PM2.5 in Northern Thailand: from raw data to a recommendation

จัดทำโดย

พิมพ์ภัทร มยุระสาคร รหัสนักศึกษา 690631022

เสนอ

ผศ. ดร. ภาสกร แช่มประเสริฐ

รายงานนี้เป็นส่วนหนึ่งของวิชา 27070 Data Science Programming

ภาคเรียนที่ 1 ปีการศึกษา 2569

มหาวิทยาลัยเชียงใหม่

1. Problem Background

ในทุกๆ ปี ระหว่างช่วงเดือนมกราคมถึงเมษายน อากาศในภาคเหนือของประเทศไทยเข้าสู่วิกฤตมลพิษทางอากาศขั้นรุนแรงจนเป็นอันตรายต่อระบบทางเดินหายใจอย่างมาก ต้นเหตุหลักเกิดจากการเผาเศษวัสดุทางการเกษตร ไฟป่า และมลพิษข้ามพรมแดน ซึ่งเมื่อเผชิญกับเงื่อนไขทางอุตุนิยมวิทยาที่มีสภาพอากาศนิ่ง (Atmospheric Stagnation) ร่วมกับลักษณะทางกายภาพของเมืองเชียงใหม่ที่เป็นหุบเขาล้อมรอบด้วยภูเขาสูง ส่งผลให้ฝุ่นละอองขนาดเล็กถูกกักขังและสะสมตัวอยู่ในแอ่งกระทะอย่างหลีกเลี่ยงไม่ได้

วิกฤตการณ์ดังกล่าวส่งผลกระทบเป็นวงกว้าง ส่งผลให้โรงเรียนต้องประกาศปิดเรียนชั่วคราว และเกิดความเสียหายต่ออุตสาหกรรมการท่องเที่ยวในจังหวัดเชียงใหม่อย่างรุนแรง เมืองเชียงใหม่ยังถูกจัดอันดับให้อยู่ในกลุ่มเมืองที่มีมลพิษทางอากาศสูงที่สุดในโลกหลายครั้ง โดยเฉพาะรายงานสถานการณ์คุณภาพอากาศจาก IQAir (IQAir, 2026) ยอดผู้ป่วยเข้ารับการรักษาในโรงพยาบาลด้วยโรคระบบทางเดินหายใจยังพุ่งสูงขึ้นอย่างมีนัยสำคัญ หลักฐานทางวิทยาศาสตร์ชี้ให้เห็นว่าการได้รับฝุ่น PM2.5 ในเชียงใหม่สัมพันธ์กับภาวะความเครียดระดับเซลล์ (Oxidative Stress) ในร่างกาย (Assi et al., 2025) รวมถึงส่งผลกระทบโดยตรงต่อสุขภาพดวงตาของประชาชนในพื้นที่อย่างเด่นชัด (Waikare et al., 2024)

2. Approach

คำถามหลักของงานนี้คือ: ฝุ่น PM2.5 ในเชียงใหม่รุนแรงแค่ไหน เกิดขึ้นเมื่อไหร่และที่ไหน สามารถทำนายคุณภาพอากาศของวันพรุ่งนี้ได้แม่นยำพอที่จะใช้เป็นระบบเตือนภัยหรือไม่ และควรมีมาตรการรับมืออย่างไร

Hypothesis: คาดว่าความเร็วลมจะเป็นปัจจัยอากาศหลักที่ควบคุมการสะสมของ PM2.5 โดยตั้งสมมติฐานว่าความเร็วลมน้อยจะทำให้มลพิษตกค้างอยู่ในแอ่งหุบเขา และคาดว่าอุณหภูมิที่สูงขึ้นในหน้าร้อนจะสัมพันธ์กับคุณภาพอากาศที่แย่ลง

3. Method

3.1. Data sources

- Open-Meteo Air Quality (PM2.5 and other pollutants)
  - latitude: 18.7904
  - longitude: 98.9847
  - hourly: ["pm10", "pm2_5", "carbon_monoxide", "dust"]
  - time zone: Asia/Bangkok
  - start date: 2023-01-01
  - end date: 2025-04-30
- Open-Meteo Archive (weather)
  - latitude: 18.7904
  - longitude: 98.9847
  - start date: "2023-01-01"
  - end date: "2025-04-30"

- hourly: ["temperature_2m", "relative_humidity_2m", "wind_speed_10m", "precipitation"]
- time zone: "Asia/Bangkok"

