Faculty of Science and Technology, Uttaradit Rajabhat University

วารสารวิชาการวิทยาศาสตร์และวิทยาศาสตร์ประยุกต์

Journal homepage: https://ajsas.uru.ac.th/

การพยากรณ์ปริมาณฝุ่นละออง PM2.5 ในจังหวัดนครสวรรค์ ด้วยเทคนิคการเรียนรู้ของเครื่อง

Forecasting of PM2.5 in Nakhon Sawan Province Using Machine Learning Techniques

วิฑูร สนธิปักษ์[1] ชม ปานตา[2] ภาสกร วรอาจ[1] ถิรภัทร มีสำราญ[1*]

Withoon Sonthipak[1], Chom Panta[2], Phassakorn Worra-arj[1], Thiraphat Meesumram[1*]

[1]*สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครสวรรค์ อำเภอเมือง จังหวัดนครสวรรค์ 60000*

[1]*Department of Computer Science and Information Technology, Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Nakhon Sawan 60000, Thailand.*

[2]*สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครสวรรค์ อำเภอเมือง จังหวัดนครสวรรค์ 60000*

[2]*Department of Mathematics and Statistics, Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Nakhon Sawan 60000, Thailand.*

ข้อมูลบทความ

ประวัติบทความ

รับเมื่อ: 15 ธันวาคม 2567

แก้ไขเมื่อ: 25 ธันวาคม 2567

ตอบรับเมื่อ: 27 ธันวาคม 2567

เผยแพร่ออนไลน์: 30 ธันวาคม 2567

คำสำคัญ

ฝุ่นละออง PM2.5

การเรียนรู้ของเครื่อง

เทคนิคการถดถอยเชิงเส้น

เทคนิคโครงข่ายประสาทเทียม

เทคนิคการสุ่มป่าไม้

***ผู้ประพันธ์บรรณกิจ**

อีเมล์:

thiraphat.m@nsru.ac.th

(ถิรภัทร มีสำราญ)

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพยากรณ์ปริมาณฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอน (PM2.5) ในจังหวัดนครสวรรค์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง 3 เทคนิค ได้แก่ เทคนิคการถดถอยเชิงเส้น (Linear Regression: LR) เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) และเทคนิคการสุ่มป่าไม้ (Random Forest: RF) ข้อมูลที่ใช้ประกอบด้วยปัจจัย 9 ตัวแปร ซึ่งเก็บรวบรวมจากสถานีตรวจคุณภาพอากาศในพื้นที่โครงการชลประทานนครสวรรค์ ระหว่างเดือนมกราคม พ.ศ. 2563 ถึงเดือนธันวาคม พ.ศ. 2565 ผลการวิจัยแสดงให้เห็นว่า เทคนิคการสุ่มป่าไม้ (RF) มีประสิทธิภาพสูงสุดในการพยากรณ์ โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เท่ากับ 18.96 ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยกำลังสอง (RMSE) เท่ากับ 4.35 และสัมประสิทธิ์การตัดสินใจ (R^2) เท่ากับ 0.83 การใช้ตัวแบบพยากรณ์นี้ช่วยให้หน่วยงานภาครัฐสามารถวางแผนเชิงรุกเพื่อจัดการมลพิษทางอากาศและลดผลกระทบต่อสุขภาพของประชาชนได้อย่างมีประสิทธิภาพ

ARTICLE INFO

Article History
Received: 15 December 2024
Revised: 25 December 2024
Accepted: 27 December 2024
Available online: 30 December 2024

Keywords:
PM2.5
Machine Learning
Linear Regression
Artificial Neural Network
Random Forest

***Corresponding author**
Email address:
thiraphat.m@nsru.ac.th
(T. Meesumrarn)

ABSTRACT

This research aims to forecast the concentration of fine particulate matter (PM2.5) in Nakhon Sawan Province using three machine learning techniques: Linear Regression (LR) Artificial Neural Network (ANN) and Random Forest (RF). The dataset consists of 9 variables, collected from the air quality monitoring station in the Nakhon Sawan Irrigation Project from January 2020 to December 2022. The findings indicate that the Random Forest (RF) technique outperforms the other models, achieving a Mean Squared Error (MSE) of 18.96, a Root Mean Squared Error (RMSE) of 4.35, and a Coefficient of Determination (R^2) of 0.83. This forecasting model provides a valuable tool for governmental agencies to proactively manage air pollution and mitigate its adverse effects on public health.

