เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกระดับคุณภาพแม่น้ำและการทำนายดัชนีชี้วัดคุณภาพแม่น้ำของประเทศไทย

MACHINE LEARNING TECHNIQUES FOR WATER QUALITY CLASSIFICATION AND WATER QUALITY INDEX FORECASTING OF THAILAND'S RIVERS

คีรีลักษณ์ ศิริคะรินทร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2566

เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกระดับคุณภาพแม่น้ำและการทำนายดัชนีชี้วัดคุณภาพแม่น้ำของประเทศไทย

คีรีลักษณ์ ศิริคะรินทร์

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2566

MACHINE LEARNING TECHNIQUES FOR WATER QUALITY CLASSIFICATION AND WATER QUALITY INDEX FORECASTING OF THAILAND'S RIVERS

KEEREELUK SIRIKARIN

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2023

ปริญญานิพนธ์

เรื่อง

เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกระดับคุณภาพแม่น้ำและการทำนายดัชนีชี้วัดคุณภาพแม่น้ำของประเทศไทย

ของ

คีรีลักษณ์ ศิริคะรินทร์

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าปริญญานิพนธ์

..... ที่ปรึกษาหลัก

(อาจารย์ ดร.ศุภร คนธภักดี)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.รัตน์ชัยนันท์ ธรรมสุจริต)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

ชื่อเรื่อง	เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกระดับคุณภาพแม่น้ำและการทำนายดัชนีชี้วัดคุณภาพแม่น้ำของประเทศไทย
ผู้วิจัย	ศีรีลักษณ์ ศิริคะรินทร์
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2566
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ศุภร คนธภักดี

ดัชนีชี้วัดคุณภาพน้ำ (Water Quality Index: WQI) เป็นค่าที่ใช้บ่งบอกคุณภาพของแม่น้ำของประเทศไทย มีค่าระหว่าง 0-100 และแบ่งเป็น 5 ระดับ ได้แก่ คุณภาพน้ำที่อยู่ในเกณฑ์ดี ดีมาก พอใช้ เสื่อมโทรม และเสื่อมโทรมมาก การใช้เทคนิคการเรียนรู้ของเครื่องเพื่อทำนายคุณภาพของน้ำเป็นวิธีการหนึ่งที่สามารถคาดการณ์คุณภาพของน้ำในอนาคต และนำข้อมูลที่ได้จากการทำนายมาใช้เป็นข้อมูลประกอบการวางแผนจัดการกับคุณภาพน้ำให้เหมาะสมต่อไป ด้วยเหตุนี้ งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อศึกษาการจำแนกระดับคุณภาพน้ำ โดยใช้ Random Forest, Extreme Gradient Boosting (XGBoost), Logistic Regression และ Support Vector Machines (SVM) ร่วมกับเทคนิคการแก้ไขปัญหาความไม่สมดุลของข้อมูลระดับคุณภาพน้ำ ได้แก่ Synthetic Minority Oversampling Technique (SMOTE) และ Random Oversampling นอกจากนี้ ยังศึกษาการทำนายดัชนีชี้วัดคุณภาพน้ำด้วยแบบจำลองอนุกรมเวลา ได้แก่ ARIMA, ARIMAX, SARIMA และ SARIMAX โดยงานวิจัยนี้ใช้ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน จากกรมควบคุมมลพิษ ระหว่างปี พ.ศ 2552 - 2564 ผลการศึกษาพบว่า แบบจำลอง XGBoost ร่วมกับ SMOTE มีประสิทธิภาพสำหรับจำแนกระดับคุณภาพน้ำดีที่สุด ค่า Accuracy เท่ากับ 91.53% Precision เท่ากับ 91.78% Recall เท่ากับ 91.53% และ F1 score เท่ากับ 91.56% และพบว่า BOD หรือปริมาณออกซิเจนที่จุลินทรีย์ใช้ย่อยสลายสารอินทรีย์เป็นพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำมากที่สุด สำหรับผลการศึกษาที่ได้จากการทำนายดัชนีชี้วัดคุณภาพน้ำด้วยแบบจำลองอนุกรมเวลา โดยใช้ข้อมูลของสถานีตรวจวัดคุณภาพน้ำ PI06 (แม่น้ำปิง), WA02 (แม่น้ำวัง), YO01 (แม่น้ำยม) และ NA02 (แม่น้ำน่าน) พบว่า ARIMAX ซึ่งกำหนดให้ ปริมาณออกซิเจนที่ละลายในน้ำ (DO) ปริมาณแบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด (TCB) ปริมาณแบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม (FCB) ปริมาณแอมโมเนียไนโตรเจน (NH_3-N) และ BOD เป็นตัวแปรภายนอก (Exogenous variable) สามารถทำนายค่า WQI ของข้อมูลสถานีตรวจวัด PI06 (MAE เท่ากับ 4.35 RMSE เท่ากับ 5.90 และ MAPE เท่ากับ 6.15%) WA02 (MAE เท่ากับ 6.36 RMSE เท่ากับ 7.55 และ MAPE เท่ากับ 9.35%) และ YO01 (MAE เท่ากับ 5.85 RMSE เท่ากับ 6.62 และ MAPE เท่ากับ 5.85%) มีค่าความคลาดเคลื่อนน้อยที่สุด สามารถสรุปได้ว่า ตัวแปรภายนอก (Exogenous variable) ส่งผลให้การทำนายค่า WQI แม่นยำเพิ่มขึ้น สำหรับข้อมูลคุณภาพน้ำของ 3 สถานีตรวจวัดดังกล่าว

คำสำคัญ : ดัชนีชี้วัดคุณภาพน้ำ, การเรียนรู้ของเครื่อง, การทำนาย, การจำแนกระดับคุณภาพน้ำ

Title	MACHINE LEARNING TECHNIQUES FOR WATER QUALITY CLASSIFICATION AND WATER QUALITY INDEX FORECASTING OF THAILAND'S RIVERS
Author	KEEREELUK SIRIKARIN
Degree	MASTER OF SCIENCE
Academic Year	2023
Thesis Advisor	Dr. Subhorn Khonthapagdee

The Water Quality Index (WQI) is a metric generally used to indicate the water quality of rivers in Thailand. The WQI scores range from 0 to 100 and can be further classified from the scores into five classes, including excellent, good, moderate, poor, and very poor. Applying machine learning techniques to the water quality data is one way to predict water quality information for developing a water quality management plan. Thus, the purpose of this study is to mainly classify water quality using four machine learning techniques: Random Forest, Extreme Gradient Boosting (XGBoost), Logistic Regression, and Support Vector Machines (SVM), together with resampling methods, such as Synthetic Minority Oversampling Technique (SMOTE) and Random Oversampling, to handle the imbalanced dataset. Moreover, time series models, including ARIMA, ARIMAX, SARIMA, and SARIMAX, were performed to forecast the WQI. In this research, the water quality data of Ping River, Wang River, Yom River, and Nan River that were collected by the Pollution Control Department between 2009 and 2021 were chosen. This study found that XGBoost with SMOTE achieved the best performance for classifying water quality with an accuracy of 91.53%, precision of 91.78%, recall of 91.53%, and F1 score of 91.56%. Additionally, Biochemical Oxygen Demand (BOD) was the most important parameter that had the highest impact on water quality classification based on this dataset. Regarding the results of WQI forecasting, the water quality data of PI06 (Ping River), WA02 (Wang River), YO01 (Yom River), and NA02 (Nan River) stations were further selected to study the time series models. The results indicated that ARIMAX (the exogenous variables were Dissolved Oxygen (DO), Total Coliform Bacteria (TCB), Fecal Coliform Bacteria (FCB), Ammonia-nitrogen (NH_3-N), and BOD) was the best model for PI06 (MAE of 4.35, RMSE of 5.90, and MAPE of 6.15%), WA02 (MAE of 6.36, RMSE of 7.55, and MAPE of 9.35%), and YO01 (MAE of 5.85, RMSE of 6.62, and MAPE of 5.85%) due to the least error for forecasting WQI. Lastly, it can be concluded that the exogenous variables improved the model performance of these three stations.

Keyword : Water quality index, Machine learning, Forecasting, Water quality classification

กิตติกรรมประกาศ

ปริญญานิพนธ์นี้สำเร็จลุล่วงได้ด้วยความช่วยเหลือ ความเอาใจใส่ และคำแนะนำที่เป็นประโยชน์ต่อปริญญานิพนธ์จากอาจารย์ ดร.ศุภร คนธภักดี อาจารย์ที่ปรึกษาปริญญานิพนธ์

ขอขอบพระคุณกรมควบคุมมลพิษ ที่ให้ความอนุเคราะห์ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ระหว่างปี พ.ศ. 2552 - 2564 สำหรับใช้ศึกษาวิจัยครั้งนี้

ขอขอบพระคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ สำหรับทุนสนับสนุนการเข้าร่วมประชุมและเสนอผลงานในระดับนานาชาติ

ขอขอบพระคุณคณะกรรมการสอบเค้าโครงปริญญานิพนธ์ และคณะกรรมการสอบปริญญานิพนธ์ สำหรับคำแนะนำและข้อเสนอแนะซึ่งเป็นประโยชน์อย่างยิ่งเพื่อนำไปแก้ไขปริญญานิพนธ์ให้มีความสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณคณาจารย์ประจำสาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่กรุณาถ่ายทอดวิชาความรู้ ให้คำปรึกษา และความช่วยเหลือ เพื่อให้สามารถนำมาประยุกต์ใช้ในการปฏิบัติงานได้

ขอขอบคุณเพื่อนสนิทของผู้วิจัยทุกคนสำหรับกำลังใจ คำปรึกษา และความช่วยเหลือ ทำให้สามารถจัดทำปริญญานิพนธ์ฉบับนี้ได้สำเร็จ และขอขอบคุณเพื่อนๆ ปริญญาโท สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ สำหรับความช่วยเหลือ และให้คำปรึกษากับผู้วิจัยมาโดยตลอด

สุดท้ายนี้ ขอกราบขอบพระคุณบิดา มารดา และครอบครัว ที่คอยสนับสนุน และเป็นกำลังใจให้ผู้วิจัยอย่างดีเสมอมา และขอบคุณตัวผู้วิจัยเองที่มีความอดทน ความพยายาม และความมุ่งมั่นที่จะจัดทำปริญญานิพนธ์ฉบับนี้ให้สำเร็จลุล่วง

คีรีลักษณ์ ศิริคะรินทร์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

ภูมิหลัง

แหล่งน้ำมีบทบาทสำคัญอย่างยิ่งกับประเทศไทย เนื่องจากประชาชนใช้น้ำเพื่อทำเกษตรกรรม อุปโภคบริโภค และอุตสาหกรรม อีกทั้งยังเป็นเส้นทางคมนาคมขนส่ง การเพิ่มขึ้นของประชากร การขยายตัวของที่อยู่อาศัย พื้นที่เกษตรกรรม และพื้นที่อุตสาหกรรมในปัจจุบัน เป็นปัจจัยที่ส่งผลต่อคุณภาพน้ำหากได้รับการจัดการที่ไม่เหมาะสมจะทำให้แหล่งน้ำเน่าเสียและไม่สามารถนำมาใช้ประโยชน์ต่อไปได้ (Taweelarp, Khebchareon, & Saenton, 2021) ปัจจัยที่มักส่งผลต่อคุณภาพแม่น้ำเกิดจากปริมาณออกซิเจนที่ละลายในน้ำมีปริมาณไม่เหมาะสม ปริมาณแอมโมเนียไนโตรเจนและปริมาณแบคทีเรียกลุ่มโคลิฟอร์มมีค่าสูง และปรากฏการณ์ยูโทรฟิเคชัน (Simachaya, 2000) และจากข้อมูลปี พ.ศ. 2563 ของสำนักงานสถิติแห่งชาติ (สสช.) ระบุว่า ดัชนีชี้วัดการจัดการน้ำ (Water Management Index: WMI) ของประเทศไทย มิติที่ 3 ความมั่นคงของน้ำเพื่อการพัฒนา เท่ากับ 2.86 ซึ่งอยู่ระดับปานกลาง เนื่องจากผลิตภาพ (Productivity) การใช้น้ำโดยรวมค่อนข้างต่ำ และมิติที่ 8 การบริหารจัดการทรัพยากรน้ำ อยู่ในระดับพอใช้ (ค่า WMI เท่ากับ 2.65) เนื่องจากแผนงาน งานวิจัยสำหรับสนับสนุนการจัดการน้ำและระบบการติดตามปริมาณและคุณภาพน้ำยังไม่มากพอ (สำนักงานสถิติแห่งชาติ, 2563) ดังนั้นควรหาแนวทางจัดการกับปัญหาคุณภาพน้ำดังกล่าว

การกำหนดดัชนีชี้วัดคุณภาพน้ำ (Water Quality Index: WQI) เป็นวิธีการหนึ่งที่ใช้บ่งชี้ระดับของคุณภาพน้ำว่าอยู่ในเกณฑ์ใด ซึ่งนิยมใช้ในหลายประเทศ เนื่องจากสามารถแสดงสถานะของคุณภาพน้ำได้ในค่าเดียว และง่ายต่อการสื่อสารให้บุคคลทั่วไปเข้าใจ ซึ่งการคำนวณ WQI มี 4 ขั้นตอน ประกอบด้วย (1) เลือกพารามิเตอร์ที่ใช้ชี้วัดคุณภาพน้ำ ซึ่งแต่ละประเทศจะใช้พารามิเตอร์เพื่อคำนวณค่า WQI ที่แตกต่างกันไป ขึ้นอยู่กับปัจจัยต่างๆ เช่น การใช้ประโยชน์จากแหล่งน้ำ ลักษณะภูมิประเทศ และปัจจัยทางสิ่งแวดล้อม เป็นต้น (2) สร้าง Sub-index ของแต่ละพารามิเตอร์ (3) กำหนดน้ำหนักการคำนวณให้แต่ละพารามิเตอร์ (4) สร้างสมการเพื่อคำนวณค่า WQI (Uddin, Nash, & Olbert, 2021; Uddin, Nash, Rahman, & Olbert, 2022) สำหรับประเทศไทย คำนวณค่า WQI จาก 5 พารามิเตอร์น้ำ ได้แก่ ปริมาณออกซิเจนที่ละลายในน้ำ (Dissolved Oxygen: DO) ปริมาณออกซิเจนที่จุลินทรีย์ใช้ย่อยสลายสารอินทรีย์ (Biochemical Oxygen Demand: BOD) ปริมาณแอมโมเนียไนโตรเจน (NH_3-N) ปริมาณแบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม (Fecal Coliform Bacteria: FCB) และปริมาณแบคทีเรีย

กลุ่มโคลิฟอร์มทั้งหมด (Total Coliform Bacteria: TCB) ค่า WQI ที่ได้จากการคำนวณมีค่าอยู่ระหว่าง 0 ถึง 100 แบ่งเป็น 5 เกณฑ์ ได้แก่ คุณภาพน้ำอยู่ในเกณฑ์ดีมาก (91 - 100) ดี (71 - 90) พอใช้ (61 - 70) เสื่อมโทรม (31 - 60) และเสื่อมโทรมมาก (0 - 30) (กรมควบคุมมลพิษ, 2565; Sillberg, Kullavanijaya, & Chavalparit, 2021; Uddin et al., 2021)

จากข้อมูลข้างต้น การคำนวณค่า WQI เพื่อบ่งชี้ระดับคุณภาพน้ำมีหลายขั้นตอน และอาศัยผู้เชี่ยวชาญเป็นผู้เลือกพารามิเตอร์และกำหนดน้ำหนักของพารามิเตอร์น้ำแต่ละตัว เพื่อนำมาใช้คำนวณค่า WQI (House, 1989; Uddin et al., 2021) การใช้การเรียนรู้ของเครื่อง (Machine learning) ซึ่งเป็นการสร้างแบบจำลองทางคณิตศาสตร์สำหรับเรียนรู้ข้อมูลในอดีตเพื่อทำนายข้อมูลในอนาคตนั้น เป็นอีกวิธีการหนึ่งที่สามารถแก้ไขปัญหาดังกล่าวได้ (Malek, Yaacob, Nasir, & Shaadan, 2022) ปัจจุบันมีงานวิจัยหลายฉบับได้ใช้การเรียนรู้ของเครื่องเพื่อทำนายค่า WQI และจำแนกระดับคุณภาพน้ำ ในแหล่งน้ำประเภทต่างๆ ตัวอย่างงานวิจัย เช่น Sillberg et al. (2021) ศึกษาการจำแนกคุณภาพแม่น้ำเจ้าพระยาในประเทศไทยด้วยเทคนิค Support Vector Machine (SVM) และ Attribute-Realization (AR) พบว่า พารามิเตอร์ที่มีความสัมพันธ์กับการจำแนกคุณภาพแม่น้ำเจ้าพระยามากที่สุด 6 พารามิเตอร์ ได้แก่ NH_3-N รองลงมาคือ TCB, FCB, BOD, DO และความเค็ม ตามลำดับ และนำพารามิเตอร์ดังกล่าวมาใช้จำแนกคุณภาพแม่น้ำด้วยเทคนิค SVM ร่วมกับ Linear kernel function พบว่า ได้ค่า Accuracy (0.94) และ F1 score (0.84) มากที่สุด และการศึกษาของ Malek et al. (2022) จำแนกคุณภาพลุ่มแม่น้ำ Kelantan ในประเทศมาเลเซีย ด้วยเทคนิคการเรียนรู้ของเครื่อง 7 เทคนิค ได้แก่ Decision Tree, Artificial Neural Networks (ANN) , K-Nearest Neighbors (K-NN), Naïve Bayes, SVM, Random Forest และ Gradient Boosting ใช้พารามิเตอร์คุณภาพน้ำจำนวน 13 พารามิเตอร์ พบว่า แบบจำลอง Gradient Boosting มีประสิทธิภาพสำหรับจำแนกคุณภาพลุ่มน้ำมากที่สุด (ค่า Accuracy เท่ากับ 94.90% และ F1 score เท่ากับ 86.49%) อย่างไรก็ตาม งานวิจัยดังกล่าวได้ให้ข้อสังเกตว่า ข้อมูลคุณภาพน้ำที่ไม่สมดุล (Imbalanced data) ส่งผลต่อความถูกต้องของการทำนาย นอกจากนี้ Singkran และคณะ (Singkran, Yenpiem, & Sasitorn, 2010) ศึกษาการทำนายค่าดัชนีชี้วัดคุณภาพน้ำ (WQI) ของแม่น้ำในภาคตะวันออกเฉียงเหนือของประเทศไทย จำนวน 5 สาย ด้วยแบบจำลองอนุกรมเวลา (Time series model) ได้แก่ Single moving average, Single exponential smoothing, Seasonal additive, Seasonal multiplicative, Double moving average, Double exponential smoothing, Holt-Winters' additive และ Holt-Winters' multiplicative พบว่า แบบจำลองที่มีประสิทธิภาพการทำนายค่า WQI ดีที่สุดของแต่ละ

สถานีสำหรับแม่น้ำทั้ง 5 สายนั้น ไม่เหมือนกัน Dastorani และคณะ (Dastorani, Mirzavand, Dastorani, & Khosravi, 2020) ใช้แบบจำลองอนุกรมเวลา 5 แบบจำลอง ประกอบด้วย Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA) และ Seasonal Auto-Regressive Integrated Moving Average (SARIMA) ทำนายค่าพารามิเตอร์น้ำ ได้แก่ Ca, Bicarbonate (HCO_3), Sulfate (SO_4), Electrical conductivity (Ec), pH, Mg, Cl, Na และ TDS โดยใช้ข้อมูลคุณภาพน้ำผิวดินบริเวณ Harmaleh ประเทศอิหร่าน ตั้งแต่ ค.ศ. 2001 ถึง 2014 พบว่า ARMA เป็นแบบจำลองที่สามารถทำนายค่าพารามิเตอร์น้ำได้แม่นยำที่สุด 6 พารามิเตอร์ (Cl, Ec, HCO3, Mg, Na และ pH) จาก 9 พารามิเตอร์ และ Fashae และ คณะ (Fashae, Olusola, Ndubuisi, & Udomboso, 2019) ทำนายอัตราการไหลของแม่น้ำ Opeki ประเทศไนจีเรีย โดยใช้ข้อมูลอัตราการไหลของแม่น้ำรายเดือนระหว่างปี ค.ศ. 1980 – 2010 (28 ปี) พบว่า แบบจำลอง ARIMA มีประสิทธิภาพการทำนายอัตราการไหลของแม่น้ำ Opeki ดีกว่า ANN

แม้ว่างานวิจัยที่ผ่านมา ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อทำนายคุณภาพน้ำ แต่การศึกษาการจำแนกระดับคุณภาพน้ำและการทำนายค่า WQI ที่ใช้ชุดข้อมูลคุณภาพแม่น้ำของประเทศไทยยังมีอยู่อย่างจำกัด ดังนั้น งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการจำแนกระดับคุณภาพน้ำและการทำนายดัชนีชี้วัดคุณภาพแม่น้ำของประเทศไทยด้วยเทคนิคการเรียนรู้ของเครื่อง และประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลองที่ศึกษาดังกล่าว ซึ่งใช้ข้อมูลคุณภาพน้ำแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 จากกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ เนื่องจากแม่น้ำทั้ง 4 สาย เป็นต้นกำเนิดของแม่น้ำเจ้าพระยา ซึ่งเป็นหนึ่งในแม่น้ำสายหลักของประเทศไทย ประชาชนใช้ประโยชน์เพื่อการเกษตรกรการประมง การคมนาคมทางน้ำ และเป็นแหล่งท่องเที่ยวทางน้ำ เป็นต้น (อดิศร์ อิศรางกูร ณ อยุธยา, 2560)

ความมุ่งหมายของงานวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ ดังนี้

1. เพื่อศึกษาการจำแนกระดับคุณภาพน้ำของประเทศไทยด้วยเทคนิคการเรียนรู้ของเครื่อง
2. เพื่อศึกษาการทำนายดัชนีชี้วัดคุณภาพน้ำโดยใช้แบบจำลองอนุกรมเวลา
3. เพื่อประเมินประสิทธิภาพของแบบจำลองที่ใช้จำแนกระดับคุณภาพของแม่น้ำและทำนายดัชนีชี้วัดคุณภาพน้ำ

ความสำคัญของการวิจัย

งานวิจัยนี้ศึกษาเทคนิคการเรียนรู้ของเครื่อง (Machine learning technique) เพื่อจำแนกระดับคุณภาพน้ำและทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ซึ่งเป็นแม่น้ำที่อยู่ทางภาคเหนือของประเทศไทย สำหรับการจำแนกระดับคุณภาพน้ำใช้แบบจำลองประเภทการเรียนรู้โดยมีผู้สอน (Supervised learning) ได้แก่ เทคนิค Random Forest เทคนิค Extreme Gradient Boosting (XGBoost) เทคนิค Logistic Regression และเทคนิค Support Vector Machine (SVM) ร่วมกับการเพิ่มปริมาณข้อมูลเพื่อแก้ไขปัญหาข้อมูลไม่สมดุล (Imbalanced data) ได้แก่ เทคนิค Synthetic Minority Oversampling Technique (SMOTE) และ Random Oversampling และงานวิจัยนี้ทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ด้วยแบบจำลองอนุกรมเวลา (Time series model) 4 แบบจำลอง ได้แก่ Auto-Regressive Integrated Moving Average (ARIMA), Auto-Regressive Integrated Moving Average with Exogenous variables (ARIMAX), Seasonal Auto-Regressive Integrated Moving Average (SARIMA) และ Seasonal Auto-Regressive Integrated Moving Average with Exogenous variables (SARIMAX) ข้อมูลแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ที่ใช้สำหรับงานวิจัยนี้ ได้รับการอนุเคราะห์จากกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ซึ่งเป็นข้อมูลคุณภาพน้ำระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 จำนวนทั้งสิ้น 2,736 แถว ประกอบด้วยดัชนีชี้วัดคุณภาพน้ำ (WQI) มีค่าระหว่าง 0 ถึง 100 สามารถแบ่งเป็นเกณฑ์ระดับคุณภาพน้ำ 5 ระดับ ได้แก่ คุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก (91 - 100) ดี (71 - 90) พอใช้ (61 - 70) เสื่อมโทรม (31 - 60) และเสื่อมโทรมมาก (0-30)

ขอบเขตของการวิจัย

ประชากรที่ใช้ในการวิจัย

ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ของกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 จำนวนทั้งสิ้น 2,736 ตัวอย่าง โดยค่าคุณภาพแม่น้ำมาจากสถานีตรวจวัดคุณภาพน้ำจำนวนทั้งสิ้น 64 สถานี แบ่งเป็น

1. แม่น้ำปิง 16 สถานี ครอบคลุมพื้นที่จังหวัดนครสวรรค์ 2 สถานี (PI01 และ PI02) กำแพงเพชร 4 สถานี (PI03, PI04, PI05 และ PI05.6) ตาก 4 สถานี (PI06, PI07, PI08 และ PI09) และเชียงใหม่ 6 สถานี (PI10, PI11, PI12, PI13, PI14 และ PI11.1)

2. แม่น้ำวัง 15 สถานี ครอบคลุมพื้นที่จังหวัดลำปาง 14 สถานี (WA1.1, WA02, WA03, WA3.4, WA3.5, WA04, WA4.1, WA4.3, WA5.1, WA5.2, WA5.3, WA5.4, WA06 และ WA07) และตาก 1 สถานี (WA01)

3. แม่น้ำยม 16 สถานี ครอบคลุมพื้นที่จังหวัดนครสวรรค์ 1 สถานี (YO0.5) พิจิตร 3 สถานี (YO01, YO02 และ YO03) พิษณุโลก 1 สถานี (YO04) สุโขทัย 4 สถานี (YO05, YO06, YO07 และ YO08) แพร่ 5 สถานี (YO09, YO10, YO11, YO12, YO12.1) และพะเยา 2 สถานี (YO13 และ YO14)

4. แม่น้ำน่าน 17 สถานี ครอบคลุมพื้นที่จังหวัดนครสวรรค์ 3 สถานี (NA0.1, NA01 และ NA1.1) พิจิตร 5 สถานี (NA02, NA03, NA04 และ NA05) พิษณุโลก 4 สถานี (NA06, NA07 และNA08) อุตรดิตถ์ 3 สถานี (NA09 NA10 และ NA11) และน่าน 4 สถานี (NA12, NA13, NA14 และ NA15)

กลุ่มตัวอย่างที่ใช้ในการวิจัย

ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ของกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 จำนวนทั้งสิ้น 2,736 ตัวอย่าง ซึ่งแต่ละตัวอย่างประกอบด้วย ข้อมูลสถานีตรวจวัด จังหวัด วันที่ตรวจวัด แม่น้ำ ค่า WQI ค่า DO ค่า BOD ค่า TCB ค่า FCB ค่า NH_3-N และระดับคุณภาพน้ำ

ตัวแปรที่ศึกษา

1. ตัวแปรอิสระ แบ่งเป็นดังนี้

1.1 ค่า DO (Dissolved Oxygen) คือ ปริมาณออกซิเจนที่ละลายน้ำ มีหน่วยเป็นมิลลิกรัมต่อลิตร

1.2 ค่า BOD (Biochemical Oxygen Demand) คือ ปริมาณออกซิเจนที่จุลินทรีย์ใช้ย่อยสลายสารอินทรีย์ มีหน่วยเป็นมิลลิกรัมต่อลิตร

1.3 ค่า TCB (Total Coliform Bacteria) คือ ปริมาณแบคทีเรียกลุ่มโคลิฟอร์มทั้งหมด มีหน่วยเป็น MPN ต่อ 100 มิลลิลิตร

1.4 ค่า FCB (Fecal Coliform Bacteria) คือ ปริมาณแบคทีเรียกลุ่มฟีคอลโคลิ-ฟอร์ม มีหน่วยเป็น MPN ต่อ100 มิลลิลิตร

1.5 ค่า NH_3-N คือ ปริมาณแอมโมเนียไนโตรเจน มีหน่วยเป็นมิลลิกรัมต่อลิตร

1.6 วันที่ตรวจวัด

2. ตัวแปรตาม ได้แก่ ค่าดัชนีชี้วัดคุณภาพน้ำ (WQI) และระดับคุณภาพน้ำ

กรอบแนวคิดในงานวิจัย

งานวิจัยนี้ศึกษาการทำนายระดับคุณภาพของแม่น้ำและการทำนายค่าดัชนีชี้วัดคุณภาพน้ำ (WQI) ซึ่งค่า WQI มีค่าระหว่าง 0 ถึง 100 และเกณฑ์คุณภาพน้ำแบ่งเป็น 5 ระดับ ได้แก่ คุณภาพน้ำ

ที่อยู่ในเกณฑ์ดีมาก คุณภาพน้ำที่อยู่ในเกณฑ์ดี คุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ คุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม และคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก โดยใช้ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ตั้งแต่ปี พ.ศ. 2552 ถึง พ.ศ. 2564 จากกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ จำนวน 2,736 ตัวอย่าง การทำนายระดับคุณภาพน้ำ ใช้ค่า DO, BOD, TCB, FCB และ NH_3-N เป็นข้อมูลลักษณะเฉพาะ (Feature) สำหรับข้อมูลนำเข้า (Input data) ให้กับแบบจำลองที่ใช้อัลกอริทึม ประกอบด้วย Random Forest, XGBoost, Logistic Regression และ SVM ร่วมกับเทคนิคการแก้ไขปัญหาความไม่สมดุลของข้อมูล (Imbalanced data) ได้แก่ SMOTE และ Random Oversampling จากนั้นประเมินประสิทธิภาพของแบบจำลองที่ใช้จำแนกระดับคุณภาพน้ำและประเมินประสิทธิภาพของแบบจำลองจากค่า Accuracy, Precision, Recall และ F1 score และสำหรับการทำนายค่า WQI ใช้แบบจำลองอนุกรมเวลา ได้แก่ ARIMA, ARIMAX, SARIMA และ SARIMAX โดยแบบจำลอง ARIMA และ SARIMA ใช้ข้อมูลวันที่ตรวจวัดและค่า WQI ช่วงเวลาก่อนหน้าเป็นข้อมูลนำเข้าแบบจำลอง แต่แบบจำลอง ARIMAX และ SARIMAX นอกจากข้อมูลวันที่ตรวจวัดและค่า WQI แล้ว ยังมีตัวแปรภายนอก (Exogenous variable) ได้แก่ ค่า DO, BOD, TCB, FCB และ NH_3-N เพื่อทำนายค่า WQI ในช่วงเวลาถัดไป และประเมินประสิทธิภาพของแบบจำลองด้วยค่า Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) และ Mean Absolute Percentage Error (MAPE)

สมมุติฐานในการวิจัย

1. เทคนิคการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) ได้แก่ Random Forest และ XGBoost เป็นแบบจำลองมีประสิทธิภาพสูงสำหรับทำนายระดับคุณภาพน้ำ

2. การใช้เทคนิคการเพิ่มปริมาณข้อมูล ได้แก่ SMOTE และ Random Oversampling สามารถแก้ไขปัญหาการไม่สมดุลของจำนวนข้อมูลได้

3. การใช้เทคนิคการสุ่มเลือก Hyper-parameter ของแบบจำลอง (Randomized search) เพื่อหาปรับหาค่าพารามิเตอร์ที่เหมาะสมกับแบบจำลอง สามารถเพิ่มประสิทธิภาพการจำแนกระดับคุณภาพน้ำได้

4. การเพิ่มตัวแปรภายนอก (Exogenous variable) ของแบบจำลองอนุกรมเวลา ARIMAX และ SARIMAX สามารถลดความคลาดเคลื่อนจากการทำนายค่า WQI ของแม่น้ำได้

บทที่ 2
ทบทวนวรรณกรรม

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

1. ดัชนีชี้วัดคุณภาพแหล่งน้ำ
2. ทฤษฎีเกี่ยวกับการเรียนรู้ของเครื่อง (Machine learning)
3. การประเมินประสิทธิภาพของแบบจำลอง
4. งานวิจัยที่เกี่ยวข้อง

ดัชนีชี้วัดคุณภาพแหล่งน้ำ

ดัชนีชี้วัดคุณภาพน้ำ (Water Quality Index) หรือ WQI เป็นเครื่องมือที่ใช้ประเมินภาพรวมคุณภาพแหล่งน้ำ มีค่าอยู่ระหว่าง 0 ถึง 100 หากมีค่า WQI มาก แสดงว่าแหล่งน้ำมีคุณภาพที่ดี ค่า WQI ได้มาจากพารามิเตอร์ที่บ่งชี้คุณภาพน้ำทั้งทางชีวภาพ กายภาพ และเคมี มาสร้างสมการทางคณิตศาสตร์ โดยการเลือกพารามิเตอร์เพื่อใช้คำนวณ WQI จะขึ้นกับประเภทของแหล่งน้ำ ลักษณะภูมิประเทศ และปัจจัยทางสิ่งแวดล้อม ด้วยเหตุผลดังกล่าว ทำให้ WQI นำมาใช้เพื่อบ่งบอกคุณภาพน้ำในหลายๆ ประเทศ รวมทั้งประเทศไทย เนื่องจากง่ายต่อการเข้าใจ โดยไม่ต้องอาศัยความรู้เกี่ยวกับพารามิเตอร์สิ่งแวดล้อมทางน้ำมากนัก (Banda & Kumarasamy, 2020; Poonam, Tanushree, & Sukalyan, 2013; Uddin et al., 2021)

ในปี ค.ศ. 1965 Horton (Horton, 1965) ได้พัฒนา WQI ขึ้นครั้งแรก ประกอบด้วยพารามิเตอร์ที่มีคามสำคัญกับคุณภาพน้ำจำนวน 10 พารามิเตอร์ เพื่อกำหนดน้ำหนักของพารามิเตอร์แต่ละตัวและใช้คำนวณค่า WQI ต่อมาในปี ค.ศ. 1970 Brown และคณะ (Brown, McClelland, Deininger, & Tozer, 1970) ได้พัฒนาต่อยอด NSF- WQI มาจาก Horton โดยใช้ความเห็นจากผู้เชี่ยวชาญจำนวน 142 คน เพื่อเลือกพารามิเตอร์จำนวน 9 พารามิเตอร์ สำหรับคำนวณค่า WQI ซึ่งการคำนวณค่า WQI ในหลายๆ ประเทศได้ดัดแปลงมาจากงานวิจัยของ Brown และคณะ เช่นเดียวกับการคำนวณค่า WQI ของประเทศไทย (Prakirake, Chaiprasert, & Tripetchkul, 2009; Uddin et al., 2021; Walsh & Wheeler, 2013) การคำนวณ WQI โดยทั่วไปประกอบด้วย 4 ขั้นตอน ดังนี้ (1) เลือกพารามิเตอร์ที่ใช้วัดชี้วัดคุณภาพน้ำ (2) สร้าง Sub-index ของแต่ละพารามิเตอร์ เพื่อให้แต่ละพารามิเตอร์อยู่ในหน่วยเดียวกัน (3) กำหนดน้ำหนักของแต่ละ

พารามิเตอร์เพื่อใช้คำนวณ (4) สร้างสมการคำนวณค่า WQI (กรมควบคุมมลพิษ, 2565; Prakirake et al., 2009)

สำหรับประเทศไทย กรมควบคุมมลพิษได้เริ่มนำค่า WQI มาใช้เพื่อบ่งชี้คุณภาพน้ำผิวดิน ในปี พ.ศ. 2538 (Prakirake et al., 2009) ใช้การคำนวณดัชนีคุณภาพน้ำจำนวน 8 พารามิเตอร์ ประกอบด้วย ความเป็นกรดด่าง (pH) ออกซิเจนละลายน้ำ (DO) ของแข็งทั้งหมด (TS) ปริมาณแบคทีเรียกลุ่มฟีคอลโคลิฟอร์ม (FCB) ไนเตรท (NO_3) ฟอสฟอรัสทั้งหมด (TP) ของแข็งแขวนลอย (SS) และปริมาณออกซิเจนที่จุลินทรีย์ที่ใช้ย่อยสลายสารอินทรีย์ (BOD) ต่อมาได้ปรับปรุงการคำนวณค่า WQI โดยใช้เพียง 5 พารามิเตอร์ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N (กรมควบคุมมลพิษ, 2554)

กรมควบคุมมลพิษ ได้กำหนดสมการสำหรับคิดคะแนนเทียบกับค่าพารามิเตอร์ ของแต่ละพารามิเตอร์ ดังตาราง 1 สำหรับการคำนวณค่า WQI ซึ่งเป็นคะแนนรวมของทั้ง 5 พารามิเตอร์ สามารถคำนวณได้จากสมการ (1) (กรมชลประทาน, 2561)

ค่า WQI = ค่าเฉลี่ยของคะแนน 5 พารามิเตอร์ – ค่าคะแนนพิเศษ ..... (1)

ค่าคะแนนพิเศษ ได้จากการเปรียบเทียบระหว่างเกณฑ์คุณภาพน้ำที่ต่ำที่สุดเทียบกับค่าเฉลี่ยของคะแนนทั้ง 5 พารามิเตอร์ เพื่อปรับให้สอดคล้องกับประเภทแหล่งน้ำผิวดิน โดยคะแนนพิเศษ เป็นดังนี้

- เกณฑ์คุณภาพน้ำไม่ต่างกัน คะแนนพิเศษ = 0
- เกณฑ์คุณภาพน้ำต่างกัน 1 ระดับ คะแนนพิเศษ = 10
- เกณฑ์คุณภาพน้ำต่างกัน 2 ระดับ คะแนนพิเศษ = 15
- เกณฑ์คุณภาพน้ำต่างกัน 3 ระดับ คะแนนพิเศษ = 20

ค่า WQI ที่ได้จากการคำนวณมีค่าระหว่าง 0 ถึง 100 ซึ่งแบ่งเป็น 5 ระดับ ดังตาราง 2

ตาราง 1 สมการสำหรับคิดคะแนนของแต่ละพารามิเตอร์น้ำ

พารามิเตอร์น้ำ	ความเข้มข้น	สมการ
1. DO (mg/l)	0.0 – 4.0	คะแนน = 15.25*(ค่า DO) + 0.1667
	4.1 – 6.0	คะแนน = 5*(ค่า DO) + 41
	6.1 – 8.4	คะแนน =12.083*(ค่า DO) - 1.5
	8.5 – 8.9	คะแนน = -78*(ค่า DO) + 755.2
	9.0 – 11.2	คะแนน = -13.043*(ค่า DO) + 177.09
	11.3 – ≥15.3	คะแนน = -7.561*(ค่า DO) + 115.68
2. BOD (mg/l)	0.0 – 1.5	คะแนน = -19.333(ค่า BOD) + 100
	1.6 – 2.0	คะแนน = -20(ค่า BOD) + 101
	2.1 – 4.0	คะแนน = -15(ค่า BOD) + 91
	4.1 – ≥8.8	คะแนน = -6.4583*(ค่า BOD) + 56.833
3. TCB (MPN/100ml)	0.0 – 5,000	คะแนน = -0.0058*(ค่า TCB) + 100
	5,001 – 20,000	คะแนน = -0.0007*(ค่า TCB) + 74.333
	20,001 – 160,000	คะแนน = -0.0002*(ค่า TCB) + 65.286
	>160,000	คะแนน = -0.000008-06*(ค่า TCB) + 32.292
4. FCB (MPN/100ml)	0.0 – 1,000	คะแนน = -0.029*(ค่า FCB) + 100
	1,001 – 4,000	คะแนน = -0.0033*(ค่า FCB) + 74.333
	4,001 – 90,000	คะแนน = -0.0003*(ค่า FCB) + 62.395
	>90,000	คะแนน = -0.00001-05*(ค่า FCB) + 32.208
5. NH_3 (mg/l)	0.0 – 0.22	คะแนน = -131.82*(ค่า NH_3) + 100
	0.23 – 0.50	คะแนน = -35.714*(ค่า NH_3) + 78.857
	0.51 – 1.83	คะแนน = -22.556*(ค่า NH_3) + 72.278
	>1.83	คะแนน = -6.1024*(ค่า NH_3) + 42.167

ที่มา: กรมชลประทาน. (2561). *รายงานความคิดเห็นของโครงการฯ เกี่ยวกับวิธีการตรวจวัดคุณภาพน้ำ*. สืบค้นจาก http://qwater.rid.go.th/report/file61/exam61/PDF/EQUIPPROB.pdf

ตาราง 2 เกณฑ์คุณภาพน้ำผิวดิน

ค่า WQI	เกณฑ์คุณภาพน้ำ	มาตรฐานคุณภาพน้ำผิวดิน
91 - 100	ดีมาก	2
71 - 90	ดี	2
61 - 70	พอใช้	3
31 - 60	เสื่อมโทรม	4
0 - 30	เสื่อมโทรมมาก	5

ที่มา: กรมชลประทาน. (2561). *รายงานความคิดเห็นของโครงการฯ เกี่ยวกับวิธีการตรวจวัดคุณภาพน้ำ*. สืบค้นจาก http://qwater.rid.go.th/report/file61/exam61/PDF/EQUIPPROB.pdf

ทฤษฎีเกี่ยวกับการเรียนรู้ของเครื่อง (Machine learning)

1. เทคนิค Ensemble Learning

Ensemble เป็นเทคนิคการรวมหลายแบบจำลองเข้าด้วยกัน เพื่อสร้างเป็นแบบจำลองที่ให้ประสิทธิภาพสำหรับการทำนายที่ดีมากขึ้น โดย Ensemble สามารถเรียนรู้แบบจำลองหลายๆ แบบจำลองพร้อมกัน (Parallel) หรือเรียนรู้แบบมีลำดับ (Sequential) กล่าวคือ เรียนรู้จากความผิดพลาดก่อนหน้าแล้วนำมาปรับปรุงเพื่อลดความผิดพลาดในครั้งถัดไป ซึ่งผลลัพธ์ที่ได้จากการทำนายด้วยวิธี Ensemble จะได้มาจากการโหวต ได้แก่ โหวตจากเสียงข้างมาก (Majority vote) และค่าเฉลี่ยของผลลัพธ์ (Bonaccorso, 2017, pp. 154-180)

วิธีการ Ensemble สามารถ แบ่งได้เป็น 3 ประเภทหลัก ดังนี้

1) Bagging (Bootstrap Aggregation) คือการสร้างหลายๆ แบบจำลอง โดยแต่ละแบบจำลองจะใช้ข้อมูลกลุ่มย่อย (Subset) ที่ได้จากการสุ่มข้อมูลแบบใส่คืนของชุดเรียนรู้ (Training data) จากนั้น ผลลัพธ์ที่ได้ของแต่ละแบบจำลองจะนำมาหาค่าฐานนิยม (Mode) เพื่อให้ได้ผลลัพธ์สุดท้าย ทั้งนี้ เทคนิค Bagging จะใช้อัลกอริทึมชนิดเดียวกันกับทุกๆ แบบจำลองสำหรับการเรียนรู้ และแต่ละแบบจำลองสามารถเรียนรู้ไปพร้อมกันได้ เนื่องจากทุกแบบจำลองเป็นอิสระต่อกัน อย่างไรก็ตาม นอกจากแต่ละแบบจำลองจะใช้ข้อมูลที่ได้จากการสุ่มข้อมูลตัวอย่างจากชุดเรียนรู้แล้ว ยังสามารถสุ่มเลือกกลุ่มย่อยของลักษณะเฉพาะ (Feature) ได้เช่นกัน ซึ่งเรียกวิธีการนี้ว่า Random Patches หากใช้ข้อมูลของชุดเรียนรู้ทั้งหมดกับทุกแบบจำลอง

แต่สุ่มเลือกกลุ่มย่อยของลักษณะเฉพาะ (Feature) จะเรียกว่า วิธี Random Subspaces สำหรับตัวอย่างอัลกอริทึมที่ใช้เทคนิค Bagging เช่น Random Forest (Bonaccorso, 2017, pp. 154-180; Geron, 2019, pp. 191-214)

2) Boosting เป็นเทคนิค Ensemble ที่ใช้อัลกอริทึมชนิดเดียว และมีลำดับการเรียนรู้จากข้อมูลของแบบจำลองก่อนหน้าที่ทำนายผิดพลาด โดยจะเพิ่มน้ำหนัก (Weight) ความน่าจะเป็นของข้อมูลที่ทำนายผิด และปรับน้ำหนักของข้อมูลที่ทำนายถูกแล้วให้น้อยลง ก่อนเข้าแบบจำลองถัดไป ซึ่งเป็นการทำงานร่วมกันของแบบจำลองที่ให้ประสิทธิภาพการทำนายต่ำ (Weak leaners) ผลการทำนายท้ายสุดที่ได้จากการโหวตเสียงข้างมากแบบถ่วงน้ำหนัก (Weighted majority vote) สำหรับตัวอย่างอัลกอริทึม ได้แก่ AdaBoost, Gradient Boosting และ XGBoost เป็นต้น (Bonaccorso, 2017, pp. 154-180; Geron, 2019, pp. 191-214)

3) Stacking เป็นวิธีที่ใช้อัลกอริทึมที่ต่างกันกับข้อมูลเรียนรู้ชุดเดียวกัน โดยแต่ละแบบจำลองที่ใช้อัลกอริทึมต่างกันจะเรียนรู้โดยเป็นอิสระต่อกัน ผลลัพธ์ที่ได้จากการทำนายมาจากการโหวตเสียงข้างมาก ค่าเฉลี่ย หรือใช้แบบจำลองอื่นมาทำนายผลลัพธ์สุดท้าย (Bonaccorso, 2017, pp. 154-180)

1.1 Random Forest

เป็นอัลกอริทึมที่ใช้เทคนิค Ensemble ประเภท Bagging ของ Decision tree โดย Decision tree แต่ละต้นเป็นอิสระต่อกัน สุ่มใช้กลุ่มข้อมูลย่อยต่างกัน และใช้กลุ่มย่อยของลักษณะเฉพาะ (Feature) ที่ต่างกันด้วย โดยทั่วไปจำนวนของลักษณะเฉพาะที่ใช้กับข้อมูลของ Decision tree หนึ่งต้น ได้มาจากการหาค่ารากที่สอง (Square Root) หรือค่า log ของจำนวนลักษณะเฉพาะทั้งหมด (Bonaccorso, 2017, pp. 154-180) สำหรับการเลือกลักษณะเฉพาะเพื่อแบ่งแต่ละโหนด (Node) ของ Decision tree ได้มาจากการคำนวณค่า Entropy ซึ่งจะเป็นค่าที่บ่งบอกความเหมือนกันของข้อมูลย่อย ถ้าค่า Entropy เท่ากับ 0 แสดงว่าข้อมูลย่อยดังกล่าวมีความเหมือนกัน โดยค่า Entropy คำนวนได้ดังสมการ (2) (Kirasich, Smith, & Sadler, 2018) เมื่อได้ผลการทำนายของ Decision tree แต่ละต้นแล้ว จะนำผลลัพธ์มารวมกันโดยใช้เสียงข้างมาก เพื่อได้เป็นผลการทำนายสุดท้าย (Bonaccorso, 2017, pp. 154-180)

$$Entropy = -p\log_2(p) - q\log_2(q) \qquad \text{..... (2)}$$

Random Forest สามารถใช้กับข้อมูลที่มีลักษณะไม่เชิงเส้น (Non-linear pattern) ได้ และสามารถใช้ได้กับข้อมูลประเภทที่จัดเป็นกลุ่ม (Categorical data) และข้อมูล

