านจุลชีพของแบคทีเรียแกรมบวก

#### 4.6.1 ชุดประสิทธิภาพทดลอง

1) เพื่อฝึกฝนโมเดลทำนายการดื้อยาต้านจุลชีพของแบคทีเรียแกรมบวกโดยใช้ชุดข้อมูลรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยา และวัดประสิทธิภาพของโมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score

2) วิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์โดยใช้ค่าสัมประสิทธิ์ครามอร์ (Cramer's V Coefficient) เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิดในกลุ่มแบคทีเรียแกรมบวก

#### 4.6.2 วิธีการทดลอง

1) เลือกยาต้านจุลชีพที่จะนำมาฝึกฝนโมเดลทำนายการดื้อยา โดยเลือกจากยาที่ได้นำไปฝึกฝนโมเดลแนะนำยาต้านจุลชีพในการทดลองตอนที่ 1 จำนวน 8 ยา ได้แก่ Amoxicillin/clavulanic acid, Cefovecin, Clindamycin, Doxycycline, Enrofloxacin, Gentamicin, Marbofloxacin, Trimethoprim/sulfamethoxazole

2) นำชุดข้อมูลที่ทำการทดสอบแล้วมาทำการจัดเตรียมข้อมูล ดังนี้  
 - จัดกลุ่มหมวดหมู่เล็ก ๆ ของฟีเจอร์ตัวอย่างที่ส่งตรวจ (submitted\_sample) ที่มีจำนวนน้อยกว่า 10 รวมกันให้เป็นหมวดหมู่ใหม่ชื่อว่า “other”  
 - หาค่าสัมประสิทธิ์ครามอร์เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิดในกลุ่มแบคทีเรียแกรมบวก

- ทำการแบ่งชุดข้อมูลสำหรับฝึกฝนโมเดลของแต่ละยา ซึ่งมี 2 คลาส คือ True (Resistant: R) และ False (Sensitive: S) โดยการฝึกฝนแต่ละโมเดลนั้นจะแบ่งออกเป็น 2 ชุดข้อมูล (80:20) คือ ชุดข้อมูลสำหรับฝึกฝน 80% และชุดข้อมูลสำหรับทดสอบ 20% โดยแบ่งให้ข้อมูลสำหรับฝึกฝน และข้อมูลสำหรับทดสอบ มีอัตราส่วนของ 2 คลาส (True/False) เท่ากัน

- ใช้วิธีการเข้ารหัสแบบวันshotในการเปลี่ยนข้อมูลให้อยู่ในรูปแบบตัวเลข  
 3) ทำการเลือกฟีเจอร์ที่ใช้ในการฝึกฝนโมเดล ได้แก่ ชนิดของสัตว์เลี้ยง (species), สกุลแบคทีเรีย (bacteria\_genus), ตัวอย่างที่ส่งตรวจ (submitted\_sample), ผลการทดสอบ cefoxitin screen และ ผลการทดสอบ inducible clindamycin resistance

4) นำชุดข้อมูลสำหรับฝึกฝนของแต่ละยามาทำการสังเคราะห์ข้อมูลของคลาสส่วนน้อยให้มีปริมาณเท่ากับคลาสส่วนมาก โดยใช้เทคนิค SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN และ R-SMOTE แล้วทำการฝึกฝน โมเดลของยาต้านจุลชีพแต่ละชนิด โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost โดยทำการทดสอบประสิทธิภาพด้วย 10-Fold Cross-Validation จากนั้นเลือกเทคนิคการสังเคราะห์ข้อมูลที่ได้ค่า F1-Score สูงที่สุดของโมเดลยาแต่ละชนิด

5) ทำการปรับจุนพารามิเตอร์ XGBoost ของโมเดลยาแต่ละชนิด โดยใช้การค้นหาแบบกริด (Grid Search) ซึ่งมีพารามิเตอร์ที่ปรับจุน ดังนี้

n_estimators	[100, 1000]
gamma	[0.5, 1, 2]
max_depth	[5, 6, 7]
subsample	[0.5, 0.7, 1]
colsample_bytree	[0.5, 0.7, 1]
learning_rate	[0.3, 0.1]

6) นำชุดข้อมูลสำหรับฝึกฝนมาทำการฝึกฝน โมเดลของยาแต่ละชนิด โดยใช้เทคนิคการสังเคราะห์ข้อมูลที่ได้เลือกไว้จากข้อที่ 4 และใช้พารามิเตอร์ XGBoost ที่ปรับจุนแล้วจากข้อที่ 5

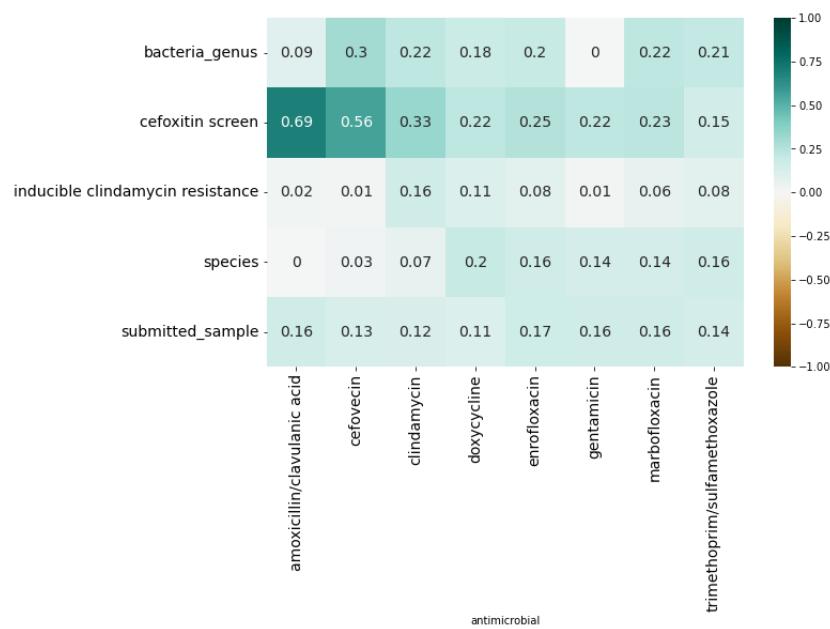
7) วัดประสิทธิภาพของ โมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score โดยใช้ชุดข้อมูลสำหรับทดสอบ

