

การวิเคราะห์ความสัมพันธ์ระหว่างระดับฝุ่น PM 2.5 และอัตราการเกิดโรคทางเดินหายใจใน เขตสุขภาพที่ 8 ของประเทศไทยด้วยเทคนิค Machine Learning

วสิน โพธิ์ทอง^{1*}, ณัฐกานต์ กลองกระโทก¹ และกัลยณัฐ กุหลาบเพชรทอง¹

¹สาขาวิทยาการคอมพิวเตอร์และนวัตกรรมข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนสุนันทา

*s65122250049@ssru.ac.th

บทคัดย่อ

บทความนี้มีวัตถุประสงค์เพื่อวิเคราะห์ความสัมพันธ์ระหว่างระดับฝุ่นละอองขนาดเล็ก PM 2.5 และอัตราการเกิดโรคทางเดินหายใจในเขตสุขภาพที่ 8 ซึ่งประกอบด้วยจังหวัดในภาคตะวันออกเฉียงเหนือของประเทศไทย ได้แก่ เลย หนองคาย หนองบัวลำภู อุดรธานี นครพนม และสกลนคร การศึกษาใช้ข้อมูลจากแหล่งทุติยภูมิที่เกี่ยวข้องกับคุณภาพอากาศและค่าฝุ่น PM 2.5 จากกรมควบคุมมลพิษ รวมถึงข้อมูลผู้ป่วยโรคทางเดินหายใจจากฐานข้อมูล Health Data Center (HDC) กระทรวงสาธารณสุข ระหว่างวันที่ 1 มกราคม 2566 ถึง 31 ธันวาคม 2566

การวิเคราะห์ข้อมูลใช้เทคนิค Machine Learning (ML) เพื่อหาความสัมพันธ์ระหว่างค่าฝุ่น PM 2.5 และอัตราการเกิดโรคทางเดินหายใจ โดยใช้สถิติเชิงพรรณนาในการอธิบายสถานการณ์ฝุ่นละอองและอัตราการเกิดโรคในแต่ละพื้นที่ และการใช้ ML ในการสร้างโมเดลเพื่อวิเคราะห์ปัจจัยที่มีผลต่อการเกิดโรคทางเดินหายใจในช่วงเวลาต่างๆ

ผลการศึกษาจะช่วยเพิ่มความเข้าใจเกี่ยวกับผลกระทบของฝุ่น PM 2.5 ต่อสุขภาพของประชาชนในพื้นที่และเป็นข้อมูลสำคัญสำหรับการวางแผนและกำหนดมาตรการในการเฝ้าระวัง ป้องกัน และแก้ไขปัญหาหมอกพิษทางอากาศในเขตสุขภาพที่ 8 อย่างมีประสิทธิภาพ

คำสำคัญ: ระดับฝุ่น PM 2.5 โรคทางเดินหายใจ คุณภาพอากาศ การเรียนรู้ของเครื่องจักร

Analyzing the Relationship between PM 2.5 Levels and Respiratory Disease Incidence in Health Zone 8 of Thailand Using Machine Learning Techniques

Wasin Phothong^{1*}, Nattakarn Klongkratok¹ and Kunyanuth Kularbphetong¹

¹Computer Science and Data Innovation Program, Faculty of Science and Technology

Suan Sunandha Rajabhat University

*s65122250049@ssru.ac.th

Abstract

This study aims to examine the correlation between fine particulate matter (PM 2.5) concentrations and the prevalence of respiratory disorders in Health Zone 8, encompassing provinces in northeastern Thailand, including Loei, Nong Khai, Nong Bua Lam Phu, Udon Thani, Nakhon Phanom, and Sakon Nakhon. This research utilizes secondary data sources, including air quality and PM 2.5 levels from the Pollution Control Department, as well as respiratory disease patient data from the Health Data Center (HDC) of the Ministry of Public Health, spanning from January 1, 2023, to December 31, 2023.

The data analysis employs Machine Learning (ML) methodologies to investigate the correlation between PM 2.5 concentrations and the prevalence of respiratory illnesses. Descriptive statistics elucidate the status of PM 2.5 pollution and respiratory disease prevalence, whereas machine learning approaches are utilized to construct models that examine the determinants influencing respiratory disease incidence across various regions and temporal contexts.

The results of this research will improve comprehension of PM 2.5 pollution's effects on public health and furnish critical data for devising and implementing effective strategies for monitoring, preventing, and mitigating air pollution challenges in Health Zone 8.

Keywords: PM 2.5, Respiratory Disease, Air Quality, machine learning

1. บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ฝุ่นละออง (Particulate Matter, PM) ที่อยู่ในบรรยากาศมีขนาดอนุภาค และองค์ประกอบของสารหลายชนิดแตกต่างกันไป [1] และมีอิทธิพลต่อการเปลี่ยนแปลงสภาพบรรยากาศ ก่อให้เกิดผลกระทบต่อสิ่งแวดล้อม รวมถึงสุขภาพของมนุษย์ และสิ่งมีชีวิตต่าง ๆ ทำให้มีการเจ็บป่วย และเสียชีวิตเพิ่มขึ้น [2-3] ความเข้มข้นของฝุ่น PM2.5 ที่เพิ่มสูงขึ้นได้กลายเป็นประเด็นทางสังคมที่มีความสัมพันธ์อย่างมีนัยสำคัญกับการขยายตัวของเมืองทั่วโลกมากกว่า 135 ประเทศ อีกทั้งยังเกี่ยวข้องกับพัฒนาการทางเศรษฐกิจ และอุตสาหกรรมที่ก่อให้เกิดการสะสมของฝุ่นละอองในบรรยากาศ [4-5] ปัจจุบันองค์การอนามัยโลก (WHO) และหน่วยงานสาธารณสุขในหลายประเทศยืนยันว่าการสัมผัส PM2.5 ถือเป็นปัจจัยเสี่ยงสำคัญที่

