การเรียนรู้ของเครื่องสำหรับการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

กชรัตน์ นฤพัฒน์ผจง[1*], นภา แซ่เบ๊[2]

บทคัดย่อ

สถานการณ์ปัจจุบันปัญหาฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5) เป็นปัญหาสำคัญของประเทศไทย งานวิจัยนี้มุ่งศึกษาการนำข้อมูลภาคอุตุนิยมวิทยามาใช้ร่วมกับเทคนิคการเรียนรู้ของเครื่อง เพื่อสร้างแบบจำลองเบื้องต้นที่ใช้สำหรับในการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ล่วงหน้า เพื่อให้มีความเข้าใจแนวโน้มของสถานการณ์ PM2.5 รวมถึงวางแผนการจัดการที่เหมาะสมในการรับมือปริมาณฝุ่นละอองขนาดเล็ก (PM2.5)ในอนาคต โดยในงานวิจัย มีการสร้างชุดข้อมูล โดยการนำข้อมูลจากแหล่งข้อมูลสาธารณะ 2 ชุดมารวมกัน ซึ่งโดยอาศัยการใช้สคริปต์ดึงข้อมูลจากหน้าเว็บไซต์ (Web Scraping) ดังนี้ 1. ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) นำมาจากเว็บไซต์ Berkeley Earth 2. ข้อมูลภาคอุตุนิยมวิทยา นำมาจากเว็บไซต์ Weather Underground โดยทำการดึงข้อมูลในช่วงวันที่ 1 มกราคม - 31 ธันวาคม 2562 และช่วง 1 มกราคม – 28 กันยายน ปี 2563 ซึ่งข้อมูลภาคอุตุนิยมวิทยานำมาจากสถานี IKRUNGTH3 ตั้งอยู่บริเวณ วิภาวดี 60 หลักสี่กรุงเทพมหานคร ประกอบด้วยตัวแปรที่สามารถส่งผลต่อค่า PM2.5 ได้แก่ อุณหภูมิ, จุดน้ำค้าง, ความชื้น, ทิศทางลม, ความเร็วลม ,ลมกระโชก และ ความกดอากาศ สำหรับแบบจำลองเบื้องต้นที่ใช้สำหรับในการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ล่วงหน้า ประกอบด้วยแบบจำลอง ทั้งหมด 4 รูปแบบ ได้แก่ LR (Linear Regression), SVR (Support Vector Regression) , XGBoost และ MLP (Multi-Layer Perceptron) โดยใช้ ค่าพารามิเตอร์เริ่มต้น จาก scikit-learn และทำการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมด จากผลการทดลองพบว่าแบบจำลอง LR- Linear Regression ผลลัพธ์ที่ดีที่สุด ทั้งในแง่ของความถูกต้องแม่นยำ และความคลาดเคลื่อนที่ต่ำลดลง โดยผลลัพธ์ที่ดีที่สุดมีค่า R2 : 0.9722 ,MAE :1.6832, RMSE : 2.4492 , MAPE(%) : 9.0302 ซึ่งเป็นแบบจำลองที่สร้างโดยอาศัยตัวแปรด้านค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 24 48 และ72 ชั่วโมงย้อนหลัง และตัวแปรด้านฤดูกาล (Season) เพิ่มจากตัวแปรด้านค่าฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง 1 6 12 และ 24 ชั่วโมงย้อนหลัง และข้อมูลภาคอุตุนิยมวิทยา

คำสำคัญ : การเรียนรู้ของเครื่อง, การทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5, ฝุ่นละอองขนาดเล็ก PM2.5

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 097-4795492 E-mail address: kojcharat.narupatpajong@g.swu.ac.th

MACHINE LEARNING MODEL FOR FORECASTING PARTICULATE MATTER CONCENTRATION (PM 2.5)

กชรัตน์ นฤพัฒน์ผจง1*, นภา แซ่เบ๊2

Abstract

The issue of particulate matter (PM2.5) pollution is escalating in Thailand. This research aims to investigate the utilization of industrial data in conjunction with machine learning techniques to create a preliminary model for predicting the concentration of PM2.5 in advance. The goal is to enhance the understanding of PM2.5 trends and to develop suitable management plans to address future PM2.5 levels. In this research, a dataset was created by combining information from two public sources using web scraping scripts, as follows: (1) the fine particulate matter (PM2.5) data extracted from the Berkeley Earth website; (2) meteorological data obtained from the Weather Underground website, specifically from the IKRUNGTH3 station, located near Vibhavadi Rangsit 60, Lak Si, Bangkok. The data spans from January 1 to December 31, 2018 and January 1 to September 28, 2019, and includes variables that may influence PM2.5 levels, such as temperature, dew point, humidity, wind direction, wind speed, gust speed, and atmospheric pressure. For the predictive model of fine particulate matter (PM2.5) levels, four models were employed: LR (Linear Regression), SVR (Support Vector Regression), XG Boost, and MLP (Multi-Layer Perceptron). The models were configured with default parameters from scikit-learn, and their performances were subsequently compared. The experimental results revealed that the LR - Linear Regression model exhibited the best outcomes in terms of accuracy and reduced errors. The optimal results included R2: 0.9722, MAE: 1.6832, RMSE: 2.4492, and MAPE (%): 9.0302. This model incorporated variables related to PM2.5 concentrations from the previous 1 6 12 and 24 hours, along with meteorological data. Additionally, it utilized variables related to the average PM2.5 concentrations in the past 24, 48, and 72 hours, as well as seasonal information (Season).

Keywords : Machine learning model, Forecasting particulate matter (PM2.5), Linear Regression, Support Vector Regression, XGBoost, Multi-Layer Perceptron

---

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 097-4795492 E-mail address: kojcharat.narupatpajong@g.swu.ac.th

บทนำ

สถานการณ์ในปัจจุบัน พบปัญหาหมอกควันและฝุ่นละอองขนาดเล็ก โดยเฉพาะฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน (PM2.5) ถือเป็นปัญหาสำคัญของประเทศไทย เนื่องจากสถานการณ์ PM2.5 เกินค่ามาตรฐานในทุกปี โดยค่ามลภาวะทางอากาศสูงติดอันดับต้นๆ ของโลก และรุนแรงมากขึ้นทุกปี โดยเฉพาะในเมืองหลวง พื้นที่กรุงเทพฯ ปริมณฑล และในบางพื้นที่ของประเทศไทย จากข้อมูลการเฝ้าระวังสถานการณ์ PM2.5 ของประเทศไทย พบว่าค่า PM2.5 สูงเกินค่ามาตรฐานของไทยและเกินค่าแนะนำขององค์การอนามัยโลกในหลายพื้นที่ [1] ซึ่งก่อให้เกิดผลกระทบต่อสุขภาพมนุษย์ได้โดยตรงและกระทบต่อภาพลักษณ์ในฐานะศูนย์กลางการท่องเที่ยวและเศรษฐกิจของเอเชียตะวันออกเฉียงใต้

กรมควบคุมมลพิษ [2] ได้กล่าวถึงเรื่อง “การพัดพาและแปรสภาพของมลพิษ (transportation and transformation of pollutants)” ปัจจัยที่ส่งผลต่อการแพร่กระจายของมลพิษ ได้แก่ สภาพอุตุนิยมวิทยาและสภาพแวดล้อม โดยสภาพอุตุนิยมวิทยา ได้แก่ ฤดูมรสุมตะวันออกเฉียงเหนือและฤดูมรสุมตะวันตกเฉียงใต้ ซึ่งฤดูมรสุมนี้ ส่งผลต่อสภาพอากาศ รวมถึงความกดอากาศ ทิศทางลมประจำฤดู อุณหภูมิ ปริมาณฝน ความชื้น ปัจจัยเหล่านี้ทำให้ระดับฝุ่นละอองในฤดูมรสุมตะวันออกเฉียงเหนือในพื้นที่กรุงเทพมหานคร มีระดับสูงขึ้น จากปัจจัยดังกล่าว ทำให้เกิดงานวิจัยนี้ โดยนำข้อมูลภาคอุตุนิยมวิทยาที่เกิดขึ้นในอดีตมาใช้รวมกับเทคนิคการเรียนรู้ของเครื่องเพื่อทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)ในอนาคต

ในงานวิจัยนี้ มีการนำเข้าข้อมูลจากเว็บไซต์สถานีตรวจวัดอากาศ ในกรุงเทพมหานคร ที่มีการนำเข้าข้อมูลในช่วงวันที่ 1 มกราคม - 31 ธันวาคม 2562 และช่วง 1 มกราคม – 28 กันยายน 2563 ซึ่งได้แก่ ค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) และตัวแปรที่สามารถส่งผลต่อความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ได้แก่ อุณหภูมิ, จุดน้ำค้าง, ความชื้น, ทิศทางลม, ความเร็วลม , ลมกระโชก และ ความกดอากาศ สำหรับแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) ที่ใช้ในงานวิจัยนี้มีทั้งหมด 4 แบบ ซึ่งได้รับความนิยมและยอมรับในปัจจุบัน ประกอบด้วย LR (Linear Regression), SVR (Support Vector Regression) , XGBoost (eXtreme Gradient Boosting) และ MLP (Multi-Layer Perceptron)

อย่างไรก็ตาม การวัดความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) นั้นต้องใช้เครื่องมือและเทคนิคต่างๆ เพื่อให้ได้ข้อมูลที่ถูกต้องและมีประสิทธิภาพ จากปัจจัยที่ส่งผล เราสามารถนำข้อมูลภาคอุตุนิยมวิทยาที่เกิดขึ้นในอดีต มาใช้ร่วมกับแบบจำลองที่สร้างขึ้น เพื่อทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคต มาใช้ประกอบการตัดสินใจร่วมได้ เพื่อเข้าใจแนวโน้มของสถานการณ์ปริมาณ PM2.5 ทำให้ผู้เกี่ยวข้อง สามารถใช้ในการวางแผนการจัดการที่เหมาะสม ช่วยลดผลกระทบต่อสุขภาพของมนุษย์ ที่อยู่ในพื้นที่ที่มีความเสี่ยงต่อฝุ่นละอองขนาดเล็ก (PM2.5) สร้างความเข้าใจในผลกระทบต่อสุขภาพ ประเมินระดับความเสี่ยงที่เกิด รวมถึงพัฒนาแนวทางการบริหารจัดการในระยะยาว ในการรับมือปริมาณฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคต

งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้อง มีงานวิจัยเกี่ยวข้องกับการทำนายค่าฝุ่นละอองขนาดเล็ก (PM2.5) มีตัวอย่างดังต่อไปนี้ (Ejohwomu et al., 2022) [3] งานวิจัยนี้ ผู้วิจัยได้สร้างแบบจำลองการทำนายที่เชื่อถือได้ เป็นเครื่องมือที่มีประโยชน์ในการทำความเข้าใจปัจจัยที่สามารถส่งผลต่อความเข้มของปริมาณ PM2.5 ซึ่งข้อมูลนี้สามารถใช้ในการพัฒนากลยุทธ์และนโยบายการลดปริมาณของมลพิษทางอากาศ ผู้วิจัยได้ใช้ข้อมูลคุณภาพอากาศ ในเมืองลากอส ประเทศไนจีเรีย มีการเปรียบเทียบใช้อัลกอริทึมทั้งหมด 7 แบบ (Prophet, XGBoost, SVM, RF, neural network, ARIMA) และแบบจำลองแบบผสม 3 แบบ โดยพบที่สำคัญ 2 ประการจากการศึกษานี้ 1. อุตุนิยมวิทยา เป็นปัจจัยที่เป็นประโยชน์สำหรับการพยากรณ์ความเข้มของ PM2.5 และ ข้อ 2.