1. บทนำ

ปัจจุบันปัญหามลพิษทางอากาศ โดยเฉพาะฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอน (PM2.5) ได้กลายเป็นปัญหาสำคัญที่ส่งผลกระทบต่อสุขภาพของประชาชนอย่างกว้างขวาง PM2.5 สามารถเข้าสู่ระบบทางเดินหายใจและกระแสเลือด ทำให้เกิดโรคเกี่ยวกับระบบทางเดินหายใจ โรคหัวใจ และโรคมะเร็งปอด โดยเฉพาะในพื้นที่ที่มีการขยายตัวของเมืองอย่างรวดเร็ว มีงานวิจัยหลายชิ้นที่ศึกษาเกี่ยวกับแหล่งที่มา ผลกระทบต่อสุขภาพ และแนวทางการจัดการฝุ่น PM2.5 เช่น สมพร จันทระ และคณะ (2561) [1] ได้ศึกษาการปล่อยฝุ่น PM2.5 จากการเผาชีวมวล 4 ชนิด ได้แก่ ฟางข้าว ต้นข้าวโพดแห้ง เศษใบไม้จากป่าเต็งรัง และเศษใบไม้จากป่าเบญจพรรณ พบว่าการเผาฟางข้าวและใบไม้จากป่าทำให้เกิด PM2.5 มากกว่าการเผาต้นข้าวโพดแห้ง วรนารา ชนะบวรสกุล และคณะ (2566) [2] ศึกษาความสัมพันธ์ระหว่างค่ามลพิษ PM2.5 กับโรคระบบทางเดินหายใจและหัวใจและหลอดเลือดในกรุงเทพมหานครและจังหวัดนครสวรรค์ พบว่าความเข้มข้นของ PM2.5 มีความสัมพันธ์กับการเพิ่มขึ้นของผู้ป่วยโรคเหล่านี้ Karimian และคณะ (2019) [3]ได้ประเมินประสิทธิภาพของวิธีการเรียนรู้ของเครื่องหลายแบบ เช่น Multiple Additive Regression Trees (MART), Deep Feedforward Neural Network (DFNN) และแบบจำลองไฮบริดที่ใช้ Long Short-Term Memory (LSTM) ในการพยากรณ์ค่าฝุ่น PM2.5 ผลการทดลองชี้ให้เห็นว่าแบบจำลอง LSTM มีความแม่นยำสูงสุดในการพยากรณ์ กฤติกา ทิพย์คำมี และคณะ (2566) [4]ได้เปรียบเทียบประสิทธิภาพของเทคนิคการเรียนรู้ของเครื่อง สำหรับการพยากรณ์ฝุ่นละอองขนาดเล็กในอากาศ (PM2.5) พบว่าเทคนิคที่มีความเหมาะสมที่สุดสำหรับการสร้างตัวแบบการพยากรณ์ ฝุ่นละอองขนาดเล็กในอากาศ คือ เทคนิคโครงข่ายประสาทเทียม (Neural Network)

ในประเทศไทย ฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอน (PM2.5) ได้ทวีความรุนแรงขึ้นอย่างต่อเนื่องในช่วงหลายปีที่ผ่านมา สาเหตุหลักมาจากการจราจรคับคั่ง การเผาในที่โล่ง การขยายตัวของภาคอุตสาหกรรม และสภาพภูมิอากาศที่เอื้อต่อการสะสมของฝุ่นละออง จังหวัดนครสวรรค์ ซึ่งเป็นพื้นที่ศูนย์กลางของภาคเหนือตอนล่างและมีบทบาทสำคัญด้านการคมนาคม การเกษตร และอุตสาหกรรม การเพิ่มขึ้นของกิจกรรมดังกล่าวทำให้เกิดปริมาณฝุ่นละออง PM2.5 สูงเกินค่ามาตรฐานในหลายช่วงเวลา จึงเป็นพื้นที่ที่ประสบปัญหาฝุ่น PM2.5 อย่างหลีกเลี่ยงไม่ได้ และยังส่งผลให้คุณภาพชีวิตของประชาชนลดลง

การพยากรณ์ปริมาณฝุ่นละออง PM2.5 เป็นวิธีการหนึ่งที่สามารถช่วยให้หน่วยงานภาครัฐและประชาชนสามารถเตรียมการรับมือกับปัญหามลพิษได้อย่างมีประสิทธิภาพ การใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการพยากรณ์ PM2.5 เป็นแนวทางที่ทันสมัยและมีประสิทธิภาพ เนื่องจากสามารถวิเคราะห์ข้อมูลขนาดใหญ่ที่มีความซับซ้อนสูง เช่น ข้อมูลสภาพอากาศ ประวัติการวัดค่า PM2.5 และปัจจัยด้านสิ่งแวดล้อมอื่น ๆ เพื่อสร้างแบบจำลองที่สามารถพยากรณ์ได้อย่างแม่นยำ ผู้วิจัยเห็นถึงปัญหาและความสำคัญดังกล่าว จึงได้ทำการวิจัยโดยมีวัตถุประสงค์เพื่อพยากรณ์ปริมาณของ PM2.5 ในจังหวัดนครสวรรค์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อแก้ไขปัญหามลพิษทางอากาศและให้ข้อมูลเชิงลึกที่มีคุณค่าสำหรับการวางแผนด้าน