ส่งผลต่อสุขภาพโดยเฉพาะในระบบทางเดินหายใจ [6] และมีแนวโน้มส่งผลต่อการเสียชีวิตในระยะยาวจากโรคเรื้อรังต่าง ๆ เช่น โรคหัวใจ โรคปอด และมะเร็งปอด [7-8]

ในประเทศไทย เขตสุขภาพที่ 8 ซึ่งครอบคลุมจังหวัดในภาคตะวันออกเฉียงเหนือตอนบน เช่น หนองคาย อุดรธานี หนองบัวลำภู และนครพนม เป็นพื้นที่ที่มีความเสี่ยงสูงต่อฝุ่นละออง PM_{2.5} โดยเฉพาะในช่วงฤดูแล้งที่มีการเผาในที่โล่ง และชีวมวลเพื่อเตรียมพื้นที่ทางการเกษตร ซึ่งส่งผลให้ระดับฝุ่นเกินค่ามาตรฐานเป็นประจำขะเดียวกัน ข้อมูลจากศูนย์ข้อมูลสุขภาพของ กระทรวงสาธารณสุข (2565) พบว่าในช่วงที่ค่าฝุ่น PM_{2.5} สูง จำนวนผู้ป่วยโรคระบบทางเดินหายใจมีแนวโน้มเพิ่มขึ้นอย่างชัดเจน ซึ่งรวมถึงโรคหลอดลมอักเสบ ปอดบวม และโรคหอบหืด ปัจจัยเหล่านี้แสดงให้เห็นถึงความจำเป็นในการศึกษาเชิงลึกเกี่ยวกับความสัมพันธ์ของมลพิษทางอากาศกับสุขภาพในพื้นที่ดังกล่าว อย่างไรก็ตาม งานวิจัยในอดีตส่วนมากมักจำกัดอยู่ที่การวิเคราะห์เชิงสถิติพื้นฐาน และยังไม่ได้นำเทคนิคด้านวิทยาการข้อมูลและการเรียนรู้ของเครื่อง (Machine Learning) มาใช้เพื่อให้ได้ผลวิเคราะห์ที่แม่นยำและมีความสามารถในการคาดการณ์แนวโน้มได้

จากความสำคัญของปัญหาดังกล่าว งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างระดับฝุ่นละออง PM_{2.5} กับอัตราการเกิดโรคทางเดินหายใจในเขตสุขภาพที่ 8 ของประเทศไทย โดยใช้ข้อมูลรายเดือนจากกรมควบคุมโรค (DDC OpenData) และข้อมูลผู้ป่วยจากกระทรวงสาธารณสุขในช่วงปี 2563–2565 มาวิเคราะห์ด้วยเทคนิค Machine Learning เพื่อค้นหารูปแบบความสัมพันธ์ที่อาจซับซ้อนและไม่สามารถตรวจจับได้ด้วยการวิเคราะห์ทางสถิติแบบดั้งเดิม ผลการวิจัยที่ได้จะสามารถประยุกต์ใช้ในการเฝ้าระวังสุขภาพของประชาชนในพื้นที่เสี่ยง และวางแผนเชิงนโยบายเพื่อป้องกันและลดผลกระทบจากมลพิษทางอากาศในอนาคตได้อย่างมีประสิทธิภาพ

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาระดับค่าฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM_{2.5}) และจำนวนผู้ป่วยโรคระบบทางเดินหายใจในพื้นที่เขตสุขภาพที่ 8 ของประเทศไทย ระหว่างปี พ.ศ. 2566–2567

1.2.2 เพื่อวิเคราะห์ความสัมพันธ์ระหว่างค่าฝุ่น PM_{2.5} กับจำนวนผู้ป่วยโรคทางเดินหายใจในแต่ละจังหวัดของเขตสุขภาพที่ 8

1.2.3 เพื่อสร้างแบบจำลองทาง Machine Learning ในการพยากรณ์แนวโน้มจำนวนผู้ป่วยโรคทางเดินหายใจจากข้อมูลค่าฝุ่น PM_{2.5}

1.2.4 เพื่อเปรียบเทียบประสิทธิภาพของโมเดลการพยากรณ์ระหว่าง Linear Regression และ K-Nearest Neighbors (KNN) และ XGBoost

1.2.5 เพื่อเสนอแนวทางในการใช้ข้อมูลจากแบบจำลองสำหรับการเฝ้าระวังและวางแผนป้องกันผลกระทบจากฝุ่น PM_{2.5} ต่อสุขภาพประชาชนในพื้นที่เสี่ยง

1.3 ขอบเขตของการวิจัย

1.3.1 ขอบเขตด้านเนื้อหา

การวิจัยมุ่งเน้นศึกษาความสัมพันธ์ระหว่างระดับค่าฝุ่นละออง PM_{2.5} กับอัตราการเกิดโรคระบบทางเดินหายใจ โดยใช้เทคนิค Machine Learning ในการวิเคราะห์ข้อมูลเชิงลึกและการพยากรณ์แนวโน้มการเกิดโรคจากค่ามลพิษฝุ่น PM_{2.5} โดยจำกัดกลุ่มโรคที่เกี่ยวข้อง เช่น โรคหอบหืด โรคหลอดลมอักเสบ ปอดบวม โรคปอดอุดกั้นเรื้อรัง และไข้หวัดใหญ่

1.3.2 ขอบเขตด้านพื้นที่ : จำกัดพื้นที่ในการศึกษาภายใน เขตสุขภาพที่ 8 ของประเทศไทย ซึ่งประกอบด้วย 7 จังหวัด ได้แก่ บึงกาฬ หนองบัวลำภู อุดรธานี เลย หนองคาย สกลนคร นครพนม