3.2. Cleaning and joining

- การปรับมาตรฐานเขตเวลา (Timezone Standardization): แปลงคอลัมน์เวลา date ของทั้งชุดข้อมูลคุณภาพอากาศ (Air Quality) และชุดข้อมูลสภาพอากาศ (Weather) ให้เป็นเขตเวลา Asia/Bangkok (+07:00) เพื่อให้สอดคล้องกับเวลาท้องถิ่นของประเทศไทย
- ตรวจสอบจำนวนแถว ประเภทข้อมูล (Data Types) และค่าที่หายไป (Missing Values) ก่อนทำการรวมข้อมูลFeatures
- ทำการรวมชุดข้อมูลทั้งสองเข้าด้วยกันโดยใช้คีย์ date

3.3. Features

- AQ
  - pm10, pm2_5, carbon_monoxide, dust
- weather
  - temperature_2m, relative_humidity_2m, wind_speed_10m, precipitation

3.4. Target definition

- Classification: ตัวบ่งชี้สภาวะเกินมาตรฐาน PM2.5 ในวันพรุ่งนี้ ค่า pm2.5 > 37.5 $\mu g/m^3$

3.5. Split strategy

- การแบ่งตามลำดับเวลา (Time-based Split): แบ่งข้อมูลตามช่วงเวลาโดยไม่สับเปลี่ยนข้อมูล (No Random Shuffle) เช่น ใช้ข้อมูลปี 2023–2024 เป็น Train Set และใช้ข้อมูลปี 2025 เป็น Test Set
- การทำ Cross-Validation: ใช้ TimeSeriesSplit ในการประเมินโมเดล เพื่อป้องกันไม่ให้โมเดลเรียนรู้ข้อมูลในอนาคตก่อนอดีต

3.6. Models and baseline

- Baseline Model: ใช้ Persistence Model (คาดการณ์ว่าค่า PM2.5 พรุ่งนี้ = ค่า PM2.5 วันนี้) สำหรับ Regression และใช้ Majority Class สำหรับ Classification
- Models: ทดสอบด้วย Linear Regression, Logistic Regression, Random Forest Regressor และ Random Forest Classifier

4. Results

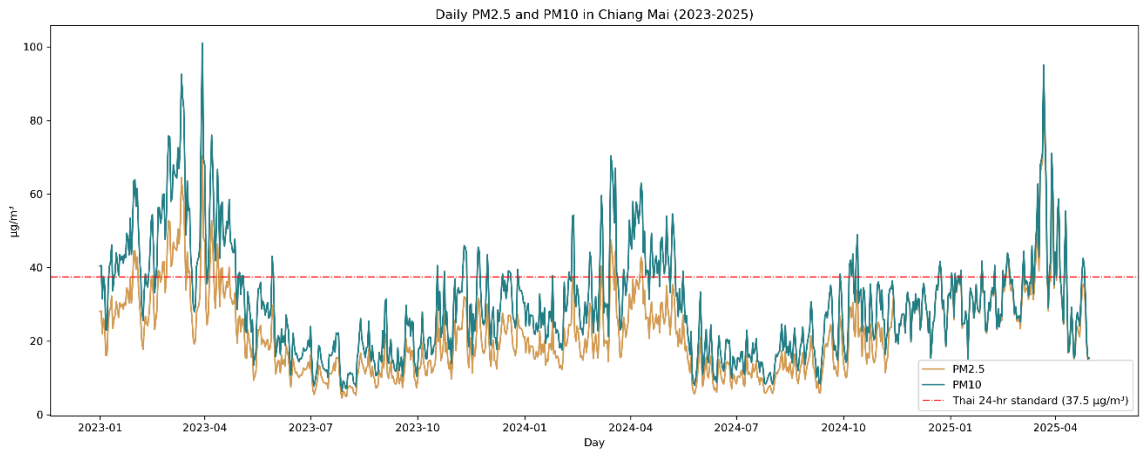
4.1. Statistical Testing Results (t-test)

ผลการทดสอบความแตกต่างของปัจจัยสภาพอากาศระหว่างวันปกติและวันฝุ่นสูง มีดังนี้

	variable	t_statistic	p_value	significant_p005
1	wind_speed_10m	-3.0243	0.002567124219722458	True
2	relative_humidity_2m	9.1962	2.802929938355764e-19	True
3	temperature_2m	-1.0276	0.30441886612514885	False