สาธารณสุขและสิ่งแวดล้อมของจังหวัด โดยการสร้างตัวแบบและเปรียบเทียบประสิทธิภาพของตัวแบบจากเทคนิคการเรียนรู้ของเครื่อง 3 เทคนิค ซึ่งได้แก่ เทคนิคการถดถอยเชิงเส้น (Linear Regression (LR)) เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network (ANN)) และเทคนิคการสุ่มป่าไม้ (Random Forest (RF))

2. วิธีดำเนินการวิจัย

ผู้วิจัยได้ศึกษาเกี่ยวกับการพยากรณ์ปริมาณฝุ่นละออง PM2.5 ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยการสร้างตัวแบบและเปรียบเทียบประสิทธิภาพของตัวแบบจากเทคนิคการเรียนรู้ของเครื่อง 3 เทคนิค ได้แก่ เทคนิคการถดถอยเชิงเส้น (LR) เทคนิคโครงข่ายประสาทเทียม (ANN) และเทคนิคการสุ่มป่าไม้ (RF) ผ่านกระบวนการ ดังนี้

2.1. การเก็บรวบรวมข้อมูล (Data Collection)

งานวิจัยนี้ได้ใช้ชุดข้อมูลที่ได้จากการเก็บรวบรวมข้อมูลปริมาณฝุ่นละออง PM2.5 จากสถานีตรวจคุณภาพอากาศ กองจัดการคุณภาพอากาศและเสียง กรมควบคุมมลพิษ ในพื้นที่โครงการชลประทานนครสวรรค์ โดยเก็บข้อมูลเป็นรายวัน ตั้งแต่เดือน มกราคม พ.ศ. 2563 ถึง ธันวาคม พ.ศ. 2565 จำนวน 1,096 วัน สำหรับข้อมูลที่นำใช้ในการวิเคราะห์นี้ คือ ปัจจัยที่ทำให้เกิด ฝุ่นละอองขนาดเล็กในอากาศ จำนวน 9 ตัวแปร ประกอบไปด้วยวัน (DATE), PM2.5, PM10, ก๊าซคาร์บอนมอนอกไซด์ (CO), ไนโตรเจนไดออกไซด์ (NO2), ก๊าซซัลฟอร์ไดออกไซด์ (SO2) , ก๊าซโอโซน (O3) , อุณหภูมิ (TEMP), และ ปริมาณฝน (RAIN)

ตารางที่ 1 ข้อมูลที่ใช้ในการสร้างตัวแบบการพยากรณ์

ลำดับ	ตัวแปร	คำอธิบาย	ประเภทข้อมูล
1	DATE	วัน/เดือน/ปี	Integer
2	PM2.5	ฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอน	Real
3	PM10	ฝุ่นละอองขนาดไม่เกิน10 ไมครอน	Real
4	CO	ก๊าซคาร์บอนมอนอกไซด์	Real
5	NO2	ไนโตรเจนไดออกไซด์	Integer
6	SO2	ก๊าซซัลฟอร์ไดออกไซด์	Integer
7	O3	ก๊าซโอโซน	Integer
8	TEMP	อุณหภูมิ	Real
9	RAIN	ปริมาณฝน	Real

2.2 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลเป็นขั้นตอนสำคัญในการสร้างโมเดลการพยากรณ์ งานวิจัยนี้ผู้วิจัยได้ดำเนินเตรียมข้อมูล ดังนี้

2.2.1 การทำความสะอาดข้อมูล (Data Cleaning)