8) สรุปผลการทดสอบ

#### 4.6.3 ผลการทดสอบ

ตาราง 4.24 ข้อมูลการดื้อยาต้านจุลชีพของแบคทีเรียแกรมบวก

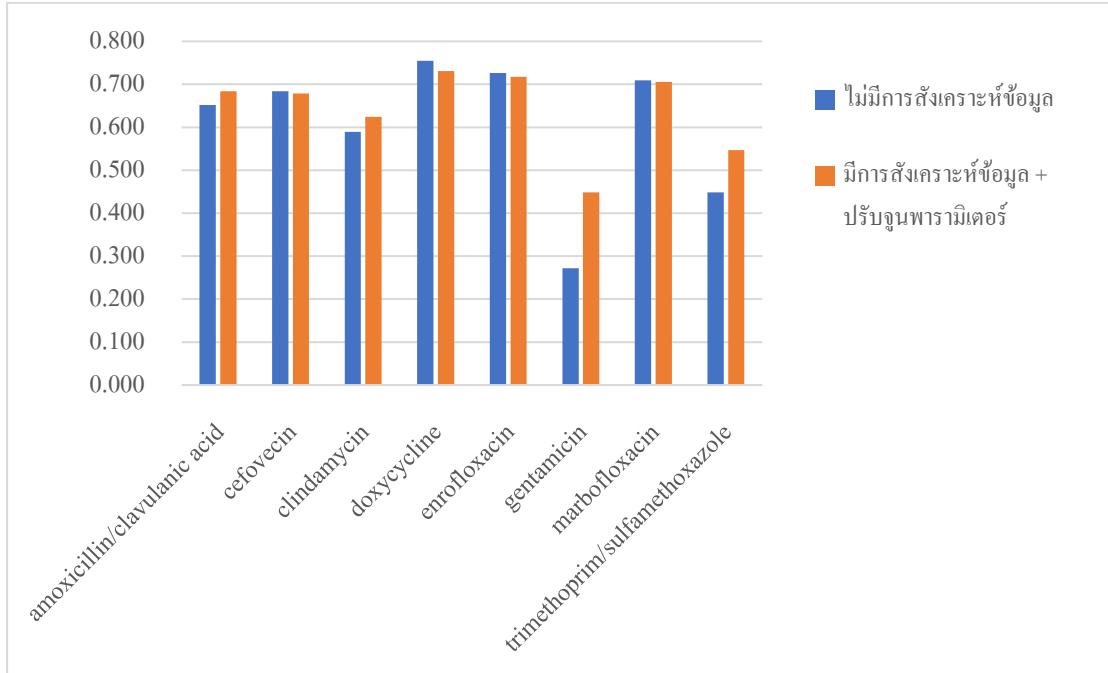
ลำดับ	ชื่อยาต้านจุลชีพ	จำนวนข้อมูลที่ไม่มีการดื้อยา (Sensitive)	จำนวนข้อมูลที่มีการดื้อยา (Resistance)	จำนวนข้อมูลทั้งหมด
1	Amoxicillin/clavulanic acid	2197	363	2560
2	Cefovecin	1765	833	2598
3	Clindamycin	1431	1175	2606
4	Doxycycline	378	500	878
5	Enrofloxacin	1143	1477	2620
6	Gentamicin	1580	603	2183
7	Marbofloxacin	1227	1638	2865
8	Trimethoprim/sulfamethoxazole	1572	1059	2631



รูปที่ 4.7 ค่าสัมประสิทธิ์相關เอมอร์ระหว่างฟิเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพในกลุ่มแบคทีเรียแกรมบวก

ตาราง 4.25 ผลการวัดประสิทธิภาพของโมเดลที่มีรายการดื้อยาต้านจุลชีพในกลุ่มแบบที่เรียกว่า  
บวกด้วยค่า Accuracy, Precision, Recall และ F1-Score

ลำดับ	โมเดล	เทคนิคการสังเคราะห์ข้อมูล	ผลการวัดประสิทธิภาพของโมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amoxicillin/clavulanic acid	R-S	0.906	0.658	0.712	0.684
2	Cefovecin	SM	0.825	0.828	0.575	0.678
3	Clindamycin	BL-S	0.670	0.641	0.609	0.624
4	Doxycycline	-	0.699	0.742	0.720	0.731
5	Enrofloxacin	-	0.647	0.653	0.797	0.718
6	Gentamicin	ADY	0.549	0.339	0.661	0.448
7	Marbofloxacin	-	0.654	0.689	0.723	0.705
8	Trimethoprim/ sulfamethoxazole	BL-S	0.632	0.542	0.552	0.547



รูปที่ 4.8 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของโมเดลกลุ่มแบบที่เรียกว่าบวก  
ระหว่างโมเดลที่ไม่มีสังเคราะห์ข้อมูลกับโมเดลที่มีการสังเคราะห์ข้อมูลพร้อมกับการปรับจูน  
พารามิเตอร์ของ XGBoost

#### 4.6.4 วิเคราะห์ผลการทดลอง

จากผลการทดลองในรูปที่ 4.7 ค่าสัมประสิทธิ์ Ramsey ระบุว่า การดื้อยาต้านจุลชีพทั้ง 8 ชนิดในกลุ่มแบคทีเรียแกรมบวกกับฟิเจอร์ต่าง ๆ พบว่า การดื้อยาต้านจุลชีพมีความสัมพันธ์ฟิเจอร์ต่าง ๆ โดยผลการทดสอบ cefoxitin screen มีความสัมพันธ์ก่อนข้างมากกับการดื้อยา Amoxicillin/clavulanic acid และ Cefovecin อีกทั้งยังมีความสัมพันธ์ปานกลางกับการดื้อยา Clindamycin ส่วนความสัมพันธ์ระหว่างการดื้อยาต้านจุลชีพทั้ง 8 ชนิดกับฟิเจอร์อื่น ๆ ได้แก่ ชนิดของสัตว์เลี้ยง (species), สกุลแบคทีเรีย (bacteria\_genus), ตัวอย่างที่ส่งตรวจ (submitted\_sample) และผลการทดสอบ inducible clindamycin resistance พบว่ามีความสัมพันธ์กันน้อย

จากผลการทดลองในตารางที่ 4.25 พบว่า มีการใช้เทคนิคสังเคราะห์ข้อมูลกับชุดข้อมูลของยาต้านจุลชีพเพียง 5 ชนิด เนื่องจากบางชุดข้อมูล ได้แก่ ยา Doxycycline, Enrofloxacin และ Marbofloxacin มีข้อมูลที่ดื้อยา (Resistance) มากกว่าข้อมูลที่ไม่ดื้อยา (Sensitive) และค่อนข้างมีความสมดุลอยู่แล้ว เมื่อใช้เทคนิคการสังเคราะห์เพื่อเพิ่มปริมาณข้อมูลจะทำให้ค่า Recall และ F1-Score ต่ำลง จึงไม่เหมาะสมที่จะใช้เทคนิคการสังเคราะห์ข้อมูล โดยผลลัพธ์ที่ได้จากการฝึกฝนโมเดลพบว่า ประสิทธิภาพของโมเดลส่วนใหญ่มีค่า F1-Score มากกว่า 0.600 ได้แก่ ยา Amoxicillin/clavulanic acid (0.684), Cefovecin (0.678), Clindamycin (0.624), Doxycycline (0.731), Enrofloxacin (0.718) และ Marbofloxacin (0.705) ส่วนโมเดลที่มีค่า F1-Score น้อยกว่า 0.600 ได้แก่ Gentamicin (0.448) และ Trimethoprim/sulfamethoxazole (0.547)