1.3.3 ขอบเขตด้านเวลา : ใช้ข้อมูลจากช่วง เดือนมกราคม 2566 ถึงเดือนเมษายน 2567 รวมระยะเวลา 16 เดือน สำหรับการวิเคราะห์สถานการณ์และการพยากรณ์

1.3.4 ขอบเขตด้านข้อมูล : ข้อมูล PM2.5 และจำนวนจุดความร้อน (hotspot) จากกรมควบคุมโรค (DDC OpenData) ข้อมูลผู้ป่วยโรคทางเดินหายใจ และรหัสโรค (ICD-10) จากระบบ Health Data Center (HDC) ของกระทรวงสาธารณสุขรูปแบบข้อมูลเป็นรายเดือนในระดับจังหวัด

ข้อมูลโรคที่ใช้ในการวิเคราะห์มีการรวมกลุ่มโรคหลายประเภท บางโรค เช่น โรคตาอักเสบ และโรคผิวหนังอักเสบ อาจไม่สัมพันธ์โดยตรงกับการได้รับ PM2.5 ซึ่งอาจทำให้ค่าความสัมพันธ์ที่ได้มีอคติ (bias) จากปัจจัยร่วม (confounding factors) จึงควรมีการวิเคราะห์แยกโรคระบบทางเดินหายใจออกจากโรคอื่นอย่างชัดเจนในงานวิจัยต่อไป

1.3.5 ขอบเขตด้านเทคนิค : ใช้เครื่องมือภาษา Python และไลบรารีที่เกี่ยวข้อง เช่น pandas, seaborn, scikit-learn เพื่อวิเคราะห์ข้อมูลและสร้างแบบจำลอง Machine Learning โดยใช้โมเดล Linear Regression และ K-Nearest Neighbors (KNN) พร้อมเปรียบเทียบประสิทธิภาพด้วยค่า RMSE

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ช่วยทำนายพื้นที่ที่มีความเสี่ยงสูงต่อการเกิดโรคทางเดินหายใจจากระดับฝุ่น PM2.5 ในปีถัดไป

1.4.2 ช่วยให้หน่วยงานสาธารณสุขสามารถเตรียมความพร้อม และวางมาตรการป้องกันโรคได้อย่างมีประสิทธิภาพ

1.4.3 ลดความรุนแรงของผลกระทบด้านสุขภาพที่เกิดจากมลพิษทางอากาศ โดยเฉพาะในกลุ่มประชากรเปราะบาง

1.4.4 สร้างฐานข้อมูลเชิงวิเคราะห์สำหรับการวางแผนนโยบาย และพัฒนากลยุทธ์การเฝ้าระวังคุณภาพอากาศในระยะยาว

2. วิธีวิจัย/วิธีการศึกษา

2.1 ประชากร/กลุ่มเป้าหมาย

2.1.1 ประชากร

ประชากรในการศึกษานี้ คือ ข้อมูลรายเดือนของระดับฝุ่นละออง PM2.5 และจำนวนผู้ป่วยโรคทางเดินหายใจในพื้นที่เขตสุขภาพที่ 8 ของประเทศไทย ซึ่งประกอบด้วย 7 จังหวัด ได้แก่ บึงกาฬ หนองบัวลำภู อุดรธานี เลย หนองคาย สกลนคร และนครพนม โดยข้อมูลฝุ่น PM2.5 มาจากฐานข้อมูลของกรมควบคุมมลพิษ และข้อมูลผู้ป่วยมาจากระบบ Health Data Center (HDC) ของกระทรวงสาธารณสุข ครอบคลุมช่วงเวลาเดือนมกราคม 2566 ถึงเมษายน 2567

2.1.2 กลุ่มเป้าหมาย

กลุ่มเป้าหมายของการศึกษานี้ ได้แก่ ข้อมูลจากพื้นที่จังหวัดในเขตสุขภาพที่ 8 ซึ่งมีการรายงานค่าฝุ่นละออง PM2.5 และจำนวนผู้ป่วยโรคระบบทางเดินหายใจอย่างต่อเนื่องและครบถ้วน โดยเน้นกลุ่มโรคที่เชื่อมโยงกับผลกระทบจากมลพิษทางอากาศ เช่น โรคหอบหืด โรคปอดอุดกั้นเรื้อรัง โรคหลอดเลือดสมอง และโรคปอดบวม ทั้งนี้เพื่อใช้เป็นฐานข้อมูลในการวิเคราะห์ความสัมพันธ์และการพยากรณ์ความเสี่ยงในพื้นที่เป้าหมาย

2.2 เครื่องมือและการหาคุณภาพเครื่องมือ

2.2.1 เครื่องมือ เครื่องมือที่ใช้ในการวิจัยครั้งนี้ประกอบด้วย:

ชุดข้อมูลทุติยภูมิ (Secondary Data Sets)

1. ข้อมูลระดับฝุ่นละออง PM2.5 และจุดความร้อนจากเว็บไซต์ DDC OpenData และกรมควบคุมมลพิษ
2. ข้อมูลจำนวนผู้ป่วยโรคทางเดินหายใจจากระบบ Health Data Center (HDC) ของกระทรวงสาธารณสุข
3. ข้อมูลจำแนกโรคตามรหัส ICD-10 เพื่อใช้วิเคราะห์เชิงลึกเฉพาะกลุ่มโรคที่เกี่ยวข้องกับฝุ่นละออง

ซอฟต์แวร์และเครื่องมือประมวลผลข้อมูล

1. ภาษา Python พร้อมไลบรารีสำคัญ เช่น pandas, numpy, matplotlib, seaborn สำหรับการจัดการข้อมูลและสร้างภาพข้อมูล