เนื่องจากข้อมูลที่ได้มาสำหรับงานวิจัยนี้ มีค่าสูญหาย (Missing Value) ในทุกตัวแปร และด้วยข้อมูลที่เกิดค่าสูญหายเป็นข้อมูลที่อยู่ในรูปแบบของข้อมูลเชิงปริมาณ ผู้วิจัยจึงได้ทำการแทนค่าสูญหาย (Replace Missing Values) โดยจะแทนค่าสูญหายด้วยการหาค่าเฉลี่ยจากข้อมูล 5 วันก่อนหน้าและ 5 วันถัดไปของวันที่เกิดค่าสูญหาย [4]โดยวิธีการนี้จะคงความต่อเนื่องของข้อมูล ช่วยลดความคลาดเคลื่อนที่อาจเกิดจากการแทนข้อมูล เนื่องจากค่าที่แทนมีพื้นฐานมาจากข้อมูลที่อยู่รอบ ๆ ทำให้ข้อมูลมีความสอดคล้องและใกล้เคียงกับความจริง ซึ่งข้อมูลอนุกรมเวลามักแสดงความสัมพันธ์ของข้อมูลในช่วงเวลาที่อยู่ใกล้กัน การใช้ค่าเฉลี่ยจากข้อมูลในช่วงใกล้เคียงช่วยรักษาความสัมพันธ์นี้ อีกทั้งค่าที่แทนด้วยวิธีนี้จะไม่เป็น Outlier หรือค่าที่ผิดปกติเมื่อเทียบกับข้อมูลจริง ทำให้โมเดลเรียนรู้ข้อมูลได้อย่างมีประสิทธิภาพ การใช้วิธีการนี้ควรใช้อย่างระมัดระวังโดยเฉพาะกับข้อมูลที่มีความแปรผันสูงหรือกรณีที่ข้อมูลรอบข้างไม่สมบูรณ์

2.2.2 การกำหนดหน้าที่ของตัวแปร (Set Role)

กำหนดตัวแปรวันเดือนปีที่เก็บข้อมูล (DATE) ให้ทำหน้าที่เป็นไอดี (ID) และกำหนดตัวแปร PM2.5 ทำหน้าที่เป็นตัวแปรตามที่ใช้สำหรับการพยากรณ์ สำหรับตัวแปรที่เหลือหน้าที่ให้เป็นตัวแปรอิสระ

2.2.3 การแบ่งชุดข้อมูล (Train-Test Split)

ผู้วิจัยได้ทำการแบ่งชุดข้อมูลเป็น 2 ชุด ชุดแรกเรียกว่า ชุดฝึก (Training Set) ตั้งแต่เดือน มกราคม พ.ศ. 2563 ถึง ธันวาคม พ.ศ. 2564 จำนวน 731 วัน อีกชุดเรียกว่า ชุดทดสอบ (Testing Set) ตั้งแต่เดือน มกราคม พ.ศ. 2565 ถึง ธันวาคม พ.ศ. 2565 จำนวน 365 วัน สัดส่วนของ Test Dataset ในที่นี้ประมาณ 33.30%เพื่อใช้ในการเปรียบเทียบประสิทธิภาพของตัวแบบ

2.3 การสร้างตัวแบบ (Modeling)

ในงานวิจัยนี้ผู้วิจัยใช้เทคนิคการเรียนรู้ของเครื่อง ในการสร้างตัวแบบและเปรียบเทียบประสิทธิภาพของตัวแบบจากเทคนิคการเรียนรู้ของเครื่อง

3 เทคนิค ได้แก่ เทคนิคการถดถอยเชิงเส้น (LR) เทคนิคโครงข่ายประสาทเทียม (ANN) และเทคนิคการสุ่มป่าไม้ (RF)

2.3.1 เทคนิคการถดถอยเชิงเส้น (Linear Regression)

เป็นเทคนิคการเรียนรู้ของเครื่อง ที่ใช้สำหรับการพยากรณ์หรือการทำนายค่าที่เป็นตัวเลข โดยสมมติว่ามีความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรอิสระ (Independent Variable) และตัวแปรตาม (Dependent Variable) สมการของการถดถอยเชิงเส้นสามารถเขียนได้เป็น [5]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + e_i \quad (1)$$

โดยที่ y คือ ตัวแปรตาม (Dependent Variable)

$x_1, x_2, x_3, \ldots, x_n$ คือ ตัวแปรอิสระ (Independent Variable)

β_0 คือ ค่าคงที่ (Intercept)

$\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ คือ ค่าสัมประสิทธิ์การถดถอย (Coefficients)

e_i คือ ค่าความคลาดเคลื่อน (Error)

ในงานวิจัยนี้ ผู้วิจัยใช้เทคนิคการถดถอยเชิงเส้น โดยกำหนดวิธีการเลือกตัวแปรอิสระเข้าในตัวแบบด้วยวิธีการเลือกแบบไปข้างหน้า (Forward Selection) และวิธีการกำจัดแบบถอยหลัง (Backward Elimination)

2.3.2 เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network)

เป็นเทคนิคการเรียนรู้ของเครื่อง ที่มีโครงสร้างเลียนแบบการทำงานของเซลล์ประสาทในสมองของมนุษย์ โดย ANN ใช้สำหรับแก้ปัญหาที่ซับซ้อน เช่น การพยากรณ์ การจัดกลุ่ม และการจำแนกประเภท [6] โดยโครงสร้างพื้นฐานของ ANN ประกอบด้วย