แบบจำลอง Ensemble (เช่น XGBoost-RF-ARIMA) สามารถทำนายความเข้มข้นของ PM2.5 ที่น่าเชื่อถือเมื่อเปรียบเทียบกับอัลกอริทึมแบบ Standalone Algorithms นอกจากนี้ สามารถสรุปได้ดังต่อไปนี้ 1. ตัวแปรทางมาตรวิทยา หรือข้อมูลที่สามารถวัดหรือวิเคราะห์ได้ตามหลักการทางวิทยาศาสตร์นั้น ไม่สามารถทำนายความเข้มข้นของ PM2.5 ในช่วงสูงสุดและต่ำสุดได้อย่างเพียงพอ ชี้ให้เห็นถึงความจำเป็นในการรวบรวมข้อมูลปัจจัยอื่นๆ เช่น จำนวนรถ ประเภทรถ และแหล่งที่มาอื่นๆ ของมลพิษทางอากาศ 2. เรื่องความก้าวหน้าด้านวิทยาศาสตร์ข้อมูล อาจจะมีเครื่องมือที่สามารถใช้ในการสร้างการทำนายความเข้มข้นของ PM2.5 เกิดขึ้น 3. ข้อจำกัดที่สำคัญที่สุดของการศึกษานี้ คือใช้ตัวแปรทางอากาศวิทยา เป็นตัวแปรเพียงอย่างเดียว ในการพัฒนาแบบจำลอง ตัวแปรอื่น ๆ อย่างเช่น ตัวแปรที่เกี่ยวกับปฏิทินและเวลา ถูกสร้างขึ้นจากองค์ประกอบเวลาเท่านั้น ไม่ได้ถูกนำมาพิจารณาในการสร้างแบบจำลอง เพื่อหาตัวแปรที่มีผลกระทบมากสุดต่อความเข้มของ PM2.5 โดยการเอาหลายตัวแปร มารวมเข้ากับแต่ละแบบจำลอง โดยทั้งหมดมีจำนวน 25 แบบจำลอง จาก 5 แบบที่ดีที่สุดเปรียบเทียบกับค่าจริง สามารถสังเกตได้ว่าค่า MAE, MASE, และ RMSE สำหรับโมเดล XGBoost_All มีค่าต่ำที่สุด เมื่อเทียบกับแบบจำลองอื่น ๆ อยู่ที่ 1.69, 0.77, และ 2.3809 ตามลำดับ

(Lee et al., 2023) [4] ผู้วิจัยได้ใช้ข้อมูลจาก Air Quality Research Centers ซึ่งดำเนินการโดยกระทรวงสิ่งแวดล้อมของเกาหลี ทั้งหมด 3 เมืองของประเทศเกาหลีใต้ ได้แก่ กรุงโซล อุลซาน และแบงเนียง ระหว่างปี 2018 ถึง 2020 Input data ถูกแบ่งออกเป็น 4 หมวด ได้แก่ 1. Chemical species 2. Time 3. Air pollutants 4. Meteorological data ในขั้นตอนการสร้างแบบจำลอง มีการเพิ่มขึ้นขั้นตอนละ 1ใน4 กลุ่มของข้อมูลนำเข้า กลุ่มข้อมูลถูกจัดประเภทตั้งแต่ Input Data #1 ถึง Input Data #4 โดยที่ตัวเลขมีค่ามากขึ้นหมายถึงมีข้อมูลนำเข้ามากขึ้นในการทำนาย และใช้ 7 คุณลักษณะที่จะทำนาย (Prediction Case) ด้วยการเพิ่มตัวแปรเกี่ยวข้องกับคุณภาพอากาศอีก 7 prediction components Case คุณลักษณะเป้าหมายที่แบบจำลองทำนาย ซึ่งข้อมูลคุณลักษณะทั้งหมดถูกทำให้เป็น min–max normalized ก่อนการฝึกแบบจำลองและแปลงกลับหลังการสร้างแบบจำลองแล้ว โดยใช้แบบจำลองการเรียนรู้ของเครื่อง ML 4 แบบ ได้แก่ Generative Adversarial Imputation Network (GAIN), Fully Connected Deep Neural Network (FCDNN), Random Forest (RF) และ k-nearest neighbor (KNN) ใช้ตัวชี้วัด ได้แก่ R2 ,RMSE , MAE ซึ่งความแม่นยำในการทำนาย หรือ ค่า R2 สูงที่สุด คือ แบบจำลอง GAIN ค่าR2 = 0.897 รองลงมา FCDNN 0.861, RF 0.785, และ KNN 0.801 ตามลำดับ บ่งบอกว่าแบบจำลองการเรียนรู้เชิงลึกมีการนำไปประยุกต์ใช้กับข้อมูลที่เพิ่มมากขึ้นได้อย่างยอดเยี่ยม

วิธีดำเนินการ

ขั้นตอนที่ 1 : ชุดข้อมูลที่ใช้ในการศึกษานี้

ในการดำเนินงานวิจัยนี้ มีการใช้งานชุดข้อมูล 2 ชุดมารวมกัน โดยนำเข้าข้อมูลจากแหล่งข้อมูลสาธารณะแบบเปิด ผ่านหน้าเว็บไซต์และวิธีการ Web Scraping ซึ่งเป็นกระบวนการในการดึงข้อมูล โดยใช้สคริปต์ดึงข้อมูลจากหน้าเว็บไซต์ รายละเอียดดังนี้

1. ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) จากเว็บไซต์ Berkeley Earth [5]ซึ่งเป็นองค์กรอิสระไม่แสวงหาผลกำไรของประเทศสหรัฐอเมริกา มุ่งเน้นวิทยาศาสตร์ข้อมูล สิ่งแวดล้อมและการวิเคราะห์ โดย Berkeley Earth รวบรวมข้อมูลที่เกี่ยวข้องกับมลพิษทางอากาศ ครอบคลุมทั่วโลก ซึ่งเป็นข้อมูลสาธารณะแบบเปิด ถูกรวบรวมไว้บนหน้าเว็บไซต์ โดยข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) ที่นำมาอยู่ในช่วงวันที่ 1 มกราคม 2562 - 31 ธันวาคม 2562 และช่วง 1 มกราคม 2563 – 28 กันยายน 2563 การเก็บค่ามีระยะห่าง 1 ชั่วโมง ที่จังหวัดกรุงเทพมหานคร ละติจูดที่ 13.754 ลองจิจูด ที่ 100.5014 ทั้งหมด 15,075 แถว

2. ข้อมูลภาคอุตุนิยมวิทยา จากเว็บไซต์ Weather Underground [6]ซึ่งเป็นอันดับที่ 1 ในการให้บริการสภาพอากาศเชิงพาณิชย์ ที่ให้ข้อมูลสภาพอากาศแบบเรียลไทม์ผ่านทางอินเทอร์เน็ต ภารกิจคือทำให้ข้อมูลสภาพอากาศ มีคุณภาพ พร้อมใช้งานสำหรับทุกคนบนโลก ซึ่งเจ้าของคือ The Weather Company ซึ่งเป็นบริษัทย่อยของ IBM โดยข้อมูลภาคอุตุนิยมวิทยา ที่นำมาผ่านวิธีการ Web Scraping ในการดึงข้อมูล จากสถานี IKRUNGTH3 บริเวณ ซอยวิภาวดี 60 เขตหลักสี่ จังหวัดกรุงเทพมหานคร ละติจูดที่ 13.865° N ลองจิจูดที่ 100.581° E ในช่วงวันที่ 1 มกราคม 2562 - 31 ธันวาคม 2562 และช่วง 1 มกราคม 2563 – 31 ธันวาคม 2563 มีการเก็บค่าระยะห่างประมาณ 15 นาที ซึ่งไม่เท่ากันขึ้นอยู่กับการอัพเดตของสถานี ภายในชุดข้อมูลมีตัวแปรที่สามารถส่งผลต่อค่า PM2.5 ได้แก่ อุณหภูมิ, จุดน้ำค้าง, ความชื้น, ทิศทางลม, ความเร็วลม ,ลมกระโชก และ ความกดอากาศ มีทั้งหมด 145,644 แถว

ขั้นตอนที่ 2 : การนำเข้าข้อมูล การสร้างชุดข้อมูล และการจัดการชุดข้อมูล

การสร้างชุดข้อมูลค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 และข้อมูลสภาพอากาศ

1. ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) ซึ่งเป็นข้อมูลสาธารณะแบบเปิด ถูกรวบรวมไว้บนหน้าเว็บไซต์ สามารถดาวน์โหลดข้อมูลได้จากเว็บไซต์ Berkeley Earth เป็นแบบรายชั่วโมง ซึ่งทางเว็บไซต์ให้คำอธิบายของการใช้งาน คือทุกเวลาถูกแสดงในรูปแบบเวลาทางโลก (UTC) และให้ทราบว่าตัวตรวจวัดคุณภาพอากาศแต่ละตัว มีกระบวนการควบคุมคุณภาพอัตโนมัติที่ใช้ตรวจสอบข้อมูลที่ผิดพลาด แต่อาจต้องทำการแก้ไขเพิ่มเติม การรายงานค่าของ PM2.5 ที่สูงหรือต่ำกว่าค่าเฉลี่ยที่รายงานไว้ รวมถึงอาจมีการเปลี่ยนแปลงในภายหลัง นอกจากนี้จำนวนของสถานีตรวจวัดและการกระจายตำแหน่ง มีโอกาสที่จะมีการเปลี่ยนแปลงในระหว่างเวลา ซึ่งข้อมูลที่นำมาอยู่ในช่วงวันที่ 1 มกราคม 2562 - 31 ธันวาคม 2562 และช่วง 1 มกราคม 2563 – 28 กันยายน 2563 การเก็บค่ามีระยะห่าง 1 ชั่วโมง ที่จังหวัดกรุงเทพมหานคร ละติจูดที่ 13.754 ลองจิจูด ที่ 100.5014 ทั้งหมด 15,075 แถว นำมาเก็บเป็นไฟล์ CSV

```
% Country: Thailand
% Region: Thailand
% Population: NaN
% Max Latitude: 20.4632
% Max Longitude: 105.6394
% Min Latitude: 5.61
% Min Longitude: 97.3456
% Year, Month, Day, UTC Hour, PM2.5,
2019    1       1       0       21.40
2019    1       1       1       19.99
2019    1       1       2       18.46
2019    1       1       3       18.95
2019    1       1       4       20.15
2019    1       1       5       17.71
2019    1       1       6       16.30
2019    1       1       7       14.15
2019    1       1       8       13.13
2019    1       1       9       12.39
2019    1       1       10      11.21
2019    1       1       11      10.42
2019    1       1       12      12.16
2019    1       1       13      15.17
2019    1       1       14      18.96
2019    1       1       15      24.77
2019    1       1       16      25.59
2019    1       1       17      28.09
2019    1       1       18      28.72
2019    1       1       19      30.18
2019    1       1       20      29.56
2019    1       1       21      26.35
2019    1       1       22      24.41
```

Year	Month	Day	UTC Hour	PM2.5
2019	1	1	0	26.5
2019	1	1	1	25.3
2019	1	1	2	24.5
2019	1	1	3	23.5
2019	1	1	4	23.9
2019	1	1	5	21
2019	1	1	6	18.7
2019	1	1	7	19.6
2019	1	1	8	18.8
2019	1	1	9	16.4
2019	1	1	10	16
2019	1	1	11	14.7
2019	1	1	12	15.7
2019	1	1	13	17.1
2019	1	1	14	18.5
2019	1	1	15	19.7
2019	1	1	16	24.5
2019	1	1	17	30.3
2019	1	1	18	31.7
2019	1	1	19	33.8

ภาพประกอบ 1 ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5)จากเว็บไซต์ Berkeley Earth

2. ข้อมูลภาคอุตุนิยมวิทยา ผ่านวิธีการ Web Scraping โดยใช้สคริปต์ดึงข้อมูลจากหน้าเว็บไซต์ Weather Underground โดย สคริปต์นำมาจาก ผู้ใช้ GitHub.com ชื่อว่า Karlheinzniebuhr/the-weather-scraper [7] ในงานวิจัยนี้ใช้ผ่าน Colab Python3 ภายใน ประกอบด้วย 5 ไฟล์ ดังนี้

1. config.py สำหรับการตั้งค่า Start Date – End Date กำหนดระยะเวลาที่จะดึงข้อมูล
2. stations.txt สำหรับ Station ID ชื่อสถานีที่ใช้ในการดึงข้อมูล
3.requirements.txt สำหรับ Install (use Python3)
4. util.rar สำหรับที่เกี่ยวข้องกับการแปลงหน่วย, การวิเคราะห์ข้อมูล, และฟังก์ชันอื่น ๆ ที่มีไว้ให้ง่ายต่อการใช้งานหรือปรับแต่งโปรแกรม ภายในประกอบด้วยไฟล์ UnitConverter.py, Parser.py, และ Utils.py ที่ถูกนำเข้ามาใช้ในสคริปต์
5. weather_scraper.py เป็นสคริปต์หลักที่ใช้ในการดึงข้อมูลหลังจากตั้งค่าในไฟล์ด้านบนทั้งหมด

เริ่มจากการนำไฟล์ทั้งหมด Upload ขึ้น Colab เนื่องจากไฟล์ util.rar มีนามสกุล rar จึงต้องใช้โปรแกรม "unrar" เพื่อแตกไฟล์จากไฟล์ที่มีนามสกุล .rar เพื่ออ่านข้อมูลภายในแฟ้มข้อมูล ดังภาพประกอบ2

```
!unrar x "util.rar"

UNRAR 5.61 beta 1 freeware      Copyright (c) 1993-2018 Alexander Roshal

Extracting from util.rar

Creating    util                                              OK
Extracting  util/Parser.py                                    OK
Extracting  util/UnitConverter.py                             OK
Extracting  util/Utils.py                                     OK
Extracting  util/__init__.py                                  OK
All OK
```

ภาพประกอบ 2 โปรแกรม "unrar" เพื่อแตกไฟล์จากไฟล์ที่มีนามสกุล .rar

จากนั้น ทำการตั้งค่า ในไฟล์ config.py กำหนดระยะเวลาที่จะดึงข้อมูล Start Date – End Date ดังภาพประกอบ 3

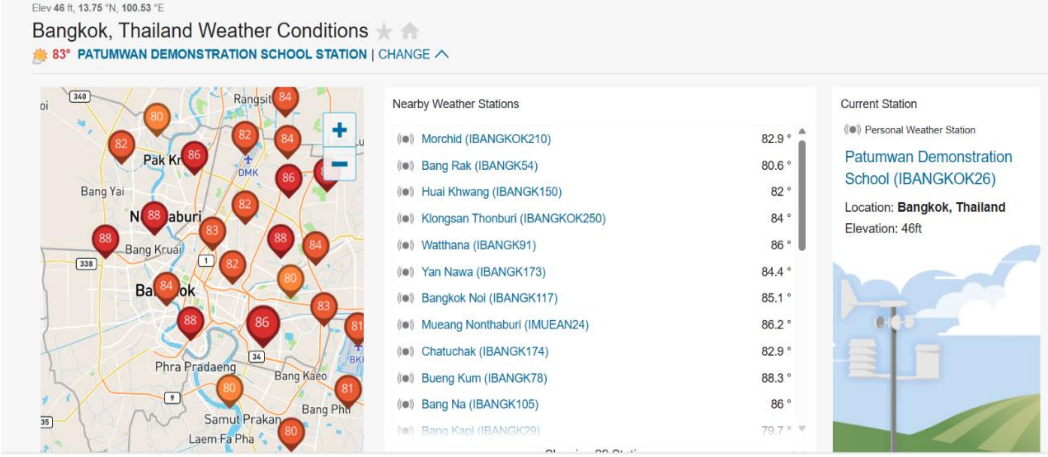
```
from datetime import date

# Set Date format like: YYYY, MM, DD
# Note that FIND_FIRST_DATE uses START_DATE as default start date
START_DATE = date(2019, 1, 1)
END_DATE = date(2020, 12, 31)

# set to "metric" or "imperial"
UNIT_SYSTEM = "metric"
# UNIT_SYSTEM = "imperial"

# Automatically find first date where data is logged
FIND_FIRST_DATE = True
```

ภาพประกอบ 3 กำหนดระยะเวลาที่จะดึงข้อมูลStart Date – End Date

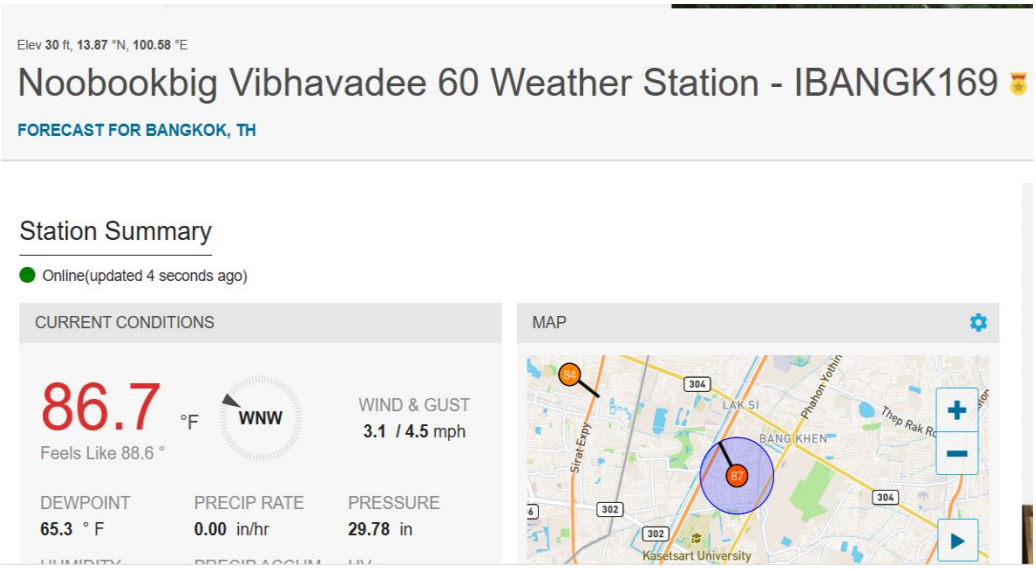
ที่มา (Karlheinzniebuhr, 2022)

โดย รูปแบบของการกำหนดวันที่ เป็น ปีค.ศ – เดือน - วันที่ และ ตั้งค่ารูปแบบหน่วยการวัดเป็น " metric" ซึ่ง Metric (เมตริก) ใช้หน่วยวัดที่มีที่มาจากระบบเมตริก เป็นระบบที่ใช้ทั่วไปในส่วนใหญ่ของโลก

จากนั้น ทำการตั้งค่าต่อที่ ไฟล์ stations.txt เป็นไฟล์ text สำหรับ วาง Station ID ชื่อสถานีที่ใช้ในการดึงข้อมูล ซึ่งการหา Station ID ได้บนเว็บไซต์ Weather Underground