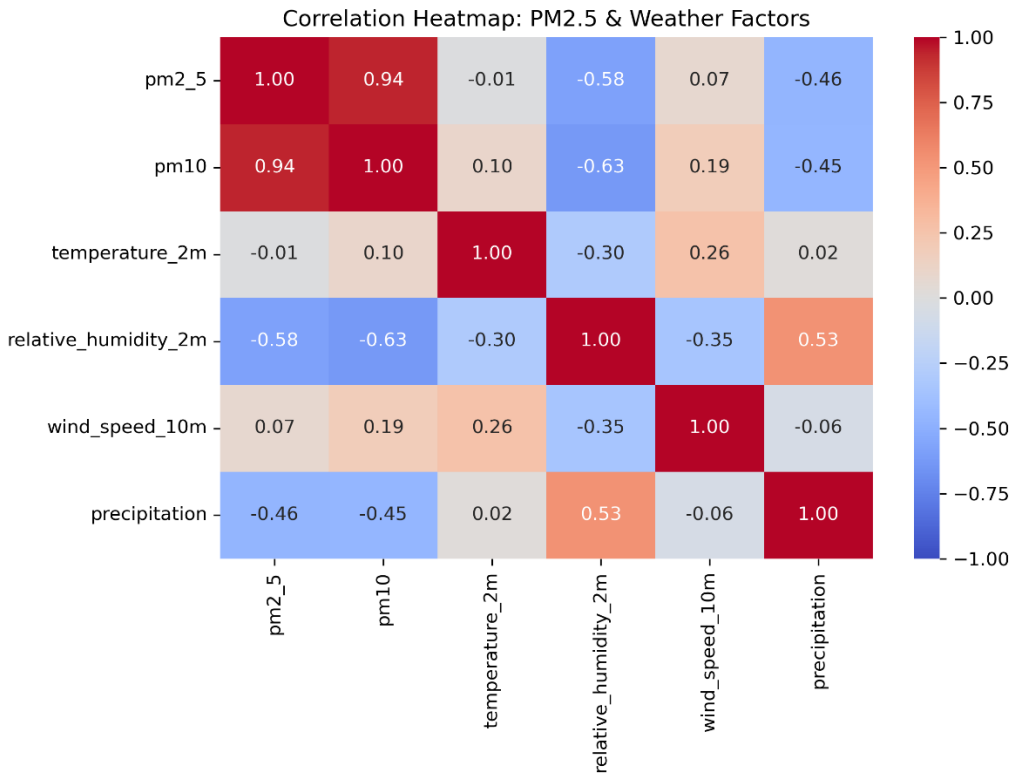
4.2. Data Visualization Insights

- Daily_PM2.5_and_PM10_in_Chiang_Mai_(2023-2025)



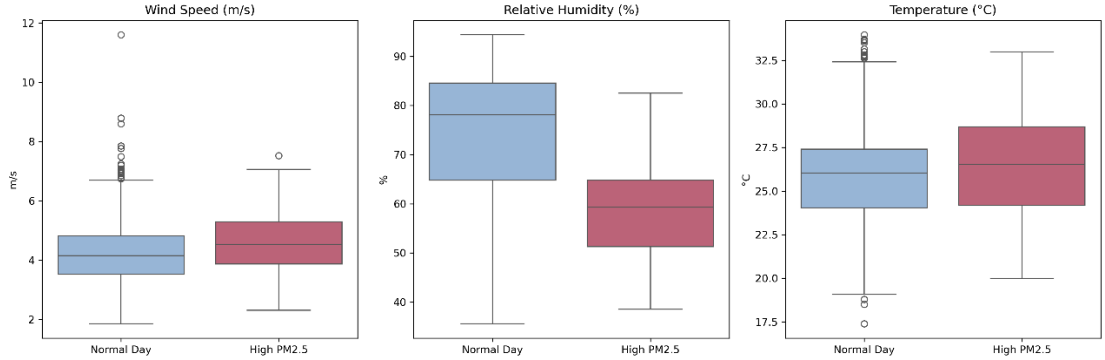
ฤดูกาลฝุ่น PM2.5 และ PM10 เริ่มต้นขึ้นในช่วงเดือนกุมภาพันธ์ และสิ้นสุดลงในเดือนเมษายนของทุกปีอย่างเด่นชัด โดยปี 2023 มีค่าความเข้มข้นสูงสุด (Peak) สูงกว่าปี 2024 และ 2025