2. โลบารรี scikit-learn สำหรับสร้างแบบจำลอง Machine Learning เช่น Linear Regression และ K-Nearest Neighbors (KNN)

3. เครื่องมือ Jupyter Notebook ใช้ในการพัฒนา วิเคราะห์ และนำเสนอผลการทดลองแบบ Interactive

เทคนิคการวิเคราะห์ข้อมูล

1. การวิเคราะห์ความสัมพันธ์ด้วยค่าสหสัมพันธ์ (Correlation Analysis)

2. การพยากรณ์ด้วยโมเดล Machine Learning โดยใช้วิธี Leave-One-Out Cross Validation (LOOCV) เพื่อประเมินคุณภาพโมเดล

2.2.2 การหาคุณภาพเครื่องมือ

เพื่อให้มั่นใจในคุณภาพของเครื่องมือที่ใช้ในการวิจัย มีการดำเนินการดังนี้:

2.2.2.1 การตรวจสอบความครบถ้วนและความถูกต้องของข้อมูล

ตรวจสอบความถูกต้องของข้อมูลด้วยการแปลงและล้างข้อมูล (Data Cleaning) เช่น การจัดรูปแบบวันที่ การเติมค่าที่หายไป (Missing Values) การลบค่าผิดปกติ และการรวมข้อมูลรายเดือนให้เป็นโครงสร้างที่วิเคราะห์ได้ง่าย

2.2.2.2 การประเมินความแม่นยำของโมเดลการพยากรณ์

ใช้ค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง (Root Mean Squared Error: RMSE) เพื่อวัดความแม่นยำของแบบจำลอง และเปรียบเทียบระหว่าง Linear Regression กับ KNN เพื่อตัดสินใจเลือกโมเดลที่เหมาะสมที่สุด

2.2.2.3 การยืนยันผลด้วยการสร้างภาพประกอบ (Visualization Validation)

สร้างกราฟเชิงเปรียบเทียบระหว่างค่าจริงและค่าที่คาดการณ์ เพื่อประเมินความแม่นยำของโมเดลเชิงภาพ รวมถึงใช้ heatmap ในการแสดงความสัมพันธ์ระหว่างตัวแปรต่าง ๆ เพื่อสนับสนุนข้อสรุปจากการวิเคราะห์เชิงตัวเลข

2.3 การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลในงานวิจัยนี้เป็นการรวบรวมข้อมูลทุติยภูมิ (Secondary Data) ที่มีอยู่แล้วจากแหล่งข้อมูลทางราชการและแหล่งข้อมูลออนไลน์ที่เชื่อถือได้ โดยมีรายละเอียดดังนี้:

1) ข้อมูลฝุ่น PM2.5 และจุดความร้อน (Hotspot)

ข้อมูลถูกนำมาจากเว็บไซต์ DDC OpenData และ กรมควบคุมมลพิษ (Pollution Control Department) ซึ่งเผยแพร่ข้อมูลรายวันของค่าฝุ่นละออง PM2.5 และจุดความร้อนในแต่ละจังหวัด ข้อมูลครอบคลุมช่วง เดือนมกราคม 2566 ถึงเดือนเมษายน 2567 โดยมีการแยกเก็บรายเดือนในรูปแบบไฟล์ .csv รวมทั้งหมด 16 ไฟล์ ตัวแปรที่เก็บ ได้แก่: วันที่เก็บข้อมูล province: จังหวัด pm25: ค่าเฉลี่ยฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน hotspot: จำนวนจุดความร้อนต่อวัน month: เดือนที่เก็บข้อมูล

ตารางที่ 1 PM 2.5 file

Attribute	คำอธิบาย	ประเภทข้อมูล
date	วันที่เก็บข้อมูล	Datetime
Province	จังหวัดที่วัดค่า	Object
Pm25	ค่า PM2.5 (ค่าฝุ่นละออง)	Float
Hotspot	จำนวนจุดความร้อน (hotspot)	Float
month	เดือนที่เก็บข้อมูล	Int

2) ข้อมูลจำนวนผู้ป่วยโรคทางเดินหายใจ

ข้อมูลนี้รวบรวมจาก ระบบ Health Data Center (HDC) ของ กระทรวงสาธารณสุข ซึ่งแสดงจำนวนผู้ป่วยรายเดือนแยกตามจังหวัดและโรคที่เกี่ยวข้องกับระบบทางเดินหายใจ ครอบคลุมช่วง ปี 2566 ถึงเมษายน 2567 โดยมีทั้งไฟล์รายจังหวัดและไฟล์รวมทั้งเขตสุขภาพที่ 8

ตัวแปรที่เก็บ ได้แก่: province: จังหวัด a_name: ชื่ออำเภอ result1 ถึง result12: จำนวนผู้ป่วยในแต่ละเดือน result: จำนวนผู้ป่วยรวม F3: อัตราป่วยต่อแสนประชากร target: จำนวนประชากรกลุ่มเป้าหมายในพื้นที่

ตารางที่ 2 respiratory file

Attribute	คำอธิบาย	ประเภทข้อมูล
a_name	ชื่ออำเภอ	Object
province	จังหวัด	Object
target (B)	จำนวนประชากรในพื้นที่	Int
result (A)	จำนวนผู้ป่วยโรคทางเดินหายใจ	Int
F3	อัตราการป่วยต่อประชากร (per 100,000)	Float
result1 - result12	จำนวนผู้ป่วยในแต่ละเดือน (ม.ค. - ธ.ค.)	Int

3) ข้อมูลจำแนกตามรหัสโรค (ICD-10)