จากผลการทดลองในรูปที่ 4.8 พบว่า เมื่อเปรียบเทียบด้วยค่า F1-Score โมเดลที่มีการใช้เทคนิคการสังเคราะห์ข้อมูลพร้อมกับการปรับจูนพารามิเตอร์ของ XGBoost จะมีประสิทธิภาพที่ดีกว่าโมเดลที่ไม่มีการใช้เทคนิคการสังเคราะห์ข้อมูล โดยเฉพาะ โมเดลยา Gentamicin และ Trimethoprim/sulfamethoxazole ซึ่งจะเห็นได้ว่าหากใช้เทคนิคการสังเคราะห์ข้อมูลพร้อมกับการปรับจูนพารามิเตอร์ของ XGBoost สามารถเพิ่มค่า F1-Score ได้มากกว่าเดิมอย่างชัดเจน โดยโมเดลยา Gentamicin เพิ่มค่า F1-Score จาก 0.272 เป็น 0.448 และ โมเดลยา Trimethoprim/sulfamethoxazole เพิ่มค่า F1-Score จาก 0.448 เป็น 0.547

#### 4.6.5 สรุปผลการทดลอง

จากผลการทดลอง ผู้จัดทำได้เลือกยาต้านจุลชีพมา 8 ชนิดเพื่อทดลองฝึกฝน โมเดลท่านายการดื้อยาสำหรับเชื้อแบคทีเรียแกรมบวก โดยทำการหาค่าสัมประสิทธิ์ Ramsey เพื่อแสดงให้เห็นถึงความสัมพันธ์ระหว่างการดื้อยาต้านจุลชีพทั้ง 8 ชนิดกับฟิเจอร์ต่าง ๆ ซึ่งจากการทดลองนี้

สรุปผลได้ว่า สามารถสร้างโมเดลทำนายการดื้อยาสำหรับเชื้อแบคทีเรียแกรมลบ ได้โดยใช้ข้อมูลผลตรวจจากห้องปฏิบัติการจุลชีววิทยา ได้แก่ ชนิดของสัตว์เลี้ยง (species), กลุ่มแบคทีเรีย (bacteria\_genus), ตัวอย่างที่ส่งตรวจ (submitted\_sample), ผลการทดสอบ cefoxitin screen และ ผลการทดสอบ inducible clindamycin resistance โดยได้มีการแก้ไขปัญหาชุดข้อมูลไม่สมดุลด้วยการใช้เทคนิคการสังเคราะห์ข้อมูลพร้อมกับการปรับจูนพารามิเตอร์ของ XGBoost ซึ่งการทดลองนี้ได้แสดงให้เห็นว่าเทคนิคนี้สามารถช่วยเพิ่มประสิทธิภาพของค่า F1-Score ให้แก่โมเดลได้ โดยประสิทธิภาพของโมเดลยาส่วนใหญ่ได้ค่า F1-Score มากกว่า 0.600 ยกเว้นยา Gentamicin (0.448) และ Trimethoprim/ sulfamethoxazole (0.547) ซึ่ง โมเดลทำนายการดื้อยาต้านจุลชีพที่มีค่า F1-Score สูงสุด คือ Doxycycline (0.731)

## 4.7 การทดลองที่ 2.2 การทำนายการดื้อยาต้านจุลชีพของแบคทีเรียแกรมลบ

### 4.7.1 จุดประสงค์การทดลอง

1) เพื่อฝึกฝน โมเดลทำนายการดื้อยาต้านจุลชีพของแบคทีเรียแกรมลบ โดยใช้ชุดข้อมูลรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อยา และวัดประสิทธิภาพของโมเดล ด้วยค่า Accuracy, Precision, Recall และ F1-Score

2) วิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์โดยใช้ค่าสัมประสิทธิ์ครามอร์ (Cramer's V Coefficient) เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิดในกลุ่มแบคทีเรียแกรมลบ

### 4.7.2 วิธีการทดลอง

1) เลือกยาต้านจุลชีพที่จะนำมาฝึกฝน โมเดลทำนายการดื้อยา โดยเลือกจากยาที่ได้นำไปฝึกฝน โมเดลแนะนำยาต้านจุลชีพในการทดลองตอนที่ 1 จำนวน 8 ยา ได้แก่ Amoxicillin/clavulanic acid, Cefalexin, Cefovecin, Enrofloxacin, Gentamicin, Marbofloxacin, Nitrofurantoin, Trimethoprim/sulfamethoxazole

2) นำชุดข้อมูลที่ทำการจำแนกตามฟีเจอร์ตัวอย่างที่ส่งตรวจ (submitted\_sample) ที่มีจำนวนน้อยกว่า 10 รวมกันให้เป็นหมวดหมู่ใหม่ชื่อว่า “other”  
- จัดกลุ่มหมวดหมู่เล็ก ๆ ของฟีเจอร์ตัวอย่างที่ส่งตรวจ (submitted\_sample) ที่มีจำนวนน้อยกว่า 10 รวมกันให้เป็นหมวดหมู่ใหม่ชื่อว่า “other”

- หาค่าสัมประสิทธิ์ความอร์เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิดในกลุ่มแบคทีเรียแกรมลบ

- ทำการแบ่งชุดข้อมูลสำหรับฝึกฝน โนเมเดลของแต่ละยา ซึ่งมี 2 คลาส คือ True (Resistant: R) และ False (Sensitive: S) โดยการฝึกฝนแต่ละ โนเมเดลนั้นจะแบ่งออกเป็น 2 ชุดข้อมูล (80:20) คือ ชุดข้อมูลสำหรับฝึกฝน 80% และชุดข้อมูลสำหรับทดสอบ 20% โดยแบ่งให้ชุดข้อมูลสำหรับฝึกฝน และชุดข้อมูลสำหรับทดสอบมีอัตราส่วนของ 2 คลาส (True/False) เท่ากัน

- ใช้วิธีการเข้ารหัสแบบวันshotในการเปลี่ยนข้อมูลให้อยู่ในรูปแบบตัวเลข

3) ทำการเลือกฟีเจอร์ที่ใช้ในการฝึกฝน โนเมเดล ได้แก่ ชนิดของสัตว์เลี้ยง (species), ถูกกลุ่มแบคทีเรีย (bacteria\_genus), ตัวอย่างที่ส่งตรวจ (submitted\_sample), และผลการทดสอบ esbl

4) นำชุดข้อมูลสำหรับฝึกฝนของแต่ละยา มาทำการสังเคราะห์ข้อมูลของคลาสส่วนน้อยให้มีปริมาณเท่ากับคลาสส่วนมาก โดยใช้เทคนิค SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN และ R-SMOTE แล้วทำการฝึกฝน โนเมเดลของยาต้านจุลชีพแต่ละชนิด โดยใช้เทคนิคการจำแนกข้อมูลด้วย XGBoost ทำการทดสอบประสิทธิภาพด้วย 10-Fold Cross-Validation จากนั้นเลือกเทคนิคการสังเคราะห์ข้อมูลที่ได้ค่า F1-Score สูงที่สุดของ โนเมเดลยาแต่ละชนิด