ภาพประกอบ 4 บนเว็บไซต์ Weather Underground



ภาพประกอบ 5 Station ID สถานีตรวจวัด

เมื่อกดเข้าไปยังสถานีตรวจวัดที่เลือก จะปรากฏ Station ID ที่ด้านขวาดังภาพประกอบ 5 IBANGK169 จากนั้น คัดลอกมาใส่ลงไฟล์ stations.txt ดังภาพประกอบ หากมีการดึงที่มากกว่า 1 สถานีให้ใส่ 1 Station ID ต่อ 1 บรรทัด URLs

```
https://www.wunderground.com/dashboard/pws/ Station ID จากบนเว็บไซต์ Weather Underground
https://www.wunderground.com/dashboard/pws/IBANGK169
```

ภาพประกอบ 6 การกรอก Station ID

เมื่อตั้งค่าทุกอย่างเรียบร้อย ทำการ RUN weather_scraper.py ซึ่งเป็นสคริปต์หลักที่ใช้ในการดึงข้อมูลจากหน้า Dashboard ของสถานีตรวจวัดอากาศที่ทำการเลือกไว้ก่อนหน้า ภายในสคริปต์ มีการดึงและสร้างคอลัมน์ ดังนี้ **['Date','Time','Temperature','Dew_Point','Humidity','Wind','Speed','Gust','Pressure','Precip_Rate','Precip_Accum','UV', 'Solar']** หลังจากดึงข้อมูลไฟล์เสร็จสิ้น จะถูกบันทึกเก็บเป็นไฟล์ CSV ดังภาพประกอบ 7

Date	Time	Temperature_C	Dew_Point_C	Humidity_%	Wind	Speed_kmh	Gust_kmh	Pressure_hPa	Precip_Rate_mm	Precip_Accum_mm	UV	Solar_w/m2
1/1/2019	12:07 AM	24.5	17.61	70	West	6.11	32.02	1013.21	NA	NA	NA	2
1/1/2019	12:17 AM	24.5	17.28	69	WSW	4.99	16.89	1013.21	NA	NA	NA	2
1/1/2019	12:26 AM	24.39	17.22	69	SW	4.99	10.14	1013.21	NA	NA	NA	2
1/1/2019	12:41 AM	24.39	17.72	71	SW	7.24	36.69	1013.21	NA	NA	NA	2
1/1/2019	12:53 AM	24.22	17.72	72	North	7.88	29.93	1013.21	NA	NA	NA	2
1/1/2019	1:04 AM	24	17.78	73	NE	11.91	47.95	1013.21	NA	NA	NA	2
1/1/2019	1:13 AM	23.89	17.89	74	NNW	10.14	47.95	1012.87	NA	NA	NA	2
1/1/2019	1:29 AM	23.78	18.11	75	NE	10.14	38.94	1012.87	NA	NA	NA	2
1/1/2019	1:39 AM	23.72	17	71	NW	6.11	33.15	1012.87	NA	NA	NA	2
1/1/2019	1:49 AM	23.72	17	71	SW	4.99	18.02	1012.87	NA	NA	NA	2
1/1/2019	1:52 AM	23.72	17	71	SW	6.11	28	1012.87	NA	NA	NA	2
1/1/2019	2:01 AM	23.61	17.61	74	WNW	11.91	32.02	1012.53	NA	NA	NA	2
1/1/2019	2:13 AM	23.5	17.78	75	SW	10.14	25.9	1012.53	NA	NA	NA	2
1/1/2019	2:27 AM	23.39	17.72	75	WSW	9.01	23.01	1012.53	NA	NA	NA	2
1/1/2019	2:37 AM	23.39	17.5	74	SW	11.91	33.15	1012.19	NA	NA	NA	2
1/1/2019	2:46 AM	23.28	16.89	72	SW	10.14	25.9	1012.19	NA	NA	NA	2
1/1/2019	2:56 AM	23.22	17.28	74	SW	9.01	29.93	1012.19	NA	NA	NA	2
1/1/2019	3:08 AM	23.11	16.89	73	WNW	10.14	33.15	1012.53	NA	NA	NA	2
1/1/2019	3:15 AM	23	17.28	75	SW	10.14	30.89	1012.19	NA	NA	NA	2
1/1/2019	3:25 AM	23	16.78	73	SW	9.01	20.92	1012.53	NA	NA	NA	2
1/1/2019	3:35 AM	23	16.39	71	North	10.14	36.04	1012.19	NA	NA	NA	2
1/1/2019	3:43 AM	22.89	16.28	71	WSW	6.11	23.01	1012.19	NA	NA	NA	2
1/1/2019	3:52 AM	22.89	16	70	WSW	9.01	30.89	1012.19	NA	NA	NA	2
1/1/2019	4:03 AM	22.89	16	70	WSW	9.01	27.03	1012.53	NA	NA	NA	2
1/1/2019	4:14 AM	22.78	16.39	72	SW	9.01	20.92	1012.53	NA	NA	NA	2

ภาพประกอบ 7 ข้อมูลภาคอุตุนิยมวิทยา ไฟล์ CSV

ข้อมูลภาคอุตุนิยมวิทยา ที่นำมา จากสถานี IKRUNGTH3 บริเวณ ซอยวิภาวดี 60 เขตหลักสี่ จังหวัดกรุงเทพมหานคร ละติจูดที่ 13.865° N ลองจิจูดที่ 100.581° E ในช่วงวันที่ 1 มกราคม 2562 - 31 ธันวาคม 2562 และช่วง 1 มกราคม 2563 – 31 ธันวาคม 2563 มีการเก็บค่าระยะห่างประมาณ 15 นาที ซึ่งไม่เท่ากันขึ้นอยู่กับการอัพเดตของสถานี ภายในชุดข้อมูลมีตัวแปรที่สามารถส่งผลต่อค่า PM2.5 ได้แก่ อุณหภูมิ, จุดน้ำค้าง, ความชื้น, ทิศทางลม, ความเร็วลม ,ลมกระโชก และ ความกดอากาศ มีทั้งหมด 145,644 แถวแถว 13 คอลัมน์ รวมคอลัมน์วันและเวลา นำมาเก็บเป็นไฟล์ CSV

การจัดการข้อมูล สำรวจและวิเคราะห์ข้อมูลเบื้องต้น

การจัดการกับข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) จากเว็บไซต์ Berkeley Earth

1. ทำการแปลง UTC เป็น Local Time +7 เนื่องจากข้อมูลดิบ Raw Data ที่ได้มาเป็น UTC ซึ่งเวลาจะต่างจากเวลาจริงอยู่ 7 ชั่วโมง พร้อมทั้งเปลี่ยนชื่อเป็น Hour

2. ทำการสร้าง คอลัมน์ใหม่ ชื่อว่า ‘Season’ เป็นคอลัมน์บอกถึงฤดูกาลของข้อมูลแต่ละแถว ซึ่งฤดูกาลในประเทศไทย ตามกรมอุตุนิยมวิทยาโดยทั่ว ๆ ไปสามารถแบ่งออกได้เป็น 3 ฤดู การเริ่มต้นและสิ้นสุดของฤดูกาล อาจผันแปรไปจากปกติได้ในแต่ละปี ซึ่งในที่นี้ใช้เงื่อนไขดังนี้

2.1. ฤดูร้อนระหว่างเดือนมีนาคมถึงเดือนพฤษภาคม

2.2. ฤดูฝนระหว่างเดือนมิถุนายนถึงเดือนตุลาคม

2.3. ฤดูหนาวระหว่างเดือนพฤศจิกายนถึงเดือนกุมภาพันธ์

3. ทำการ One-Hot Encoding ในการจัดการกับคุณลักษณะที่ไม่มีลำดับ ในคอลัมน์ “ Season”โดยแต่ละตัวแปรจะแทนค่าของหมวดหมู่ด้วยตัวเลข 0 หรือ 1 เพื่อให้แบบจำลองเรียนรู้และวิเคราะห์ข้อมูลได้ถูกต้อง ผลลัพธ์ที่ได้จะเป็นตารางใหม่มี 3 คอลัมน์ ดังนี้ Season_Rainy, Season_Summer, Season_Winter ดังภาพประกอบ 8

Season	Season_Rainy	Season_Summer	Season_Winter
Winter	0	0	1
Winter	0	0	1
Winter	0	0	1
Winter	0	0	1
Winter	0	0	1
...	...	...	...
Rainy	1	0	0
Rainy	1	0	0
Rainy	1	0	0
Rainy	1	0	0
Rainy	1	0	0

ภาพประกอบ 8 การ One-Hot Encoding ในคอลัมน์ “ Season”

4. ทำการ Rolling Mean ข้อมูล 24 48 72 ชั่วโมง เพื่อสร้างคอลัมน์ใหม่ เป็นคอลัมน์ Last24hrs_mean, Last48hrs_mean, Last72hrs_mean การ Rolling เป็นการนำข้อมูลในช่วงเวลาหนึ่งมาประมวลผล เช่น หาค่าเฉลี่ย หาผลรวม เฉลี่ยรายเดือน วัน หรือ ชั่วโมง

5. สร้างคอลัมน์ใหม่ จากข้อมูลคอลัมน์ “PM 2.5” โดยการ shift (1),(6),(12),(24) ตามลำดับ เพื่อเลื่อนข้อมูลมาใช้ในชั่วโมงย้อนหลัง ซึ่งใช้เป็นค่า PM2.5 ของ 1,6,12,24 ชั่วโมงย้อนหลัง ได้คอลัมน์ ดังนี้ PM2.5(h-1), PM2.5(h-6), PM2.5(h-12), PM2.5(h-24)

6. สร้างคอลัมน์ใหม่ จากข้อมูลคอลัมน์ “PM2.5” ด้วยเช่นกัน โดยการ shift (-1), (-6), (-12), (-24) ตามลำดับ เพื่อเลื่อนข้อมูลมาใช้ในชั่วโมงล่วงหน้า 1,6,12,24 ชั่วโมงตามลำดับ ซึ่งคอลัมน์ใหม่ที่ได้จากการขยับคอลัมน์ “PM 2.5” ดังนี้ PM2.5(h+1), PM2.5(h+6), PM2.5(h+12), PM2.5(h+24) เพื่อใช้เป็นข้อมูล Target ที่จะวิเคราะห์ PM 2.5 ในชั่วโมงล่วงหน้าถัดๆไป

