PRAWARUN AGRICULTURAL JOURNAL

https://li01.tci-thaijo.org/index.php/pajrmu/index

บทความวิจัย

การพัฒนาแบบจำลองเพื่อพยากรณ์ปริมาณฝุ่นละออง PM2.5 โดยใช้เทคนิคการเรียนรู้ของเครื่อง: กรณีศึกษาประเทศไทย

ณิฏะญาร์ บรรเทา[1] พูนศักดิ์ ศิริโสม[1] ปรมาภรณ์ แสงภารา[1] วริดา พลาศรี[1] อุเทน จิณโรจน์[3] และ รดา สมเขื่อน[2*]

[1]สาขาวิชาสถิติศาสตร์ประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม อำเภอเมือง จังหวัดมหาสารคาม ประเทศไทย 44000

[2]กลุ่มวิชาคณิตศาสตร์ สาขาวิทยาศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยีการเกษตร มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา อำเภอเมือง จังหวัดเชียงใหม่ ประเทศไทย 50300

[3]สำนักงานสาธารณสุขอำเภอเชียงยืน อำเภอเชียงยืน จังหวัดมหาสารคาม ประเทศไทย 44160

ข้อมูลบทความ

Article history

รับ: 30 กรกฎาคม 2568

แก้ไข: 17 ตุลาคม 2568

ตอบรับการตีพิมพ์: 12 ธันวาคม 2568

ตีพิมพ์ออนไลน์: 30 มกราคม 2569

คำสำคัญ

ฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5)

การเรียนรู้ของเครื่อง

แบบจำลอง

สิ่งแวดล้อม

การเกษตร

บทคัดย่อ

ฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5) เป็นปัญหาสำคัญที่ส่งผลกระทบต่อสุขภาพ สิ่งแวดล้อม และภาคการเกษตรของประเทศไทย โดยเฉพาะในช่วงฤดูแล้งที่ระดับฝุ่นมักเกินค่ามาตรฐาน งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองสำหรับการพยากรณ์ (Forecasting Model) ค่าฝุ่น PM2.5 โดยเปรียบเทียบวิธีการพยากรณ์อนุกรมเวลา (Time Series Forecasting Methods) ได้แก่ วิธีโฮลท์–วินเทอร์ส (Holt–Winters) และเออาร์ไอม่า (ARIMA) กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning Techniques) ได้แก่ เรนดอมฟอเรสต์ (Random Forest) และซัพพอร์ตเวกเตอร์รีเกรสชัน (Support Vector Regression: SVR) ใช้ข้อมูลรายเดือนจากกรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา ครอบคลุมช่วงปี พ.ศ. 2561–2567 โดยแบ่งข้อมูลตามลำดับเวลาเป็นชุดฝึก (Training Set) และชุดทดสอบ (Test Set) ในสัดส่วน 80:20 และประเมินความแม่นยำด้วยตัวชี้วัด รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error: RMSE) และค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) ผลการศึกษาพบว่าเรนดอมฟอเรสต์ (Random Forest) ให้ค่าความคลาดเคลื่อนต่ำที่สุดในทุกพื้นที่ โดยเฉพาะจังหวัดที่มีความผันผวนสูง เช่น เชียงใหม่และขอนแก่น ขณะที่เอสวีอาร์ (SVR) ให้ผลลัพธ์แม่นยำน้อยกว่า แบบจำลองอนุกรมเวลา (Time Series Models) ยังคงให้ผลลัพธ์ที่ดีในพื้นที่ที่ข้อมูลมีความต่อเนื่อง เช่น สงขลา โดยปัจจัยสำคัญที่มีผลต่อความแม่นยำของแบบจำลอง ได้แก่ ตัวแปรแบบหน่วงเวลา (Lag Variables) และค่าเฉลี่ยเคลื่อนที่ (Moving Average) แบบจำลองที่พัฒนาขึ้น โดยเฉพาะเรนดอมฟอเรสต์ (Random Forest) แสดงศักยภาพสูงในการนำไปประยุกต์ใช้เป็นระบบแจ้งเตือนคุณภาพอากาศ (Air Quality Alert System) และสนับสนุนการตัดสินใจเชิงนโยบายด้านสิ่งแวดล้อมและภาคการเกษตรอย่างยั่งยืน

บทนำ

ตลอดช่วงระยะปี พ.ศ.2554 - 2567 ที่ผ่านมา ประเทศไทยต้องเผชิญกับปัญหาฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน หรือที่เรียกกันโดยทั่วไปว่า PM2.5 ซึ่งส่งผลกระทบต่อหลายด้าน ทั้งสุขภาพของประชาชน คุณภาพสิ่งแวดล้อม และระบบเศรษฐกิจโดยรวม ฝุ่น PM2.5 มีขนาดเล็กมากจนสามารถเข้าสู่ระบบทางเดินหายใจได้ลึกถึงถุงลมปอด ทำให้เพิ่มความเสี่ยงต่อโรคเรื้อรังต่าง ๆ เช่น โรคปอดอุดกั้นเรื้อรัง มะเร็งปอด และโรคหัวใจขาดเลือด โดยเฉพาะอย่างยิ่งในกลุ่มประชากรเปราะบาง เช่น ผู้สูงอายุ เด็กเล็ก และผู้ป่วยโรคประจำตัว (World Health Organization, 2021) ผลกระทบของ PM2.5 ไม่ได้จำกัดเฉพาะต่อสุขภาพของมนุษย์เท่านั้น แต่ยังแผ่ขยายไปสู่ภาคการเกษตรอย่างชัดเจน หลายงานวิจัยพบว่าฝุ่นละอองสามารถลดประสิทธิภาพของกระบวนการสังเคราะห์แสง ทำให้พืชเจริญเติบโตช้าลงและส่งผลให้ผลผลิตลดลงในพื้นที่เปิดโล่ง อีกทั้งยังอาจทำให้แรงงานภาคเกษตรมีปัญหาด้านสุขภาพ ส่งผลต่อแรงงานผลิตโดยรวมในระยะยาว สำหรับประเทศไทย แหล่งกำเนิด PM2.5 มีความหลากหลายและขึ้นอยู่กับบริบทของแต่ละพื้นที่ เช่น ในเขตเมืองใหญ่มักเกิดจากการคมนาคมขนส่งและกิจกรรมก่อสร้าง ขณะที่ในเขตชนบท โดยเฉพาะภาคเหนือและภาคตะวันออกเฉียงเหนือ ปัญหาหลักมักมาจากการเผาเศษวัสดุทางการเกษตรหลังฤดูเก็บเกี่ยว ไม่ว่าจะเป็นการเผาตอซังข้าว อ้อย หรือซังข้าวโพด แม้ว่ารัฐจะมีนโยบายควบคุมการเผาในที่โล่งและมีการรณรงค์อย่างต่อเนื่อง แต่การเปลี่ยนแปลงพฤติกรรมในระดับพื้นที่ยังคงเป็นความท้าทายที่ต้องใช้เวลา จังหวัดเชียงใหม่เป็นตัวอย่างชัดเจนของพื้นที่ที่ได้รับผลกระทบจากการเผาในภาคเกษตร โดยเฉพาะในช่วงฤดูแล้งระหว่างเดือนมีนาคมถึงเมษายน ซึ่งค่าฝุ่น PM2.5 มักสูงเกินมาตรฐานติดต่อกันนานหลายวัน ส่วนจังหวัดขอนแก่นซึ่งตั้งอยู่ในภาคตะวันออกเฉียงเหนือก็มีปัญหาฝุ่นจากกิจกรรมทางการเกษตรผสมกับการขยายตัวของเมืองอย่างรวดเร็ว ด้านกรุงเทพมหานครซึ่งเป็นศูนย์กลางทางเศรษฐกิจของประเทศ ประสบกับฝุ่นจากการจราจรหนาแน่นและมลพิษทางอุตสาหกรรม ขณะที่จังหวัดสงขลาในภาคใต้

*Corresponding author

E-mail address: rada@rmutl.ac.th (R. Somkhuean)

 https://doi.org/10.14456/paj.2026.1

ของประเทศไทย แม้จะไม่มีแหล่งกำเนิดมลพิษทางอากาศขนาดใหญ่ภายในพื้นที่เมื่อเทียบกับเขตเมืองหรือพื้นที่เกษตรกรรมในภาคเหนือและภาคตะวันออกเฉียงเหนือ แต่ยังคงพบปัญหาฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5) ในบางช่วงเวลา โดยเฉพาะในฤดูมรสุม ซึ่งมีรายงานการเคลื่อนย้ายของควันไฟป่าข้ามพรมแดนจากประเทศเพื่อนบ้าน ส่งผลให้ระดับความเข้มข้นของฝุ่นละอองเพิ่มสูงขึ้นเป็นระยะ รายงานสถานการณ์คุณภาพอากาศของประเทศไทยระบุว่าปรากฏการณ์ดังกล่าวเป็นหนึ่งในปัจจัยสำคัญที่ส่งผลต่อคุณภาพอากาศในพื้นที่ภาคใต้ แม้จะไม่มีแหล่งกำเนิดมลพิษหลักในพื้นที่โดยตรง (Pollution Control Department (PCD), 2023) ด้วยความซับซ้อนของปัจจัยที่เกี่ยวข้องกับ PM2.5 ทั้งด้านภูมิอากาศ พื้นที่ และพฤติกรรมมนุษย์ การเฝ้าระวังและการพยากรณ์ล่วงหน้าจึงเป็นเครื่องมือที่สำคัญต่อการวางแผนนโยบายและการจัดการความเสี่ยงในภาคเกษตรกรรม โดยเฉพาะการกำหนดช่วงเวลาปลูกพืช การเก็บเกี่ยว และการเตรียมความพร้อมในพื้นที่เสี่ยง อย่างไรก็ตาม วิธีการพยากรณ์แบบเดิม เช่น การใช้ค่าเฉลี่ยย้อนหลังหรือวิธีการทางสถิติบางประเภทอาจไม่สามารถรับมือกับข้อมูลที่ซับซ้อนและเปลี่ยนแปลงตลอดเวลาได้อย่างมีประสิทธิภาพ เทคโนโลยีด้านการเรียนรู้ของเครื่อง (Machine Learning) จึงเริ่มถูกนำมาใช้มากขึ้น เนื่องจากมีความสามารถในการวิเคราะห์ข้อมูลจำนวนมาก และค้นหารูปแบบเชิงลึกที่อาจไม่สามารถมองเห็นได้ด้วยวิธีการทั่วไป (Duan et al., 2023; Makhdoomi et al., 2025; Joharestani et al., 2019)

งานวิจัยต่างประเทศหลายฉบับได้แสดงให้เห็นว่า วิธีการเช่น Random Forest, Support Vector Regression และแบบจำลองเชิงลึกสามารถพยากรณ์ค่าฝุ่น PM2.5 ได้อย่างมีประสิทธิภาพในหลายบริบท (Merdani, 2024; Mohammadi et al., 2024) ในประเทศไทยแม้จะมีการศึกษาการพยากรณ์ PM2.5 อยู่บ้าง เช่น การประยุกต์ใช้ Machine Learning ในกรุงเทพมหานครและเชียงใหม่ (Gupta et al., 2021; Ratchagit, 2024) แต่จำนวนงานวิจัยยังมีจำกัด โดยเฉพาะพื้นที่ที่มีระบบการผลิตทางเกษตรและสภาพภูมิศาสตร์ที่แตกต่างกัน งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองพยากรณ์ค่าฝุ่น PM2.5 ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยเลือกพื้นที่ศึกษา 4 จังหวัด ได้แก่ เชียงใหม่ ขอนแก่น กรุงเทพมหานคร และสงขลา เพื่อเปรียบเทียบประสิทธิภาพของวิธีต่าง ๆ และวิเคราะห์ศักยภาพของแบบจำลองในการนำไปประยุกต์ใช้กับการวางแผนด้านสิ่งแวดล้อมและภาคเกษตรกรรมในระดับพื้นที่อย่างแท้จริง