5) ทำการปรับจุนพารามิเตอร์ XGBoost ของ โนเมเดลยาแต่ละชนิด โดยใช้การค้นหาแบบกริด (Grid Search) ซึ่งมีพารามิเตอร์ที่ปรับจุน ดังนี้

n_estimators	[100, 1000]
gamma	[0.5, 1, 2]
max_depth	[5, 6, 7]
subsample	[0.5, 0.7, 1]
colsample_bytree	[0.5, 0.7, 1]
learning_rate,	[0.3, 0.1]

6) ทำการฝึกฝน โนเมเดลของยาแต่ละชนิด โดยใช้เทคนิคการสังเคราะห์ข้อมูลที่ได้เลือกไว้จากข้อที่ 4 และใช้พารามิเตอร์ XGBoost ที่ปรับจุนแล้วจากข้อที่ 5

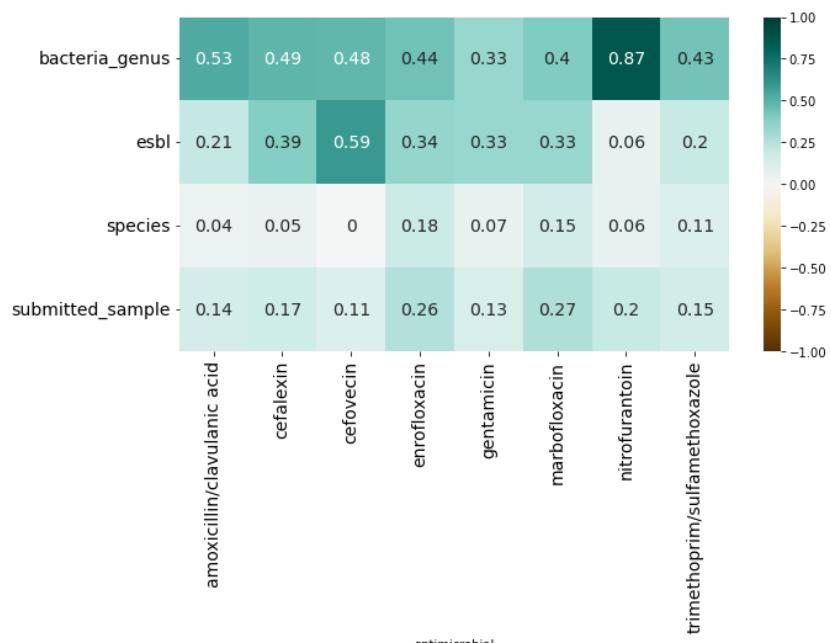
7) วัดประสิทธิภาพของ โนเมเดลด้วยค่า Accuracy, Precision, Recall และ F1-Score โดยใช้ชุดข้อมูลสำหรับทดสอบ

8) สรุปผลการทดลอง

#### 4.7.3 ผลการทดสอบ

ตาราง 4.26 ข้อมูลการดื้อยาต้านจุลชีพของแบคทีเรียแกรมลบ

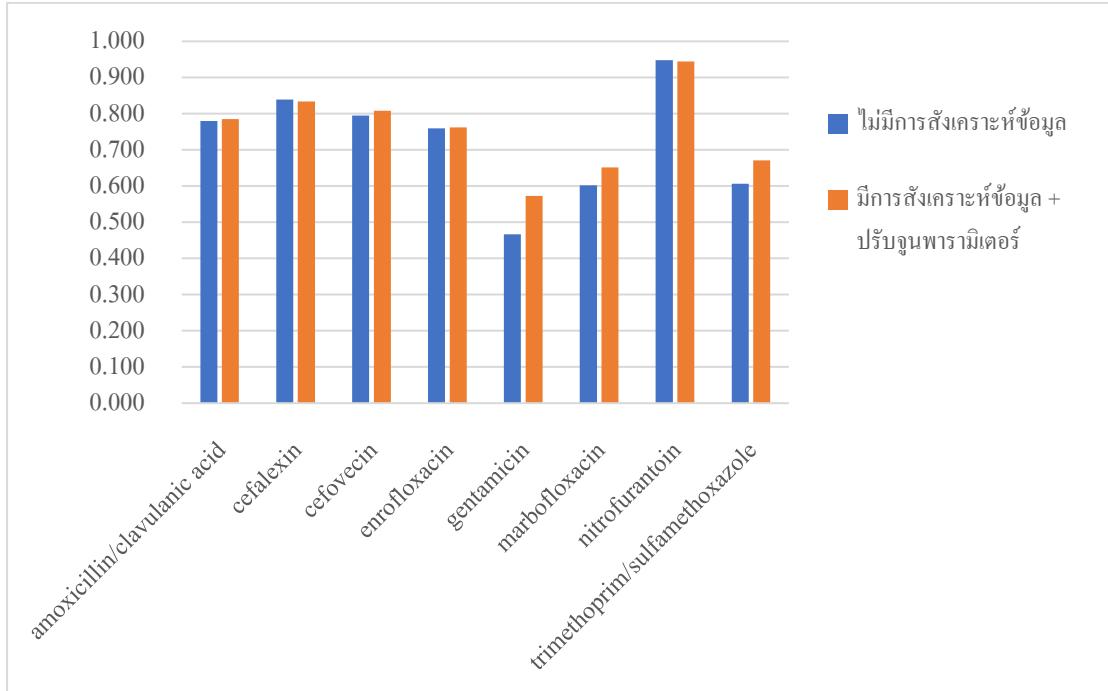
ลำดับ	ชื่อยาต้านจุลชีพ	จำนวนข้อมูลที่ไม่มีการดื้อยา (Sensitive)	จำนวนข้อมูลที่มีการดื้อยา (Resistance)	จำนวนข้อมูลทั้งหมด
1	Amoxicillin/clavulanic acid	995	1389	2384
2	Cefalexin	793	1592	2385
3	Cefovecin	1222	1122	2344
4	Enrofloxacin	1125	1584	2709
5	Gentamicin	2004	1193	3197
6	Marbofloxacin	1751	1469	3220
7	Nitrofurantoin	1075	994	2069
8	Trimethoprim/sulfamethoxazole	1532	1311	2843



รูปที่ 4.9 ค่าสัมประสิทธิ์ correlation ระหว่างพิจารณาต่าง ๆ กับการดื้อยาต้านจุลชีพในกลุ่มแบคทีเรียแกรมลบ

ตาราง 4.27 ผลการวัดประสิทธิภาพของโมเดลที่มีแนวโน้มในการคือยาต้านจุลชีพในกลุ่มแบนคทีเรียแกรมลบด้วยค่า Accuracy, Precision, Recall และ F1-Score