- Correlation Heatmap: PM2.5 & Weather Factors



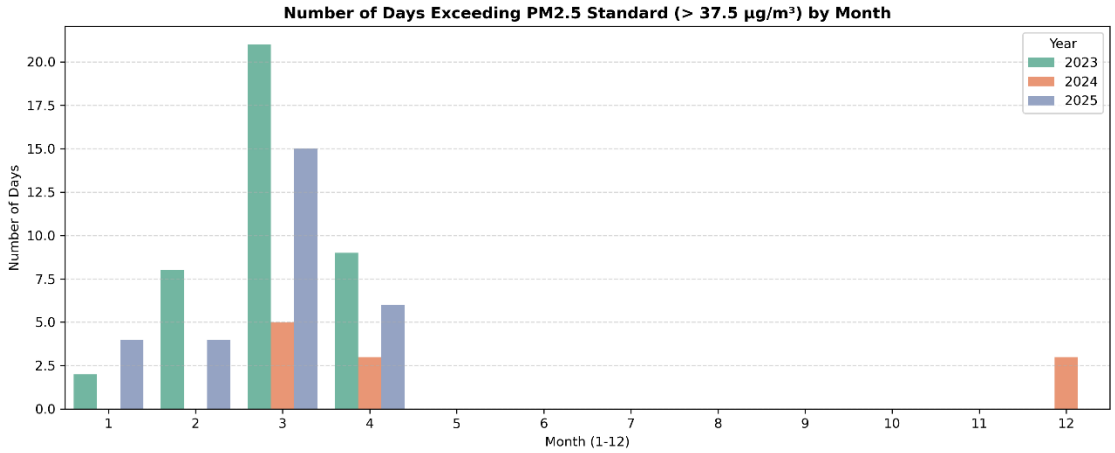
ความชื้นสัมพัทธ์ (relative_humidity_2m) มีความสัมพันธ์ทิศทางตรงกันข้ามกับ PM2.5 สูงที่สุด (r = -0.58) ตามด้วย ปริมาณฝน (r = -0.46) ในขณะที่อุณหภูมิแทบไม่มีความสัมพันธ์เชิงเส้นกับระดับฝุ่น (r = -0.01)

- Weather Conditions on Bad Days vs Normal Days



ในวันฝุ่นสูง (High PM2.5) ค่ามัธยฐานของความชื้นสัมพัทธ์จะลดลงอย่างเห็นได้ชัด (59%) เมื่อเทียบกับวันปกติ (78%) สอดคล้องกับผลการทดสอบ t-test ที่มีนัยสำคัญทางสถิติ ($p < 0.0001$)

- Number of Days Exceeding PM2.5 Standard (> 37.5 μg/m³) by Mont



จำนวนวันที่ฝุ่นเกินเกณฑ์มาตรฐาน (> 37.5 μg/m³) กระจุกตัวสูงสุดในเดือนมีนาคมของทุกปี โดยรวมแล้วปี 2023 มีจำนวนวันที่ฝุ่นเกินมาตรฐานมากที่สุด และมีแนวโน้มลดลงในปี 2024–2025

4.3. Model Performance

- Regression (ทำนายค่าเฉลี่ย PM2.5 ของวันพรุ่งนี้):

Model	Test MAE	Test RMSE	R^2	beats_baseline
Baseline	5.679	7.752	0.530	-
Linear Regression	5.135	7.307	0.582	Yes
Random Forest	5.909	8.238	0.469	no

- Classification (ทำนายว่าพรุ่งนี้เกิน 37.5 μg/m³ หรือไม่):

Model	Recall	Precision	F1	beats_baseline_recall
Baseline	0.000	0.000	0.000	-
Logistic Regression	0.906	0.362	0.518	yes
Random Forest Classifier	0.781	0.510	0.617	yes

เกิดจากปัญหา Overfitting ร่วมกับ Time-series Concept Drift เนื่องจาก Random Forest เป็นโมเดลที่มีความซับซ้อนสูง จึงจดจำรูปแบบ Noise ในข้อมูลอดีต (Train Set) มากเกินไป เมื่อนำมาทดสอบกับข้อมูลช่วงเวลาใหม่ใน Test Set (ปี 2025) ที่มีสภาพอากาศและรูปแบบฝุ่นเปลี่ยนไป โมเดล

จึงไม่สามารถปรับตัวได้ดี (Poor Generalization) ในขณะที่ Linear Regression เป็นโมเดลที่ไม่ซับซ้อน (Simpler Model) จึงไม่เกิด Overfitting และให้ผลลัพธ์บน Test Set ที่มั่นคงและดีกว่า Baseline ชัดเจน