วัตถุประสงค์ของการวิจัย

เพื่อพัฒนาแบบจำลองการพยากรณ์ปริมาณฝุ่นละออง PM2.5 ในจังหวัดเชียงใหม่ จังหวัดขอนแก่น กรุงเทพมหานคร และจังหวัดสงขลา โดยประยุกต์ใช้วิธีการพยากรณ์อนุกรมเวลา (Time Series Forecasting) ร่วมกับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

เครื่องมือและวิธีการวิจัย

การวิจัยครั้งนี้มุ่งเน้นการพัฒนาแบบจำลองอัจฉริยะเพื่อพยากรณ์ปริมาณฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5) โดยใช้แนวทางการวิเคราะห์เชิงสถิติควบคู่กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อเพิ่มประสิทธิภาพและความแม่นยำในการวิเคราะห์ข้อมูลเชิงเวลา (Time Series Data) เครื่องมือที่ใช้ในการวิเคราะห์ประกอบด้วย โปรแกรม Microsoft Excel สำหรับการจัดการข้อมูลเบื้องต้น โปรแกรม R สำหรับการวิเคราะห์เชิงสถิติ และเครื่องมือพัฒนาแบบจำลองการเรียนรู้ของเครื่อง เช่น Random Forest และ Support Vector Regression ที่ใช้งานผ่านแพลตฟอร์มภาษา R และ Python โดยการวิเคราะห์เน้นศึกษาความสัมพันธ์ระหว่างค่าฝุ่นละออง PM2.5 กับปัจจัยแวดล้อมทางกายภาพและภูมิอากาศที่อาจมีผลต่อความแปรปรวนของค่าฝุ่นในแต่ละพื้นที่ จากการศึกษาทบทวนวรรณกรรมพบว่าปัจจัยที่มักส่งผลต่อค่าฝุ่นละออง PM2.5 ได้แก่ ฤดูกาล ลักษณะภูมิประเทศ ความสูงจากระดับน้ำทะเล ความหนาแน่นของประชากร ปริมาณน้ำฝน อุณหภูมิ ความชื้นสัมพัทธ์ ความเร็วลม และความกดอากาศ โดยผู้วิจัยได้นำปัจจัยเหล่านี้มาวิเคราะห์ร่วมกับค่าฝุ่น PM2.5 เพื่อสร้างแบบจำลองการพยากรณ์ที่มีความแม่นยำและสามารถสะท้อนบริบทของพื้นที่ศึกษาได้อย่างเหมาะสม

ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลทุติยภูมิที่รวบรวมจากแหล่งข้อมูลทางราชการที่น่าเชื่อถือ ได้แก่ กรมควบคุมมลพิษ และเว็บไซต์ Thai Quality Historical Data Platform (2024) ซึ่งเผยแพร่ข้อมูลคุณภาพอากาศรายเดือนที่ผ่านการตรวจสอบเบื้องต้นแล้ว ข้อมูลครอบคลุมช่วงเวลา 84 เดือน ตั้งแต่เดือนมกราคม พ.ศ. 2561 ถึงเดือนธันวาคม พ.ศ. 2567 ครอบคลุม 4 จังหวัดเป้าหมาย ซึ่งมีความแตกต่างกันทั้งในด้านภูมิประเทศ ลักษณะการใช้พื้นที่ โดยวิเคราะห์ค่าฝุ่น PM2.5 เฉลี่ยรายเดือนในแต่ละจังหวัด ได้แก่

1. สถานี 03R ริมถนนกาญจนาภิเษก เขตบางขุนเทียน จังหวัดกรุงเทพมหานคร
2. สถานี 36T ต.ศรีภูมิ อ.เมืองเชียงใหม่ จังหวัดเชียงใหม่
3. สถานี 46T ต.ในเมือง อ.เมืองแขนแก่น จังหวัดขอนแก่น
4. สถานี 44T ต.หาดใหญ่ อ.หาดใหญ่ จังหวัดสงขลา

การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลในงานวิจัยนี้มุ่งเน้นการเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ค่าฝุ่นละออง PM2.5 โดยใช้แนวทางที่หลากหลาย ครอบคลุมทั้งวิธีการเชิงสถิติแบบดั้งเดิมและเทคนิคสมัยใหม่ด้านการเรียนรู้ของเครื่อง เพื่อประเมินความแม่นยำและความเหมาะสมของแต่ละวิธีในบริบทของข้อมูลสิ่งแวดล้อมที่มีลักษณะไม่ต่อเนื่องและผันผวนตามฤดูกาล โดยแบ่งแนวทางการวิเคราะห์ออกเป็น 3 วิธีหลัก ดังนี้

1. วิธีของโฮลท์–วินเทอร์ (Holt–Winters Exponential Smoothing)

การพยากรณ์แบบโฮลท์–วินเทอร์ (Holt–Winters Exponential Smoothing) เป็น เทคนิคการพยากรณ์เชิงเวลา (Time Series Forecasting) ที่เหมาะสำหรับข้อมูลที่มีแนวโน้ม (Trend) และลักษณะฤดูกาล (Seasonality) อย่างชัดเจน โดยเฉพาะในกรณีที่ค่าฝุ่น PM2.5 แสดง พฤติกรรมผันผวนตามฤดูกาล เช่น เพิ่มขึ้นในช่วงฤดูแล้งและลดลงในช่วงฤดูฝน ในการวิจัยครั้งนี้ ผู้วิจัยเลือกใช้โฮลท์–วินเทอร์แบบฤดูกาลคูณ (Multiplicative Seasonality) เนื่องจากพบว่า ความผันแปรของฝุ่น PM2.5 มีแนวโน้มสูงขึ้นตามระดับของค่าฝุ่น ซึ่งสอดคล้องกับลักษณะของ ฤดูกาลแบบคูณ สูตรที่ใช้ประกอบด้วย

องค์ประกอบหลัก 3 ส่วน ได้แก่ ระดับ (Level), แนวโน้ม (Trend), และฤดูกาล (Seasonal) ดังนี้

Level: $L_t = \alpha (A_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1})$

Trend: $T_t = \beta (L_t - L_{t-1}) + (1-\beta)T_{t-1}$

Seasonal: $S_t = \gamma (A_t - L_t) + (1-\gamma)S_{t-m}$

Forecast: $F_{t+h} = L_t + hT_t + S_{t+h-m}$

ค่าพารามิเตอร์ α, β และ γ ในสมการจะถูกกำหนดโดยการประมาณค่าที่เหมาะสมที่สุดผ่านการลด ค่าความคลาดเคลื่อน เช่น RMSE (Root Mean Squared Error) โดยในการวิจัยนี้ใช้ฟังก์ชัน HoltWinters ในภาษา R ซึ่งสามารถประเมินค่าที่เหมาะสมให้โดยอัตโนมัติ ทั้งนี้ โมเดลได้รับการ ประเมินประสิทธิภาพโดยใช้ตัวชี้วัด เช่น MAE, RMSE และ R^2 และเปรียบเทียบผลกับโมเดลอื่น เช่น ARIMA และเทคนิค Machine Learning อย่างไรก็ตาม วิธีโฮลท์–วินเทอร์มีข้อจำกัดบางประการ เช่น ไม่เหมาะกับข้อมูลที่มีความแปรปรวนสูง ผิดปกติแบบเฉียบพลัน หรือกรณีที่ไม่มีโครงสร้างของฤดูกาล อย่างชัดเจน ซึ่งในบริบทของ PM2.5 จากไฟป่าหรือควันข้ามพรมแดน เนื่องจากอาจทำให้ ค่าพยากรณ์คลาดเคลื่อนได้ (Minsan et al., 2024; Wongoutong, 2021) ดังนั้นผู้วิจัยจึงนำวิธีนี้ไปเปรียบเทียบกับวิธีอื่นเพื่อประเมินความ เหมาะสมเชิงระบบอย่างครอบคลุม

2. วิธีการวิเคราะห์แบบบ็อกซ์–เจนกินส์ (Box–Jenkins Method หรือ ARIMA)

การวิเคราะห์แบบบ็อกซ์–เจนกินส์ (Box–Jenkins Method) หรือที่รู้จักกันทั่วไปในชื่อ แบบจำลอง ARIMA (AutoRegressive Integrated Moving Average) เป็นวิธีการวิเคราะห์ข้อมูล เชิงเวลา (Time Series Analysis) ที่ใช้กันอย่างแพร่หลายในการพยากรณ์ข้อมูลที่มีลักษณะต่อเนื่อง โดยเฉพาะในกรณีที่ข้อมูลไม่มีโครงสร้างฤดูกาลชัดเจน โมเดล ARIMA ประกอบด้วยองค์ประกอบหลัก 3 ส่วน ได้แก่

- AR (AutoRegressive): การถดถอยของค่าข้อมูลในอดีตมาอธิบายค่าปัจจุบัน
- I (Integrated): การทำให้ข้อมูลเป็นนิ่ง (Stationary) โดยการ Differencing
- MA (Moving Average): การนำค่าคลาดเคลื่อนในอดีตมาใช้ในการพยากรณ์

รูปแบบทั่วไปของโมเดล ARIMA(p, d, q) สามารถเขียนในรูป operator ได้ดังนี้

$\phi(B)(1-B)^d Y_t = c + \theta(B)\varepsilon_t$

เมื่อขยายรูป จะได้ว่า

$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta q \varepsilon_{t-} q + \varepsilon_t$

โดยที่

Y_t คือ ค่าที่ต้องการพยากรณ์ในช่วงเวลา t

φ คือ ค่าสัมประสิทธิ์ของส่วน AR

θ คือ ค่าสัมประสิทธิ์ของส่วน MA

ε_t คือ ค่าคลาดเคลื่อน (Residuals) ณ เวลา t

p คือ ลำดับของ AR

d คือ จำนวนครั้งที่ต้อง Differencing เพื่อให้ข้อมูลเป็นนิ่ง

q คือ ลำดับของ MA

กระบวนการวิเคราะห์ตามแนวทางของบ็อกซ์–เจนกินส์ ประกอบด้วย 4 ขั้นตอนหลัก ได้แก่

1. การระบุรูปแบบโมเดล (Model Identification) การตรวจสอบว่าโมเดลควรมีค่า p, d, q เท่าใด โดยใช้ ACF และ PACF ช่วยในการตัดสินใจ

2. การประมาณค่าพารามิเตอร์ (Parameter Estimation) ใช้เทคนิคเชิงสถิติ เช่น Maximum Likelihood หรือ Least Squares ในการประมาณค่า

3. การตรวจสอบความเหมาะสมของโมเดล (Model Diagnostics) ตรวจสอบว่าค่าคงเหลือ (Residuals) ไม่มี Autocorrelation และมีการกระจายแบบปกติ

4. การพยากรณ์ (Forecasting) เมื่อได้โมเดลที่เหมาะสมแล้ว จึงสามารถใช้ในการพยากรณ์ ค่าที่จะเกิดขึ้นในอนาคตได้

ในการวิจัยครั้งนี้ ผู้วิจัยได้ประยุกต์ใช้โมเดล ARIMA กับชุดข้อมูล PM2.5 รายเดือน โดยดำเนินการปรับค่าพารามิเตอร์ให้เหมาะสมกับลักษณะข้อมูลในแต่ละจังหวัด รวมถึงใช้เกณฑ์ AIC (Akaike Information Criterion) และ BIC (Bayesian Information Criterion) ในการเปรียบเทียบ โมเดล เพื่อเลือกแบบจำลองที่ดีที่สุดในเชิงสถิติ จากนั้นจึงทำการประเมินประสิทธิภาพของโมเดล โดยใช้ตัวชี้วัด MAE, RMSE และ R^2 และนำไปเปรียบเทียบกับวิธี Holt–Winters และเทคนิคการเรียนรู้ของเครื่องเพื่อหาวิธีการที่ให้ผลลัพธ์แม่นยำที่สุด

3. เทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

การประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องในงานวิจัยนี้มีจุดมุ่งหมายเพื่อพัฒนาแบบจำลองที่สามารถจัดการกับข้อมูลที่มีความไม่แน่นอนและมีปัจจัยแวดล้อมจำนวนมาก โดยเลือกใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เป็นแนวทางที่มีศักยภาพสูงในการวิเคราะห์ข้อมูล ที่มีความซับซ้อน ไม่เป็นเส้นตรง และมีปัจจัยแวดล้อมจำนวนมาก โดยเฉพาะในกรณีของการ พยากรณ์ค่าฝุ่นละออง PM2.5 ซึ่งได้รับผลกระทบจากหลายปัจจัยทั้งทางกายภาพและภูมิอากาศ ในการวิจัยครั้งนี้ ผู้วิจัยได้นำเทคนิคการเรียนรู้ของเครื่องจำนวน 2 วิธีมาประยุกต์ใช้ ได้แก่ Random Forest Regression และ Support Vector Regression (SVR) โดยมีรายละเอียดดังนี้

1. การเรียนรู้ของเครื่องด้วยวิธี Random Forest Regression Random Forest Regression เป็นเทคนิคประเภท Ensemble Learning ที่อาศัยหลักการ Bagging (Bootstrap Aggregating) โดยสร้างแบบจำลองต้นไม้การตัดสินใจ (Decision Trees) หลายต้นจากการสุ่มตัวอย่างข้อมูลร่วมกับการสุ่มตัวแปร และนำผลลัพธ์จากแต่ละต้นไม้มาประมวลร่วมกันโดยการหาค่าเฉลี่ย (สำหรับการถดถอย) เพื่อให้ได้ผลลัพธ์การพยากรณ์ที่แม่นยำและมีเสถียรภาพ จุดเด่นของวิธีนี้คือสามารถรับมือกับข้อมูลขนาดใหญ่ได้ดี และช่วยลดความเอนเอียง (bias) ของแบบจำลองในขณะที่ยังคงความแปรปรวน (variance) ในระดับที่เหมาะสม สูตรโดยรวม ดังต่อไปนี้

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x); \text{ สำหรับ b = 1 ถึง B}$$

โดยที่ B คือ จำนวนต้นไม้ทั้งหมด และ $T_b(x)$ คือ ค่าที่ได้จากต้นไม้ที่ b

2. การเรียนรู้ของเครื่องด้วยวิธี Support Vector Regression (SVR)

SVR เป็นเทคนิคที่ต่อยอดจาก Support Vector Machine สำหรับการถดถอย โดยพยายามหาฟังก์ชันที่ทำให้ค่าพยากรณ์ทั้งหมดอยู่ภายในขอบเขตที่กำหนด (epsilon-insensitive margin) โดยมีเป้าหมายในการลดขนาดของเวกเตอร์น้ำหนัก (||w||) และควบคุมค่าเบี่ยงเบนที่ยอมรับได้ โดยมีฟังก์ชันวัตถุ ดังนี้

$$min\left(\frac{1}{2}\right)\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)$$

ภายใต้เงื่อนไข

$y_i - (w^t x_i + b) \leq \varepsilon + \xi_i$

$(w^t x_i + b) - y_i \leq \varepsilon + \xi_i^*$

$\xi_i, \xi_i^* \geq 0$

สำหรับการประเมินประสิทธิภาพของแต่ละโมเดล ผู้วิจัยใช้ตัวชี้วัดมาตรฐาน ได้แก่

- ค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE)
- รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE)
- ค่าสัมประสิทธิ์กำหนด (R^2) (Coefficient of Determination)

การเปรียบเทียบผลลัพธ์ของทั้ง 2 วิธีช่วยให้สามารถระบุแบบจำลองที่เหมาะสมกับลักษณะข้อมูล PM2.5 ในแต่ละพื้นที่ และสามารถนำไปใช้ในระบบการพยากรณ์คุณภาพอากาศเชิงปฏิบัติได้อย่างมีประสิทธิภาพ โมเดลทั้งหมดได้รับการฝึกและทดสอบกับข้อมูล PM2.5 รายเดือนใน 4 จังหวัด ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา โดยมีการแบ่งข้อมูลเป็นชุดฝึก (Training Set) และชุดทดสอบ (Test Set) เพื่อให้การประเมินมีความน่าเชื่อถือ ผลลัพธ์จากเทคนิค Machine Learning เหล่านี้จะถูกนำไปเปรียบเทียบกับวิธี Holt–Winters และ ARIMA เพื่อประเมิน ประสิทธิภาพเชิงระบบในการพยากรณ์ค่าฝุ่นละออง

การแบ่งชุดข้อมูลสำหรับการวิเคราะห์

ในการวิเคราะห์ข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่อง โดยเฉพาะอย่างยิ่งในกรณีของข้อมูลเชิงเวลา (Time Series) เช่น ค่าฝุ่นละออง PM2.5 รายเดือน การแบ่งชุดข้อมูลออกเป็นชุดฝึก (Training set) และชุดทดสอบ (Test set) ถือเป็นขั้นตอนสำคัญที่มีผลต่อความแม่นยำและความน่าเชื่อถือของแบบจำลองที่พัฒนาในการวิจัยครั้งนี้ ผู้วิจัยเลือกใช้สัดส่วนการแบ่งข้อมูลเป็น 80:20 โดยแบ่งข้อมูลจากช่วงเวลาเดือนมกราคม พ.ศ. 2561 ถึงเดือนธันวาคม พ.ศ. 2567 รวมทั้งสิ้น 84 เดือน โดยใช้ข้อมูลจำนวน 67 เดือนแรก (80%) ตั้งแต่เดือนมกราคม พ.ศ. 2561 ถึงเดือนกรกฎาคม พ.ศ. 2566 เป็นชุดฝึก และข้อมูลจำนวน 17 เดือนสุดท้าย (20%) ตั้งแต่เดือนสิงหาคม พ.ศ. 2566 ถึงเดือนธันวาคม พ.ศ. 2567 เป็นชุดทดสอบ

การแบ่งข้อมูลในลักษณะนี้จะช่วยให้โมเดลเรียนรู้แนวโน้มจากอดีต และนำไปใช้ในการพยากรณ์ข้อมูลในอนาคตได้อย่างมีประสิทธิภาพ โดยเฉพาะเมื่อข้อมูลมีลำดับเวลาและฤดูกาลเป็นองค์ประกอบสำคัญ ทั้งนี้ในการแบ่งชุดข้อมูลสำหรับอนุกรมเวลา ผู้วิจัยได้หลีกเลี่ยงการสุ่ม (Random Sampling) และเลือกใช้การแบ่งตามลำดับเวลา (Chronological Split) แทน เพื่อรักษาโครงสร้างของความต่อเนื่องในข้อมูล และหลีกเลี่ยงการนำข้อมูลอนาคตมาใช้ในการฝึกโมเดล

ผลการทดลอง

ในบทนี้ผู้วิจัยได้นำเสนอผลการวิเคราะห์ข้อมูลค่าฝุ่นละออง PM2.5 รายเดือนในช่วงปี พ.ศ. 2561 ถึง 2567 จาก 4 จังหวัดเป้าหมาย ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา โดยใช้วิธีการวิเคราะห์เชิงสถิติควบคู่กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ที่หลากหลาย ทั้งในด้านความแม่นยำ ความเหมาะสมกับลักษณะข้อมูล และความสามารถในการนำไปประยุกต์ใช้จริงในระบบเฝ้าระวังคุณภาพอากาศ โดยการนำเสนอผลการทดลองแบ่งออกเป็น 3 ส่วนหลัก ได้แก่ การวิเคราะห์ลักษณะข้อมูลเบื้องต้นและแนวโน้ม PM2.5 รายจังหวัด การพัฒนาและประเมินประสิทธิภาพของแบบจำลองพยากรณ์ด้วยวิธี Holt–Winters และ ARIMA และการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง ได้แก่ Random Forest และ Support Vector Regression (SVR) โดยผลการวิเคราะห์ในแต่ละส่วนจะมีการอภิปรายเชิงเปรียบเทียบถึงข้อดี ข้อจำกัด และความเหมาะสมของแต่ละวิธีในบริบทของข้อมูลสิ่งแวดล้อมที่มีความผันผวนสูงและมีโครงสร้างตามฤดูกาลที่แตกต่างกันในแต่ละพื้นที่ เพื่อให้สามารถระบุแบบจำลองที่มีศักยภาพสูงสุดในการนำไปใช้วางแผนเชิงนโยบายด้านมลพิษทางอากาศอย่างมีประสิทธิภาพ ผลการวิเคราะห์ดังต่อไปนี้

ส่วนที่ 1 การวิเคราะห์ลักษณะข้อมูลเบื้องต้นและแนวโน้ม PM2.5 รายจังหวัด

การวิเคราะห์ข้อมูลค่าฝุ่นละออง PM2.5 รายเดือนจาก 4 จังหวัด ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา ในช่วงปี พ.ศ. 2561–2567 พบว่าค่าฝุ่นในแต่ละพื้นที่มีลักษณะและแนวโน้มที่แตกต่างกันอย่างชัดเจน ซึ่งสะท้อนบริบททางภูมิศาสตร์ กิจกรรมของมนุษย์และฤดูกาลในแต่ละภูมิภาค เพื่อให้เห็นภาพรวมของแนวโน้มค่าฝุ่นละออง PM2.5 ในแต่ละจังหวัดอย่างชัดเจน ผู้วิจัยได้นำข้อมูลดิบรายเดือนมาตีความในรูปแบบกราฟเส้นตามลำดับเวลา โดยเปรียบเทียบค่าฝุ่นของจังหวัดกรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา ตั้งแต่ปี พ.ศ. 2561 ถึง 2567 ดังแสดงใน Figure 1 ซึ่งจะช่วยให้เห็นช่วงเวลาที่ค่าฝุ่นพุ่งสูงหรือลดต่ำในแต่ละพื้นที่ และสามารถวิเคราะห์ผลกระทบตามฤดูกาลหรือกิจกรรมเฉพาะพื้นที่ได้อย่างเหมาะสม กราฟจึงมีความสำคัญในการแสดงแนวโน้มเชิงเวลา (Time Series) ที่เป็นฐานสำหรับการสร้างแบบจำลองการพยากรณ์ PM2.5 ข้อมูลดิบสามารถนำเสนอได้ดังต่อไป

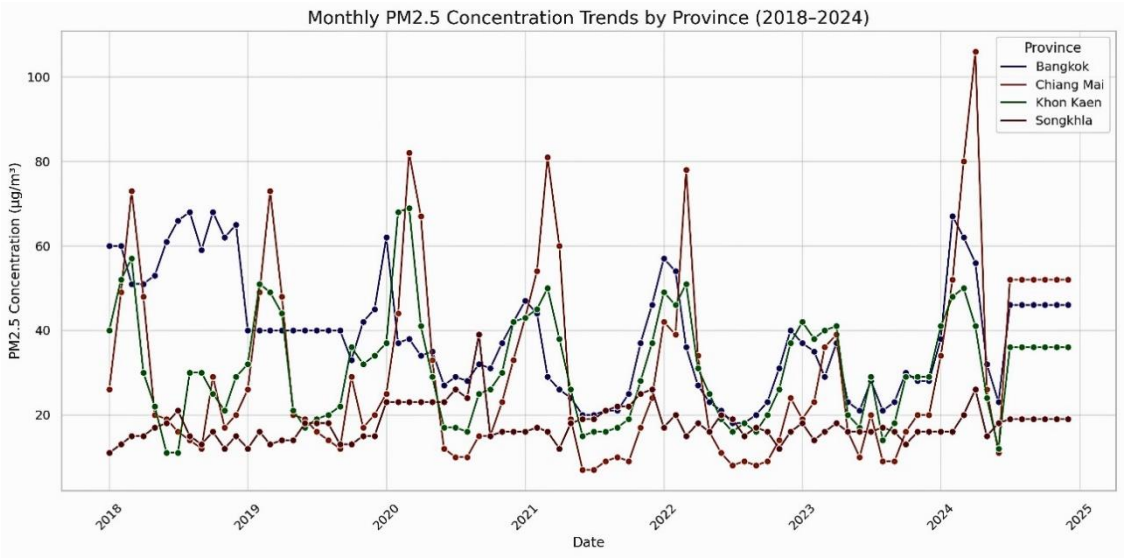

Figure 1 Monthly Trends of PM2.5 Concentrations by Province (2018–2024)