ลำดับ	โภเมเดล	เทคนิคการสังเคราะห์ข้อมูล	ผลการวัดประสิทธิภาพของโภเมเดล			
			Accuracy	Precision	Recall	F1-Score
1	Amoxicillin/clavulanic acid	-	0.753	0.794	0.777	0.785
2	Cefalexin	-	0.774	0.816	0.852	0.834
3	Cefovecin	SVM-S	0.825	0.848	0.772	0.808
4	Enrofloxacin	-	0.694	0.698	0.839	0.762
5	Gentamicin	SM	0.664	0.545	0.603	0.573
6	Marbofloxacin	SVM-S	0.644	0.590	0.728	0.651
7	Nitrofurantoin	SVM-S	0.944	0.911	0.980	0.944
8	Trimethoprim/sulfamethoxazole	BL-S	0.641	0.581	0.794	0.671



รูปที่ 4.10 การเปรียบเทียบประสิทธิภาพด้วยค่า F1-Score ของโภเมเดลกลุ่มแบนคทีเรียแกรมลบ ระหว่างโภเมเดลที่ไม่มีสังเคราะห์ข้อมูลกับโภเมเดลที่มีการสังเคราะห์ข้อมูลพร้อมกับการปรับจูนพารามิเตอร์ของ XGBoost

#### 4.7.4 วิเคราะห์ผลการทดลอง

จากผลการทดลองในรูปที่ 4.9 ค่าสัมประสิทธิ์รวมอัตราหัวงการคือยาต้านจุลชีพทั้ง 8 ชนิดในกลุ่มแบคทีเรียแกรมลบกับฟิเจอร์ต่าง ๆ พบว่า การคือยาต้านจุลชีพมีความสัมพันธ์ฟิเจอร์ต่าง ๆ โดยสกุลแบคทีเรีย (bacteria\_genus) มีความสัมพันธ์ค่อนข้างมากกับการคือยา Nitrofurantoin อีกทั้งยังมีความสัมพันธ์ปานกลางกับการคือยา Amoxicillin/clavulanic acid, Cefalexin, Cefovecin, Enrofloxacin, Gentamicin, Marbofloxacin และ Trimethoprim/sulfamethoxazole ส่วนผลการทดสอบ ESBL ความสัมพันธ์ค่อนข้างมากกับการคือยา Cefovecin อีกทั้งยังมีความสัมพันธ์ปานกลางกับการคือยา Cefalexin, Enrofloxacin, Gentamicin และ Marbofloxacin นอกจากนี้ตัวอย่างที่ส่งตรวจ (submitted\_sample) ยังมีความสัมพันธ์ปานกลางกับการคือยา Enrofloxacin และ Marbofloxacin ส่วนชนิดของสัตว์เดี่ยว (species) มีความสัมพันธ์น้อยกับการคือยาต้านจุลชีพทั้ง 8 ชนิด

จากผลการทดลองในตารางที่ 4.27 พบว่า มีการใช้เทคนิคสังเคราะห์ข้อมูลกับชุดข้อมูลของยาต้านจุลชีพเพียง 5 ชนิด เนื่องจากบางชุดข้อมูล ได้แก่ ยา Amoxicillin/clavulanic acid, Cefalexin และ Enrofloxacin มีข้อมูลที่ดื้อยา (Resistance) มากกว่า ข้อมูลที่ไม่ดื้อยา (Sensitive) เมื่อใช้เทคนิคการสังเคราะห์เพื่อเพิ่มปริมาณข้อมูลจะทำให้ค่า Recall และ F1-Score ต่างลง จึงไม่เหมาะสมที่จะใช้เทคนิคการสังเคราะห์ข้อมูล โดยผลลัพธ์ที่ได้จากการฝึกฝนโมเดล พบว่า ประสิทธิภาพของโมเดลส่วนใหญ่มีค่า F1-Score มากกว่า 0.700 ได้แก่ Amoxicillin/clavulanic acid (0.785), Cefalexin (0.834), Cefovecin (0.808), Enrofloxacin (0.762) และ Nitrofurantoin (0.944) ส่วนโมเดลที่มีค่า F1-Score น้อยกว่า 0.700 ได้แก่ Gentamicin (0.573), Trimethoprim/sulfamethoxazole (0.671) และ Marbofloxacin (0.651)

จากผลการทดลองในรูปที่ 4.10 พบว่า เมื่อเปรียบเทียบด้วยค่า F1-Score โมเดลที่มีการใช้เทคนิคการสังเคราะห์ข้อมูลพร้อมกับการปรับจูนพารามิเตอร์ของ XGBoost จะมีประสิทธิภาพที่ดีกว่าโมเดลที่ไม่มีการใช้เทคนิคการสังเคราะห์ข้อมูล โดยเฉพาะยา Gentamicin, Marbofloxacin และ Trimethoprim/sulfamethoxazole ซึ่งจะเห็นได้ว่าหากใช้เทคนิคการสังเคราะห์ข้อมูลพร้อมกับการปรับจูนพารามิเตอร์ของ XGBoost สามารถเพิ่มค่า F1-Score ได้มากกว่าเดิมอย่างชัดเจน โดยโมเดลยา Gentamicin เพิ่มค่า F1-Score จาก 0.467 เป็น 0.573, โมเดลยา Marbofloxacin เพิ่มค่า F1-Score จาก 0.602 เป็น 0.651 และ โมเดลยา Trimethoprim/sulfamethoxazole เพิ่มค่า F1-Score จาก 0.607 เป็น 0.671

#### 4.7.5 สรุปผลการทดลอง

จากผลการทดลอง ผู้จัดทำได้เลือกยาต้านจุลชีพมา 8 ชนิดเพื่อทดลองฝึกฝนโมเดล นำมายกตัวอย่างคือยาสำหรับเชื้อแบคทีเรียแกรมลบ โดยทำการหาค่าสัมประสิทธิ์รามอร์เพื่อแสดงให้เห็นถึงความสัมพันธ์ระหว่างการคือยาต้านจุลชีพทั้ง 8 ชนิดกับฟีเจอร์ต่าง ๆ ซึ่งจากการทดลองนี้ สรุปผลได้ว่า สามารถสร้างโมเดลนำมายกตัวอย่างคือยาสำหรับเชื้อแบคทีเรียแกรมลบได้โดยใช้ข้อมูลผลตรวจจากห้องปฏิบัติการจุลชีววิทยา ได้แก่ ชนิดของสัตว์เลี้ยง (species), กลุ่มแบคทีเรีย (bacteria\_genus), ตัวอย่างที่ส่งตรวจ (submitted\_sample) และ ผลการทดสอบ ESBL โดยได้มีการแก้ไขปัญหาชุดข้อมูลไม่สมดุลด้วยการใช้เทคนิคการสังเคราะห์ข้อมูลพร้อมกับการปรับจูนพารามิเตอร์ของ XGBoost ซึ่งการทดลองนี้ได้แสดงให้เห็นว่าเทคนิคนี้สามารถช่วยเพิ่มประสิทธิภาพของค่า F1-Score ให้แก่โมเดลได้ โดยประสิทธิภาพของโมเดลยาส่วนใหญ่ได้ค่า F1-Score มากกว่า 0.700 ยกเว้นยา Gentamicin (0.573), Trimethoprim/sulfamethoxazole (0.671) และ Marbofloxacin (0.651) ซึ่ง โมเดลนำมายกตัวอย่างคือยาต้านจุลชีพที่มีค่า F1-Score สูงสุด คือ Nitrofurantoin (0.944)