7. จัดการข้อมูลสูญหาย โดยทำการตรวจสอบค่า NaN หรือ ค่าว่างในข้อมูล โดยใช้ .isnull().sum() และจัดการค่า NaN ค่าว่างในข้อมูล โดยการ Fill ซึ่งเป็นเทคนิคในการจัดการข้อมูลที่สูญหายในชุดข้อมูล โดยการเติมค่าข้อมูลที่ขาดหายไป ในที่นี้ใช้ 2 วิธี คือ “ .fillna (method='bfill') และ .fillna (method='ffill')” ซึ่งการ Fill แบบ bfill เป็นการ Fill Value Backward เมื่อใช้บนคอลัมน์ใด ๆ ของชุดข้อมูล จะเป็นการกรอกค่าสูญหายด้วยค่าที่อยู่ในแถวถัดไป ที่ไม่ใช่ค่าว่าง ย้อนกลับมาเติม และ การ Fill แบบ ffill เป็นการ forward fill เมื่อใช้บนคอลัมน์ใด ๆ ของชุดข้อมูล จะเป็นการกรอกค่าสูญหายด้วยค่าที่อยู่ในแถวก่อนหน้า ที่ไม่ใช่ค่าว่างมาเติม เนื่องจากคอลัมน์ Last24hrs_mean, Last48hrs_mean, Last72hrs_mean, PM2.5(h-1), PM2.5(h-6), PM2.5(h-12), PM2.5(h-24), PM2.5(h+24) ที่ได้สร้างขึ้น ในแถวแรกของข้อมูลนั้น เป็นค่าว่าง หากใช้แบบ “ ffill Method ” ที่นำแถวก่อนหน้ามาใส่ จะไม่สามารถหาข้อมูลมาใส่ได้ ทำการตรวจข้อมูลที่ขาดหายไปตรวจพบว่ามีจำนวน 47, 95, 143 , 1 ,6 ,12 ,24 ตามลำดับ จึงทำการ bfill เพื่อจัดการกับค่าว่างเหล่านี้ ถัดมาเป็น คอลัมน์ PM2.5(h+1), PM2.5(h+6), PM2.5(h+12), PM2.5(h+24) ที่ได้สร้างขึ้นจัดการโดยใช้วิธี “ ffill Method ” เนื่องจากแถวสุดท้ายของแต่ละคอลัมน์เป็นค่าว่างจึงต้องใช้แบบ “ ffill”เพื่อนำค่าที่อยู่แถวก่อนหน้ามากรอกค่าสูญหาย ดังภาพประกอบ 9

Year	0
Month	0
Day	0
Hour	0
PM2.5	0
Date	0
DayName	0
Date_Time1	0
Season	0
Season_Rainy	0
Season_Summer	0
Season_Winter	0
dayofweek	0
dayofyear	0
weekofyear	0
quarter	0
Last24hrs_mean	47
Last48hrs_mean	95
Last72hrs_mean	143
PM2.5(h-1)	1
PM2.5(h-6)	6
PM2.5(h-12)	12
PM2.5(h-24)	24
PM2.5(h+1)	1
PM2.5(h+6)	6
PM2.5(h+12)	12
PM2.5(h+24)	25

ภาพประกอบ 9 แสดงค่าว่างในข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5)

8. ใช้. duplicated() ตรวจสอบข้อมูลซ้ำ ไม่พบข้อมูลซ้ำ

การจัดการกับข้อมูลภาคอุตุนิยมวิทยา จากเว็บไซต์ Weather Underground

1. ทำการ read csv และ parse_dates=[['Date','Time']] เนื่องจากข้อมูลดิบเป็นคอลัมน์แยก Date กับ Time นำมารวมใน Data Frame เป็นคอลัมน์ใหม่ เพื่อใช้เป็น Index
2. จัดการข้อมูลสูญหาย โดยตรวจสอบค่า NaN หรือค่าว่างในข้อมูล พบค่า NaN ดังภาพ ประกอบ 10

```
Date_Time              0
Temperature_C          0
Dew_Point_C           66
Humidity_%             0
Wind                   6
Speed_kmh           8581
Gust_kmh            8581
Pressure_hPa           0
Precip_Rate_mm     135275
Precip_Accum_mm    112611
UV                 145644
Solar_w/m2             0
dtype: int64
```

ภาพประกอบ 10 แสดงค่าว่างในข้อมูลภาคอุตุนิยมวิทยา

ทำการจัดการค่า NaN โดยการ Fill ค่า NaN ด้วย “.fillna(method='ffill')” Fill Value Forward ที่คอลัมน์ Dew_Point_c, Wind, Speed_kmh, และ Gust_kmh เป็นวิธีการเติมด้วยค่าที่อยู่ในแถวก่อนหน้า ในแถวที่ไม่ใช่ค่าว่างมาเติมลงไป ซึ่งเหมาะสำหรับข้อมูลที่แนวโน้มค่าถัดไป จะมีค่าใกล้เคียงของเดิม

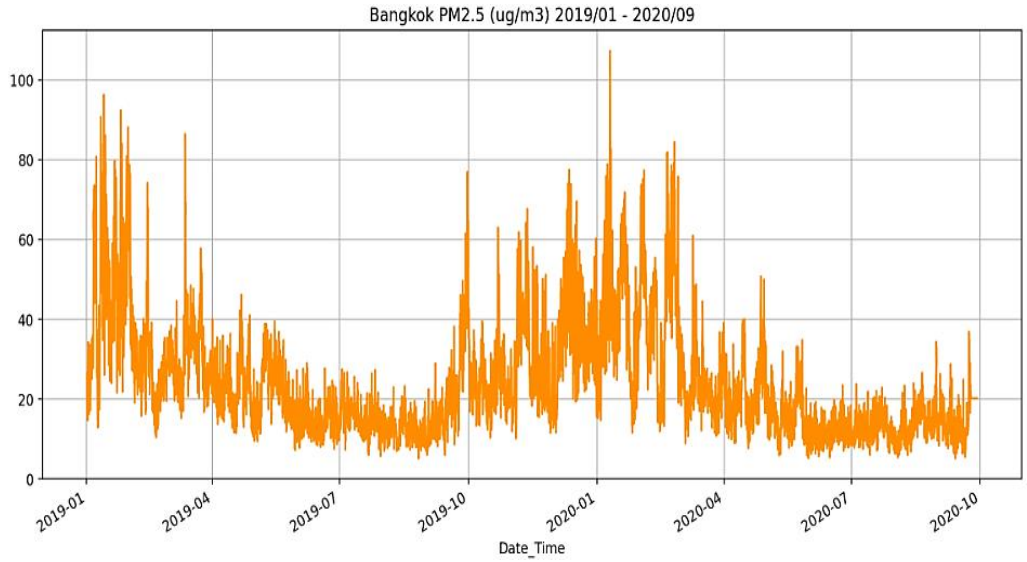
3. สำหรับค่า NaN หรือค่าว่างในข้อมูลคอลัมน์ Precip_Rate_mm ,Precip_Accum_mm และ UV มีจำนวนมากเกือบเท่ากับจำนวนแถวปกติ ให้ทำการ Drop คอลัมน์ทิ้ง เนื่องจากข้อมูลที่ขาดหายจำนวนมาก ส่งผลต่อประสิทธิภาพของแบบจำลอง
4. ทำการ One-Hot Encoding ในการจัดการกับคุณลักษณะที่ไม่มีลำดับ ในคอลัมน์ " Wind " เนื่องจากการประมวลผลของแบบจำลอง ต้องใช้ข้อมูลที่เป็นตัวเลขเท่านั้น ไม่สามารถอ่านค่าที่เป็นตัวอักษรในการคำนวณได้ โดยในคอลัมน์ “Wind” แต่ละตัวแปรจะแทนค่าของหมวดหมู่ด้วยตัวเลข 0 หรือ 1 เพื่อให้แบบจำลองเรียนรู้และวิเคราะห์ข้อมูลได้ถูกต้อง
5. ตรวจสอบข้อมูลซ้ำ ไม่พบข้อมูลซ้ำ
6. ทำการ Resample ข้อมูล เป็นช่วงเวลารายชั่วโมง H คือการสรุปหาค่าเฉลี่ยหรือผลรวมช่วงเวลาที่สนใจ เนื่องจากข้อมูลดิบในภาคอุตุนิยมวิทยา ที่เก็บได้จากทางสถานี มีระยะเวลาในการเก็บค่า ที่ไม่เท่ากัน รวมถึงค่า PM2.5 มีการเก็บค่าเป็นรายชั่วโมง เพื่อให้สอดคล้องกัน จึงต้อง resample('H'). mean() เฉลี่ยข้อมูลเป็นรายชั่วโมง
7. ตรวจสอบค่า NaN หรือค่าว่างในข้อมูลอีกครั้ง พบค่า NaN ดังภาพ ทุกคอลัมน์มีจำนวนเท่ากัน 653 ค่าว่าง
8. ทำการ Interpolate ข้อมูลที่ขาดหาย แบบ linear Interpolation คือการสร้างค่าใหม่จากค่าที่มีอยู่ โดยอาศัยอย่างน้อย 2 ค่าของข้อมูลระหว่างค่าเหล่านั้น เพื่อให้ได้ค่าตัวกลางที่อาจไม่มีอยู่ในข้อมูลต้นฉบับ นำมาใช้ในการจัดการข้อมูลที่ขาดหายไปหรือการสร้างข้อมูลใหม่ที่ต้องการค่าต่อเนื่อง เช่น การใช้ในกราฟแสดงความแตกต่างของอุณหภูมิตามเวลาหรือการสร้างข้อมูลเสมือนในกรณีข้อมูลที่หายไปบางส่วนในช่วงเวลาที่สนใจ

```
Temperature_C    653
Dew_Point_C      653
Humidity_%       653
Speed_kmh        653
Gust_kmh         653
Pressure_hPa     653
Solar_w/m2       653
Wind_ENE         653
Wind_ESE         653
Wind_East        653
Wind_NE          653
Wind_NNE         653
Wind_NNW         653
Wind_NW          653
Wind_North       653
Wind_SE          653
Wind_SSE         653
Wind_SSW         653
Wind_SW          653
Wind_South       653
Wind_WNW         653
Wind_WSW         653
Wind_West        653
dtype: int64
```