4.4. What Didn't Work

Random Forest แพ้ทั้ง persistence baseline และ Linear Regression ในงาน regression (test MAE 5.909 vs. 5.135) ทั้งที่ตอน cross-validation ทำคะแนนดีกว่า (CV MAE 4.824 vs. 4.859) สัญญาณของ overfitting เพราะ Random Forest ซับซ้อนเกินไป จดจำ noise ของช่วง train (2023-2024) แล้ว generalize ไปปี 2025 ที่รูปแบบอากาศเปลี่ยนไปได้ไม่ดี

Linear Regression เองก็มี CV MAE std สูง (3.145) แสดงว่ายังไม่เสถียรเต็มที่ ส่วน Logistic Regression มี precision ต่ำ (0.362)

4.5. Discussion on Checkpoint Questions

C1 · Time

Show that your daily averages correspond to Thailand local days. Describe how you verified this rather than asserting it. If you fetched from two endpoints, show that their time axes agree.

C1 · answer in the report; note your evidence here

`-------- AQ --------`
`First timestamp: 2023-01-01 00:00:00+07:00`
`Last timestamp: 2025-04-30 23:00:00+07:00`
`Timezone info: Asia/Bangkok`
 	
`-------- Weather --------`
`First timestamp: 2023-01-01 00:00:00+07:00`
`Last timestamp: 2025-04-30 23:00:00+07:00`
`Timezone info: Asia/Bangkok`
 	
`-------- After Join --------`
`First timestamp: 2023-01-01 00:00:00+07:00`
`Last timestamp: 2025-04-30 23:00:00+07:00`
`Timezone info: Asia/Bangkok`

C2 · The join

How many rows did each source have before joining, and how many after? If the number changed, account for every row. If it did not change, show how you checked.

C2

ก่อน join

Air Quality: 20424 แถว , Weather: 20424 แถว

หลัง join – 20424 แถว

โดยเช็คจากการ ใช้ print(df_pm5_processed.shape)

```
-------- After Join --------

First timestamp: 2023-01-01 00:00:00+07:00
Last timestamp: 2025-04-30 23:00:00+07:00
Timezone info: Asia/Bangkok
<class 'pandas.DataFrame'>
RangeIndex: 20424 entries, 0 to 20423
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   date                  20424 non-null  datetime64[us, Asia/Bangkok]
 1   pm10                  20424 non-null  float64
 2   pm2_5                 20424 non-null  float64
 3   carbon_monoxide       20424 non-null  float64
 4   dust                  20424 non-null  float64
 5   temperature_2m        20424 non-null  float64
 6   relative_humidity_2m  20424 non-null  float64
 7   wind_speed_10m        20424 non-null  float64
 8   precipitation         20424 non-null  float64
dtypes: datetime64[us, Asia/Bangkok](1), float64(8)
memory usage: 1.4 MB
None
(20424, 9)
```

C3 · Missing values

Report the percentage missing for every column. If any column is exactly 0.00% missing across years of data, explain what that implies about where the data came from.

C3

ทุกดาต้ามี missing 0% เป็นข้อมูลจาก Reanalysis/Atmospheric Model (Copernicus CAMS & ERA5) ซึ่งใช้คำนวณจากแบบจำลองทางคณิตศาสตร์และดาวเทียม จึงมีการประมวลผลเติมเต็มช่องว่าง (Grid Imputation) มาเรียบร้อยแล้วตั้งแต่ต้นทาง

	missing_count	missing_pct (%)
date	0	0.0
pm10	0	0.0
pm2_5	0	0.0
carbon_monoxide	0	0.0
dust	0	0.0

	missing_count	missing_pct (%)
date	0	0.0
temperature_2m	0	0.0
relative_humidity_2m	0	0.0
wind_speed_10m	0	0.0
precipitation	0	0.0

```
-------- C3: Missing Values Inspection --------
```

	missing_count	missing_pct (%)
date	0	0.0
pm10	0	0.0
pm2_5	0	0.0
carbon_monoxide	0	0.0
dust	0	0.0
temperature_2m	0	0.0
relative_humidity_2m	0	0.0
wind_speed_10m	0	0.0
precipitation	0	0.0

C4 · Comparing places

If your analysis compares two or more locations, demonstrate that they actually return different data before you draw any spatial conclusion. State the evidence.

C4

```
- ศูนย์ราชการจังหวัดเชียงใหม่  (ต.ช้างเผือก อ.เมือง, เชียงใหม่)
  วันที่: 2026-09-05 เวลา: 16:00  PM2.5 = 18.4 µg/m³

- โรงเรียนยุพราชวิทยาลัย  (ต.ศรีภูมิ อ.เมือง, เชียงใหม่)
  วันที่: 2026-09-05 เวลา: 16:00  PM2.5 = 11.2 µg/m³
```

จากค่าจริงจะเห็นได้ว่าแม้ว่าอยู่ในอำเภอเมืองเชียงใหม่เหมือนกัน ค่าฝุ่นก็ต่างกัน

C5 · Model versus baseline

State the baseline score and your model score side by side, on the same test set. If your cross-validation score and your test score disagree substantially, explain why.

