



การทำนายคุณภาพอากาศในจังหวัดปทุมธานี ประเทศไทย โดยใช้ข้อมูลอุตุนิยมวิทยา
ด้วยเทคนิคทางสถิติและการเรียนรู้ของเครื่อง

Predicting Air Quality in Pathum Thani, Thailand, Based on Meteorological Data Using Statistical and Machine Learning Methods

นันทรัตน์ โอภาช, นาซีน แก้วประดับ, ปามอนปน พุมิปามอร์*, บุญยฤทธิ์ ชูประดิษฐ์

สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ปทุมธานี 12120

Nantharat Aophat, Nasneen Kaewpradab, Pamornpon Phumipamorn*, Boonyarit Choopradit

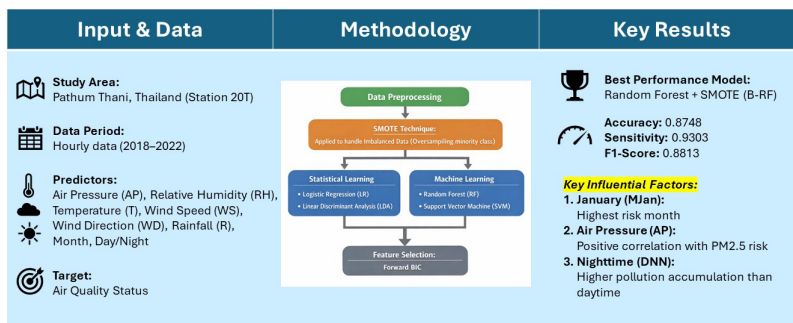
Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University,

Pathum Thani 12120

Received 25 April 2025; Received in revised 18 February 2026; Accepted 9 April 2026

GRAPHICAL ABSTRACT

ABSTRACT



Air pollution, particularly fine particulate matter (PM_{2.5}), is a significant environmental issue affecting Thailand. This study compares the effectiveness of various air quality prediction methods in Pathum Thani Province and investigates the factors influencing local air quality forecasts. Statistical and machine learning methods, including binary logistic regression, linear discriminant analysis, random forest, and support vector

machine, are employed. The models are evaluated using performance metrics such as accuracy, sensitivity, specificity, precision, and F1-score. The results indicate that the random forest model, using data balanced with the synthetic minority over-sampling technique (SMOTE), achieves the best performance, with an accuracy of 0.8748, sensitivity of 0.9303, precision of 0.8373, and an F1-score of 0.8813. The most influential factors in predicting air quality are the month of January, atmospheric pressure, and nighttime intervals, followed by relative humidity, temperature, and wind speed, in decreasing order of importance.

คำสำคัญ

ข้อมูลไม่สมดุล; ปทุมธานี;
การเรียนรู้เชิงสถิติ;
การเรียนรู้ของเครื่อง

Keywords

Imbalanced data;
Pathum Thani;
Statistical learning;
Machine learning

บทคัดย่อ

ปัญหามลพิษทางอากาศโดยเฉพาะฝุ่นละออง PM_{2.5} เป็นปัญหาสิ่งแวดล้อมที่มีผลกระทบอย่างรุนแรงในประเทศไทย งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการในการทำนายคุณภาพอากาศในจังหวัดปทุมธานี และศึกษาปัจจัยที่มีผลในการทำนายคุณภาพอากาศในจังหวัดปทุมธานี โดยใช้การเรียนรู้เชิงสถิติและการเรียนรู้ของเครื่อง ได้แก่ การถดถอยลอจิสติกทวิภาค (Binary logistic regression) และการวิเคราะห์จำแนกกลุ่มเชิงเส้น (Linear discriminant analysis) ป่าสุ่ม (Random forest) และเครื่องเวกเตอร์ค้ำจุน (Support vector machine) โดยใช้เกณฑ์ในการประเมินแบบจำลองประกอบด้วยค่า ความแม่นยำ (Accuracy) ความไว (Sensitivity) ความจำเพาะ (Specificity) ความเที่ยง (Precision) และ F1-score ผลการวิจัย พบว่า แบบจำลองที่ใช้เทคนิคป่าสุ่มบนข้อมูลที่มีการปรับสมดุลด้วยเทคนิค SMOTE ให้ผลลัพธ์แม่นยำสูงสุด โดยมีค่าความแม่นยำเท่ากับ 0.8748 ค่าความไวเท่ากับ 0.9303 ค่าความเที่ยงเท่ากับ 0.8373 และ F1-score เท่ากับ 0.8813 โดยปัจจัยที่ส่งผลกระทบต่อทำนายคุณภาพอากาศมากที่สุด ได้แก่ เดือนมกราคม ความกดอากาศ ช่วงเวลากลางคืน รวมถึงตัวแปรอื่นๆ เช่น ความชื้นสัมพัทธ์ อุณหภูมิ และความเร็วลม ซึ่งช่วยให้แบบจำลองสามารถคาดการณ์คุณภาพอากาศได้อย่างมีประสิทธิภาพ

*ผู้รับผิดชอบบทความ: newreborn66@hotmail.com

DOI: 10.14456/tstj.2026.25

1. บทนำ

มลพิษทางอากาศนับเป็นปัญหาสิ่งแวดล้อมสำคัญของประเทศไทยอย่างต่อเนื่อง โดยเฉพาะฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5: Particulate matter $\leq$ 2.5 microns) ซึ่งส่งผลกระทบต่อสุขภาพและคุณภาพชีวิตของประชาชนในวงกว้าง ปัจจุบันกรุงเทพมหานครและพื้นที่ปริมณฑลกำลังเผชิญกับสถานการณ์มลพิษทางอากาศในระดับรุนแรง โดยข้อมูลจากกรมควบคุมมลพิษประจำปี พ.ศ. 2563 ระบุว่าค่าความเข้มข้นของ PM2.5 ในหลายพื้นที่มีค่าสูงเกินเกณฑ์มาตรฐานอย่างต่อเนื่อง โดยเฉพาะในพื้นที่ภาคกลางของประเทศไทย สถานการณ์ดังกล่าวสะท้อนถึงความซับซ้อนของปัญหามลพิษทางอากาศและความจำเป็นในการพัฒนาแนวทางการเฝ้าระวังและบริหารจัดการอย่างมีประสิทธิภาพ

ในช่วงทศวรรษที่ผ่านมาวิธีการเรียนรู้ของเครื่องได้รับการพัฒนาและนำมาใช้สร้างแบบจำลองคุณภาพอากาศอย่างแพร่หลาย เนื่องจากมีศักยภาพสูงในการจัดการความสัมพันธ์ที่ซับซ้อนของข้อมูลและสามารถพยากรณ์ความเข้มข้นของฝุ่นละออง PM2.5 ได้อย่างมีประสิทธิภาพ [1, 2] Fang et al. [3] ได้ศึกษาการพยากรณ์ PM2.5 โดยใช้วิธีการเรียนรู้ของเครื่อง 4 แบบ คือ การถดถอยเชิงเส้นพหุคูณ, Backpropagation neural network (BPNN), เครื่องเวกเตอร์ค้ำจุน และ ป่าสุ่ม พบว่า วิธีป่าสุ่มมีความแม่นยำในการพยากรณ์สูงสุดและสามารถอธิบายการเปลี่ยนแปลงของความเข้มข้น PM2.5 ตามชั่วโมงภายใต้สภาพอากาศที่แตกต่างกันได้อย่างแม่นยำ Makhdoomi et al. [4] พยากรณ์ความเข้มข้นของ PM2.5 โดยใช้การเรียนรู้ของเครื่อง 4 วิธี ได้แก่ Light gradient boosting machine (LGBM), Extreme gradient boosting regressor (XGBR), ป่าสุ่ม และ Gradient boosting regressor (GBR) ใช้ข้อมูลช่วงปี 2016-2022 พบว่าค่า PM2.5 มีความผันผวนสูงโดยแบบจำลองทั้งหมดให้ผลการพยากรณ์ที่แม่นยำสูง 95% ของค่าที่พยากรณ์ใกล้เคียงค่าจริง ใน