Input Layer (ชั้นข้อมูลนำเข้า) รับข้อมูลนำเข้าจากฟีเจอร์ต่าง ๆ ในชุดข้อมูล จำนวนเซลล์ประสาทในชั้นนี้เท่ากับจำนวนฟีเจอร์

Hidden Layer (ชั้นซ่อน) ประมวลผลข้อมูลผ่านการคำนวณที่ซับซ้อน สามารถมีได้หลายชั้น (Deep Neural Network)

Output Layer (ชั้นผลลัพธ์) ให้ผลลัพธ์ที่ต้องการ เช่น ค่าพยากรณ์ หรือการจัดประเภท

การออกแบบ ANN จะแบ่งออกเป็น Node เล็ก ๆ โดย Node แต่ละตัวถูกออกแบบมาโดยมีแรงบันดาลใจมาจากเซลล์ประสาท โดยแต่ละ Node มีหลักการทำงานคือการรับข้อมูลจำนวนมากเข้าสู่ช่องทาง Input หรือเทียบได้กับ Axon ของเซลล์ประสาท ข้อมูลแต่ละตัวจะถูกคำนวณโดยการคูณน้ำหนักของ Axon แต่ละเส้นก่อนที่กระแสสัญญาณจากทุกเส้นจะมารวมกัน และนำผลลัพธ์นี้เข้าสู่ Activation Function เพื่อคำนวณผลลัพธ์ของ Node โดยการคำนวณ Activation Function สามารถทำได้โดยการเลือกใช้ Activation Function ให้เหมาะกับงาน โดย Activation Function มีให้เลือกหลากหลาย แต่ที่นิยมใช้ และเกี่ยวข้องกับงานนี้มีด้วยกัน 4 ฟังก์ชัน ได้แก่ Identity, Rectifier (ReLU), Sigmoid และ Hyperbolic Tangent (tanh) ในงานวิจัยนี้ผู้วิจัยได้วิเคราะห์โดยใช้อัลกอริทึม Multilayer Perceptron จำนวนชั้นปกปิด เท่ากับ 5 ชั้น แต่ละชั้น มีจำนวน Node 20 โหนด ใช้ Identity เป็น Activation function

2.3.3 เทคนิคการสุ่มป่าไม้ (Random Forest)

เป็นเทคนิคที่พัฒนามาจากตัวแบบต้นไม้ตัดสินใจ (Decision Trees) โดยจะใช้วิธีการแบ่งจำนวนต้นไม้ตัดสินใจออกเป็นหลาย ๆ ต้น (Tree) โดยแต่ละต้นสร้างมาจากการสุ่มคุณลักษณะ (Feature) และข้อมูล (Data) บางส่วนด้วยวิธีบูตสแทร็ป (Bootstrap Method) จากคุณลักษระและข้อมูลชุดฝึกฝนทั้งหมด ซึ่งทำให้ได้ต้นไม้ตัดสินใจที่มีความเป็นอิสระต่อกันมากขึ้น จากนั้นหาค่าพยากรณ์จากต้นไม้ตัดสินใจที่ได้ในแต่ละต้น ทำซ้ำจนกว่าจะได้จำนวนต้นไม้ตัดสินใจครบตามกำหนด หากตัวแปรกลุ่มเป้าหมายที่สนใจศึกษาเป็นตัวแปรเชิงปริมาณ ซึ่งถือว่าเป็นปัญหาวิเคราะห์การถดถอย ในการหาค่าพยากรณ์สุดท้ายสามารถทำได้โดยการคำนวณหาค่าเฉลี่ยของค่าพยากรณ์ของต้นไม้ตัดสินใจทุกต้น [7] โดยหลักการทำงานของ Random Forest ประกอบด้วย

Bootstrap Aggregating (Bagging) สุ่มตัวอย่างข้อมูลจากชุดข้อมูลต้นฉบับ (Bootstrap Sampling) หลายครั้ง โดยอาจมีการสุ่มซ้ำ สร้างต้นไม้ตัดสินใจ (Decision Trees) หลายต้น โดยแต่ละต้นใช้ข้อมูลที่สุ่มมา

Random Feature Selection ในแต่ละการสร้างต้นไม้ จะสุ่มเลือกเฉพาะบางฟีเจอร์ (Features) แทนการใช้ทุกฟีเจอร์

การรวมผลลัพธ์ (Aggregation) หากเป็นการจำแนก จะใช้การโหวตข้างมาก (Majority Voting) จากผลลัพธ์ของทุกต้นไม้ และถ้าเป็นการพยากรณ์ค่าเชิงปริมาณ จะใช้ค่าเฉลี่ยของผลลัพธ์จากทุกต้นไม้ ผู้วิจัยได้วิเคราะห์โดยใช้อัลกอริทึม Random Forest โดยกำหนดจำนวนต้นไม้ตัดสินใจ เท่ากับ100 ต้น และจำนวนกิ่งของต้นไม้ตัดสินใจแต่ละต้นไม่เกิน 100 กิ่ง