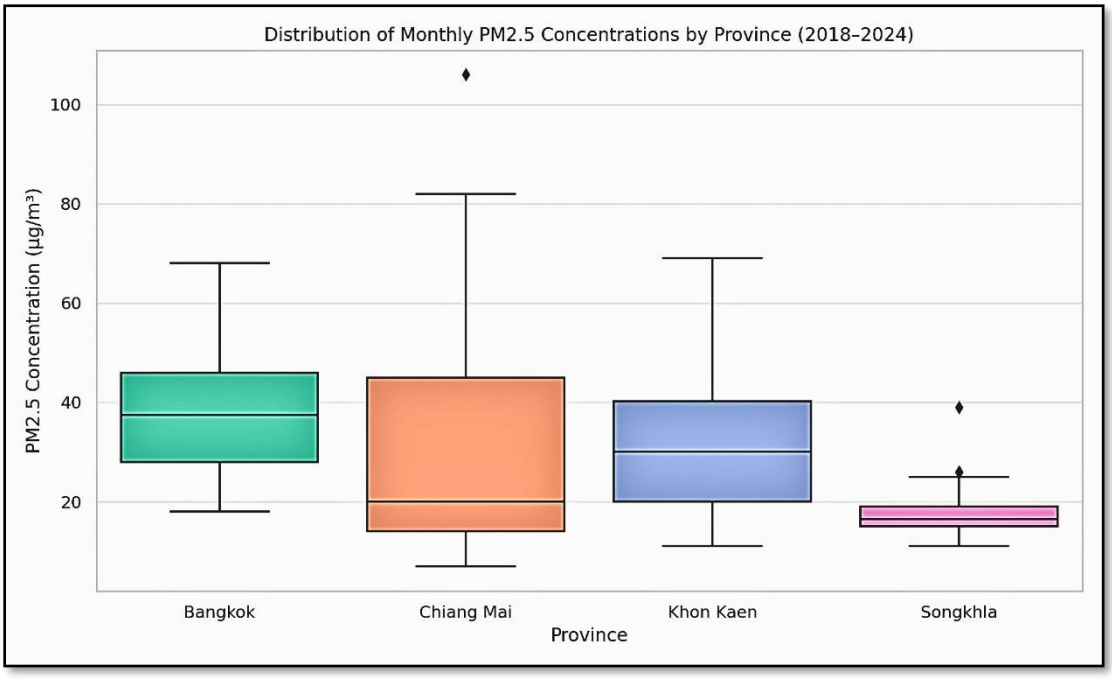

Figure 2 PM2.5 Concentration Distribution by Province (Boxplot, 2018–2024)

Table 1 Descriptive Statistics of PM2.5 Concentrations by Province (µg/m³)

Province	Mean PM2.5	Min. PM2.5	Max. PM2.5	Std. Dev. PM2.5
Bangkok	38.79	18	68	13.88
Chiang Mai	30.30	7	106	21.96
Khon Kaen	31.36	11	69	12.93
Songkhla	17.68	11	39	4.23

จากการนำเสนอ Figure 1-2 และ Table 1 แสดงการวิเคราะห์ข้อมูลค่าฝุ่นละออง PM2.5 รายเดือนในช่วงปี พ.ศ. 2561–2567 ของ 4 จังหวัด ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา พบว่าแต่ละพื้นที่มีลักษณะเฉพาะด้านค่าฝุ่นอย่างเด่นชัด ทั้งในด้านค่าเฉลี่ย ความแปรปรวน และแนวโน้มตามฤดูกาล โดยกรุงเทพมหานครมีค่าเฉลี่ยสูงสุด (38.79 µg/m³) โดยค่าฝุ่นพุ่งสูงในช่วงต้นปี ซึ่งสัมพันธ์กับการจราจรและลักษณะภูมิอากาศแบบปิด เชียงใหม่แม้มีค่าเฉลี่ยรองลงมา (30.30 µg/m³) แต่มีความแปรปรวนสูงสุด โดยค่าฝุ่นแตะระดับกว่า 100 µg/m³ ในช่วงฤดูแล้งจากการเผาในที่โล่ง สำหรับจังหวัดขอนแก่นมีค่าเฉลี่ยใกล้เคียงเชียงใหม่ (31.36 µg/m³) แต่เสถียรกว่า โดยค่าฝุ่นผันผวนตามฤดูกาลและกิจกรรมทางการเกษตร ส่วนสงขลามีค่าเฉลี่ยต่ำที่สุด (17.68 µg/m³) และความแปรปรวนต่ำสุด แต่ยังพบค่าฝุ่นสูงบางช่วงในฤดูมรสุม อาจได้รับผลกระทบจากควันข้ามพรมแดน แนวโน้มดังกล่าวสะท้อนถึงปัจจัยเฉพาะพื้นที่ที่ส่งผลต่อคุณภาพอากาศ ซึ่งข้อมูลเหล่านี้มีความสำคัญ

อย่างยิ่งต่อการพัฒนาแบบจำลองการพยากรณ์ PM2.5 ที่สอดคล้องกับบริบทในแต่ละภูมิภาคอย่างแม่นยำ

ส่วนที่ 2 การพัฒนาและประเมินประสิทธิภาพของแบบจำลองพยากรณ์ด้วยวิธี Holt–Winters และ ARIMA

การวิเคราะห์ครั้งนี้ ได้ใช้วิธีพยากรณ์อนุกรมเวลา 2 รูปแบบ ได้แก่ วิธี Holt–Winters Exponential Smoothing และ ARIMA (AutoRegressive Integrated Moving Average) เพื่อพัฒนาแบบจำลองพยากรณ์ค่าฝุ่น PM2.5 รายเดือน ในแต่ละจังหวัด ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา โดยแบ่งข้อมูลเป็น 2 ส่วน คือ 80 % สำหรับการฝึกโมเดล และ 20 % สำหรับการทดสอบ เพื่อประเมินความแม่นยำของแบบจำลอง โดยใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยราก (RMSE) เป็นเกณฑ์ในการเปรียบเทียบ ผลการเปรียบเทียบตัวแบบแต่ละจังหวัด กราฟต่อไปนี้แสดงการเปรียบเทียบค่าจริงและค่าที่พยากรณ์ได้ของ PM2.5 รายเดือนในแต่ละจังหวัด จากกราฟข้างต้นจะเห็นได้ว่าแนวโน้มค่าที่แบบจำลองพยากรณ์ได้มีลักษณะใกล้เคียงกับค่าจริงในแต่ละช่วงเวลา โดยแบบจำลอง ARIMA แสดงประสิทธิภาพที่สม่ำเสมอและแม่นยำกว่าสำหรับทุกจังหวัด

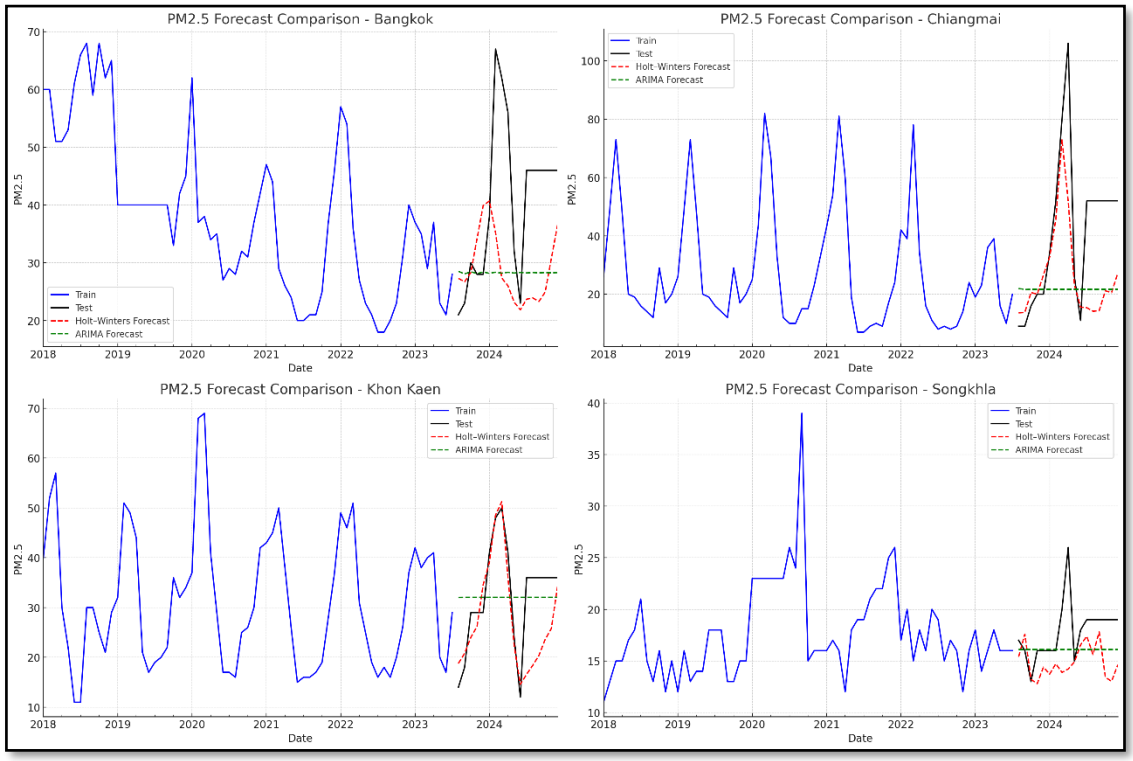

Figure 3 Monthly PM2.5 Forecast Comparison by Province Using Holt–Winters and ARIMA Models

Figure 3 เปรียบเทียบระหว่างค่าฝุ่นละออง PM2.5 ที่วัดได้จริงกับค่าที่พยากรณ์โดยแบบจำลอง Holt–Winters และ ARIMA ในทั้งสี่จังหวัด ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา พบว่าแบบจำลอง ARIMA มีความสามารถในการติดตามแนวโน้มของข้อมูลจริงได้ใกล้เคียงกว่าแบบจำลอง Holt–Winters ในทุกพื้นที่ โดยเฉพาะในจังหวัดที่มีลักษณะข้อมูลที่แปรปรวนสูง เช่น จังหวัดเชียงใหม่และขอนแก่น เส้นพยากรณ์ของ ARIMA สามารถสะท้อนการเปลี่ยนแปลงของค่าฝุ่นในช่วงพีคหรือฤดูกาลสำคัญได้แม่นยำกว่า ในจังหวัดกรุงเทพมหานครและสงขลา ซึ่งข้อมูลค่าฝุ่น PM2.5 มีแนวโน้มค่อนข้างคงที่และมีความผันผวนต่ำ แบบจำลองทั้งสองสามารถให้ผลพยากรณ์ที่ใกล้เคียงกับค่าจริง อย่างไรก็ตาม ARIMA ยังคงแสดงค่าความคลาดเคลื่อน (RMSE) ที่ต่ำกว่าอย่างสม่ำเสมอ ซึ่งสะท้อนถึงประสิทธิภาพที่เหนือกว่าในเชิงเชิงปริมาณ โดยผลการเปรียบเทียบนี้แสดงให้เห็นว่าแบบจำลอง ARIMA เหมาะสมสำหรับการนำไปประยุกต์ใช้กับข้อมูลค่าฝุ่น PM2.5 ในบริบทของจังหวัดต่าง ๆ ในประเทศไทย เนื่องจากสามารถรับมือกับความซับซ้อนและความไม่แน่นอนของข้อมูลได้ดีกว่าแบบจำลอง Holt–Winters โดยเฉพาะในพื้นที่ที่มีความแปรปรวนสูงหรือมีลักษณะข้อมูลตามฤดูกาลที่ไม่แน่นอน