ข้อมูลนี้ใช้สำหรับการวิเคราะห์โรคเฉพาะกลุ่ม เช่น โรคระบบทางเดินหายใจ โรคหัวใจ โรคตาอักเสบ และโรคผิวหนัง โดยใช้รหัส ICD-10 ที่ได้จากระบบ HDC และแยกเก็บเป็นรายจังหวัด รวมทั้งมีไฟล์รวมทั้งมดดังตารางที่ 3

ตารางที่ 3 แสดงข้อมูลจำแนกตามรหัสโรค

Attribute	คำอธิบาย	ประเภทข้อมูล
กลุ่มโรค	ชื่อของกลุ่มโรค	Object
ทั้งหมด (รายโรค)	จำนวนผู้ป่วยทั้งหมด	Int
Jan	จำนวนผู้ป่วยในเดือน มกราคม	Int
Feb	จำนวนผู้ป่วยในเดือน กุมภาพันธ์	Int
Mar	จำนวนผู้ป่วยในเดือน มีนาคม	Int
Apr	จำนวนผู้ป่วยในเดือน เมษายน	Int
May	จำนวนผู้ป่วยในเดือน พฤษภาคม	Int
Jun	จำนวนผู้ป่วยในเดือน มิถุนายน	Int
Jul	จำนวนผู้ป่วยในเดือน กรกฎาคม	Int
Aug	จำนวนผู้ป่วยในเดือน สิงหาคม	Int
Sep	จำนวนผู้ป่วยในเดือน กันยายน	Int
Oct	จำนวนผู้ป่วยในเดือน ตุลาคม	Int
Nov	จำนวนผู้ป่วยในเดือน พฤศจิกายน	Int
Dec	จำนวนผู้ป่วยในเดือน ธันวาคม	Int

4) ขั้นตอนการจัดการข้อมูล

1. ดาวน์โหลดไฟล์ .csv จากแหล่งข้อมูลที่กล่าวถึงข้างต้น
2. รวมไฟล์รายเดือนเข้าเป็นชุดข้อมูลเดียวด้วยไลบรารี pandas
3. ทำความสะอาดข้อมูล เช่น ลบข้อมูลที่ซ้ำหรือว่าง, แปลงรูปแบบวันที่, แปลงค่าที่มีหน่วย (เช่น ,) เป็นตัวเลข
4. สร้างตัวแปรใหม่ เช่น year, month, month year เพื่อใช้ในการวิเคราะห์และพยากรณ์
5. จัดเก็บไฟล์ข้อมูลที่พร้อมใช้งานไว้ในโฟลเดอร์ clean data สำหรับการวิเคราะห์ต่อไป

2.4 การวิเคราะห์ข้อมูล

ในการวิเคราะห์ข้อมูลครั้งนี้ ใช้เทคนิคทางสถิติร่วมกับการเรียนรู้ของเครื่อง (Machine Learning) เพื่อศึกษาความสัมพันธ์ระหว่างค่าฝุ่น PM2.5 กับจำนวนผู้ป่วยโรคทางเดินหายใจในเขตสุขภาพที่ 8 โดยมีขั้นตอนการวิเคราะห์ดังนี้:

1) การวิเคราะห์เชิงพรรณนา (Descriptive Analysis)

ใช้สถิติเบื้องต้นในการสรุปลักษณะของข้อมูล เช่น ค่ามัธยฐาน ค่าเฉลี่ย ค่าสูงสุด ต่ำสุด และส่วนเบี่ยงเบนมาตรฐาน เพื่ออธิบายแนวโน้มของระดับ PM2.5 และจำนวนผู้ป่วยในแต่ละเดือน รวมถึงการสร้างกราฟแท่งและกราฟเส้น แสดงการเปรียบเทียบค่าเฉลี่ย PM2.5 และจำนวนผู้ป่วยรายเดือน

2) การหาความสัมพันธ์ (Correlation Analysis)

ใช้การคำนวณค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) เพื่อตรวจสอบระดับความสัมพันธ์ระหว่าง ค่า PM2.5 และจำนวนผู้ป่วยรวม จำนวนจุดความร้อน (Hotspot) และจำนวนผู้ป่วย ค่า PM2.5 กับแต่ละกลุ่มโรค เช่น โรคปอดบวม หอบหืด และหลอดเลือดสมอง โดยผลลัพธ์จะแสดงออกมาในรูปแบบ Heatmap เพื่อให้เข้าใจง่ายและชัดเจนยิ่งขึ้น

3) การสร้างแบบจำลอง Machine Learning เพื่อวิเคราะห์และพยากรณ์แนวโน้มจำนวนผู้ป่วยจากระดับ PM2.5 โดยใช้ 2 โมเดลหลัก ได้แก่:

Linear Regression: ใช้สำหรับจำลองความสัมพันธ์เชิงเส้นระหว่างตัวแปร

K-Nearest Neighbors (KNN) Regression: ใช้สำหรับจำลองความสัมพันธ์เชิงไม่เชิงเส้น โดยพิจารณาข้อมูลใกล้เคียง

ขั้นตอนที่ดำเนินการ:

1. แบ่งชุดข้อมูลเป็นข้อมูลฝึก (Training) และข้อมูลทดสอบ (Testing)
2. ประเมินผลลัพธ์ของแต่ละโมเดลด้วยค่า RMSE (Root Mean Squared Error)
3. เลือกโมเดลที่มีค่าความคลาดเคลื่อนต่ำที่สุดเพื่อใช้ในการพยากรณ์

4) การพยากรณ์แนวโน้ม

ใช้โมเดลที่ดีที่สุด (Best Model) จากการประเมิน เพื่อนำไปพยากรณ์จำนวนผู้ป่วยและค่า PM2.5 สำหรับช่วงเดือนพฤษภาคม-ธันวาคม ปี 2567 โดยแสดงผลผ่านกราฟเส้นพยากรณ์ พร้อมเปรียบเทียบกับข้อมูลจริง (เดือนมกราคม-เมษายน 2567)