2.4 การทดสอบประสิทธิภาพของตัวแบบ (Evaluation Metrics)

สำหรับงานวิจัยนี้ ผู้วิจัยเลือกใช้เกณฑ์ในการวัดค่าความแม่นยำของตัวแบบ เพื่อทำการวัดประสิทธิภาพก่อนที่จะนำตัวแบบไปใช้ในการพยากรณ์ ซึ่งมี 3 ค่า ประกอบด้วย

2.4.1 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error: MSE)

เป็นค่าที่ใช้วัดความคลาดเคลื่อนระหว่างค่าจริง (y_i) กับค่าที่พยากรณ์ ($\hat{y}_i$) โดยการหาความคลาดเคลื่อนกำลังสองและเฉลี่ยออกมา โดยมีสูตรในการคำนวณ ดังนี้ [8]

$$MSE = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n} \quad (2)$$

โดยที่ y_i คือ ค่าจริง

$\hat{y}_i$ คือ ค่าพยากรณ์

n คือ จำนวนข้อมูล

2.4.2 ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Squared Error: RMSE)

เป็นค่าที่ได้จากการถอดรากที่สองของ MSE เพื่อทำให้ค่าความคลาดเคลื่อนกลับมาอยู่ในหน่วยเดียวกับข้อมูลเดิม ทำให้อ่านและตีความได้ง่ายขึ้น

2.4.3 สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2)

เป็นตัวชี้วัดว่าโมเดลสามารถอธิบายความแปรปรวนของข้อมูลได้ดีเพียงใด โดยมีสูตรในการคำนวณ ดังนี้ [9]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

โดยที่ y_i คือ ค่าจริง

$\hat{y}_i$ คือ ค่าพยากรณ์

$\bar{y}$ คือ ค่าเฉลี่ยของ y_i

n คือ จำนวนข้อมูล

3. ผลการวิจัยและอภิปรายผล

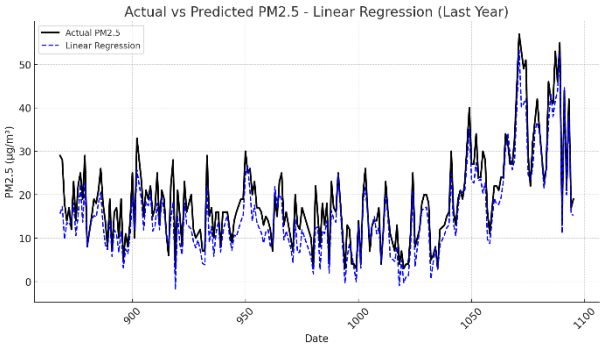
ผู้วิจัยสร้างตัวแบบจากเทคนิคการเรียนรู้ของเครื่อง 3 เทคนิค ได้แก่ เทคนิคการถดถอยเชิงเส้น (LR) เทคนิคโครงข่ายประสาทเทียม (ANN) และเทคนิคการสุ่มป่าไม้ (RF) และเปรียบเทียบประสิทธิภาพของตัวแบบ โดยการทดสอบประสิทธิภาพของตัวแบบ 3 ค่า ได้แก่ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยกำลังสอง (RMSE) และสัมประสิทธิ์การตัดสินใจ (R^2) โดยใช้ Python และโปรแกรม Weka ซึ่งได้ผลดังตารางที่ 2

ตารางที่ 2 ตารางเปรียบเทียบประสิทธิภาพของตัวแบบ

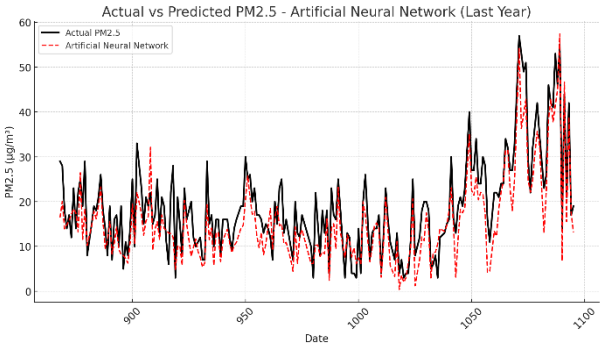
ตัวแบบ	MSE	RMSE	R^2
Linear Regression	29.45	5.43	0.74
Artificial Neural Network	25.79	5.08	0.77
Random Forest*	18.96	4.35	0.83