บรรดาแบบจำลองทั้งหมด GBR ให้สมรรถนะดีที่สุดด้วยค่า MSE และ RMSE ต่ำสุด Warangkhan [5] ได้ศึกษาข้อมูลเกี่ยวกับประสิทธิภาพของเกณฑ์การคัดเลือกตัวแบบสำหรับตัวแบบการถดถอยพหุคูณ การศึกษาเปรียบเทียบเกณฑ์การคัดเลือกที่สำคัญ เช่น Akaike information criterion (AIC), Bayesian information criterion (BIC), และ Hannan-Quinn information criterion (HQIC) ผลการศึกษาพบว่า AIC และ HQIC มีประสิทธิภาพสูงในกรณีที่ตัวอย่างมีขนาดเล็กและตัวแบบยากที่จะระบุ ในขณะที่ BIC อาจมีประสิทธิภาพมากกว่าเมื่อจำนวนตัวแปรในตัวแบบมากขึ้นและตัวอย่างมีขนาดใหญ่

จากผลกระทบของฝุ่นละอองขนาดเล็ก PM2.5 ต่อสุขภาพและคุณภาพชีวิตของประชาชน จึงมีความจำเป็นอย่างยิ่งที่จะต้องมีความก้าวหน้าในการทำนายคุณภาพอากาศล่วงหน้า เพื่อใช้เป็นข้อมูลในการแจ้งเตือนและช่วยให้ประชาชนในพื้นที่สามารถเตรียมรับมือกับสถานการณ์มลพิษที่อาจเกิดขึ้นได้อย่างเหมาะสม ด้วยเหตุนี้ คณะผู้วิจัยจึงพัฒนาแบบจำลองทำนายคุณภาพอากาศโดยใช้ 2 แนวทางหลัก ได้แก่ เทคนิคทางสถิติ (Statistical learning) คือ การถดถอยลอจิสติกทวิภาค (Binary logistic regression) และการวิเคราะห์จำแนกกลุ่มเชิงเส้น (Linear discriminant analysis) รวมถึงเทคนิคการเรียนรู้ของเครื่อง (Machine learning) ได้แก่ ป่าสุ่ม และเครื่องเวกเตอร์ค้ำจุน ทั้งนี้ เพื่อแก้ไขปัญหาข้อมูลไม่สมดุล ได้มีการใช้เทคนิคการเพิ่มข้อมูลในคลาสขนาดเล็กด้วยการสังเคราะห์ข้อมูลใหม่จากข้อมูลเดิม (Synthetic minority oversampling technique: SMOTE) จากนั้นทำการประเมินประสิทธิภาพของแบบจำลองโดยใช้ตารางเมทริกซ์ความสับสน (Confusion matrix) ซึ่งพิจารณาจากเกณฑ์การวัด 5 ค่า ได้แก่ ความแม่นยำ (Accuracy) ความไว (Sensitivity) ความจำเพาะ (Specificity) ความเที่ยง (Precision) และ F1-score เพื่อสะท้อนสมรรถนะของแบบจำลองอย่างรอบด้าน

2. วิธีการดำเนินงานวิจัย

2.1 ขอบเขตการศึกษา

ผู้วิจัยกำหนดพื้นที่ในการศึกษาการทำนายคุณภาพอากาศในจังหวัดปทุมธานี โดยกรมควบคุมมลพิษ กำหนดสถานีตรวจวัดเพียงแห่งเดียวคือ สถานีมหาวิทยาลัยกรุงเทพ วิทยาเขตรังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี ละติจูด 14°02'14.7"N ลองจิจูด 100°36'18.8"E รหัสสถานี 20T โดยเก็บข้อมูลเป็นรายชั่วโมงระหว่าง 1 มกราคม 2561 เวลา 00:00 น. ถึง 31 ธันวาคม 2565 เวลา 23:00 น. [6]

2.2 ตัวแปรที่ใช้ในการศึกษา

Table 1 แสดงตัวแปรในการศึกษา แบ่งเป็นตัวแปรอิสระ ได้แก่ 1) ความกดอากาศ (Air pressure; AP) 2) ความชื้นสัมพัทธ์ (Relative humidity; RH) 3) อุณหภูมิ (Temperature; T) 4) ความเร็วลม (Wind speed; WS) 5) ทิศทางลม (Wind direction; WD) 6) ปริมาณน้ำฝน (Rainfall; R) 7) เดือน (Month; M) 8) ช่วงเวลา (Day & night; DN) ตัวแปรตาม ได้แก่

คุณภาพอากาศ (Air quality) ในจังหวัดปทุมธานี โดยแบ่งเป็น 2 กลุ่มคือ Negative class : คุณภาพอากาศไม่มีผลต่อสุขภาพ (PM2.5 มีค่า 0-37.5 มคก./ลบ.ม.) และ Positive class : คุณภาพอากาศมีผลต่อสุขภาพ (PM2.5 มีค่า มากกว่า 37.5 มคก./ลบ.ม.) โดยอ้างอิงจากประกาศกรมควบคุมมลพิษ เรื่องดัชนีคุณภาพอากาศของประเทศไทย พ.ศ. 2566 [7]

2.3 ขั้นตอนการศึกษา

1. รวบรวมชุดข้อมูลคุณภาพอากาศและข้อมูลอุตุนิยมวิทยาจากสถานีตรวจวัด เพื่อสร้างชุดข้อมูลเริ่มต้น (Dataset)
2. เตรียมข้อมูลโดยตรวจสอบและจัดการข้อมูลสูญหาย (Missing data) ให้มีความสมบูรณ์ก่อนการวิเคราะห์
3. แบ่งกระบวนการวิเคราะห์ออกเป็น 2 กรณี ได้แก่ กรณีที่ไม่มีการจัดการข้อมูลไม่สมดุล และกรณีที่มีการจัดการข้อมูลไม่สมดุลด้วยเทคนิค SMOTE
4. แบ่งข้อมูลออกเป็นชุดฝึกฝน (Training set) ร้อยละ 80 และชุดทดสอบ (Testing set) ร้อยละ 20

Table 1 Dependent variable and independent variable

Attributes	Description	Attribute types
Air pressure (AP)	at 2 m (mbar)	Continuous
Relative humidity (RH)	at 2 m (%RH)	Continuous
Temperature (T)	at 2 m (Celsius)	Continuous
Wind speed (WS)	at 10 m (m/s)	Continuous
Rainfall (R)	at 3 m (mm)	Continuous
Wind direction (WD)	at 10 m (Deg.M)	Categorical
Month (M)	-	Categorical
Day/Night (DN)	-	Categorical
Air Quality	Dependent variable	Categorical

โดยมีจำนวนวันที่คุณภาพอากาศไม่มีผลต่อสุขภาพต่อจำนวนวันที่คุณภาพอากาศมีผลต่อสุขภาพคิดเป็นอัตราส่วน 16.1% ต่อ 83.9% และเป็นอัตราส่วนที่เท่ากันทั้งชุดฝึกฝนและชุดทดสอบ