ที่เป็นตัวเลข (Numerical data) โดยไม่จำเป็นต้องปรับข้อมูลให้อยู่ในช่วง (Scale) เดียวกันก่อนเข้าอัลกอริทึม นอกจากนี้ Random Forest สามารถแก้ไขปัญหา Overfitting ของ Decision tree ได้ (Overfitting คือแบบจำลองถูกฝึกฝนให้เข้ากับข้อมูลชุดเรียนรู้ (Training data) มากเกินไป เมื่อนำแบบจำลองไปทดสอบกับชุดทดสอบ (Test data) จะเกิดความผิดพลาดสูง) (Archana, Savita, & Raj, 2016) เนื่องจาก Decision tree จะก่อให้เกิดความแปรปรวน (Variance) มาก Random Forest ช่วยลดความแปรปรวนได้ โดยรวมการทำงานของ Decision tree หลายต้นอาจนำมาซึ่งความเอนเอียง (Bias) ที่เพิ่มขึ้น

1.2 XGBoost

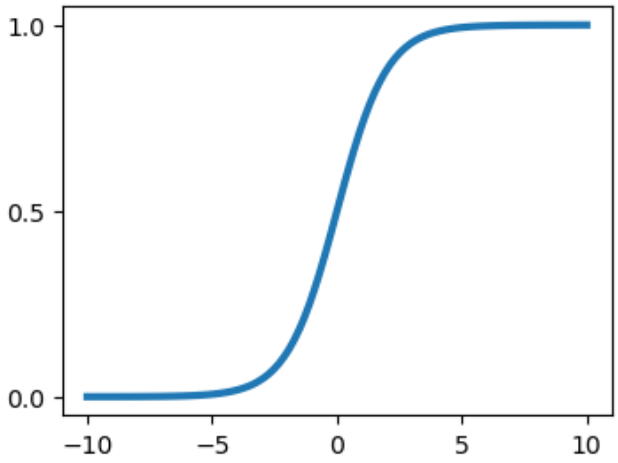
XGBoost หรือ Extreme Gradient Boosting เป็นหนึ่งในอัลกอริทึมของ Gradient Boosting โดย XGBoost สามารถทำงานได้อย่างรวดเร็ว เนื่องจากสามารถสร้างโหนดของ Decision tree ไปพร้อมๆ กันได้ และเก็บข้อมูลไว้ใน Block ซึ่งเป็นหน่วยความจำ (In-memory units) ข้อมูลแต่ละ Block จะถูกเก็บไว้ในรูปแบบ Compressed column (CSC) แต่ละคอลัมน์ใน Block จะเรียงลำดับข้อมูลตามค่าของลักษณะเฉพาะที่เหมือนกัน (Corresponding feature value) นอกจากนี้ XGBoost สามารถใช้กับข้อมูลที่มีค่าว่างอยู่ในข้อมูลได้เช่นกัน (Chen & Guestrin, 2016)

Gradient Boosting เป็นการเพิ่มจำนวน Decision tree โดยมีลำดับแบบจำลองถัดไปเรียนรู้จากข้อมูลที่ทำนายผิดพลาดของแบบจำลองก่อนหน้า เพื่อทำให้ Cost function หรือ Loss function มีค่าน้อยที่สุด Learning rate เป็นพารามิเตอร์ที่ควรพิจารณาสำหรับ Decision tree แต่ละต้น เนื่องจากค่า Learning rate ต่ำ จำเป็นต้องใช้จำนวน Decision tree มากขึ้น แต่หากจำนวน Decision tree มากเกินไป อาจทำให้เกิด Overfitting ได้ (Bonaccorso, 2017, pp. 154-180)

2. เทคนิค Logistic regression

Logistic regression เป็นเทคนิคการสร้างแบบจำลองเชิงเส้นสำหรับใช้จำแนกประเภท โดยอาศัย Logistic function หรือ Sigmoid function ดังสมการ (3) และผลลัพธ์ที่ได้จาก Logistic function เป็นค่าความน่าจะเป็นที่มีค่าระหว่าง 0 ถึง 1 ซึ่งกราฟของ Logistic function จะมีลักษณะเป็น S-curve (ภาพประกอบ 1) (Amazon Web Services, 2022; Scikit-learn developers, 2022a)

$$f(x) = \frac{1}{1+e^{-(X_i w + w_0)}}$$ โดย w คือ ค่าสัมประสิทธิ์ ..... (3)



ภาพประกอบ 1 กราฟ Logistic function

Logistic regression จะทำ Regularization โดยเพิ่มตัวแปร Penalty $(r(w))$ เพื่อลดความซับซ้อนของแบบจำลอง ทำให้แบบจำลองนำไปใช้กับข้อมูลชุดทดสอบแล้วทำนายได้แม่นยำมากขึ้น โดยพยายามทำให้ค่าที่ได้จาก Cost function มีค่าน้อยที่สุด ดังสมการ (4)

$$\min_{w} C \sum_{i=1}^{n}(-y_i log(\hat{p}(X_i)) - (1-y_i)log(1-(\hat{p}(X_i))) + r(w) \quad \text{..... (4)}$$

ประเภทของ Logistic regression มี 3 ประเภท (Amazon Web Services, 2022) ได้แก่

1) Binary logistic regression ใช้จำแนกประเภทเพียง 2 ประเภท เช่น กลุ่ม 0 และกลุ่ม 1 หากค่าความน่าจะเป็นที่ได้มีค่าน้อยกว่า 0.5 จะจำแนกเป็นกลุ่ม 0 (กลุ่มลบ) แต่ถ้าค่าความน่าจะเป็นที่ได้มีค่ามากกว่า 0.5 จะจำแนกเป็นกลุ่ม 1 (กลุ่มบวก)

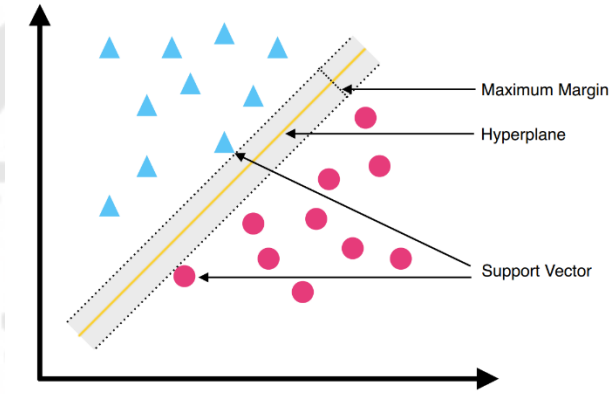
2) Multinomial logistic regression ใช้จำแนกกลุ่มที่มีจำนวนมากกว่า 2 กลุ่ม โดยค่าความน่าจะเป็นที่ได้มีค่าอยู่ระหว่าง 0 ถึง 1 ถ้าค่าความน่าจะเป็นของกลุ่มใดมีค่ามากที่สุด ผลลัพธ์ที่ได้จากการทำนายจะเป็นกลุ่มนั้น

3) Ordinal logistic regression ใช้จำแนกกลุ่มที่มีจำนวนมากกว่า 2 กลุ่ม เหมาะกับปัญหาที่ใช้ตัวเลขที่ได้จะแสดงลำดับ เช่น ดีมาก พอใช้ และควรปรับปรุง

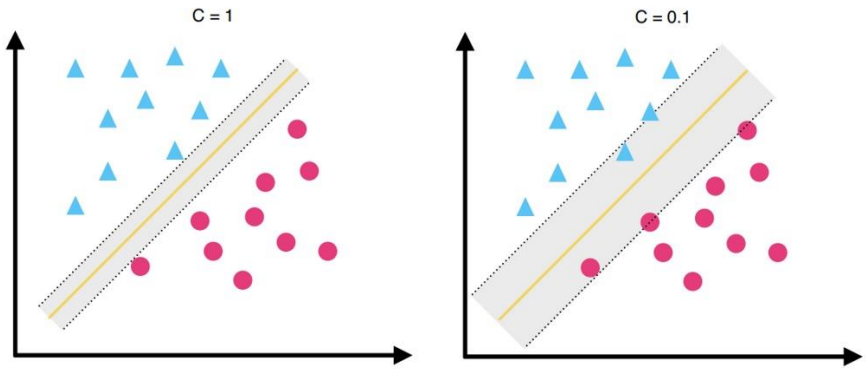
3. เทคนิค Support Vector Machine

Support Vector Machine หรือ SVM อาศัยหลักการสร้างเส้นแบ่งหรือ Hyperplane เพื่อแบ่งแยกกลุ่มของข้อมูลออกจากกัน จากนั้นหาว่าเส้นแบ่งหรือ Hyperplane ใด ที่สามารถแบ่งแยกกลุ่มของข้อมูลได้ดีที่สุด (Optimal hyperplane) โดยเลือก Hyperplane ที่ให้ Maximum

margin คือผลรวมของระยะระหว่าง Hyperplane กับเส้นขอบที่ผ่านข้อมูลที่ใกล้ที่สุด (Margin) ของแต่ละฝั่งมากที่สุด โดยข้อมูลที่อยู่บน Margin จะเรียกว่า Support vector (ภาพประกอบ 2) การกำหนดค่า Support vector ขึ้นอยู่กับพารามิเตอร์ C (Penalty parameter หรือ Regularization parameter) (รัสรินทร์ เมธาเฉลิมพัฒน์, 2565) โดยค่า C มีค่าน้อย Margin จะกว้าง ทำให้จัดการกับข้อมูลที่เป็น Outlier ได้ แต่ถ้าค่า C มีค่ามาก Margin จะแคบ ทำให้เกิด Overfitting และไม่ทนต่อข้อมูล Outlier ทำให้เส้น Margin เพี้ยน ซึ่งจะเรียก Margin ประเภทนี้ว่า Hard margin แก้ได้โดยยอมให้มีบางข้อมูลอยู่ระหว่าง Margin ได้ ซึ่งเรียก Margin ลักษณะนี้ว่า Soft margin (Saini, 2021) (ภาพประกอบ 3)



ภาพประกอบ 2 การแบ่งเขตข้อมูลเป็น 2 กลุ่ม ของ SVM



ภาพประกอบ 3 ตัวอย่างผลลัพธ์ที่ได้จากการปรับค่าพารามิเตอร์ C ของ SVM

การทำให้ Margin กว้างมากที่สุดและลดความผิดพลาดที่ได้จากการทำนาย โดยทำให้สมการ (5) มีค่าน้อยที่สุด (Malek et al., 2022)

$$\frac{1}{2}\|w\|^2 \quad \text{..... (5)}$$

กำหนดให้

$$y_i \cdot (w \cdot x + b) \geq 1 \quad \text{เมื่อ} \quad i = 1, 2, \ldots, n$$

SVM สามารจัดการกับข้อมูลที่มีหลายมิติได้ และใช้ Kernel function $(k(x_i \cdot x))$ ช่วยจำแนกข้อมูลที่ไม่เชิงเส้นตรงได้อย่างมีประสิทธิภาพ สำหรับ SVM ที่ไม่เป็นเชิงเส้นตรง แสดงได้ดังสมการ (6)

$$f(x) = sign \sum_{i=1}^{n} y_i \alpha_i k(x_i \cdot x) + b \quad \text{..... (6)}$$

SVM มี Kernel Function ที่เลือกใช้ให้เหมาะกับลักษณะของข้อมูล ดังตาราง 3

ตาราง 3 Kernel Function ของ SVM

Kernel Function	สมการ
Linear	$K(x_k, x) = x_k^T x$
Polynomial	$K(x_k, x) = (x_k^T x/\sigma^2 + \gamma)^d$
RBF	$K(x_k, x) = exp(-\|x_k - x\|/\sigma^2)$
Sigmoid	$K(x_k, x) = tanh(\gamma x_k^T x + \gamma)$

ที่มา: Malek, N. H. A., Yaacob, W. F. W., Nasir, S. A. M., & Shaadan, N. (2022). Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. *Water, 14*(7), 1067.

4. เทคนิคการจัดการกับข้อมูลที่ไม่สมดุล (Imbalance data)

ชุดข้อมูลที่ไม่สมดุลคือ ชุดข้อมูลที่มีจำนวนข้อมูลในแต่ละกลุ่มเป้าหมาย (Class) ไม่สมดุลกัน โดยมีบางกลุ่มเป้าหมายที่มีจำนวนข้อมูลมากกว่า (Majority class) และมีบางกลุ่มเป้าหมายที่มีจำนวนข้อมูลน้อย (Minority class) ซึ่งการจำแนกประเภทของข้อมูลที่ไม่สมดุลจะส่งผลต่อการประเมินประสิทธิภาพของแบบจำลอง เช่น ค่า Accuracy Precision และ Recall เนื่องจากจะมีความเอนเอียงไปยังกลุ่มเป้าหมายส่วนมากมากกว่า (Malek et al., 2022)

การทำให้ข้อมูลของแต่ละกลุ่มเป้าหมายมีขนาดใกล้เคียงกันสามารถทำได้หลายวิธี โดยในที่นี้จะยกตัวอย่างมาเพียง 2 กรณี ดังนี้

4.1 Random Oversampling คือ เทคนิคการเพิ่มปริมาณของข้อมูลในกลุ่มเป้าหมายที่มีจำนวนน้อย โดยสุ่มเลือกตัวอย่างข้อมูลแล้วคืน (Random with replacement) (Lemaître, Nogueira, & Aridas, 2017)

4.2 SMOTE (The Synthetic Minority Oversampling Technique) เป็นเทคนิคการสังเคราะห์ข้อมูลของกลุ่มเป้าหมายที่มีจำนวนน้อยขึ้นใหม่ โดยใช้หลักการ K Neareast-Neighbors ดังสมการ (7) ซึ่งเป็นการประมาณค่าของตัวอย่างใหม่จะอยู่บนเส้นตรงระหว่าง X_i และ X_{zi} (Lemaître et al., 2017)

$$X_{new} = X_i + \lambda \times (X_i - X_{zi}) \quad \text{..... (7)}$$

โดย X_{new} คือ ตัวอย่างใหม่ได้จากการสังเคราะห์ข้อมูล

λ คือ จำนวนที่สุ่ม ซึ่งอยู่ในช่วง 0 ถึง 1

ทั้งนี้ แบบจำลองที่ใช้เทคนิค Ensemble เช่น Random Forest และ Gradient Boosting สามารถจัดการกับข้อมูลที่ไม่สมดุลได้ ทำให้ค่า Precision และ Recall ของแบบจำลองมีค่าเพิ่มขึ้น (Malek et al., 2022)

5. เทคนิค Features Importance

5.1 Gini importance

Gini importance (Breiman, 2001) วัดจากการลดลงของค่าดัชนี Gini (Gini index) หรือความไม่บริสุทธิ์ (Impurity) ที่เกิดขึ้นหลังจากแบ่งโหนดแล้วจะมีค่าลดลง Gini importance ใช้บ่งบอกถึงความเกี่ยวข้องของลักษณะเฉพาะของแบบจำลองที่ใช้ทำนายได้ อีกทั้งยังสามาถใช้จัดอันดับความสำคัญของลักษณะเฉพาะได้ ซึ่งเป็นผลพลอยได้ของแบบจำลอง Random Forest โดยการแบ่งโหนด τ ของต้นไม้ T ใน Random Forest จะใช้ Gini impurity $i(\tau)$ (สมการ (8)) เพื่อวัดว่าสามารถจำแนกประเภทของตัวอย่างได้ดีเพียงใด (Menze et al., 2009)

$$i(\tau) = 1 - p_1^2 - p_0^2 \quad \text{..... (8)}$$

โดย $p_k = \frac{n_k}{n}$

n_k คือ จำนวนตัวอย่างที่อยู่ในกลุ่ม k สมมติให้ กลุ่ม $k = \{0,1\}$

n คือ จำนวนตัวอย่างทั้งหมดของโหนด τ_i

การลดลงของ Gini impurity (Δi) จากการแบ่งและส่งตัวอย่างไปยัง 2 โหนดย่อย ได้แก่ τ_l และ τ_r ($p_l = \frac{n_l}{n}$ และ $p_r = \frac{n_r}{n}$ ตามลำดับ) ตามเกณฑ์ t_θ ของลักษณะเฉพาะ θ (สมการ (9)) (Menze et al., 2009)

$$\Delta i(\tau) = i(\tau) - p_l i(\tau_l) - p_r i(\tau_r) \quad \text{..... (9)}$$

การลดลงของ Gini impurity จากการแบ่งที่เหมาะสม $\Delta i_\theta(\tau, T)$ ของโหนดทั้งหมด τ ในต้นไม้ T ใน Forest ซึ่งจะถูกเก็บและสะสมไว้โดยแยกเป็นของแต่ละลักษณะเฉพาะ θ (สมการ (10)) (Menze et al., 2009)

$$I_G(\theta) = \sum_T \sum_\tau \Delta i_\theta(\tau, T) \quad \text{..... (10)}$$

Gini importance ใช้บ่งชี้ว่าลักษณะเฉพาะ (Feature) θ นำมาใช้เพื่อแบ่งโหนดบ่อยเพียงใด และบ่งบอกว่าค่าการจำแนกประเภทโดยรวมของลักษณะเฉพาะนั้นมีมากเพียงใด สำหรับการศึกษาปัญหาการจำแนกประเภท (Menze et al., 2009)

นอกจาก Random Forest แล้ว อัลกอริทึมที่สร้างบนพื้นฐานของ Tree (Tree-based algorithm) อื่นๆ เช่น XGBoost สามารถจัดลำดับความสำคัญของลักษณะเฉพาะโดยอาศัย Gini importance ได้เช่นกัน (Shi, Wong, Li, Palanisamy, & Chai, 2019)

5.2 Permutation importance

Breiman (2001) ได้นำเสนออัลกอริทึม Permutation ใน Random Forest ซึ่ง Permutation importance เป็นค่าเฉลี่ยความแม่นยำที่ลดลงของแบบจำลองเมื่อแต่ละลักษณะเฉพาะถูกสุ่มสับเปลี่ยนขณะที่อยู่ในขั้นตอนการฝึกฝนของแบบจำลอง วิธีการดังกล่าวเป็นการรบกวนความสัมพันธ์ระหว่างลักษณะเฉพาะ (Feature) และกลุ่มเป้าหมาย (Target) หากความถูกต้องของแบบจำลองลดลงจากการสุ่มสับเปลี่ยนลักษณะเฉพาะใดๆ สามารถบ่งชี้ได้ว่าลักษณะเฉพาะนั้นมีผลต่อการจำแนกประเภทของข้อมูล (Arunthavanathan, Khan, Ahmed, & Imtiaz, 2022) สามารถหาค่า Permutation importance ได้ดังสมการ (11) (Scikit-learn developers, 2022b)

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \quad \text{..... (11)}$$

กำหนดให้ i_j คือ ค่า Permutation importance ของลักษณะเฉพาะ j

s คือ คะแนนอ้างอิงของแบบจำลอง m ที่ได้จากข้อมูล D (เช่น ค่า Accuracy)

K คือ จำนวนข้อมูล

$s_{k,j}$ คือ คะแนนของแบบจำลองที่ถูกรบกวนจากการสุ่มและสับเปลี่ยนลักษณะเฉพาะ j

5.3 SHAP

SHAP (SHapely Additive exPlanations) ใช้เพื่ออธิบายผลการทำนายที่ได้จากแบบจำลอง และบ่งบอกถึงการมีส่วนร่วมของแต่ละลักษณะเฉพาะ (Feature) ที่สอดคล้องกันเพื่อการทำนาย อีกทั้งอธิบายความสัมพันธ์ของลักษณะเฉพาะในรูปแบบภาพรวมและการส่งผลของลักษณะเฉพาะในแต่ละตัวอย่าง (Ekanayake, Meddage, & Rathnayake, 2022)

Lundberg and Lee (2017) เสนอเทคนิค SHAP เพื่อใช้อธิบายผลที่ได้จากการทำนายของแบบจำลอง ซึ่งอ้างอิงตามทฤษฎีเกม ตัวอย่างเช่น ข้อมูลนำเข้าเปรียบเสมือนผู้เล่นเกม การทำนายเปรียบเสมือนเงินรางวัล และ SHAP คือการมีส่วนร่วมของผู้เล่นแต่ละคนในเกม ต่อมา Lundberg, Erion, and Lee (2019) ได้เสนอเทคนิค SHAP รูปแบบอื่นๆ ที่มีความเฉพาะกับประเภทของแบบจำลอง ได้แก่ DeepSHAP, Kernel SHAP, LinearSHAP และ TreeSHAP ซึ่ง TreeSHAP ใช้เพื่ออธิบายการทำนายของแบบจำลองที่สร้างบนพื้นฐานของ Tree (Ekanayake et al., 2022)

การอธิบายการทำนายของแบบจำลองโดยอาศัยค่า Shapely นั้นเป็นส่วนหนึ่งของวิธีการ Additive feature attribution ซึ่งอธิบายผลการทำนายของแบบจำลองที่ได้จากผลรวมของค่าจริงของคุณสมบัติแต่ละลักษณะเฉพาะ การคำนวณค่า Shapley ดังสมการ (12) กำหนดให้แบบจำลองที่ใช้อธิบาย g เป็นฟังก์ชันเชิงเส้นของตัวแปรที่มีค่าได้เพียง 2 ค่า (Binary variable) (Lundberg & Lee, 2017)

$$g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z'_i \quad \text{..... (12)}$$

โดย $z' \in \{0,1\}^M$

g คือ แบบจำลองที่ใช้อธิบาย

M คือ จำนวนของข้อมูลนำเข้า

$\emptyset_i$ คือ ค่าคุณสมบัติของลักษณะเฉพาะ (Feature) โดย $\emptyset_i \in \mathbb{R}$

z'_i คือ ลักษณะเฉพาะที่สนใจ $(z'_i = 1)$ หรือไม่ทราบ $(z'_i = 0)$

ค่าคุณสมบัติของแต่ละลักษณะเฉพาะ คำนวณได้จากสมการ (13)

$$\emptyset_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \quad \text{..... (13)}$$

โดย $f_x(S) = E[f(x)|x_s]$

S คือ เซตของ Non-zero index ที่อยู่ใน z' (ภาพประกอบ 4)

N คือ เซตของลักษณะเฉพาะทั้งหมด

M คือจำนวนของลักษณะเฉพาะ

การคำนวณหาค่า SHAP ที่แสดงปฏิสัมพันธ์ระหว่างลักษณะเฉพาะ i และ j (Lundberg et al., 2019) คำนวณได้ดังสมการ (14)

$$\emptyset_{i,j} = \sum_{S \subseteq N\{i,j\}} \frac{|S|!(M-|S|-2)!}{2(M-1)!} \nabla_{ij}(S) \quad \text{..... (14)}$$

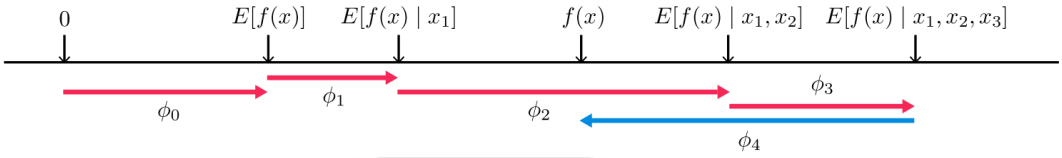
โดย $i \neq j$ และ

$$\nabla_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \quad \text{..... (15)}$$

$$= f_x(S \cup \{i,j\}) - f_x(S \cup \{j\}) - [f_x(S \cup \{i\}) - f_x(S)] \quad \text{..... (16)}$$

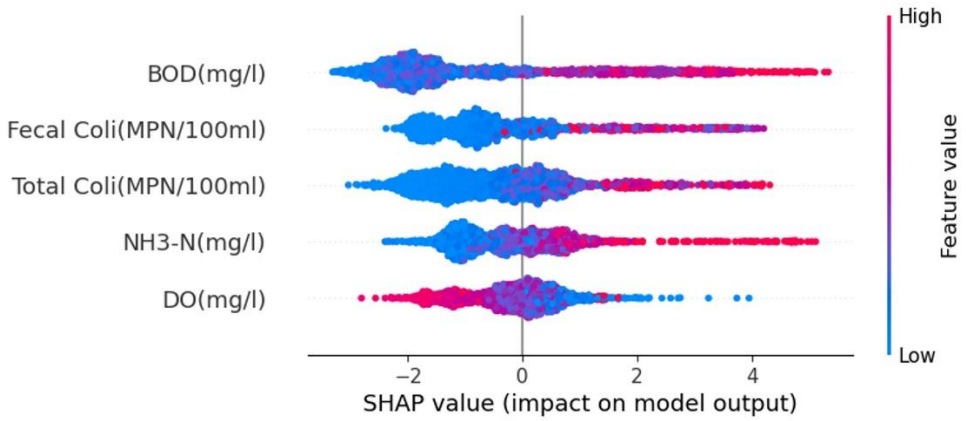
กำหนดให้ $\emptyset_{i,j} = \emptyset_{j,i}$ และผลกระทบของปฏิสัมพันธ์ทั้งหมดได้จาก $\emptyset_{i,j} + \emptyset_{j,i}$

การอธิบายความสำคัญของลักษณะเฉพาะในลักษณะของแผนภูมิจะใช้ SHAP summary plot (Lundberg et al., 2019) ดังตัวอย่างในภาพประกอบ 5



ภาพประกอบ 4 ค่า SHAP ใช้อธิบายผลลัพธ์ของฟังก์ชั่น f ซึ่งได้จากผลรวมของ ϕ_i ของแต่ละลักษณะเฉพาะ (Feature) ที่นำมาใช้

ที่มา: Lundberg, S. M., & Lee, S. (2017). *A Unified Approach to Interpreting Model Predictions.* Paper presented at the 31st International Conference on Neural Information Processing Systems, Long Beach, CA.



ภาพประกอบ 5 ตัวอย่าง SHAP summary plot

จากภาพประกอบ 5 จะเห็นว่า แกน y แสดงลักษณะเฉพาะซึ่งเรียงลำดับจากลักษณะเฉพาะที่มีผลต่อการทำนายโดยภาพรวมของแบบจำลอง (Global impact) $\sum_{j=1}^{N} |\emptyset_i^{(j)}|$ มากที่สุดไปยังน้อยที่สุด และแต่ละจุดที่กระจายตามแนวนอนของแผนภูมิแสดงค่า SHAP $\emptyset_i^{(j)}$ โดยสีของจุดบ่งบอกถึงค่าลักษณะเฉพาะ (Feature value) ที่มีตั้งแต่ค่าต่ำ (สีน้ำเงิน) ไปจนถึงค่าสูง (สีแดง)

6. เทคนิคอนุกรมเวลา (Time series algorithm)

อนุกรมเวลาแบ่งได้เป็น 2 ประเภท (กิติ์สุชาต พสุภา, 2564) ดังนี้

1) อนุกรมเวลาแบบคงที่ (Stationary Time series) เป็นอนุกรมเวลาที่มีค่าเฉลี่ย (Mean) ค่าความแปรปรวน (Variance) และค่าความแปรปรวนร่วม (Covariance) ไม่เปลี่ยนแปลงตามเวลา

2) อนุกรมเวลาแบบไม่คงที่ (Non-stationary Time series) เป็นอนุกรมเวลาที่มีค่าเฉลี่ย (Mean) ค่าความแปรปรวน (Variance) และค่าความแปรปรวนร่วม (Covariance) เปลี่ยนแปลงตามเวลา

ข้อมูลอนุกรมเวลามี 4 องค์ประกอบ (กิติ์สุชาต พสุภา, 2564; ภูมิฐาน รังคกูลนุวัฒน์, 2562; Lazzeri, 2020, pp. 1-26) ดังนี้ (ภาพประกอบ 6)

1) แนวโน้ม (Trend) เป็นรูปแบบของค่าอนุกรมเวลาเพิ่มขึ้นหรือลดลงเรื่อยๆ เมื่อระยะเวลาผ่านไป

2) ฤดูกาล (Seasonality) เป็นรูปแบบการเปลี่ยนแปลงที่เกิดขึ้นซ้ำของอนุกรมเวลาโดยมีรูปแบบการเปลี่ยนแปลงเหมือนเดิมในช่วงเวลาเดียวกัน

3) วัฏจักร (Cyclicity) เป็นรูปแบบการเปลี่ยนของอนุกรมเวลาที่เกิดขึ้นซ้ำ แต่ไม่ได้เกิดในช่วงเวลาเดิม ซึ่งการนับระยะเวลาของวัฏจักรนั้น นับจากจุดสูงสุด (Peak) ของอนุกรมเวลาจุดหนึ่งไปยังจุดสูงสุดอีกจุดหนึ่ง หรือจากจุดต่ำสุด (Trough) จุดหนึ่งไปยังจุดต่ำสุดอีกจุดหนึ่ง โดยมักใช้ระยะเวลามากกว่า 1 ปี

4) ความผิดปกติ (Irregularity) เป็นรูปแบบการเปลี่ยนแปลงของอนุกรมเวลาที่ไม่สามารถคาดเดาได้ หรือไม่มีรูปแบบที่แน่นอน

รูปแบบความสัมพันธ์ของอนุกรมเวลามี 2 รูปแบบ ได้แก่

1) รูปแบบผลบวก (Additive decomposition)

$$Y = T + S + C + I \qquad \text{..... (17)}$$

2) รูปแบบผลคูณ (Multiplicative decomposition)

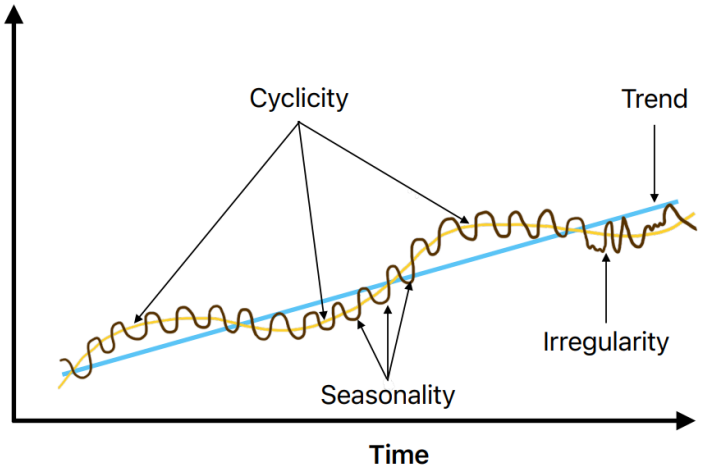
$$Y = T \times S \times C \times I \qquad \text{..... (18)}$$

เมื่อ Y คือ ข้อมูลอนุกรมเวลา

T คือ แนวโน้ม

S คือ ฤดูกาล

I คือ ความผิดปกติ



ภาพประกอบ 6 องค์ประกอบของอนุกรมเวลา

แบบจำลองที่ใช้ทำนายข้อมูลอนุกรมเวลา

6.1 ARIMA หรือ Auto-Regressive Integrated Moving Average ใช้ทำนายข้อมูลอนุกรมเวลาที่เป็น Non-seasonality ซึ่งข้อมูลไม่เปลี่ยนแปลงไปตามฤดูกาล (Su & Ye, 2020) ARIMA ประกอบด้วย 3 พารามิเตอร์ ได้แก่ p (ลำดับของ Auto-Regressive (AR) หรือลำดับของข้อมูลอนุกรมเวลาในอดีตเพื่อทำนาย), d (จำนวนครั้งของผลต่าง (Differencing) อนุกรมเวลา เพื่อใช้ข้อมูลมีลักษณะ Stationary) และ q (ลำดับของ Moving Average (MA) แสดงจำนวน lag ที่ใช้ข้อมูลความผิดพลาดในอดีตมาทำนาย) โดย ARIMA สามารถเขียนได้ดังสมการ (19)

6.2 ARIMAX หรือ Auto-Regressive Integrated Moving Average with Exogenous variables ต่างจาก ARIMA ตรงที่เพิ่มตัวแปรภายนอกเข้าไป ในแบบจำลอง (Exogenous variables) (Su & Ye, 2020) ดังสมการ (20)

$$(1-B)^d Y_t = \frac{\Theta(B)}{\Phi(B)} Z_t \quad \text{..... (19)}$$

$$Y_t = \mu + \sum_{i=1}^{k} \frac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} X_{it} + \varepsilon_t \quad \text{..... (20)}$$

โดย $\Phi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p$

$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$

$\varepsilon_t = \frac{\Theta(B)}{\Phi(B)} Z_t$

เมื่อ d คือ จำนวนครั้งของผลต่าง (Differencing) อนุกรมเวลา

B คือ Backward shift operator โดย $BY_t = Y_{t-1}$

Y_t คือ ค่าอนุกรมเวลาที่เวลา t

Y_{t-1} คือ ค่าอนุกรมเวลาที่เวลา t -1

Z_t คือ ค่าความคลาดเคลื่อนที่เวลา t

$\Phi(B)$ คือ Auto-Regressive

$\Theta(B)$ คือ Moving Average

φ คือ พารามิเตอร์ของ Auto-Regressive

θ คือ พารามิเตอร์ของ Moving Average

X_{it} คือ ลำดับของตัวแปรนำเข้า

l_i คือ Delay order ของตัวแปรนำเข้าที่ i

μ คือ ค่าเฉลี่ย

ε_t คือ ค่า White noise หรือค่าความคลาดเคลื่อนของแบบจำลอง ณ เวลา t

6.3 SARIMA หรือ Seasonal Auto-Regressive Integrated Moving Average ใช้ทำนายข้อมูลอนุกรมเวลาที่มี Seasonality โดย SARIMA สามารถเขียนในรูป SARIMA (p, d, q) × $(P, D, Q)_S$ ซึ่ง (p, d, q) แทนส่วนที่เป็น Non-seasonality และ $(P, D, Q)_S$ แทนส่วน Seasonality ซึ่ง S คือช่วงเวลาใน 1 Season (Vagropoulos, Chouliaras, Kardakos, Simoglou, & Bakirtzis, 2016) SARIMA สามารถเขียนได้ดังสมการ (21)

6.4 SARIMAX หรือ Seasonal Auto-Regressive Integrated Moving Average with Exogenous variables ต่างจาก SARIMA ตรงที่เพิ่มตัวแปรภายนอกเข้าไปในแบบจำลอง (Exogenous variables) สามารถเขียนในรูป SARIMAX (p, d, q) × $(P, D, Q)_S$ (Vagropoulos et al., 2016) ดังสมการ (22)

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad \text{..... (21)}$$

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D Y_t = \beta_k x_{k,t}{}' + \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad \text{..... (22)}$$

โดย

$$\nabla^d = (1-B)^d$$
$$\nabla_s^D = (1-B^s)^D$$

เมื่อ Y_t คือ ค่าอนุกรมเวลาที่เวลา t

$\varphi_p(B)$ คือ Non-seasonal Autoregressive ที่ลำดับ p

$\Phi_P(B^s)$ คือ Seasonal Autoregressive ที่ลำดับ P

$\theta_q(B)$ คือ Non-seasonal Moving Average ที่ลำดับ q

$\Theta_Q(B^s)$ คือ Seasonal Moving Average ที่ลำดับ Q

∇^d คือ จำนวนครั้งของผลต่าง (Differencing) ของ Non-seasonal

∇_s^D คือ จำนวนครั้งของผลต่าง (Differencing) ของ Seasonal

ε_t คือ ค่า White noise หรือค่าความคลาดเคลื่อนของแบบจำลอง ณ เวลา t

$x_{k,t}{}'$ คือ เวกเตอร์ที่ประกอบด้วยตัวแปร Exogenous ลำดับที่ k ณ เวลา t

β_k คือ ค่าสัมประสิทธิ์ของตัวแปร Exogenous ลำดับที่ k

การประเมินประสิทธิภาพของแบบจำลอง

การประเมินประสิทธิภาพของแบบจำลอง ใช้เพื่อวัดประสิทธิภาพและเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง และนำข้อมูลที่ได้มาพิจารณาว่าแบบจำลองที่สร้างมีความเหมาะสมสำหรับนำมาใช้งานต่อไปหรือไม่

การประเมินประสิทธิภาพของแบบจำลองแบ่งได้เป็น 2 กรณี

1. การจำแนกประเภท (Classification)

โดยการจำแนกประเภทจะแบบกลุ่มเป็น 2 กลุ่มเพื่อใช้ประเมินประสิทธิภาพ ได้แก่ กลุ่มบวก (Positive class) และกลุ่มลบ (Negative class) สามารถพิจารณาโดยแบ่งเป็น 2 กรณี ดังนี้

1) กรณีการจำแนกแบบทวินาม (Binary classification) สามารถแบ่งได้ 2 กลุ่ม ได้แก่ กลุ่มที่สนใจจะกำหนดให้เป็นกลุ่มบวก หรือ Positive class และกลุ่มที่ไม่ได้สนใจจะกำหนดเป็นกลุ่มลบ หรือ Negative class ทั้งนี้ สามารถใช้เลขแทนความหมายได้ กรณีกำหนด [-1, 1] ให้ -1 เป็นกลุ่มลบ และ 1 กลุ่มบวก สำหรับกรณี [0, 1] จะให้ 0 เป็นกลุ่มลบ และ 1 เป็นกลุ่มบวก

2) กรณีการจำแนกประเภทแบบหลายประเภท (Multiclass classification) ซึ่งมีจำนวนกลุ่มเป้าหมาย (Class) มากกว่า 2 กลุ่ม กลุ่มที่สนใจกำหนดเป็นกลุ่มบวกหรือ Positive class และกลุ่มที่เหลือจะจัดเป็นกลุ่มลบ หรือ Negative class

การประเมินประสิทธิภาพของแบบจำลองจะใช้วิธีการประเมินประสิทธิภาพดังต่อไปนี้

1.1 Confusion Matrix

เป็นตารางที่ใช้แสดงความสามารถในการทำนายของแบบจำลอง (ตาราง 4) โดยตารางมีขนาดความกว้างและความยาวเท่ากับจำนวนกลุ่มเป้าหมาย (Class) แต่ละแถวจะแสดงผลที่ได้จากการทำนาย (Predicted class) และแต่ละคอลัมน์แสดงคำตอบจริง (Actual class) ซึ่งค่าที่ได้จาก Confusion matrix ประกอบด้วย 4 ค่า (Hossin, 2015) ดังนี้

1) ค่าบวกจริง (True Positive: TP) คือ แบบจำลองทำนายว่าเป็นบวก ซึ่งเป็นจริงตามคำตอบจริงที่เป็นผลบวก นั่นคือ แบบจำลองทำนายผลบวกถูกต้อง

2) ผลบวกลอง (False Positive: FP) คือ แบบจำลองทำนายกลุ่มลบว่าเป็นกลุ่มบวก ซึ่งเป็นความผิดพลาดชนิดที่ 1 (Type 1 error) สมมติฐานว่าง (Null hypothesis) เป็นจริง แต่ปฏิเสธสมมติฐานว่าง (Null hypothesis: H_0)

3) ผลลบจริง (True Negative: TN) คือ แบบจำลองทำนายกลุ่มลบได้ถูกต้อง

4) ผลลบลวง (False Negative: FN) คือ แบบจำลองทำนายกลุ่มบวกเป็นกลุ่มลบ ซึ่งเป็นความผิดพลาดชนิดที่ 2 (Type 2 error) ยอมรับสมมติฐานว่าง (Null hypothesis) แต่สมมติฐานว่าง (Null hypothesis: H_0) ไม่เป็นจริง

ตาราง 4 Confusion Matrix

	Actual Positive Class	Actual Nagative Class
Predicted Positive Class	True Positive (TP)	False Negative (FN)
Predicted Negative Class	False Positive (FP)	True Negative (TN)

ที่มา: Hossin, M., Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 1-11.

1.2 Accuracy

Accuracy คือ การวัดความแม่นยำของแบบจำลอง เป็นสัดส่วนระหว่างจำนวนตัวอย่างที่แบบจำลองทำนายถูกกับจำนวนตัวอย่างทั้งหมด ดังสมการ (23) Accuracy เหมาะกับข้อมูลที่มีจำนวนของกลุ่มเป้าหมายเท่าๆ กัน โดยมีค่า Accuracy มากที่สุดเท่ากับ 1 และน้อยที่สุดเท่ากับ 0

$$Accuracy = \frac{TP + TN}{P + N} \quad \text{..... (23)}$$

1.3 Precision

ความแม่นยำของแบบจำลองที่สามารถทำนายกลุ่มบวกได้ถูกต้อง โดยเป็นอัตราส่วนของกลุ่มบวกจริง (True positive) ต่อผลการทำนายว่าเป็นกลุ่มบวก ดังสมการ (24)

$$Precision = \frac{TP}{TP + FP} \quad \text{..... (24)}$$

1.4 Recall

การวัดประสิทธิภาพของแบบจำลองว่ามีความสามารถสำหรับทำนายผลบวกจริงได้ถูกต้องเพียงใด โดยเป็นอัตราส่วนของผลบวกจริง (True Positive) ต่อจำนวนของกลุ่มบวกทั้งหมด ดังสมการ (25)

$$Recall = \frac{TP}{TP + FN} \quad \text{..... (25)}$$

1.5 F1 score

เป็นค่าเฉลี่ยระหว่าง (Harmonic mean) ระหว่างค่า Recall และ Precision ดังสมการ (26)

$$F1\ Score = \frac{2}{\frac{1}{Recall} \times \frac{1}{Precision}} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{..... (26)}$$

เมื่อพิจารณากรณีการจำแนกหลายประเภท (Multiclass classification) สามารถคำนวณค่าเฉลี่ยของค่า Precision, Recall และ F1 score ได้ (Khalusova, 2022; Scikit-learn developers, 2022c; Shamsuddin, Othman, & Sani, 2022) ดังนี้

1) Micro average คำนวณโดยนับจากจำนวนตัวอย่าง เพื่อนำมาใช้หาค่าเฉลี่ย

2) Macro average คำนวณจากการนำค่าที่ได้จาก Class ทุก Class มาหาค่าเฉลี่ยของ Metric

3) Weighted average คำนวณค่าของแต่ละ Class และถ่วงน้ำหนักของแต่ละ Class ด้วยจำนวนสมาชิก แล้วจึงนำมาหาค่าเฉลี่ย

สูตรที่ใช้คำนวณ เป็นไปดังตาราง 5 (กำหนดให้ n คือ จำนวน Class และ N คือ จำนวนตัวอย่าง)

2. อนุกรมเวลา (Time series)

การวัดประสิทธิภาพของแบบจำลองอนุกรมเวลา สามารถใช้วิธีการต่างๆ (กิติ์สุชาต พสุภา, 2564) ดังนี้

ตาราง 5 สูตรการคำนวณเพื่อวัดประสิทธิภาพของแบบจำลองสำหรับการจำแนกหลายประเภท (Multiclass classification)

การวัดประสิทธิภาพ	สูตร
Micro Averaged Precision	$\frac{\sum_{k=1}^{n} TP_k}{\sum_{k=1}^{n}(TP_k + FP_k)}$
Micro Averaged Recall	$\frac{\sum_{k=1}^{n} TP_k}{\sum_{k=1}^{n}(TP_k + FN_k)}$
Micro Averaged F1 score	$2 \times \frac{Micro\ averaged\ Recall \times Micro\ averaged\ Precision}{Micro\ averaged\ Recall + Micro\ averaged\ Precision}$
Macro Averaged Precision	$\frac{\sum_{k=1}^{n} \frac{TP_k}{TP_k + FP_k}}{n}$
Macro Averaged Recall	$\frac{\sum_{k=1}^{n} \frac{TP_k}{TP_k + FN_k}}{n}$
Macro Averaged F1 score	$\frac{\sum_{k=1}^{n}(F1\ score_k)}{n}$
Weighted Averaged Precision	$\frac{\sum_{k=1}^{n}\left(\frac{TP_k}{TP_k + FP_k} \times N_k\right)}{N}$
Weighted Averaged Recall	$\frac{\sum_{k=1}^{n}\left(\frac{TP_k}{TP_k + FN_k} \times N_k\right)}{N}$
Weighted Averaged F1 score	$\frac{\sum_{k=1}^{n}(F1\ score_k \times N_k)}{N}$

2.1 MAE (Mean Absolute Error)

หรือ L1 Loss คือค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ย มีค่าอยู่ระหว่าง [0, ∞) (สมการ (27))

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \qquad \text{..... (27)}$$

กำหนดให้ y_i คือ ค่าจริง

$\hat{y}$ คือ ค่าที่ได้จากการทำนาย

2.2 RMSE (Root Mean Squared Error)

ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง มีค่าอยู่ระหว่าง [0, ∞) (สมการ (28))

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad \text{..... (28)}$$

กำหนดให้ y_i คือ ค่าจริง

$\hat{y}$ คือ ค่าที่ได้จากการทำนาย

2.2 MAPE (Mean Absolute Percentage Error)

ค่าร้อยละความคลาดเคลื่อนสมบูรณ์เฉลี่ย มีค่าระหว่างร้อยละ 0 – 100 (สมการ (29))

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}|}{y_i} \times 100 \quad \text{..... (29)}$$

กำหนดให้ y_i คือ ค่าจริง

$\hat{y}$ คือ ค่าที่ได้จากการทำนาย

งานวิจัยที่เกี่ยวข้อง

1. Sillberg et al. (2021) ศึกษาการจำแนกคุณภาพแม่น้ำเจ้าพระยาในประเทศไทยด้วยเทคนิค SVM (Support Vector Machine) โดยใช้ข้อมูลแม่น้ำเจ้าพระยาระหว่าง พ.ศ. 2551 – 2562 ค่าพารามิเตอร์คุณภาพน้ำที่นำมาใช้เป็นข้อมูลลักษณะเฉพาะ (Feature) ประกอบด้วย 12 พารามิเตอร์ ได้แก่ BOD, การนำไฟฟ้า, DO, FCB, TCB, แอมโมเนีย (NH_3-N), ไนเตรท, ความเค็ม, ของแข็งแขวนลอย, ไนโตรเจนทั้งหมด, ปริมาณของแข็งที่ละลายได้ทั้งหมด (TDS) และความขุ่น พบว่า พารามิเตอร์ที่มีความสัมพันธ์กับการจำแนกคุณภาพแม่น้ำเจ้าพระยามากที่สุดจากการใช้อัลกอริทึม Apriori คือ NH_3-N รองลงมาคือ TCB, FCB, BOD, DO และความเค็มตามลำดับ และเมื่อนำ 6 พารามิเตอร์ดังกล่าวมาใช่สำหรับจำแนกคุณภาพแม่น้ำด้วยเทคนิค SVM ร่วมกับ Kernel function ที่เป็น Linear พบว่าได้ค่า Accuracy มากที่สุด เท่ากับ 0.94 และได้ค่า Precision เท่ากับ 0.84 ค่า Recall เท่ากับ 0.84 และค่า F1 score เท่ากับ 0.84 อย่างไรก็ตาม เมื่อนำแบบจำลองดังกล่าวมาใช้จำแนกคุณภาพแม่น้ำท่าจีนตั้งแต่ปี พ.ศ. 2560 - 2562 เพื่อตรวจสอบความถูกต้องของแบบจำลอง (Validation) พบว่า ค่า Accuracy เท่ากับ 0.95 เมื่อเลือกใช้เพียง 6 พารามิเตอร์ ข้างต้น