5) การวิเคราะห์เชิงสรุปและอภิปรายผล

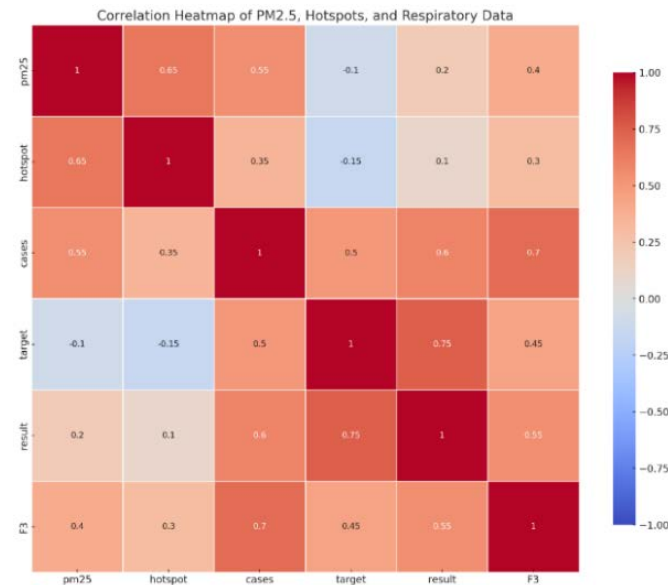
สรุปผลการวิเคราะห์จากค่า Correlation และผลการพยากรณ์ เช่น จังหวัดหรือช่วงเวลาที่มีความสัมพันธ์สูง กลุ่มโรคที่ได้รับผลกระทบจากฝุ่น PM2.5 มากที่สุด รูปแบบแนวโน้มการเปลี่ยนแปลงของผู้ป่วยตามระดับ PM2.5

3. ผลการวิจัยและอภิปรายผล

3.1 การวิเคราะห์ความสัมพันธ์ด้วย Heatmap

การวิเคราะห์ความสัมพันธ์เชิงสถิติเบื้องต้นได้ใช้เทคนิค การคำนวณค่าสหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) เพื่อตรวจสอบความสัมพันธ์ระหว่างตัวแปรสำคัญ ได้แก่ ค่า PM2.5, จำนวนจุดความร้อน (Hotspot), จำนวนผู้ป่วย (cases), จำนวนประชากรกลุ่มเป้าหมาย (target), จำนวนผู้ป่วยรวม(result) และอัตราป่วยต่อแสน (F3) โดยใช้ไลบรารี pandas, seaborn และ matplotlib ของภาษา Python ในการประมวลผลข้อมูลและสร้างภาพข้อมูล

จากการรวบรวมและจัดรูปแบบข้อมูลในแต่ละเดือนและจังหวัดในเขตสุขภาพที่ 8 สามารถสร้าง Heatmap ที่แสดงความสัมพันธ์ระหว่างตัวแปรทั้งหมดได้ดังนี้:

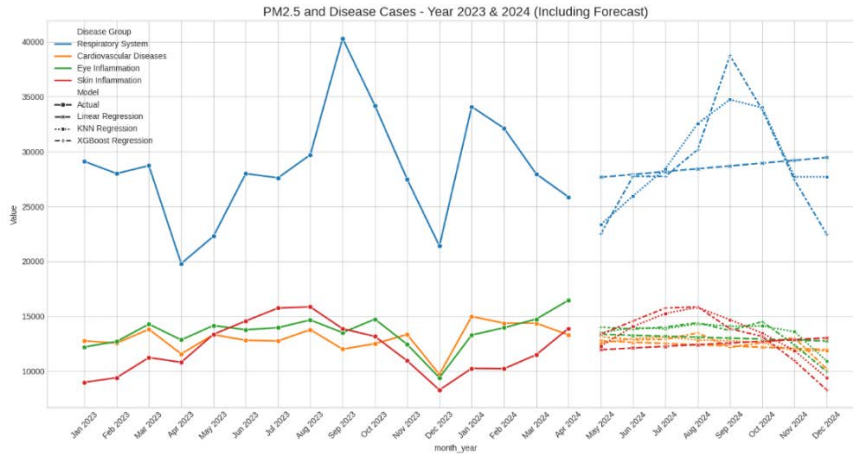


ภาพที่ 1 Heatmap แสดงค่าสหสัมพันธ์ระหว่างระดับฝุ่นละออง PM2.5 จุดความร้อน จำนวนผู้ป่วย และอัตราป่วยในพื้นที่เขตสุขภาพที่ 8 ที่มา: ผู้วิจัยประมวลผลข้อมูลโดยใช้เครื่องมือ Google Colab (2025)

3.2 การวิเคราะห์แนวโน้มและการพยากรณ์โรคด้วยโมเดล Machine Learning

ในการศึกษานี้ได้ใช้ข้อมูลจำนวนผู้ป่วยใน 4 กลุ่มโรคหลัก ได้แก่ โรคระบบทางเดินหายใจ โรคหัวใจและหลอดเลือด โรคตาอักเสบและโรคผิวหนังอักเสบ จากปี 2566 ถึงเดือนเมษายน 2567 และนำข้อมูลดังกล่าวมาสร้างแบบจำลองเพื่อพยากรณ์จำนวนผู้ป่วยล่วงหน้าในช่วงเดือนพฤษภาคม-ธันวาคม 2567 ด้วยเทคนิคการเรียนรู้ของเครื่อง 2 แบบ ได้แก่ Linear Regression และ K-Nearest Neighbors