5. สร้างแบบจำลองจากชุดฝึกฝนโดยใช้เทคนิคเชิงสถิติ ได้แก่ การถดถอยลอจิสติกทวิภาค และการวิเคราะห์จำแนกกลุ่มเชิงเส้น สร้างแบบจำลองจากชุดฝึกฝนโดยใช้เทคนิคการเรียนรู้ของเครื่อง ได้แก่ ปาสุมและเครื่องเวกเตอร์ค้ำจุน (Support vector machine: SVM)

6. นำแบบจำลองที่ได้ไปทำนายข้อมูลในชุดทดสอบ

7. สร้างตารางเมทริกซ์ความสับสนเพื่อเปรียบเทียบค่าจริงกับค่าที่แบบจำลองทำนาย

8. ประเมินประสิทธิภาพของแบบจำลองด้วยตัวชี้วัด ได้แก่ ความแม่นยำ ความไว ความจำเพาะ ความเที่ยง และ F1-score

9. เปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมด เพื่อสรุปวิธีที่ให้ผลการทำนายคุณภาพอากาศดีที่สุด

2.4 แบบจำลอง

2.4.1 การวิเคราะห์การถดถอยลอจิสติกทวิภาค

แบบจำลองการถดถอยลอจิสติกซึ่งเสนอโดย Cox [8] และถูกนำไปประยุกต์ใช้อย่างกว้างขวางในหลายสาขาวิชา เป็นเทคนิคทางสถิติที่ใช้ในการทำนายความน่าจะเป็นของการเกิดเหตุการณ์ การถดถอยลอจิสติกและการวิเคราะห์การถดถอยทั่วไปมีความคล้ายกันที่ตัวแปรตามถูกอธิบายด้วยการรวมเชิงเส้นของตัวแปรอิสระ การถดถอยลอจิสติกจัดเป็นเทคนิคการจำแนกประเภทเนื่องจากตัวแปรตามเป็นข้อมูลเชิงจัดประเภท ข้อมูลที่ใช้ในการศึกษานี้ถูกจัดประเภทตามแบบจำลองการถดถอยลอจิสติก โดยสามารถแสดงความสัมพันธ์ระหว่างตัวแปรต่างๆ ได้อย่างชัดเจน การวิเคราะห์

ด้วยการถดถอยลอจิสติกมีความเรียบง่ายและตีความค่าสัมประสิทธิ์ได้ง่าย แนวคิดพื้นฐานของการถดถอยลอจิสติกคือการใช้ตัวทำนายเชิงเส้นในลักษณะเดียวกับการถดถอยเชิงเส้น การถดถอยลอจิสติกจะทำให้ค่าผลลัพธ์ของตัวแปรตามอยู่ในช่วงความน่าจะเป็น $[0,1]$ โดยไม่ขึ้นกับช่วงค่าของตัวแปรอิสระและส่วนมากจะใช้ค่าเกณฑ์ (Threshold) เท่ากับ 0.5 ในการจำแนกผลลัพธ์ออกเป็น 0 หรือ 1 [9] โดยงานวิจัยนี้ใช้ฟังก์ชัน `glm()` ในภาษาอาร์ กำหนดค่าไฮเปอร์พารามิเตอร์ (Hyperparameter) ดังนี้ 1) Random seed = 1234 2) Link function = Logit 3) Estimation = MLE (IRLS) 4) Max iterations = 25 5) Convergence tolerance = $1e-8$ 6) Classification threshold = 0.5 หรือ 0.16

2.4.2 การวิเคราะห์จำแนกกลุ่มเชิงเส้น

การวิเคราะห์จำแนกกลุ่มเชิงเส้นเป็นวิธีการจำแนกประเภทข้อมูลแบบมีผู้สอนที่อาศัยการสร้างเส้นแบ่งเชิงเส้นเพื่อแยกกลุ่มข้อมูลออกจากกัน หลักการสำคัญคือการเพิ่มความแตกต่างระหว่างกลุ่มให้มากที่สุด ขณะเดียวกันลดความแปรปรวนภายในกลุ่มให้น้อยที่สุด วิธีการนี้มีพื้นฐานทางสถิติที่ชัดเจนและสามารถตีความผลลัพธ์ได้โดยตรง การวิเคราะห์จำแนกกลุ่มเชิงเส้นมีประสิทธิภาพดีเมื่อข้อมูลในแต่ละกลุ่มมีการกระจายใกล้เคียงแบบปกติและมีความแปรปรวนร่วมเท่ากัน จุดเด่นของการวิเคราะห์จำแนกกลุ่มเชิงเส้น คือความเรียบง่าย ความเร็วในการคำนวณ อย่างไรก็ตามประสิทธิภาพอาจลดลงเมื่อสมมติฐานของแบบจำลองไม่สอดคล้องกับลักษณะของข้อมูลจริง [10] โดยงานวิจัยนี้ใช้แพ็คเกจ `MASS::lda` ในภาษาอาร์ กำหนดค่าไฮเปอร์พารามิเตอร์ (Hyperparameter) ดังนี้ 1) Random seed = 1234 2) Prior probabilities = Empirical (จากข้อมูลฝึกสอน) 3) Covariance estimation = Sample covariance 4) Tolerance (tol) = 0.0001 5) Classification threshold = 0.5 หรือ 0.16

2.4.3 ป่าสุ่ม

ป่าสุ่มเป็นเทคนิคการเรียนรู้แบบกลุ่ม (Ensemble learning) ที่ถูกเสนอโดย Breiman [11] โดยอาศัยการรวมผลการทำนายจากต้นไม้ตัดสินใจหลายต้นซึ่งมีลักษณะที่แตกต่างกัน ส่งผลให้ประสิทธิภาพในการทำนายโดยรวมของแบบจำลองเพิ่มขึ้น แบบจำลองป่าสุ่มประกอบด้วยโครงสร้างแบบต้นไม้จำนวนมาก ซึ่งแต่ละต้นถูกฝึกจากข้อมูลตัวอย่างที่ได้จากกระบวนการบูตสแตรปแบบรวมผล (Bootstrap aggregating: Bagging) หรือการสุ่มตัวอย่างแบบมีการแทนที่ ในกระบวนการเรียนรู้การแตกกิ่งของแต่ละต้นไม้จะพิจารณาจากชุดของตัวแปรที่ถูกสุ่มเลือกมาเพียงบางส่วนในแต่ละโหนด ซึ่งช่วยเพิ่มความหลากหลายระหว่างต้นไม้ ผลลัพธ์สุดท้ายของแบบจำลองจะถูกกำหนดจากเสียงข้างมากของผลการทำนายเหล่านั้น นอกจากนี้แบบจำลองป่าสุ่มยังสามารถใช้ในการประเมินความสำคัญของตัวแปรอิสระ (Feature importance) ภายในแบบจำลองได้อีกด้วย [11, 12] โดยงานวิจัยนี้ใช้แพ็คเกจ Random forest ในภาษาอาร์ กำหนดค่าไฮเปอร์พารามิเตอร์ ดังนี้ 1) Random seed = 1234 2) Number of trees (ntree) = 500 3) Number of variables tried at each split (mtry) = $\lfloor p/3 \rfloor$ 4) ไม่ได้กำหนดค่า Maximum tree depth 5) Sampling method = Bootstrap sampling 6) Variable importance คำนวณด้วย IncNodePurity

2.4.4 เครื่องเวกเตอร์ค้ำจุน

เครื่องเวกเตอร์ค้ำจุนเป็นอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอน ที่สามารถนำมาใช้ในการแก้ปัญหการจำแนกประเภท โดยเครื่องเวกเตอร์ค้ำจุนจะพิจารณาข้อมูลแต่ละรายการเป็นจุดข้อมูลในปริภูมิแบบ n มิติ ซึ่ง n แทนจำนวนตัวแปรหรือคุณลักษณะ (Features) อัลกอริทึมเครื่องเวกเตอร์ค้ำจุนจะสร้างไฮเปอร์เพลน (Hyperplane) เพื่อใช้เป็นเส้นแบ่งในการจำแนกข้อมูลใหม่ จุดข้อมูลที่อยู่ใกล้ไฮเปอร์เพลน