C5

Regression (ทำนายค่าเฉลี่ย PM2.5 ของวันพรุ่งนี้):

Model	Test MAE	Test RMSE	R^2	beats_baseline
Baseline	5.679	7.752	0.530	-
Linear Regression	5.135	7.307	0.582	Yes
Random Forest	5.909	8.238	0.469	no

Classification (ทำนายว่าพรุ่งนี้เกิน 37.5 μg/m³ หรือไม่):

Model	Recall	Precision	F1	beats_baseline_recall
Baseline	0.000	0.000	0.000	-
Logistic Regression	0.906	0.362	0.518	yes
Random Forest Classifier	0.781	0.510	0.617	yes

เกิดจากปัญหา Overfitting ร่วมกับ Time-series Concept Drift เนื่องจาก Random Forest เป็นโมเดลที่มีความซับซ้อนสูง จึงจดจำรูปแบบ Noise ในข้อมูลอดีต (Train Set) มากเกินไป เมื่อนำมาทดสอบกับข้อมูลช่วงเวลาใหม่ใน Test Set (ปี 2025) ที่มีสภาพอากาศและรูปแบบฝุ่นเปลี่ยนไป โมเดลจึงไม่สามารถปรับตัวได้ดี (Poor Generalization) ในขณะที่ Linear Regression เป็นโมเดลที่ไม่ซับซ้อน (Simpler Model) จึงไม่เกิด Overfitting และให้ผลลัพธ์บน Test Set ที่มั่นคงและดีกว่า Baseline ชัดเจน

C6 · Ground truth

Compare your data against Air4Thai station readings for at least one point in time. Report the difference. If they disagree, which one is right, and what does that mean for your conclusions?

C6
เมื่อเวลา 16:00 น. ของวันที่ 5 กันยายน 2569 Open-Meteo รายงานค่า PM2.5 ที่ 5.7 μg/m³ ในขณะที่สถานีตรวจวัดภาคพื้นดินของ Air4Thai บันทึกค่าได้ 18.4 μg/m³ (ช้างเผือก) และ 11.2 μg/m³ (โรงเรียนยุพราช) Air4Thai มีความแม่นยำมากกว่าสำหรับสภาพในระดับท้องถิ่น เนื่องจากเป็นการตรวจวัดทางกายภาพโดยตรง Open-Meteo แสดงค่าเฉลี่ยของตารางเชิงพื้นที่ ทำให้เกิดการประเมินค่าต่ำกว่าความเป็นจริงอย่างเป็นระบบ แม้ว่าค่าเกณฑ์สัมบูรณ์ (> 37.5 μg/m³) ใน Open-Meteo อาจประเมินจำนวนวันที่มลพิษรุนแรงต่ำกว่าความเป็นจริง แต่ความสัมพันธ์เชิงสัมพัทธ์ที่ระบุ (เช่น ความสำคัญของความชื้นและความเร็วลม) และแนวโน้มตามฤดูกาลยังคงใช้ได้สำหรับการวิเคราะห์ทางสถิติ - ศูนย์ราชการจังหวัดเชียงใหม่ (ต.ช้างเผือก อ.เมือง, เชียงใหม่) วันที่: 2026-09-05 เวลา: 16:00 PM2.5 = 18.4 μg/m³ - โรงเรียนยุพราชวิทยาลัย (ต.ศรีภูมิ อ.เมือง, เชียงใหม่) วันที่: 2026-09-05 เวลา: 16:00 PM2.5 = 11.2 μg/m³ time pm2_5 16 2026-09-05T16:00 5.7

5. Conclusion

ความชื้นสัมพัทธ์ต่ำ (อากาศแห้ง r = -0.58) เป็นปัจจัยทางอุตุนิยมวิทยาหลักที่เอื้อต่อการสะสมตัวของฝุ่น PM2.5 ในเชียงใหม่ ส่วนความเร็วลมแม้มีความแตกต่างอย่างมีนัยสำคัญทางสถิติ (p = 0.0026) แต่ทิศทางกลับตรงข้ามกับสมมติฐาน คือลมมีแนวโน้มแรงขึ้นเล็กน้อยในวันฝุ่นหนัก ไม่ใช่ลมสงบ จึงปฏิเสธสมมติฐานเรื่องลมสงบ เช่นเดียวกับที่ปฏิเสธสมมติฐานเรื่องอุณหภูมิ (p > 0.05)