(KNN) Regression ผลลัพธ์จากการประเมินด้วยค่ารากของค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Squared Error: RMSE) พบว่า KNN Regression มีค่าความคลาดเคลื่อนต่ำกว่าสำหรับกลุ่มโรคระบบทางเดินหายใจและกลุ่มโรคตาอักเสบ ในขณะที่ Linear Regression ให้ผลที่แม่นยำกว่าในบางกลุ่มโรคเช่น โรคหัวใจโดยได้แสดงแนวโน้มของจำนวนผู้ป่วยจริงร่วมกับผลลัพธ์จากแบบจำลองทั้งสองในรูปแบบกราฟเส้น ดังนี้:



ภาพที่ 2 กราฟแสดงแนวโน้มจำนวนผู้ป่วยรายเดือนในกลุ่มโรคทางเดินหายใจ โรคหัวใจ โรคตาอักเสบ และโรคผิวหนังอักเสบ ระหว่างปี 2566–2567 พร้อมการพยากรณ์ล่วงหน้าด้วยโมเดล Linear Regression และ KNN Regression ที่มา: ผู้วิจัยประมวลผลข้อมูลโดยใช้เครื่องมือ Google Colab (2025)

ตารางที่ 4 เปรียบโมเดลที่ใช้กับกลุ่มโรคต่างๆ

Disease Group	Model	RMSE
Respiratory System	Linear Regression	4635.85
Respiratory System	KNN Regression	3660.53
Respiratory System	XGBoost Regression	4169.90
Cardiovascular Diseases	Linear Regression	1322.76
Cardiovascular Diseases	KNN Regression	1470.55
Cardiovascular Diseases	XGBoost Regression	1672.10
Eye Inflammation	Linear Regression	1649.86
Eye Inflammation	KNN Regression	1797.23
Eye Inflammation	XGBoost Regression	1933.07
Skin Inflammation	Linear Regression	1345.85
Skin Inflammation	KNN Regression	1709.84
Skin Inflammation	XGBoost Regression	1724.88

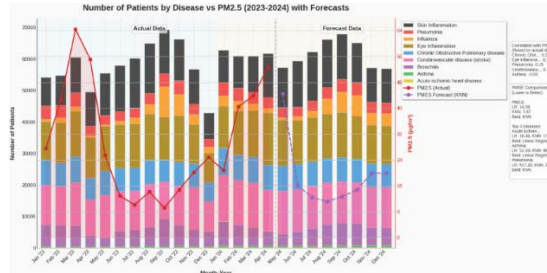
BestModelSummary

PM2.5: Best Model = KNN (6.30), Linear=14.58, KNN=6.30, XGBoost=7.59

KNN สรุปแนวโน้ม: ส่วนใหญ่ KNN ชอบใช้เพื่อนข้างเคียงน้อย ๆ (2–3) และใช้ weights='distance' เพื่อให้เพื่อนที่ใกล้กว่ามีผลมากกว่า สรุปแนวโน้ม: XGBoost ชอบใช้ colsample_bytree=0.8, learning_rate ต่ำ (0.05) และ max_depth ค่อนข้างตื้น (3–4) KNN ส่วนใหญ่เลือกจำนวนเพื่อนค่อนข้างน้อย (2–3) และนิยมใช้ weights='uniform' ซึ่งให้น้ำหนักเพื่อนทุกตัวเท่ากัน XGBoost มักเลือก learning_rate ต่ำ (0.05) ร่วมกับ max_depth ตื้น (4–5) และใช้การสุ่มคอลัมน์/แถว บางส่วนเพื่อลด overfitting

3.3 การพยากรณ์แนวโน้มผู้ป่วยและระดับ PM2.5 ด้วยโมเดล Machine Learning

จากการรวบรวมข้อมูลจำนวนผู้ป่วยแยกตามกลุ่มโรคในพื้นที่เขตสุขภาพที่ 8 และข้อมูลค่าฝุ่นละออง PM2.5 รายเดือนในช่วงปี พ.ศ. 2566 ถึงเดือนเมษายน พ.ศ. 2567 ผู้วิจัยได้ดำเนินการสร้างแบบจำลองเพื่อพยากรณ์แนวโน้มในช่วงเดือน พฤษภาคม-ธันวาคม พ.ศ. 2567 โดยใช้โมเดลการเรียนรู้ของเครื่อง ได้แก่ Linear Regression และ K-Nearest Neighbors (KNN) Regression โดยทำการประเมินความแม่นยำด้วยค่ารากของค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Squared Error: RMSE)



ภาพที่ 3 กราฟแสดงจำนวนผู้ป่วยรายเดือนตามกลุ่มโรค เปรียบเทียบกับค่า PM2.5 และผลการพยากรณ์ระหว่างปี พ.ศ. 2566–2567 ที่มา: ผู้วิจัยประมวลผลจากข้อมูลกรมควบคุมโรคและกรมควบคุมมลพิษ โดยใช้เครื่องมือ Google Colab (2025)

ตารางที่ 5 เปรียบโมเดลที่ใช้กับรายโรคต่างๆ

Disease	Best Model	Best RMSE	Linear RMSE	KNN RMSE	XGBoost RMSE
Chronic Obstructive Pulmonary disease	Linear Regression	408.31	408.31	412.47	440.95
Asthma	XGBoost Regression	48.79	52.03	68.3	48.79
Pneumonia	KNN Regression	356.29	517.33	356.29	399.75
Influenza	KNN Regression	1080.64	1817.92	1080.64	1205.47
Bronchitis	KNN Regression	783.41	1398.98	783.41	857.65
Acute ischemic heart disease	KNN Regression	15.27	16.46	15.27	15.34
Cerebrovascular disease (stroke)	XGBoost Regression	823.97	876.1	858.46	823.97
Eye Inflammation	XGBoost Regression	925.22	1393.15	989.86	925.22
Skin Inflammation	KNN Regression	1001.15	2247.3	1001.15	1151.94