มากที่สุดและมีอิทธิพลต่อการกำหนดตำแหน่งของไฮเปอร์เพลนเรียกว่าเวกเตอร์ค้ำจุน (Support vectors) หลักการสำคัญของเครื่องเวกเตอร์ค้ำจุนคือการสร้างไฮเปอร์เพลนที่มีระยะขอบ (Margin) มากที่สุด ซึ่งหมายถึงการเพิ่มระยะห่างระหว่างกลุ่มข้อมูลให้มากที่สุดเท่าที่เป็นไปได้เพื่อลดความคลาดเคลื่อนในการทำนายประสิทธิภาพของเครื่องเวกเตอร์ค้ำจุนในงานจำแนกประเภทขึ้นอยู่กับกำหนัดค่าพารามิเตอร์อย่างเหมาะสม โดยเฉพาะ พารามิเตอร์ค่าปรับ (Penalty factor) C ซึ่งควบคุมสมดุลระหว่างการลดความผิดพลาดจากข้อมูลฝึก ความกว้างของระยะขอบและพารามิเตอร์เคอร์เนล (Kernel parameter) ซึ่งกำหนดลักษณะการแปลงข้อมูลในกรณีใช้ Gaussian (RBF) kernel [13] โดยงานวิจัยนี้ใช้แพ็คเกจ e1071 ในภาษาอาร์ กำหนดค่าไฮเปอร์พารามิเตอร์ ดังนี้ 1) Random seed = 1234 2) Kernel = Radial (RBF) 3) Cost (C) = 1 4) Gamma = $1 / p$ 5) Epsilon = 0.1 6) Tolerance = 0.001

2.5 การจัดการข้อมูลไม่สมดุลด้วยวิธี SMOTE

SMOTE เป็นวิธีการทางสถิติที่ใช้สร้างตัวอย่างข้อมูลใหม่ เพื่อเพิ่มจำนวนข้อมูลของคลาสส่วนน้อยในชุดข้อมูล โดยวิธีนี้จะพิจารณาตำแหน่งของข้อมูลในพื้นที่คุณลักษณะ (Feature space) ของคลาสเป้าหมายร่วมกับข้อมูลเพื่อนบ้านที่ใกล้ที่สุด จากนั้นจึงสร้างข้อมูลสังเคราะห์ใหม่ด้วยการผสมค่าคุณลักษณะระหว่างตัวอย่างเป้าหมายกับเพื่อนบ้านเหล่านั้น ส่งผลให้ข้อมูลที่สร้างขึ้นมีลักษณะอยู่ระหว่างจุดข้อมูลเดิม ไม่ใช่การคัดลอกข้อมูลของคลาสส่วนน้อยที่มีอยู่โดยตรง วิธีนี้ช่วยเพิ่มความหลากหลายและกระจายตัวของข้อมูลส่วนน้อย พร้อมทั้งหลีกเลี่ยงปัญหาภาวะฟิตเกิน (Overfitting) ที่เกิดจากการคัดลอกข้อมูลซ้ำๆ ได้ [14]

2.6 การประเมินประสิทธิภาพตัวแบบ

การศึกษานี้ประเมินความสามารถในการทำนายของตัวแบบ โดยใช้ตัวชี้วัดที่สำคัญได้แก่ ความแม่นยำ ความไว ความจำเพาะ ความเที่ยง และ F1-score โดยมีรูปแบบทางคณิตศาสตร์แสดงไว้ด้านล่าง [10]

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

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

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

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

$$\text{F1-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

โดยที่ TP แทนจำนวนผลลัพธ์บวกจริง (True positive) TN แทนจำนวนผลลัพธ์ลบจริง (True negative) FP แทนจำนวนผลลัพธ์บวกเท็จ (False positive) และ FN แทนจำนวนผลลัพธ์ลบเท็จ (False negative)

3. ผลการวิจัย

3.1 ผลการวิเคราะห์กรณีที่ไม่มีการจัดการข้อมูลไม่สมดุล

Table 2 และ Table 3 แสดงผลการประเมินประสิทธิภาพของแบบจำลองที่สร้างจากข้อมูลไม่สมดุล โดยยังไม่มี การคัดเลือกตัวแปร และผลการประเมินประสิทธิภาพของแบบจำลองที่สร้างจากข้อมูลไม่สมดุล โดยมีการคัดเลือกตัวแปร ตามลำดับ โดยแต่ละตาราง แสดงการเปรียบเทียบพฤติกรรมของวิธีเชิงสถิติและการเรียนรู้ของเครื่อง

Table 2 พบว่าแบบจำลองการถดถอยลอจิสติก พหุภาคีกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลใช้ค่าเกณฑ์เป็น 0.5 (IM-LR (0.5)) และการวิเคราะห์จำแนกกลุ่มเชิงเส้นกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลใช้ค่าเกณฑ์เป็น 0.5 (IM-LDA (0.5)) ให้ค่าความแม่นยำและความจำเพาะสูงมาก แสดงถึงความสามารถในการจำแนกกลุ่มคุณภาพอากาศที่ไม่มีผลต่อสุขภาพได้อย่างแม่นยำ อย่างไรก็ตาม ค่าความไวของทั้งสองตัวแบบมีค่าต่ำมาก (0.2466 และ

Table 2 Model evaluation without variable selection

Models	Accuracy	Sensitivity	Specificity	Precision	F1-Score
IM-LR (0.5) ¹	0.8481	0.2466	0.9635	0.5643	0.3432
IM-LR (0.16) ²	0.7341	0.8085	0.7198	0.3562	0.4946
IM-LDA (0.5) ³	0.8450	0.3395	0.9419	0.5286	0.4134
IM-LDA (0.16) ⁴	0.7655	0.6984	0.7783	0.3766	0.4894
IM-RF	0.8511	0.6904	0.8819	0.5285	0.5987
IM-SVM	0.7837	0.6101	0.8170	0.3900	0.4759

IM : Imbalanced data

¹ : Binary logistic regression with a classification threshold of 0.5

² : Binary logistic regression with a classification threshold of 0.16

³ : Linear discriminant analysis with a classification threshold of 0.5

⁴ : Linear discriminant analysis with a classification threshold of 0.16

Table 3 Model evaluation with forward BIC variable selection

Models	Accuracy	Sensitivity	Specificity	Precision	F1-Score
IM-F-LR (0.5) ¹	0.8485	0.2477	0.9637	0.5669	0.3448
IM-F-LR (0.16) ²	0.7346	0.8096	0.7203	0.3569	0.4954
IM-F-LDA (0.5) ³	0.8437	0.3395	0.9404	0.5221	0.4114
IM-F-LDA (0.16) ⁴	0.7627	0.6984	0.7750	0.3732	0.4864
IM-F-RF	0.8544	0.6227	0.8988	0.5414	0.5792
IM-F-SVM	0.7837	0.6101	0.8170	0.3900	0.4759