ภาพประกอบ 11 จำนวนข้อมูลที่ขาดหาย ก่อนทำการ Interpolate

หลังจากจัดการกับข้อมูล เตรียมข้อมูลที่ได้มาให้อยู่ในรูปแบบที่เหมาะสม สำหรับการนำเข้าสู่การสร้างแบบจำลอง ถัดมาทำการรวม Data Frame จาก 2 ตารางข้อมูลดิบ CSV ที่ได้ ซึ่งมี Index ของตารางที่อยู่ในช่วงค่าข้อมูลที่เท่ากัน คือ ระยะห่างข้อมูล 1 ชั่วโมง โดยใช้วิธีการ Merge Data Frame ใช้ key เป็นตัวเชื่อม ซึ่งก็คือ คอลัมน์ 'Date_Time' แบบ 'inner' ทั้งหมด 38 คอลัมน์

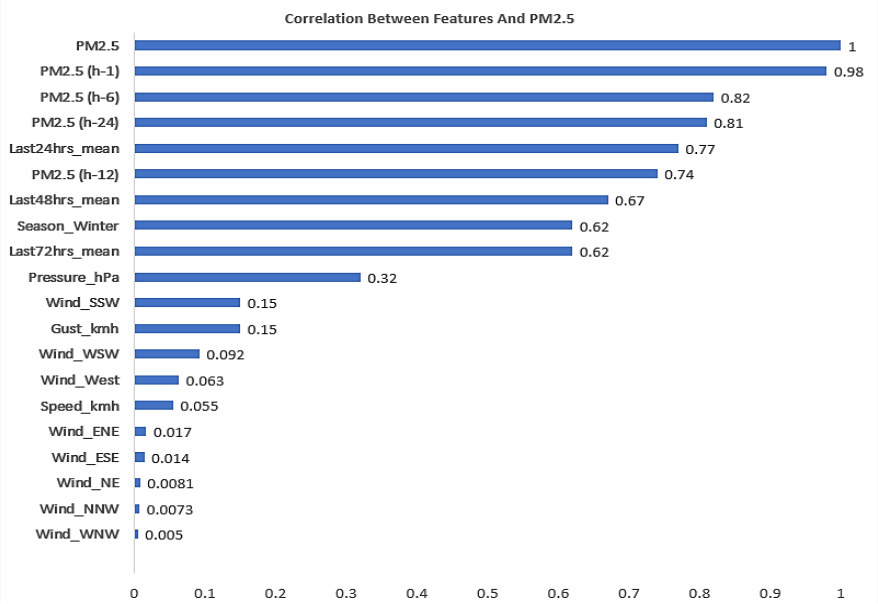
ขั้นตอนที่ 3 : สำรวจและวิเคราะห์ข้อมูลเบื้องต้น



ภาพประกอบ 12 ปริมาณฝุ่นละอองขนาดเล็ก PM2.5 ในปี2019-2020

ทำการวิเคราะห์ปริมาณของฝุ่นละลอง PM2.5 ในช่วง 2 ปี เริ่มตั้งแต่วันที่ มกราคม 2019 - ตุลาคม 2020 จะเห็นได้ว่าในช่วงต้นปี เดือน มกราคม ของ 2019 และ 2020 มีค่าฝุ่นละอองที่เพิ่มสูงขึ้น และค่อยๆลดลงในช่วงกลางปี และ เพิ่มสูงขึ้นอีกครั้งเมื่อเข้าสู่ปลายปีจนถึงต้นปี

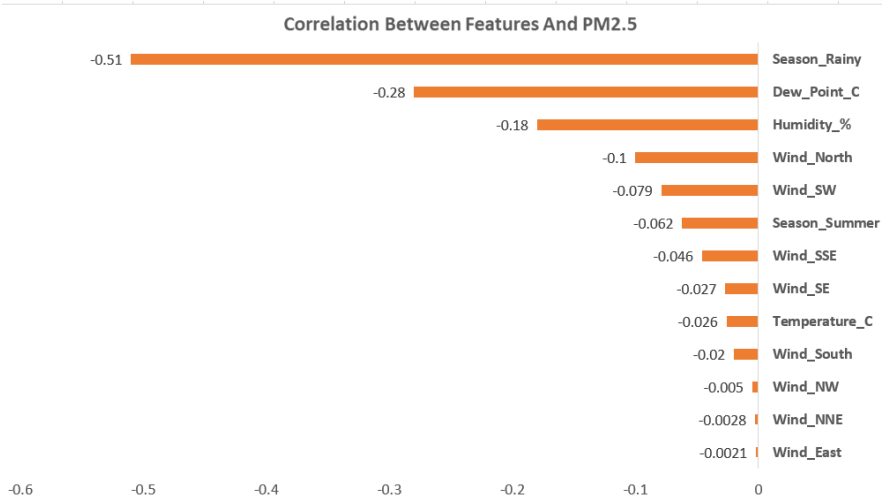
การสำรวจความสัมพันธ์ ระหว่างข้อมูลฝุ่นละอองขนาดเล็ก PM2.5 กับ คุณลักษณะอื่น ประกอบด้วย Season_Rainy, Season_Summer, Season_Winter, Last24hrs_mean, Last48hrs_mean, Last72hrs_mean, PM2.5(h-1), PM2.5(h-6), PM2.5(h-12), PM2.5(h-24), Temperature_C (อุณหภูมิ), Dew_Point_C (จุดน้ำค้าง), Humidity_% (ความชื้น), Wind (ทิศทางลม), Speed_kmh (ความเร็วลม), Gust_kmh (ลมกระโชก), Pressure_hPa (ความกดอากาศ) และทิศทางลมต่างๆ ดังนี้ Wind_ENE, Wind_ESE, Wind_East, Wind_NE, Wind_NNE, Wind_NNW, Wind_NW, Wind_North, Wind_SE, Wind_SSE, Wind_SSW, Wind_SW, Wind_South, Wind_WNW, Wind_WSW, Wind_West West ซึ่งมีทั้งความสัมพันธ์ในทางบวกและทางลบ ดังภาพประกอบ 13-14



ภาพประกอบ 13 แสดงความสัมพันธ์ระหว่างชุดข้อมูลฝุ่น PM2.5 กับ คุณลักษณะอื่นในทางบวก

เรียงลำดับความสัมพันธ์ทางบวกดังนี้

PM2.5(h-1) > PM2.5(h-6) > PM2.5(h-24) > Last24hrs_mean > PM2.5(h-12) > Last48hrs_mean > Last72hrs_mean เท่ากันกับ Season_Winter > Pressure_hPa (ความกดอากาศ) > Gust_kmh (ลมกระโชก) เท่ากันกับ ทิศทางลม Wind_SSW > ทิศทางลม Wind_WSW > ทิศทางลม Wind_West > Speed_kmh (ความเร็วลม) > ทิศทางลม Wind_ENE > ทิศทางลม Wind_ESE > ทิศทางลม Wind_NE > ทิศทางลม Wind_NNW > ทิศทางลม Wind_WNW



ภาพประกอบ 14 แสดงความสัมพันธ์ระหว่างชุดข้อมูลฝุ่น PM2.5 กับ คุณลักษณะอื่นในทางลบ

เรียงลำดับความสัมพันธ์ทางลบดังนี้

Season_Rainy > Dew_Point_C (จุดน้ำค้าง) > Humidity_% (ความชื้น) > ทิศทางลม Wind_North >ทิศทางลม Wind_SW > Season_Summer > ทิศทางลม Wind_ SSE > ทิศทางลม Wind_SE > ทิศทางลม Wind_South >> ทิศทางลม Wind_NW > ทิศทางลม Wind_NNE > ทิศทางลม Wind_East

ขั้นตอนที่ 4 : การสร้างแบบจำลอง

ในการวิจัยนี้ ได้ใช้ แบบจำลอง ทั้งหมด 4 แบบ ที่ได้รับความนิยมและยอมรับในปัจจุบัน ได้แก่ LR (Linear Regression), SVR (Support Vector Regression), XGBoost (eXtreme Gradient Boosting) และ MLP(Multi-Layer Perceptron) จากนั้นเริ่มสร้างแบบจำลองในการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 ซึ่งแบ่งเป็น 4 แบบย่อย โดยแต่ละแบบนั้นครอบคลุมการทำนายค่า PM2.5 หรือใช้ Target PM2.5 ของชั่วโมงที่ต่างกัน ในช่วงเวลา +1 ชั่วโมง, +6 ชั่วโมง, +12 ชั่วโมง, และ +24 ชั่วโมงล่วงหน้าตามลำดับ ซึ่งแต่ละแบบจำลองได้ ทำการ Scaling ข้อมูล ใน 2 รูปแบบคือ Standard Scaling และ Min-Max Scaling เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมสำหรับการสร้างและฝึกสอนแบบจำลอง และยังได้ทดลองเลือกใช้คุณลักษณะที่รวมทั้งค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ใน 24,36,72 ชั่วโมงย้อนหลัง และไม่ได้รวมค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง เพื่อทดสอบสมมุติฐานที่เกี่ยวข้อง นอกจากนี้ยังได้ทดลองเลือกใช้คุณลักษณะข้อมูลฤดูกาล (Season) และแบบไม่รวมข้อมูลฤดูกาล เพื่อทดสอบความสามารถของแบบจำลองในการทำนายค่า PM2.5 ในเรื่องข้อมูลสภาพแวดล้อมที่เปลี่ยนแปลงตามฤดูกาล ผลลัพธ์ที่ได้จะช่วยในการเลือกและนำเสนอแบบจำลองที่ดีที่สุดสำหรับการทำนายค่า PM2.5 ในบริบทที่ต่าง ๆ

ภายในข้อมูลจะถูกแบ่งออกเป็น 2 ชุด ได้แก่ Train Dataset คือชุดที่ใช้สำหรับให้แบบจำลองเรียนรู้ โดยจะใช้ข้อมูลทั้งหมดที่เกิดขึ้นในปี 2019 ซึ่งมีทั้งหมด 8,634 ตัวอย่าง และอีกชุด Test Dataset เป็นชุดข้อมูลสำหรับทดสอบประสิทธิภาพของแบบจำลอง โดยจะใช้ข้อมูลทั้งหมดที่เกิดขึ้นในปี 2020 ซึ่งมีทั้งหมด 6,441 ตัวอย่าง สัดส่วนของ Test Dataset ในที่นี้คือประมาณ 42.71% ในงานวิจัยนี้ ได้ทดลองสร้างแบบจำลองเบื้องต้น โดยใช้ ค่าพารามิเตอร์เริ่มต้น จาก scikit-learn ดังนี้ (scikit-learn)