*คือ ตัวแบบที่มีประสิทธิภาพในการพยากรณ์ปริมาณของ PM2.5 ในจังหวัดนครสวรรค์

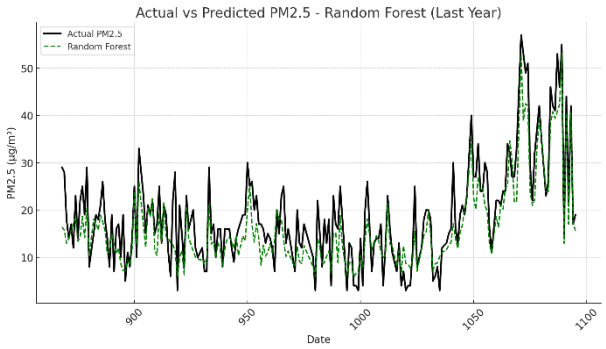
จากตารางที่ 2 พบว่า ตัวแบบที่มีประสิทธิภาพในการพยากรณ์ปริมาณของ PM2.5 ในจังหวัดนครสวรรค์ คือ เทคนิคการสุ่มป่าไม้ (RF) มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย เท่ากับ 18.96 ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยกำลังสอง เท่ากับ 4.35 ซึ่งเป็นค่าที่น้อยที่สุดเมื่อเทียบกับเทคนิคอื่น และให้สัมประสิทธิ์การตัดสินใจ เท่ากับ 0.83 ซึ่งมากที่สุดใน 3 เทคนิคที่นำเสนอในงานวิจัยนี้ และจากการเปรียบเทียบประสิทธิภาพการพยากรณ์ได้ผลดัง ภาพที่ 1 ภาพที่ 2 และภาพที่ 3



ภาพที่ 1 เปรียบเทียบข้อมูลระหว่างค่าจากข้อมูลจริงและค่าจากการพยากรณ์ ด้วยเทคนิคการถดถอยเชิงเส้น



ภาพที่ 2 เปรียบเทียบข้อมูลระหว่างค่าจากข้อมูลจริงและค่าจากการพยากรณ์ ด้วยเทคนิคโครงข่ายประสาทเทียม



ภาพที่ 3 เปรียบเทียบข้อมูลระหว่างค่าจากข้อมูลจริงและค่าจากการพยากรณ์ ด้วยเทคนิคการสุ่มป่าไม้

จากภาพที่ 3 เมื่อเปรียบเทียบปริมาณปริมาณฝุ่น PM2.5 ทั้งค่าจริง (Actual) และค่าพยากรณ์ (Predicted) แสดงให้เห็นว่าข้อมูลปริมาณฝุ่น PM2.5 และข้อมูลพยากรณ์ของปริมาณฝุ่น PM2.5 มีค่าใกล้เคียงกันมาก สอดคล้องกับค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยกำลังสอง และสัมประสิทธิ์การตัดสินใจ ซึ่งแสดงให้เห็นว่า เทคนิคที่มีประสิทธิภาพในการนำไปใช้ในการพยากรณ์ปริมาณของ PM2.5 ในจังหวัดนครสวรรค์ มากที่สุดคือ เทคนิคการสุ่มป่าไม้ (RF) ซึ่งเป็นเทคที่เหมาะสำหรับข้อมูลที่มีความซับซ้อนและความสัมพันธ์ที่ไม่สามารถจำลองด้วยโมเดลเชิงเส้นได้