F : Forward

0.3395 ตามลำดับ) ส่งผลให้ไม่สามารถตรวจจับเหตุการณ์คุณภาพอากาศที่มีผลต่อสุขภาพได้ดี ซึ่งสะท้อนข้อจำกัดของการใช้ค่าเกณฑ์เป็น 0.5 ในบริบทข้อมูลไม่สมดุล เมื่อปรับค่าเกณฑ์เป็น 0.16 ทั้งในตัวอย่างการถดถอยลอจิสติกทวิภาคกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลและตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุล พบว่าค่าความไวเพิ่มขึ้นอย่างชัดเจน (สูงถึง 0.8085 และ 0.6984) แต่แลกมาด้วยการลดลงของค่าความจำเพาะและค่าความเที่ยง ทำให้แบบจำลองมีแนวโน้มทำนายกลุ่มคุณภาพอากาศที่มีผลต่อสุขภาพเกินจริงและมีความผิดพลาดเพิ่มขึ้น เมื่อพิจารณาแบบจำลองการเรียนรู้ของเครื่อง พบว่าตัวแบบป่าสุ่มกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุล (IM-RF) ให้ผลลัพธ์โดยรวมสมดุลที่สุด โดยมีค่าความแม่นยำสูงสุด (0.8511) พร้อมกับค่าความไว (0.6904) และค่าความจำเพาะ (0.8819) ที่อยู่ในระดับค่อนข้างสูงพร้อมๆ กัน ส่งผลให้ค่า F1-score สูงที่สุด (0.5987) ด้วยเช่นกัน ซึ่งสะท้อนความสามารถในการจัดการระหว่างการตรวจจับกลุ่มคุณภาพอากาศที่มีผลต่อสุขภาพและการลดความผิดพลาดในการทำนายกลุ่มคุณภาพอากาศที่ไม่มีผลต่อสุขภาพได้ดีกว่าวิธีอื่นๆ ขณะที่เครื่องเวกเตอร์ค้ำจุนกรณีที่ไม่มีการจัดการข้อมูล

ไม่สมดุล (IM-SVM) ให้ผลลัพธ์ในระดับปานกลาง โดยมีค่าความไวและค่าความจำเพาะอยู่ในช่วงใกล้เคียงกัน แต่ค่าความเที่ยงและ F1-score ต่ำกว่าตัวแบบป่าสุ่มแสดงถึงความสามารถในการจำแนกที่ยังไม่เหมาะสมเท่าตัวแบบป่าสุ่มภายใต้ข้อมูลที่ไม่สมดุล

จาก Table 3 พบว่าแบบจำลองเชิงสถิติการถดถอยลอจิสติกทวิภาคกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือกตัวแปรใช้ค่าเกณฑ์เป็น 0.5 (IM-F-LR (0.5)) และการวิเคราะห์จำแนกกลุ่มเชิงเส้นกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือกตัวแปรใช้ค่าเกณฑ์เป็น 0.5 (IM-F-LDA (0.5)) ยังคงมีพฤติกรรมใกล้เคียงกับกรณีไม่คัดเลือกตัวแปร กล่าวคือให้ค่าความแม่นยำและค่าความจำเพาะสูงมาก (มากกว่า 0.94) แต่มีค่าความไวต่ำมาก (ประมาณ 0.25-0.34) แสดงว่าการใช้เทคนิค Forward BIC จะช่วยลดจำนวนตัวแปรได้แต่ยังไม่สามารถแก้ปัญหาการตรวจจับเหตุการณ์คุณภาพอากาศที่มีผลต่อสุขภาพได้เมื่อใช้ค่าเกณฑ์มาตรฐาน 0.5 แต่เมื่อปรับค่าเกณฑ์เป็น 0.16 ทั้งในการถดถอยลอจิสติกทวิภาคกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือกตัวแปรและการวิเคราะห์จำแนกกลุ่มเชิงเส้นกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือกตัวแปร พบว่าค่าความไวเพิ่ม

ขึ้นอย่างชัดเจน (0.8096 และ 0.6984 ตามลำดับ) สะท้อนความสามารถในการตรวจจับกลุ่มคุณภาพอากาศที่มีผลต่อสุขภาพที่ดีขึ้น อย่างไรก็ตามค่าความจำเพาะและค่าความเที่ยงลดลงอย่างเห็นได้ชัด ส่งผลให้ค่า F1-score อยู่ในระดับปานกลาง และบ่งชี้ถึงการแลกเปลี่ยน (Trade-off) ระหว่างการเพิ่มอัตราการตรวจจับกับความผิดพลาดในการทำนายที่เพิ่มขึ้น สำหรับแบบจำลองการเรียนรู้ของเครื่องตัวแบบป่าสุ่มกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือกตัวแปร (IM-F-RF) ให้ผลลัพธ์ที่ดีที่สุดโดยรวมในตารางนี้ โดยมีค่าความแม่นยำสูงสุด (0.8544) และรักษาความสมดุลระหว่างค่าความไว (0.6227) และค่าความจำเพาะ (0.8988) ได้ดีกว่าวิธีอื่น ส่งผลให้ค่า F1-score สูงที่สุด (0.5792) แสดงว่าวิธีป่าสุ่มยังคงมีประสิทธิภาพเมื่อมีการลดจำนวนตัวแปรและสามารถคงประสิทธิภาพในการจำแนกข้อมูลที่ไม่สมดุลได้ดี ในขณะที่เครื่องเวกเตอร์ค้ำจุนกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือกตัวแปร IM-F-SVM ให้ผลลัพธ์ใกล้เคียงกับกรณีไม่คัดเลือกตัวแปร โดยมีค่าความไวและค่าความจำเพาะอยู่ในระดับปานกลาง แต่ค่าความเที่ยงและค่า F1-score ต่ำกว่าวิธีป่าสุ่มแสดงถึงข้อจำกัดของเครื่องเวกเตอร์ค้ำจุนภายใต้ข้อมูลไม่สมดุลแม้มีการคัดเลือกตัวแปรแล้ว

3.2 ผลการวิเคราะห์กรณีที่มีการจัดการข้อมูลไม่สมดุลด้วยเทคนิค SMOTE

Table 4 และ Table 5 แสดงผลการประเมินแบบจำลองหลังจากจัดการปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE โดยยังไม่มีคัดเลือกตัวแปรและการคัดเลือกตัวแปร ตามลำดับ เพื่อวิเคราะห์ผลของการปรับสมดุลข้อมูลต่อประสิทธิภาพการทำนายคุณภาพอากาศ โดยแสดงการเปรียบเทียบพฤติกรรมของวิธีเชิงสถิติและการเรียนรู้ของเครื่อง

Table 4 พบว่าเมื่อมีการปรับสมดุลข้อมูลด้วยเทคนิค SMOTE ประสิทธิภาพของแบบจำลองทุกวิธีดีขึ้นอย่างชัดเจนในมิติของการตรวจจับกลุ่มคุณภาพอากาศที่มีผลต่อสุขภาพ โดยเฉพาะค่าความไวซึ่งเพิ่มขึ้นมาอยู่ในระดับสูงกว่า 0.82 สำหรับทุกตัวแบบ แสดงให้เห็นว่าแบบจำลองสามารถตรวจจับเหตุการณ์คุณภาพอากาศที่เริ่มส่งผลต่อสุขภาพได้ดีขึ้นอย่างมีนัยสำคัญ สำหรับแบบจำลองเชิงสถิติการถดถอยลอจิสติกทวิภาคใช้ค่าเกณฑ์เป็น 0.5 ปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-LR (0.5)) และการวิเคราะห์จำแนกกลุ่มเชิงเส้นใช้ค่าเกณฑ์เป็น 0.5 ปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-LDA (0.5)) พบว่ามีค่าความแม่นยำในระดับใกล้เคียงกัน (ประมาณ 0.77) และให้ค่าความไวสูง (0.8245 และ

Table 4 Model evaluation without variable selection and SMOTE for class balancing