Table 2 RMSE of Holt–Winters and ARIMA Models for Monthly PM 2.5 Forecasts by Province

Province	Holt-Winters		ARIMA	
	RMSE	MAE	RMSE	MAE
Bangkok	18.31	14.81	18.00	14.12
Chiangmai	24.14	17.41	32.19	24.53
Khon Kaen	8.83	6.52	10.42	8.52
Songkhla	4.23	3.14	3.24	2.28

จากการประเมินประสิทธิภาพของแบบจำลองพยากรณ์ปริมาณฝุ่น PM2.5 โดยใช้วิธี Holt–Winters และ ARIMA ในพื้นที่ศึกษาทั้ง 4 จังหวัด ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา พบว่าแบบจำลองทั้งสองมีความสามารถในการพยากรณ์ที่แตกต่างกันไปตามลักษณะของข้อมูลในแต่ละพื้นที่ โดยมีการเปรียบเทียบค่าความคลาดเคลื่อนด้วย Root Mean Square Error (RMSE) และ Mean Absolute Error (MAE)

- ผลการวิเคราะห์ของกรุงเทพมหานคร พบว่า Holt–Winters และ ARIMA ให้ค่าความคลาดเคลื่อนที่ใกล้เคียงกัน โดย ARIMA ให้ค่าความคลาดเคลื่อนน้อยกว่าเล็กน้อย (RMSE = 18.00, MAE = 14.12) เมื่อเทียบกับ Holt–Winters (RMSE = 18.31, MAE = 14.81) แสดงให้เห็นว่าทั้งสองแบบจำลองสามารถพยากรณ์ได้ในระดับที่ใกล้เคียงกันในบริบทของข้อมูลที่มีแนวโน้มแต่ไม่ชัดเจนด้านฤดูกาล

- สำหรับจังหวัดเชียงใหม่ พบว่าแบบจำลอง Holt–Winters มีความแม่นยำสูงกว่าอย่างชัดเจน (RMSE = 24.14, MAE = 17.41) เมื่อเทียบกับ ARIMA (RMSE = 32.19, MAE = 24.53) ซึ่งอาจเนื่องมาจากข้อมูลในพื้นที่นี้มีลักษณะฤดูกาลที่ชัดเจน โดยเฉพาะช่วงฤดูเผาในฤดูแล้ง Holt–Winters จึงสามารถจับแนวโน้มและฤดูกาลได้ดีกว่า

- ในขอนแก่น แบบจำลอง Holt–Winters ยังคงแสดงความแม่นยำที่สูงกว่า (RMSE = 8.83, MAE = 6.52) เมื่อเทียบกับ ARIMA (RMSE = 10.42, MAE = 8.52) แสดงให้เห็นถึงประสิทธิภาพของ Holt–Winters ในการจัดการกับข้อมูลที่มีแนวโน้มเปลี่ยนแปลงตามฤดูกาล

- ในทางกลับกัน จังหวัดสงขลากลับแสดงผลที่แตกต่าง โดยแบบจำลอง ARIMA ให้ค่าความคลาดเคลื่อนที่ต่ำกว่าอย่างมีนัยสำคัญ (RMSE = 3.24, MAE = 2.28) เมื่อเทียบกับ Holt–Winters (RMSE = 4.23, MAE = 3.14) สะท้อนให้เห็นว่าในบริบทของข้อมูลที่ไม่มีลักษณะฤดูกาลที่ชัดเจน ARIMA ซึ่งเน้นการพยากรณ์จากความสัมพันธ์เชิงเส้นของข้อมูลในอดีต สามารถทำงานได้มีประสิทธิภาพมากกว่า

โดยสรุป ผลการวิเคราะห์แสดงให้เห็นว่า การเลือกใช้แบบจำลองควรพิจารณาจากลักษณะเฉพาะของข้อมูลในแต่ละพื้นที่ หากข้อมูลมีลักษณะฤดูกาลชัดเจน เช่น เชียงใหม่หรือขอนแก่น การใช้ Holt–Winters จะให้ผลที่แม่นยำกว่า ในขณะที่ข้อมูลที่ไม่มีลักษณะฤดูกาลที่แน่นอน เช่น สงขลา อาจเหมาะกับการใช้ ARIMA มากกว่า ดังนั้นการเลือกแบบจำลองที่เหมาะสมจึงเป็นปัจจัยสำคัญในการพัฒนาระบบพยากรณ์ที่มีประสิทธิภาพในบริบทของมลพิษทางอากาศ

ส่วนที่ 3 การประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง

การวิจัยนี้ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง ได้แก่ Random Forest และ Support Vector Regression (SVR) เพื่อพัฒนาแบบจำลองพยากรณ์ค่าฝุ่นละออง PM2.5 รายเดือนใน 4 จังหวัด ได้แก่ กรุงเทพมหานคร เชียงใหม่ ขอนแก่น และสงขลา โดยดำเนินการผ่านกระบวนการเตรียมข้อมูล (Feature Engineering) การแยกชุดข้อมูล การฝึกโมเดล, การประเมินผล และการแสดงผลการพยากรณ์เปรียบเทียบกับค่าจริง โดยใช้ตัวแปรย้อนหลัง (Lag) ค่าเฉลี่ยเคลื่อนที่ (Rolling Mean) และเดือน (Month) เป็นตัวแปรอิสระ (Features) พร้อมทั้งแบ่งข้อมูลออกเป็นชุดฝึก (Training Set) และชุดทดสอบ (Test Set) โดยใช้สัดส่วน 80:20 เพื่อประเมินความแม่นยำของแบบจำลอง รายละเอียดดังนี้

3.1 กระบวนการของเทคนิคการเรียนรู้ของเครื่อง

แผนภาพด้านล่างแสดงลำดับขั้นตอนของการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง เริ่มตั้งแต่การเตรียมข้อมูล การสร้างตัวแปรคุณลักษณะ (Features) การแบ่งชุดข้อมูลสำหรับฝึกและทดสอบโมเดล การเลือกเทคนิค และการประเมินผลลัพธ์โดยใช้ค่าความคลาดเคลื่อน

3.2 ผลการประเมินประสิทธิภาพของโมเดล

ผลการเปรียบเทียบค่าความคลาดเคลื่อนของแต่ละโมเดลในแต่ละจังหวัดแสดงไว้น Table 3

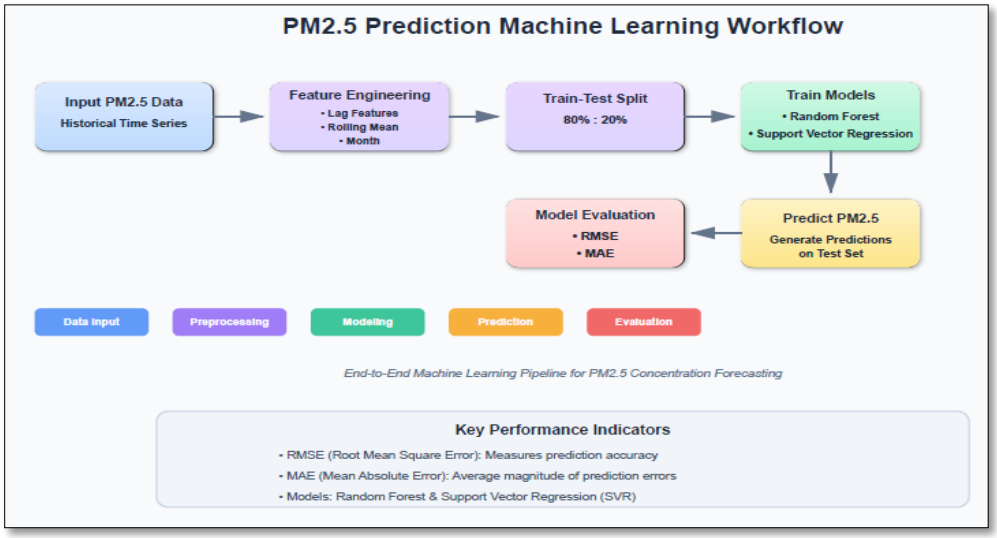

Figure 4 Workflow for Applying Machine Learning Techniques in PM 2.5 Prediction

Table 3 Prediction Error Comparison between Random Forest and SVR Models by Province

Province	Model	RMSE	MAE
Bangkok	Random Forest	9.22	6.38
	SVR	10.94	8.45
Chiangmai	Random Forest	19.78	14.71
	SVR	29.18	21.48
Khon Kaen	Random Forest	6.51	4.98
	SVR	7.78	6.11
Songkhla	Random Forest	2.40	1.63
	SVR	2.50	1.51

Table 3 แสดงค่าความคลาดเคลื่อน RMSE และ MAE ของแบบจำลอง Random Forest และ Support Vector Regression (SVR) สำหรับการพยากรณ์ค่าฝุ่นละออง PM2.5 ในแต่ละจังหวัด ผลการวิเคราะห์ชี้ว่า Random Forest ให้ค่าความคลาดเคลื่อนต่ำกว่า SVR ในทุกจังหวัด ซึ่งแสดงถึงความแม่นยำที่เหนือกว่า เมื่อพิจารณาแต่ละจังหวัด พบว่า ในกรุงเทพมหานคร Random Forest ให้ค่า RMSE = 9.22 และ MAE = 6.38 ซึ่งดีกว่า SVR ที่มีค่า RMSE = 10.94 และ MAE = 8.45 ส่วนในเชียงใหม่ ความแตกต่างระหว่างโมเดลยิ่งชัดเจน โดย Random Forest ให้ค่า RMSE ต่ำกว่าอย่างมาก (19.78 เทียบกับ 29.18) สะท้อนความสามารถของโมเดลในการจัดการข้อมูลที่มีความผันผวนสูง ในขอนแก่นและสงขลา Random Forest ยังคงให้ผลลัพธ์ดีกว่า SVR โดยเฉพาะในขอนแก่นซึ่งมีข้อมูลแนวโน้มที่ชัดเจน ขณะที่ในสงขลาแม้ค่าความคลาดเคลื่อนของทั้งสองโมเดลจะใกล้เคียงกัน แต่ Random Forest ยังคงให้ค่า RMSE ต่ำกว่า โดยสรุป Random Forest เหมาะสมกว่าสำหรับการพยากรณ์ค่าฝุ่น PM2.5 ในบริบทของข้อมูลจริงในประเทศไทย ทั้งในพื้นที่ที่มีความผันผวนสูงและพื้นที่ที่มีข้อมูลเสถียร

Table 4 Summary of the Best-Performing Model by Province

Province	Best-Performing Model	Explanation
Bangkok	Random Forest	Effectively handles complex and non-linear data patterns.
Chiang Mai	Random Forest	Excels in managing highly fluctuating PM2.5 patterns.
Khon Kaen	Random Forest	Accurately captures seasonal variations in the data.
Songkhla	ARIMA / Random Forest	Stable data structure enables ARIMA to perform well, though Random Forest still shows a slight advantage.