2. Ahmed et al. (2019) จำแนกคุณภาพน้ำและทำนายค่าดัชนีชี้วัดคุณภาพน้ำ (WQI) ของทะเลสาบ Rawal ประเทศปากีสถาน โดยเลือกใช้พารามิเตอร์คุณภาพน้ำจำนวน 4 พารามิเตอร์ จากทั้งหมด 12 พารามิเตอร์ ได้แก่ อุณหภูมิของน้ำ ความขุ่น pH และ TDS ซึ่งพารามิเตอร์ทั้ง 4 นี้ ได้มาจากการคำนวณค่า Pearson correlation จากนั้น นำทั้ง 4 พารามิเตอร์ มาเป็นข้อมูลนำเข้าแบบจำลองเพื่อการทำนายค่า WQI พบว่า Gradient Boosting (ใช้ Loss function เป็น least squares regression จำนวนตัวอย่างขั้นต่ำที่จะแบ่งโหนดภายใน (Internal node) เท่ากับ 2 ตัวอย่าง และใช้ Learning rate เท่ากับ 0.1) เป็นแบบจำลองที่ให้ประสิทธิภาพดีที่สุด โดยค่า MAE เท่ากับ 1.9642 ค่า MSE เท่ากับ 7.2011 ค่า RMSE เท่ากับ 2.6835 และ RSE เท่ากับ 0.7485 รองลงมาได้แก่ Polynomial Regression และ Random Forest ตามลำดับ และการสร้างแบบจำลองเพื่อทำนายระดับคุณภาพน้ำ หรือ WQC (Water Quality Class) พบว่า แบบจำลองที่ใช้เทคนิค Multi-layer perceptron (MLP) (จำนวน 3 Hidden layer 7 Element ใช้ 200 epochs และ lbfgs optimizer) ให้ประสิทธิภาพดีที่สุด มีค่า Accuracy, Precision, Recall,และ F1 score เท่ากับ 0.8507, 0.5659, 0.5640 และ 0.5649 ตามลำดับ รองลงมาได้แก่ Logistic Regression และ Stochastic Gradient Descent ตามลำดับ

3. Kouadri, Elbeltagi, Islam, and Kateb (2021) ทำการศึกษาประสิทธิภาพของการเรียนรู้ของเครื่องสำหรับการทำนายดัชนีชี้วัดคุณภาพน้ำใต้ดิน (WQI) ของภูมิภาค Illizi ของประเทศแอลจีเรีย พบว่า แบบจำลองที่ใช้เทคนิค MLP มีประสิทธิภาพสูงสุด ได้ค่า Correlation coefficient (R), MAE และ RMSE เท่ากับ 1, 1.4572×10^{-08} และ 2.1418×10^{-08} ตามลำดับ สำหรับทำนายค่า WQI จากพารามิเตอร์คุณภาพน้ำจำนวน 12 พารามิเตอร์ (การนำไฟฟ้า (EC), ความกระด้างน้ำ (TH), pH, TDS (Total dissolved salts), HCO_3^-, Ca^{2+}, Mg^{2+}, Na^+, K^+, SO_4^{2-}, Cl^- และ NO_3^-) แต่เมื่อเลือกใช้เพียง 2 พารามิเตอร์ที่มีความสัมพันธ์กับค่า WQI มากที่สุด (จากการหาค่า Pearson correlation coefficient) ได้แก่ ความกระด้างน้ำ (TH) และ TDS พบว่าแบบจำลอง Random Forest มีประสิทธิภาพสูงสุด ซึ่งค่า R, MAE และ RMSE เท่ากับ 0.9984, 1.9942 และ 3.2488 ตามลำดับ โดยแบบจำลอง Random Forest ดังกล่าว กำหนดค่าพารามิเตอร์ Batch size เท่ากับ 100 Bag size percent เท่ากับ 100 Max depth เท่ากับ 0 จำนวน Executions slots เท่ากับ 1, จำนวนของ Iterations เท่ากับ 100 และ Random seed เท่ากับ 1

4. Malek et al. (2022) ศึกษาการจำแนกคุณภาพลุ่มแม่น้ำ Kelantan ในประเทศมาเลเซีย ด้วยเทคนิคการเรียนรู้ของเครื่อง 7 เทคนิค ได้แก่ Decision Tree, Artificial Neural Networks, K-Nearest Neighbors (K-NN), Naïve Bayes, SVM, Random Forest และ Gradient Boosting

ใช้พารามิเตอร์คุณภาพน้ำจำนวน 13 พารามิเตอร์ ได้แก่ DO, BOD, COD, TSS, pH, NH_3-N, อุณหภูมิของน้ำ, ค่าการนำไฟฟ้า, ความเค็ม, ความขุ่น, ไนโตรเจน, ฟอสฟอรัส และ *E. coli* พบว่า แบบจำลอง Gradient Boosting (maximal depth = 5 จำนวน Tree = 50 และ learning rate = 0.1) เป็นแบบจำลองที่ให้ค่า Balanced Accuracy มากที่สุด เท่ากับ 89.36% Accuracy เท่ากับ 94.90% Precision เท่ากับ 94.12% Recall เท่ากับ 98.72% F1 score เท่ากับ 86.49% และค่า AUC เท่ากับ 0.9811 ลำดับถัดมา ได้แก่ แบบจำลอง Random Forest และ ANN ตามลำดับ แต่เมื่อพิจารณาเฉพาะค่า Recall พบว่า แบบจำลอง ANN มีค่า Recall (เท่ากับ 82.50%) สูงที่สุด นอกจากนี้ การศึกษาดังกล่าวแสดงให้เห็นว่า TSS (Total Suspended Solids), NH_3-N, BOD, COD, ความขุ่น และ DO เป็นเป็นพารามิเตอร์ที่สำคัญเพื่อใช้จำแนกคุณภาพน้ำมากที่สุด ทั้งนี้ คณะผู้วิจัยได้เสนอว่า ข้อมูลคุณภาพน้ำที่ไม่สมดุล (Imbalanced data) จะส่งผลต่อความถูกต้องของการทำนาย

5. Suphawan and Chaisee (2021) ศึกษาการทำนายดัชนีชี้วัดคุณภาพลุ่มแม่น้ำปิงในประเทศไทย จากตัวแปรที่มีผลต่อสภาพภูมิอากาศ ประกอบด้วย อุณหภูมิเฉลี่ยต่อเดือน ความชื้นเฉลี่ยต่อเดือน ปริมาณน้ำฝนต่อเดือน และการระเหยของน้ำเฉลี่ยต่อเดือน ด้วยแบบจำลอง GPR (Gaussian Process Regression), MLR (Multiple Linear Regression) และ ANN (Artificial Neural Network) พบว่า ปริมาณน้ำฝนเป็นตัวแปรที่มีผลต่อค่า WQI ของลุ่มแม่น้ำปิงมากที่สุด และแบบจำลอง GPR มีประสิทธิภาพสำหรับทำนายค่า WQI มากกว่า MLP และ ANN ทั้งกรณีที่ใช้ตัวแปรทั้ง 4 ตัวแปร และใช้เพียงปริมาณน้ำฝนเป็นข้อมูลนำเข้าแบบจำลอง

6. Northep, Srijiranon, and Eiamkanitchat (2020) จำแนกคุณภาพของค่า DO ของแม่น้ำวัง ในประเทศไทย ใช้ข้อมูลคุณภาพน้ำตั้งแต่ พ.ศ. 2544 – 2562 ซึ่งเป็นข้อมูลคุณภาพน้ำที่เก็บประมาณ 4 ต่อปี และจำแนกคุณภาพของค่า DO ว่าอยู่ในระดับดีหรือไม่ดี พบว่า เทคนิค MLP-kNN สามารถจำแนกคุณภาพของ DO ได้ดีที่สุด โดยค่าพารามิเตอร์น้ำของแต่ละสถานีที่ใช้เป็นของมูลนำเข้าแบบจำลองจะไม่เหมือนกันขึ้นอยู่กับค่า Pearson correlation coefficient เช่น สถานี WA01 พบว่า อุณหภูมิ pH, DO, BOD และ TS (Total Solids) มีค่า Coefficient มากที่สุด และสถานี WA4.1 พบว่า อุณหภูมิ pH, DO, BOD และ TCB มีค่า Coefficient มากที่สุด

7. Uyun and Sulistyowati (2020) จำแนกคุณภาพของแม่น้ำ Brantas ในประเทศอินโดนีเซีย ด้วยเทคนิคการเรียนรู้ของเครื่อง ร่วมกับเทคนิค SMOTE เพื่อแก้ปัญหาความไม่สมดุลของข้อมูล และเทคนิคการเลือกลักษณะเฉพาะ (Feature selection) 5 เทคนิค ได้แก่ Chi square, Correlation, Derivation, Information gain และ Rule จากผลการศึกษาพบว่า เมื่อใช้เทคนิค

SMOTE ค่า Accuracy เพิ่มขึ้นจาก 83.3% เป็น 98.8% และเมื่อใช้เทคนิคการเลือกลักษณะเฉพาะ (Feature selection) ร่วมด้วย พบว่า โดยเฉลี่ยแล้ว Rule และ Information gain ทำให้ได้ค่า Accuracy มากที่สุด นอกจากนี้ เมื่อใช้ Information gain ร่วมกับอัลกอริทึม Decision tree เพื่อจำแนกคุณภาพน้ำ ทำให้ได้ค่า Accuracy เพิ่มขึ้นเป็น 99.5%

8. Singkran et al. (2010) ศึกษาการทำนายค่าดัชนีชี้วัดคุณภาพน้ำ (WQI) ของแม่น้ำในภาคตะวันออกเฉียงเหนือของประเทศไทยจำนวน 5 สาย ด้วยแบบจำลองอนุกรมเวลา (Time series model) 8 แบบจำลอง ได้แก่ Single moving average, Single exponential smoothing, Seasonal additive, Seasonal multiplicative, Double moving average, Double exponential smoothing, Holt-Winters' additive และ Holt-Winters' multiplicative ซึ่งใช้ข้อมูลปี พ.ศ. 2537 – 2550 ซึ่งแบ่งเป็น 2 ฤดู ได้แก่ ฤดูฝน (เดือนมิถุนายนถึงพฤศจิกายน) และฤดูแล้ง (เดือนธันวาคมถึงพฤษภาคม) เป็นข้อมูล WQI เพื่อให้แบบจำลองเรียนรู้ และข้อมูลปี พ.ศ. 2551 – 2555 (5 ปี) เป็นข้อมูล WQI อนาคต โดยค่า WQI คำนวณจาก DO, BOD, NO_3–N (ไนเตรท-ไนโตรเจน), TP (Total Phosphorus), FCB และ SS (Suspended Solids) พบว่า แบบจำลองที่มีประสิทธิภาพดีที่สุดสำหรับการทำนายค่า WQI ของแต่ละสถานีของแม่น้ำทั้ง 5 สายนั้น ไม่เหมือนกัน และพบว่าค่าเฉลี่ย WQI ของแม่น้ำลำชีและแม่น้ำเลยในอีก 5 ปีข้างหน้า มีแนวโน้มลดลง อย่างไรก็ตาม งานวิจัยฉบับนี้ เสนอให้ลองใช้แบบจำลองอนุกรมเวลาอื่นๆ เพิ่มเติม เพื่อเปรียบเทียบเทียบประสิทธิภาพและหาแบบจำลองที่เหมาะสมที่สุด

9. Dastorani et al. (2020) ใช้แบบจำลองอนุกรมเวลา 5 แบบจำลอง ประกอบด้วย Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA) และ Seasonal Auto-Regressive Integrated Moving Average (SARIMA) ซึ่งมี 12 รูปแบบ ได้แก่ AR(1), AR(2), MA(1), MA(2), ARMA(1,1), ARMA(1,2), ARMA(2,1), ARMA(2,2), ARIMA(1,1,2), ARIMA(1,2,1), SARIMA$(1,1,0)(1,1,1)_{12}$ และ SARIMA$(1,1,1)(1,1,1)_{12}$ ทำนายค่าพารามิเตอร์น้ำ ได้แก่ Ca, Bicarbonate (HCO_3), Sulfate (SO_4), Electrical conductivity (Ec), pH, Mg, Cl, Na และ TDS โดยใช้ข้อมูลคุณภาพน้ำผิวดินบริเวณ Harmaleh ประเทศอิหร่าน ตั้งแต่ ค.ศ. 2001 ถึง 2014 พบว่า เมื่อใช้แบบจำลองอนุกรมเวลาทั้ง 5 ทำนายค่าพารามิเตอร์น้ำทั้ง 9 พารามิเตอร์ ARMA เป็นแบบจำลองที่สามารถทำนายค่าพารามิเตอร์น้ำได้แม่นยำที่สุด 6 พารามิเตอร์ (Cl, Ec, HCO_3, Mg, Na และ pH) จาก 9 พารามิเตอร์

10. Fashae et al. (2019) ทำนายอัตราการไหลของแม่น้ำ Opeki ประเทศไนจีเรีย โดยใช้ข้อมูลอัตราการไหลของแม่น้ำรายเดือนระหว่างปี ค.ศ. 1980 – 2010 (28 ปี) ด้วยแบบจำลอง Artificial Neural Network (ANN) และ ARIMA พบว่า แบบจำลอง ARIMA ได้ค่า Correlation coefficient (r) เท่ากับ 0.97 และ RMSE เท่ากับ 0.57 สำหรับแบบจำลอง ANN ค่า Correlation coefficient (r) เท่ากับ 0.93 และ RMSE เท่ากับ 15.06 ซึ่งสามารถสรุปได้ว่า ARIMA มีประสิทธิภาพการทำนายอัตราการไหลของแม่น้ำ Opeki ดีกว่า ANN

บทที่ 3
วิธีดำเนินการวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอน ดังนี้

1. การกำหนดประชากรและการสุ่มตัวอย่าง
2. การสร้างเครื่องมือที่ใช้ในการวิจัย
3. การเก็บรวบรวมข้อมูล
4. การจัดกระทำและการวิเคราะห์ข้อมูล

การกำหนดกลุ่มประชากรและการสุ่มตัวอย่าง

ประชากร

การวิจัยนี้ใช้ชุดข้อมูลจากระบบฐานข้อมูลสารสนเทศสาธารณะของกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 จำนวนทั้งสิ้น 2,736 ตัวอย่าง และ 11 คอลัมน์ ได้แก่ สถานีตรวจวัด จังหวัด วันที่ตรวจวัด แม่น้ำ ค่า WQI ค่า DO ค่า BOD ค่า TCB ค่า FCB ค่า NH_3-N และระดับคุณภาพน้ำ

ค่าคุณภาพแม่น้ำมาจากสถานีตรวจวัดคุณภาพน้ำจำนวนทั้งสิ้น 64 สถานี ดังภาพประกอบ 7 ดังนี้

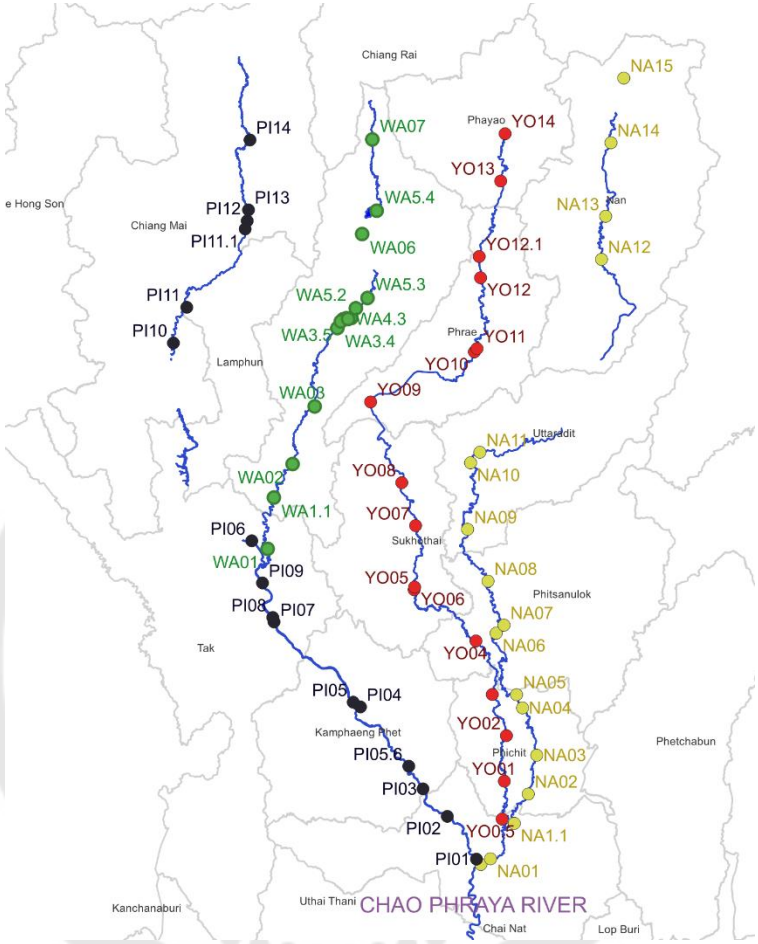
1. แม่น้ำปิง 16 สถานี ครอบคลุมพื้นที่จังหวัดนครสวรรค์ 2 สถานี (PI01 และ PI02) กำแพงเพชร 4 สถานี (PI03, PI04, PI05 และ PI05.6) ตาก 4 สถานี (PI06, PI07, PI08 และ PI09) และเชียงใหม่ 6 สถานี (PI10, PI11, PI12, PI13, PI14 และ PI11.1)

2. แม่น้ำวัง 15 สถานี ครอบคลุมพื้นที่จังหวัดลำปาง 14 สถานี (WA1.1, WA02, WA03, WA3.4, WA3.5, WA04, WA4.1, WA4.3, WA5.1, WA5.2, WA5.3, WA5.4, WA06 และ WA07) และตาก 1 สถานี (WA01)

3. แม่น้ำยม 16 สถานี ครอบคลุมพื้นที่จังหวัดนครสวรรค์ 1 สถานี (YO0.5) พิจิตร 3 สถานี (YO01, YO02 และ YO03) พิษณุโลก 1 สถานี (YO04) สุโขทัย 4 สถานี (YO05, YO06, YO07 และ YO08) แพร่ 5 สถานี (YO09, YO10, YO11, YO12 และ YO12.1) และพะเยา 2 สถานี (YO13 และ YO14)

4. แม่น้ำน่าน 17 สถานี ครอบคลุมพื้นที่จังหวัดนครสวรรค์ 3 สถานี (NA0.1, NA01 และ NA1.1) พิจิตร 5 สถานี (NA02, NA03, NA04 และ NA05) พิษณุโลก 3 สถานี (NA06,

NA07 และ NA08) อุตรดิตถ์ 3 สถานี (NA09, NA10 และ NA11) และน่าน 4 สถานี (NA12, NA13, NA14 และ NA15)



ภาพประกอบ 7 สถานีตรวจวัดคุณภาพน้ำ

การเลือกกลุ่มตัวอย่าง

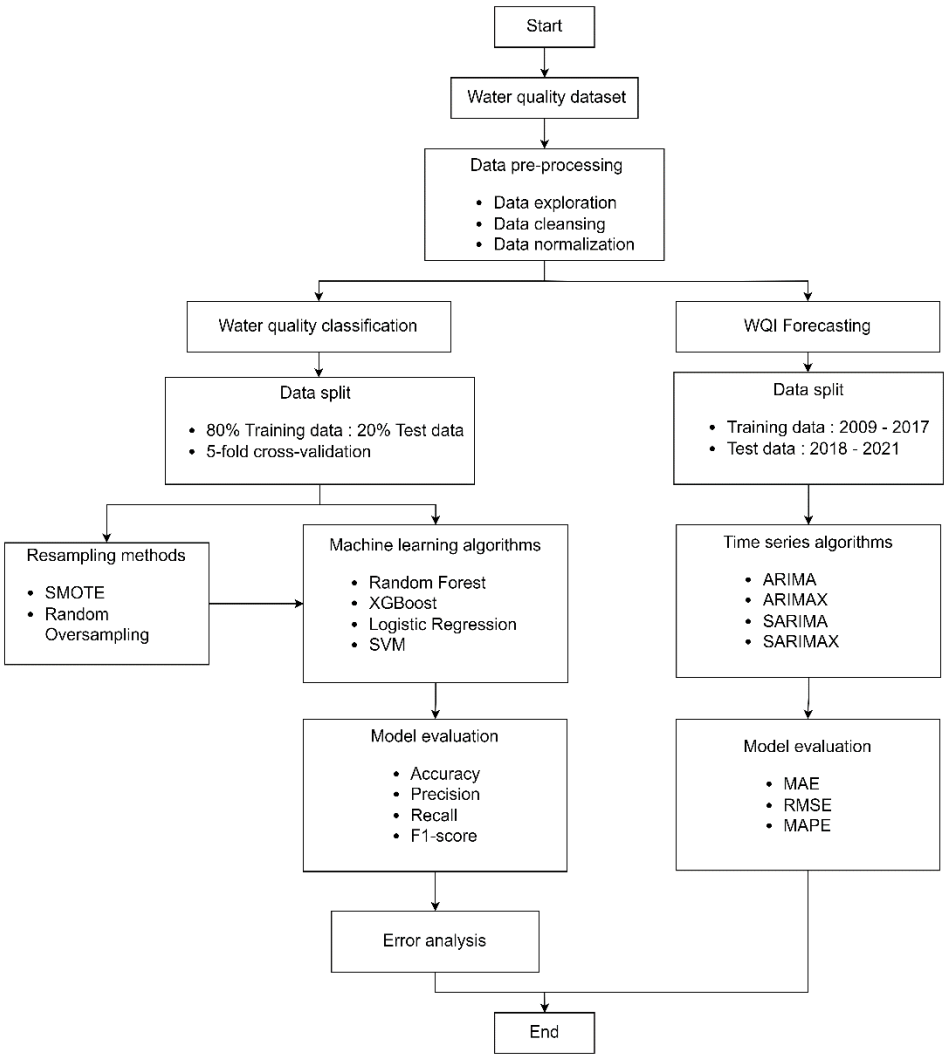
แบ่งชุดข้อมูลเป็นจำนวน 2 ชุด ประกอบด้วย ข้อมูลสำหรับเรียนรู้ (Training data) และข้อมูลสำหรับทดสอบ (Test data) โดยแบ่งในอัตราส่วน 80 ต่อ 20 เพื่อใช้จำแนกระดับคุณภาพน้ำ และสำหรับการทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) กำหนดให้ข้อมูลระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2560 เป็นข้อมูลชุดฝึกฝน และข้อมูลระหว่างปี พ.ศ 2561 ถึง พ.ศ. 2564 ใช้เป็นข้อมูลชุดทดสอบ

การสร้างเครื่องมือที่ใช้ในการวิจัย

การวิจัยนี้สร้างแบบจำลองเพื่อจำแนกและทำนายข้อมูลคุณภาพแม่น้ำด้วยการเขียนโปรแกรมภาษา Python ซึ่งกระบวนการสร้างแบบจำลองดังภาพประกอบ 8 เริ่มจากการนำเข้าข้อมูล ทำความสะอาดข้อมูล ได้แก่ การลบข้อมูลว่าง เปลี่ยนชนิดของข้อมูลให้เหมาะสมต่อการนำไปวิเคราะห์ และปรับขนาดของข้อมูลให้อยู่ในช่วง (Scale) เดียวกัน จากนั้นแสดงผลข้อมูลที่ได้หลังจากการทำความสะอาดข้อมูลเบื้องต้น เพื่อดูแนวโน้ม การกระจายตัวของข้อมูล และความสัมพันธ์ของข้อมูล ขั้นตอนถัดมานำข้อมูลเข้าแบบจำลอง โดยงานวิจัยนี้ทำการจำแนกประเภทของระดับคุณภาพน้ำและทำนายดัชนีชี้วัดคุณภาพน้ำ จึงใช้เทคนิคสำหรับการเรียนรู้ของเครื่อง 2 ประเภท ดังนี้

1. การจำแนกประเภท (Classification) แบ่งข้อมูล 2 ส่วนก่อนเข้าแบบจำลอง ได้แก่ ข้อมูลชุดฝึกฝน (Training data) และข้อมูลชุดทดสอบ (Test data) ในอัตราส่วน 80 ต่อ 20 และนำข้อมูลเข้าแบบจำลอง ได้แก่ Random Forest, XGBoost, Logistic Regression และ SVM โดยใช้เทคนิคการแก้ไขปัญหาข้อมูลไม่สมดุล 2 เทคนิค ได้แก่ SMOTE และ Random Oversampling นอกจากนี้ ได้ใช้เทคนิคการสุ่มเลือก Hyper-parameter ของแบบจำลอง (Randomized search) กับข้อมูลชุดฝึกฝน เพื่อหาพารามิเตอร์ที่ทำให้ผลการทำนายของแบบจำลองมีประสิทธิภาพมากที่สุด จากนั้นทำการประเมินประสิทธิภาพของแบบจำลอง เปรียบเทียบผลที่ได้ระหว่างแบบจำลอง และวิเคราะห์ความผิดพลาดที่เกิดจากการทำนายที่ได้จากแบบจำลอง

2. อนุกรมเวลา (Time series) แบ่งข้อมูลชุดฝึกฝน (Training data) เป็นข้อมูลระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2560 (9 ปี) และข้อมูลชุดทดสอบ (Test data) ข้อมูลระหว่างปี พ.ศ. 2561 ถึง พ.ศ. 2564 (4 ปี) จากนั้นนำข้อมูลเข้าแบบจำลองอนุกรมเวลา ได้แก่ ARIMA, ARIMAX, SARIMA และ SARIMAX และนำผลการทำนายมาเปรียบเทียบกับข้อมูลชุดทดสอบเพื่อประเมินประสิทธิภาพของแบบจำลอง



ภาพประกอบ 8 ขั้นตอนการดำเนินการวิจัย

การเก็บรวบรวมข้อมูล

งานวิจัยนี้ใช้ชุดข้อมูลจากระบบฐานข้อมูลของกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 จำนวนทั้งสิ้น 2,736 ตัวอย่าง โดยเก็บข้อมูล 4 ครั้งต่อปี แบ่งได้เป็น 2 ช่วง ประกอบด้วยช่วงฤดูแล้ง (ปริมาณน้ำน้อย) ได้แก่ ครั้งที่ 1 เดือนมกราคมถึงเดือนมีนาคม ครั้งที่ 2 เดือนเมษายนถึงเดือนมิถุนายน และช่วงฤดูฝน (ปริมาณน้ำเยอะ) ได้แก่ ครั้งที่ 3 เดือนกรกฎาคมถึงเดือนกันยายน และครั้งที่ 4 เดือนตุลาคมถึงเดือนธันวาคม (กรมควบคุมมลพิษ, 2561)

ข้อมูลคุณภาพน้ำประกอบด้วย 11 คอลัมน์ ซึ่งระบุข้อมูลเกี่ยวกับสถานีตรวจวัด จังหวัด วันที่ตรวจวัด แม่น้ำ ค่า WQI ค่า DO ค่า BOD ค่า TCB ค่า FCB ค่า NH_3-N และระดับคุณภาพน้ำ โดยแสดงรายละเอียดของข้อมูลคุณภาพน้ำดังตาราง 6

ตาราง 6 รายละเอียดของข้อมูลคุณภาพน้ำ

ลำดับ	ตัวแปร	คำอธิบาย
1	Station	สถานีตรวจวัด
2	Province	จังหวัด
3	River	แม่น้ำ
4	Date	วันที่ตรวจวัด
5	WQI	ค่าดัชนีชี้วัดคุณภาพแม่น้ำ
6	DO (mg/l)	ค่า DO
7	BOD (mg/l)	ค่า BOD
8	Total Coli (MPN/100ml)	ค่า TCB
9	Fecal Coli (MPN/100ml)	ค่า FCB
10	NH3-N (mg/l)	ค่า NH_3-N
11	WQ Class	ระดับคุณภาพน้ำ

การจัดกระทำและการวิเคราะห์ข้อมูล

1. เลือกข้อมูลที่เกี่ยวข้องสำหรับใช้วิเคราะห์ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น่าน ได้แก่ สถานีตรวจวัด (Station) จังหวัด (Province) แม่น้ำ (River) วันที่ตรวจวัด (Date) ค่า DO (DO (mg/l)) ค่า BOD (BOD (mg/l)) ค่า TCB (Total Coli (MPN/100ml)) ค่า FCB (Fecal Coli (MPN/100ml)) ค่า NH_3-N (NH3-N (mg/l)) และค่า WQI ดังภาพประกอบ 9 แสดงให้เห็นว่า มีข้อมูลทั้งสิ้น 2,736 แถว และ 10 คอลัมน์

	Station	Province	River	Date	DO(mg/l)	BOD(mg/l)	Total Coli(MPN/100ml)	Fecal Coli(MPN/100ml)	NH3-N(mg/l)	WQI
0	NA01	Nakhon Sawan	Nan	2021-11-15	3.4	2.0	1100.0	700.0	0.07	66.0
1	NA02	Phichit	Nan	2021-09-11	4.0	2.1	1700.0	1300.0	0.10	59.0
2	NA03	Phichit	Nan	2021-09-11	4.2	2.0	1700.0	700.0	0.07	67.0
3	NA04	Phichit	Nan	2021-09-11	4.0	1.5	2800.0	2200.0	0.10	64.0
4	NA05	Phichit	Nan	2021-09-11	4.8	1.7	490.0	230.0	0.07	73.0
...	...	...	...	...	...	...	...	...	...	...
2731	WA02	Lampang	Wang	2009-02-18	6.3	1.2	500.0	170.0	0.20	84.0
2732	WA03	Lampang	Wang	2009-02-18	6.1	2.0	800.0	40.0	0.30	69.0
2733	WA4.1	Lampang	Wang	2009-02-17	5.4	1.0	50000.0	3400.0	0.30	57.0
2734	WA5.1	Lampang	Wang	2009-02-17	5.4	1.4	1300.0	170.0	0.30	69.0
2735	WA06	Lampang	Wang	2009-02-17	6.3	0.7	30000.0	400.0	0.30	60.0

2736 rows × 10 columns

ภาพประกอบ 9 ข้อมูลที่นำมาใช้วิเคราะห์ข้อมูลคุณภาพแม่น้ำ

2. เพิ่มคอลัมน์ระดับคุณภาพน้ำ (WQ Class) ที่ได้จากการแบ่งเกณฑ์ระดับคุณภาพน้ำเป็น 5 ระดับ ได้แก่ คุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก (ค่า WQI ระหว่าง 91 - 100) คุณภาพน้ำที่อยู่ในเกณฑ์ดี (ค่า WQI ระหว่าง 71 - 90) คุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ (ค่า WQI ระหว่าง 61 - 70) คุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม (ค่า WQI ระหว่าง 31 - 60) และคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก (ค่า WQI ระหว่าง 0 - 30) ทำให้มีจำนวนข้อมูลทั้งสิ้น 2,736 แถว และ 11 คอลัมน์ ดังภาพประกอบ 10

	Station	Province	River	Date	DO(mg/l)	BOD(mg/l)	Total Coli(MPN/100ml)	Fecal Coli(MPN/100ml)	NH3-N(mg/l)	WQI	WQ Class
0	NA01	Nakhon Sawan	Nan	2021-11-15	3.4	2.0	1100.0	700.0	0.07	66.0	moderate
1	NA02	Phichit	Nan	2021-09-11	4.0	2.1	1700.0	1300.0	0.10	59.0	poor
2	NA03	Phichit	Nan	2021-09-11	4.2	2.0	1700.0	700.0	0.07	67.0	moderate
3	NA04	Phichit	Nan	2021-09-11	4.0	1.5	2800.0	2200.0	0.10	64.0	moderate
4	NA05	Phichit	Nan	2021-09-11	4.8	1.7	490.0	230.0	0.07	73.0	good
...	...	...	...	...	...	...	...	...	...	...	...
2731	WA02	Lampang	Wang	2009-02-18	6.3	1.2	500.0	170.0	0.20	84.0	good
2732	WA03	Lampang	Wang	2009-02-18	6.1	2.0	800.0	40.0	0.30	69.0	moderate
2733	WA4.1	Lampang	Wang	2009-02-17	5.4	1.0	50000.0	3400.0	0.30	57.0	poor
2734	WA5.1	Lampang	Wang	2009-02-17	5.4	1.4	1300.0	170.0	0.30	69.0	moderate
2735	WA06	Lampang	Wang	2009-02-17	6.3	0.7	30000.0	400.0	0.30	60.0	poor

2736 rows × 11 columns

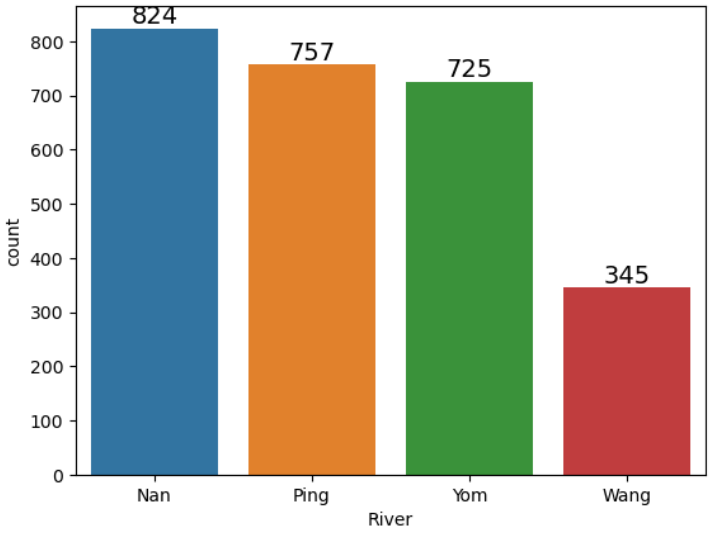
ภาพประกอบ 10 ข้อมูลที่ได้หลังจากการเพิ่มคอลัมน์ระดับคุณภาพน้ำ (WQ Class)

3. ทำความสะอาดข้อมูล ได้แก่ เปลี่ยนชนิดของข้อมูลให้เหมาะสม จัดการกับแถวที่มีค่าว่างและแถวที่มีข้อมูลซ้ำกันโดยลบข้อมูลดังกล่าวทิ้ง ทำให้ข้อมูลที่ได้หลังจากทำความสะอาดแล้วมีจำนวนข้อมูล 2,651 ตัวอย่าง 11 คอลัมน์ และมีชนิดของข้อมูลของแต่ละคอลัมน์ ดังภาพประกอบ 11