Models	Accuracy	Sensitivity	Specificity	Precision	F1-Score
B-LR (0.5) ¹	0.7715	0.8245	0.7185	0.7455	0.7830
B-LDA (0.5) ²	0.7702	0.8399	0.7005	0.7371	0.7852
B-RF	0.8748	0.9303	0.8192	0.8373	0.8813
B-SVM	0.7612	0.8335	0.6888	0.7381	0.7773

B : Balanced data

¹ : Binary logistic regression with a classification threshold of 0.5

² : Linear discriminant analysis with a classification threshold of 0.5

0.8399) ควบคู่กับค่าความจำเพาะที่ลดลงมาอยู่ในช่วง 0.70-0.72 สะท้อนว่าการปรับสมดุลข้อมูลช่วยลดอคติในการทำนายกลุ่มคุณภาพอากาศที่ไม่มีผลต่อสุขภาพ และทำให้ตัวแบบมีความสมดุลมากขึ้น อย่างไรก็ตามค่าความเที่ยงและค่า F1-score ของทั้งสองวิธียังคงอยู่ในระดับปานกลาง แสดงถึงข้อจำกัดในการจำแนกกลุ่มคุณภาพอากาศที่มีผลต่อสุขภาพ ส่วนในกลุ่มแบบจำลองการเรียนรู้ของเครื่องวิธีป่าสุ่มปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-RF) ให้ผลลัพธ์ที่ดีที่สุดอย่างชัดเจนในทุกตัวชี้วัด โดยมีค่าความแม่นยำสูงสุด (0.8748) และค่าความไวสูงมากถึง 0.9303 แสดงถึงความสามารถในการตรวจจับเหตุการณ์คุณภาพอากาศที่มีผลต่อสุขภาพได้อย่างมีประสิทธิภาพ พร้อมทั้งยังคงค่าความจำเพาะ (0.8192) และค่าความเที่ยง (0.8373) ในระดับสูง ส่งผลให้ค่า F1-score สูงที่สุด (0.8813) ซึ่งสะท้อนความสมดุลที่ดีเยี่ยมระหว่างการตรวจจับกลุ่มคุณภาพอากาศที่มีผลต่อสุขภาพและการลดความผิดพลาดในการทำนาย ขณะที่เครื่องเวกเตอร์ค้ำจุนปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-SVM) แม้จะมีค่าความไว สูง (0.8335) แต่ค่าความจำเพาะและค่าความเที่ยงต่ำกว่าวิธีป่าสุ่ม ส่งผลให้ค่า F1-score ต่ำกว่า ซึ่งแสดงถึงข้อจำกัดของเครื่องเวกเตอร์ค้ำจุนเมื่อเทียบกับวิธีป่าสุ่มภายใต้ข้อมูลที่ถูกปรับสมดุลแล้ว

จาก Table 5 พบว่าภายหลังการปรับสมดุลข้อมูลด้วยเทคนิค SMOTE แบบจำลองเชิงสถิติการถด

ถอยลอจิสติกทวิภาคใช้ค่าเกณฑ์เป็น 0.5 โดยมีการคัดเลือกตัวแปรและปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-F-LR (0.5)) และการวิเคราะห์จำแนกกลุ่มเชิงเส้นใช้ค่าเกณฑ์เป็น 0.5 โดยมีการคัดเลือกตัวแปรและปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-F-LDA (0.5)) ให้ผลลัพธ์ใกล้เคียงกันมาก โดยมีค่าความแม่นยำประมาณ 0.77 และให้ค่าความไวสูง (0.8247 และ 0.8399) สะท้อนว่าสามารถตรวจจับเหตุการณ์คุณภาพอากาศที่มีผลต่อสุขภาพได้ดี อย่างไรก็ตามค่าความจำเพาะอยู่ในระดับปานกลาง (ประมาณ 0.70-0.72) แสดงถึงการแลกเปลี่ยนระหว่างการเพิ่มความสามารถในการตรวจจับกลุ่มคุณภาพอากาศที่มีผลต่อสุขภาพกับการทำนายกลุ่มคุณภาพอากาศที่ไม่มีผลต่อสุขภาพที่แม่นยำน้อยลง ค่าความเที่ยงและค่า F1-score อยู่ในช่วง 0.74-0.79 ซึ่งบ่งชี้ว่าแม้ตัวแบบเชิงสถิติจะมีเสถียรภาพดีขึ้นเมื่อใช้การปรับสมดุลข้อมูลด้วยเทคนิค SMOTE และมีการคัดเลือกตัวแปรร่วมกัน แต่ยังคงมีข้อจำกัดด้านความสมดุลของการจำแนกเมื่อเทียบกับตัวแบบที่มีการจัดการกับข้อมูลที่ซับซ้อนน้อยกว่า สำหรับแบบจำลองการเรียนรู้ของเครื่องเมื่อมีการคัดเลือกตัวแปรแล้ว พบว่าวิธีป่าสุ่มที่มีการคัดเลือกตัวแปรและปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-F-RF) ยังคงเป็นตัวแบบที่ให้ผลลัพธ์โดยรวมดีที่สุด โดยมีค่าความแม่นยำสูงถึง 0.8621 และค่าความไว 0.9112 แสดงถึงความสามารถในการตรวจจับเหตุการณ์คุณภาพอากาศที่มีผลต่อสุขภาพได้อย่างมี

Table 5 Model evaluation with forward BIC variable selection and SMOTE for class balancing

Models	Accuracy	Sensitivity	Specificity	Precision	F1-Score
B-F-LR (0.5) ¹	0.7716	0.8247	0.7185	0.7455	0.7831
B-F-LDA (0.5) ²	0.7701	0.8399	0.7002	0.7370	0.7851
B-F-RF	0.8621	0.9112	0.8131	0.8298	0.8686
B-F-SVM	0.7612	0.8335	0.6888	0.7381	0.7773

ประสิทธิภาพ พร้อมทั้งคงค่าความจำเพาะ (0.8131) และค่าความเที่ยง (0.8298) ในระดับสูง ส่งผลให้ค่า F1-score เท่ากับ 0.8686 ซึ่งสะท้อนความสมดุลที่ดีมาก ระหว่างการตรวจจับและความแม่นยำในการทำนาย ผลลัพธ์นี้ชี้ให้เห็นว่าวิธีป่าสุ่มมีความทนทานต่อการคัด เลือกตัวแปรและยังคงมีประสิทธิภาพสูงแม้ข้อมูลมีปรับ สมดุล ขณะที่เครื่องเวกเตอร์คำจูนที่มีการคัดเลือกตัวแปร และปรับสมดุลข้อมูลด้วยเทคนิค SMOTE (B-F-SVM) ให้ค่าความไวค่อนข้างสูง (0.8335) แต่ค่าความจำเพาะ และค่าความเที่ยงต่ำกว่าวิธีป่าสุ่มส่งผลให้ค่า F1-score ต่ำกว่า และสะท้อนข้อจำกัดของเครื่องเวกเตอร์คำจูนใน การสร้างสมดุลระหว่างตัวชี้วัดภายใต้กรอบการวิจัยนี้

3.3 ผลการศึกษาปัจจัยที่มีผลในการทำนายคุณภาพ อากาศในจังหวัดปทุมธานี

Table 6 สรุปได้ว่าตัวแบบการถดถอยลอจิสติก ทวิภาคกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการ คัดเลือกตัวแปร (IM-F-LR) แสดงให้เห็นว่าในเดือน มกราคม (MJan) มีโอกาสที่คุณภาพอากาศเริ่มส่งผลต่อ