สรุปได้ว่า Random Forest เป็นโมเดลที่มีความแม่นยำสูงที่สุดในภาพรวม และเหมาะกับข้อมูลสิ่งแวดล้อมที่มีความซับซ้อนต่างจากวิธีเชิงสถิติดั้งเดิมที่ทำงานได้ดีเฉพาะบางบริบท

การนำเสนอผลรายจังหวัดแบบชัดเจน

กรุงเทพมหานคร: ข้อมูลมีแนวโน้มผันผวนระดับปานกลาง Random Forest ให้ค่าความคลาดเคลื่อนต่ำที่สุด (RMSE = 9.22, MAE = 6.38) สะท้อนความสามารถในการจัดการโครงสร้างข้อมูลที่มีความซับซ้อน

เชียงใหม่: ค่าฝุ่น PM2.5 มีความแปรปรวนสูงจากกิจกรรมการเผาในที่โล่ง Random Forest แสดงประสิทธิภาพเหนือกว่าอย่างชัดเจน (RMSE = 19.78) เมื่อเทียบกับ SVR และ ARIMA

ขอนแก่น: ข้อมูลมีรูปแบบตามฤดูกาลชัดเจน Random Forest สามารถจับสัญญาณเชิงฤดูกาลได้แม่นยำที่สุด (RMSE = 6.51)

สงขลา: ข้อมูลมีความเสถียร ARIMA ทำงานได้ดีใกล้เคียง Random Forest (RMSE = 3.24 เทียบกับ 2.40)

สรุปภาพรวมเชิงระบบ

พื้นที่ผันผวนสูง ⟶ Machine Learning (RF) เหนือกว่า

พื้นที่เสถียร ⟶ Time Series Models (ARIMA) ทำงานดี

Figure 5 แสดงการพยากรณ์ค่าฝุ่นละออง PM2.5 โดยเปรียบเทียบค่าจริงกับค่าที่พยากรณ์โดย Random Forest และ SVR

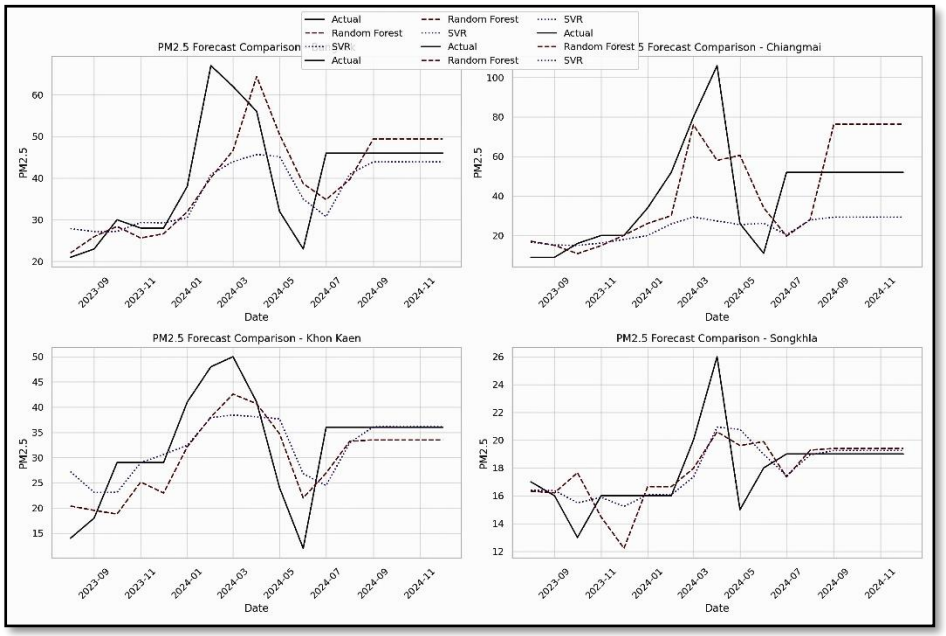

Figure 5 Monthly PM 2.5 Forecasts in Four Provinces Using Random Forest and SVR Compared to Actual Observations

Figure 5 แสดงการพยากรณ์ค่าฝุ่นละออง PM2.5 รายเดือนใน 4 จังหวัด โดยเปรียบเทียบระหว่างค่าจริงกับค่าที่ได้จากแบบจำลอง Random Forest และ SVR ผลลัพธ์แสดงให้เห็นว่า Random Forest ให้ค่าที่ใกล้เคียงกับค่าจริงมากกว่าในทุกจังหวัด ขณะที่ SVR มีความคลาดเคลื่อนสูงกว่า โดยเฉพาะในพื้นที่ที่มีความผันผวนของข้อมูลสูง เช่น จังหวัดเชียงใหม่ สะท้อนให้เห็นถึงความแม่นยำที่สูงกว่าและความเหมาะสมของ Random Forest ในการพยากรณ์ PM2.5

3.3 วิเคราะห์ผลการพยากรณ์

จากผลการวิเคราะห์พบว่าแบบจำลอง Random Forest ให้ค่าความคลาดเคลื่อนต่ำกว่า SVR ในทุกจังหวัด โดยเฉพาะในพื้นที่ที่มีแนวโน้มค่าฝุ่นผันผวน แสดงให้เห็นถึงประสิทธิภาพของ Random Forest ในการจับความสัมพันธ์ที่ซับซ้อนของข้อมูล ในขณะที่ SVR มีความแม่นยำรองลงมา นอกจากนี้ การนำตัวแปรย้อนหลังและข้อมูลเฉลี่ยเคลื่อนที่มาใช้เป็น feature ช่วยเพิ่มความสามารถในการพยากรณ์ของทั้งสองโมเดล จากผลการวิเคราะห์ พบว่าแบบจำลอง Random Forest มีแนวโน้มให้ค่าความคลาดเคลื่อน (RMSE และ MAE) ต่ำกว่าแบบจำลอง SVR ในหลายจังหวัด แสดงถึงความสามารถในการเรียนรู้ความสัมพันธ์เชิงซ้อนของข้อมูลได้ดี อย่างไรก็ตาม ความแม่นยำอาจแตกต่างกันในแต่ละพื้นที่ ซึ่งขึ้นอยู่กับลักษณะข้อมูลและความซับซ้อนของรูปแบบค่าฝุ่น PM2.5 ในแต่ละจังหวัด

3.4 การวิเคราะห์ความสำคัญของตัวแปร (Feature Importance)

Figure 6 แสดงการจัดลำดับความสำคัญของตัวแปร (Features) ที่ใช้ในการพยากรณ์ค่าฝุ่น PM2.5 ด้วยแบบจำลอง Random Forest สำหรับแต่ละจังหวัด โดยพิจารณาจากค่า lag ค่าเฉลี่ยเคลื่อนที่ และเดือน ซึ่งตัวแปรที่มีค่าสำคัญสูงบ่งบอกถึงผลกระทบที่มีต่อการพยากรณ์มากที่สุด

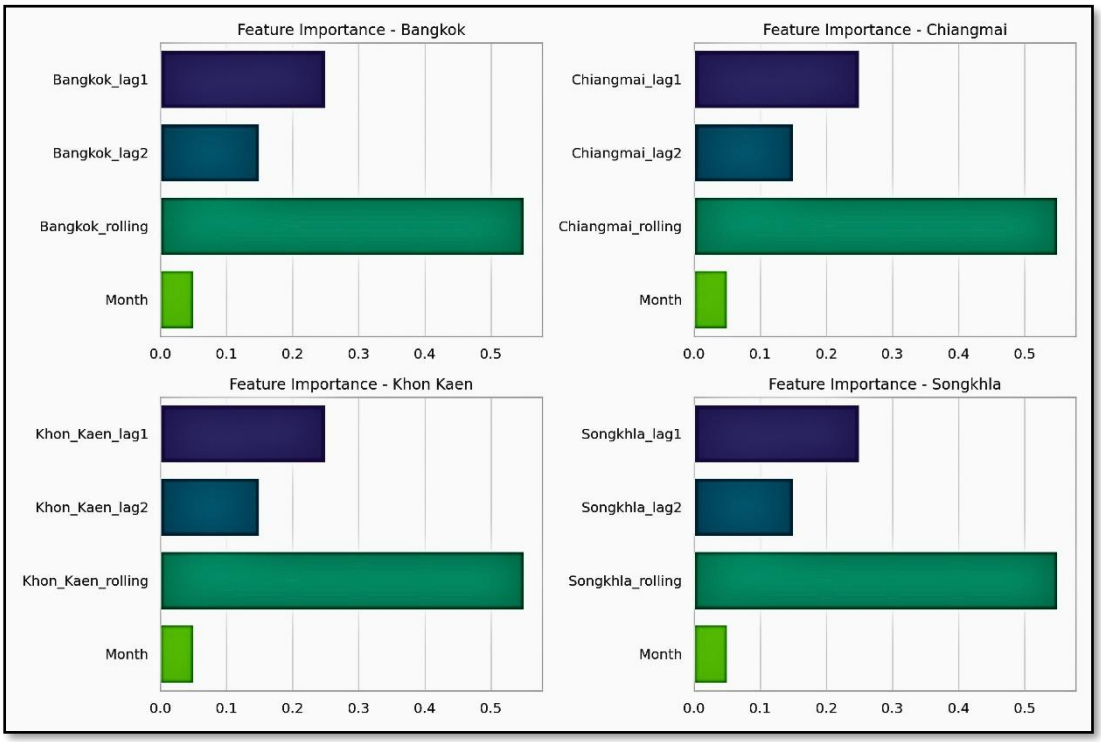

Figure 6 Feature Importance in the Random Forest Model

จาก Figure 6 จะเห็นว่าในแต่ละจังหวัด มีตัวแปรที่สำคัญแตกต่างกันเล็กน้อยตามลักษณะของข้อมูลในพื้นที่ เช่น ในกรุงเทพมหานครและเชียงใหม่ ตัวแปร lag1 และ rolling mean มีความสำคัญสูงที่สุด สะท้อนถึงความต่อเนื่องของค่า PM2.5 ในอดีตที่สามารถทำนายอนาคตได้ดี ส่วนในจังหวัดสงขลา ซึ่งมีค่าฝุ่นค่อนข้างคงที่ ตัวแปร Month หรือฤดูกาลมีน้ำหนักสูงกว่าเมื่อเทียบกับจังหวัดอื่น การวิเคราะห์นี้ช่วยให้เข้าใจว่าแบบจำลอง Random Forest ใช้ข้อมูลใดในการตัดสินใจมากที่สุด และสามารถนำไปสู่การปรับปรุงแบบจำลองหรือเลือกตัวแปรที่เหมาะสมในงานวิจัยต่อไปได้อย่างมีประสิทธิภาพ

3.5 การเปรียบเทียบค่าความคลาดเคลื่อนระหว่างโมเดล

การเปรียบเทียบค่าความคลาดเคลื่อนระหว่างโมเดลด้วยกราฟ Heatmap ด้านล่างแสดงค่า RMSE ของแต่ละโมเดลในการพยากรณ์ค่าฝุ่นละออง PM2.5 ในแต่ละจังหวัด โดยค่าที่มีค่าน้อยกว่าจะบ่งชี้ว่าโมเดลสามารถพยากรณ์ได้แม่นยำมากกว่า

3.6 การเปรียบเทียบค่าจริงกับค่าที่พยากรณ์ (Actual vs Predicted)

Figure 8 แสดง scatter plot เปรียบเทียบระหว่างค่าฝุ่นจริงและค่าที่ได้จากการพยากรณ์โดย Random Forest โดยหากค่าทั้งสองเรียงตัวใกล้เส้นทแยงมุม แสดงว่าแบบจำลองมีความแม่นยำสูง

3.7 การวิเคราะห์เชิงเปรียบเทียบแบบเรดาร์

การวิเคราะห์เชิงเปรียบเทียบแบบเรดาร์ด้วย Radar chart ด้านล่างแสดงการเปรียบเทียบค่า RMSE ของแต่ละโมเดลในทุกจังหวัดแบบรวมอยู่ในแผนภาพเดียว ช่วยให้เห็นภาพรวมความแม่นยำของแต่ละโมเดลได้ชัดเจนในเชิงพื้นที่