อย่างไรก็ตาม ความสัมพันธ์เชิงสถิตินี้ไม่ควรถูกตีความเป็นความสัมพันธ์เชิงเหตุ-ผลที่นำไปสู่มาตรการแทรกแซงสภาพอากาศได้โดยตรง กล่าวคือ ความชื้นสัมพัทธ์ที่พบว่าสัมพันธ์กับ PM2.5 ต่ำเป็นความชื้นระดับมวลอากาศ (synoptic scale) ที่มากับระบบอากาศ ไม่ใช่ความชื้นระดับพื้นที่ที่มาตรการเชิงกล เช่น การฉีดพ่นละอองน้ำ จะสามารถเปลี่ยนแปลงได้ในสเกลที่มีนัยสำคัญ ข้อค้นพบนี้จึงถูกนำไปใช้ในการตัดสินใจเลือกทิศทางของข้อเสนอแนะในหัวข้อถัดไป

5.1. ข้อเสนอแนะเชิงนโยบายและการนำไปใช้จริง

- กลุ่มเป้าหมาย (Target Audience) : ผู้ว่าราชการจังหวัด และ ส่วนสำนักงานสาธารณสุข

- สิ่งที่จะแนะนำ: เสนอให้ผู้ว่าฯ สั่งการจัดตั้ง "ระบบแจ้งเตือนภัยและมาตรการควบคุมฝุ่นล่วงหน้า 24 ชั่วโมง (24-Hour Preemptive Dust Control System)" โดยใช้ผลการพยากรณ์จากโมเดล classification เป็นตัวสั่งการ (ไม่ใช่ threshold ของค่าอากาศดิบ):
  - เมื่อโมเดล Logistic Regression (recall 0.906) ทำนายว่าพรุ่งนี้มีความเสี่ยงเกินมาตรฐาน ให้สาธารณสุขจังหวัดออกประกาศเตือนกลุ่มเสี่ยงทันที (soft alert) พร้อมกระจายหน้ากาก N95
  - เมื่อโมเดล Random Forest Classifier (precision 0.510) ยืนยันความเสี่ยงซ้ำ ให้ผู้ว่าฯ สั่งมาตรการที่กระทบวงกว้างกว่า เช่น ปิดโรงเรียนหรือเปิด Clean Air Shelter (hard alert)
  - การใช้สองโมเดลร่วมกันเป็นการถ่วงดุลระหว่าง recall สูง (ไม่พลาดวันอันตราย) กับ precision ที่น่าเชื่อถือกว่าในการสั่งมาตรการหนัก
- สิ่งที่พิจารณาแล้วตัดออก: ในช่วงแรกของการวิเคราะห์ พิจารณามาตรการเพิ่มความชื้น เช่น ฝนเทียมและการฉีดพ่นละอองน้ำ โดยอิงจากความสัมพันธ์ผกผันระหว่างความชื้นสัมพัทธ์กับ PM2.5 ($r = -0.58$) แต่ตัดออกจากข้อเสนอแนะ เนื่องจาก (1) ฝนเทียมต้องการความชื้นสัมพัทธ์สูงเป็นเงื่อนไขเบื้องต้น ซึ่งขัดแย้งกับสภาพอากาศแล้งที่เกิดวันฝุ่นหนักพอดี และ (2) การฉีดพ่นละอองน้ำระดับถนนไม่มีสเกลเพียงพอที่จะเปลี่ยนความชื้นสัมพัทธ์ระดับหุบเขา/เมืองได้จริง
- สิ่งที่ analysis สนับสนุนจริง กับสิ่งที่เป็นการ extrapolate: ข้อมูลสนับสนุนว่า (1) ความชื้นสัมพัทธ์ต่ำสัมพันธ์กับ PM2.5 สูงอย่างมีนัยสำคัญ ($r = -0.58$, $p < 0.0001$) และ (2) โมเดล classification พยากรณ์วันเกินมาตรฐานล่วงหน้าได้ดีกว่า baseline อย่างชัดเจน (recall 0.906 เทียบกับ 0.000 ของ majority-class baseline) อย่างไรก็ตาม การตีความว่าความชื้นเป็น "สาเหตุ" ที่แทรกแซงได้ เป็นการขยายผลเกินกว่าที่ข้อมูล correlational จะรับรอง จึงใช้ข้อค้นพบนี้เพียงเป็นตัวบ่งชี้ ไม่ใช่ lever ของมาตรการ เช่นเดียวกัน ตัวเลข precision/recall คำนวณจากข้อมูลปี 2023–2025 เท่านั้น การนำไปใช้ในปีที่รูปแบบการเผา/ไฟป่าเปลี่ยนไปมาก ยังไม่มีหลักฐานยืนยันว่าโมเดลจะแม่นยำในระดับเดียวกัน
- what happens on the days your model is wrong?
  - หากใช้ Logistic Regression เป็นตัวเตือนหลัก (recall 0.906, precision 0.362) วันที่โมเดลผิดส่วนใหญ่จะเป็น false positive คือเตือนทั้งที่ฝุ่นไม่เกินจริง (ประมาณ 64% ของครั้งที่เตือน) ซึ่งมีต้นทุนคือประชาชนเริ่มไม่เชื่อการแจ้งเตือน (alert fatigue) แต่ยังคงจับวันอันตรายจริงได้เกือบทั้งหมด (พลาดเพียง ~9.4%) ในบริบทสุขภาพของกลุ่มเสี่ยง (เด็ก ผู้สูงอายุ ผู้ป่วยระบบทางเดินหายใจ) ต้นทุนของการพลาดวันอันตรายจริง (false negative) สูงกว่าต้นทุนของการเตือนเกินจำเป็น (false positive) มาก จึงเลือกโมเดลที่เอนไปทาง recall สูง แม้ precision จะต่ำ และใช้ Random Forest Classifier (precision 0.510) เป็นตัวกรองชั้นที่สองก่อนสั่งมาตรการที่มีต้นทุนสูง เช่น การปิดโรงเรียน