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2651 entries, 0 to 2650
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Station               2651 non-null   object
 1   Province              2651 non-null   object
 2   River                 2651 non-null   object
 3   Date                  2651 non-null   datetime64[ns]
 4   DO(mg/l)              2651 non-null   float64
 5   BOD(mg/l)             2651 non-null   float64
 6   Total Coli(MPN/100ml) 2651 non-null   float64
 7   Fecal Coli(MPN/100ml) 2651 non-null   float64
 8   NH3-N(mg/l)           2651 non-null   float64
 9   WQI                   2651 non-null   float64
 10  Class                 2651 non-null   category
dtypes: category(1), datetime64[ns](1), float64(6), object(3)
memory usage: 210.0+ KB
```

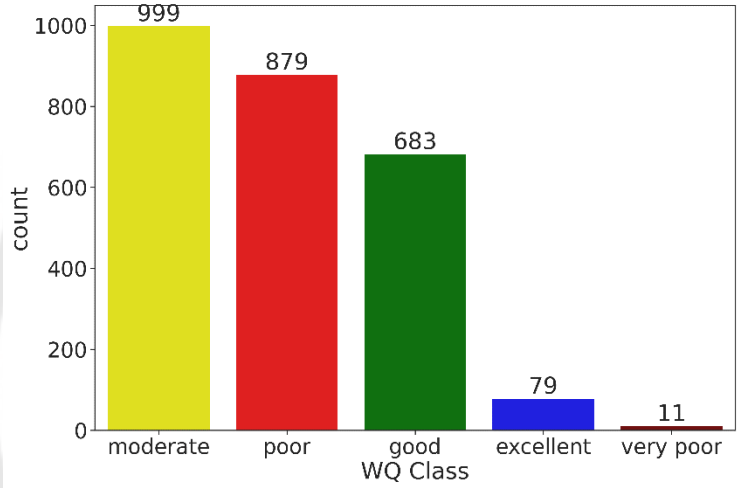
ภาพประกอบ 11 รายละเอียดข้อมูลที่ได้หลังจากทำความสะอาดข้อมูล

4. แสดงจำนวนข้อมูลของแม่น้ำแต่ละสาย (ภาพประกอบ 12) พบว่า ในข้อมูลชุดนี้มีข้อมูลคุณภาพน้ำของแม่น้ำน่านมากที่สุด เท่ากับ 824 ตัวอย่าง รองลงมา ได้แก่ แม่น้ำปิง เท่ากับ 757 ตัวอย่าง แม่น้ำยม เท่ากับ 725 ตัวอย่าง และแม่น้ำวัง จำนวน 345 ตัวอย่าง ตามลำดับ

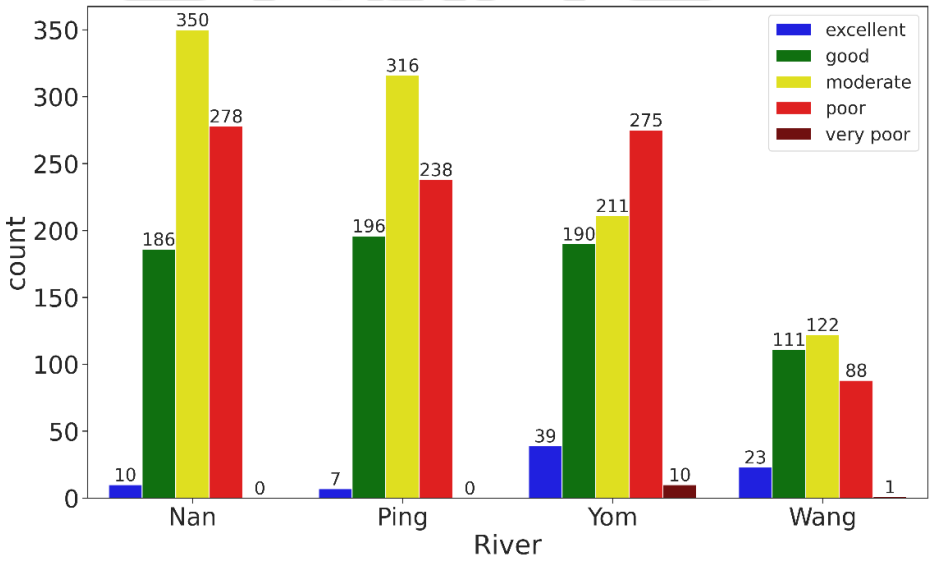


ภาพประกอบ 12 จำนวนข้อมูลของแม่น้ำแต่ละสาย

5. แสดงการกระจายตัวของระดับคุณภาพแม่น้ำทั้ง 4 สาย (ภาพประกอบ 13) พบว่าระดับคุณภาพน้ำส่วนใหญ่อยู่ในเกณฑ์พอใช้ (Moderate) จำนวน 999 ตัวอย่าง รองลงมา ได้แก่ เสื่อมโทรม (Poor) ดี (Good) ดีมาก (Excellent) และเสื่อมโทรมมาก (Very poor) เท่ากับ 879, 683, 79 และ 11 ตัวอย่าง ตามลำดับ แสดงให้เห็นว่า ระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากและเสื่อมโทรมมากมีจำนวนข้อมูลเพียง 2.98% และ 0.41% ของจำนวนตัวอย่างทั้งหมด ตามลำดับ ดังนั้น งานวิจัยนี้จึงนำเทคนิคการเพิ่มจำนวนตัวอย่างด้วย SMOTE และ Random Oversampling มาให้เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล



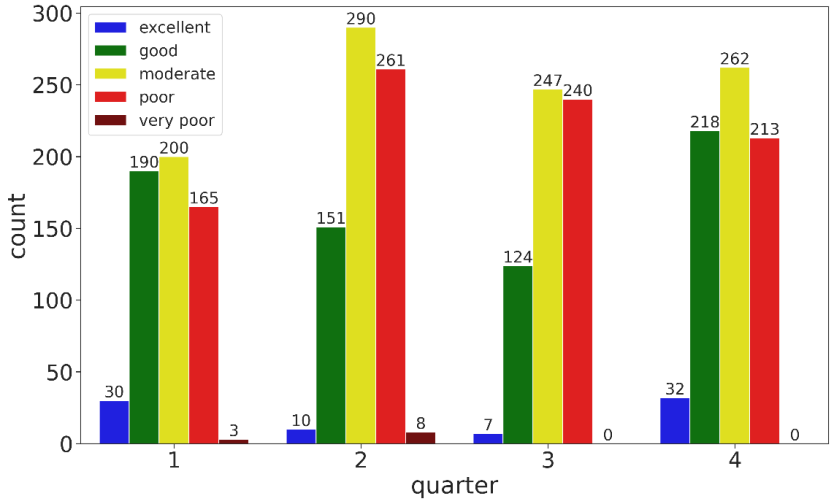
ภาพประกอบ 13 การกระจายตัวของข้อมูลระดับคุณภาพแม่น้ำ



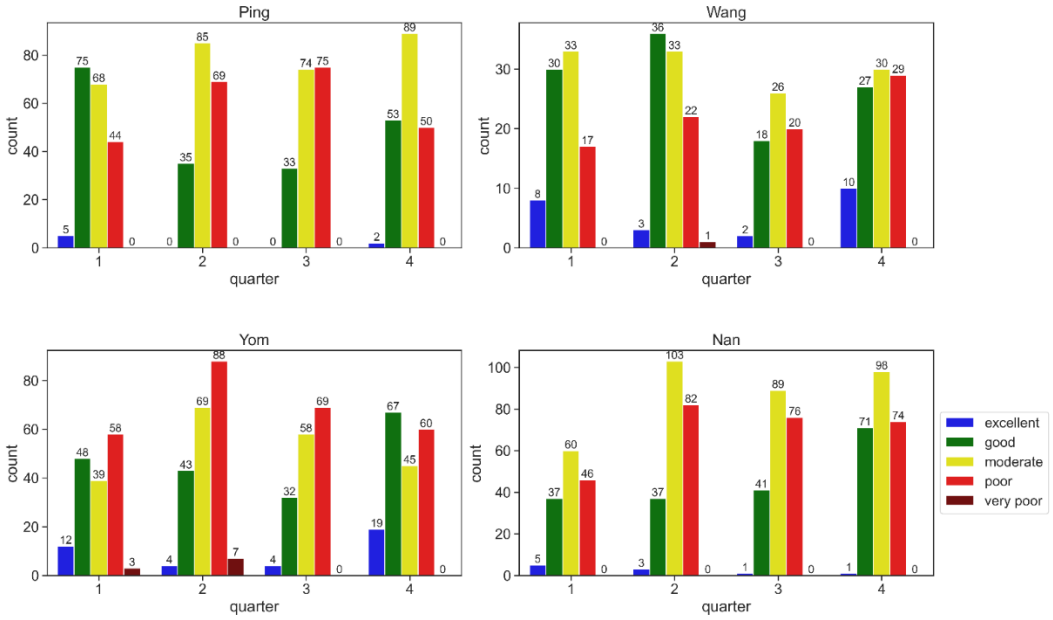
ภาพประกอบ 14 การกระจายตัวของข้อมูลระดับคุณภาพน้ำของแม่น้ำน่าน ปิง ยม และวัง

เมื่อพิจารณาระดับคุณภาพน้ำโดยแยกเป็นข้อมูลของแม่น้ำแต่ละสาย (ภาพประกอบ 14) พบว่า แม่น้ำน่าน มีระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้มากที่สุด (350 ตัวอย่าง) รองลงมา ได้แก่ เสื่อมโทรม (278 ตัวอย่าง) ดี (186 ตัวอย่าง) และดีมาก (10 ตัวอย่าง) ตามลำดับ แม่น้ำปิง มีระดับคุณภาพน้ำอยู่เกณฑ์พอใช้มากที่สุด (316 ตัวอย่าง) รองลงมา ได้แก่ เสื่อมโทรม (238 ตัวอย่าง) ดี (196 ตัวอย่าง) และดีมาก (7 ตัวอย่าง) ตามลำดับ แม่น้ำวังมีระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้มากที่สุด (122 ตัวอย่าง) รองลงมา ได้แก่ ดี (111 ตัวอย่าง) เสื่อมโทรม (88 ตัวอย่าง) ดีมาก (23 ตัวอย่าง) และเสื่อมโทรมมาก (1 ตัวอย่าง) ตามลำดับ และแม่น้ำยมมีระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมสูงที่สุด (275 ตัวอย่าง) รองลงมา ได้แก่ พอใช้ (211 ตัวอย่าง) ดี (190 ตัวอย่าง) ดีมาก (39 ตัวอย่าง) และเสื่อมโทรมมาก (10 ตัวอย่าง) ตามลำดับ โดยแม่น้ำน่านและแม่น้ำปิงไม่พบระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก จากข้อมูลระดับคุณภาพน้ำของแม่น้ำแต่ละสาย สามารถสรุปได้ว่า แม่น้ำน่าน แม่น้ำปิง และแม่น้ำวัง ระดับคุณภาพน้ำส่วนใหญ่อยู่ในเกณฑ์พอใช้ ส่วนแม่น้ำยมมีระดับคุณภาพน้ำส่วนใหญ่อยู่ในเกณฑ์เสื่อมโทรม

6. พิจารณาระดับคุณภาพน้ำโดยแบ่งตามครั้งที่เก็บตัวอย่างน้ำ ซึ่งเก็บทั้งหมด 4 ครั้งต่อปี พบว่า ระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้จำนวนมากที่สุดทั้ง 4 ครั้ง โดยข้อมูลคุณภาพน้ำที่เก็บครั้งที่ 1 (เดือนมกราคมถึงเดือนมีนาคม) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้มากที่สุด เท่ากับ 200 ตัวอย่าง ลำดับถัดมา ได้แก่ ดี (190 ตัวอย่าง) เสื่อมโทรม (165 ตัวอย่าง) ดีมาก (30 ตัวอย่าง) และเสื่อมโทรมมาก (3 ตัวอย่าง) ตามลำดับ ข้อมูลคุณภาพน้ำที่เก็บครั้งที่ 2 (เดือนเมษายนถึงเดือนมิถุนายน) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้มากที่สุด เท่ากับ 290 ตัวอย่าง ลำดับถัดมา ได้แก่ เสื่อมโทรม (261 ตัวอย่าง) ดี (151 ตัวอย่าง) ดีมาก (10 ตัวอย่าง) และเสื่อมโทรมมาก (8 ตัวอย่าง) ตามลำดับ ข้อมูลคุณภาพน้ำที่เก็บครั้งที่ 3 (เดือนกรกฎาคมถึงเดือนกันยายน) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้มากที่สุด เท่ากับ 247 ตัวอย่าง ลำดับถัดมาได้แก่ เสื่อมโทรม (240 ตัวอย่าง) ดี (124 ตัวอย่าง) และดีมาก (7 ตัวอย่าง) ตามลำดับ และข้อมูลคุณภาพน้ำที่เก็บครั้งที่ 4 (เดือนตุลาคมถึงเดือนธันวาคม) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้มากที่สุด เท่ากับ 262 ตัวอย่าง ลำดับถัดมา ได้แก่ ดี (218 ตัวอย่าง) เสื่อมโทรม (213 ตัวอย่าง) และดีมาก (32 ตัวอย่าง) ตามลำดับ (ภาพประกอบ 15)



ภาพประกอบ 15 ระดับคุณภาพน้ำแบ่งตามช่วงเวลาการเก็บตัวอย่างน้ำ



ภาพประกอบ 16 ข้อมูลระดับคุณภาพน้ำแบ่งตามช่วงเวลาการเก็บตัวอย่างน้ำของแม่น้ำแต่ละสาย

จากภาพประกอบ 16 แสดงข้อมูลระดับคุณภาพน้ำที่อยู่ในเกณฑ์ต่างๆ แบ่งตามช่วงเวลาการเก็บตัวอย่างน้ำของแม่น้ำแต่ละสาย พบว่า แม่น้ำปิง มีเพียงช่วงเวลาการเก็บตัวอย่างครั้งที่ 1 ที่ระดับคุณภาพน้ำอยู่ในเกณฑ์ดีสูงที่สุด สำหรับการเก็บตัวอย่างครั้งที่ 2 - 4 ระดับคุณภาพน้ำส่วนใหญ่อยู่ในเกณฑ์พอใช้ และแม่น้ำวัง ระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีและพอใช้มีสัดส่วนสูง

ในช่วงเวลาการเก็บครั้งที่ 1 และ 2 เมื่อพิจารณาแม่น้ำน่าน ระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้มีจำนวนมากที่สุดในทุกช่วงเวลาการเก็บตัวอย่าง อย่างไรก็ตาม ระดับคุณภาพน้ำส่วนใหญ่ของแม่น้ำยมอยู่ในเกณฑ์เสื่อมโทรม 3 ช่วงเวลาการเก็บตัวอย่าง ได้แก่ ครั้งที่ 1, 2 และ 3 แต่ช่วงเวลาการเก็บตัวอย่างคุณภาพน้ำครั้งที่ 4 ระดับคุณภาพน้ำอยู่ในเกณฑ์ดีสูงที่สุด

7. วิเคราะห์ค่าสถิติเบื้องต้นของค่า DO, BOD, TCB, FCB, NH_3-N และ WQI ในแต่ละเกณฑ์ของระดับคุณภาพน้ำ ดังแสดงในตาราง 7

ตาราง 7 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของพารามิเตอร์น้ำทั้ง 5 พารามิเตอร์ในแต่ละเกณฑ์ของระดับคุณภาพน้ำ

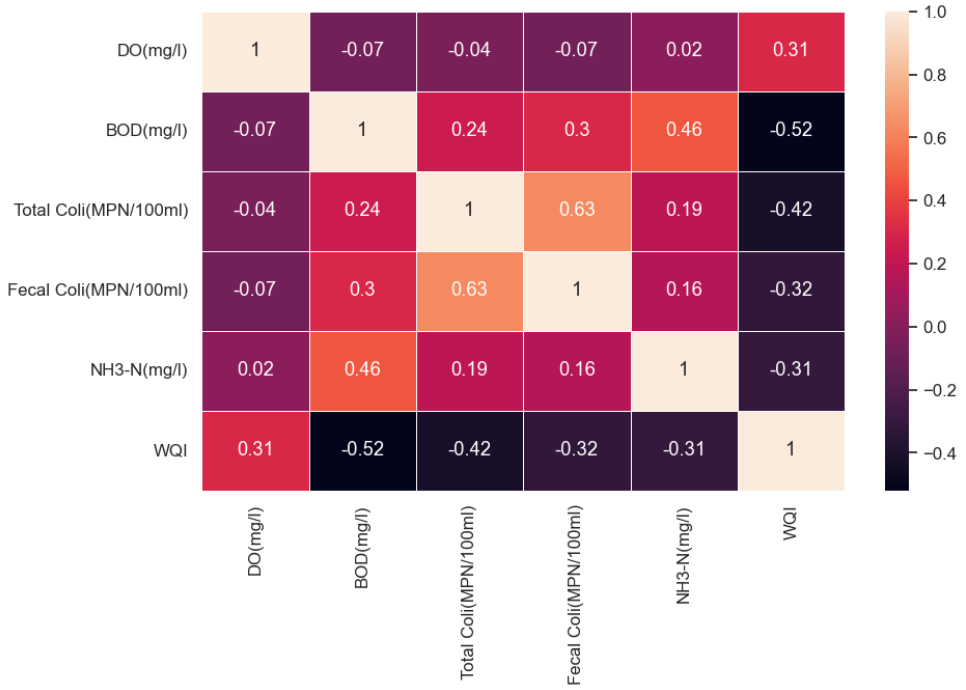
พารามิเตอร์น้ำ	ค่าสถิติ	เกณฑ์คุณภาพน้ำ				
		ดีมาก	ดี	พอใช้	เสื่อมโทรม	เสื่อมโทรมมาก
DO (mg/l)	ค่าเฉลี่ย	7.63	6.67	6.13	5.71	5.64
	S.D.	0.57	1.24	1.29	1.42	6.64
BOD (mg/l)	ค่าเฉลี่ย	0.91	1.20	1.54	2.54	16.94
	S.D.	0.32	0.40	0.61	1.54	11.92
TCB (MPN/100ml)	ค่าเฉลี่ย	641.62	2,552.71	7,339.94	25,442.06	118,272.73
	S.D.	784.51	4,498.05	9,282.26	36,463.25	52,195.96
FCB (MPN/100ml)	ค่าเฉลี่ย	91.19	492.29	1,403.33	7,482.19	68,172.73
	S.D.	75.82	1,244.17	1,719.01	16,566.70	64,607.87
NH_3-N (mg/l)	ค่าเฉลี่ย	0.02	0.12	0.19	0.34	4.37
	S.D.	0.03	0.11	0.15	0.77	4.37
WQI	ค่าเฉลี่ย	92.68	79.68	65.55	55.90	24.82
	S.D.	1.70	6.23	2.84	4.51	3.43

8. ศึกษาความสัมพันธ์ของข้อมูลระหว่างค่า WQI กับลักษณะเฉพาะต่างๆ (Feature) ได้แก่ ค่า DO, BOD, TCB, FCB และ NH_3-N โดยแสดงในรูป Pearson Correlation Coefficient ซึ่งมีค่าอยู่ระหว่าง -1.00 ถึง +1.00 หากค่า Correlation Coefficient มีค่าเข้าใกล้ +1.00 แสดงว่าค่า WQI และลักษณะเฉพาะนั้นมีความสัมพันธ์ในทิศทางเดียวกัน กรณี Correlation Coefficient มีค่าเข้าใกล้ -1.00 แสดงว่า ค่า WQI และลักษณะเฉพาะนั้นมีความสัมพันธ์ในทิศทางตรงข้าม

และกรณี Correlation Coefficient มีค่าเท่ากับ 0 หมายความว่าทั้ง 2 ตัวแปรดังกล่าว ไม่มีความสัมพันธ์กัน (Schober, Boer, & Schwarte, 2018)

	DO(mg/l)	BOD(mg/l)	Total Coli(MPN/100ml)	Fecal Coli(MPN/100ml)	NH3-N(mg/l)	WQI
DO(mg/l)	1.00	-0.07	-0.04	-0.07	0.02	0.31
BOD(mg/l)	-0.07	1.00	0.24	0.30	0.46	-0.52
Total Coli(MPN/100ml)	-0.04	0.24	1.00	0.63	0.19	-0.42
Fecal Coli(MPN/100ml)	-0.07	0.30	0.63	1.00	0.16	-0.32
NH3-N(mg/l)	0.02	0.46	0.19	0.16	1.00	-0.31
WQI	0.31	-0.52	-0.42	-0.32	-0.31	1.00

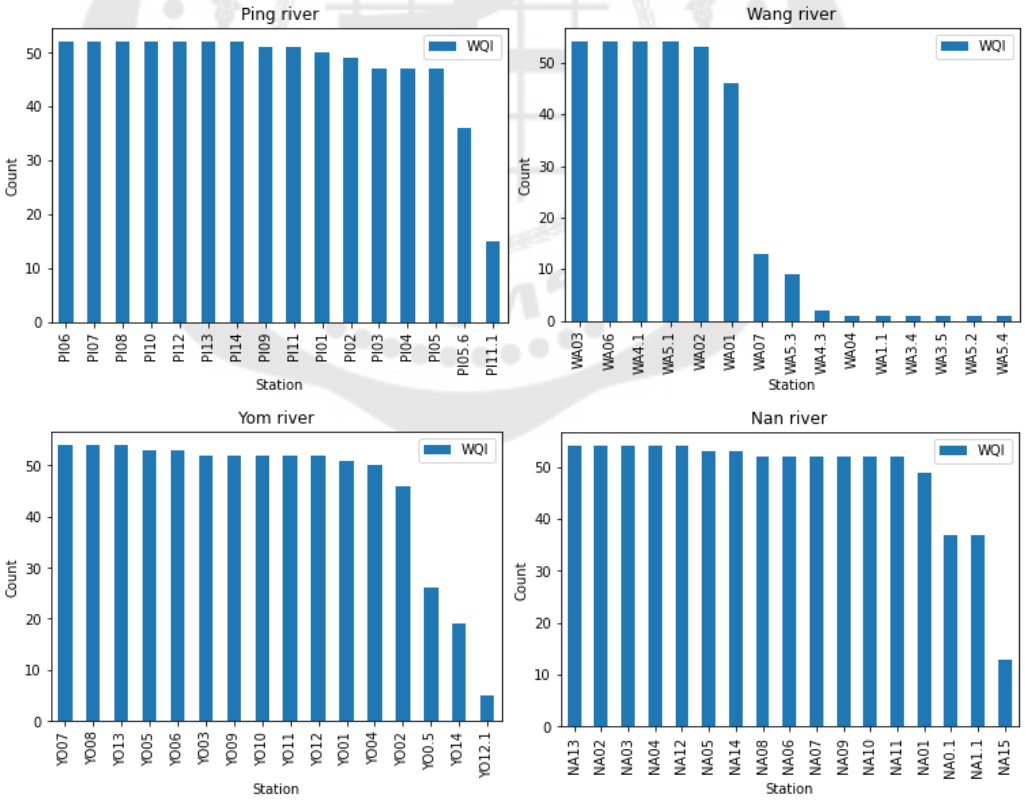
ภาพประกอบ 17 แสดงค่าความสัมพันธ์ระหว่างค่า WQI กับลักษณะเฉพาะ ได้แก่ ค่า DO, BOD, TCB, FCB และ NH_3-N



ภาพประกอบ 18 แสดงค่าความสัมพันธ์ระหว่างค่า WQI กับลักษณะเฉพาะ (Feature) ได้แก่ ค่า DO, BOD, TCB, FCB และ NH_3-N ในรูปแบบ Heatmap

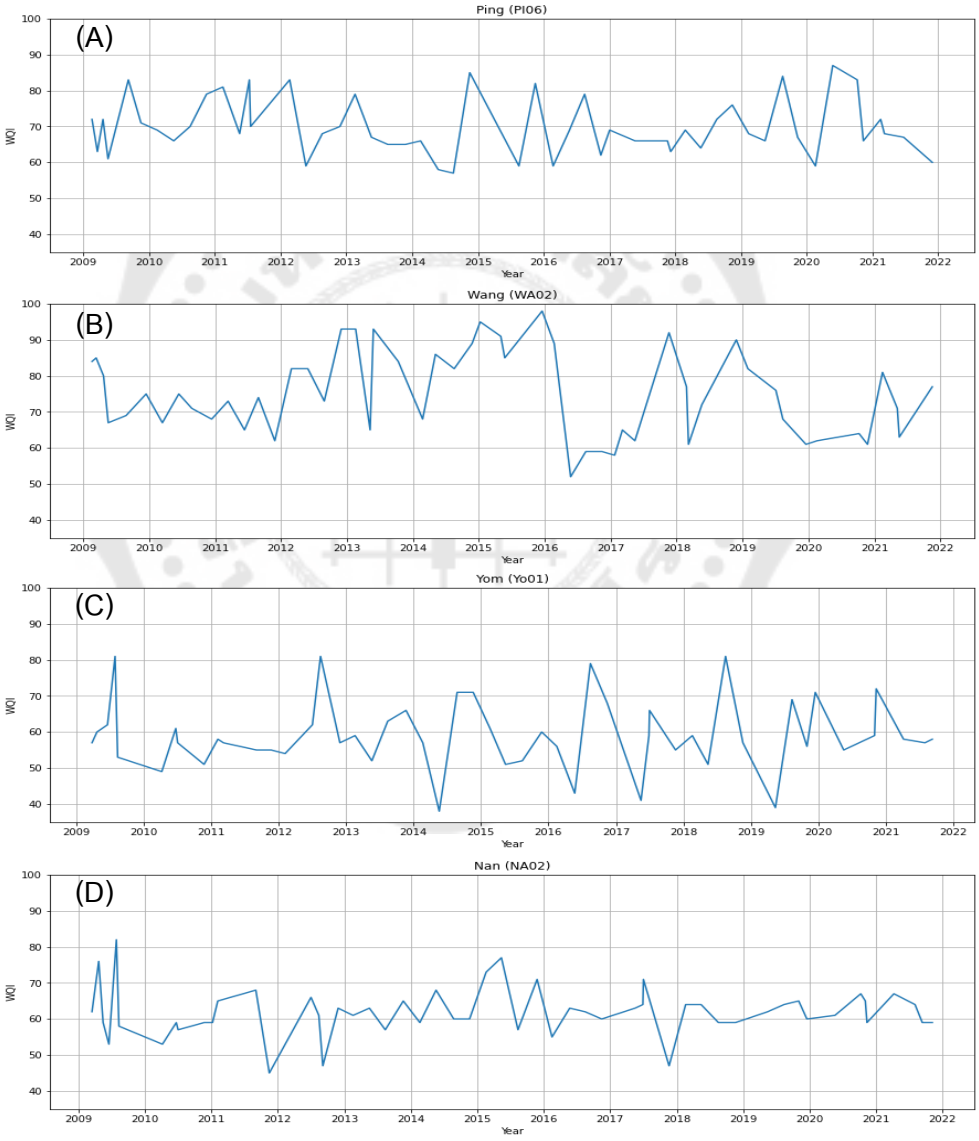
จากภาพประกอบ 17 และ 18 แสดงให้เห็นว่า ค่า WQI มีความสัมพันธ์เชิงบวกกับค่า DO ในทางกลับกัน ค่า WQI มีความสัมพันธ์เชิงลบกับค่า BOD มากที่สุด รองลงมา ได้แก่ TCB, FCB และ NH_3-N ตามลำดับ

9. พิจารณาจำนวนข้อมูลคุณภาพน้ำในแต่ละสถานีของแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน พบว่า สถานีตรวจวัดคุณภาพแม่น้ำปิงที่มีจำนวนข้อมูลมากที่สุด ได้แก่ สถานี PI06, PI07, PI08, PI10, PI12, PI13 และ PI14 จำนวน 52 ตัวอย่าง ลำดับถัดมา ได้แก่ PI09 และ PI11 จำนวน 51 ตัวอย่าง และสถานี PI01 จำนวน 50 ตัวอย่าง สถานีตรวจวัดคุณภาพแม่น้ำวังที่มีจำนวนข้อมูลมากที่สุด ได้แก่ WA03, WA06, WA4.1 และ WA5.1 จำนวน 54 ตัวอย่าง ลำดับถัดมา ได้แก่ WA02 จำนวน 53 ตัวอย่าง และสถานี WA01 จำนวน 46 ตัวอย่าง สำหรับสถานีตรวจวัดคุณภาพแม่น้ำยมที่มีจำนวนข้อมูลมากที่สุด ได้แก่ YO07, YO08 และ YO13 จำนวน 54 ตัวอย่าง ลำดับถัดมา ได้แก่ สถานี YO05 และ YO06 จำนวน 53 ตัวอย่าง และสถานี YO03, YO09, YO10, YO11 และ YO12 จำนวน 52 ตัวอย่าง และสถานีตรวจวัดคุณภาพแม่น้ำน่านที่มีจำนวนข้อมูลมากที่สุด ได้แก่ NA13, NA02, NA03, NA04 และ NA12 จำนวน 54 ตัวอย่าง ลำดับถัดมา ได้แก่ สถานี NA05 และ NA14 จำนวน 53 ตัวอย่าง และสถานี NA08, NA06, NA07, NA09, NA10 และ NA11 จำนวน 52 ตัวอย่าง (ภาพประกอบ 19)



ภาพประกอบ 19 จำนวนข้อมูลคุณภาพน้ำในแต่ละสถานีของแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน

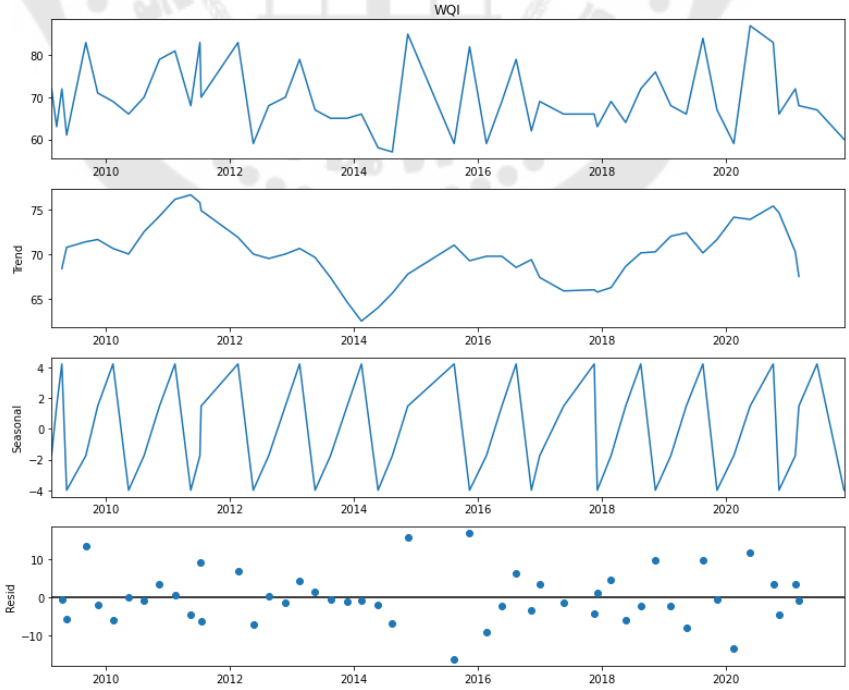
10. เลือกข้อมูลคุณภาพน้ำจากสถานีตรวจวัดคุณภาพแม่น้ำแต่ละสาย เพื่อใช้เป็นข้อมูลเข้าแบบจำลองอนุกรมเวลา โดยพิจารณาจากสถานีที่อยู่ก่อนจุดรวมกันของแม่น้ำ 2 สาย (ภาพประกอบ 7) ร่วมกับการพิจารณาจำนวนข้อมูลคุณภาพน้ำในสถานีนั้นๆ โดยแม่น้ำปิงเลือกใช้ข้อมูลจากสถานี PI06 ซึ่งมีข้อมูล 52 ตัวอย่าง แม่น้ำวัง สถานี WA02 มีจำนวนข้อมูล 53 ตัวอย่าง แม่น้ำยม สถานี YO01 มีจำนวนข้อมูล 51 ตัวอย่าง และแม่น้ำน่าน สถานี NA02 ข้อมูลจำนวน 54 ตัวอย่าง โดยแต่ละสถานีดังกล่าว มีแนวโน้มของค่า WQI ดังภาพประกอบ 20



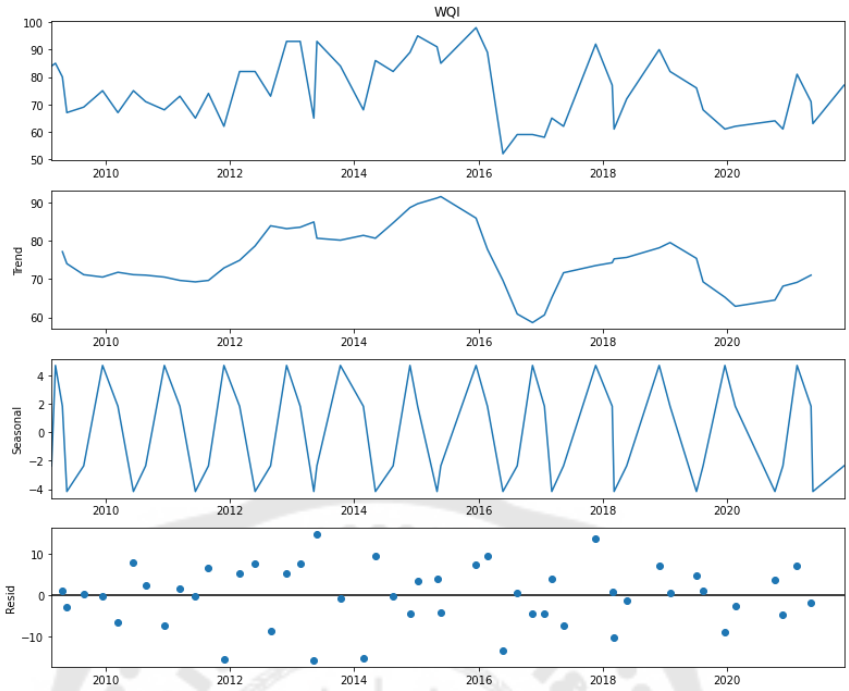
ภาพประกอบ 20 แนวโน้มค่า WQI ของสถานีตรวจวัดคุณภาพน้ำ (A) แม่น้ำปิง PI06 (B) แม่น้ำวัง WA02 (C) แม่น้ำยม YO01 และ (D) แม่น้ำน่าน NA02

จากภาพประกอบ 20 แนวโน้มค่า WQI ตั้งแต่ปี พ.ศ. 2552 (2009) ถึงปี พ.ศ. 2564 (2021) ของแม่น้ำปิง มีค่า WQI อยู่ระหว่าง 57 ถึง 87 สำหรับแม่น้ำวัง ค่า WQI อยู่ระหว่าง 52 ถึง 98 โดยค่า WQI ค่อนข้างสูง ในช่วงปี พ.ศ. 2558 (2015) ถึงปี พ.ศ. 2559 (2016) ทำให้ระดับคุณภาพน้ำอยู่ในเกณฑ์ดีถึงดีมาก และข้อมูลคุณภาพน้ำที่เก็บในครั้งที่ 2 ของปี พ.ศ. 2559 (2016) ค่า WQI มีค่าต่ำที่สุด เท่ากับ 52 เมื่อพิจารณาแม่น้ำยม ค่า WQI อยู่ระหว่าง 38 ถึง 81 ซึ่งค่า WQI มีค่าระหว่าง 31 ถึง 50 อยู่ 4 ช่วง ได้แก่ ช่วงปี พ.ศ. 2557 (2014), 2559 (2016), 2560 (2017) และ พ.ศ. 2562 (2019) ทำให้ระดับคุณภาพน้ำอยู่ในเกณฑ์เสื่อมโทรม และสำหรับแม่น้ำน่าน มีค่า WQI อยู่ระหว่าง 45 ถึง 82 โดยค่า WQI มีค่าอยู่ระหว่าง 31 - 50 ที่ทำให้ระดับคุณภาพน้ำอยู่ในเกณฑ์เสื่อมโทรม 3 ช่วง ได้แก่ ช่วงปี พ.ศ. 2554 (2011), 2555 (2012) และ พ.ศ. 2560 (2017) และค่า WQI สูงสุดใน พ.ศ. 2552 (2009)

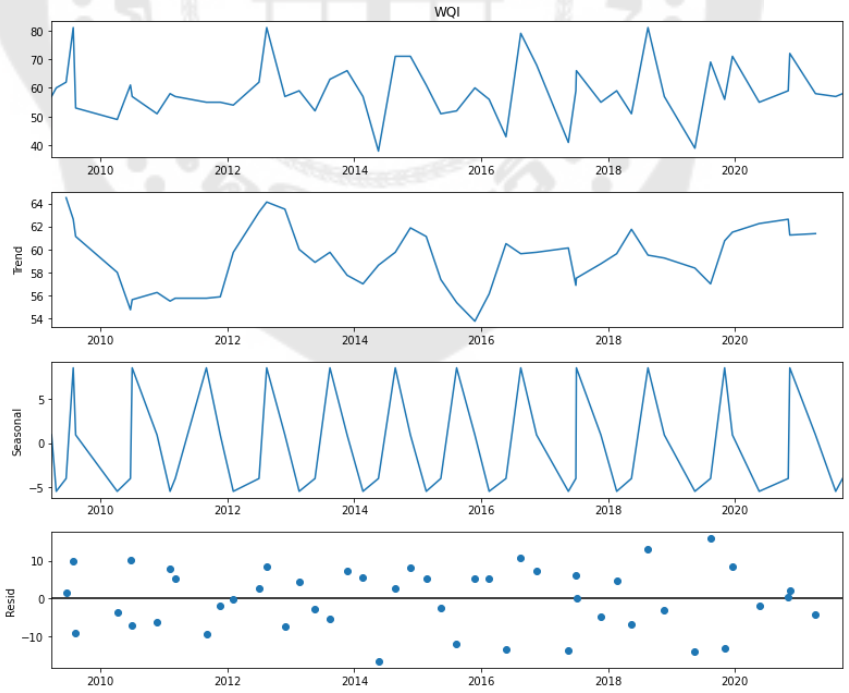
12. ศึกษาแนวโน้ม (Trend) ฤดูกาล (Seasonal) และค่าส่วนที่เหลือ (Residual) โดยใช้ฟังก์ชั่น seasonal_decompose() เพื่อแยกองค์ประกอบของข้อมูลอนุกรมเวลา ซึ่งกำหนดตัวแปร model = “additive” และกำหนด period = 4 เนื่องจากข้อมูลที่ใช้วิเคราะห์เป็นข้อมูลรายไตรมาส ซึ่งจากผลการศึกษาข้อมูลของสถานี PI06, WA02, YO01 และ NA02 พบว่า มีรูปแบบของฤดูกาล (Seasonal) แต่รูปแบบของแนวโน้มของข้อมูล (Trend) ไม่ชัดเจน ดังภาพประกอบ 21 – 24



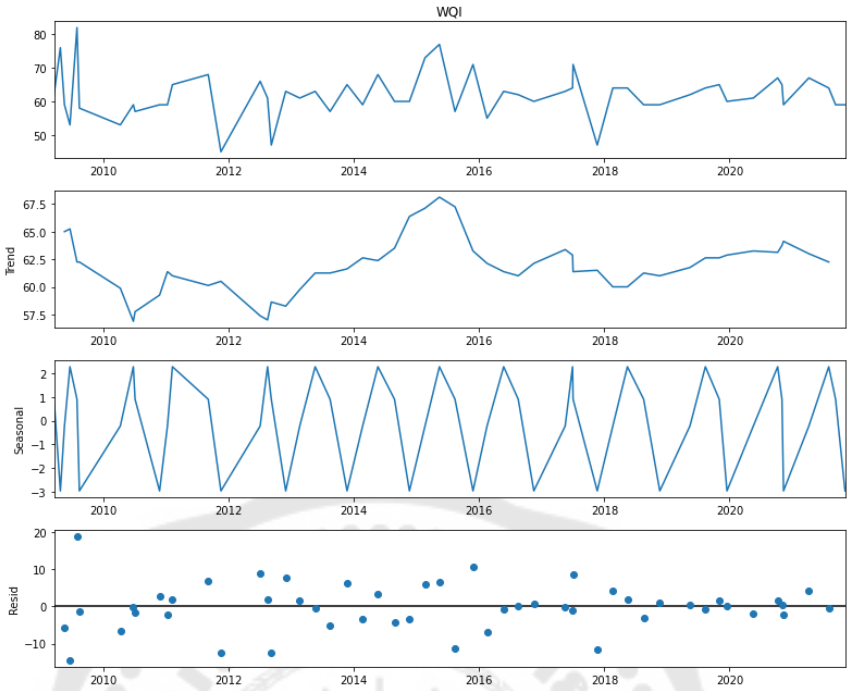
ภาพประกอบ 21 Trend, Seasonal และ Residual ของข้อมูล WQI แม่น้ำปิง สถานี PI06



ภาพประกอบ 22 Trend, Seasonal และ Residual ของข้อมูล WQI แม่น้ำวัง สถานี WA02



ภาพประกอบ 23 Trend, Seasonal และ Residual ของข้อมูล WQI แม่น้ำยม สถานี YO01



ภาพประกอบ 24 Trend, Seasonal และ Residual ของข้อมูล WQI แม่น้ำน่าน สถานี NA02

13. วิเคราะห์ Unit root test เพื่อตรวจสอบความนิ่งของข้อมูล (Stationary) ด้วย ADF test (Dickey & Fuller, 1981) และ KPSS test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992)

ตาราง 8 ผลการวิเคราะห์ ADF test และ KPSS test เพื่อตรวจสอบความนิ่งของข้อมูล (Stationary)

สถานีตรวจวัด	ADF test					KPSS test				
	Test Statistic	p-value	Critical Value (1%)	Critical Value (5%)	Stationary	Test Statistic	p-value	Critical Value (1%)	Critical Value (5%)	Stationary
แม่น้ำปิง สถานี PI06	-8.51	0.00	-3.57	-2.92	Stationary	0.12	0.10	0.74	0.46	Stationary
แม่น้ำวัง สถานี WA02	-2.94	0.04	-3.57	-2.92	Stationary	0.19	0.10	0.74	0.46	Stationary
แม่น้ำยม สถานี YO01	-2.37	0.15	-3.59	-2.93	Non-stationary	0.29	0.10	0.74	0.46	Stationary
แม่น้ำน่าน สถานี NA02	-9.08	0.00	-3.56	-2.92	Stationary	0.10	0.10	0.74	0.46	Stationary

จากผลการทดสอบด้วยวิธี ADF test (ตาราง 8) กับข้อมูล WQI ของสถานีตรวจวัดคุณภาพน้ำ PI06, WA02, YO01 และ NA02 ซึ่งเป็นตัวแทนของแม่น้ำปิง วัง ยม และน่าน ตามลำดับ พบว่า ข้อมูลจากสถานี PI06 (แม่น้ำปิง) WA02 (แม่น้ำวัง) และ NA02 (แม่น้ำน่าน) มีลักษณะนิ่ง (Stationary) เนื่องจากค่าทดสอบทางสถิติน้อยกว่าค่าวิกฤติที่ระดับนัยสำคัญ 0.05 (Critical Value (5%)) และค่า p-value น้อยกว่า 0.05 แสดงให้เห็นว่าปฏิเสธสมมติฐานหลัก (H_0) ยอมรับสมมติฐานรอง (H_1) กล่าวคือ ข้อมูลมีลักษณะนิ่ง (Stationary) และไม่มี Unit root แต่ในทางกลับกันข้อมูลของสถานี YO01 (แม่น้ำยม) มีลักษณะไม่นิ่ง (Non-stationary) เนื่องจากยอมรับสมมติฐาน (H_0) สังเกตได้จากค่าทดสอบทางสถิติมากกว่าค่าวิกฤติที่ระดับนัยสำคัญ 0.05 (Critical Value (5%)) และค่า p-value มากกว่า 0.05

สำหรับการทดสอบด้วยวิธี KPSS test (ตาราง 8) ข้อมูลของทั้ง 4 สถานี มีลักษณะนิ่ง (Stationary) เนื่องจากค่าทดสอบทางสถิติน้อยกว่าค่าวิกฤติที่ระดับนัยสำคัญ 0.05 (Critical Value (5%)) และค่า p-value มากกว่า 0.05 แสดงว่ายอมรับสมมติฐานหลัก (H_0) กล่าวคือ ข้อมูลมีลักษณะนิ่ง (Stationary) และไม่มี Unit root

เนื่องจากข้อมูลของสถานี YO01 มีลักษณะไม่นิ่งเมื่อทดสอบด้วย ADF test ดังนั้นจึงควรนำข้อมูลดังกล่าวมาหาผลต่างลำดับที่ 1 (1^{st} Difference) เพื่อปรับให้ข้อมูลมีลักษณะนิ่ง และผลที่ได้จากการหาผลต่างลำดับที่ 1 พบว่า ค่าการทดสอบทางสถิติของ ADF test เท่ากับ -5.64 ค่าวิกฤติที่ระดับนัยสำคัญ 0.05 (Critical Value (5%)) เท่ากับ -2.93 และค่า p-value เท่ากับ 0.00 ซึ่งค่าทดสอบทางสถิติน้อยกว่าค่าวิกฤติที่ระดับนัยสำคัญ 0.05 (Critical Value (5%)) และค่า p-value น้อยกว่า 0.05 แสดงให้เห็นว่า ข้อมูลมีลักษณะนิ่ง (Stationary) หลังจากหาผลต่างลำดับที่ 1

14. แบ่งชุดข้อมูลเป็น 2 ส่วน ได้แก่ ข้อมูลชุดฝึกฝน (Training data) และข้อมูลชุดทดสอบ (Test data) โดยแบบจำลองที่ใช้สำหรับการจำแนกระดับคุณภาพน้ำ ข้อมูลชุดฝึกฝนใช้ข้อมูล 80% ของข้อมูลทั้งหมด เท่ากับ 2,120 ตัวอย่าง และข้อมูลชุดทดสอบใช้ข้อมูล 20% ของข้อมูลทั้งหมด เท่ากับ 531 ตัวอย่าง และสำหรับการทำนายค่า WQI ด้วยแบบจำลองอนุกรมเวลา ข้อมูลชุดฝึกฝน (Training data) ใช้ข้อมูลระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2560 (9 ปี) และข้อมูลชุดทดสอบ (Test data) ข้อมูลระหว่างปี พ.ศ. 2561 ถึง พ.ศ. 2564 (4 ปี) ดังตาราง 9

ตาราง 9 จำนวนข้อมูลที่ใช้สำหรับใช้กับแบบจำลองอนุกรมเวลา

สถานีตรวจวัด	ข้อมูลชุดฝึกฝน	ข้อมูลชุดทดสอบ	ข้อมูลทั้งหมด
แม่น้ำปิง สถานี PI06	36	16	52
แม่น้ำวัง สถานี WA02	38	15	53
แม่น้ำยม สถานี YO01	37	14	51
แม่น้ำน่าน สถานี NA02	38	16	54

15. สร้างแบบจำลองสำหรับจำแนกระดับคุณภาพน้ำ ได้แก่ Random Forest, XGBoost, Logistic Regression และ SVM โดยกำหนดให้ค่า DO, BOD, TCB, FCB และ NH_3-N เป็นข้อมูลลักษณะเฉพาะ (Feature) เพื่อใช้จำแนกระดับคุณภาพน้ำ 5 ระดับ ได้แก่ คุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก ดี พอใช้ เสื่อมโทรม และเสื่อมโทรมมาก ทั้งนี้ เกณฑ์ระดับคุณภาพน้ำทั้ง 5 ดังกล่าว จะถูกแทนค่าจากข้อความเป็นตัวเลข (Label encoding) โดยเลข 0 แทนคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก เลข 1 แทนคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม เลข 2 แทนคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ เลข 3 แทนคุณภาพน้ำที่อยู่ในเกณฑ์ดี และเลข 4 แทนคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก

สำหรับแบบจำลอง Logistic Regression และ SVM จำเป็นต้องทำการปรับขอบเขตของข้อมูลลักษณะเฉพาะให้อยู่ในช่วงเดียวกันก่อนเข้าแบบจำลอง โดยงานวิจัยนี้ใช้ RobustScaler เนื่องจากข้อมูลลักษณะเฉพาะมีค่าผิดปกติ (Outlier) (Scikit-learn developers, 2022d) โดยข้อมูลที่ได้หลังจากการปรับข้อมูลให้อยู่ในช่วงเดียวกันมีค่าตั้งแต่ -2.88 ถึง 78.81

งานวิจัยนี้ใช้เทคนิคที่ใช้แก้ไขปัญหาข้อมูลไม่สมดุล 2 เทคนิค เพื่อเพิ่มปริมาณข้อมูลในชุดฝึกฝน ได้แก่ SMOTE และ Random Oversampling ทำให้จำนวนข้อมูลในระดับคุณภาพน้ำในเกณฑ์ต่างๆ ของข้อมูลชุดฝึกฝน (Training data) มีจำนวน 799 ตัวอย่าง เท่ากับจำนวนข้อมูลของระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ที่มีจำนวนข้อมูลมากที่สุด ดังแสดงในตาราง 10

ตาราง 10 การกระจายตัวของข้อมูลระดับคุณภาพน้ำของชุดฝึกฝน (Training data) และชุดทดสอบ (Test data) ในระดับคุณภาพน้ำทั้ง 5 ระดับ

ระดับคุณภาพน้ำ	จำนวนข้อมูลในชุดฝึกฝน			จำนวนข้อมูลในชุดทดสอบ
	ข้อมูลเดิม	Random Oversampling	SMOTE	
เสื่อมโทรมมาก	63	799	799	2
เสื่อมโทรม	703	799	799	176
พอใช้	799	799	799	200
ดี	546	799	799	137
ดีมาก	9	799	799	16

งานวิจัยนี้ใช้เทคนิค Randomized search เพื่อสุ่มเลือก Hyper-parameter ของแบบจำลอง โดยกำหนด k-fold cross-validation ให้ k เท่ากับ 5 และใช้ค่าเฉลี่ยถ่วงน้ำหนักของ F1 score (f1_weighted) เป็นตัวชี้วัดประสิทธิภาพของแบบจำลอง

16. สร้างแบบจำลองอนุกรมเวลา (Time series model) เพื่อทำนายค่า WQI ในอนาคต ซึ่งใช้แบบจำลอง ARIMA, ARIMAX, SARIMA และ SARIMAX สำหรับการหารูปแบบพารามิเตอร์ (p, d, q) และ $(P, D, Q)_S$ จะพิจารณาจากรูปแบบพารามิเตอร์ที่ให้ค่า Akaike Information Criterion (AIC) ที่ต่ำที่สุด กำหนดให้พารามิเตอร์ p, q, P, Q มีค่าอยู่ในช่วง 0 - 3 พารามิเตอร์ d, D มีค่าอยู่ในช่วง 0 - 2 และพารามิเตอร์ S มีค่าเท่ากับ 4 ซึ่งสอดคล้องกับภาพประกอบ 21 – 24 ที่แสดงให้เห็นว่า ข้อมูลมีรูปแบบของฤดูกาล (Seasonal) สำหรับแบบจำลอง ARIMAX และ SARIMAX จำเป็นต้องทำการปรับขอบเขตของข้อมูลให้อยู่ในช่วงเดียวกันก่อนเข้าแบบจำลอง เนื่องจากทั้ง 2 แบบจำลอง นำพารามิเตอร์น้ำตัวอื่นๆ เช่น DO, BOD, TCB, FCB และ NH_3-N มาใช้เป็นตัวแปรภายนอก (Exogenous variable) เพื่อทำนายค่า WQI โดยในงานวิจัยนี้ใช้ MinMaxScaler ปรับขอบเขตของข้อมูลให้มีค่าอยู่ระหว่าง 0 ถึง 1

17. ประเมินประสิทธิภาพของแบบจำลอง เปรียบเทียบผลที่ได้ระหว่างแบบจำลองสำหรับการจำแนกระดับคุณภาพแม่น้ำ ใช้ค่า Accuracy, Precision, Recall และ F1 score และแบบจำลองอนุกรมเวลา ใช้ MAE, RMSE และ MAPE

18. วิเคราะห์ผลการทำนายที่ผิดพลาดของแบบจำลองที่ใช้สำหรับจำแนกระดับคุณภาพแม่น้ำด้วยเทคนิค SHAP

บทที่ 4
ผลการศึกษา

การศึกษาการจำแนกระดับคุณภาพแม่น้ำและการทำนายดัชนีชี้วัดคุณภาพแม่น้ำด้วยการใช้แบบจำลอง ผู้วิจัยได้ดำเนินการวิจัยโดยศึกษาตามขบวนการและขั้นตอนต่างๆ ตลอดจนประเมินประสิทธิภาพของแบบจำลอง และสอดคล้องกับวัตถุประสงค์ที่ได้กำหนดไว้ ดังนี้

1. ผลลัพธ์ของแบบจำลองที่ใช้จำแนกระดับคุณภาพแม่น้ำ
2. ผลลัพธ์ของแบบจำลองอนุกรมเวลาสำหรับทำนายคุณภาพแม่น้ำ

ผลลัพธ์ของแบบจำลองที่ใช้จำแนกระดับคุณภาพแม่น้ำ

การศึกษาเทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกระดับคุณภาพแม่น้ำของประเทศไทย ใช้ข้อมูลจากกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 จำนวนทั้งสิ้น 2,736 ตัวอย่าง หลังจากผ่านขั้นตอนทำความสะอาดข้อมูลแล้ว เหลือข้อมูลคุณภาพน้ำจำนวน 2,651 ตัวอย่าง โดยกำหนดให้ค่า DO, BOD, TCB, FCB และ NH_3-N เป็นลักษณะเฉพาะ สำหรับจำแนกระดับคุณภาพน้ำแบ่งเป็น 5 เกณฑ์ ได้แก่ คุณภาพน้ำอยู่ในเกณฑ์ดีมาก ดี พอใช้ เสื่อมโทรม และเสื่อมโทรมมาก งานวิจัยนี้แบ่งชุดข้อมูลสำหรับใช้จำแนกระดับคุณภาพน้ำเป็น 2 ส่วน ได้แก่ ข้อมูลชุดฝึกฝน (Training data) และข้อมูลชุดทดสอบ (Test data) ในอัตราส่วน 80 ต่อ 20 และใช้แบบจำลองทั้งสิ้น 12 แบบจำลอง โดยใช้ Random Forest, XGBoost, Logistic Regression และ SVM ร่วมกับเทคนิค SMOTE และ Random Oversampling เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูลระดับคุณภาพน้ำ

งานวิจัยนี้ ใช้เทคนิค RandomizedSearchCV เพื่อสุ่มหา Hyper-parameter ที่ทำให้แบบจำลองมีประสิทธิภาพในการจำแนกได้ถูกต้องมากที่สุด กำหนด k-fold cross-validation ให้ k เท่ากับ 5 และใช้ค่าเฉลี่ยถ่วงน้ำหนักของ F1 score (f1_weighted) เป็นตัวชี้วัดประสิทธิภาพของแบบจำลอง พบว่า Hyper-parameter ที่ทำให้ได้ค่าเฉลี่ยถ่วงน้ำหนักของ F1 score ใน cross-validation ของแต่ละแบบจำลอง เป็นดังตาราง 11

ผลลัพธ์ที่ได้จากแบบจำลองที่ใช้จำแนกระดับคุณภาพแม่น้ำของประเทศไทย (ตาราง 12) เป็นไปดังต่อไปนี้

ตาราง 11 Hyper-parameter และค่าเฉลี่ย cross-validation ที่ได้จาก RandomizedSearchCV ของแบบจำลอง

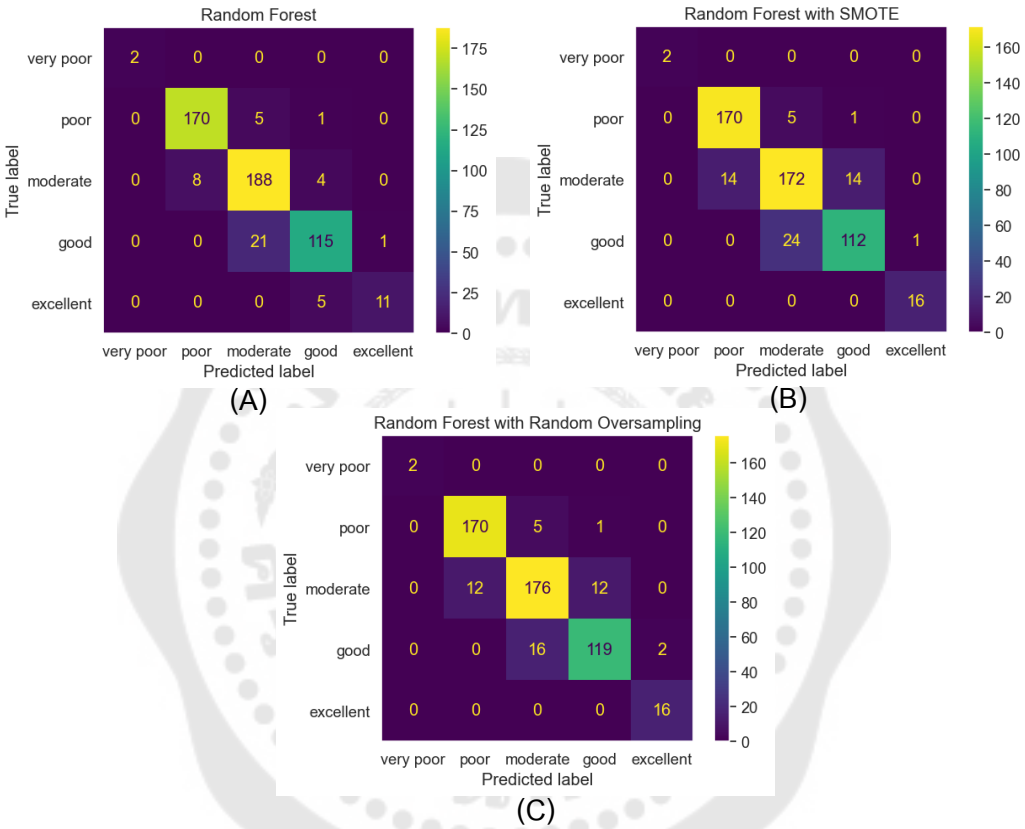
แบบจำลอง	Hyper-parameter tuning	**ค่าเฉลี่ยถ่วงน้ำหนักของ F1 score ใน cross-validation**
Random Forest	n_estimators: 200, min_samples_split: 14, max_features: log2, max_depth: 14	88.91%
Random Forest with SMOTE	smote__k_neighbors: 6, rf__min_samples_split: 33, rf__max_leaf_nodes: 66, rf__max_features: sqrt, rf__max_depth: 47, n_estimators: 200	87.01%
Random Forest with Random Oversampling	rf__min_samples_split: 19, rf__max_leaf_nodes: 90, rf__max_features: sqrt, rf__max_depth: 71, n_estimators: 200	87.87%
XGBoost	n_estimators: 200, subsample: 0.75, min_child_weight: 1, max_depth: 3, learning_rate: 0.3, colsample_bytree: 1	88.37%
XGBoost with SMOTE	subsample: 0.5, smote__k_neighbors: 3, min_child_weight: 1, max_depth: 8, learning_rate: 0.155, colsample_bytree: 1, n_estimators: 200	89.27%
XGBoost with Random Oversampling	subsample: 1, min_child_weight: 5, max_depth: 3, learning_rate: 0.227, colsample_bytree: 0.75, n_estimators: 200	88.48%
Logistic Regression	lr__C: 57.90, penalty= l2	80.97%
Logistic Regression with SMOTE	smote__k_neighbors: 6, lr__C: 100.0, penalty= l2	80.73 %
Logistic Regression with Random Oversampling	lr__C: 57.90, penalty= l2	80.62 %
SVM	svc__kernel: linear, svc__gamma: 2.0, svc__degree: 1, svc__C: 47.37	82.19 %
SVM with SMOTE	svc__kernel: linear, svc__gamma: 0.0002, svc__degree: 2, svc__C: 5.27, smote__k_neighbors: 1	81.24 %
SVM with Random Oversampling	svc__kernel: linear, svc__gamma: 0.0002, svc__degree: 1, svc__C: 42.11	81.24 %

ตาราง 12 ประสิทธิภาพการจำแนกระดับคุณภาพแม่น้ำของแบบจำลอง

แบบจำลอง	การเพิ่มปริมาณข้อมูลในชุดทดสอบ	ค่า F1 score แบ่งตามระดับคุณภาพน้ำ					ค่า F1 score (%)	ค่า Accuracy (%)	ค่า Precision (%)	ค่า Recall (%)
		เสื่อมโทรมมาก (%)	เสื่อมโทรม (%)	พอใช้ (%)	ดี (%)	ดีมาก (%)				
Random Forest	-	100	96.05	90.82	87.79	78.57	91.44	91.53	91.62	91.53
	SMOTE	100	94.44	85.79	84.85	96.97	88.80	88.89	88.89	88.82
	Random Oversampling	100	94.97	88.66	88.48	94.12	90.91	90.96	90.92	90.96
XGBoost	-	66.67	94.29	89.21	87.36	77.42	89.97	90.02	90.22	90.02
	SMOTE	100	94.49	90.38	90.15	84.85	91.56	91.53	91.78	91.53
	Random Oversampling	66.67	93.71	89.47	88.46	77.42	90.17	90.21	90.46	90.21
Logistic Regression	-	100	85.89	77.37	80.28	64.00	80.63	80.6	81.32	80.60
	SMOTE	100	85.8	74.49	75.89	65.22	78.42	78.15	79.23	78.15
	Random Oversampling	100	85.88	74.55	76.33	65.22	78.58	78.34	79.37	78.34
SVM	-	100	85.71	78.16	82.11	64.00	81.34	81.36	81.89	81.36
	SMOTE	100	85.29	73.52	76.76	71.11	78.29	78.15	78.97	78.15
	Random Oversampling	100	85.88	73.90	77.78	74.42	78.99	78.91	79.61	78.91

1. **อัลกอริทึม** Random Forest

ผลลัพธ์ที่ได้จาก Random Forest แบ่งเป็น 3 ส่วน ได้แก่ ผลลัพธ์ที่ได้จากการใช้ Random Forest กับข้อมูลที่ไม่สมดุล Random Forest ร่วมกับเทคนิค SMOTE และ Random Forest ร่วมกับเทคนิค Random Oversampling มีรายละเอียด ดังนี้

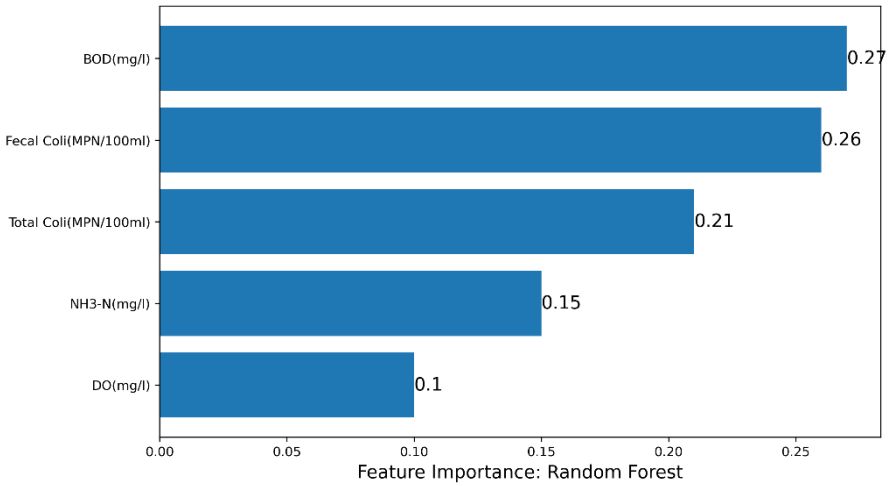


ภาพประกอบ 25 Confusion Matrix ของ (A) แบบจำลอง Random Forest กับข้อมูลที่ไม่สมดุล (B) แบบจำลอง Random Forest ร่วมกับ SMOTE และ (C) แบบจำลอง Random Forest ร่วมกับ Random Oversampling

1.1 Random Forest กับข้อมูลที่ไม่สมดุล

ผลการศึกษา พบว่า แบบจำลองสามารถทำนายข้อมูลชุดทดสอบได้ถูกต้อง 486 จาก 531 ตัวอย่าง ค่า Accuracy, Precision, Recall และ F1 score จากการจำแนกด้วยแบบจำลอง Random Forest กับข้อมูลที่ไม่สมดุล เท่ากับ 91.53%, 91.62%, 91.53% และ 91.44% ตามลำดับ (ตาราง 12) โดยแบบจำลองสามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก ซึ่งมีปริมาณข้อมูลในชุดทดสอบจำนวน 2 ตัวอย่าง ได้ถูกต้องทั้งหมด ทำให้ค่า F1 score ของ

ระดับคุณภาพน้ำดังกล่าวเท่ากับ 100% แต่เมื่อพิจารณาค่า F1 score ของคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 78.57% เนื่องจากแบบจำลองจำแนกตัวอย่างได้ถูกต้อง 11 จาก 16 ตัวอย่าง ซึ่งที่จำแนกผิดเป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดี จำนวน 5 ตัวอย่าง ดังแสดงใน Confusion Matrix (ภาพประกอบ 25A)



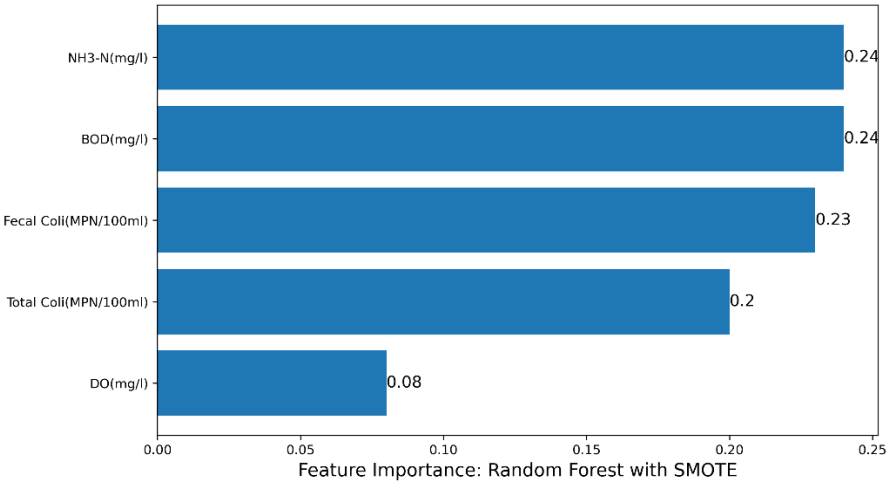
ภาพประกอบ 26 Feature Importance ของแบบจำลอง Random Forest กับข้อมูลที่ไม่สมดุล

เมื่อพิจารณาความสำคัญของพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำด้วยการหาค่า Feature Importance ซึ่งได้มาจากการวิเคราะห์ค่า Gini impurity ของ Random Forest ดังแสดงในภาพประกอบ 26 พบว่า BOD (mg/l) มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองดังกล่าวมากที่สุด โดยค่า Feature Importance เท่ากับ 0.27 รองลงมา ได้แก่ FCB เท่ากับ 0.26 TCB เท่ากับ 0.21 NH_3-N เท่ากับ 0.15 และ DO เท่ากับ 0.10 ตามลำดับ

1.2 Random Forest ร่วมกับเทคนิค SMOTE

ผลการศึกษา พบว่า ค่า Accuracy, Precision, Recall และ F1 score ของแบบจำลอง Random Forest ร่วมกับเทคนิค SMOTE เท่ากับ 88.89%, 88.89%, 88.82% และ 88.80% ตามลำดับ (ตาราง 12) จำแนกระดับคุณภาพของชุดทดสอบได้ถูกต้อง 472 จาก 531 ตัวอย่าง โดยสามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก ได้ถูกต้องทั้ง 2 ตัวอย่าง ทำให้ค่า F1 score ของคุณภาพน้ำระดับดังกล่าวเท่ากับ 100% แต่เมื่อพิจารณาผลการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก แบบจำลองดังกล่าวสามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากถูกต้องทั้ง 16 ตัวอย่าง แต่มีเพียง 1 ตัวอย่าง ที่จำแนกระดับคุณภาพน้ำที่อยู่ใน

เกณฑ์ดีเป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก (ภาพประกอบ 25B) ซึ่งทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 96.97%

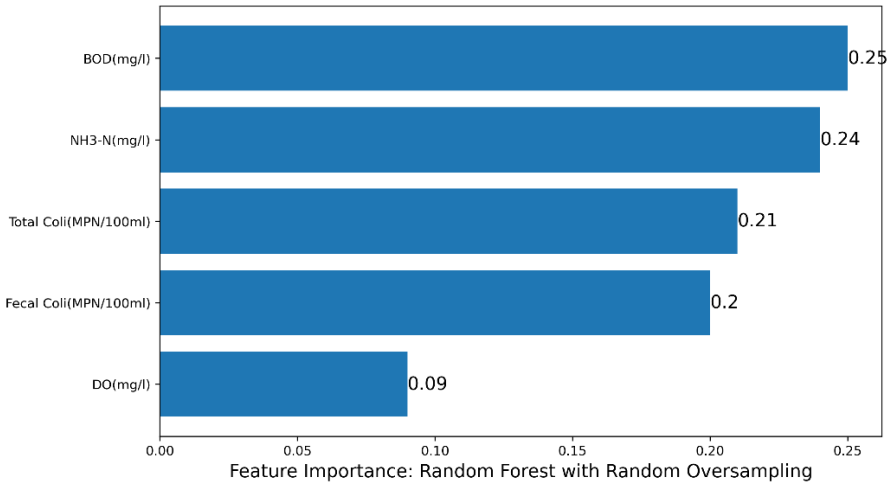


ภาพประกอบ 27 Feature Importance ของแบบจำลอง Random Forest ร่วมกับเทคนิค SMOTE

เมื่อพิจารณาพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลอง Random Forest ร่วมกับเทคนิค SMOTE (ภาพประกอบ 27) พบว่า NH_3-N และ BOD มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองดังกล่าวมากที่สุด โดยค่า Feature Importance เท่ากับ 0.24 รองลงมา ได้แก่ FCB เท่ากับ 0.23 TCB เท่ากับ 0.20 และ DO เท่ากับ 0.08 ตามลำดับ

1.3 Random Forest ร่วมกับเทคนิค Random Oversampling

จากผลการศึกษาการจำแนกระดับคุณภาพน้ำของแบบจำลอง Random Forest ร่วมกับเทคนิค Random Oversampling พบว่า ค่า Accuracy, Precision, Recall และ F1 score เท่ากับ 90.96%, 90.92%, 90.96% และ 90.91% ตามลำดับ (ตาราง 12) โดยจำแนกระดับคุณภาพน้ำของข้อมูลในชุดทดสอบถูกต้อง 483 จาก 531 ตัวอย่าง ซึ่งแบบจำลองจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก ถูกต้องทั้ง 2 ตัวอย่าง (F1 score เท่ากับ 100%) เช่นเดียวกับแบบจำลอง Random Forest กับข้อมูลที่ไม่สมดุล และแบบจำลอง Random Forest ร่วมกับเทคนิค SMOTE นอกจากนี้ แบบจำลอง Random Forest ร่วมกับเทคนิค Random Oversampling สามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากถูกต้อง 16 ตัวอย่าง แต่มีเพียง 2 ตัวอย่าง ที่จำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมาก (ภาพประกอบ 25C) ซึ่งทำให้ค่า F1 score ของระดับคุณภาพน้ำดีมาก เท่ากับ 94.12%

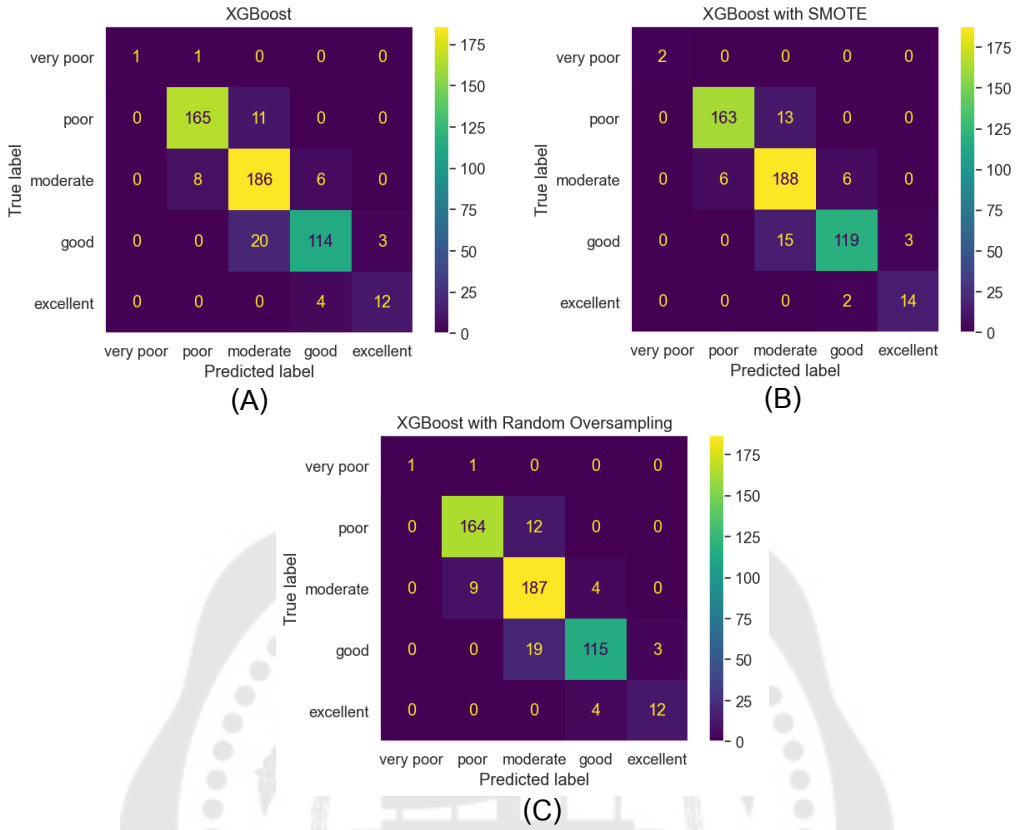


ภาพประกอบ 28 Feature Importance ของแบบจำลอง Random Forest ร่วมกับเทคนิค Random Oversampling

เมื่อพิจารณาพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลอง Random Forest ร่วมกับ Random Oversampling (ภาพประกอบ 28) โดยพิจารณาจากค่า Feature Importance พบว่า BOD มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองมากที่สุด เท่ากับ 0.25 รองลงมา ได้แก่ NH_3-N เท่ากับ 0.24 TCB เท่ากับ 0.21 FCB เท่ากับ 0.20 และ DO เท่ากับ 0.09 ตามลำดับ

2. อัลกอริทึม XGBoost

ผลลัพธ์ที่ได้จาก XGBoost แบ่งเป็น 3 ส่วน ได้แก่ ผลลัพธ์ที่ได้จากการใช้ XGBoost ร่วมกับข้อมูลที่ไม่สมดุล Random Forest ร่วมกับเทคนิค SMOTE และ XGBoost ร่วมกับเทคนิค Random Oversampling โดยมีรายละเอียด ดังนี้

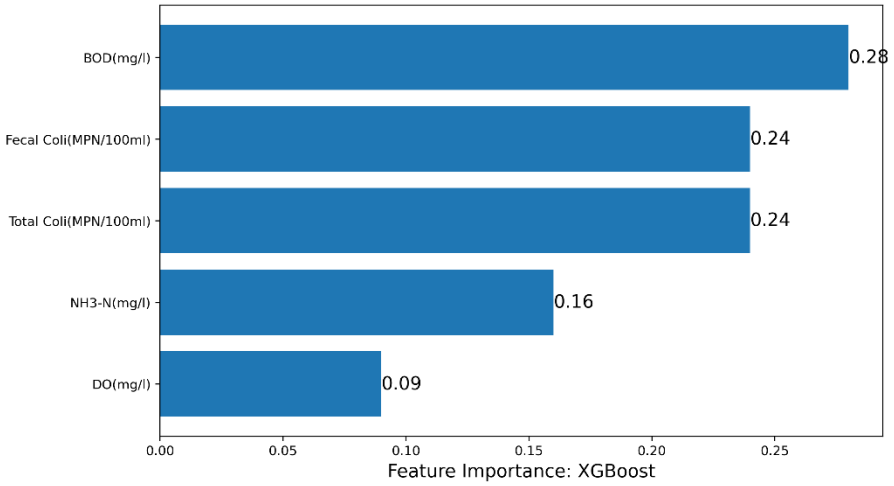


ภาพประกอบ 29 Confusion Matrix ของ (A) แบบจำลอง XGBoost กับข้อมูลที่ไม่สมดุล (B) แบบจำลอง XGBoost ร่วมกับ SMOTE และ (C) แบบจำลอง XGBoost ร่วมกับ Random Oversampling

2.1 XGBoost กับข้อมูลที่ไม่สมดุล

ผลการศึกษา XGBoost กับข้อมูลที่ไม่สมดุล พบว่า แบบจำลอง XGBoost สามารถจำแนกระดับคุณภาพน้ำถูกต้อง 478 ตัวอย่าง จากข้อมูลชุดทดสอบทั้งสิ้น 531 ตัวอย่าง ซึ่งค่า Accuracy, Precision, Recall และ F1 score เท่ากับ 90.02%, 90.02%, 90.22% และ 89.97% ตามลำดับ (ตาราง 12) เมื่อพิจารณาประสิทธิภาพการจำแนกระดับคุณภาพน้ำในแต่ละเกณฑ์ พบว่า แบบจำลองจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมได้ดีที่สุด (F1 score เท่ากับ 94.29%) อย่างไรก็ตาม แบบจำลองสามารถจำแนกข้อมูลระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมากถูกต้องจำนวน 1 จาก 2 ตัวอย่าง โดยจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมากเป็นเกณฑ์เสื่อมโทรม (ภาพประกอบ 29A) ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก เท่ากับ 66.67% และเมื่อพิจารณาระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก (F1 score เท่ากับ 77.42%) พบว่า สามารถจำแนกระดับคุณภาพน้ำได้ถูกต้อง 12 ตัวอย่าง

ซึ่งจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดีจำนวน 4 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 3 ตัวอย่าง

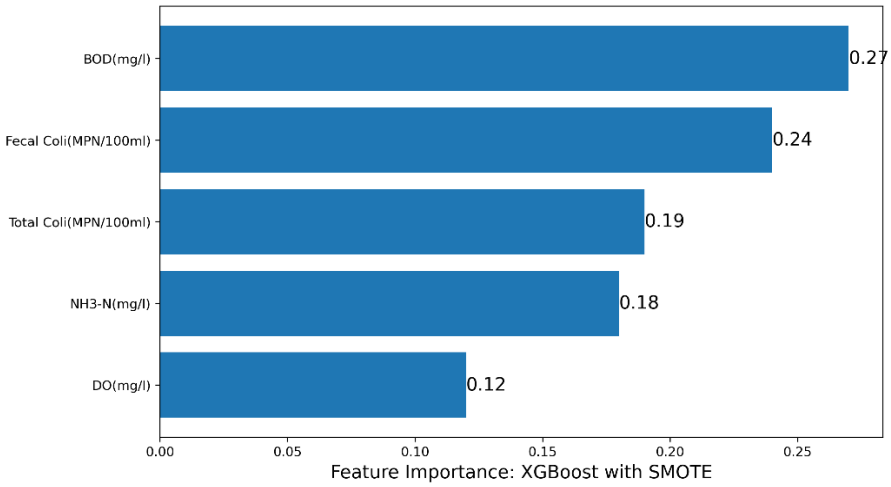


ภาพประกอบ 30 Feature Importance ของแบบจำลอง XGBoost กับข้อมูลที่ไม่สมดุล

เมื่อพิจารณาพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลอง XGBoost (ภาพประกอบ 30) พบว่า BOD มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองมากที่สุด โดยค่า Feature Importance เท่ากับ 0.28 ลำดับถัดไป ได้แก่ FCB และ TCB เท่ากับ 0.24 NH_3-N เท่ากับ 0.16 และ DO เท่ากับ 0.09 ตามลำดับ

2.2 XGBoost ร่วมกับเทคนิค SMOTE

ผลการศึกษาแบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE สำหรับจำแนกระดับคุณภาพน้ำ พบว่า ค่า Accuracy, Precision, Recall และ F1 score เท่ากับ 91.53%, 91.78%, 91.53% และ 91.56% ตามลำดับ (ตาราง 12) โดยแบบจำลองสามารถจำแนกข้อมูลชุดทดสอบได้ถูกต้อง 486 จาก 531 ตัวอย่าง และเมื่อพิจารณาผลการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก แบบจำลองสามารถจำแนกระดับคุณภาพน้ำถูกต้องทั้ง 2 ตัวอย่าง อย่างไรก็ตาม แบบจำลองสามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากได้ถูกต้อง 14 ตัวอย่าง ซึ่งจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดีจำนวน 2 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 3 ตัวอย่าง ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 84.85% ดังภาพประกอบ 29B และตาราง 12

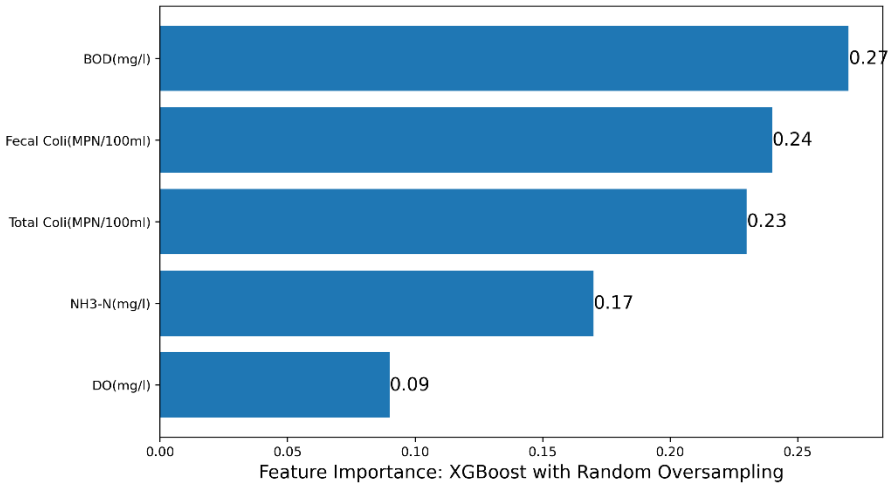


ภาพประกอบ 31 Feature Importance ของแบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE

จากภาพประกอบ 31 แสดงค่า Feature Importance ที่ได้จากแบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE พบว่า BOD มีค่ามากที่สุด เท่ากับ 0.27 แสดงให้เห็นว่า BOD มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองมากที่สุด รองลงมา ได้แก่ FCB (0.24), TCB (0.19), NH_3-N (0.18) และ DO (0.12) ตามลำดับ

2.3 XGBoost ร่วมกับเทคนิค Random Oversampling

จากผลการศึกษาแบบจำลอง XGBoost ร่วมกับเทคนิค Random Oversampling พบว่า จำแนกระดับคุณภาพน้ำถูกต้องจำนวน 479 ตัวอย่าง จากข้อมูลชุดทดสอบ 531 ตัวอย่าง ทำให้ค่า Accuracy, Precision, Recall และ F1 score ของแบบจำลองเท่ากับ 90.21%, 90.46%, 90.21% และ 90.17% ตามลำดับ (ตาราง 12) เมื่อพิจารณาประสิทธิภาพการจำแนกระดับคุณภาพน้ำในแต่ละเกณฑ์ พบว่า แบบจำลองมีสามารถจำแนกข้อมูลที่อยู่ในเกณฑ์เสื่อมโทรมมากถูกจำนวน 1 จาก 2 ตัวอย่าง (F1 score เท่ากับ 66.67%) โดยจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมากเป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม (ภาพประกอบ 29C) และเมื่อพิจารณาระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากพบว่า สามารถจำแนกข้อมูลได้ถูกต้อง 12 ตัวอย่าง ซึ่งจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดีจำนวน 4 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 3 ตัวอย่าง ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 77.42%

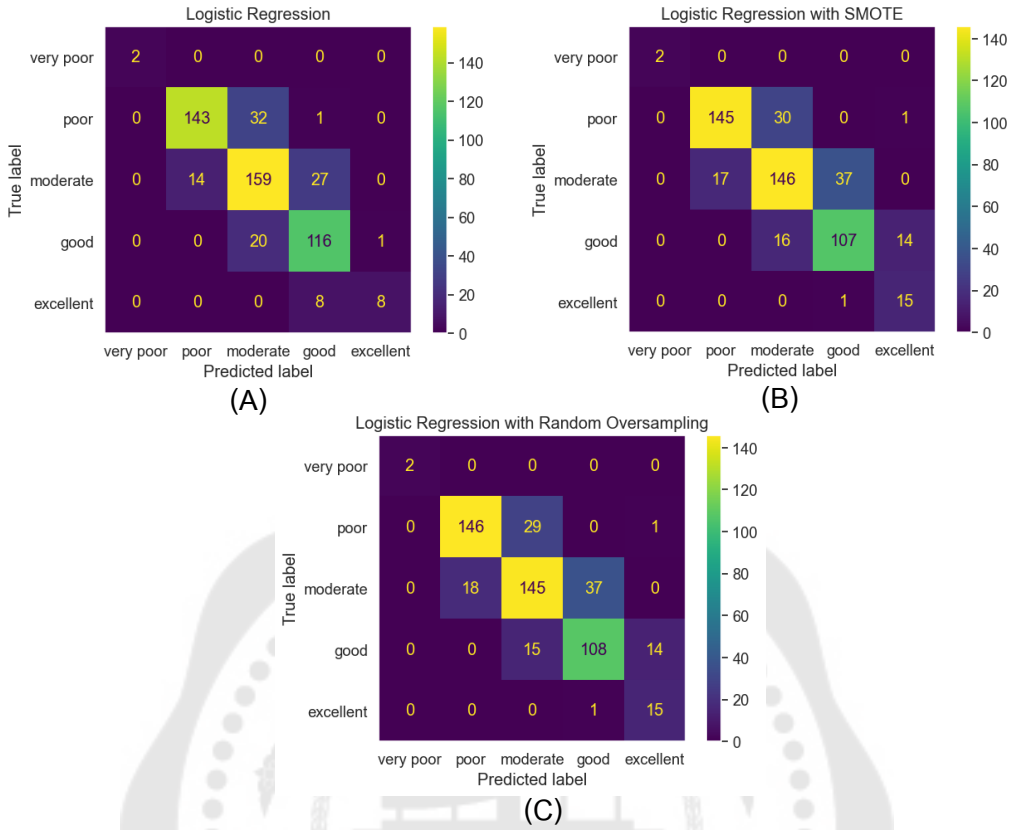


ภาพประกอบ 32 Feature Importance ของแบบจำลอง XGBoost ร่วมกับเทคนิค Random Oversampling

จากภาพประกอบ 32 แสดงค่า Feature Importance เพื่อบ่งบอกพารามิเตอร์น้ำที่มีผลต่อแบบจำลอง XGBoost ร่วมกับเทคนิค Random Oversampling ที่ใช้จำแนกระดับคุณภาพน้ำ พบว่า BOD มีค่ามากที่สุด เท่ากับ 0.27 รองลงมา ได้แก่ FCB (0.24), TCB (0.23), NH_3-N (0.18) และ DO (0.09) ตามลำดับ

3. **อัลกอริทึม** Logistic Regression

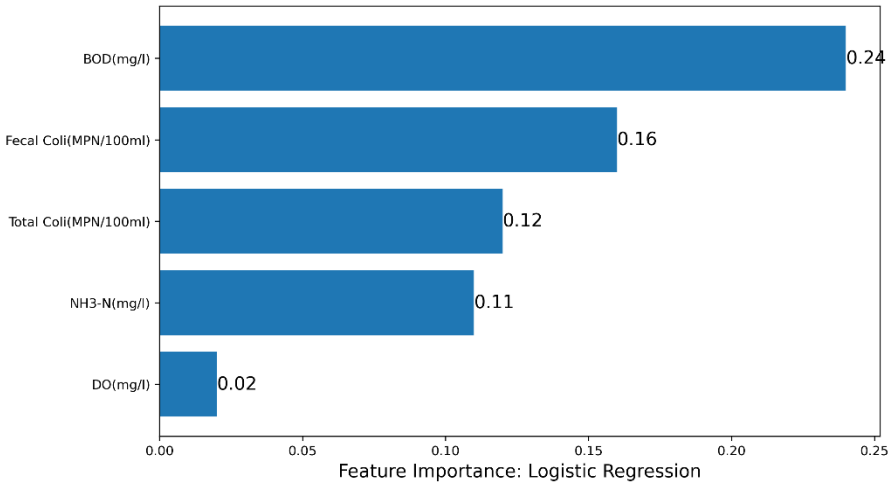
ผลลัพธ์ที่ได้จาก Logistic Regression แบ่งเป็น 3 ส่วน ได้แก่ ผลลัพธ์ที่ได้จากการใช้ Logistic Regression ร่วมกับข้อมูลที่ไม่สมดุล Logistic Regression ร่วมกับเทคนิค SMOTE และ Random Forest ร่วมกับเทคนิค Random Oversampling โดยมีรายละเอียด ดังนี้



ภาพประกอบ 33 Confusion Matrix ของ (A) แบบจำลอง Logistic Regression กับข้อมูลที่ไม่สมดุล (B) แบบจำลอง Logistic Regression ร่วมกับ SMOTE และ (C) แบบจำลอง Logistic Regression ร่วมกับ Random Oversampling

3.1 Logistic Regression กับข้อมูลที่ไม่สมดุล

จากการศึกษาการจำแนกระดับคุณภาพน้ำด้วย Logistic Regression กับข้อมูลที่ไม่สมดุล พบว่า แบบจำลองมีประสิทธิภาพสำหรับจำแนกข้อมูลชุดทดสอบถูกต้อง 428 ตัวอย่าง จาก 531 ตัวอย่าง ทำให้ค่า Accuracy, Precision, Recall และ F1 score เท่ากับ 80.60%, 81.32%, 80.60% และ 80.63% ตามลำดับ (ตาราง 12) เมื่อพิจารณาข้อมูลระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมากพบว่า แบบจำลองสามารถจำแนกข้อมูลได้ถูกต้องทั้ง 2 ตัวอย่าง ดังแสดงในภาพประกอบ 33A แต่เมื่อพิจารณาระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก สามารถจำแนกข้อมูลได้ถูกต้อง 8 ตัวอย่าง ซึ่งจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดีจำนวน 8 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 1 ตัวอย่าง ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 65.22%

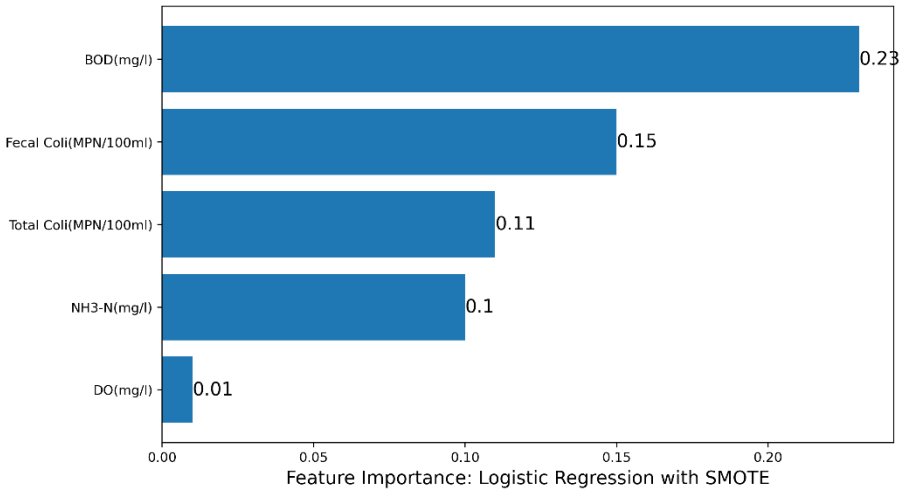


ภาพประกอบ 34 Feature Importance ของแบบจำลอง Logistic Regression กับข้อมูลที่ไม่สมดุล

เมื่อพิจารณาพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลอง Logistic Regression พบว่า BOD มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองมากที่สุด เนื่องจากค่า Permutation Importance ของแบบจำลองมีมากที่สุด เท่ากับ 0.24 ลำดับถัดมา ได้แก่ FCB (0.16), TCB (0.12), NH_3-N (0.11) และ DO (0.02) ตามลำดับ (ภาพประกอบ 34)

3.2 Logistic Regression ร่วมกับเทคนิค SMOTE

ผลการศึกษาการจำแนกระดับคุณภาพน้ำด้วย Logistic Regression ร่วมกับเทคนิค SMOTE พบว่า แบบจำลองสามารถจำแนกข้อมูลชุดทดสอบถูกต้อง 415 ตัวอย่าง จาก 531 ตัวอย่าง และค่า Accuracy, Precision, Recall และ F1 score ของแบบจำลองเท่ากับ 78.15%, 79.23%, 78.15% และ 78.42% ตามลำดับ ดังตาราง 12 เมื่อพิจารณาความถูกต้องของการจำแนกข้อมูลระดับคุณภาพน้ำที่ได้จากแบบจำลองดังกล่าว พบว่าสามารถจำแนกข้อมูลที่อยู่ในเกณฑ์เสื่อมโทรมมากถูกต้องทั้ง 2 ตัวอย่าง นอกจากนี้ สามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากได้ถูกต้อง 15 ตัวอย่าง ซึ่งจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดีจำนวน 1 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 14 ตัวอย่าง (ภาพประกอบ 33B) ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 65.22%

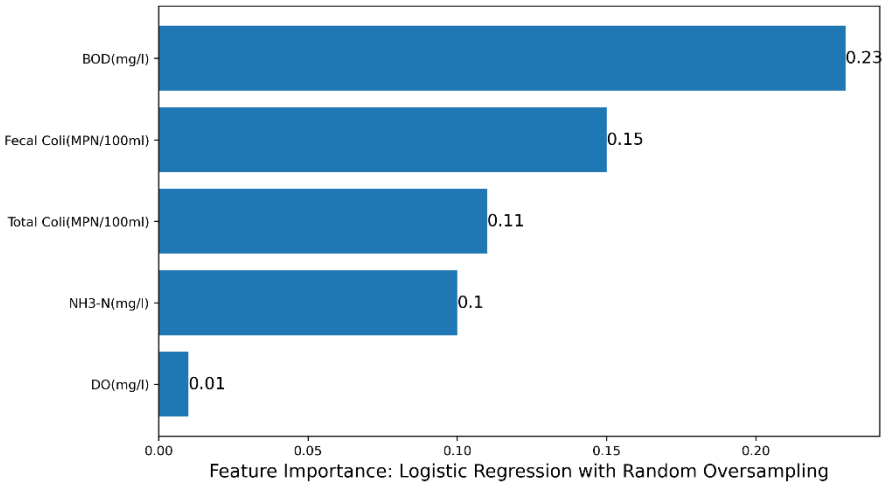


ภาพประกอบ 35 Feature Importance ของแบบจำลอง Logistic Regression ร่วมกับเทคนิค SMOTE

จากภาพประกอบ 35 แสดงความสำคัญของพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำสำหรับแบบจำลอง Logistic Regression ร่วมกับเทคนิค SMOTE พบว่า BOD มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองมากที่สุด เนื่องจากค่า Permutation Importance ของแบบจำลองมีมากที่สุด เท่ากับ 0.23 ลำดับถัดมา ได้แก่ FCB (0.15), TCB (0.11), NH_3-N (0.10) และ DO (0.01) ตามลำดับ

3.3 Logistic Regression ร่วมกับเทคนิค Random Oversampling

ผลการศึกษาการจำแนกระดับคุณภาพน้ำด้วย Logistic Regression ร่วมกับเทคนิค Random Oversampling พบว่า ค่า Accuracy, Precision, Recall และ F1 score ของแบบจำลองเท่ากับ 78.34%, 79.37%, 78.34% และ 78.58% ตามลำดับ (ตาราง 12) ซึ่งสามารถจำแนกข้อมูลถูกต้อง 416 ตัวอย่างจากจำนวนข้อมูลชุดทดสอบทั้งสิ้น 531 ตัวอย่าง เมื่อพิจารณาความถูกต้องของการจำแนกข้อมูลระดับคุณภาพน้ำที่ได้จากแบบจำลองดังกล่าวพบว่า สามารถจำแนกข้อมูลที่อยู่ในเกณฑ์เสื่อมโทรมมากได้ถูกต้องทั้ง 2 ตัวอย่าง แต่เมื่อพิจารณาระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากพบว่า สามารถจำแนกระดับคุณภาพน้ำได้ถูกต้อง 15 ตัวอย่าง ซึ่งจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดีจำนวน 1 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 14 ตัวอย่าง (ภาพประกอบ 33C) ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 65.22%

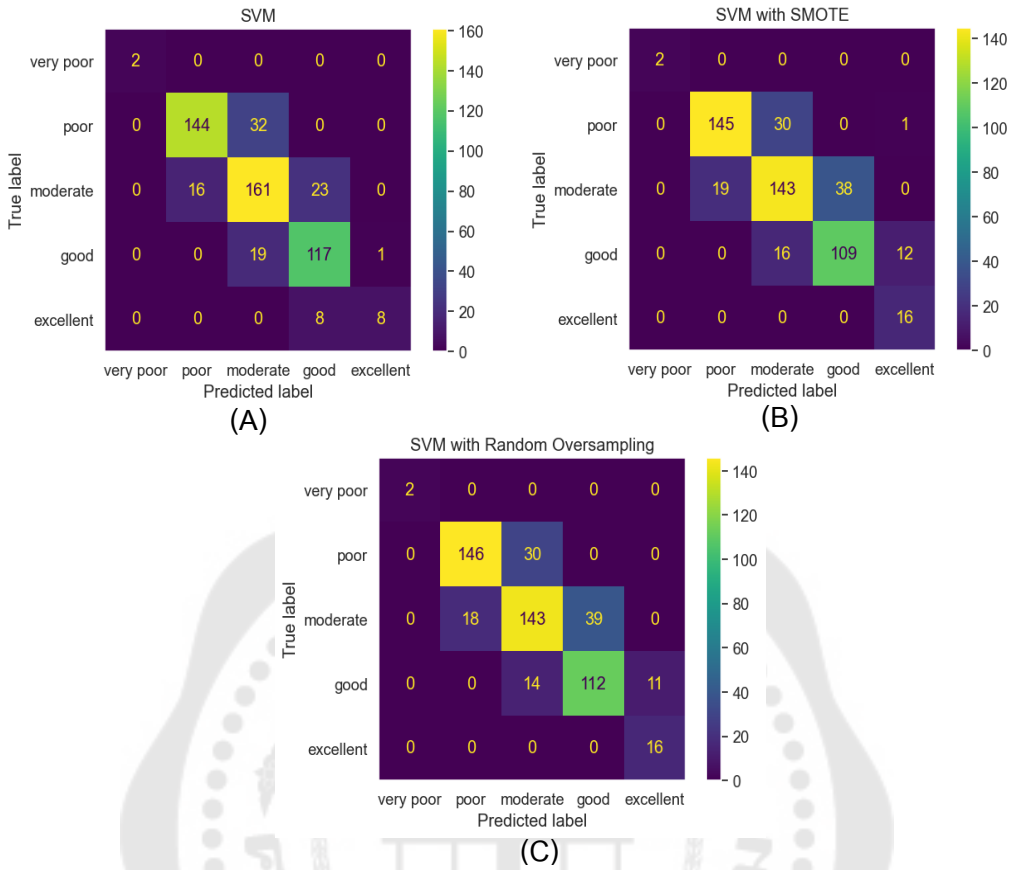


ภาพประกอบ 36 Feature Importance ของแบบจำลอง Logistic Regression ร่วมกับเทคนิค Random Oversampling

จากภาพประกอบ 36 แสดงความสำคัญของพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลอง Logistic Regression ร่วมกับเทคนิค Random Oversampling พบว่า BOD มีผลต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองมากที่สุด เนื่องจากค่า Permutation Importance ของแบบจำลองมีมากที่สุด เท่ากับ 0.23 ลำดับถัดมา ได้แก่ FCB (0.15), NH_3-N (0.11), TCB (0.10) และ DO (0.01) ตามลำดับ

4. อัลกอริทึม SVM

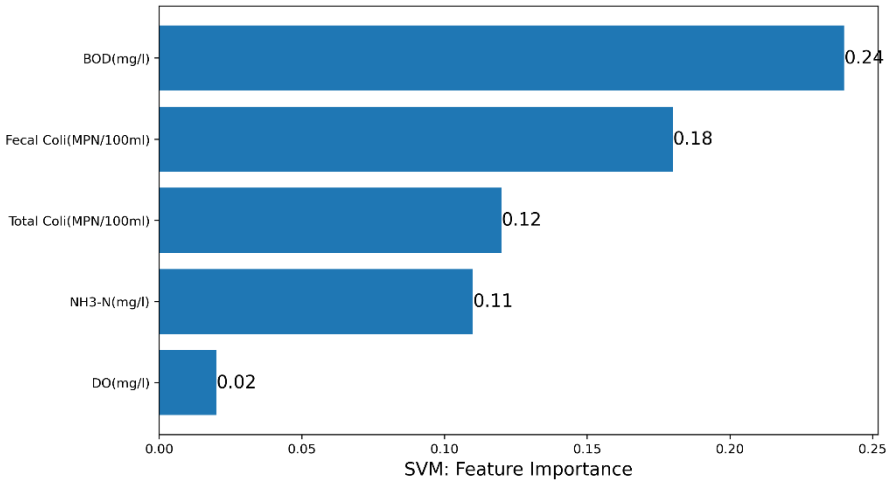
ผลลัพธ์ที่ได้จาก SVM แบ่งเป็น 3 ส่วน ได้แก่ ผลลัพธ์ที่ได้จากการใช้ SVM ร่วมกับข้อมูลที่ไม่สมดุล SVM ร่วมกับเทคนิค SMOTE และ SVM ร่วมกับเทคนิค Random Oversampling โดยมีรายละเอียด ดังนี้



ภาพประกอบ 37 Confusion Matrix ของ (A) แบบจำลอง SVM กับข้อมูลที่ไม่สมดุล (B) แบบจำลอง SVM ร่วมกับ SMOTE และ (C) แบบจำลอง SVM ร่วมกับ Random Oversampling

4.1 SVM กับข้อมูลที่ไม่สมดุล

จากผลการศึกษาการจำแนกระดับคุณภาพน้ำด้วยแบบจำลอง SVM กับข้อมูลที่ไม่สมดุล พบว่า แบบจำลองสามารถจำแนกข้อมูลได้ถูกต้อง 432 จากข้อมูลชุดทดสอบทั้งสิ้น 531 ตัวอย่าง ทำให้ค่า Accuracy, Precision, Recall และ F1 score ของแบบจำลองเท่ากับ 81.36%, 81.89%, 81.36% และ 81.34% ตามลำดับ (ตาราง 12) เมื่อพิจารณาความถูกต้องของการจำแนกข้อมูลระดับคุณภาพน้ำ พบว่า สามารถจำแนกข้อมูลที่อยู่ในเกณฑ์เสื่อมโทรมมากได้ถูกต้องทั้ง 2 ตัวอย่าง และสำหรับระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากนั้น สามารถจำแนกระดับคุณภาพน้ำได้ถูกต้อง 8 ตัวอย่าง โดยจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากเป็นดีจำนวน 8 ตัวอย่าง อย่างไรก็ตาม แบบจำลองจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 1 ตัวอย่าง (ภาพประกอบ 37A) ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 64.00%

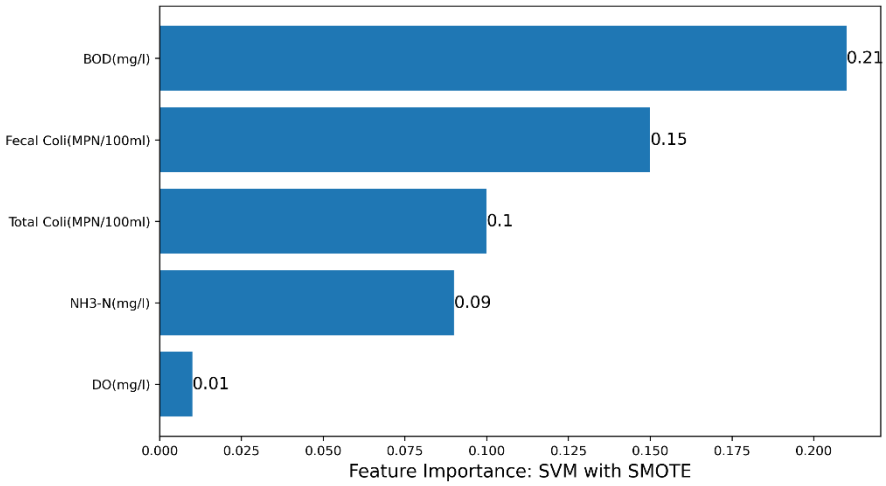


ภาพประกอบ 38 Feature Importance ของแบบจำลอง SVM กับข้อมูลที่ไม่สมดุล

เมื่อพิจารณาความสำคัญของพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำด้วยแบบจำลอง SVM พบว่า BOD เป็นพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำมากที่สุด เนื่องจากค่า Permutation Importance ของแบบจำลอง เท่ากับ 0.24 ซึ่งมีค่ามากที่สุดลำดับถัดมา ได้แก่ FCB (0.18), TCB (0.12), NH_3-N (0.11) และ DO (0.03) ตามลำดับ (ภาพประกอบ 38)

4.2 SVM ร่วมกับเทคนิค SMOTE

จากผลการศึกษาการจำแนกระดับคุณภาพน้ำด้วยแบบจำลอง SVM ร่วมกับเทคนิค SMOTE พบว่า สามารถจำแนกข้อมูลได้ถูกต้อง 415 จากข้อมูลชุดทดสอบทั้งสิ้น 531 ตัวอย่าง ให้ค่า Accuracy, Precision, Recall และ F1 score ของแบบจำลองเท่ากับ 78.15%, 78.97%, 78.15% และ 78.29% ตามลำดับ (ตาราง 12) เมื่อพิจารณาความถูกต้องของการจำแนกข้อมูลระดับคุณภาพน้ำ พบว่า สามารถจำแนกข้อมูลที่อยู่ในเกณฑ์เสื่อมโทรมมากได้ถูกต้องทั้ง 2 ตัวอย่าง และแบบจำลองสามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากได้ถูกต้อง 16 ตัวอย่าง อย่างไรก็ตาม แบบจำลองจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีผิดพลาดเป็นดีมากจำนวน 12 ตัวอย่าง (ภาพประกอบ 37B) ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 71.11%

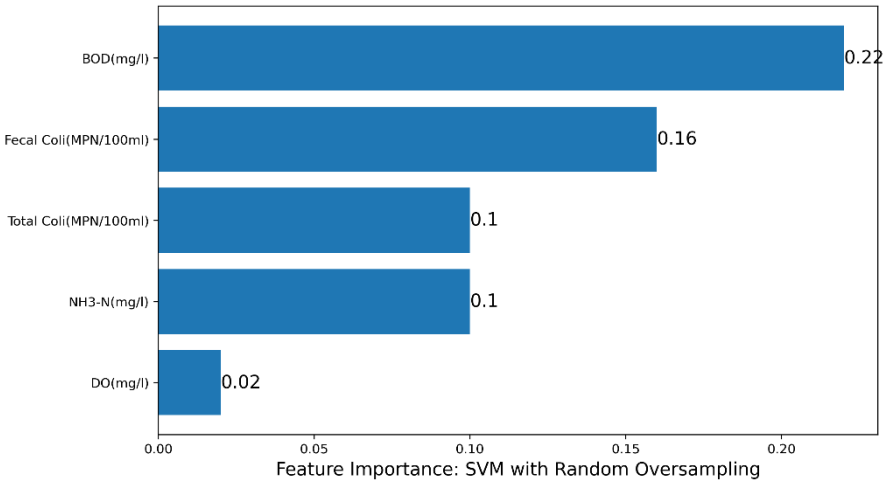


ภาพประกอบ 39 Feature Importance ของแบบจำลอง SVM ร่วมกับเทคนิค SMOTE

เมื่อพิจารณาพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำด้วยแบบจำลอง SVM ร่วมกับเทคนิค SMOTE พบว่า BOD เป็นพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำมากที่สุด โดยค่า Permutation Importance ของแบบจำลองมีค่ามากที่สุดเท่ากับ 0.21 รองลงมา ได้แก่ FCB เท่ากับ 0.15 TCB เท่ากับ 0.10 NH_3-N เท่ากับ 0.09 และ DO เท่ากับ 0.01 ตามลำดับ (ภาพประกอบ 39)

4.3 SVM ร่วมกับเทคนิค Random Oversampling

จากผลการศึกษาการจำแนกระดับคุณภาพน้ำด้วยแบบจำลอง SVM ร่วมกับเทคนิค Random Oversampling พบว่า สามารถจำแนกข้อมูลได้ถูกต้อง 419 จากข้อมูลชุดทดสอบทั้งสิ้น 531 ตัวอย่าง ทำให้ค่า Accuracy, Precision, Recall และ F1 score เท่ากับ 78.91%, 79.61%, 78.91% และ 78.99% ตามลำดับ (ตาราง 12) เมื่อพิจารณาความถูกต้องของการจำแนกข้อมูลระดับคุณภาพน้ำ พบว่า สามารถจำแนกข้อมูลที่อยู่ในเกณฑ์เสื่อมโทรมมากได้ถูกต้องทั้ง 2 ตัวอย่าง และสำหรับระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก แบบจำลองสามารถจำแนกได้ถูกต้อง 16 ตัวอย่าง อย่างไรก็ตาม แบบจำลองจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีผิดพลาดเป็นดีมากจำนวน 11 ตัวอย่าง (ภาพประกอบ 37C) ทำให้ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 74.42%

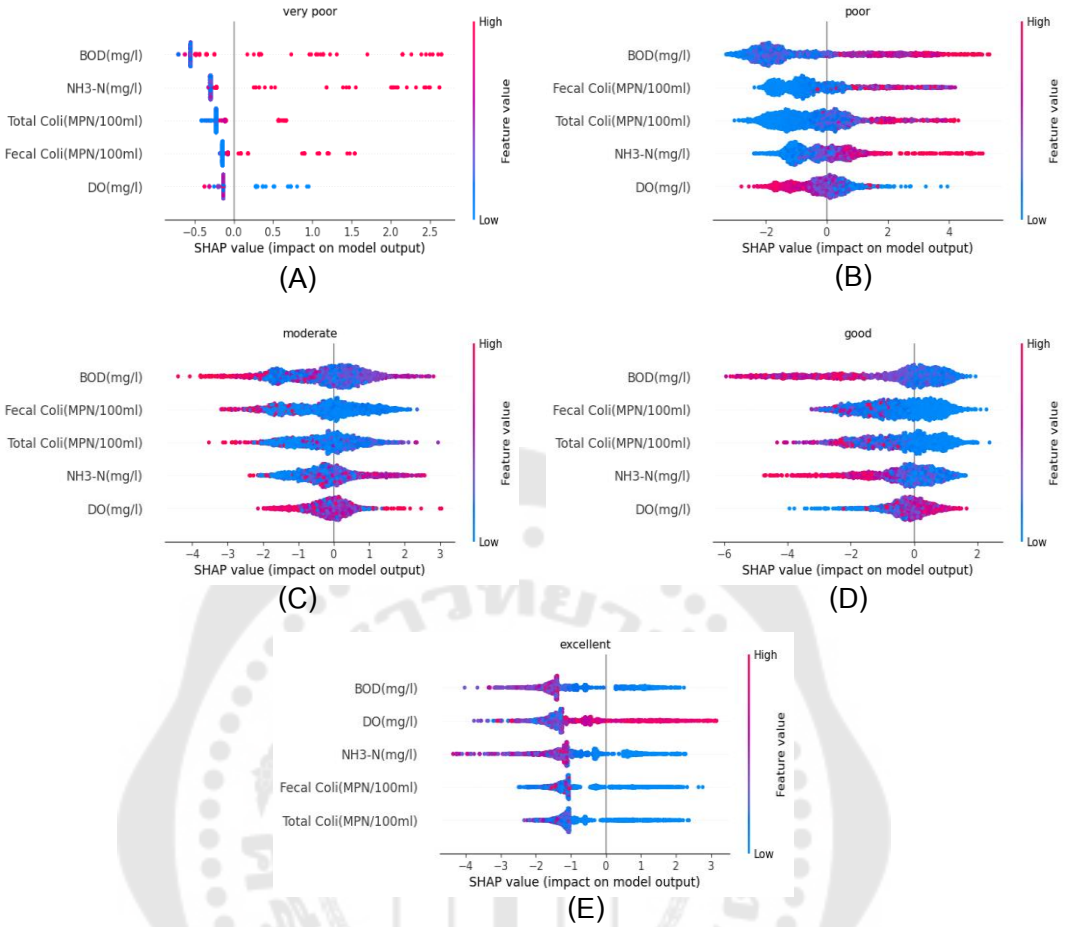


ภาพประกอบ 40 Feature Importance ของแบบจำลอง SVM ร่วมกับเทคนิค Random Oversampling

เมื่อพิจารณาพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำด้วยแบบจำลอง SVM ร่วมกับเทคนิค Random Oversampling พบว่า BOD เป็นพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำมากที่สุด โดยค่า Permutation Importance ของแบบจำลองมีค่ามากที่สุด เท่ากับ 0.22 รองลงมา ได้แก่ FCB เท่ากับ 0.16 TCB และ NH_3-N เท่ากับ 0.10 และ DO เท่ากับ 0.02 ซึ่ง DO เป็นพารามิเตอร์น้ำที่มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำของแบบจำลองน้อยที่สุด (ภาพประกอบ 40)

5. วิเคราะห์ผลการทำนายที่ผิดพลาดของแบบจำลอง

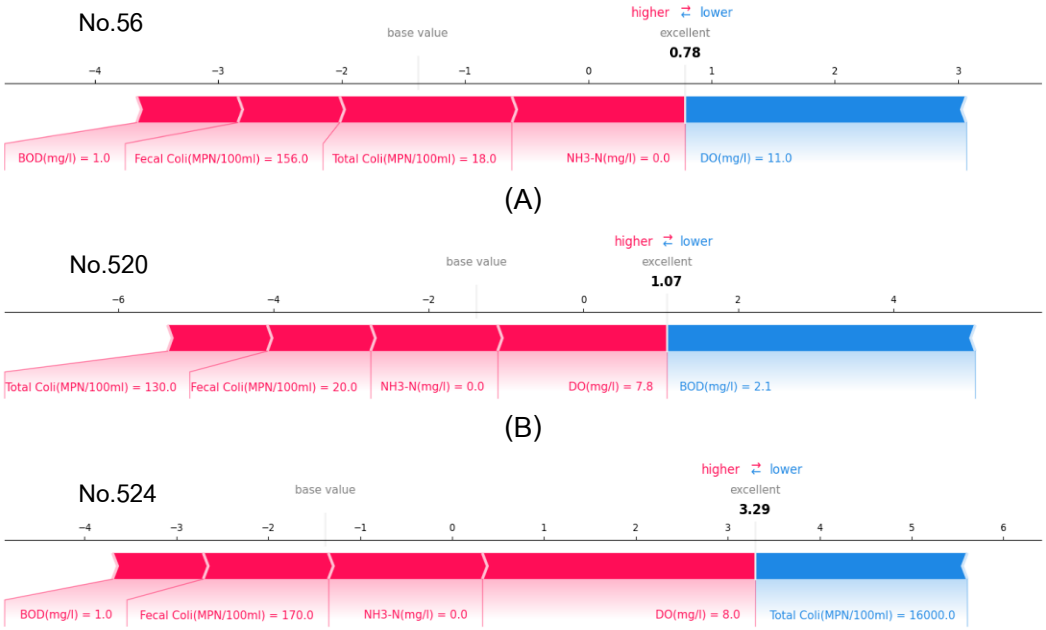
แบบจำลอง XGBoost ร่วมกับการแก้ไขปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE มีประสิทธิภาพสูงที่สุดสำหรับจำแนกระดับคุณภาพน้ำ สังเกตได้จากตาราง 12 ซึ่งค่า F1 score (91.56%), Accuracy (91.53%), Precision (91.53%) และ Recall (91.78%) สูงที่สุด เมื่อพิจารณา Confusion Matrix ของแบบจำลองดังภาพประกอบ 29B พบว่า ระดับคุณภาพน้ำที่มีจำนวนตัวอย่างน้อย ได้แก่ ระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก (2 ตัวอย่าง) และอยู่ในเกณฑ์ดีมาก (16 ตัวอย่าง) แบบจำลองสามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมากได้ถูกต้องทั้ง 2 ตัวอย่าง อย่างไรก็ตาม แบบจำลองจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดี จำนวน 2 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีผิดพลาดเป็นดีมากจำนวน 3 ตัวอย่าง ด้วยเหตุนี้ จึงนำเทคนิค SHAP ที่ช่วยอธิบายปัจจัยที่ส่งผลให้แบบจำลองทำนายบางตัวอย่างผิดพลาด มาใช้กับงานวิจัยนี้



ภาพประกอบ 41 ความสำคัญของพารามิเตอร์น้ำที่ใช้จำแนกระดับคุณภาพน้ำในแต่ละเกณฑ์ของแบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE ได้แก่ (A) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก (B) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม (C) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ (D) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดี และ (E) ระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก โดยสีของแต่ละจุดบ่งบอกถึงพารามิเตอร์มีค่าสูง (จุดสีแดง) และต่ำ (จุดสีน้ำเงิน)

จากภาพประกอบ 41 แสดงความสำคัญของพารามิเตอร์น้ำที่ใช้จำแนกระดับคุณภาพน้ำในเกณฑ์ต่างๆ ของแบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE ซึ่งแสดงการกระจายตัวของค่า SHAP ของแต่ละพารามิเตอร์น้ำ พบว่า พารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก โดยพิจารณาจากค่า SHAP สัมบูรณ์เฉลี่ย (Mean absolute SHAP value) และเรียงลำดับจากมากไปน้อย ได้แก่ BOD, NH_3-N, TCB, FCB และ DO ตามลำดับ (ภาพประกอบ 41A) สำหรับพารามิเตอร์น้ำที่มีผลต่อการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม พอใช้ และดี โดยเรียงลำดับจากมากไปน้อย ได้แก่ BOD, FCB, TCB,

NH_3-N และ DO ตามลำดับ (ภาพประกอบ 41B,C,D) และจากระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมและเสื่อมโทรมมาก BOD, FCB, TCB และ NH_3-N มีความสัมพันธ์เชิงบวกต่อการจำแนกระดับคุณภาพน้ำของทั้ง 2 เกณฑ์ดังกล่าว สังเกตจากค่าพารามิเตอร์หรือลักษณะเฉพาะ (Feature value) มีค่าสูง (จุดสีแดง) ค่า SHAP มีค่าเป็นบวก ในทางกลับกัน DO มีความสัมพันธ์เชิงลบ สังเกตจากค่าพารามิเตอร์หรือลักษณะเฉพาะ (Feature value) มีค่าสูง (สีแดง) ค่า SHAP มีค่าติดลบ นอกจากนี้ เมื่อพิจารณาพารามิเตอร์ที่มีผลต่อการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก (ภาพประกอบ 41E) พบว่า BOD มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากมากที่สุด รองลงมา ได้แก่ DO, NH_3-N, FCB และ TCB ตามลำดับ โดย BOD, FCB, TCB และ NH_3-N มีความสัมพันธ์เชิงลบ แต่ DO มีความสัมพันธ์เชิงบวกต่อการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เช่นเดียวกับระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดี

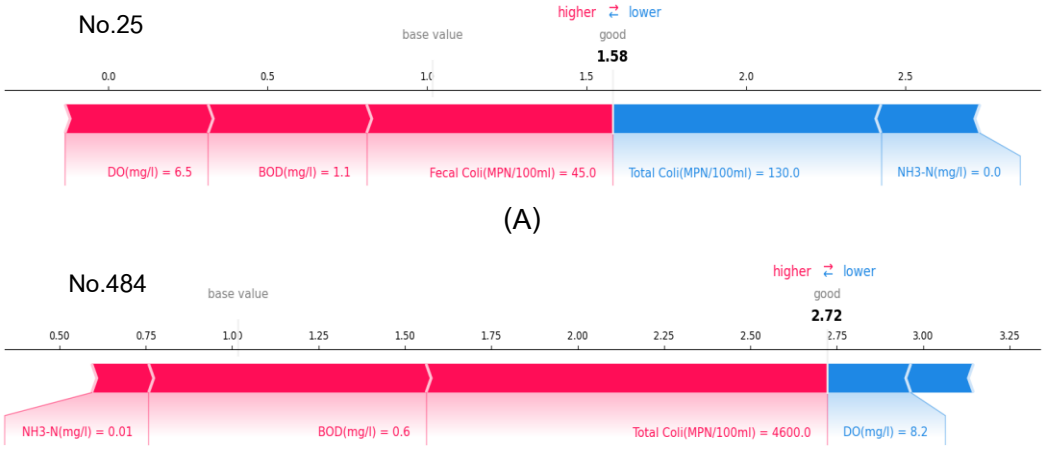


ภาพประกอบ 42 ค่า SHAP ของตัวอย่างที่แบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE ที่จำแนกระดับคุณภาพน้ำผิดพลาดจากระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมาก ได้แก่ (A) ตัวอย่างที่ 56 (B) ตัวอย่างที่ 520 และ (C) ตัวอย่างที่ 524 (สีของลูกศรแสดงถึงพารามิเตอร์ทำให้แบบจำลองจำแนกระดับคุณภาพน้ำอยู่ในเกณฑ์ดีมากได้มากขึ้น (สีแดง) หรือน้อยลง (สีน้ำเงิน) และความกว้างของลูกศรแสดงถึงค่าความสำคัญของพารามิเตอร์)

เมื่อพิจารณาตัวอย่างที่แบบจำลองจำแนกผิดพลาดจากระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 3 ตัวอย่าง (ภาพประกอบ 42) พบว่า ตัวอย่างที่ 56 พารามิเตอร์น้ำ DO (11.00 mg/l) ส่งผลต่อการจำแนกระดับคุณภาพน้ำของตัวอย่างนี้มากที่สุด ซึ่ง DO มีความสัมพันธ์เชิงลบกับการจำแนกระดับคุณภาพน้ำ พารามิเตอร์ที่มีผลลำดับถัดมา ได้แก่ NH_3-N (0.00 mg/l) และ TCB (18.00 MPN/100ml) ซึ่งพารามิเตอร์น้ำดังกล่าว ส่งผลให้แบบจำลองจำแนกตัวอย่างนี้เป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก สำหรับตัวอย่างที่ 520 พารามิเตอร์น้ำที่ส่งผลต่อการจำแนกระดับคุณภาพน้ำของตัวอย่างนี้มากที่สุด คือ BOD (2.10 mg/l) ซึ่งมีความสัมพันธ์เชิงลบกับการจำแนกระดับคุณภาพน้ำ พารามิเตอร์ที่มีผลลำดับถัดมา ได้แก่ DO (7.80 mg/l) และ NH_3-N (0.00 mg/l) โดยทั้ง 2 พารามิเตอร์ส่งผลให้แบบจำลองจำแนกตัวอย่างนี้เป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก และตัวอย่างที่ 524 พารามิเตอร์น้ำ DO (8.00 mg/l) ส่งผลต่อการจำแนกระดับคุณภาพน้ำของตัวอย่างนี้มากที่สุด รองลงมา ได้แก่ TCB (16,000.00 MPN/100ml) และ NH_3-N (0.0 mg/l) ตามลำดับ โดย DO และ NH_3-N ส่งผลให้แบบจำลองจำแนกตัวอย่างนี้เป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก

จากตัวอย่างที่จำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากทั้ง 3 ตัวอย่าง จะเห็นได้ว่า DO และ NH_3-N เป็นพารามิเตอร์ที่มีความสำคัญสูงอยู่ใน 3 อันดับแรก สำหรับการจำแนกระดับคุณภาพน้ำเป็นเกณฑ์ดังกล่าวสูงที่สุดของทั้ง 3 ตัวอย่าง ซึ่งสอดคล้องกับภาพประกอบ 41E ที่ BOD, DO และ NH_3-N มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำในเกณฑ์ดีมาก

สำหรับตัวอย่างที่แบบจำลองจำแนกผิดพลาดจากระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากเป็นดีจำนวน 2 ตัวอย่าง (ภาพประกอบ 43) พบว่า ตัวอย่างที่ 25 พารามิเตอร์น้ำที่มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำมากที่สุด ได้แก่ TCB (130.00 MPN/100ml), FCB (45.00 MPN/100ml) และ BOD (1.10 mg/l) ตามลำดับ โดย FCB และ BOD ส่งผลให้แบบจำลองจำแนกตัวอย่างนี้เป็นให้อยู่ในเกณฑ์ดี แต่ TCB มีความสัมพันธ์เชิงลบกับการจำแนกระดับคุณภาพน้ำ และตัวอย่างที่ 484 พารามิเตอร์น้ำที่มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำมากที่สุด ได้แก่ TCB (4,600.00 MPN/100ml), BOD (0.60 mg/l) และ DO (8.20 mg/l) ตามลำดับ ซึ่ง BOD และ FCB ส่งผลให้แบบจำลองจำแนกตัวอย่างนี้เป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดี ในขณะที่ DO มีความสัมพันธ์เชิงลบกับการจำแนกระดับคุณภาพน้ำ



ภาพประกอบ 43 ค่า SHAP ของตัวอย่างที่แบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE จำแนกระดับคุณภาพน้ำผิดพลาดจากระดับคุณภาพที่อยู่ในเกณฑ์ดีมากเป็นดี ได้แก่ (A) ตัวอย่างที่ 25 และ (B) ตัวอย่างที่ 484 (สีของลูกศรแสดงถึงพารามิเตอร์ทำให้แบบจำลองจำแนกระดับคุณภาพน้ำอยู่ในเกณฑ์ดีได้มากขึ้น (สีแดง) หรือน้อยลง (สีน้ำเงิน) และความกว้างของลูกศรแสดงถึงค่าความสำคัญของพารามิเตอร์)

จากตัวอย่างที่จำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากเป็นดี จะเห็นว่า TCB และ BOD เป็นพารามิเตอร์ที่มีความสำคัญสูงอยู่ใน 3 อันดับแรกทั้ง 2 ตัวอย่าง โดยสอดคล้องกับภาพประกอบ 41D ซึ่งปรากฏว่า BOD, FCB และ TCB อยู่ใน 3 อันดับแรกของพารามิเตอร์ที่มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำในเกณฑ์ดี

ผลลัพธ์ของแบบจำลองอนุกรมเวลาสำหรับทำนายคุณภาพแม่น้ำ

การศึกษาการทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของประเทศไทยด้วยแบบจำลองอนุกรมเวลา ใช้ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ตั้งแต่ปี พ.ศ. 2552 ถึง พ.ศ. 2564 โดยเก็บข้อมูลดังกล่าวทุกๆ 3 เดือน หรือ 4 ครั้งต่อปี ด้วยเหตุนี้ การแบ่งข้อมูลสำหรับนำเข้าแบบจำลองอนุกรมเวลา จึงประกอบด้วยข้อมูลชุดฝึกฝน (Training data) ใช้ข้อมูลระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2560 และข้อมูลชุดทดสอบ (Test data) ใช้ข้อมูลระหว่างปี 2561 ถึง พ.ศ. 2564

งานวิจัยนี้เลือกข้อมูลของสถานีตรวจวัดคุณภาพน้ำ 1 สถานี ใช้เป็นตัวแทนของแม่น้ำแต่ละสาย โดยมีข้อพิจารณา ได้แก่ เป็นสถานีตรวจวัดคุณภาพน้ำที่อยู่ก่อนแม่น้ำ 2 สาย

มาบรรจบกัน (ภาพประกอบ 7) และจำนวนข้อมูลต้องมีปริมาณไม่น้อยเกินไป จากเงื่อนไขดังกล่าว ทำให้เลือกสถานี PI06, WA02, YO01 และ NA02 เป็นตัวแทนของแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ตามลำดับ