ขั้นตอนที่ 5 : การประเมินผล

ตัวชี้วัดประสิทธิภาพของแบบจำลอง

การประเมินผล (Evaluation) ของแบบจำลอง เป็นขั้นตอนที่สำคัญในการตรวจสอบความแม่นยำของแบบจำลอง โดยจะใช้ข้อมูลที่ไม่เคยใช้ ในการให้แบบจำลองเรียนรู้ เพื่อดูว่าแบบจำลองสามารถทำนายผลได้แม่นยำมากน้อยเพียงใด การประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องสำหรับทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 ในแบบจำลอง Regression มีวิธีการประเมินผลหลักๆ 4 วิธี [8]

1. R-squared (R2)

ค่าระดับความใกล้เคียงระหว่างผลการทำนายกับข้อมูลจริง หรือ การบ่งชี้ระดับความถูกต้องแม่นยำ Accuracy มีน่าเชื่อถือมากเพียงใด สามารถอธิบายข้อมูลได้ดีแค่ไหน โดยค่า R2 จะอยู่ในช่วง 0-1 โดยค่าที่ใกล้เคียง 1 หมายถึงแบบจำลองทำนายได้ดี

2. Mean Absolute Error (MAE)

Mean Absolute Error (MAE): MAE คือค่าเฉลี่ยของความคลาดเคลื่อนในค่าการทำนายกับข้อมูลจริง ทุกรายการในชุดข้อมูลทดสอบ หรือ Test Set ยิ่งค่า MAE น้อยแสดงว่า คลาดเคลื่อนน้อย มีแม่นยำสูง

3. Root Mean Squared Error (RMSE)

นำ MSE ไปหารากที่สอง หลักการคือ นำ MSE มาถอดรากที่สอง Square root โดยค่าที่ได้ จะเป็นหน่วยเดียวกันกับค่า y ยิ่งค่า RMSE น้อย แสดงว่า คลาดเคลื่อนน้อย มีแม่นยำสูง

4. Mean Absolute Percentage Error (MAPE)

MAPE ใช้ในงานที่ต้องการวัดประสิทธิภาพของการทำนาย ซึ่งบอกถึงความคลาดเคลื่อนในรูปแบบร้อยละ หรือ เปอร์เซ็นต์

ผลการวิจัยและอภิปรายผลการวิจัย

ในแต่ละการทดลองจะมีการสร้างแบบจำลองใน 4 รูปแบบ เพื่อทำการศึกษาเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลอง

ตารางที่ 1 คุณลักษณะพื้นฐานที่ใช้ในการสร้างแบบจำลอง

หมวดหมู่	คุณลักษณะ
ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5)	PM2.5, PM2.5(h-1), PM2.5(h-6), PM2.5(h-12), PM2.5(h-24)
ข้อมูลอุตุนิยมวิทยาและสภาพอากาศ	Temperature_C, Dew_Point_C, Humidity_%, Speed_kmh, Gust_kmh, Pressure_hPa, Wind_North, Wind_NNE, Wind_NE, Wind_ENE, Wind_East, Wind_ESE, Wind_SE, Wind_SSE, Wind_South, Wind_SSW, Wind_SW, Wind_WSW, Wind_West, Wind_WNW, Wind_NW, Wind_NNW

โดยมีการเพิ่มตัวแปรด้านค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง และตัวแปรด้านฤดูกาล (Season) ในข้อมูลนำเข้า Input Data เพื่อใช้ในการสร้างแบบจำลอง

1. Without Mean_PM and Season

การสร้างแบบจำลองโดยใช้คุณลักษณะพื้นฐาน ไม่เพิ่มคุณลักษณะค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง และ ไม่เพิ่มคุณลักษณะข้อมูลฤดูกาล (Season)

2.Mean_PM Without Season

การสร้างแบบจำลองโดยการเพิ่มคุณลักษณะค่าเฉลี่ยPM2.5 ย้อนหลัง 24,48,72 ชั่วโมงย้อนหลัง

3.Only Season

การสร้างแบบจำลองโดยการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Season)

4.Mean_PM and Season

การสร้างแบบจำลองโดยการเพิ่มคุณลักษณะค่าเฉลี่ยPM2.5 ย้อนหลัง และข้อมูลฤดูกาล(Season)

ตาราง 2 คุณลักษณะเพิ่มเติมที่ใช้ในการสร้างแบบจำลอง

คุณลักษณะเพิ่มเติมที่ใช้ในการสร้างแบบจำลอง	**ชื่อคุณลักษณะเพิ่มเติม**	**จำนวนคุณลักษณะทั้งหมดที่ใช้ในการสร้างแบบจำลอง**
Without Mean_PM and Season แบบจำลองโดยใช้คุณลักษณะพื้นฐาน	คุณลักษณะพื้นฐาน	27 คุณลักษณะ
Mean_PM Without Season แบบจำลองโดยการเพิ่มคุณลักษณะค่าเฉลี่ย PM2.5 ย้อนหลัง	Last24hrs_mean Last48hrs_mean Last72hrs_mean	30 คุณลักษณะ
Only Season แบบจำลองโดยการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Season)	Season_Rainy Season_Summer Season_Winter	30 คุณลักษณะ
Mean_PM and Season แบบจำลองโดยการเพิ่มคุณลักษณะค่าเฉลี่ย **PM**2.5 ย้อนหลัง และข้อมูลฤดูกาล(**Season)**	Last24hrs_mean Last48hrs_mean Last72hrs_mean Season_Rainy Season_Summer Season_Winter	33 คุณลักษณะ

การเปรียบเทียบผลการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 ล่วงหน้าที่ช่วงเวลาต่างๆ

ผลเปรียบเทียบการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 1,6 ชั่วโมงล่วงหน้า ตามลำดับ

1 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 1 ชั่วโมงล่วงหน้า

จากผลลัพธ์ประสิทธิภาพของแบบจำลอง 1 ชั่วโมงล่วงหน้า จะเห็นได้ว่าแบบจำลอง Linear Regression (LR+1) โดยการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Only Season) ให้ประสิทธิภาพที่ดีที่สุดในการทำนายฝุ่นละอองขนาดเล็ก PM2.5 โดยมีค่า R2 Score สูงที่สุดและค่า MAE MSE RMSE และ MAPE ต่ำที่สุด ด้วยค่า R2 Score 0.9722, MAE: 1.6832, RMSE: 2.4492, MAPE (%): 9.0302 แบบจำลอง LR+1 ให้ผลลัพธ์สูงที่สุด รองลงมาเป็น MLP+1, XBG+1 และ SVR+1 ตามลำดับ สังเกตได้ว่าใช้ Mean_PM Without Season และ Without Mean_PM and Season ในการทำนายของทุกแบบจำลอง ได้ผลลัพธ์น้อยที่สุด แสดงให้เห็นการใช้ข้อมูลฤดูกาลเป็นองค์ประกอบที่สำคัญในการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 ในชุดข้อมูลนี้ แบบจำลองที่ใช้ข้อมูลฤดูกาล เข้ามาเกี่ยวข้อง มีประสิทธิภาพที่ดีกว่าในการทำนาย ดังตารางผลแบบจำลองค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 1 ชั่วโมงล่วงหน้า

ตาราง 3 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 1 ชั่วโมงล่วงหน้า

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
LR+1	Training	R2	0.9583	0.9585	0.9585	0.9586
	Training	MAE	1.9209	1.9139	1.9179	1.9150
	Training	RMSE	2.8337	2.8268	2.8278	2.8256
	Training	MAPE (%)	8.2574	8.2262	8.2442	8.2372
	Test	R2	0.9722	0.972	0.9722	0.9721
	Test	MAE	1.6835	1.6929	1.6832	1.6891
	Test	RMSE	2.4523	2.4616	2.4492	2.4557
	Test	MAPE (%)	9.0376	9.0408	9.0302	9.051
SVR+1	Training	R2	0.9422	0.9417	0.9410	0.9398
	Training	MAE	2.2633	2.2672	2.3251	2.3425
	Training	RMSE	3.3371	3.3507	3.3728	3.4074
	Training	MAPE (%)	9.5282	9.5284	9.8089	9.8627
	Test	R2	0.931	0.9316	0.941	0.9384
	Test	MAE	2.5526	2.5338	2.4346	2.4744
	Test	RMSE	3.8621	3.8454	3.5850	3.6489
	Test	MAPE (%)	14.2835	13.9404	13.3705	13.493

ตาราง 3 (ต่อ)

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
MLP+1	Training	R2	0.9591	0.9593	0.9599	0.9590
	Training	MAE	1.9164	1.9060	1.8897	1.9182
	Training	RMSE	2.8084	2.8010	2.7816	2.8124
	Training	MAPE (%)	8.1553	8.1519	8.0995	8.2630
	Test	R2	0.9708	0.9711	0.9717	0.9713
	Test	MAE	1.7437	1.7345	1.7110	1.7384
	Test	RMSE	2.5101	2.4984	2.4709	2.491
	Test	MAPE (%)	9.1794	9.1034	9.0232	9.2409
XGB+1	Training	R2	0.9904	0.9891	0.9900	0.9896
	Training	MAE	1.0021	1.0676	1.0214	1.0539
	Training	RMSE	1.3578	1.4496	1.3876	1.4158
	Training	MAPE (%)	4.7157	4.9731	4.8139	4.9478
	Test	R2	0.9607	0.9577	0.9608	0.9571
	Test	MAE	2.0122	2.0781	2.0218	2.0865
	Test	RMSE	2.913	3.0221	2.9105	3.0451
	Test	MAPE (%)	10.7124	11.4137	10.8274	11.2541

2 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 6 ชั่วโมงล่วงหน้า

จากผลลัพธ์ประสิทธิภาพของแบบจำลอง 6 ชั่วโมงล่วงหน้า จะเห็นได้ว่าแบบจำลองLinear Regression (LR+6) โดยการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Only Season) ให้ประสิทธิภาพที่ดีที่สุดในการทำนายฝุ่นละอองขนาดเล็ก PM2.5 โดยมีค่า R2 Score สูงที่สุดและค่า MAE MSE RMSE และ MAPE ต่ำที่สุด ด้วยค่า R2 Score 0.7735, MAE: 4.9834, RMSE: 6.9925, MAPE (%): 26.0489 แบบจำลองการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Only Season) ในการทำนาย ได้ผลลัพธ์ที่ดีที่สุด เรียงอันดับ แบบจำลองได้ดังนี้ แบบจำลอง LR+6 ให้ผลลัพธ์สูงที่สุด รองลงมาเป็น MLP+6, SVR+6 และ XGB+6 ตามลำดับ