Best Model Summary

PM2.5: Best Model = KNN (6.30), Linear=14.58, KNN=6.30, XGBoost=7.59

KNN สรุปแนวโน้ม: ส่วนใหญ่ KNN ชอบใช้เพื่อนข้างเคียงน้อย ๆ (2–3) และใช้ weights='distance' เพื่อให้เพื่อนที่ใกล้กว่ามีผลมากกว่า สรุปแนวโน้ม: XGBoost ชอบใช้ colsample_bytree=0.8, learning_rate ต่ำ (0.05) และ max_depth ค่อนข้างตื้น (3–4)

4. สรุปผล

การศึกษานี้วิเคราะห์ความสัมพันธ์ระหว่างระดับฝุ่นละออง PM2.5 กับจำนวนผู้ป่วยโรคระบบทางเดินหายใจและโรคอื่น ๆ ในเขตสุขภาพที่ 8 โดยใช้เทคนิค Machine Learning ได้แก่ Linear Regression, K-Nearest Neighbors (KNN) และ XGBoost Regression เพื่อพยากรณ์แนวโน้มในช่วงเดือนพฤษภาคม-ธันวาคม 2567 ผลการวิเคราะห์พบว่า: ค่า PM2.5 มีความสัมพันธ์ปานกลางกับโรคบางกลุ่ม เช่น COPD ($r = 0.32$), โรคตาอักเสบ ($r = 0.18$) โมเดล KNN ให้ผลแม่นยำที่สุดในการพยากรณ์ค่า PM2.5 (RMSE = 6.30) รองลงมาคือ XGBoost (RMSE = 7.59) และ Linear Regression (RMSE = 14.58) สำหรับโรคต่าง ๆ Linear Regression เหมาะสมกับโรค COPD และบางโรคหัวใจ KNN เหมาะสมกับโรค ปอดบวม, ไข้หวัดใหญ่, โรคผิวหนังอักเสบ, และ โรคหัวใจขาดเลือดเฉียบพลัน XGBoost เหมาะสมกับโรค หิด, โรคหลอดเลือดสมอง (Stroke) และโรคตาอักเสบ

แนวโน้มพารามิเตอร์ KNN มักเลือก $n_neighbors = 2-3$ และใช้ $weights='distance'$ XGBoost มักเลือก $colsample_bytree = 0.8$, $learning_rate = 0.05$ และ $max_depth = 3-4$ บางโรค เช่น โรคตาอักเสบ และโรคผิวหนังอักเสบ อาจไม่มีความสัมพันธ์เชิงสาเหตุโดยตรงกับ PM2.5 และอาจได้รับอิทธิพลจากปัจจัยอื่น ๆ การตีความค่าความสัมพันธ์ของโรคเหล่านี้จึงควรทำด้วยความระมัดระวัง ภาพรวมการพยากรณ์แสดงแนวโน้มผู้ป่วยเพิ่มขึ้นในช่วงที่ PM2.5 สูง โดยเฉพาะในฤดูแล้ง การใช้โมเดล Machine Learning ทั้ง 3 ช่วยให้คาดการณ์ล่วงหน้าได้อย่างมีประสิทธิภาพ และสามารถนำไปใช้ในการเฝ้าระวังและวางแผนเชิงนโยบายด้านสาธารณสุข

5. เอกสารอ้างอิง (References)

- [1] P. K. Rai, "Impacts of particulate matter pollution on plants: Implications for environmental biomonitoring," *Ecotoxicology and Environmental Safety*, vol. 129, pp. 120–136, 2016.
- [2] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: A review," *Frontiers in Public Health*, vol. 8, p. 14, 2020.
- [3] X. Wang, H. Ding, and M. Ryan, "The health effects of fine particulate air pollution: What do we know in 2020?" *Environment International*, vol. 120, pp. 245–266, 2018.
- [4] D. Yang, M. Gao, and S. Ji, "The global landscape of urban air pollution: China's comparative advantage," *Sustainability*, vol. 10, no. 5, p. 1430, 2018.
- [5] Q. Wang and M. Su, "A preliminary assessment of the impact of air pollution on China's economic development," *Journal of Cleaner Production*, vol. 206, pp. 118–129, 2019.
- [6] S. Chowdhury, S. Dey, and K. R. Smith, "Ambient PM2.5 exposure and risk of respiratory diseases in India: A comprehensive assessment," *The Lancet Planetary Health*, vol. 6, no. 3, pp. e156–e167, 2022.
- [7] S. E. Alexeeff, N. S. Liao, X. Liu, S. K. Van Den Eeden, and S. Sidney, "Long-term PM2.5 exposure and risks of ischemic heart disease and stroke events in the California Teachers Study," *Environmental Health Perspectives*, vol. 129, no. 4, p. 047003, 2021.
- [8] X. Yang, Y. Wu, H. Ma, et al., "Exposure to PM2.5 and lung cancer incidence and mortality: A global assessment," *Environment International*, vol. 158, p. 106962, 2022.
- [9] Geo-Informatics and Space Technology Development Agency (GISTDA). (2022). Report on Hotspot and Fine Particulate Matter (PM2.5) Situation in the Upper Northeastern Region of Thailand. [Online]. Available: <https://gistda.or.th/>



- [10] Department of Disease Control. (2023). Air Quality and PM2.5 Database for Health Zone 8 (Open Data). [Online]. Available: <https://opendata.ddc.moph.go.th/th/dataset/odpc8-01>
- [11] Ministry of Public Health. (2023). Respiratory Disease Patient Data for Health Zone 8, 2023–2024. [Online]. Available: <https://hdcservice.moph.go.th/hdc/>
- [12] Ministry of Public Health. (2023). Patient Data by Disease Group Classified by ICD-10 for Health Zone 8, 2023–2024. [Online]. Available: <https://hdcservice.moph.go.th/hdc/>