แบบจำลองอนุกรมเวลาที่ใช้ทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ได้แก่ ARIMA, ARIMAX, SARIMA และ SARIMAX ซึ่งการเลือกรูปแบบพารามิเตอร์ (p, d, q) ของแบบจำลอง ARIMA และ ARIMAX และ (p, d, q)(P, D, Q) ของแบบจำลอง SARIMA และ SARIMAX พิจารณารูปแบบพารามิเตอร์ที่ให้ค่า AIC ต่ำที่สุด สำหรับเลือกพารามิเตอร์ S ของ SARIMA และ SARIMAX มีค่าเท่ากับ 4 ซึ่งสอดคล้องกับข้อมูลที่ได้จากการแยกองค์ประกอบของอนุกรมเวลาที่มีรูปแบบของฤดูกาล (Seasonal) สำหรับตัวแปรภายนอก (Exogenous variable) ของแบบจำลอง ARIMAX และ SARIMAX นั้น แบ่งเป็น 2 กรณี ได้แก่ กรณีที่ 1 กำหนดให้ตัวแปรภายนอก คือ BOD (Exog = BOD) เนื่องจาก BOD มีผลต่อการจำแนกระดับคุณภาพน้ำมากที่สุด และกรณีที่ 2 กำหนดให้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N (Exog = DO, BOD, TCB, FCB, NH_3-N) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองของทั้ง 2 กรณีดังกล่าว

ผลการศึกษาการทำนายค่า WQI ของข้อมูลคุณภาพน้ำที่ได้จากสถานีตรวจวัดน้ำ PI06, WA02, YO01 และ NA02 ด้วยแบบจำลองอนุกรมเวลา เป็นดังนี้

1. การทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของแม่น้ำปิง สถานี PI06

สถานีตรวจวัดคุณภาพน้ำ PI06 ของแม่น้ำปิง มีข้อมูลทั้งสิ้น 52 ตัวอย่าง ประกอบด้วย ข้อมูลชุดฝึกฝน 36 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2552 - 2560 และข้อมูลชุดทดสอบ 16 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2561 - 2564 ผลลัพธ์ที่ได้จากการศึกษาเป็นดังตาราง 13

จากการเลือกรูปแบบพารามิเตอร์ (p, d, q) ของแบบจำลอง ARIMA และ ARIMAX และ (p, d, q)(P, D, Q) ของแบบจำลอง SARIMA และ SARIMAX โดยพิจารณาจากรูปแบบที่ทำให้ค่า AIC มีค่าต่ำที่สุด ผลที่ได้เป็นดังนี้ ARIMA(1, 1, 1), ARIMAX(1, 1, 1) เมื่อกำหนดให้ตัวแปร Exogenous คือ BOD, ARIMAX(0, 1, 1) เมื่อกำหนดให้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5 (DO, BOD, TCB, FCB และ NH_3-N), SARIMA(2, 1, 1)(1, 2, 1)$_4$, SARIMAX(1, 1, 1)(0, 0, 0)$_4$ เมื่อกำหนดให้ตัวแปร Exogenous คือ BOD และ SARIMAX(3, 1, 2)(2, 0, 0)$_4$ เมื่อกำหนดให้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5

ตาราง 13 ประสิทธิภาพการการทำนายค่า WQI ของแบบจำลองอนุกรมเวลา

สถานีตรวจวัด	แบบจำลอง	ตัวชี้วัดประสิทธิภาพ		
		MAE	RMSE	MAPE (%)
แม่น้ำปิง (PI06)	ARIMA(1,1,1)	5.85	8.60	7.73
	ARIMAX(1, 1, 1) (exog= BOD)	6.01	7.62	8.32
	ARIMAX(0, 1, 1) (exog=5 water parameters)	4.35	5.90	6.15
	$SARIMA(2, 1, 1)(1, 2, 1)_4$	28.68	32.31	40.36
	$SARIMAX(1, 1, 1)(0, 0, 0)_4$ (exog= BOD)	6.01	7.62	8.32
	$SARIMAX(3, 1, 2)(2, 0, 0)_4$ (exog=5 water parameters)	6.51	7.70	9.56
แม่น้ำวัง (WA02)	ARIMA(3, 2, 1)	17.91	19.82	26.94
	ARIMAX(0, 1, 1) (exog= BOD)	7.47	9.19	10.88
	ARIMAX(2, 1, 2) (exog=5 water parameters)	6.36	7.55	9.35
	$SARIMA(2, 1, 2)(2, 2, 0)_4$	21.95	26.17	30.45
	$SARIMAX(2, 1, 2)(0, 0, 1)_4$ (exog= BOD)	8.54	9.49	12.66
	$SARIMAX(2, 1, 1)(0, 0, 2)_4$ (exog=5 water parameters)	6.57	7.93	9.66
แม่น้ำยม (YO01)	ARIMA(0, 1, 3)	7.15	10.15	12.15
	ARIMAX(0, 1, 1) (exog= BOD)	7.33	8.33	12.78
	ARIMAX(0, 1, 1) (exog=5 water parameters)	5.85	6.62	10.04
	$SARIMA(3, 0, 2)(3, 2, 0)_4$	11.68	13.38	19.08
	$SARIMAX(0, 1, 1)(0, 0, 1)_4$ (exog= BOD)	7.82	9.04	13.54
	$SARIMAX(0, 1, 1)(0, 0, 1)_4$ (exog=5 water parameters)	6.95	8.27	11.85
แม่น้ำน่าน (NA02)	ARIMA(2, 1, 3)	1.99	2.48	3.14
	ARIMAX(0, 1, 1) (exog= BOD)	2.43	2.85	3.89
	ARIMAX(3, 1, 3) (exog=5 water parameters)	2.04	2.59	3.21
	$SARIMA(3, 1, 0)(1, 2, 1)_4$	14.66	18.46	23.82
	$SARIMAX(0, 1, 1)(0, 0, 0)_4$ (exog= BOD)	2.43	2.85	3.89
	$SARIMAX(2, 1, 0)(2, 0, 3)_4$ (exog=5 water parameters)	2.29	2.93	3.62

ตาราง 14 ผลการทดสอบ Ljung-Box ของแบบจำลองอนุกรมเวลา

สถานีตรวจวัด	แบบจำลอง	Ljung-Box Q	
		ค่า Ljung-Box	p-value
แม่น้ำปิง (PI06)	ARIMA(1,1,1)	0.01	0.93
	ARIMAX(1, 1, 1) (exog= BOD)	0.04	0.85
	ARIMAX(0, 1, 1) (exog=5 water parameters)	0.19	0.66
	SARIMA(2, 1, 1)(1, 2, 1)$_4$	0.01	0.94
	SARIMAX(1, 1, 1)(0, 0, 0)$_4$ (exog= BOD)	0.04	0.85
	SARIMAX(3, 1, 2)(2, 0, 0)$_4$ (exog=5 water parameters)	0.00	0.98
แม่น้ำวัง (WA02)	ARIMA(3, 2, 1)	0.02	0.88
	ARIMAX(0, 1, 1) (exog= BOD)	0.35	0.56
	ARIMAX(2, 1, 2) (exog=5 water parameters)	0.23	0.63
	SARIMA(2, 1, 2)(2, 2, 0)$_4$	0.15	0.70
	SARIMAX(2, 1, 2)(0, 0, 1)$_4$ (exog= BOD)	1.06	0.30
	SARIMAX(2, 1, 1)(0, 0, 2)$_4$ (exog=5 water parameters)	1.43	0.23
แม่น้ำยม (YO01)	ARIMA(0, 1, 3)	0.11	0.74
	ARIMAX(0, 1, 1) (exog= BOD)	0.31	0.58
	ARIMAX(0, 1, 1) (exog=5 water parameters)	0.01	0.94
	SARIMA(3, 0, 2)(3, 2, 0)$_4$	0.01	0.93
	SARIMAX(0, 1, 1)(0, 0, 1)$_4$ (exog= BOD)	0.11	0.74
	SARIMAX(0, 1, 1)(0, 0, 1)$_4$ (exog=5 water parameters)	0.35	0.55
แม่น้ำน่าน (NA02)	ARIMA(2, 1, 3)	0.11	0.74
	ARIMAX(0, 1, 1) (exog= BOD)	0.10	0.75
	ARIMAX(3, 1, 3) (exog=5 water parameters)	0.27	0.61
	SARIMA(3, 1, 0)(1, 2, 1)$_4$	0.18	0.68
	SARIMAX(0, 1, 1)(0, 0, 0)$_4$ (exog= BOD)	0.10	0.75
	SARIMAX(2, 1, 0)(2, 0, 3)$_4$ (exog=5 water parameters)	0.36	0.55

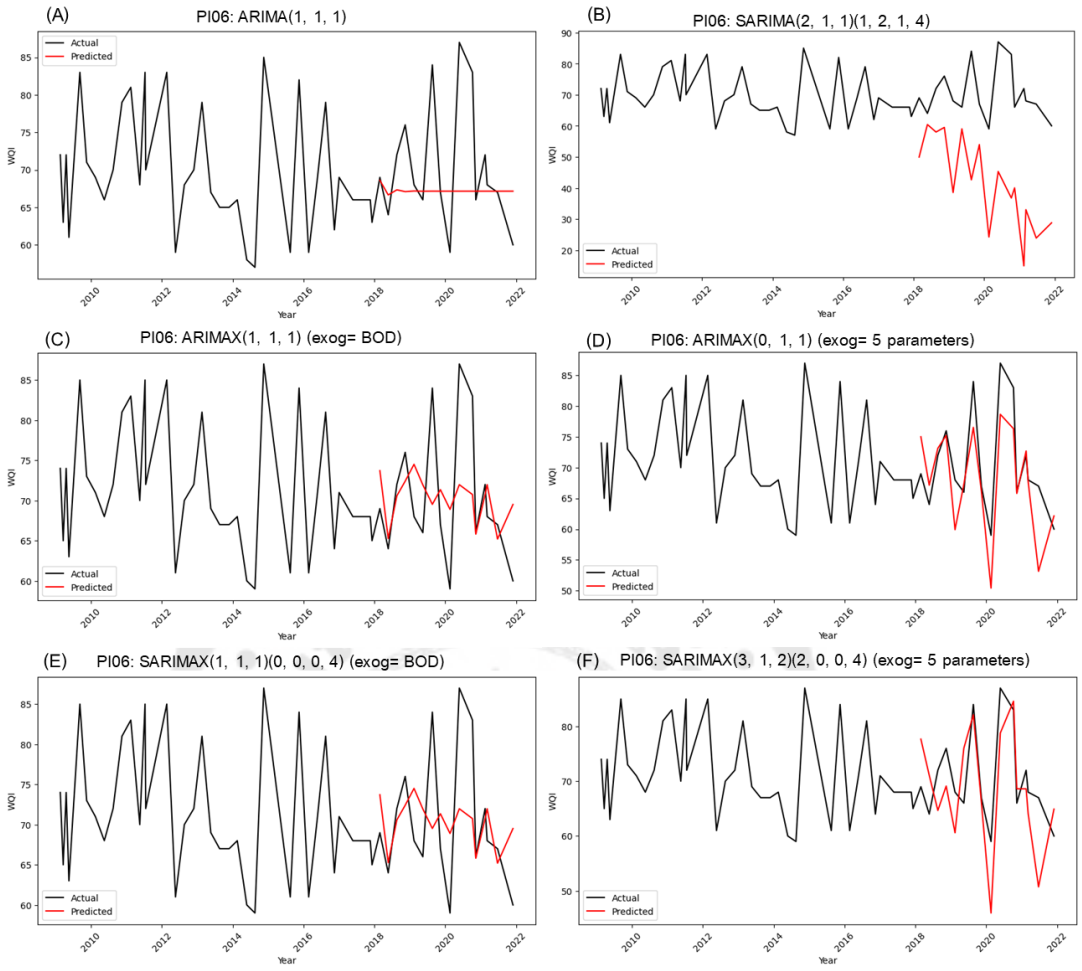
ตาราง 15 แบบจำลองที่ให้ค่าความคลาดเคลื่อนจากการทำนายน้อยที่สุดของแม่น้ำแต่ละสาย

สถานีตรวจวัด	แบบจำลอง	ตัวชี้วัดประสิทธิภาพ		
		MAE	RMSE	MAPE (%)
แม่น้ำปิง (PI06)	ARIMAX(0, 1, 1) (exog = 5 water parameters)	4.35	5.90	6.15
แม่น้ำวัง (WA02)	ARIMAX(2, 1, 2) (exog = 5 water parameters)	6.36	7.55	9.35
แม่น้ำยม (YO01)	ARIMAX(0, 1, 1) (exog = 5 water parameters)	5.85	6.62	10.04
แม่น้ำน่าน (NA02)	ARIMA(2, 1, 3)	1.99	2.48	3.14

ผลจากการทำนายค่า WQI ที่ได้จากแบบจำลอง ARIMA(1, 1, 1) เป็นไปดังภาพประกอบ 44A เห็นได้ว่า ผลการทำนายข้อมูลชุดทดสอบตั้งแต่ปี พ.ศ. 2562 (2019) เส้นกราฟสีแดงแสดงถึงค่าที่ได้จากการทำนายค่อนข้างเป็นเส้นตรง ทำให้ค่าความคลาดเคลื่อนของแบบจำลอง ได้แก่ ค่า MAE เท่ากับ 5.85 ค่า RMSE เท่ากับ 8.60 และค่า MAPE เท่ากับ 7.73% (ตาราง 13)

ผลการศึกษาแบบจำลอง ARIMAX สำหรับทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ PI06 พบว่า ARIMAX(0, 1, 1) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5 (ค่า MAE เท่ากับ 4.35 ค่า RMSE เท่ากับ 5.90 และค่า MAPE เท่ากับ 6.15%) มีค่าความคลาดเคลื่อนจากการทำนายน้อยกว่า ARIMAX(1, 1, 1) ที่ตัวแปร Exogenous เป็น BOD (ค่า MAE เท่ากับ 6.01 ค่า RMSE เท่ากับ 7.62 และค่า MAPE เท่ากับ 8.32%) (ตาราง 13) เมื่อพิจารณาจากภาพประกอบ 44C,D แสดงให้เห็นว่า เมื่อใช้ตัวแปร Exogenous เป็นทั้ง 5 พารามิเตอร์น้ำ เส้นกราฟสีแดงที่ได้จากการทำนายมีค่าใกล้เคียงกับค่าที่ได้จากข้อมูลจริงมากกว่ากำหนดตัวแปร Exogenous เป็น BOD

จากภาพประกอบ 44B พบว่า ค่าที่ได้จากการทำนายของชุดทดสอบของแบบจำลอง SARIMA(2, 1, 1)(1, 2, 1)$_4$ มีค่าน้อยกว่าข้อมูลจริงทุกตัวอย่างอย่างเห็นได้ชัด ทำให้ค่า MAE, RMSE และ MAPE ของแบบจำลองดังกล่าว เท่ากับ 28.68, 32.31 และ 40.36% ตามลำดับ (ตาราง 13)



ภาพประกอบ 44 การเปรียบเทียบระหว่างค่า WQI ที่ได้จากทำนายกับข้อมูลจริงของสถานีตรวจวัดคุณภาพน้ำ PI06 ที่ได้จากแบบจำลอง (A) ARIMA(1, 1, 1) (B) SARIMA(2, 1, 1)(1, 2, 1)$_4$ (C) ARIMAX(1, 1, 1) ตัวแปร Exogenous เป็น BOD (D) ARIMAX(0, 1, 1) ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N (E) SARIMAX(1, 1, 1)(0, 0, 0)$_4$ ตัวแปร Exogenous เป็น BOD และ (F) SARIMAX(3, 1, 2)(2, 0, 0)$_4$ ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N

สำหรับประสิทธิภาพการทำนายค่า WQI ของแบบจำลอง SARIMAX พบว่า SARIMAX(1, 1, 1)(0, 0, 0)$_4$ ตัวแปร Exogenous เป็น BOD เกิดความผิดพลาดเมื่อทำนายค่า WQI ของข้อมูลชุดทดสอบน้อยกว่า SARIMAX(3, 1, 2)(2, 0, 0)$_4$ ที่ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N โดยค่า MAE, RMSE และ MAPE ของ SARIMAX(1, 1, 1)(0, 0, 0)$_4$ ที่ตัวแปร Exogenous คือ BOD เท่ากับ 6.01, 7.62 และ 8.32% ตามลำดับ และค่า MAE, RMSE

และ MAPE ของ SARIMAX(3, 1, 2)(2, 0, 0)$_4$ ที่ตัวแปร Exogenous เป็นทั้ง 5 พารามิเตอร์น้ำ เท่ากับ 6.51, 7.70 และ 9.56% ตามลำดับ (ตาราง 13) และเมื่อพิจารณาพารามิเตอร์ (P, D, Q)$_S$ แทนส่วน Seasonality ของ SARIMAX(1, 1, 1)(0, 0, 0)$_4$ ตัวแปร Exogenous เป็น BOD พบว่า (P, D, Q)$_S$ เท่ากับ (0, 0, 0)$_4$ แสดงว่า Seasonality หรือฤดูกาลไม่มีผลต่อการทำนายค่า WQI ของแบบจำลอง

เมื่อพิจารณาความเป็นอิสระต่อกันของความคลาดเคลื่อน (Residual) ด้วยการทดสอบ Ljung-Box Q ของแบบจำลองที่ใช้ทำนายค่า WQI สถานีตรวจวัด PI06 พบว่า ARIMA(1, 1, 1) (Ljung-Box = 0.01 และ p-value = 0.93), ARIMAX(1, 1, 1) (Exog = BOD) (Ljung-Box = 0.04 และ p-value = 0.85), ARIMAX(0, 1, 1) (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 0.19 และ p-value = 0.66), SARIMA(2, 1, 1)(1, 2, 1)$_4$ (Ljung-Box = 0.01 และ p-value = 0.94), SARIMAX(1, 1, 1)(0, 0, 0)$_4$ (Exog = BOD) (Ljung-Box = 0.04 และ p-value = 0.85) และ SARIMAX(3, 1, 2)(2, 0, 0)$_4$ (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 0.00 และ p-value = 0.98) มีค่า p-value มากกว่า 0.05 (ที่ระดับนัยสำคัญ 0.05) (ตาราง 14) กล่าวคือ ยอมรับสมมติฐานหลัก (H_0) ที่ว่า ความคลาดเคลื่อนเป็นอิสระต่อกัน ดังนั้นแบบจำลองจึงมีความเหมาะสมกับการนำไปใช้ทำนายค่า WQI

จากผลการศึกษาดังกล่าว แสดงให้เห็นว่า ARIMAX(0, 1, 1) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N มีความแม่นยำสำหรับการทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ PI06 มากที่สุด (ตาราง 15) เนื่องจากให้ค่า MAE และ MAPE น้อยที่สุด รองลงมา ได้แก่ ARIMA(1, 1, 1), SARIMAX (1, 1, 1)(0, 0, 0)$_4$ (Exog = BOD) และ SARIMA(2, 1, 1)(1, 2, 1)$_4$ ตามลำดับ

สำหรับค่าพารามิเตอร์ของแบบจำลอง ARIMAX(0, 1, 1) ที่ใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N ซึ่งเป็นแบบจำลองที่ทำนายค่า WQI ของข้อมูลแม่น้ำปิง PI06 ได้ความคลาดเคลื่อนน้อยที่สุด (ตาราง 16) พบว่า BOD (ค่าสัมประสิทธิ์เท่ากับ -0.3782) TCB (ค่าสัมประสิทธิ์ เท่ากับ -0.7496) FCB (ค่าสัมประสิทธิ์ เท่ากับ -0.4263) และ NH_3-N (ค่าสัมประสิทธิ์ เท่ากับ -0.4046) มีค่า p-value เท่ากับ 0.011, 0.001, 0.042 และ 0.001 ตามลำดับ จะเห็นได้ว่าค่า p-value น้อยกว่า 0.05 ทำให้ปฏิเสธสมมติฐานหลัก (H_0) กล่าวคือ ค่าพารามิเตอร์มีค่าแตกต่างจาก 0 ที่ระดับนัยสำคัญ 0.05 จึงสรุปว่า BOD, TCB, FCB, และ NH_3-N มีความสำคัญต่อการทำนายค่า WQI ของแบบจำลองดังกล่าว

ตาราง 16 การประมาณค่าพารามิเตอร์ของข้อมูลแม่น้ำปิง PI06 แบบจำลอง ARIMAX (0, 1, 1) (exog = 5 water parameters)

พารามิเตอร์	ค่าสัมประสิทธิ์	ค่าความคลาดเคลื่อนมาตรฐาน (Standard Error)	Z-score	p-value
DO	-0.0653	0.158	-0.413	0.680
BOD	-0.3782	0.148	-2.556	0.011
TCB	-0.7496	0.225	-3.334	0.001
FCB	-0.4263	0.209	-2.036	0.042
NH_3-N	-0.4046	0.122	-3.328	0.001
MA(1)	-0.9980	7.929	-0.126	0.900

2. การทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของแม่น้ำวัง สถานี WA02

สถานีตรวจวัดคุณภาพน้ำ WA02 ของแม่น้ำวัง มีข้อมูลทั้งสิ้น 53 ตัวอย่าง ประกอบด้วย ข้อมูลชุดฝึกฝน 38 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2552 - 2560 และข้อมูลชุดทดสอบ 15 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2561 - 2564 ผลลัพธ์ที่ได้จากการศึกษาเป็นดังตาราง 13

จากการเลือกรูปแบบพารามิเตอร์ (p, d, q) ของแบบจำลอง ARIMA และ ARIMAX และ (p, d, q)(P, D, Q) ของแบบจำลอง SARIMA และ SARIMAX โดยพิจารณาจากรูปแบบที่ทำให้ค่า AIC มีค่าต่ำที่สุด ผลที่ได้เป็นดังนี้ ARIMA(3, 2, 1), ARIMAX(0, 1, 1) ที่กำหนดให้ตัวแปร Exogenous คือ BOD, ARIMAX(2, 1, 2) เมื่อตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5 (DO, BOD, TCB, FCB และ NH_3-N), SARIMA(2, 1, 2)(2, 2, 0)$_4$, SARIMAX(2, 1, 2)(0, 0, 1)$_4$ เมื่อตัวแปร Exogenous คือ BOD และ SARIMAX(2, 1, 1)(0, 0, 2)$_4$ เมื่อกำหนดให้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5

ผลจากการทำนายค่า WQI ที่ได้จากแบบจำลอง ARIMA(3, 2, 1) เป็นไปดังภาพประกอบ 45A และตาราง 13 พบว่า ผลการทำนายข้อมูลชุดทดสอบส่วนใหญ่ทำนายคลาดเคลื่อนจากข้อมูลจริง เป็นผลให้ค่าความคลาดเคลื่อนของแบบจำลอง ได้แก่ ค่า MAE เท่ากับ 17.91 ค่า RMSE เท่ากับ 19.82 และค่า MAPE เท่ากับ 26.94%

ผลการศึกษาแบบจำลอง ARIMAX เพื่อทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ PI06 พบว่า ARIMAX(2, 1, 2) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5 (ค่า MAE เท่ากับ 6.36 ค่า RMSE เท่ากับ 7.55 และค่า MAPE เท่ากับ 9.35%) ให้ค่า

ความคลาดเคลื่อนน้อยกว่า ARIMAX(0, 1, 1) ที่ตัวแปร Exogenous เป็น BOD (ค่า MAE เท่ากับ 7.47 ค่า RMSE เท่ากับ 9.19 และค่า MAPE เท่ากับ 10.88%) (ตาราง 13) เมื่อพิจารณาจากภาพประกอบ 45D พบว่า เมื่อใช้ตัวแปร Exogenous เป็นทั้ง 5 พารามิเตอร์น้ำ แบบจำลองสามารถทำนายข้อมูลตั้งแต่ปี พ.ศ. 2563 - 2564 (2011 - 2012) ได้ใกล้เคียงกับข้อมูลจริง

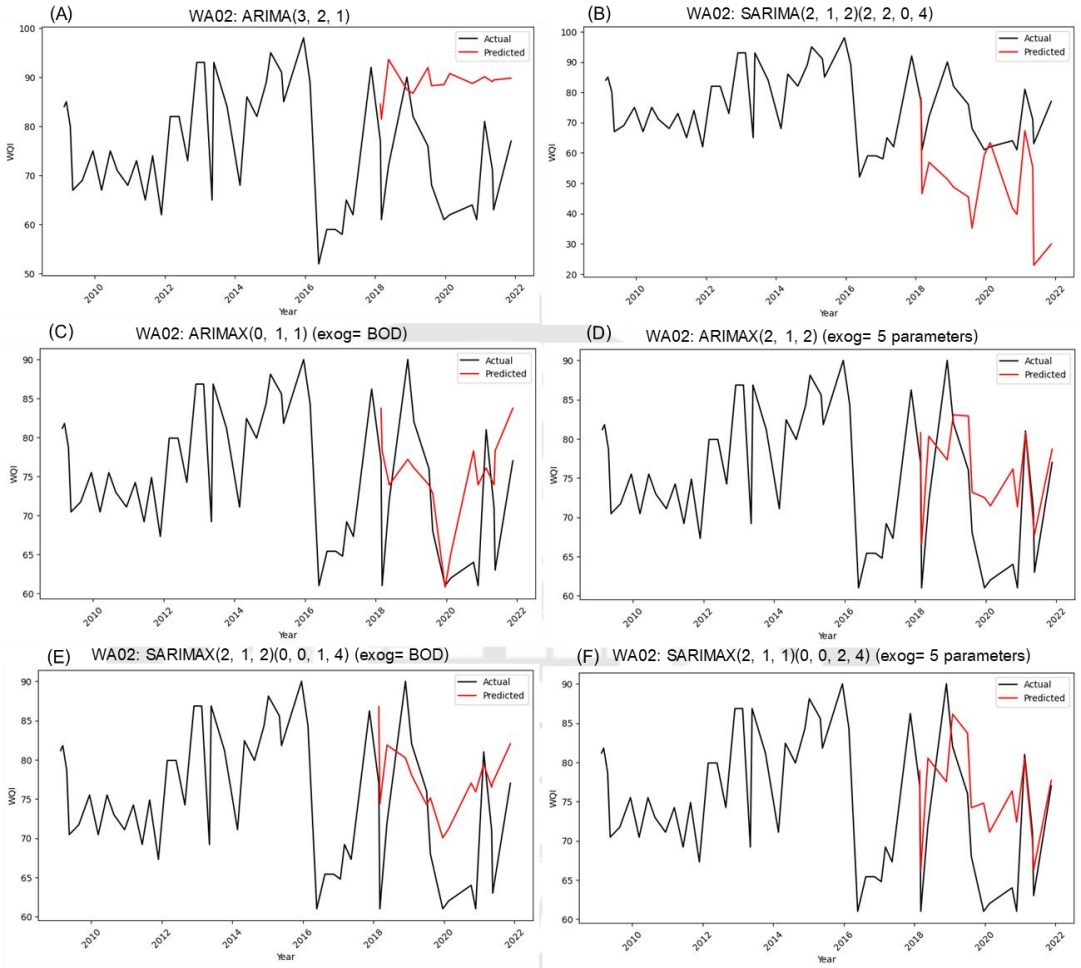
จากภาพประกอบ 45B และตาราง 13 พบว่า ค่าที่ได้จากการทำนายของชุดทดสอบของแบบจำลอง SARIMA(2, 1, 2)(2, 2, 0)$_4$ มีความคลาดเคลื่อนค่อนข้างสูง ทำให้ค่า MAE, RMSE และ MAPE ของแบบจำลอง เท่ากับ 21.95, 26.17 และ 30.45% ตามลำดับ

สำหรับประสิทธิภาพการทำนายค่า WQI ของแบบจำลอง SARIMAX พบว่า SARIMAX(2, 1, 1)(0, 0, 2)$_4$ ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N มีความผิดพลาดที่เกิดจากการทำนายค่า WQI ของข้อมูลชุดทดสอบน้อยกว่า SARIMAX(2, 1, 2)(0, 0, 1)$_4$ ที่ตัวแปร Exogenous เป็น BOD โดยค่า MAE, RMSE และ MAPE ของ SARIMAX (2, 1, 1)(0, 0, 2)$_4$ ที่ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N เท่ากับ 6.57, 7.93 และ 9.66% ตามลำดับ และค่า MAE, RMSE และ MAPE ของ SARIMAX(2, 1, 2)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น BOD เท่ากับ 8.54, 9.49 และ 12.66% ตามลำดับ (ตาราง 13 และภาพประกอบ 45E,F)

เมื่อพิจารณาความเป็นอิสระต่อกันของความคลาดเคลื่อน (Residual) ด้วยการทดสอบ Ljung-Box Q ของแบบจำลองที่ใช้ทำนายค่า WQI สถานีตรวจวัด WA02 พบว่า ARIMA(3, 2, 1) (Ljung-Box = 0.02 และ p-value = 0.88), ARIMAX(0, 1, 1) (Exog = BOD) (Ljung-Box = 0.35 และ p-value = 0.56), ARIMAX(2, 1, 2) (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 0.23 และ p-value = 0.63), SARIMA(2, 1, 2)(2, 2, 0)$_4$ (Ljung-Box = 0.15 และ p-value = 0.70), SARIMAX(2, 1, 2)(0, 0, 1)$_4$ (Exog = BOD) (Ljung-Box = 1.06 และ p-value = 0.30) และ SARIMAX(2, 1, 1)(0, 0, 2)$_4$ (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 1.43 และ p-value = 0.23) มีค่า p-value มากกว่า 0.05 (ที่ระดับนัยสำคัญ 0.05) (ตาราง 14) กล่าวคือ ยอมรับสมมติฐานหลัก (H_0) ที่ว่า ความคลาดเคลื่อนเป็นอิสระต่อกัน ดังนั้นแบบจำลองจึงมีความเหมาะสมกับการนำไปใช้ทำนายค่า WQI

จากผลการศึกษาดังกล่าว แสดงให้เห็นว่า ARIMAX(2, 1, 2) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N มีความแม่นยำสำหรับการทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ WA02 มากที่สุด (ตาราง 15) เนื่องจาก

ให้ค่า MAE, RMSE และ MAPE น้อยที่สุด รองลงมา ได้แก่ SARIMAX(2, 1, 1)(0, 0, 2)$_4$ (Exog = DO, BOD, TCB, FCB, NH_3-N), ARIMA(3, 2, 1) และ SARIMA(2, 1, 2)(2, 2, 0)$_4$ ตามลำดับ



ภาพประกอบ 45 การเปรียบเทียบระหว่างค่า WQI ที่ได้จากทำนายกับข้อมูลจริงของสถานีตรวจวัดคุณภาพน้ำ WA02 ที่ได้จากแบบจำลอง (A) ARIMA(3, 2, 1) (B) SARIMA(2, 1, 2)(2, 2, 0)$_4$ (C) ARIMAX(0, 1, 1) ตัวแปร Exogenous เป็น BOD (D) ARIMAX(2, 1, 2) ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N (E) SARIMAX(2, 1, 2)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น BOD และ (F) SARIMAX(2, 1, 1)(0, 0, 2)$_4$ ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N

สำหรับค่าพารามิเตอร์ของแบบจำลอง ARIMAX(2, 1, 2) ที่ใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N ซึ่งเป็นแบบจำลองที่ทำนายค่า WQI ของข้อมูลแม่น้ำวัง WA02 ได้ความคลาดเคลื่อนน้อยที่สุด (ตาราง 17) พบว่า DO (ค่าสัมประสิทธิ์

เท่ากับ 0.1759) BOD (ค่าสัมประสิทธิ์ เท่ากับ -0.5661) TCB (ค่าสัมประสิทธิ์ เท่ากับ 0.3221) FCB (ค่าสัมประสิทธิ์ เท่ากับ -0.4676) และ NH_3-N (ค่าสัมประสิทธิ์ เท่ากับ -0.4025) ค่า p-value เท่ากับ 0.039, 0.000, 0.005, 0.000 และ 0.006 ตามลำดับ ซึ่งค่า p-value น้อยกว่า 0.05 ทำให้ปฏิเสธสมมติฐานหลัก (H_0) กล่าวคือ ค่าพารามิเตอร์มีค่าแตกต่างจาก 0 ที่ระดับนัยสำคัญ 0.05 จึงสรุปว่า DO, BOD, TCB, FCB และ NH_3-N มีความสำคัญต่อการทำนายค่า WQI ของแบบจำลอง

ตาราง 17 การประมาณค่าพารามิเตอร์ของข้อมูลแม่น้ำวัง WA02 แบบจำลอง ARIMAX(2, 1, 2) (exog = 5 water parameters)

พารามิเตอร์	ค่าสัมประสิทธิ์	ค่าความคลาดเคลื่อนมาตรฐาน (Standard Error)	Z-score	p-value
DO	0.1759	0.085	2.065	0.039
BOD	-0.5661	0.097	-5.817	0.000
TCB	0.3221	0.114	2.826	0.005
FCB	-0.4676	0.120	-3.892	0.000
NH_3-N	-0.4025	0.145	-2.773	0.006
AR(1)	-1.0390	0.219	-4.753	0.000
AR(2)	-0.6796	0.230	-2.961	0.003
MA(1)	0.6134	3.740	0.164	0.870
MA(2)	-0.3794	1.449	-0.262	0.793

3. การทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของแม่น้ำยม สถานี YO01

สถานีตรวจวัดคุณภาพน้ำ YO01 ของแม่น้ำยม มีข้อมูลทั้งสิ้น 51 ตัวอย่าง ประกอบด้วย ข้อมูลชุดฝึกฝน 37 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2552 - 2560 และข้อมูลชุดทดสอบ 14 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2561 - 2564 ผลลัพธ์ที่ได้จากการศึกษาเป็นไปดังตาราง 13

จากการเลือกรูปแบบพารามิเตอร์ (p, d, q) ของแบบจำลอง ARIMA และ ARIMAX และ (p, d, q)(P, D, Q) ของแบบจำลอง SARIMA และ SARIMAX โดยพิจารณาจากรูปแบบที่ทำให้ค่า AIC มีค่าต่ำที่สุด ผลที่ได้เป็นดังนี้ ARIMA(0, 1, 3), ARIMAX(0, 1, 1) ที่กำหนดให้ตัวแปร Exogenous คือ BOD, ARIMAX(0, 1, 1) เมื่อตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5

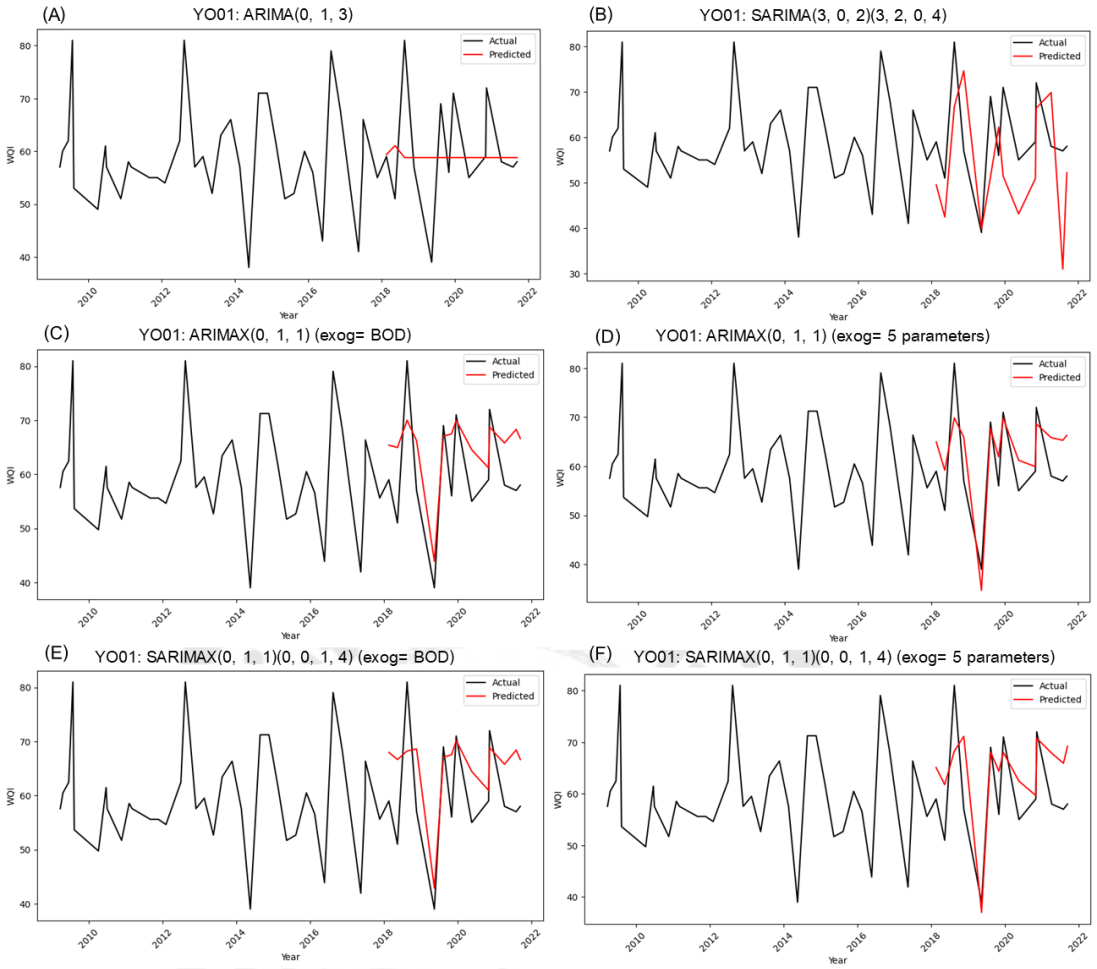
(DO, BOD, TCB, FCB และ NH_3-N), SARIMA(3, 0, 2)(3, 2, 0)$_4$, SARIMAX(0, 1, 1)(0, 0, 1)$_4$ เมื่อตัวแปร Exogenous คือ BOD และ SARIMAX(0, 1, 1)(0, 0, 1)$_4$ เมื่อกำหนดให้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5

ผลจากการทำนายค่า WQI ที่ได้จาก ARIMA(0, 1, 3) เป็นไปดังภาพประกอบ 46A เห็นได้ว่า ผลการทำนายข้อมูลชุดทดสอบตั้งแต่ปี พ.ศ. 2562 (2019) เส้นกราฟสีแดงแสดงถึงค่าที่ได้จากการทำนายค่อนข้างเป็นเส้นตรง ทำให้ค่าความคลาดเคลื่อนของแบบจำลอง ได้แก่ ค่า MAE เท่ากับ 7.15 ค่า RMSE เท่ากับ 10.15 และค่า MAPE เท่ากับ 12.15% (ตาราง 13)

ผลการศึกษาแบบจำลอง ARIMAX เพื่อทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ YO01 พบว่า ARIMAX(0, 1, 1) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5 (ค่า MAE เท่ากับ 5.85 ค่า RMSE เท่ากับ 6.62 และค่า MAPE เท่ากับ 10.04%) สามารถทำนายได้ค่าคลาดเคลื่อนน้อยกว่า ARIMAX(0, 1, 1) ที่ตัวแปร Exogenous เป็น BOD (ค่า MAE เท่ากับ 7.33 ค่า RMSE เท่ากับ 8.33 และค่า MAPE เท่ากับ 12.78%) (ตาราง 13) และจากภาพประกอบ 46C,D พบว่า เมื่อใช้ตัวแปร Exogenous เป็นทั้ง 5 พารามิเตอร์น้ำ แบบจำลองสามารถทำนายค่า WQI ได้ใกล้เคียงกับข้อมูลจริงมากกว่ากำหนดให้ตัวแปร Exogenous เป็น BOD

จากตาราง 13 และภาพประกอบ 46B พบว่า ผลที่ได้จากการทำนายของชุดทดสอบของแบบจำลอง SARIMA(3, 0, 2)(3, 2, 0)$_4$ ให้ค่า MAE, RMSE และ MAPE ของแบบจำลองเท่ากับ 11.68, 13.38 และ 19.07% ตามลำดับ

สำหรับประสิทธิภาพการทำนายค่า WQI ของแบบจำลอง SARIMAX พบว่า SARIMAX(0, 1, 1)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N มีความผิดพลาดที่เกิดจากการทำนายค่า WQI ของข้อมูลชุดทดสอบน้อยกว่า SARIMAX(0, 1, 1)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น BOD โดยค่า MAE, RMSE และ MAPE ของ SARIMAX(0, 1, 1)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N เท่ากับ 6.95, 8.28 และ 11.85% ตามลำดับ และค่า MAE, RMSE และ MAPE ของ SARIMAX(0, 1, 1)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น BOD เท่ากับ 7.82, 9.04 และ 13.54% ตามลำดับ (ตาราง 13 และภาพประกอบ 46E,F)



ภาพประกอบ 46 การเปรียบเทียบระหว่างค่า WQI ที่ได้จากทำนายกับข้อมูลจริงของสถานีตรวจวัดคุณภาพน้ำ YO01 ที่ได้จากแบบจำลอง (A) ARIMA(0, 1, 3) (B) SARIMA(3, 0, 2)(3, 2, 0)$_4$ (C) ARIMAX(0, 1, 1) ตัวแปร Exogenous เป็น BOD (D) ARIMAX(0, 1, 1) ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N (E) SARIMAX(0, 1, 1)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น BOD และ (F) SARIMAX(0, 1, 1)(0, 0, 1)$_4$ ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N

เมื่อพิจารณาความเป็นอิสระต่อกันของความคลาดเคลื่อน (Residual) ด้วยการทดสอบ Ljung-Box Q ของแบบจำลองที่ใช้ทำนายค่า WQI สถานีตรวจวัด YO01 พบว่า ARIMA(0, 1, 3) (Ljung-Box = 0.11 และ p-value = 0.74), ARIMAX(0, 1, 1) (Exog = BOD) (Ljung-Box = 0.31 และ p-value = 0.58), ARIMAX(0, 1, 1) (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 0.01 และ p-value = 0.94), SARIMA(3, 0, 2)(3, 2, 0)$_4$ (Ljung-Box =

0.01 และ p-value = 0.93), SARIMAX(0, 1, 1)(0, 0, 1)$_4$ (Exog = BOD) (Ljung-Box = 0.11 และ p-value = 0.74) และ SARIMAX(0, 1, 1)(0, 0, 1)$_4$ (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 0.35 และ p-value = 0.55) มีค่า p-value มากกว่า 0.05 (ที่ระดับนัยสำคัญ 0.05) (ตาราง 14) กล่าวคือ ยอมรับสมมติฐานหลัก (H_0) ที่ว่า ความคลาดเคลื่อนเป็นอิสระต่อกัน ดังนั้น แบบจำลองจึงมีความเหมาะสมกับการนำไปใช้ทำนายค่า WQI

จากผลการศึกษาดังกล่าว แสดงให้เห็นว่า ARIMAX(0, 1, 1) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N มีความแม่นยำสำหรับการทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ YO01 มากที่สุด (ตาราง 15) เนื่องจากมีค่า MAE, RMSE และ MAPE น้อยที่สุด รองลงมา ได้แก่ SARIMAX(0, 1, 1)(0, 0, 1)$_4$ (Exog = DO, BOD, TCB, FCB, NH_3-N), ARIMA(0, 1, 3) และ SARIMA(3, 0, 2)(3, 2, 0)$_4$ ตามลำดับ

สำหรับค่าพารามิเตอร์ของแบบจำลอง ARIMAX(0, 1, 1) ที่ใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N ซึ่งเป็นแบบจำลองที่ทำนายค่า WQI ของข้อมูลแม่น้ำยม YO01 ได้ความคลาดเคลื่อนน้อยที่สุด (ตาราง 18) พบว่า BOD (ค่าสัมประสิทธิ์เท่ากับ -0.5227) มีค่า p-value เท่ากับ 0.000 ซึ่งน้อยกว่า 0.05 ทำให้ปฏิเสธสมมติฐานหลัก (H_0) กล่าวคือ ค่าพารามิเตอร์มีค่าแตกต่างจาก 0 ที่ระดับนัยสำคัญ 0.05 จึงสรุปว่า BOD มีความสำคัญต่อการทำนายค่า WQI ของแบบจำลองดังกล่าว

ตาราง 18 การประมาณค่าพารามิเตอร์ของข้อมูลแม่น้ำยม YO01 แบบจำลอง ARIMAX (0, 1, 1) (exog = 5 water parameters)

พารามิเตอร์	**ค่าสัมประสิทธิ์**	**ค่าความคลาดเคลื่อนมาตรฐาน (Standard Error)**	Z-score	p-value
DO	0.0504	0.148	0.341	0.733
BOD	-0.5227	0.079	-6.589	0.000
TCB	-0.2273	0.445	-0.511	0.609
FCB	-0.0742	0.809	-0.092	0.927
NH_3-N	-0.0739	0.177	-0.418	0.676
MA(1)	-0.9999	148.905	-0.007	0.995

4. การทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของแม่น้ำน่าน สถานี NA02

สถานีตรวจวัดคุณภาพน้ำ NA02 ของแม่น้ำน่าน มีข้อมูลทั้งสิ้น 54 ตัวอย่าง ประกอบด้วย ข้อมูลชุดฝึกฝน 38 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2552 - 2560 และข้อมูลชุดทดสอบ 16 ตัวอย่าง ซึ่งเป็นข้อมูลตั้งแต่ปี พ.ศ. 2561 - 2564 ผลลัพธ์ที่ได้จากการศึกษาเป็นดังตาราง 13

จากการเลือกรูปแบบพารามิเตอร์ (p, d, q) ของแบบจำลอง ARIMA และ ARIMAX และ (p, d, q)(P, D, Q) ของแบบจำลอง SARIMA และ SARIMAX โดยพิจารณาจากรูปแบบที่ทำให้ค่า AIC มีค่าต่ำที่สุด ผลที่ได้เป็นดังนี้ ARIMA(2, 1, 3), ARIMAX(0, 1, 1) ที่กำหนดให้ตัวแปร Exogenous คือ BOD, ARIMAX(3, 1, 3) เมื่อตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5 (DO, BOD, TCB, FCB และ NH_3-N), SARIMA(3, 1, 0)(1, 2, 1)$_4$, SARIMAX(0, 1, 1)(0, 0, 0)$_4$ เมื่อตัวแปร Exogenous คือ BOD และ SARIMAX(2, 1, 0)(2, 0, 3)$_4$ เมื่อกำหนดให้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5

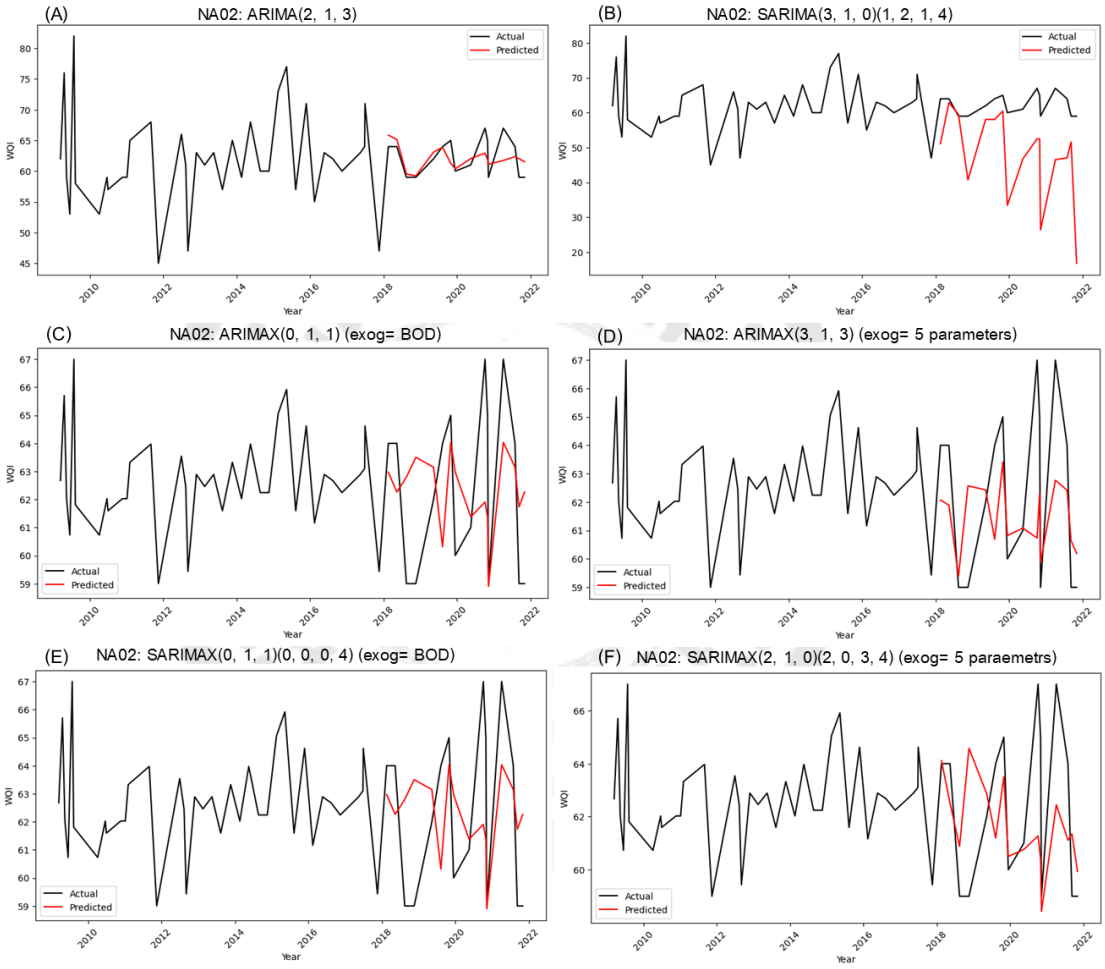
ผลจากการทำนายค่า WQI ที่ได้จาก ARIMA(2, 1, 3) เป็นไปดังภาพประกอบ 47A และตาราง 13 พบว่า ค่า MAE เท่ากับ 1.99 ค่า RMSE เท่ากับ 2.48 และค่า MAPE เท่ากับ 3.14%

ผลการศึกษาแบบจำลอง ARIMAX เพื่อทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ YO01 พบว่า ARIMAX(3, 1, 3) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำทั้ง 5 (ค่า MAE เท่ากับ 2.04 ค่า RMSE เท่ากับ 2.59 และค่า MAPE เท่ากับ 3.21%) ให้ค่าคลาดเคลื่อนน้อยกว่า ARIMAX(0, 1, 1) ที่ตัวแปร Exogenous เป็น BOD (ค่า MAE เท่ากับ 2.43 ค่า RMSE เท่ากับ 2.85 และค่า MAPE เท่ากับ 3.89%) (ตาราง 13) และจากภาพประกอบ 47C,D พบว่าเมื่อใช้ตัวแปร Exogenous เป็นทั้ง 5 พารามิเตอร์น้ำ แบบจำลองสามารถทำนายค่า WQI ได้ใกล้เคียงกับข้อมูลจริงมากกว่ากำหนดให้ตัวแปร Exogenous เป็น BOD

จากตาราง 13 และภาพประกอบ 47B พบว่า ค่าที่ได้จากการทำนายของชุดทดสอบของแบบจำลอง SARIMA(3, 1, 0)(1, 2, 1)$_4$ มีความคลาดเคลื่อนโดยค่าที่ได้จากการทำนายมีค่าน้อยกว่าข้อมูลมูลจริง ทำให้ค่า MAE, RMSE และ MAPE ของแบบจำลอง เท่ากับ 14.66, 18.46 และ 23.82% ตามลำดับ

สำหรับประสิทธิภาพการทำนายค่า WQI ของแบบจำลอง SARIMAX พบว่า SARIMAX(2, 1, 0)(2, 0, 3)$_4$ ที่ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N (ค่า MAE เท่ากับ 2.29 ค่า RMSE เท่ากับ 2.93 และค่า MAPE เท่ากับ 3.62%) มีความผิดพลาดที่

เกิดจากการทำนายค่า WQI ของข้อมูลชุดทดสอบน้อยกว่า SARIMAX(0, 1, 1)$(0, 0, 0)_4$ ที่ตัวแปร Exogenous เป็น BOD (ค่า MAE เท่ากับ 2.43 ค่า RMSE เท่ากับ 2.85 และค่า MAPE เท่ากับ 3.89%) และเมื่อพิจารณาพารามิเตอร์ $(P, D, Q)_S$ แทนส่วน Seasonality ของ SARIMAX(0, 1, 1)$(0, 0, 0)_4$ ตัวแปร Exogenous เป็น BOD พบว่า $(P, D, Q)_S$ เท่ากับ $(0, 0, 0)_4$ แสดงว่า Seasonality หรือฤดูกาลไม่มีผลต่อการทำนายค่า WQI ของแบบจำลอง