## บทที่ 5

### การสรุปผล และข้อเสนอแนะ

#### 5.1 สรุปผลที่ได้จากการวิเคราะห์

โครงการนี้ได้ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อสร้างโมเดลแนวนำการใช้ยาต้านจุลชีพที่เหมาะสมตามคำแนะนำของสัตวแพทย์ โดยใช้ชุดข้อมูลจากรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อของห้องปฏิบัติการจุลชีววิทยา โดยได้ทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้เทคนิคการจำแนกข้อมูล 4 เทคนิค ได้แก่ Decision Tree, Random Forest, SVM และ XGBoost และศึกษาวิธีการแก้ไขปัญหาชุดข้อมูลไม่สมดุล โดยใช้เทคนิคการสังเคราะห์ข้อมูลแบบต่าง ๆ ได้แก่ SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN และ R-SMOTE รวมทั้งการปรับพารามิเตอร์ของโมเดล XGBoost เพื่อเพิ่มประสิทธิภาพให้แก่โมเดล นอกจากนี้ยังมีการทดลองฝึกฝนโมเดลทำนายการดื้อยาต้านจุลชีพ พร้อมกับวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์เพื่อหาความสัมพันธ์ระหว่างฟีเจอร์ต่าง ๆ กับการดื้อยาต้านจุลชีพแต่ละชนิด อีกด้วย ได้ดำเนินการออกแบบ และพัฒนาเว็บแอปพลิเคชันสำหรับทดสอบการใช้งานโมเดล แนะนำยาต้านจุลชีพ และการอัปโหลดชุดข้อมูลเข้าสู่ระบบเพื่อฝึกฝนโมเดลใหม่ ให้รองรับกับข้อมูลชุดใหม่ที่เพิ่มเข้ามาในอนาคต ได้อย่างมีประสิทธิภาพ

#### 5.2 ปัญหาและอุปสรรค

- เนื่องจากชุดข้อมูลรายงานผลการระบุชนิดจุลชีพและการทดสอบความไวรับต่อของห้องปฏิบัติการจุลชีววิทยามีการจัดเก็บข้อมูลเป็นไฟล์ .docx จึงต้องทำการแปลงข้อมูลให้อยู่ในรูปของไฟล์ .csv ก่อนนำไปใช้งาน ซึ่งข้อมูลที่ได้จากการแปลงไฟล์นั้นอาจไม่ถูกต้องทั้งหมด จึงได้ทำการตรวจสอบข้อมูลจากไฟล์ .docx พบว่าไฟล์ที่เป็นผลการตรวจด้วยวิธี Disk diffusion Test มีการใส่ผลทดสอบความไวต่อยาต้านจุลชีพ (S/I/R) ไว้ในส่วนของการแนะนำยา ซึ่งปกติจะต้องใส่ไว้ในตารางผลการทดสอบ จึงทำให้แปลงข้อมูลได้ไม่ครบสมบูรณ์ และหากจะแปลงให้สมบูรณ์ก็ทำได้ยาก จึงได้ทำการลบข้อมูลที่ได้แปลงมาจากไฟล์ที่เป็นผลการตรวจด้วยวิธี Disk diffusion Test ออกไปจากชุดข้อมูล

- 2) ข้อมูลการแนะนำยาต้านจุลชีพบางชนิด เช่น ยา Doxycycline ของชุดข้อมูลแบคทีเรีย แกรมลบมีข้อมูลการแนะนำยาค่อนข้างน้อย และมีความไม่สมดุลของชุดข้อมูลสูง ทำให้ประสิทธิภาพของโมเดล (วัดด้วยค่า F1-Score) ต่ำกว่ามาตรฐานอื่น
- 3) ในโครงการนี้ใช้ข้อมูลจากห้องปฏิบัติการจุลชีววิทยาในการฝึกฝนโมเดล จึงอาจมีข้อจำกัดในการนำไปใช้งานหากมีปัจจัยอื่นร่วมด้วย เช่น โรคประจำตัวของสัตว์ หรือประวัติการใช้ยาต่าง ๆ ก่อนหน้านี้ของสัตว์