4. สรุปผล

การวิจัยนี้มีวัตถุประสงค์เพื่อพยากรณ์ปริมาณของ PM2.5 ในจังหวัดนครสวรรค์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อแก้ไขปัญหามลพิษทางอากาศและให้ข้อมูลเชิงลึกที่มีคุณค่าสำหรับการวางแผนด้านสาธารณสุขและสิ่งแวดล้อมของจังหวัด โดยการสร้างตัวแบบและเปรียบเทียบประสิทธิภาพของตัวแบบจากเทคนิคการเรียนรู้ของเครื่องโดยใช้ชุดข้อมูลที่ได้จากการเก็บรวบรวมข้อมูลปริมาณฝุ่นละออง PM2.5 จากสถานีตรวจคุณภาพอากาศ กองจัดการคุณภาพอากาศและเสียง กรมควบคุมมลพิษ ในพื้นที่โครงการชลประทานนครสวรรค์ โดยเก็บข้อมูลเป็นรายวัน จากผลการวิจัยสรุปได้ว่า เทคนิคการสุ่มป่าไม้ (RF) เป็นเทคนิคที่มีประสิทธิภาพดีที่สุด สอดคล้องกับ [10][11]ที่กล่าวว่า เทคนิคการสุ่มป่าไม้ เหมาะสำหรับข้อมูลที่มีความซับซ้อนและความสัมพันธ์ที่ไม่สามารถจำลองด้วยโมเดลเชิงเส้นได้ แสดงให้เห็นว่า เทคนิคการเรียนรู้ของเครื่อง โดยเฉพาะเทคนิคการสุ่มป่าไม้ สามารถพยากรณ์ปริมาณ PM2.5 ของจังหวัดนครสวรรค์ได้อย่างมีประสิทธิภาพ โดยให้ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยกำลังสอง เท่ากับ 4.35 นั่นหมายความว่า ถ้าแบบจำลองพยากรณ์ว่าค่า PM2.5 จะอยู่ที่ 50 ไมโครกรัมต่อลูกบาศก์เมตร จะมีค่าความคลาดเคลื่อนที่อาจเกิดขึ้นได้ในช่วง ±4.35 ไมโครกรัมต่อลูกบาศก์ ซึ่งหน่วยงานสาธารณสุขควรเตรียมการรับมือและคำนึงถึง สามารถใช้เป็นเครื่องมืออันมีค่าสำหรับผู้กำหนดนโยบายในการดำเนินมาตรการเชิงรุกสำหรับการจัดการคุณภาพอากาศและการปรับปรุงด้านสาธารณสุข เพื่อลดผลกระทบจากมลพิษทางอากาศในระยะยาวได้อย่างมีประสิทธิภาพ การวิจัยในอนาคตอาจสำรวจแบบจำลองไฮบริด การบูรณาการแหล่งข้อมูลภาพถ่ายดาวเทียมและข้อมูลด้านสิ่งแวดล้อม เพิ่มเติมเพื่อเพิ่มความแม่นยำให้สูงขึ้น

5. ข้อเสนอแนะ

ในการวิจัยครั้งต่อไปควรศึกษาเพิ่มเติมเกี่ยวกับการทดลองใช้โมเดลที่ออกแบบมาเฉพาะสำหรับข้อมูลอนุกรมเวลา เช่น LSTM, ARIMA, SARIMA หรือแบบผสม ควรพิจารณาเพิ่มการวิเคราะห์ความเป็นฤดูกาลและแนวโน้มของข้อมูล พิจารณาตัวแปรอื่นที่เกี่ยวข้องกับปริมาณ PM2.5 นอกเหนือจากที่นำเสนอในงานวิจัยนี้ หรือศึกษาข้อมูลจากหลายสถานีตรวจวัด เพื่อปรับปรุงความแม่นยำในการพยากรณ์

เอกสารอ้างอิง

[1] สมพร จันทระ, ชาคริต โชติอมรศักดิ์, & ว่าน วิริยา. (2561). การติดตามตรวจสอบการเผาในที่โล่งใน ภาคเหนือของประเทศไทย สำหรับการประเมินการปล่อยและการเคลื่อนที่ของมลพิษทาง อากาศ เพื่อการวางแผนการจัดการปัญหาหมอกควัน. รายงานวิจัยฉบับสมบูรณ์. สำนักงานกองทุนสนับสนุนการวิจัย (สกว.).

[2] วรนารา ชนะบวรสกุล, เสรีย์ ตู้ประกาย, ปิยะรัตน์ ปรีย์มาโนช, & มงคล รัชชะ. (2566). ความสัมพันธ์ระหว่างค่ามลพิษฝุ่นละออง PM2.5กับโรคระบบทางเดินหายใจ และโรคหัวใจหลอดเลือด : กรณีศึกษาพื้นที่กรุงเทพมหานครและจังหวัดนครสวรรค์. วารสารวิจัย วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา, 8(1), 61-72.

[3] Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., & Sachdeva, S. (2019). Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations. Aerosol Air Qual. Res. 19: 1400-1410. https://doi.org/10.4209/aaqr.2018.12.0450

[4] กฤติกา ทิพย์คำมี, อนุพงศ์ สุขประเสริฐ, สุพัตรา กอผจญ, & ณัฐกานต์ ชุติมารังสรรค. (2566). ประสิทธิภาพของเทคนิคการเรียนรู้ของเครื่อง สำหรับการพยากรณ์ฝุ่นละอองขนาดเล็กในอากาศ. วารสารวิทยาศาสตร์ เทคโนโลยี และนวัตกรรม มหาวิทยาลัยกาฬสินธุ, 2(1), 58-74.

[5] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). Wiley.

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[7] Granata, F., Saroli, M., de Marinis, G., & Gargano, R. (2018). Machine learning models for spring discharge forecasting. Geofluids, 2018, 1-15.

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

[9] Draper, N. R., & Smith, H. (1998). Applied regression analysis (3rd ed.). Wiley.

[10] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.