ภาพประกอบ 47 การเปรียบเทียบระหว่างค่า WQI ที่ได้จากทำนายกับข้อมูลจริงของสถานีตรวจวัดคุณภาพน้ำ NA02 ที่ได้จากแบบจำลอง (A) ARIMA(2, 1, 3) (B) SARIMA(3, 1, 0)$(1, 2, 1)_4$ (C) ARIMAX(0, 1, 1) ตัวแปร Exogenous เป็น BOD (D) ARIMAX(3, 1, 3) ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N (E) SARIMAX(0, 1, 1)$(0, 0, 0)_4$ ตัวแปร Exogenous เป็น BOD และ (F) SARIMAX(2, 1, 0)$(2, 0, 3)_4$ ตัวแปร Exogenous เป็น DO, BOD, TCB, FCB และ NH_3-N

เมื่อพิจารณาความเป็นอิสระต่อกันของความคลาดเคลื่อน (Residual) ด้วยการทดสอบ Ljung-Box Q พบว่า ARIMA(2, 1, 3) (Ljung-Box = 0.11 และ p-value = 0.74), ARIMAX(0, 1, 1) (Exog = BOD) (Ljung-Box = 0.10 และ p-value = 0.75), ARIMAX(3, 1, 3) (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 0.27 และ p-value = 0.61), SARIMA(3, 1, 0)$(1, 2, 1)_4$ (Ljung-Box = 0.18 และ p-value = 0.68), SARIMAX(0, 1, 1)$(0, 0, 0)_4$ (Exog = BOD) (Ljung-Box = 0.10 และ p-value = 0.75) และ SARIMAX(2, 1, 0)$(2, 0, 3)_4$ (Exog = DO, BOD, TCB, FCB, NH_3-N) (Ljung-Box = 0.36 และ p-value = 0.55) มีค่า p-value มากกว่า 0.05 (ที่ระดับนัยสำคัญ 0.05) (ตาราง 14) กล่าวคือ ยอมรับสมมติฐานหลัก (H_0) ที่ว่า ความคลาดเคลื่อนเป็นอิสระต่อกัน ดังนั้น แบบจำลองจึงมีความเหมาะสมกับการนำไปใช้ทำนายค่า WQI

จากผลการศึกษาดังกล่าว แสดงให้เห็นว่า ARIMA(2, 1, 3) มีความแม่นยำสำหรับการทำนายค่า WQI ของข้อมูลสถานีตรวจวัดคุณภาพน้ำ NA02 มากที่สุด (ตาราง 15) เนื่องจากมีค่า MAE และ MAPE น้อยที่สุด รองลงมา ได้แก่ ARIMAX(3, 1, 3) เมื่อใช้ตัวแปร Exogenous เป็นพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N แบบจำลอง SARIMAX(2, 1, 0)(2, 0, 3)$_4$ (Exog = DO, BOD, TCB, FCB, NH_3-N) และ SARIMA(3, 1, 0)$(1, 2, 1)_4$ ตามลำดับ

สำหรับค่าพารามิเตอร์ของแบบจำลอง ARIMA(2, 1, 3) ซึ่งเป็นแบบจำลองที่ทำนายค่า WQI ของข้อมูลแม่น้ำน่าน NA02 ได้ความคลาดเคลื่อนน้อยที่สุด (ตาราง 19) พบว่า AR(2) (ค่าสัมประสิทธิ์ เท่ากับ -0.6986) มีค่า p-value เท่ากับ 0.000 ซึ่งน้อยกว่า 0.05 ทำให้ปฏิเสธสมมติฐานหลัก (H_0) กล่าวคือ ค่าพารามิเตอร์มีค่าแตกต่างจาก 0 ที่ระดับนัยสำคัญ 0.05

ตาราง 19 การประมาณค่าพารามิเตอร์ของข้อมูลแม่น้ำน่าน (NA02) แบบจำลอง ARIMA(2, 1, 3)

พารามิเตอร์	**ค่าสัมประสิทธิ์**	**ค่าความคลาดเคลื่อนมาตรฐาน** (Standard Error)	Z-score	p-value
AR(1)	0.1449	0.214	0.679	0.497
AR(2)	-0.6986	0.200	-3.496	0.000
MA(1)	-1.5389	3.874	-0.397	0.691
MA(2)	1.5363	7.552	0.203	0.839
MA(3)	-0.9953	6.730	-0.148	0.882

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้ศึกษาเทคนิคการเรียนรู้ของเครื่องสำหรับจำแนกระดับคุณภาพแม่น้ำและทำนายดัชนีชี้วัดคุณภาพแม่น้ำของประเทศไทย โดยใช้ข้อมูลจากกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2564 สำหรับเป็นข้อมูลชุดฝึกฝน (Training data) และข้อมูลชุดทดสอบ (Test data) จากนั้นประเมินประสิทธิภาพของแบบจำลองที่ใช้จำแนกระดับคุณภาพของแม่น้ำและแบบจำลองที่ใช้ทำนายดัชนีชี้วัดคุณภาพน้ำ โดยหลังจากได้ผลดำเนินงานแล้ว สามารถสรุปผลการดำเนินงาน ซึ่งแบ่งหัวข้อการสรุปผล ดังต่อไปนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

สรุปผลการวิจัย

การศึกษาเทคนิคการเรียนรู้ของเครื่องสำหรับจำแนกระดับคุณภาพแม่น้ำและทำนายดัชนีชี้วัดคุณภาพแม่น้ำของประเทศไทย ใช้ข้อมูลคุณภาพแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน จากกองจัดการคุณภาพน้ำ กรมควบคุมมลพิษ ระหว่างปี พ.ศ. 2552 ถึง 2564 จำนวนทั้งสิ้น 2,736 ตัวอย่าง ซึ่งข้อมูลที่ได้หลังจากผ่านกระบวนการทำความสะอาดข้อมูลแล้วมีจำนวน 2,651 ตัวอย่าง ประกอบด้วยระดับคุณภาพน้ำแบ่งเป็น 5 เกณฑ์ ได้แก่ คุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก ดี พอใช้ เสื่อมโทรม และเสื่อมโทรมมาก

การจำแนกระดับคุณภาพแม่น้ำด้วยเทคนิคการเรียนรู้ของเครื่องแบ่งชุดข้อมูลสำหรับใช้จำแนกระดับคุณภาพแม่น้ำเป็น 2 ส่วน ได้แก่ ข้อมูลชุดฝึกฝน (Training data) และข้อมูลชุดทดสอบ (Test data) ในอัตราส่วน 80 ต่อ 20 ก่อนเข้าแบบจำลอง โดยงานวิจัยนี้ใช้ 12 แบบจำลอง ประกอบด้วย 4 อัลกอริทึม ได้แก่ Random Forest, XGBoost, Logistic Regression และ SVM และเปรียบเทียบประสิทธิภาพของแบบจำลองเมื่อทำงานร่วมกับเทคนิค SMOTE และ Random Oversampling ซึ่งนำมาใช้ในงานวิจัยนี้เพื่อเพิ่มปริมาณข้อมูลในแต่ละระดับคุณภาพน้ำของชุดฝึกฝนให้มีจำนวนเท่ากัน ผลการศึกษาพบว่า แบบจำลอง XGBoost ร่วมกับ SMOTE มีประสิทธิภาพสำหรับจำแนกระดับคุณภาพน้ำได้ดีที่สุด เนื่องจากได้ค่า Accuracy (91.53%), Precision (91.78%), Recall (91.53%) และ F1 score (91.56%) มากที่สุด ลำดับรองลงมาได้แก่ Random Forest ที่ใช้กับข้อมูลที่ไม่สมดุล ค่า Accuracy เท่ากับ 91.53% Precision เท่ากับ

91.62% Recall เท่ากับ 91.53% และ F1 score เท่ากับ 91.44% และเมื่อเปรียบเทียบค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ต่างๆ ของทั้ง 2 แบบจำลองดังกล่าว พบว่า Random Forest ให้ค่า F1 score ในระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม (96.05%) และพอใช้ (90.82%) มากกว่าแบบจำลอง XGBoost ร่วมกับ SMOTE (ระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม เท่ากับ 94.49% และระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ เท่ากับ 90.38%) ในทางกลับกัน XGBoost ร่วมกับ SMOTE สามารถจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดี (90.15%) และดีมาก (84.85%) ได้ดีกว่า Random Forest (ระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดี เท่ากับ 87.79% และระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก เท่ากับ 78.57%) และสำหรับระดับคุณภาพที่อยู่ในเกณฑ์เสื่อมโทรมมาก ทั้ง 2 แบบจำลองให้ค่า F1 score เท่ากับ 100% เท่ากัน

เมื่อเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้อัลกอริทึม XGBoost พบว่า การแก้ไขปัญหาข้อมูลที่ไม่สมดุลของ XGBoost ด้วยการเพิ่มปริมาณตัวอย่างในชุดฝึกฝน มีประสิทธิภาพสำหรับจำแนกระดับคุณภาพน้ำได้ดีกว่าการใช้ข้อมูลดั้งเดิมที่ไม่ได้แก้ไขปัญหาข้อมูลไม่สมดุล ทั้งนี้ XGBoost ร่วมกับ SMOTE สามารถจำแนกระดับคุณภาพน้ำที่อยู่ในแต่ละเกณฑ์ได้ดีกว่า XGBoost และ XGBoost ร่วมกับ Random Oversampling โดยเฉพาะระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมาก XGBoost ร่วมกับ SMOTE สามารถจำแนกได้ถูกต้องทั้งหมด ทำให้ค่า F1 score เท่ากับ 100% ในขณะที่ค่า F1 score ของ XGBoost และ XGBoost ร่วมกับ Random Oversampling เท่ากับ 66.67%

สำหรับแบบจำลองที่ใช้อัลกอริทึม Random Forest, Logistic Regression และ SVM นั้น การแก้ไขปัญหาข้อมูลไม่สมดุลด้วยการใช้ SMOTE และ Random Oversampling ไม่มีผลทำให้แบบจำลองมีประสิทธิภาพการจำแนกระดับคุณภาพน้ำได้ดีขึ้น อย่างไรก็ตาม หากเปรียบเทียบระหว่างเทคนิค SMOTE และ Random Oversampling ของทั้ง 3 อัลกอริทึมดังกล่าว พบว่าค่า Accuracy, Precision, Recall และ F1 score ของ Random Oversampling มีค่ามากกว่าการใช้เทคนิค SMOTE เล็กน้อย นอกจากนี้ เมื่อพิจารณาเฉพาะระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมาก การเพิ่มปริมาณข้อมูลในชุดฝึกฝนของอัลกอริทึมทั้ง 3 ด้วยเทคนิค SMOTE และ Random Oversampling ทำให้จำแนกระดับคุณภาพน้ำในเกณฑ์ดังกล่าวได้ถูกต้องมากขึ้น

เมื่อพิจารณาผลการศึกษาที่ได้จากการทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของประเทศไทยด้วยแบบจำลองอนุกรมเวลา ได้แก่ ARIMA, ARIMAX, SARIMA และ SARIMAX โดยใช้ข้อมูลคุณภาพน้ำของสถานีตรวจวัดคุณภาพน้ำ PI06, WA02, YO01 และ NA02 เป็นตัวแทนของแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ตามลำดับ และแบ่งข้อมูลสำหรับเข้าแบบจำลองออกเป็น

2 ชุด ได้แก่ ข้อมูลชุดฝึกฝนเป็นข้อมูลระหว่างปี พ.ศ. 2552 ถึง พ.ศ. 2560 (9 ปี) และข้อมูลชุดทดสอบเป็นข้อมูลระหว่างปี 2561 ถึง พ.ศ. 2564 (4 ปี) พบว่า ARIMAX โดยกำหนดตัวแปรภายนอก (Exogenous variable) ได้แก่ DO, BOD, TCB, FCB และ NH_3-N เป็นแบบจำลองที่สามารถทำนายค่า WQI สถานีตรวจวัด PI06, WA02 และ YO01 ของแม่น้ำปิง แม่น้ำวัง และแม่น้ำยม ตามลำดับ ได้แม่นยำที่สุด ซึ่ง ARIMAX(0, 1, 1) ของสถานี PI06 มีค่า MAE, RMSE และ MAPE เท่ากับ 4.35 5.90 และ 6.15% ตามลำดับ ARIMAX(2, 1, 2) ของสถานี WA02 มีค่า MAE, RMSE และ MAPE เท่ากับ 6.36 7.55 และ 9.35% ตามลำดับ และ ARIMAX(0, 1, 1) ของสถานี YO01 มีค่า MAE, RMSE และ MAPE เท่ากับ 5.85 6.62 และ 5.85% ตามลำดับ อย่างไรก็ตาม เมื่อพิจารณาสถานี NA02 ของแม่น้ำน่าน พบว่า ARIMA(2, 1, 3) เป็นแบบจำลองที่สามารถทำนายค่า WQI ได้ดีที่สุด (MAE เท่ากับ 1.99 RMSE เท่ากับ 2.48 และ MAPE เท่ากับ 3.14%) รองลงมาเป็น ARIMAX(3, 1, 3) (MAE เท่ากับ 2.04 RMSE เท่ากับ 2.59 และ MAPE เท่ากับ 3.21%) ในทางกลับกัน งานวิจัยนี้แสดงให้เห็นว่า SARIMA เป็นแบบจำลองที่มีความแม่นยำสำหรับใช้ทำนายค่า WQI น้อยที่สุดของทุกสถานีตรวจวัดคุณภาพน้ำที่ใช้ศึกษา

อภิปรายผลการวิจัย

งานวิจัยนี้จำแนกระดับคุณภาพแม่น้ำของประเทศไทยด้วย Random Forest, XGBoost, Logistic Regression และ SVM ร่วมกับเทคนิคการแก้ไขปัญหาข้อมูลที่ไม่สมดุล (SMOTE และ Random Oversampling) เนื่องจากปริมาณข้อมูลระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมากเท่ากับ 2.98% และระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากเท่ากับ 0.41% ของจำนวนตัวอย่างทั้งหมด ซึ่งผลการศึกษาแสดงให้เห็นว่า มีเพียงอัลกอริทึม XGBoost ที่การแก้ไขปัญหาความไม่สมดุลของข้อมูลด้วย SMOTE และ Random Oversampling ทำให้การจำแนกระดับคุณภาพน้ำมีประสิทธิภาพเพิ่มขึ้นเมื่อเทียบกับข้อมูลดั้งเดิม แต่สำหรับการแก้ไขปัญหาข้อมูลไม่สมดุลด้วยการเพิ่มปริมาณข้อมูลในชุดฝึกฝนด้วย SMOTE และ Random Oversampling ของอัลกอริทึม Random Forest, Logistic Regression และ SVM นั้น ไม่สามารถทำให้แบบจำลองมีประสิทธิภาพในการจำแนกเพิ่มขึ้น

เมื่อพิจารณาพารามิเตอร์น้ำหรือลักษณะเฉพาะ (Feature) ที่มีผลต่อการจำแนกระดับคุณภาพน้ำ พบว่า BOD เป็นพารามิเตอร์ที่มีผลต่อการจำแนกระดับคุณภาพน้ำสำหรับชุดข้อมูลนี้มากที่สุด ซึ่งให้ผลเหมือนกันทั้ง 12 แบบจำลอง และพารามิเตอร์ที่มาความสำคัญต่อการจำแนกระดับคุณภาพแม่น้ำลำดับถัดมาสำหรับแบบจำลองส่วนใหญ่ ได้แก่ FCB, TCB, NH_3-N และ DO ตามลำดับ โดยผลการศึกษาที่ได้จากงานวิจัยนี้ต่างจากงานวิจัยของ Sillberg et al.

(2021) ซึ่งจำแนกระดับคุณภาพน้ำของแม่น้ำเจ้าพระยาด้วย SVM พบว่า NH_3-N มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำของแม่น้ำเจ้าพระยามากที่สุด รองลงมา ได้แก่ TCB, FCB, BOD, DO และความเค็ม ตามลำดับ

เมื่อวิเคราะห์ความสำคัญของพารามิเตอร์ที่มีผลต่อการจำแนกระดับคุณภาพแม่น้ำในแต่ละเกณฑ์ของแบบจำลอง XGBoost ร่วมกับเทคนิค SMOTE ซึ่งเป็นแบบจำลองที่สามารถจำแนกระดับคุณภาพน้ำได้ดีที่สุด ทำให้ทราบว่า ระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมและเสื่อมโทรมมาก ค่า BOD, NH_3-N, TCB และ FCB มีความสัมพันธ์เชิงบวกกับการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดังกล่าว กล่าวคือ BOD, NH_3-N, TCB และ FCB มีค่าสูง จะส่งผลต่อการจำแนกเป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมและเสื่อมโทรมมากเพิ่มขึ้น ในขณะที่ค่า DO มีความสัมพันธ์เชิงลบกับระดับคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมและเสื่อมโทรมมาก และสำหรับระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีและดีมาก ค่า DO มีความสัมพันธ์เชิงบวกกับระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดังกล่าว กล่าวคือ DO มีค่าสูง จะส่งผลให้แบบจำลองจำแนกเป็นระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีและดีมากเพิ่มขึ้น แต่ BOD, NH_3-N, TCB และ FCB มีความสัมพันธ์เชิงลบกับระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีและดีมาก ซึ่งผลดังกล่าวสอดคล้องกับค่า Pearson Correlation Coefficient แสดงความสัมพันธ์ระหว่างค่า WQI กับพารามิเตอร์น้ำต่างๆ ที่ระบุว่าค่า WQI มีความสัมพันธ์เชิงบวกกับค่า DO แต่ค่า WQI มีความสัมพันธ์เชิงลบกับค่า BOD, TCB, FCB และ NH_3-N ดังนั้น จากความสัมพันธ์ระหว่างพารามิเตอร์น้ำกับการจำแนกระดับคุณภาพน้ำของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเหมือนระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากนั้น จึงเป็นสาเหตุหนึ่งที่ทำให้แบบจำลองจำแนกระดับคุณภาพน้ำทั้ง 2 ผิดพลาดได้ ซึ่งเห็นได้จากการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากผิดพลาดเป็นดีจำนวน 2 ตัวอย่าง และจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีเป็นดีมากจำนวน 3 ตัวอย่าง ทำให้ ค่า F1 score ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์ดีมากเท่ากับ 84.85% นอกจากนี้ เมื่อพิจารณาความสำคัญของพารามิเตอร์ต่อการจำแนกระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ ทำให้ทราบว่าพารามิเตอร์น้ำได้แก่ BOD, NH_3-N และ DO มีการกระจายตัวของข้อมูลที่ไม่สามารถบอกแนวโน้มความสัมพันธ์ระหว่างพารามิเตอร์น้ำดังกล่าวกับการจำแนกระดับคุณภาพน้ำในเกณฑ์นี้ได้ว่ามีความสัมพันธ์เชิงบวกหรือลบ และจากการวิเคราะห์ค่า SHAP สัมบูรณ์เฉลี่ยของพารามิเตอร์น้ำ (ใช้บ่งบอกความสำคัญของพารามิเตอร์น้ำที่ส่งผลการจำแนกระดับคุณภาพน้ำ) ของระดับคุณภาพน้ำที่อยู่ในเกณฑ์พอใช้ ดี เสื่อมโทรม และเสื่อมโทรมมาก มีแนวโน้มเหมือนกัน กล่าวคือ BOD มากกว่า

FCB, TCB, NH_3-N และ DO ตามลำดับ จึงอาจเป็นอีกสาเหตุหนึ่งที่ทำให้แบบจำลองจำแนกระดับคุณภาพน้ำคลาดเคลื่อนได้

นอกจากนี้ งานวิจัยนี้ได้ทำการศึกษาการทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ของประเทศไทยด้วยแบบจำลองอนุกรมเวลา ได้แก่ ARIMA, ARIMAX, SARIMA และ SARIMAX โดยใช้ข้อมูลคุณภาพน้ำของสถานีตรวจวัดคุณภาพน้ำ PI06, WA02, YO01 และ NA02 เป็นตัวแทนของแม่น้ำปิง แม่น้ำวัง แม่น้ำยม และแม่น้ำน่าน ตามลำดับ จากผลการศึกษาพบว่า ARIMAX ที่ใช้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์ (DO, BOD, TCB, FCB และ NH_3-N) เป็นแบบจำลองที่สามารถทำนายค่า WQI ได้แม่นยำที่สุด เมื่อใช้กับข้อมูลของสถานีตรวจวัด PI06 (แม่น้ำปิง) WA02 (แม่น้ำวัง) และ YO01 (แม่น้ำยม) แสดงให้เห็นว่า ตัวแปรภายนอก (Exogenous variable) ส่งผลให้การทำนายค่า WQI แม่นยำเพิ่มขึ้น และเมื่อพิจารณาค่าที่ได้จากการทดสอบทางสถิติของค่าพารามิเตอร์ของแบบจำลอง ARIMAX ที่ใช้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์ (DO, BOD, TCB, FCB และ NH_3-N) ของสถานี PI06, WA02 และ YO01 พบว่า พารามิเตอร์ BOD, TCB, FCB และ NH_3-N ส่งผลต่อการทำนายค่า WQI ของสถานี PI06 พารามิเตอร์ DO, BOD, TCB, FCB และ NH_3-N ส่งผลต่อการทำนายค่า WQI ของสถานี WA02 และพารามิเตอร์ BOD ส่งผลต่อการทำนายค่า WQI ของสถานี YO01 แสดงให้เห็นว่า BOD มีผลต่อการทำนายค่า WQI ทั้ง 3 สถานี ซึ่งสอดคล้องกับผลการศึกษาการจำแนกระดับคุณภาพน้ำที่ BOD เป็นพารามิเตอร์ที่มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำมากที่สุด

เมื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง ARIMAX ของสถานี PI06, WA02, YO01 และ NA02 ระหว่างใช้ตัวแปรภายนอกเพียงค่า BOD ซึ่งเป็นพารามิเตอร์ที่มีผลต่อการจำแนกระดับคุณภาพน้ำมากที่สุด และใช้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N พบว่า การใช้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์ให้ค่าความคลาดเคลื่อนจากการทำนายน้อยกว่าการใช้ตัวแปรภายนอกเพียง BOD อาจกล่าวได้ว่า ถึงแม้ BOD จะส่งผลต่อการทำนายค่า WQI แต่การใช้ทั้ง 5 พารามิเตอร์ เป็นตัวแปรภายนอกยังให้ผลการทำนายแม่นยำกว่าการใช้เพียง BOD

การเปรียบเทียบประสิทธิภาพของแบบจำลอง SARIMAX ระหว่างใช้ตัวแปรภายนอกเป็น BOD กับใช้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์ (DO, BOD, TCB, FCB และ NH_3-N) พบว่า สถานี WA02, YO01 และ NA02 ที่ใช้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์มีค่าความคลาดเคลื่อนจากการทำนาย (ค่า MAE และ MAPE) น้อยกว่าการใช้ตัวแปรภายนอก

เพียงค่า BOD แต่สถานี PI06 ที่ใช้ตัวแปรภายนอกเป็น BOD ให้ค่าความคลาดเคลื่อนจากการทำนายน้อยกว่าการใช้ตัวแปรภายนอกเป็นพารามิเตอร์น้ำ 5 พารามิเตอร์

อีกทั้ง จากการศึกษาการทำนายดัชนีชี้วัดคุณภาพน้ำ (WQI) ยังแสดงให้เห็นว่า SARIMA เป็นแบบจำลองที่มีประสิทธิภาพสำหรับทำนายค่า WQI น้อยที่สุด ทั้ง 4 สถานีตรวจวัดคุณภาพน้ำ ดังนั้น แบบจำลองดังกล่าวจึงไม่เหมาะสมกับการนำมาใช้ทำนายค่า WQI ของสถานี PI06, WA02, YO01 และ NA02 นอกจากนี้ ผลการศึกษาแสดงให้เห็นว่า ARIMAX เป็นแบบจำลองที่มีประสิทธิภาพการทำนายค่า WQI มากกว่า SARIMAX สามารถสรุปได้ว่า Seasonality หรือฤดูกาล อาจไม่ส่งผลต่อการเพิ่มประสิทธิภาพการทำนายค่า WQI ที่ใช้กับข้อมูลของสถานีตรวจวัดคุณภาพน้ำ PI06, WA02, YO01 และ NA02 ในงานวิจัยนี้

ข้อเสนอแนะ

1. นำแบบจำลองที่ใช้ศึกษาการจำแนกระดับคุณภาพน้ำในงานวิจัยนี้ ไปใช้ร่วมกับชุดข้อมูลคุณภาพน้ำของประเทศอื่นที่มีพารามิเตอร์น้ำ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N เป็นลักษณะเฉพาะ เพื่อใช้จำแนกระดับคุณภาพน้ำและทดสอบประสิทธิภาพของแบบจำลอง

2. งานวิจัยในอนาคต ควรนำพารามิเตอร์น้ำอื่นๆ เช่น ค่า pH ค่าของแข็งแขวนลอย (Suspended Solids) ค่าฟอสฟอรัสรวม (Total Phosphorus) เป็นต้น ร่วมกับค่า DO, BOD, TCB, FCB และ NH_3-N เป็นลักษณะเฉพาะนำเข้าแบบจำลอง และใช้เทคนิคการเลือกลักษณะเฉพาะ (Feature selection) เพื่อให้ทราบว่าลักษณะเฉพาะใดมีความสำคัญต่อการจำแนกระดับคุณภาพของประเทศไทยมากที่สุด และสอดคล้องกับพารามิเตอร์น้ำ 5 พารามิเตอร์ ได้แก่ DO, BOD, TCB, FCB และ NH_3-N ที่ใช้คำนวณค่า WQI ของประเทศไทยหรือไม่

3. เนื่องจากจำนวนข้อมูลในแต่ละสถานีตรวจวัดคุณภาพน้ำมีจำนวนน้อย จึงทำให้ผลการทำนายดัชนีชี้วัดคุณภาพน้ำอาจยังไม่ได้ประสิทธิภาพเท่าที่ควร หากในอนาคตมีจำนวนข้อมูลมากขึ้น อาจทำให้แบบจำลองสามารถทำนายข้อมูลได้แม่นยำมากขึ้น

4. งานวิจัยในอนาคต ควรศึกษาเพิ่มเติมเกี่ยวกับการทำนายค่า WQI ด้วยแบบจำลองอนุกรมเวลา โดยเปรียบเทียบประสิทธิภาพของผลการทำนายค่า WQI ล่วงหน้าที่ช่วงเวลาต่างๆ เพื่อให้ทราบว่าแบบจำลองสามารถทำนายค่า WQI ล่วงหน้าได้แม่นยำที่สุด ณ ช่วงเวลาใด

5. ในอนาคตสามารถนำชุดข้อมูลที่ใช้ในงานวิจัยนี้ ไปใช้ทำนายดัชนีชี้วัดคุณภาพน้ำด้วยแบบจำลองประเภทอื่นๆ เช่น Support Vector Regression (SVR), Random Forest, XGBoost และ Neural network เป็นต้น

6. ข้อมูลที่ได้จากการศึกษาในงานวิจัยนี้ สามารถนำไปประยุกต์ใช้และเป็นข้อมูลประกอบการพิจารณาวางแผนการบริหารจัดการคุณภาพน้ำเพื่อเป็นประโยชน์ต่อหน่วยงานที่เกี่ยวข้อง เช่น กรมควบคุมมลพิษ สำนักงานทรัพยากรน้ำแห่งชาติ หรือนักวิจัยที่เกี่ยวข้อง เป็นต้น ตัวอย่างประโยชน์ที่ได้จากการศึกษานี้ ทำให้ทราบว่าข้อมูลชุดนี้มีคุณภาพน้ำส่วนใหญ่อยู่ในเกณฑ์พอใช้คิดเป็น 37.68% รองลงมา ได้แก่ คุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรม 33.16% ดี 25.76% ดีมาก 2.98% และเสื่อมโทรม 0.41% ตามลำดับ ซึ่งจากข้อมูลดังกล่าวสามารถใช้เป็นข้อมูลประกอบการวางแผนบริหารจัดการน้ำ พร้อมทั้งหามาตรการป้องกันแก้ไข เพื่อทำให้คุณภาพน้ำมีแนวโน้มที่ดีขึ้นและคุณภาพน้ำที่อยู่ในเกณฑ์เสื่อมโทรมมีแนวโน้มลดลง นอกจากนี้ ผลการศึกษายังแสดงให้เห็นว่า BOD เป็นพารามิเตอร์น้ำที่มีความสำคัญต่อการจำแนกระดับคุณภาพน้ำมากที่สุด และจากข้อมูลในรายงานการดำเนินงานของกองจัดการคุณภาพน้ำ พ.ศ. 2565 พบว่า ค่า BOD ของแหล่งน้ำหลายแห่งไม่เป็นไปตามมาตรฐานคุณภาพน้ำผิวดิน (กรมควบคุมมลพิษ, 2566) ด้วยเหตุนี้ ควรนำข้อมูลความสำคัญของพารามิเตอร์ที่ได้จากการสร้างแบบจำลองมาใช้ประกอบการพิจารณาสำหรับหาแนวทางป้องกันไม่ให้ค่า BOD เกินค่ามาตรฐาน ทำให้คุณภาพน้ำอยู่ในเกณฑ์ที่ดีขึ้น นำมาซึ่งปัญหาแหล่งน้ำเน่าเสียมีแนวโน้มลดลง อีกทั้ง งานวิจัยนี้ศึกษาการจำแนกระดับคุณภาพแม่น้ำและการทำนายดัชนีชี้วัดคุณภาพน้ำโดยใช้แบบจำลอง ซึ่งเป็นวิธีหนึ่งที่ช่วยคาดการณ์คุณภาพน้ำล่วงหน้าและใช้เป็นข้อมูลประกอบการจัดทำแผนบริหารจัดการคุณภาพน้ำในอนาคตได้

บรรณานุกรม

กรมควบคุมมลพิษ. (2554). *ค่าคะแนนรวมของคุณภาพน้ำ 5 พารามิเตอร์.* สืบค้นจาก http://iwis.pcd.go.th/officer/document/download/2/2.pdf

กรมควบคุมมลพิษ. (2561). *คู่มือการปฏิบัติงานการดำเนินการติดตามตรวจสอบคุณภาพน้ำในแหล่งน้ำผิวดิน.* สืบค้นจาก https://www.pcd.go.th/wp-content/uploads/2020/04/pcdnew-2020-04-21_09-32-00_948434.pdf

กรมควบคุมมลพิษ. (2565). *ดัชนีคุณภาพน้ำแหล่งน้ำผิวดิน (Water Quality Index: WQI).* สืบค้นจาก https://www.pcd.go.th/wp-content/uploads/2022/08/pcdnew-2022-08-23_03-47-16_304672.pdf

กรมควบคุมมลพิษ. (2566). *รายงานการดำเนินงาน กองจัดการคุณภาพน้ำ พ.ศ. 2565.* สืบค้นจาก https://www.pcd.go.th/wp-content/uploads/2023/04/pcdnew-2023-04-04_08-17-30_089786.pdf

กรมชลประทาน. (2561). *รายงานความคิดเห็นของโครงการฯ เกี่ยวกับวิธีการตรวจวัดคุณภาพน้ำ.* สืบค้นจาก http://qwater.rid.go.th/report/file61/exam61/PDF/EQUIPPROB.pdf

กิติ์สุชาต พสุภา. (2564). ระบบอัจฉริยะเพื่อทำนายข้อมูลอนุกรมเวลา. ใน *ระบบอัจฉริยะขั้นสูง: ทฤษฎี อัลกอริทึม และการประยุกต์ใช้* (น. 205-224). กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

ภูมิฐาน รังคกูลนุวัฒน์. (2562). *การวิเคราะห์อนุกรมเวลาสำหรับเศรษฐศาสตร์และธุรกิจ.* สืบค้นจาก https://economics.utcc.ac.th/wp-content/uploads/Time-Series-for-Econ-and-Bus_Poomthan.pdf

รัสรินทร์ เมธาเฉลิมพัฒน์. (2565). *การประยุกต์ใช้* Machine Learning *กับงานในภาคอุตสาหกรรม (ตอนที่ 1).* สืบค้นจาก https://www.nectec.or.th/wp-content/uploads/2022/08/CPS-ML-manufacturing.pdf

สำนักงานสถิติแห่งชาติ. (2563). *ดัชนีชี้วัดการจัดการน้ำเพื่อการบริการจัดการทรัพยากรน้ำอย่างยั่งยืน.* สืบค้นจาก http://www.nso.go.th/sites/2014/DocLib14/Press_Release/2563/P14-07-63.pdf

อดิศร์ อิศรางกูร ณ อยุธยา. (2560). *แม่น้ำเจ้าพระยาขาดคนดูแล.* สืบค้นจาก https://tdri.or.th/2017/07/chao-phraya-river-management-bodies/

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. *Water,*

11(11), 2210.

Amazon Web Services. (2022). *What is Logistic Regression?* Retrieved from https://aws.amazon.com/what-is/logistic-regression/

Archana, C., Savita, K., & Raj, K. (2016). An improved random forest classifier for multi-class classification. *Information Processing in Agriculture, 3*(4), 215-222.

Arunthavanathan, R., Khan, F., Ahmed, S., & Imtiaz, S. (2022). Autonomous Fault Diagnosis and Root Cause Analysis for the Processing System Using One-Class SVM and NN Permutation Algorithm. *Industrial & Engineering Chemistry Research, 61*(3), 1408-1422.

Banda, T. D., & Kumarasamy, M. V. (2020). Development of Water Quality Indices (WQIs): A Review. *Polish Journal of Environmental Studies, 29*(3), 2011-2021.

Bonaccorso, G. (2017). Decision Trees and Ensemble Learning. In *Machine Learning Algorithms* (pp. 154-180). Birmingham: Packt Publishing.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Brown, R. M., McClelland, N. I., Deininger, R. A., & Tozer, R. G. (1970). A Water Quality Index - Do We Dare? *Water and Sewage Works, 117*(10), 339- 343.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA.

Dastorani, M., Mirzavand, M., Dastorani, M. T., & Khosravi, H. (2020). Simulation and prediction of surface water quality using stochastic models. *Sustainable Water Resources Management, 6*(4), 74.

Dickey, D. A., & Fuller, W. A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometric Society, 49*(4), 1057-1072.

Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials, 16*, e01059.

Fashae, O. A., Olusola, A. O., Ndubuisi, I., & Udomboso, C. G. (2019). Comparing ANN

and ARIMA model in predicting the discharge of River Opeki from 2010 to 2020. *River Research and Applications, 35*(2), 169-177.

Geron, A. (2019). Ensemble Learning and Random Forests. In *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed., pp. 191-214). Sebastopol, CA: O'Reilly Media.

Horton, R. K. (1965). An Index Number System for Rating Water Quality. *Journal of the Water Pollution Control Federation, 37*(3), 300-306.

Hossin, M., Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 1-11.

House, M. A. (1989). A Water Quality Index for River Management. *Water and Environment Journal, 3*(4), 336-344.

Khalusova, M. (2022). *Machine Learning Model Evaluation Metrics Part 2*: *Multi-Class Classification.* Retrieved from https://www.mariakhalusova.com/posts/2019-04-17-ml-model-evaluation-metrics-p2/

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review, 1*(3).

Kouadri, S., Elbeltagi, A., Islam, A. R. M. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Applied Water Science, 11*(190), 1-20.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics, 54*(1-3), 159-178.

Lazzeri, F. (2020). Overview of Time Series Forecasting. In *Machine Learning for Time Series Forecasting with Python* (pp. 1-26). Hoboken, NJ: John Wiley & Sons.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research, 18*(1), 559-563.

Lundberg, S. M., Erion, G. G., & Lee, S. (2019). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint arXiv:1802.03888v3*.

Lundberg, S. M., & Lee, S. (2017). *A Unified Approach to Interpreting Model Predictions.* Paper presented at the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA.

Malek, N. H. A., Yaacob, W. F. W., Nasir, S. A. M., & Shaadan, N. (2022). Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. *Water, 14*(7), 1067.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics, 10*(213).

Northep, K., Srijiranon, K., & Eiamkanitchat, N. (2020). *Water Quality Classification Using Data Mining Techniques: A Case Study on Wang River in Thailand.* Paper presented at the 2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan.

Poonam, T., Tanushree, B., & Sukalyan, C. (2013). Water Quality Indices-Important Tools for Water Quality Assessment: A Review. *International Journal of Advances in Chemistry, 1*(1), 15-28.

Prakirake, C., Chaiprasert, P., & Tripetchkul, S. (2009). Development of specific water quality index for water supply in Thailand. *Songklanakarin Journal of Science & Technology, 31*(1), 91-104.

Saini, A. (2021). *Support Vector Machine (SVM): A Complete guide for beginners.* Retrieved from https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768.