### 5.3 แนวทางในการพัฒนาต่อ

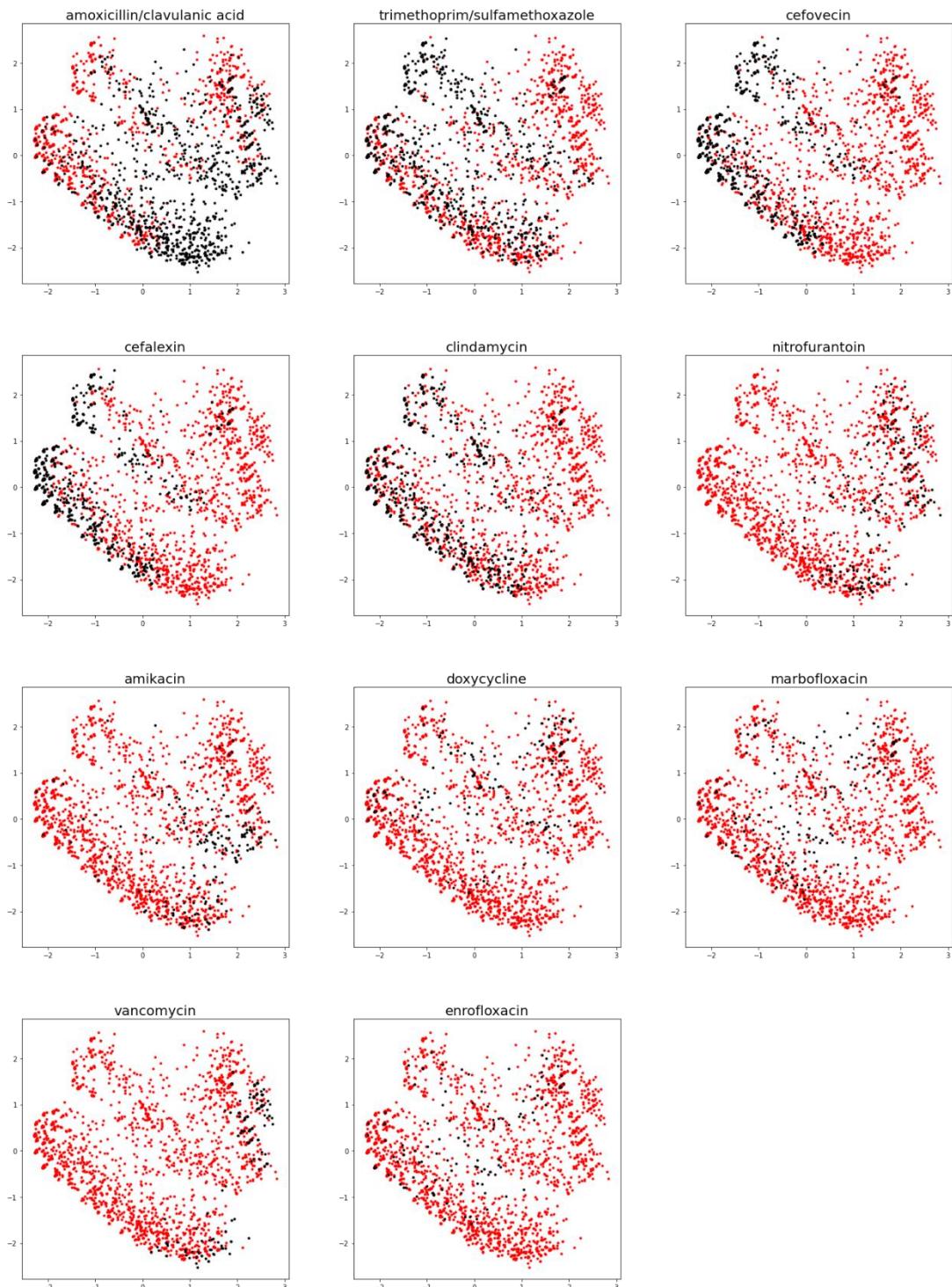
- 1) การนำเทคนิคการเรียนรู้เชิงลึก (Deep Learning) มาใช้ในการฝึกฝนโมเดล เช่น Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) เป็นต้น
- 2) ทดลองใช้วิธีการแก้ไขปัญหาชุดข้อมูลไม่สมดุล (Imbalanced Dataset) ด้วยวิธีอื่น ๆ เช่น Generative Adversarial Networks (GANs) ซึ่งเป็นการสร้างโมเดลเพื่อสร้าง (Generate) ข้อมูลขึ้นมาใหม่โดยใช้เทคนิคการเรียนรู้เชิงลึก
- 3) ในอนาคตการดื้อยาต้านจุลชีพจะเพิ่มขึ้นเรื่อยๆ ซึ่งข้อมูลเก่า ๆ อาจจะไม่เหมาะสมในการนำมาใช้ฝึกฝนโมเดล และในอนาคตอาจมีการเลิกใช้ยาบางชนิดไป ดังนั้นในการอัปเดตโมเดลจึงควรมีการคัดเลือกเอาข้อมูลเก่าออกแล้วนำข้อมูลใหม่มาทำการฝึกฝนโมเดลเพื่อประสิทธิภาพของโมเดลที่ดีขึ้น

## បរចាំនូករម

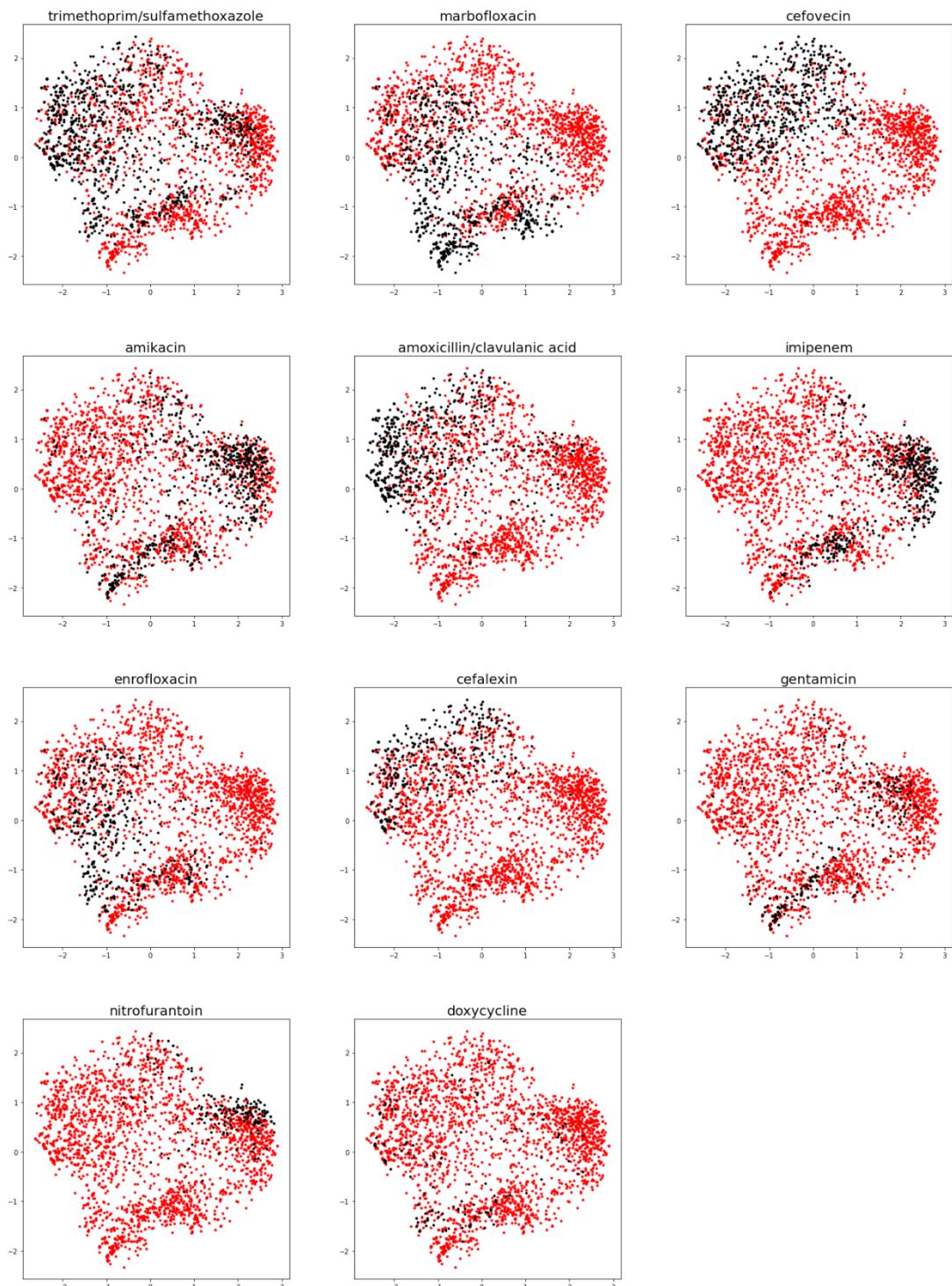
- [1] Feretzakis G, Loupelis E, Sakagianni A, et al. 2020. **Using Machine Learning Techniques to Aid Empirical Antibiotic Therapy Decisions in the Intensive Care Unit of a General Hospital in Greece.** *Antibiotics* 9(2). [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7167935/>
- [2] Martínez-Agüero S, Mora-Jiménez I et al. 2019. **Machine Learning Techniques to Identify Antimicrobial Resistance in the Intensive Care Unit.** *Entropy* 21. [Online]. Available: <https://www.mdpi.com/1099-4300/21/6/603>
- [3] Feretzakis G, Loupelis E, Sakagianni A, et al. 2021. **Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy.** [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/34384203/>
- [4] Farnaz H. Foomani, Shahzad Mirza, Sahjid Mukhida, et al. 2022. **Machine learning techniques to identify antibiotic resistance in patients diagnosed with various skin and soft tissue infections.** [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/2202/2202.13496.pdf>
- [5] Eyad Elyan, Amir Hussain, Aziz Sheikh, wt al. 2022. **Antimicrobial Resistance and Machine Learning: Challenges and Opportunities.** [Online]. Available: <https://ieeexplore.ieee.org/document/9737114>
- [6] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. **SMOTE: Synthetic Minority Over-Sampling Technique.** *Journal of Artificial Intelligence Research*, vol. 16, 321-357. [Online]. Available: <https://arxiv.org/pdf/1106.1813.pdf>