สุขภาพเพิ่มขึ้นเกือบ 16 เท่า เมื่อเทียบกับเดือนตุลาคม (Baseline) สะท้อนว่าช่วงต้นปีมีแนวโน้มอากาศแย่กว่า ทิศลมตะวันออก (WDE) เพิ่มโอกาสเกิดมลพิษ 1.4771 เท่า เมื่อเทียบกับทิศตะวันตก (Baseline) ในช่วงกลาง คืน (DNN) มีโอกาสมลพิษเพิ่มขึ้น 1.247 เท่า เมื่อเทียบ กับช่วงกลางวัน (Baseline) และเมื่อความกดอากาศ (AP) เพิ่มขึ้น 1 mbar จะเพิ่มโอกาสมลพิษอีกเกือบ ร้อยละ 20 โดยรวมแสดงให้เห็นว่าปัจจัยด้านฤดูกาล ทิศทางลม ช่วงเวลา และความกดอากาศมีความสัมพันธ์ เชิงบวกกับความเสี่ยงต่อคุณภาพอากาศที่แย่ลง ซึ่งเป็น ข้อมูลสำคัญต่อการวางแผนจัดการคุณภาพอากาศใน พื้นที่ต่างๆ ในส่วนของตัวแบบการถดถอยลอจิสติก ทวิภาคโดยมีการคัดเลือกตัวแปรและปรับสมดุลข้อมูล ด้วยเทคนิค SMOTE (B-F-LR) การสรุปผลเป็นไปใน แนวทางเดียวกับตัวแบบการถดถอยลอจิสติกทวิภาค กรณีที่ไม่มีการจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือก ตัวแปร

Table 6 Odds ratios (OR) and 95% confidence intervals for two models (IM-F-LR and B-F-LR)

Attributes	Methods	Odd ratio (OR)	
		IM-F-LR	B-F-LR
January		15.9239 (12.8844, 19.8299)	13.7324 (12.1899, 15.4871)
February		8.6414 (6.9464, 10.8241)	7.2894 (6.4574, 8.2359)
December		7.4902 (6.0560, 9.3322)	6.3202 (5.6468, 7.0802)
March		5.4656 (4.3398, 6.9220)	5.1657 (4.5691, 5.8446)
April		4.1329 (3.2673, 5.2533)	3.8464 (3.4031, 4.3506)
November		3.5773 (2.8803, 4.4722)	3.0244 (2.7050, 3.3840)
WDE		1.4771 (1.3150, 1.6595)	1.7228 (1.5999, 1.8552)
DNN		1.2470 (1.1191, 1.3900)	1.2568 (1.1761, 1.3431)
AP		1.1821 (1.1102, 1.2588)	1.1327 (1.0897, 1.1774)

สำหรับเดือนมกราคมที่มีค่าอัตราส่วนออกดส์ (Odds ratio) สูงที่สุดในตัวแบบการถดถอยลอจิสติก ทวิภาคกรณีที่ไม่มีจัดการข้อมูลไม่สมดุลโดยมีการคัดเลือกตัวแปรและตัวแบบการถดถอยลอจิสติกทวิภาค โดยมีการคัดเลือกตัวแปรและปรับสมดุลข้อมูลด้วยเทคนิค SMOTE สามารถอธิบายเชิงบริบทพื้นที่ได้จากการเป็นช่วงฤดูหนาวที่มีสภาพอากาศนิ่ง ลมอ่อน และชั้นผกผันอุณหภูมิ (Temperature inversion) เกิดได้บ่อย ส่งผลให้การระบายมลพิษในบรรยากาศต่ำลง นอกจากนี้ยังเป็นช่วงที่มีกิจกรรมภายนอกเพิ่มขึ้น เช่น มีการเผาในที่โล่งทางกิจกรรมการเกษตรในพื้นที่ใกล้เคียง การสะสมฝุ่นจากการจราจรหลังช่วงเทศกาลปีใหม่ และการลดลงของปริมาณน้ำฝน ซึ่งทำให้กลไกการชะล้างมลพิษตามธรรมชาติอ่อนลงไป ปัจจัยเหล่านี้ทำให้ความ

เข้มข้นของ PM2.5 มีแนวโน้มสะสมต่อเนื่อง ส่งผลให้โอกาสที่คุณภาพอากาศเริ่มส่งผลกระทบต่อสุขภาพในเดือนมกราคมสูงกว่าเดือนตุลาคมอย่างมีนัยสำคัญ

จากการพิจารณาใน 9 อันดับแรกพบว่า ปัจจัยที่มีอิทธิพลต่อการทำนายคุณภาพอากาศมากที่สุดคือ เดือนมกราคม (MJan), ความกดอากาศ (AP) และช่วงเวลากลางคืน (DNN) ส่วนปัจจัยอื่น เช่น ความชื้นสัมพัทธ์ (RH), อุณหภูมิ (T) และ ความเร็วลม (WS) แม้มีอิทธิพลน้อยกว่า แต่ยังช่วยเพิ่มความแม่นยำของการทำนายเมื่อใช้ร่วมกับปัจจัยหลัก

นอกจากนั้น พบว่าตัวแบบการถดถอยลอจิสติกทวิภาคให้ความสำคัญกับตัวแปรเดือนมกราคม (MJan) และทิศทางลม (WD) เนื่องจากโครงสร้างของตัวแบบเน้นความสัมพันธ์เชิงเส้นต่อ Log-odds และเหมาะกับ

Table 7 Feature importance

Feature importance			
IM-RF	IM-F-RF	B-RF	B-F-RF
RH	T	RH	RH
MJan	WS	T	T
T	MJan	WS	WS
WS	AP	MJan	AP
AP	WDS	AP	MJan
MDec	DNN	MDec	DNN
MFeb	WDE	MNov	WDE
DNN	MDec	MFeb	MMar
WDE	MFeb	DNN	MNov
Remark MJan : January	MDec : December	WDE : Wind direction (East)	
MFeb : February	AP : Air Pressure	WDS : Wind direction (South)	
MMar : March	RH : Relative humidity	DNN : Night	
MApr : April	T : Temperature		
MNov : November	WS : Wind speed		

ตัวแปรเชิงกลุ่มที่สะท้อนการเปลี่ยนแปลงเชิงโครงสร้างของระบบ เช่น ฤดูกาลและทิศทางแหล่งกำเนิดมลพิษ อีกทั้งผลของตัวแปรต่อเนื่องอย่างอุณหภูมิและความชื้น อาจถูกลดบทบาทลงจากปัญหาความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) กับตัวแปรฤดูกาล ในทางตรงกันข้ามเทคนิคป่าสุ่มเป็นวิธีการไม่เชิงเส้นที่อาศัยการแบ่งข้อมูลซ้ำๆ ทำให้ตัวแปรต่อเนื่องที่มีความแปรปรวนสูงชัดเจน เช่น ความชื้นและอุณหภูมิ สามารถลดความไม่บริสุทธิ์ (Impurity) ได้มากและถูกเลือกใช้บ่อยกว่า ความสำคัญของตัวแปร (Feature importance) จึงสะท้อนทั้งสมมติฐานของตัวแบบ กลไกการเรียนรู้และอคติของวิธีคำนวณมากกว่าความสำคัญเชิงกายภาพโดยตรงของตัวแปรนั้นๆ