Scikit-learn developers. (2022a). *Logistic regression.* Retrieved from https://scikit-

learn.org/stable/modules/linear_model.html?highlight=logistic+regression#logistic-regression

Scikit-learn developers. (2022b). *Permutation feature importance.* Retrieved from https://scikit-learn.org/stable/modules/permutation_importance.html#

Scikit-learn developers. (2022c). *Metrics and scoring: quantifying the quality of predictions.* Retreived from https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics

Scikit-learn developers. (2022d). *Preprocessing data.* Retrieved from https://scikit-learn.org/stable/modules/preprocessing.html#

Shamsuddin, I. I. S., Othman, Z., & Sani, N. S. (2022). Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. *Water, 14*(19), 2939.

Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention, 129*, 170-179.

Sillberg, C. V., Kullavanijaya, P., & Chavalparit, O. (2021). Water Quality Classification by Integration of Attribute-Realization and Support Vector Machine for the Chao Phraya River. *Journal of Ecological Engineering, 22*(9), 70-86.

Simachaya, W. (2000). *Water quality management in Thailand.* Paper presented at the Workshop on “Environmentally sound technology on water quality management” UNEP, Mekong River Commission, Bangkok, Thailand.

Singkran, N., Yenpiem, A., & Sasitorn, P. (2010). Determining Water Conditions in the Northeastern Rivers of Thailand Using Time Series and Water Quality Index Models. *Journal of Sustainable Energy & Environment, 1*, 47-58.

Su, Y., & Ye, Y. (2020). *Daily Passenger Volume Prediction in the Bus Transportation System using ARIMAX Model with Big Data*. Paper presented at the 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chongqing, China.

Suphawan, K., & Chaisee, K. (2021). Gaussian process regression for predicting water

quality index: A case study on ping river basin, thailand. *AIMS Environmental Science, 8*(3), 268-282.

Taweelarp, S., Khebchareon, M., & Saenton, S. (2021). Evaluation of Groundwater Potential and Safe Yield of Heterogeneous Unconsolidated Aquifers in Chiang Mai Basin, Northern Thailand. *Water, 13*(4), 558.

Uddin, M. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators, 122*, 107218.

Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2022). A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. *Water Research, 219*, 118532.

Uyun, S., & Sulistyowati, E. (2020). Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes. *International Journal of Electrical and Computer Engineering, 10*(4), 4331-4339.

Vagropoulos, S. I., Chouliaras, G. I., Kardakos, E. G., Simoglou, C. K., & Bakirtzis, A. G. (2016). *Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting.* Paper presented at the 2016 IEEE International Energy Conference (ENERGYCON), Leuven, Belgium.

Walsh, P., & Wheeler, W. (2013). Water Quality Index Aggregation and Cost Benefit Analysis. *Journal of Benefit-Cost Analysis, 4*(4), 81-106

ภาคผนวก

1. ค่าพารามิเตอร์และค่าทางสถิติของแบบจำลองอนุกรมเวลา แม่น้ำปิง สถานี PI06

PI06: ARIMA(1, 1, 1)

Dep. Variable:		WQI	No. Observations:			36
Model:		SARIMAX(1, 1, 1)	Log Likelihood			-122.878
Date:		Fri, 27 Oct 2023	AIC			251.756
Time:		14:41:38	BIC			256.422
Sample:		0	HQIC			253.366
		- 36				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3487	0.210	-1.657	0.098	-0.761	0.064
ma.L1	-0.8253	0.103	-8.025	0.000	-1.027	-0.624
sigma2	62.3599	20.565	3.032	0.002	22.053	102.667
Ljung-Box (L1) (Q):		0.01	Jarque-Bera (JB):			2.06
Prob(Q):		0.93	Prob(JB):			0.36
Heteroskedasticity (H):		1.29	Skew:			0.47
Prob(H) (two-sided):		0.67	Kurtosis:			2.28

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

PI06: ARIMAX(1, 1, 1) (exog = BOD)

Dep. Variable:		WQI	No. Observations:			36
Model:		SARIMAX(1, 1, 1)	Log Likelihood			-4.047
Date:		Fri, 29 Sep 2023	AIC			16.094
Time:		11:28:59	BIC			22.315
Sample:		0	HQIC			18.241
		- 36				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.3299	0.151	-2.179	0.029	-0.627	-0.033
ar.L1	-0.3916	0.242	-1.620	0.105	-0.865	0.082
ma.L1	-0.9093	0.143	-6.377	0.000	-1.189	-0.630
sigma2	0.0686	0.025	2.759	0.006	0.020	0.117
Ljung-Box (L1) (Q):		0.04	Jarque-Bera (JB):			1.92
Prob(Q):		0.85	Prob(JB):			0.38
Heteroskedasticity (H):		1.26	Skew:			0.33
Prob(H) (two-sided):		0.69	Kurtosis:			2.06

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

PI06: ARIMAX(0, 1, 1) (exog = 5 parameters)

Dep. Variable:		WQI	No. Observations:			36
Model:		SARIMAX(0, 1, 1)	Log Likelihood			11.368
Date:		Fri, 27 Oct 2023	AIC			-8.736
Time:		16:22:10	BIC			2.151
Sample:		0	HQIC			-4.978
		- 36				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	-0.0653	0.158	-0.413	0.680	-0.375	0.245
BOD(mg/l)	-0.3782	0.148	-2.556	0.011	-0.668	-0.088
Total Coli(MPN/100ml)	-0.7496	0.225	-3.334	0.001	-1.190	-0.309
Fecal Coli(MPN/100ml)	-0.4263	0.209	-2.036	0.042	-0.837	-0.016
NH3-N(mg/l)	-0.4046	0.122	-3.328	0.001	-0.643	-0.166
ma.L1	-0.9980	7.929	-0.126	0.900	-16.539	14.543
sigma2	0.0277	0.225	0.123	0.902	-0.413	0.468
Ljung-Box (L1) (Q):		0.19	Jarque-Bera (JB):			2.40
Prob(Q):		0.66	Prob(JB):			0.30
Heteroskedasticity (H):		0.88	Skew:			0.17
Prob(H) (two-sided):		0.83	Kurtosis:			1.76

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

PI06: SARIMA(2, 1, 1)(1, 2, 1, 4)

Dep. Variable:		WQI	No. Observations:			36
Model:		SARIMAX(2, 1, 1)x(1, 2, 1, 4)	Log Likelihood			-103.509
Date:		Fri, 29 Sep 2023	AIC			219.017
Time:		11:36:03	BIC			226.792
Sample:		0	HQIC			221.329
		- 36				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.7287	0.307	-5.628	0.000	-2.331	-1.127
ar.L2	-0.8515	0.267	-3.191	0.001	-1.374	-0.329
ma.L1	0.9982	43.180	0.023	0.982	-83.633	85.630
ar.S.L4	-0.6730	0.195	-3.457	0.001	-1.055	-0.291
ma.S.L4	-0.9965	51.190	-0.019	0.984	-101.328	99.335
sigma2	58.9273	3225.700	0.018	0.985	-6263.328	6381.182
Ljung-Box (L1) (Q):		0.01	Jarque-Bera (JB):			1.07
Prob(Q):		0.94	Prob(JB):			0.59
Heteroskedasticity (H):		1.97	Skew:			0.47
Prob(H) (two-sided):		0.33	Kurtosis:			3.29

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

PI06: SARIMAX(1, 1, 1)(0, 0, 0, 4) (exog = BOD)

Dep. Variable:		WQI	No. Observations:			36
Model:		SARIMAX(1, 1, 1)	Log Likelihood			-4.047
Date:		Wed, 18 Oct 2023	AIC			16.094
Time:		00:49:00	BIC			22.315
Sample:		0	HQIC			18.241
		- 36				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.3299	0.151	-2.179	0.029	-0.627	-0.033
ar.L1	-0.3916	0.242	-1.620	0.105	-0.865	0.082
ma.L1	-0.9093	0.143	-6.377	0.000	-1.189	-0.630
sigma2	0.0686	0.025	2.759	0.006	0.020	0.117
Ljung-Box (L1) (Q):		0.04	Jarque-Bera (JB):			1.92
Prob(Q):		0.85	Prob(JB):			0.38
Heteroskedasticity (H):		1.26	Skew:			0.33
Prob(H) (two-sided):		0.69	Kurtosis:			2.06

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

PI06: SARIMAX(3, 1, 2)(2, 0, 0, 4) (exog = 5 parameters)

Dep. Variable:		WQI	No. Observations:			36
Model:		SARIMAX(3, 1, 2)x(2, 0, [], 4)	Log Likelihood			18.418
Date:		Wed, 18 Oct 2023	AIC			-10.836
Time:		00:50:00	BIC			9.384
Sample:		0	HQIC			-3.856
		- 36				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	-0.0294	0.122	-0.240	0.810	-0.269	0.210
BOD(mg/l)	-0.5177	0.138	-3.755	0.000	-0.788	-0.247
Total Coli(MPN/100ml)	-0.9185	0.166	-5.520	0.000	-1.245	-0.592
Fecal Coli(MPN/100ml)	-0.3016	0.161	-1.870	0.062	-0.618	0.015
NH3-N(mg/l)	-0.4039	0.085	-4.755	0.000	-0.570	-0.237
ar.L1	-0.1470	0.380	-0.387	0.699	-0.891	0.597
ar.L2	-0.5900	0.195	-3.032	0.002	-0.971	-0.209
ar.L3	-0.5890	0.267	-2.208	0.027	-1.112	-0.066
ma.L1	-1.3127	0.616	-2.130	0.033	-2.521	-0.105
ma.L2	0.8852	0.689	1.285	0.199	-0.465	2.235
ar.S.L4	-0.2413	0.315	-0.767	0.443	-0.858	0.375
ar.S.L8	-0.6779	0.212	-3.192	0.001	-1.094	-0.262
sigma2	0.0140	0.008	1.863	0.062	-0.001	0.029
Ljung-Box (L1) (Q):		0.00	Jarque-Bera (JB):			10.40
Prob(Q):		0.98	Prob(JB):			0.01
Heteroskedasticity (H):		0.85	Skew:			-0.96
Prob(H) (two-sided):		0.78	Kurtosis:			4.87

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

2. ค่าพารามิเตอร์และค่าทางสถิติของแบบจำลองอนุกรมเวลา แม่น้ำวัง สถานี WA02

WA02: ARIMA(3, 2, 1)

Dep. Variable:		WQI	No. Observations:			38
Model:		SARIMAX(3, 2, 1)	Log Likelihood			-140.887
Date:		Fri, 29 Sep 2023	AIC			291.774
Time:		18:38:10	BIC			299.692
Sample:		0	HQIC			294.538
		- 38				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3428	0.312	-1.099	0.272	-0.955	0.269
ar.L2	-0.1552	0.463	-0.335	0.737	-1.062	0.752
ar.L3	0.3312	0.389	0.850	0.395	-0.432	1.094
ma.L1	-0.9998	86.354	-0.012	0.991	-170.250	168.250
sigma2	129.2369	1.11e+04	0.012	0.991	-2.17e+04	2.2e+04
Ljung-Box (L1) (Q):		0.02	Jarque-Bera (JB):			11.54
Prob(Q):		0.88	Prob(JB):			0.00
Heteroskedasticity (H):		5.51	Skew:			-0.56
Prob(H) (two-sided):		0.01	Kurtosis:			5.54

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

WA02: ARIMAX(0, 1, 1) (exog = BOD)

Dep. Variable:		WQI	No. Observations:			38
Model:		SARIMAX(0, 1, 1)	Log Likelihood			8.757
Date:		Fri, 29 Sep 2023	AIC			-11.514
Time:		08:48:24	BIC			-6.681
Sample:		0	HQIC			-9.810
		- 38				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.7902	0.150	-5.274	0.000	-1.084	-0.497
ma.L1	-0.7971	0.119	-6.688	0.000	-1.031	-0.563
sigma2	0.0355	0.007	5.318	0.000	0.022	0.049
Ljung-Box (L1) (Q):		0.35	Jarque-Bera (JB):			5.17
Prob(Q):		0.56	Prob(JB):			0.08
Heteroskedasticity (H):		3.62	Skew:			-0.71
Prob(H) (two-sided):		0.03	Kurtosis:			4.15

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

WA02: ARIMAX(2, 1, 2) (exog = 5 parameters)

Dep. Variable:		WQI	No. Observations:			38
Model:		SARIMAX(2, 1, 2)	Log Likelihood			29.252
Date:		Fri, 29 Sep 2023	AIC			-38.505
Time:		08:49:45	BIC			-22.396
Sample:		0	HQIC			-32.826
		- 38				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	0.1759	0.085	2.065	0.039	0.009	0.343
BOD(mg/l)	-0.5661	0.097	-5.817	0.000	-0.757	-0.375
Total Coli(MPN/100ml)	0.3221	0.114	2.826	0.005	0.099	0.545
Fecal Coli(MPN/100ml)	-0.4676	0.120	-3.892	0.000	-0.703	-0.232
NH3-N(mg/l)	-0.4025	0.145	-2.773	0.006	-0.687	-0.118
ar.L1	-1.0390	0.219	-4.753	0.000	-1.467	-0.611
ar.L2	-0.6796	0.230	-2.961	0.003	-1.129	-0.230
ma.L1	0.6134	3.740	0.164	0.870	-6.717	7.944
ma.L2	-0.3794	1.449	-0.262	0.793	-3.220	2.461
sigma2	0.0107	0.039	0.273	0.785	-0.066	0.087
Ljung-Box (L1) (Q):		0.23	Jarque-Bera (JB):			0.71
Prob(Q):		0.63	Prob(JB):			0.70
Heteroskedasticity (H):		0.61	Skew:			0.07
Prob(H) (two-sided):		0.40	Kurtosis:			3.67

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

WA02: SARIMA(2, 1, 2)(2, 2, 0, 4)

Dep. Variable:		WQI	No. Observations:			38
Model:		SARIMAX(2, 1, 2)x(2, 2, [], 4)	Log Likelihood			-121.171
Date:		Fri, 29 Sep 2023	AIC			256.342
Time:		08:51:55	BIC			265.913
Sample:		0	HQIC			259.339
		- 38				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7555	0.086	-8.752	0.000	-0.925	-0.586
ar.L2	-0.9559	0.074	-12.927	0.000	-1.101	-0.811
ma.L1	0.3531	7.533	0.047	0.963	-14.411	15.117
ma.L2	0.9982	41.918	0.024	0.981	-81.160	83.157
ar.S.L4	-1.4191	0.164	-8.633	0.000	-1.741	-1.097
ar.S.L8	-0.7708	0.151	-5.089	0.000	-1.068	-0.474
sigma2	122.4857	5110.375	0.024	0.981	-9893.664	1.01e+04
Ljung-Box (L1) (Q):		0.15	Jarque-Bera (JB):			0.39
Prob(Q):		0.70	Prob(JB):			0.82
Heteroskedasticity (H):		3.06	Skew:			-0.19
Prob(H) (two-sided):		0.09	Kurtosis:			2.57

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

WA02: SARIMAX(2, 1, 2)(0, 0, 1, 4) (exog = BOD)

Dep. Variable:		WQI	No. Observations:			38
Model:		SARIMAX(2, 1, 2)x(0, 0, [1], 4)	Log Likelihood			13.261
Date:		Fri, 29 Sep 2023	AIC			-12.521
Time:		09:00:46	BIC			-1.245
Sample:		0	HQIC			-8.546
		- 38				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.4884	0.096	-5.085	0.000	-0.677	-0.300
ar.L1	-0.0171	0.064	-0.267	0.790	-0.142	0.108
ar.L2	-0.9797	0.075	-12.999	0.000	-1.127	-0.832
ma.L1	-0.6143	14.366	-0.043	0.966	-28.771	27.543
ma.L2	0.9995	46.692	0.021	0.983	-90.515	92.514
ma.S.L4	-0.8416	0.251	-3.359	0.001	-1.333	-0.350
sigma2	0.0226	1.055	0.021	0.983	-2.044	2.090
Ljung-Box (L1) (Q):		1.06	Jarque-Bera (JB):			0.39
Prob(Q):		0.30	Prob(JB):			0.82
Heteroskedasticity (H):		3.31	Skew:			-0.03
Prob(H) (two-sided):		0.05	Kurtosis:			2.50

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

WA02: SARIMAX(2, 1, 1)(0, 0, 2, 4) (exog = 5 parameters)

Dep. Variable:		WQI	No. Observations:			38
Model:		SARIMAX(2, 1, 1)x(0, 0, [1, 2], 4)	Log Likelihood			30.617
Date:		Fri, 29 Sep 2023	AIC			-39.235
Time:		09:04:16	BIC			-21.515
Sample:		0	HQIC			-32.988
		- 38				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	0.1793	0.079	2.273	0.023	0.025	0.334
BOD(mg/l)	-0.5444	0.076	-7.133	0.000	-0.694	-0.395
Total Coli(MPN/100ml)	0.3038	0.126	2.418	0.016	0.058	0.550
Fecal Coli(MPN/100ml)	-0.4567	0.133	-3.437	0.001	-0.717	-0.196
NH3-N(mg/l)	-0.4786	0.137	-3.493	0.000	-0.747	-0.210
ar.L1	-1.3817	0.230	-5.999	0.000	-1.833	-0.930
ar.L2	-0.9821	0.246	-3.992	0.000	-1.464	-0.500
ma.L1	0.9580	0.474	2.023	0.043	0.030	1.886
ma.S.L4	0.4854	0.943	0.515	0.607	-1.362	2.333
ma.S.L8	-0.4012	0.538	-0.745	0.456	-1.456	0.654
sigma2	0.0089	0.006	1.379	0.168	-0.004	0.022
Ljung-Box (L1) (Q):		1.43	Jarque-Bera (JB):			0.66
Prob(Q):		0.23	Prob(JB):			0.72
Heteroskedasticity (H):		0.61	Skew:			0.19
Prob(H) (two-sided):		0.40	Kurtosis:			3.54

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

3. ค่าพารามิเตอร์และค่าทางสถิติของแบบจำลองอนุกรมเวลา แม่น้ำยม สถานี YO01

YO01: ARIMA(0, 1, 3)

Dep. Variable:	WQI	No. Observations:	37
Model:	SARIMAX(0, 1, 3)	Log Likelihood	-130.223
Date:	Fri, 29 Sep 2023	AIC	268.446
Time:	14:43:27	BIC	274.780
Sample:	0 - 37	HQIC	270.657
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-1.3903	325.055	-0.004	0.997	-638.487	635.707
ma.L2	-0.2107	128.362	-0.002	0.999	-251.796	251.374
ma.L3	0.6058	197.094	0.003	0.998	-385.691	386.903
sigma2	61.2968	2e+04	0.003	0.998	-3.9e+04	3.92e+04

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	0.64
Prob(Q):	0.74	Prob(JB):	0.73
Heteroskedasticity (H):	1.04	Skew:	0.25
Prob(H) (two-sided):	0.95	Kurtosis:	2.59

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

YO01: ARIMAX(0, 1, 1) (exog = 5 parameters)

Dep. Variable:	WQI	No. Observations:	37
Model:	SARIMAX(0, 1, 1)	Log Likelihood	21.502
Date:	Thu, 28 Sep 2023	AIC	-37.003
Time:	21:53:00	BIC	-32.252
Sample:	0 - 37	HQIC	-35.345
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.6212	0.064	-9.663	0.000	-0.747	-0.495
ma.L1	-0.9969	2.724	-0.366	0.714	-6.335	4.341
sigma2	0.0161	0.042	0.378	0.705	-0.067	0.099

Ljung-Box (L1) (Q):	0.31	Jarque-Bera (JB):	1.40
Prob(Q):	0.58	Prob(JB):	0.50
Heteroskedasticity (H):	0.74	Skew:	0.48
Prob(H) (two-sided):	0.60	Kurtosis:	3.04

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

YO01: ARIMAX(0, 1, 1) (exog = BOD)

Dep. Variable:	WQI	No. Observations:	37
Model:	SARIMAX(0, 1, 1)	Log Likelihood	24.993
Date:	Thu, 28 Sep 2023	AIC	-35.986
Time:	21:52:29	BIC	-24.902
Sample:	0 - 37	HQIC	-32.118
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	0.0504	0.148	0.341	0.733	-0.239	0.340
BOD(mg/l)	-0.5227	0.079	-6.589	0.000	-0.678	-0.367
Total Coli(MPN/100ml)	-0.2273	0.445	-0.511	0.609	-1.099	0.644
Fecal Coli(MPN/100ml)	-0.0742	0.809	-0.092	0.927	-1.660	1.511
NH3-N(mg/l)	-0.0739	0.177	-0.418	0.676	-0.421	0.273
ma.L1	-0.9999	148.905	-0.007	0.995	-292.849	290.849
sigma2	0.0132	1.966	0.007	0.995	-3.840	3.866

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	4.89
Prob(Q):	0.94	Prob(JB):	0.09
Heteroskedasticity (H):	0.62	Skew:	0.81
Prob(H) (two-sided):	0.42	Kurtosis:	3.81

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

YO01: SARIMA(3, 0, 2)(3, 2, 0, 4)

Dep. Variable:	WQI	No. Observations:	37
Model:	SARIMAX(3, 0, 2)x(3, 2, [], 4)	Log Likelihood	-109.079
Date:	Thu, 28 Sep 2023	AIC	236.157
Time:	21:49:59	BIC	248.463
Sample:	0 - 37	HQIC	240.011
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8520	0.378	-2.255	0.024	-1.592	-0.112
ar.L2	-0.4957	0.488	-1.015	0.310	-1.453	0.461
ar.L3	0.4009	0.371	1.081	0.280	-0.326	1.128
ma.L1	1.2117	0.568	2.135	0.033	0.099	2.324
ma.L2	0.9112	0.969	0.940	0.347	-0.989	2.811
ar.S.L4	-0.8316	0.233	-3.569	0.000	-1.288	-0.375
ar.S.L8	-0.6734	0.238	-2.829	0.005	-1.140	-0.207
ar.S.L12	-0.7429	0.155	-4.805	0.000	-1.046	-0.440
sigma2	64.0822	48.788	1.313	0.189	-31.540	159.704

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	7.11
Prob(Q):	0.93	Prob(JB):	0.03
Heteroskedasticity (H):	0.78	Skew:	0.92
Prob(H) (two-sided):	0.70	Kurtosis:	4.58

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

YO01: SARIMAX(0, 1, 1)(0, 0, 1, 4) (exog = BOD)

Dep. Variable:	WQI	No. Observations:	37
Model:	SARIMAX(0, 1, 1)x(0, 0, 1, 4)	Log Likelihood	23.364
Date:	Thu, 28 Sep 2023	AIC	-38.728
Time:	21:57:25	BIC	-32.394
Sample:	0 - 37	HQIC	-36.517
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.6505	0.050	-12.936	0.000	-0.749	-0.552
ma.L1	-0.9140	0.169	-5.406	0.000	-1.245	-0.583
ma.S.L4	-0.4133	0.219	-1.890	0.059	-0.842	0.015
sigma2	0.0146	0.003	4.196	0.000	0.008	0.021

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	1.52
Prob(Q):	0.74	Prob(JB):	0.47
Heteroskedasticity (H):	1.12	Skew:	0.50
Prob(H) (two-sided):	0.85	Kurtosis:	3.14

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

YO01: SARIMAX(0, 1, 1)(0, 0, 1, 4) (exog = 5 parameters)

Dep. Variable:	WQI	No. Observations:	37
Model:	SARIMAX(0, 1, 1)x(0, 0, 1, 4)	Log Likelihood	28.112
Date:	Thu, 28 Sep 2023	AIC	-40.223
Time:	22:01:17	BIC	-27.555
Sample:	0 - 37	HQIC	-35.802
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	-0.0921	0.101	-0.908	0.364	-0.291	0.107
BOD(mg/l)	-0.5431	0.066	-8.210	0.000	-0.673	-0.413
Total Coli(MPN/100ml)	-0.1865	0.244	-0.764	0.445	-0.665	0.292
Fecal Coli(MPN/100ml)	-0.0489	0.316	-0.155	0.877	-0.668	0.570
NH3-N(mg/l)	-0.1335	0.089	-1.493	0.135	-0.309	0.042
ma.L1	-0.9704	0.766	-1.266	0.205	-2.472	0.532
ma.S.L4	-0.6202	0.277	-2.235	0.025	-1.164	-0.076
sigma2	0.0103	0.006	1.719	0.086	-0.001	0.022

Ljung-Box (L1) (Q):	0.35	Jarque-Bera (JB):	1.36
Prob(Q):	0.55	Prob(JB):	0.51
Heteroskedasticity (H):	1.21	Skew:	0.40
Prob(H) (two-sided):	0.74	Kurtosis:	3.51

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

4. ค่าพารามิเตอร์และค่าทางสถิติของแบบจำลองอนุกรมเวลา แม่น้ำน่าน สถานี NA02

NA02: ARIMA(2, 1, 3)

Dep. Variable:	WQI	No. Observations:	38
Model:	SARIMAX(2, 1, 3)	Log Likelihood	-126.158
Date:	Fri, 29 Sep 2023	AIC	264.317
Time:	11:58:09	BIC	273.982
Sample:	0	HQIC	267.725
	- 38		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1449	0.214	0.679	0.497	-0.274	0.563
ar.L2	-0.6986	0.200	-3.496	0.000	-1.090	-0.307
ma.L1	-1.5389	3.874	-0.397	0.691	-9.133	6.055
ma.L2	1.5363	7.552	0.203	0.839	-13.266	16.338
ma.L3	-0.9953	6.730	-0.148	0.882	-14.187	12.196
sigma2	43.4373	289.155	0.150	0.881	-523.296	610.170

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	0.02
Prob(Q):	0.74	Prob(JB):	0.99
Heteroskedasticity (H):	1.11	Skew:	-0.06
Prob(H) (two-sided):	0.86	Kurtosis:	2.98

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

NA02: ARIMAX(0, 1, 1) (exog = BOD)

Dep. Variable:	WQI	No. Observations:	38
Model:	SARIMAX(0, 1, 1)	Log Likelihood	15.010
Date:	Fri, 29 Sep 2023	AIC	-24.020
Time:	19:49:59	BIC	-19.187
Sample:	0	HQIC	-22.316
	- 38		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.6423	0.181	-3.550	0.000	-0.997	-0.288
ma.L1	-0.9981	4.633	-0.215	0.829	-10.078	8.082
sigma2	0.0236	0.107	0.221	0.825	-0.186	0.233

Ljung-Box (L1) (Q):	0.10	Jarque-Bera (JB):	2.34
Prob(Q):	0.75	Prob(JB):	0.31
Heteroskedasticity (H):	0.84	Skew:	0.61
Prob(H) (two-sided):	0.77	Kurtosis:	2.82

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

NA02: ARIMAX(3, 1, 3) (exog = 5 parameters)

Dep. Variable:	WQI	No. Observations:	38
Model:	SARIMAX(3, 1, 3)	Log Likelihood	33.545
Date:	Thu, 28 Sep 2023	AIC	-43.091
Time:	21:13:03	BIC	-23.760
Sample:	0	HQIC	-36.275
	- 38		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	0.2140	0.104	2.058	0.040	0.010	0.418
BOD(mg/l)	-0.5333	0.088	-6.085	0.000	-0.705	-0.362
Total Coli(MPN/100ml)	-0.0500	0.087	-0.572	0.567	-0.221	0.121
Fecal Coli(MPN/100ml)	-0.3570	0.109	-3.269	0.001	-0.571	-0.143
NH3-N(mg/l)	-0.1453	0.118	-1.230	0.219	-0.377	0.086
ar.L1	-0.1706	0.439	-0.389	0.697	-1.030	0.689
ar.L2	-0.3602	0.395	-0.911	0.362	-1.135	0.415
ar.L3	0.4750	0.379	1.254	0.210	-0.267	1.217
ma.L1	-0.9587	0.790	-1.214	0.225	-2.507	0.589
ma.L2	0.7311	0.351	2.081	0.037	0.042	1.420
ma.L3	-0.7220	0.407	-1.772	0.076	-1.521	0.077
sigma2	0.0085	0.005	1.690	0.091	-0.001	0.018

Ljung-Box (L1) (Q):	0.27	Jarque-Bera (JB):	1.29
Prob(Q):	0.61	Prob(JB):	0.52
Heteroskedasticity (H):	0.58	Skew:	0.41
Prob(H) (two-sided):	0.35	Kurtosis:	2.61

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

NA02: SARIMA(3, 1, 0)(1, 2, 1, 4)

Dep. Variable:	WQI	No. Observations:	38
Model:	SARIMAX(3, 1, 0)x(1, 2, [1], 4)	Log Likelihood	-112.245
Date:	Thu, 28 Sep 2023	AIC	236.489
Time:	21:14:42	BIC	244.693
Sample:	0	HQIC	239.058
	- 38		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.2143	0.198	-6.137	0.000	-1.602	-0.827
ar.L2	-1.0830	0.256	-4.238	0.000	-1.584	-0.582
ar.L3	-0.4893	0.227	-2.152	0.031	-0.935	-0.044
ar.S.L4	-0.8005	0.186	-4.310	0.000	-1.164	-0.436
ma.S.L4	-0.7618	0.496	-1.536	0.124	-1.734	0.210
sigma2	89.0126	34.142	2.607	0.009	22.095	155.930

Ljung-Box (L1) (Q):	0.18	Jarque-Bera (JB):	0.41
Prob(Q):	0.68	Prob(JB):	0.81
Heteroskedasticity (H):	1.74	Skew:	-0.28
Prob(H) (two-sided):	0.39	Kurtosis:	2.81

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

NA02: SARIMAX(0, 1, 1)(0, 0, 0, 4) (exog = BOD)

Dep. Variable:	WQI	No. Observations:	38
Model:	SARIMAX(0, 1, 1)	Log Likelihood	15.010
Date:	Wed, 18 Oct 2023	AIC	-24.020
Time:	01:00:43	BIC	-19.187
Sample:	0	HQIC	-22.316
	- 38		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
BOD(mg/l)	-0.6423	0.181	-3.550	0.000	-0.997	-0.288
ma.L1	-0.9981	4.633	-0.215	0.829	-10.078	8.082
sigma2	0.0236	0.107	0.221	0.825	-0.186	0.233

Ljung-Box (L1) (Q):	0.10	Jarque-Bera (JB):	2.34
Prob(Q):	0.75	Prob(JB):	0.31
Heteroskedasticity (H):	0.84	Skew:	0.61
Prob(H) (two-sided):	0.77	Kurtosis:	2.82

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

NA02: SARIMAX(2, 1, 0)(2, 0, 3, 4) (exog = 5 paraemetrs)

Dep. Variable:	WQI	No. Observations:	38
Model:	SARIMAX(2, 1, 0)x(2, 0, [1, 2, 3], 4)	Log Likelihood	40.264
Date:	Wed, 18 Oct 2023	AIC	-54.528
Time:	00:56:21	BIC	-33.586
Sample:	0	HQIC	-47.145
	- 38		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
DO(mg/l)	0.1520	0.044	3.459	0.001	0.066	0.238
BOD(mg/l)	-0.6918	0.044	-15.757	0.000	-0.778	-0.606
Total Coli(MPN/100ml)	0.0637	0.035	1.827	0.068	-0.005	0.132
Fecal Coli(MPN/100ml)	-0.5284	0.085	-6.205	0.000	-0.695	-0.362
NH3-N(mg/l)	0.0159	0.052	0.308	0.758	-0.085	0.117
ar.L1	-1.0774	0.178	-6.067	0.000	-1.425	-0.729
ar.L2	-0.7865	0.153	-5.151	0.000	-1.086	-0.487
ar.S.L4	-0.1651	0.353	-0.468	0.640	-0.857	0.527
ar.S.L8	-0.6011	0.205	-2.938	0.003	-1.002	-0.200
ma.S.L4	-0.7743	8.726	-0.089	0.929	-17.878	16.329
ma.S.L8	-0.9035	8.976	-0.101	0.920	-18.496	16.689
ma.S.L12	0.8119	4.506	0.180	0.857	-8.020	9.644
sigma2	0.0026	0.013	0.196	0.845	-0.024	0.029

Ljung-Box (L1) (Q):	0.36	Jarque-Bera (JB):	3.51
Prob(Q):	0.55	Prob(JB):	0.17
Heteroskedasticity (H):	0.51	Skew:	0.72
Prob(H) (two-sided):	0.26	Kurtosis:	3.47

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

ประวัติผู้เขียน