- [7] Han H, Wang W, Mao B. 2005. **Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning.** ICIC 2005: Advances in Intelligent Computing, pp. 878-887. [Online]. Available: [https://sci2s.ugr.es/keel/pdf/specific/congreso/han\\_borderline\\_smote.pdf](https://sci2s.ugr.es/keel/pdf/specific/congreso/han_borderline_smote.pdf)
- [8] Haibo He, Yang Bai, Edwardo A. Garcia, Shutao Li. 2008. **ADASYN: Adaptive synthetic sampling approach for imbalanced learning.** IEEE International Joint Conference on Neural Networks [Online]. Available: <https://ieeexplore.ieee.org/document/4633969>
- [9] Nguyen Hien M, Cooper Eric W, Kamei Katsuari. 2009. **Borderline Over-sampling for Imbalanced Data Classification.** [Online]. Available: [https://ousar.lib.okayama-u.ac.jp/files/public/1/19617/20160528004522391723/IWCIA2009\\_A1005.pdf](https://ousar.lib.okayama-u.ac.jp/files/public/1/19617/20160528004522391723/IWCIA2009_A1005.pdf)
- [10] Baiyun Chena, Shuyin Xiaa, Zizhong Chenb, Binggui Wanga, Guoyin Wanga. 2020. **RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise.** [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0020025520310045>

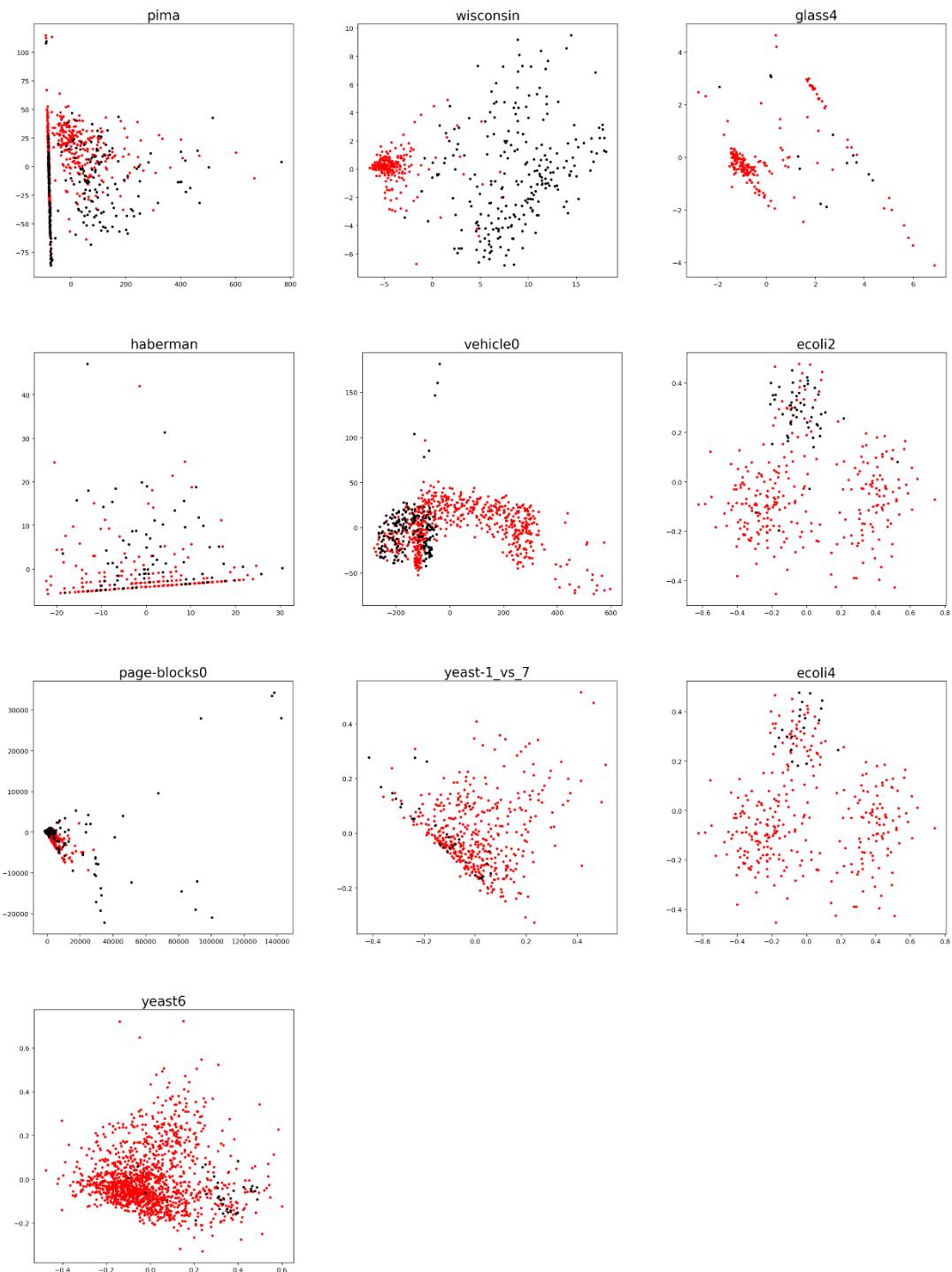
ภาคนวก ก  
แผนภาพการกระจายของชุดข้อมูล



รูปที่ ก.1 แผนภาพการกระจายของชุดข้อมูลแบบที่เรียกว่ารุ่นนวก



รูปที่ ก.2 แผนภูมิการกระจายของชุดข้อมูลแบบที่เรียกว่ากลุ่ม



รูปที่ ก.3 แผนภาพการกระจายของชุดข้อมูลที่ใช้ในการทดลองที่ 1.3