- What would you need that you do not have, in order to give a stronger recommendation?
  - หากในอนาคตต้องการประเมินมาตรการแทรกแซงสภาพอากาศ (เช่น ฝนเทียมหรือการเพิ่มความชื้นเชิงกล) อย่างจริงจัง จำเป็นต้องมีข้อมูลเชิงทดลอง (ก่อน-หลังการแทรกแซงจริง) ไม่ใช่เพียงค่าสหสัมพันธ์จากสถานีตรวจวัดสภาพอากาศ เนื่องจากค่าสหสัมพันธ์ที่พบในงานนี้ไม่สามารถยืนยันความสัมพันธ์เชิงเหตุ-ผลได้ นอกจากนี้ยังต้องการข้อมูลตรวจวัดจริงภาคพื้นดินหลายพิกัด (multi-station ground-truth) เพื่อแก้ปัญหา micro-climate ในเขตเมือง และข้อมูลจุดความร้อน (hotspot) ย้อนหลังมากกว่า 2.5 ปี

6. Use of AI tools

Where you used it	What you asked for	What you changed or rejected
src/analyse.py — กราฟเปรียบเทียบวันฝุ่นสูง vs วันปกติ	คำแนะนำว่าควรใช้กราฟแบบไหน	แนะนำให้ใช้ scatterplot เอา แต่รู้สึกว่ามันดูไม่ค่อยรู้เรื่อง เลยเปลี่ยนมาใช้เป็น boxplot แทน
src/model.py — ออกแบบ feature	คำแนะนำเรื่อง feature ที่ใช้ได้กับงานทำนาย	ปรับ feature list
Checkpoint C6 (ground truth)	วิธีเทียบ Open-Meteo กับ Air4Thai	แก้ไขเองว่าต้องดึง Open-Meteo ของวันเดียวกับที่ดึง Air4Thai ไม่ใช่ใช้ข้อมูลเก่าที่มีอยู่แล้ว (ตอนแรกเข้าใจผิดว่าใช้ข้อมูลเดิมได้)

7. REFERENCES

Assi, S., et al. (2025). Assessment of urinary oxidative stress biomarkers associated with fine particulate matter (PM2.5) exposure in Chiang Mai, Thailand. PeerJ, 13, e19047. https://doi.org/10.7717/peerj.19047

IQAir. (2026, April 16). Chiang Mai among top 10 most polluted cities in the world. IQAir Newsroom. https://www.iqair.com/th/newsroom/chiang-mai-among-top-10-most-polluted-cities-in-the-world-4-16-2026

Waikare, A., et al. (2024). Impact of fine particulate matter (PM2.5) on ocular health among people living in Chiang Mai, Thailand. Scientific Reports, 14, Article 27288. https://doi.org/10.1038/s41598-024-77288-8