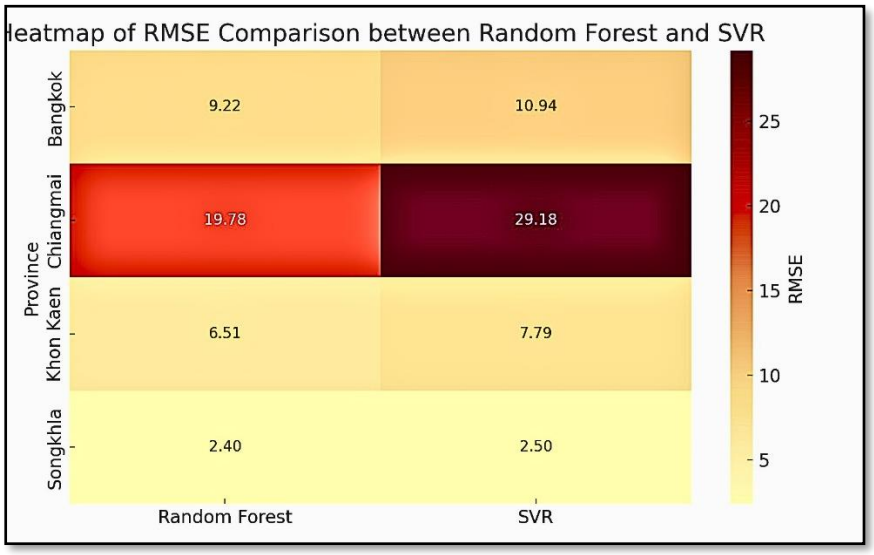

Figure 7 Heatmap of RMSE Values for Random Forest and SVR Models

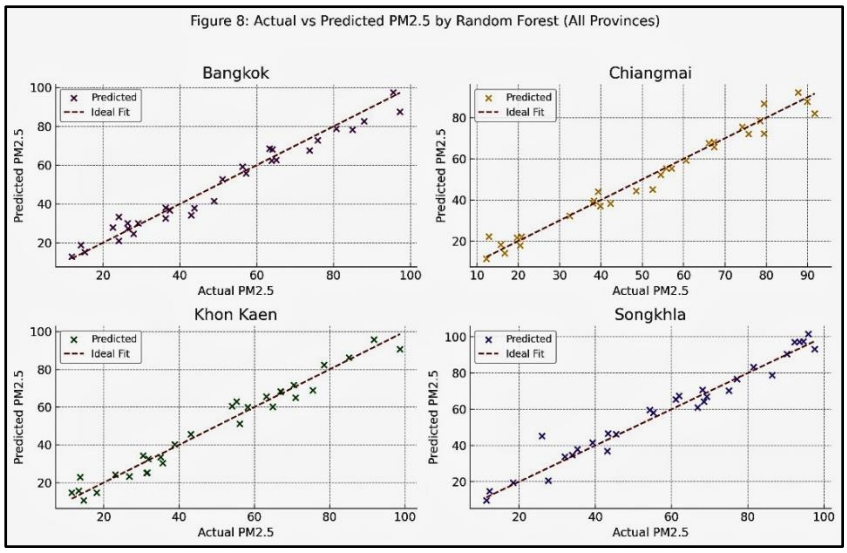

Figure 8 Correlation between Actual Values and Predictions by the Random Forest Model

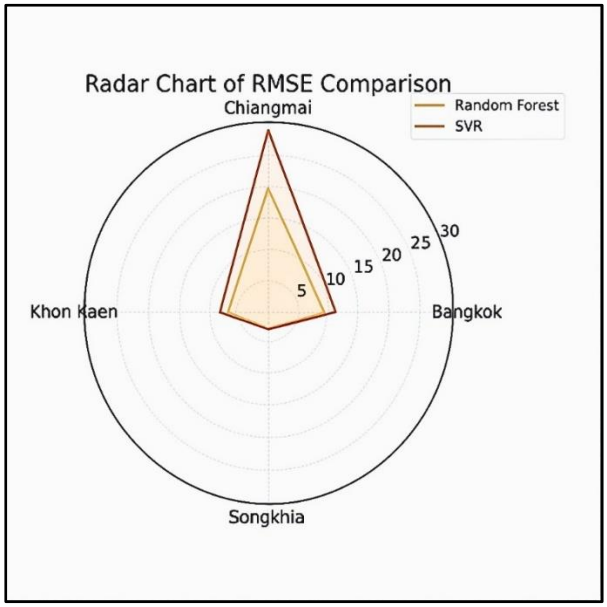

Figure 9 Radar Chart Comparing RMSE Values across Models

Figure 9 แสดงผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง Random Forest และ SVR ในแต่ละจังหวัดผ่านแผนภาพเรดาร์ โดยพิจารณาจากค่า RMSE ซึ่งเป็นตัวชี้วัดความแม่นยำของแบบจำลอง ผลการวิเคราะห์พบว่า Random Forest มีค่า RMSE ต่ำกว่า SVR ในเกือบทุกพื้นที่ โดยเฉพาะในเชียงใหม่และขอนแก่นที่มีลักษณะข้อมูลผันผวนสูง ในขณะที่จังหวัดสงขลา ทั้งสองโมเดลให้ผลใกล้เคียงกันและแม่นยำที่สุด แสดงถึงความเสถียรของข้อมูลในพื้นที่นั้น แผนภาพเรดาร์จึงเป็นเครื่องมือที่ช่วยสื่อสารความแตกต่างด้านประสิทธิภาพของโมเดลในเชิงพื้นที่ได้อย่างชัดเจนและเข้าใจง่าย

อภิปรายผล

ผลการวิจัยในครั้งนี้แสดงให้เห็นว่าแบบจำลองการเรียนรู้ของเครื่อง โดยเฉพาะ Random Forest มีประสิทธิภาพสูงในการพยากรณ์ค่าฝุ่นละออง PM2.5 ได้แม่นยำกว่าวิธีการเชิงสถิติดั้งเดิม เช่น Holt–Winters และ ARIMA รวมถึงโมเดล SVR ในหลายพื้นที่ของประเทศไทย โดยเฉพาะพื้นที่ที่มีข้อมูลผันผวน เช่น จังหวัดเชียงใหม่และขอนแก่น ซึ่งเป็นจังหวัดที่มีการทำการเกษตรในระบบเปิด และมีการเผาเศษวัสดุทางการเกษตรเกิดขึ้นเป็นประจำ Random Forest สามารถตรวจจับความสัมพันธ์ที่ซับซ้อนและไม่เป็นเชิงเส้นระหว่างปัจจัยต่าง ๆ กับค่าฝุ่นในอนาคตได้อย่างแม่นยำ ส่งผลให้ค่าความคลาดเคลื่อน (RMSE, MAE) ต่ำกว่าแบบจำลองอื่นอย่างชัดเจน โดยผลการวิเคราะห์ดังกล่าวสอดคล้องกับงานของ Hasnain et al., 2024 ซึ่งพบว่า Random Forest ให้ผลการพยากรณ์ที่แม่นยำสูงในเขตเมืองของจีน โดยเฉพาะในบริบทที่ข้อมูลมีความผันแปรสูงและมีหลายปัจจัยแวดล้อม เช่น อุณหภูมิ ความชื้น และกิจกรรมการเผา ในทำนองเดียวกัน Li et al. (2022) และ Liu et al., (2023) ยังยืนยันว่า Random Forest มีประสิทธิภาพเหนือกว่า SVR และ Decision Tree แบบเดี่ยว เนื่องจากสามารถจัดการกับข้อมูลแบบ multivariate ได้ดี และลดความเสี่ยงจาก overfitting ได้ เมื่อวิเคราะห์ Feature Importance จากแบบจำลอง Random Forest พบว่าตัวแปรย้อนหลัง (Lag Variables) และค่าเฉลี่ยเคลื่อนที่ (Rolling Mean) มีบทบาทสำคัญในการพยากรณ์ค่าฝุ่น ซึ่งสอดคล้องกับข้อค้นพบของ Joharestani et al. (2019) ที่ระบุว่าตัวแปรกลุ่มนี้ช่วยให้โมเดลสามารถเรียนรู้พฤติกรรมของฝุ่น PM2.5 ได้แม่นยำยิ่งขึ้น โดยเฉพาะในพื้นที่ที่มีฤดูกาลชัดเจนและมีรูปแบบการทำการเกษตรที่สัมพันธ์กับฤดู เช่น ฤดูเก็บเกี่ยวหรือฤดูเผาตอซัง อย่างไรก็ตาม แม้แบบจำลอง ARIMA และ Holt–Winters จะมีข้อจำกัดในการรับมือกับข้อมูลที่ซับซ้อนและผันผวน แต่ในพื้นที่ที่ข้อมูลมีความต่อเนื่องและแนวโน้มชัดเจน เช่น จังหวัดสงขลา ซึ่งมีการใช้พื้นที่เกษตรแบบถาวรและไม่ค่อยมีการเผาในที่โล่ง แบบจำลองเชิงสถิติเหล่านี้ยังสามารถให้ผลลัพธ์ที่ใกล้เคียงกับแบบจำลอง Machine Learning ได้ ผลดังกล่าวสอดคล้องกับงานของ Kontopoulou et al. (2023) และ Jiang (2025) ที่ชี้ว่า ARIMA ยังคงเหมาะสมในพื้นที่ที่มีแนวโน้มข้อมูลสม่ำเสมอและไม่แปรปรวนตามฤดูกาลมากนัก

ในภาพรวม งานวิจัยนี้ชี้ให้เห็นว่าการเลือกใช้แบบจำลองในการพยากรณ์ PM2.5 ควรอิงตามลักษณะของพื้นที่เกษตรเป็นสำคัญ เช่น ในพื้นที่ที่มีการเปลี่ยนแปลงตามฤดูกาลสูง มีปัจจัยแวดล้อมซับซ้อน หรือมีการจัดการหลังฤดูเก็บเกี่ยวด้วยการเผา Random Forest จะเป็นทางเลือกที่เหมาะสมและให้ความแม่นยำสูง ขณะที่ในพื้นที่เกษตรที่มีการควบคุมหรือใช้ระบบเรือนเพาะชำ หรือมีข้อมูลคุณภาพอากาศสม่ำเสมอ โมเดลเชิงสถิติอาจเพียงพอสำหรับการประเมินเบื้องต้น ดังนั้น การผสมผสานระหว่างแบบจำลอง Machine Learning และแบบจำลองเชิงสถิติ อาจเป็นแนวทางที่เหมาะสมสำหรับการพัฒนาเครื่องมือพยากรณ์ PM2.5 เพื่อใช้สนับสนุนการตัดสินใจด้านการวางแผนการเพาะปลูก ปรับปรุงระบบเตือนภัยล่วงหน้า และลดผลกระทบจากฝุ่นละอองต่อผลผลิตทางการเกษตรได้อย่างมีประสิทธิภาพและยั่งยืนในระยะยาว

ข้อเสนอแนะ

จากผลการศึกษานี้สามารถสรุปได้ว่า การประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง โดยเฉพาะแบบจำลอง Random Forest มีประสิทธิภาพในการพยากรณ์ค่าฝุ่นละออง PM2.5 ได้แม่นยำกว่าวิธีการดั้งเดิม เช่น Holt–Winters และ ARIMA รวมถึงแบบจำลอง SVR โดยเฉพาะในพื้นที่เกษตรกรรมที่มีข้อมูลผันผวนและลักษณะไม่เป็นเชิงเส้น เช่น จังหวัดเชียงใหม่และขอนแก่น Random Forest สามารถลดค่าความคลาดเคลื่อนในการพยากรณ์ได้อย่างมีนัยสำคัญ และแสดงศักยภาพในการตรวจจับรูปแบบข้อมูลที่ซับซ้อนได้ดีกว่าแบบจำลองอื่น ในขณะที่แบบจำลองเชิงสถิติดั้งเดิมยังคงให้ผลลัพธ์ที่ดีในพื้นที่ที่มีข้อมูลต่อเนื่องและแนวโน้มชัดเจน เช่น จังหวัดสงขลา ซึ่งเป็นพื้นที่เกษตรถาวรและไม่ได้มีพฤติกรรมการเผาในที่โล่งมากนัก จากผลการวิเคราะห์ Feature Importance จากแบบจำลอง Random Forest ยังแสดงให้เห็นว่า ตัวแปรเชิงเวลา เช่น ค่าฝุ่นย้อนหลัง (Lag) และค่าเฉลี่ยเคลื่อนที่ (Rolling Mean) มีบทบาทสำคัญอย่างยิ่ง ซึ่งสอดคล้องกับรูปแบบการทำเกษตรที่สัมพันธ์กับฤดูกาลและสภาพภูมิอากาศ การนำข้อมูลในอดีตมาใช้ประกอบการพยากรณ์จึงมีส่วนสำคัญต่อความแม่นยำของแบบจำลอง