4. สรุปผลการศึกษา

งานวิจัยนี้เปรียบเทียบแบบจำลองเชิงสถิติและการเรียนรู้ของเครื่องภายใต้ข้อมูล PM2.5 ที่ไม่สมดุล พร้อมทั้งแสดงให้เห็นว่าการจัดการข้อมูลให้สมดุลด้วยเทคนิค SMOTE มีบทบาทสำคัญต่อการเพิ่มประสิทธิภาพการจำแนกคุณภาพอากาศ โดยเฉพาะในตัวแบบป่าสุ่มซึ่งเหมาะสมกว่าการเลือกใช้อัลกอริทึมที่ซับซ้อนเพียงอย่างเดียว งานวิจัยนี้ใช้ข้อมูลรายชั่วโมงระยะยาว 5 ปี ร่วมกับตัวแปรอุตุนิยมวิทยาและตัวแปรเชิงเวลา ทำให้สามารถวิเคราะห์ได้ทั้งเชิงทำนายและเชิงการอธิบายผ่านการใช้อัตราส่วนออกไซด์และความสำคัญของตัวแปรควบคู่กัน ผลการวิจัยชี้ให้เห็นบทบาทของฤดูกาล (เดือน) ช่วงเวลากลางคืน และความกดอากาศต่อความเสี่ยงด้านคุณภาพอากาศ ซึ่งมีนัยสำคัญต่อการวางแผนนโยบายและการจัดการคุณภาพอากาศระดับพื้นที่ ผลการศึกษาการทำนายคุณภาพอากาศในจังหวัดปทุมธานี โดยใช้เทคนิคการวิเคราะห์การถดถอยลอจิสติกทวิภาค การวิเคราะห์จำแนกกลุ่มเชิงเส้น ป่าสุ่ม และเครื่องเวกเตอร์ค้ำจุน พบว่าตัวแบบป่าสุ่มที่จัดการข้อมูลไม่สมดุลด้วย SMOTE ให้ผลดีที่สุด โดยมีค่าความแม่นยำเท่ากับ 0.8748

ค่าความไวเท่ากับ 0.9303 ค่าความเที่ยงเท่ากับ 0.8373 ค่า F1-Score เท่ากับ 0.8813 และค่าความจำเพาะเท่ากับ 0.8192 แสดงถึงความแม่นยำและความน่าเชื่อถือสูงสุด ปัจจัยสำคัญจากการวิเคราะห์การถดถอยลอจิสติกทวิภาค ได้แก่ เดือนมกราคม-เมษายน, พฤศจิกายน-ธันวาคม, ทิศทางลมตะวันออก, ช่วงกลางวัน และความกดอากาศ ขณะที่ผลจากความสำคัญของตัวแปรของป่าสุ่มชี้ว่า ความชื้นสัมพัทธ์, อุณหภูมิ, ความเร็วลม, ความกดอากาศ, ช่วงกลางวัน และเดือนมกราคม-กุมภาพันธ์, พฤศจิกายน-ธันวาคม เป็นปัจจัยหลักที่มีผลต่อคุณภาพอากาศ

เนื่องจากข้อมูลที่ใช้ในการพัฒนาแบบจำลองมาจากสถานีตรวจวัดคุณภาพอากาศเพียงแห่งเดียว คือ สถานีมหาวิทยาลัยกรุงเทพ วิทยาเขตรังสิต ซึ่งมีบริบทเชิงพื้นที่เฉพาะ แบบจำลองที่ได้อาจสะท้อนรูปแบบความสัมพันธ์ระหว่าง PM2.5 และตัวแปรอุตุนิยมวิทยาที่เหมาะสมกับพื้นที่เป็นหลัก การนำแบบจำลองไปประยุกต์ใช้กับพื้นที่อื่นในจังหวัดปทุมธานีที่มีลักษณะแตกต่างกัน เช่น พื้นที่เมืองหนาแน่นหรือพื้นที่เกษตรกรรม อาจทำให้ความแม่นยำลดลง เนื่องจากแหล่งกำเนิดมลพิษ โครงสร้างการใช้ที่ดิน และพลวัตของการสะสมฝุ่นแตกต่างกัน นอกจากนี้แบบจำลองไม่ได้รวมปัจจัยเชิงพื้นที่เฉพาะ เช่น การจราจรหรือกิจกรรมอุตสาหกรรม จึงมีข้อจำกัดด้านความสามารถในการทำนายเชิงพื้นที่

5. ข้อเสนอแนะ

1. ผู้วิจัยอาจพิจารณาเทคนิคการเรียนรู้ของเครื่องอื่นๆ เพิ่มเติม เช่น K-Nearest neighbors (KNN), Artificial neural network (ANN) หรือ XGboost
2. ผู้วิจัยอาจนำปัจจัยด้านอื่นๆ มาพิจารณาเพิ่มเติม เพื่อให้การทำนายคุณภาพอากาศแม่นยำมากยิ่งขึ้น เช่น จำนวนรถที่ผ่านทางแยกต่างระดับต่างๆ รายชั่วโมง

3. ผู้วิจัยอาจมีการปรับข้อมูลสมมูลด้วยเทคนิคอื่นๆ เช่น Under-sampling, Over-sampling และ Combination sampling

6. กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณกองจัดการคุณภาพอากาศ และเสียง กรมควบคุมมลพิษ ที่ได้ให้ความอนุเคราะห์ข้อมูล ในการดำเนินงานวิจัยครั้งนี้

7. References

- [1] Hardini, M., Sunarjo, R. A., Asfi, M., Riza Chakim, M. H. and Ayu Sanjaya, Y. P., 2023, Predicting air quality index using ensemble machine learning, ADI J. Recent. Innov. (AJRI). 5 (1Sp): 78-86.
- [2] Yang, H., Wang, W. and Li, G., 202, Prediction method of PM2.5 concentration based on decomposition and integration, Measurement. 216(8): 112954.
- [3] Fang, Z., Yang, H., Li, C., Cheng, L., Zhao, M. and Xie, C., 2021, Prediction of PM2.5 hourly concentrations in Beijing based on machine learning algorithm and ground-based LiDAR, Arch. Environ. Prot. 47(3): 98-107.
- [4] Makhdoomi, A., Sarkhosh, M. and Ziaei, S., 2025, PM2.5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations, Sci. Rep. 15: 8076.
- [5] Warangkhan, R., 2023, A study of the effectiveness of model selection criteria for multiple regression model, Rajamangala Univ. Technol. Srivijaya Res. J (RST). 15(1), 198-212. (in Thai).
- [6] Pollution Control Department, 2024, Regional Air Quality and Situation Reports, Available Source: <http://air4thai.pcd.go.th/webV3/#/StationDetail>, Feb 2, 2024. (in Thai).
- [7] Pollution Control Department, 2025, "Pollution Control Department Announces Thailand's 2023 Air Quality Index.", Available Source: <https://www.pcd.go.th/laws/29909>, Jan 25, 2025. (in Thai).
- [8] Cox, D. R., 1958, The regression analysis of binary sequences, J. R. Stat. Soc. Ser. B Methodol. 20(2): 215-232.
- [9] Choi, J., Gu, B., Chin, S. and Lee, J.-S., 2020, Machine learning predictive model based on national data for fatal accidents of construction workers, Autom. Constr. 110(3): 102974.
- [10] Adebiyi, M.O., Arowolo, M.O., Mshelia, M.D. and Olugbara, O.O., 2022, A linear discriminant analysis and classification model for breast cancer diagnosis, Appl. Sci. 12(22): 11455.
- [11] Breiman L., 2001, Random forests, Mach. Learn. 45: 5-32.
- [12] Ali, Y., Hussain, F. and Haque, M. M., 2024, Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review, Accid. Anal. Prev. 194: 107378.

- [13] Liu, W., Guo, G., Chen, F. and Chen, Y., 2019, Meteorological pattern analysis assisted daily PM2.5 grades prediction using SVM optimized by PSO algorithm, *Atmos. Pollut. Res.* 10(5): 1482-1491.
- [14] Fiorentini, N. and Losa, M., 2020, Handling imbalanced data in road crash severity prediction by machine learning algorithms, *Infrastructures.* 5(7): 61.