จากผลการศึกษาที่ดำเนินการในงานวิจัยนี้ พบว่าแบบจำลอง Random Forest มีความแม่นยำสูงที่สุดเมื่อเปรียบเทียบกับวิธีการพยากรณ์เชิงสถิติดั้งเดิม (Holt–Winters และ ARIMA) และแบบจำลองเชิงเครื่องจักรกลแบบเส้นตรงอย่าง SVR ในหลายพื้นที่ของประเทศไทย โดยเฉพาะจังหวัดที่มีพฤติกรรมข้อมูลผันผวนสูง อย่างไรก็ตาม เพื่อให้เกิดความสมบูรณ์เชิงวิชาการและเพิ่มความน่าเชื่อถือของข้อสรุป งานวิจัยควรขยายกรอบการวิเคราะห์เพิ่มเติม ดังนี้

1. ควรเพิ่มการเปรียบเทียบประสิทธิภาพกับแบบจำลองขั้นสูงที่เป็นมาตรฐานสากลในงานพยากรณ์ข้อมูลเชิงเวลา เช่น LightGBM และ LSTM ซึ่งมีศักยภาพสูงในการจัดการข้อมูลที่ไม่เป็นเชิงเส้นและข้อมูลที่มีรูปแบบสลับซับซ้อน การเปรียบเทียบเพิ่มเติมจะช่วยยืนยันข้อได้เปรียบหรือข้อจำกัดของ Random Forest อย่างมีน้ำหนักมากขึ้น

2. ควรพิจารณาวิเคราะห์ผลต่างระหว่างการใช้ข้อมูลสิ่งแวดล้อมแบบดั้งเดิม (เช่น ปริมาณน้ำฝน อุณหภูมิ ความชื้น) และข้อมูลที่มีความละเอียดสูง เช่น ข้อมูลรายวันหรือรายชั่วโมง เพื่อประเมินว่าความละเอียดของข้อมูลส่งผลต่อความแม่นยำของแบบจำลองในระดับใด

3. ควรพัฒนาการวิเคราะห์เชิงลึกเกี่ยวกับ Feature Engineering โดยเฉพาะการออกแบบตัวแปรประเภท lag และตัวแปร

ดัชนีเชิงพฤติกรรม เช่น ช่วงฤดูการเผา หรือดัชนีความหนาแน่นของกิจกรรมเกษตรกรรม ซึ่งอาจมีผลต่อความสามารถของแบบจำลองเชิงเรียนรู้ของเครื่องอย่างมีนัยสำคัญ

4. ควรขยายพื้นที่ศึกษาครอบคลุมหลายประเภทของระบบเกษตร เช่น พื้นที่ปลูกพืชไร่ พืชสวน เกษตรแบบเรือนเพาะชำ และพื้นที่เกษตรกรรมที่มีการบริหารจัดการเชิงกลยุทธ์แตกต่างกัน เพื่อตรวจสอบว่าโครงสร้างเชิงพื้นที่มีผลต่อรูปแบบค่าฝุ่นและประสิทธิภาพของแบบจำลองมากน้อยเพียงใด

5. ควรพิจารณาสร้าง Hybrid Model ที่ผสานข้อดีของแบบจำลองทางสถิติเชิงเวลาเข้ากับความสามารถด้านการเรียนรู้ของเครื่อง เช่น ARIMA–LSTM หรือ Holt–Winters ร่วมกับ LightGBM เพื่อเพิ่มประสิทธิภาพการจับโครงสร้างข้อมูลทั้งเชิงเส้นและไม่เป็นเชิงเส้นพร้อมกัน ซึ่งสอดคล้องกับลักษณะของข้อมูล PM2.5 ที่มีทั้งฤดูกาล แนวโน้ม และสัญญาณสุ่มขั้นสูง

โดยสรุป การพัฒนาแบบจำลองพยากรณ์ PM2.5 ในงานวิจัยนี้ควรมีการเปรียบเทียบเพิ่มเติมกับแบบจำลองสมัยใหม่ที่ได้รับการยอมรับในระดับสากล รวมถึงการเพิ่มมุมมองด้าน Feature Engineering และการเลือกใช้ข้อมูลความละเอียดสูง เพื่อให้ข้อเสนอเชิงนโยบายและการใช้งานจริงในภาคเกษตรกรรมและสิ่งแวดล้อมมีความน่าเชื่อถือและพร้อมนำไปใช้ได้อย่างแท้จริง ซึ่งแบบจำลอง Random Forest ถือเป็นเครื่องมือที่มีศักยภาพสูงในการพยากรณ์ค่าฝุ่นละออง PM2.5 ในบริบทของการเกษตรไทย สามารถนำไปประยุกต์ใช้เป็นระบบสนับสนุนการตัดสินใจเพื่อวางแผนเพาะปลูกและจัดการพื้นที่เกษตรได้อย่างมีประสิทธิภาพ ทั้งในระดับไร่นา ระดับชุมชน และระดับนโยบาย โดยมีเป้าหมายสำคัญคือการลดผลกระทบจากมลพิษทางอากาศต่อผลผลิตและสุขภาพแรงงานในภาคเกษตรกรรมอย่างยั่งยืน

References

Gupta, P., Zhan, S., Mishra, V., Aekakkararungroj, A., Markert, A., Paibong, S., & Chishtie, F. (2021). Machine learning algorithm for estimating surface PM2.5 in Thailand. *Aerosol and Air Quality Research, 21*(11), 210105. doi:10.4209/aaqr.210105.

Duan, M., Sun, Y., Zhang, B., Chen, C., Tan, T., & Zhu, Y. (2023). *PM2.5 concentration prediction in six majorChinese urban agglomerations: A comparative study of various machine learning methods based on meteorological data*. *Atmosphere, 14*(5), 903. doi: 10.3390/atmos14050903.

Hasnain, A., Hashmi, M. Z., Khan, S., Bhatti, U. A., Min, X., Yue, Y., He, Y., & Wei, G. (2024). Predicting ambient *PM2.5* concentrations via time series models in Anhui Province, China. *Environmental Monitoring and Assessment, 196*(487). doi: 10.1007/s10661-024-12644-9.

Jiang, C. (2025). Comparative Analysis of ARIMA and Deep Learning Models for Time Series Prediction.*Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)* (p306–310). Kuala Lumpur: Science and Technology Publications.

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet, 15*(8), 255. doi: 10.3390/fi15080255.

Li, X., Li, L., Chen, L., Zhang, T., Xiao, J., & Chen, L. (2022). Random Forest Estimation and Trend Analysis of PM2. 5 Concentration over the Huaihai Economic Zone, China (2000– 2020). *Sustainability, 14*(14), 8520. doi: 10.3390/SU14148520.

Liu, R., Pang, L., Yang, Y., Gao, Y., Gao, B., Liu, F., & Wang, L. (2023). Air Quality—Meteorology CorrelationModeling Using Random Forest and Neural Network. *Sustainability*, *15*(5), 4531. doi: 10.3390/su15054531.

Makhdoomi, A., Sarkhosh, M., & Ziaei, S. (2025). PM2.5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations. *Scientific Reports, 15*, 14775. doi: 10.1038/s41598-025-92019-3.

Merdani, A. (2024). Comparative machine learning analysis of PM2.5 and PM10 forecasting in Albania. In*International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. 1-7. doi: 10.23919/SoftCOM62040.2024.10721971.

Minsan, W., Minsan, P., & Panichkitkosolkul, W. (2024). Enhancing Decomposition and Holt–Winters WeeklyForecasting of PM2.5 Concentrations in Thailand's Eight Northern Provinces Using the Cuckoo Search Algorithm. *Thailand Statistician*, *22*(4), 963–985.

Mohammadi, F., Teiri, H., Hajizadeh, Y., Abdolahnejad, A., & Ebrahimi, A. (2024). Prediction of atmospheric PM2.5 level by machine learning techniques in Isfahan Iran. *Scientific Reports, 14*, 2109. doi:10.1038/s41598-024-52617-z.

Pollution Control Department (PCD). (2023). Thailand air quality situation report 2023. Accessed 15 March2024, Retrieved from http://www.pcd.go.th.

Ratchagit, M. (2024). Forecasting PM2.5 concentrations in Chiang Mai using machine learning models.*International Journal on Robotics, Automation and Sciences, 6*(2), 37–41. doi:10.33093/ijoras.2024.6.2.6.

Thai Quality Historical Data Platform. (2024). Historical air quality data and monthly PM2.5 statistics.Accessed 12 March 2024. Retrieved from https://aqicn.org/historical.

Wongoutong, C. (2021). The effect on forecasting accuracy of the Holt–Winters method when using the incorrect model on a non-stationary time series. *Thailand Statistician*, *19*(3), 565–582.

World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10),ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Accessed 10 March 2024, Retrieved from https://www.who.int/publications/i/item/9789240034228.

Joharestani, M. Z., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere, 10*(7), 373. doi: 10.3390/atmos10070373.

Research article

Development of a Forecasting Model for PM2.5 Concentrations Using Machine Learning Techniques: A Case Study in Thailand

Nitaya Buntao[1] Poonsak Sirisom[1] Paramaporn Sangpara[1] Warida palasri[1]

Auten Gennaroj[3] and Rada Somkhuean[2*]

[1]*Department of Applied Statistics, Faculty of Science and Technology, Mahasarakham Rajabhat University, Mueang District, Maha Sarakham Province, Thailand 44000*
[2]*Department of Mathematics, Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna, Mueang District, Chiang Mai Province, Thailand 50300*
[3]*Chiang Yuen District Public Health Office, Chiang Yuen District, Maha Sarakham Province, Thailand, 44160*

ARTICLE INFO

Article history
Received: 30 July 2025
Revised: 17 October 2025
Accepted: 12 December 2025
Online published: 30 January 2026

Keyword
PM2.5
machine learning
intelligent model
environment
agriculture

ABSTRACT

Fine particulate matter smaller than 2.5 micrometers (PM2.5) poses a persistent threat to public health, environmental quality, and agricultural productivity in Thailand, particularly during the dry season, when concentrations frequently exceed national standards. This study aims to develop a PM2.5 forecasting model by comparing two time-series forecasting approaches—Holt–Winters exponential smoothing and the ARIMA model—with machine learning techniques, namely Random Forest and Support Vector Regression (SVR). Monthly PM2.5 data from Bangkok, Chiang Mai, Khon Kaen, and Songkhla for the period 2018–2024 were utilized. The dataset was chronologically divided into an 80% training set and a 20% test set, and model performance was evaluated using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The findings indicate that the Random Forest model consistently achieved the lowest prediction errors across all provinces, particularly in areas with highly volatile PM2.5 patterns, such as Chiang Mai and Khon Kaen. In contrast, SVR yielded relatively low predictive accuracy. Traditional time-series models performed well in provinces with more stable air quality patterns, such as Songkhla. Lag variables and moving averages were identified as key predictors contributing to model accuracy. Overall, the Random Forest model demonstrates strong potential for application in air quality alert systems and for supporting evidence-based environmental and agricultural policy planning toward long-term sustainability.

*Corresponding author
E–mail address: rada@rmutl.ac.th (R. Somkhuean)
Online print: 30 January 2026 https://doi.org/10.14456/paj.2026.1