ตาราง 4 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 6 ชั่วโมงล่วงหน้า

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
LR+6	Training	R2	0.6848	0.6957	0.6922	0.6983
	Training	MAE	5.5284	5.3850	5.4550	5.3667
	Training	RMSE	7.7955	7.6593	7.7034	7.6272
	Training	MAPE (%)	23.2778	22.4870	22.8702	22.4579
	Test	R2	0.7655	0.7664	0.7735	0.7702
	Test	MAE	5.0893	5.0061	4.9834	4.9592
	Test	RMSE	7.1146	7.1015	6.9925	7.0270
	Test	MAPE (%)	26.8894	25.6720	26.0489	25.2962
SVR+6	Training	R2	0.6889	0.6998	0.6973	0.7035
	Training	MAE	5.3598	5.2281	5.2617	5.1769
	Training	RMSE	7.7449	7.6079	7.6389	7.5604
	Training	MAPE (%)	21.5709	21.0216	21.2401	20.8481
	Test	R2	0.7301	0.7312	0.7451	0.7423
	Test	MAE	5.2463	5.2170	5.1296	5.1187
	Test	RMSE	7.6330	7.6180	7.4175	7.4587
	Test	MAPE (%)	26.8390	26.3808	26.4585	25.9228
MLP+6	Training	R2	0.7109	0.7202	0.7225	0.7321
	Training	MAE	5.3224	5.1970	5.1893	5.0988
	Training	RMSE	7.4652	7.3447	7.3137	7.1870
	Training	MAPE (%)	22.0569	21.3085	21.4707	21.3095
	Test	R2	0.7501	0.7435	0.7594	0.7526
	Test	MAE	5.1294	5.1847	4.9916	5.0364
	Test	RMSE	7.3452	7.4416	7.2074	7.3080
	Test	MAPE (%)	24.7874	24.4267	24.2252	24.2278

ตาราง 4 (ต่อ)

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
XGB+6	Training	R2	0.9574	0.9696	0.9593	0.9714
	Training	MAE	2.1432	1.8217	2.0968	1.7509
	Training	RMSE	2.8670	5.8575	2.8003	2.3469
	Training	MAPE (%)	10.0522	8.6662	9.9132	8.2816
	Test	R2	0.6827	0.6992	0.6964	0.7029
	Test	MAE	5.9327	5.5039	5.6648	5.5025
	Test	RMSE	8.2759	8.0583	8.0953	8.0085
	Test	MAPE (%)	31.6585	27.5374	29.3036	27.8301

ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ทั้งหมด ใน 2 ช่วงเวลา ได้แก่ +1, +6 ชั่วโมงล่วงหน้า พบว่าแบบจำลอง LR- Linear Regression ผลลัพธ์ที่ดีที่สุด ทั้งในแง่ของความถูกต้องแม่นยำ และความคลาดเคลื่อนที่ต่ำลดลง รองลงมาเป็น แบบจำลอง MLP Multi-Layer Perceptron , SVR- Support Vector Regression และ แบบจำลอง XGBoost ตามลำดับ แบบจำลองทุกแบบ ผลลัพธ์จะลดลงตามช่วงเวลาที่มากขึ้น

จากการสร้างแบบจำลอง 4 กรณี โดยมีการเพิ่มตัวแปรด้านค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง และตัวแปรด้านฤดูกาล (Season) ในข้อมูลนำเข้า Input Data เพื่อใช้ในการสร้างแบบจำลอง พบว่าแบบจำลอง Only Season ที่มีการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Season) ให้ผลลัพธ์ที่ดีที่สุด ทั้งในแง่ของความถูกต้องแม่นยำ และความคลาดเคลื่อนที่ต่ำลดลง จากในทุกๆช่วงเวลา และจาก 4 แบบจำลองที่เพิ่มคุณลักษณะของข้อมูล การใช้ Only Season และ Mean_PM and Season แบบจำลองที่มีฤดูกาลเข้ามาเกี่ยวข้อง จะอยู่ในกลุ่มที่ให้ผลลัพธ์สูงกว่า แบบ Mean_PM Without Season และ Without Mean_PM and Season ที่ไม่ได้ใช้ข้อมูลอุตุนิยมวิทยา ฤดูกาลเข้ามาเกี่ยวข้อง

สรุปผลการวิจัย

ในงานวิจัยนี้เป็นการศึกษาโดยนำข้อมูลภาคอุตุนิยมวิทยาที่เกิดขึ้นในอดีตมาใช้ร่วมกับเทคนิคการเรียนรู้ของเครื่อง เพื่อสร้างแบบจำลองเบื้องต้นโดยใช้ ค่าพารามิเตอร์เริ่มต้น จาก scikit-learn ที่ใช้สำหรับในการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ล่วงหน้า สรุปได้ว่า ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ทั้งหมด พบว่าแบบจำลอง LR- Linear Regression ผลลัพธ์ที่ดีที่สุด ทั้งในแง่ของความถูกต้องแม่นยำ และความคลาดเคลื่อนที่ต่ำลดลง รองลงมาเป็น แบบจำลอง MLP Multi-Layer Perceptron , SVR- Support Vector Regression และ แบบจำลอง XGBoost ตามลำดับ ปัจจัยที่ทำให้แบบจำลอง LR- Linear Regression ได้ผลลัพธ์ที่ดีที่สุด มีได้หลายปัจจัย เช่น Linear Relationship มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรต้น

และตัวแปรตาม จากในชุดข้อมูล ความสัมพันธ์ของข้อมูลส่วนใหญ่ มีลักษณะเป็น Linear ส่งผลให้ตัวแบบจำลอง LR- Linear Regression สามารถเรียนรู้และทำนายผลได้แม่นยำที่สุด ส่วนแบบจำลอง MLP- Multi-Layer Perceptron เป็น แบบจำลอง Neural Network โครงข่ายเซลล์ประสาท ทำให้สามารถเรียนรู้ข้อมูลที่ซับซ้อนมากขึ้นได้ ทำให้ทำนายผลออกมาได้อย่างมีประสิทธิภาพ ส่วนแบบจำลอง SVR- Support Vector Regression และ แบบจำลอง XGBoost เนื่องจากในวิจัยนี้ได้สร้างแบบจำลองเบื้องต้นที่ใช้ ค่าพารามิเตอร์เริ่มต้น จาก scikit-learn ไม่ได้ตั้งค่า ซึ่งแบบจำลองทั้งสองนั้น สามารถกำหนดค่าพารามิเตอร์ ได้หลากหลายรูปแบบ ให้เหมาะสมกับข้อมูล เช่นค่า Kernel Functions เมื่อไม่ได้กำหนด อาจทำให้การเรียนรู้อาจจะไม่ได้เหมาะสมกับข้อมูลที่มี ส่งผลให้แบบจำลองเรียนรู้และทำนายได้อย่างไม่ค่อยมีประสิทธิภาพเท่าที่ควร แบบจำลองทุกแบบผลลัพธ์จะลดลงตามช่วงเวลาที่มากขึ้น และจากการสร้างแบบจำลอง 4 กรณี โดยมีการเพิ่มตัวแปรด้านค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง และตัวแปรด้านฤดูกาล (Season) ในข้อมูลนำเข้า Input Data เพื่อใช้ในการสร้างแบบจำลอง พบว่าแบบจำลอง Only Season ที่มีการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Season) ให้ผลลัพธ์ที่ดีที่สุด ทั้งในแง่ของความถูกต้องแม่นยำ และความคลาดเคลื่อนที่ต่ำลดลง จากในทุกๆช่วงเวลา และจาก 4 กรณีที่เพิ่มคุณลักษณะของข้อมูล การเพิ่มคุณลักษณะแบบที่ใช้ Only Season และ Mean_PM and Season ซึ่งเป็นแบบจำลองที่มีฤดูกาลเข้ามาเกี่ยวข้อง อยู่ในกลุ่มที่ให้ผลลัพธ์สูงกว่า แบบ Mean_PM Without Season และ Without Mean_PM and Season ที่ไม่ใช้ข้อมูลอุตุนิยมวิทยา ฤดูกาลเข้ามาเกี่ยวข้องในงานวิจัยนี้ ยังไม่ได้มีการปรับค่าพารามิเตอร์ ใช้ค่าเริ่มต้นจาก scikit-learn อาจมีผลต่อผลลัพธ์การทำนาย ในอนาคตหากมีการปรับจูน Model (Model Tuning) ค้นหาพารามิเตอร์ที่เหมาะสมที่สุด อาจช่วยทำให้ได้ผลของการประเมินแบบจำลองที่สูงขึ้นได้

กิตติกรรมประกาศ

การจัดทำวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีจากการสนับสนุน การให้ความช่วยเหลือ คำแนะนำ ตลอดจนแนวทางในการทำการวิจัย ของ ผศ.ดร.นภา แซ่เบ๊ ผู้เป็นอาจารย์ที่ปรึกษา ที่ได้กรุณาให้ความรู้ ข้อแนะนำและผลักดันมาโดยตลอด ตลอดจนขอขอบคุณ คณาจารย์ทุกท่านในภาควิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ และบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย

เอกสารอ้างอิง

[1] กระทรวงสาธารณสุข, ก. (2566). รายงานสถานการณ์และผลการดำเนินงานด้านการแพทย์และสาธารณสุข กรณี ฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน ปี 2565. กองประเมินผลกระทบต่อสุขภาพ กรมอนามัย กระทรวงสาธารณสุข. สืบค้นจาก https://hia.anamai.moph.go.th/th/handbook/3912#wow-book/.

[2] กรมควบคุมมลพิษ กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม. (2561). โครงการศึกษาแหล่งกำเนิดและแนวทางการจัดการฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน ในพื้นที่กรุงเทพและปริมณฑล. กรมควบคุมมลพิษ กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม. สืบค้นจาก

https://www.pcd.go.th/airandsound/%e0%b9%82%e0%b8%84%e0%b8%a3%e0%b8%87%e0%b8%81%e0%b8%b2%e0%b8%a3%e0%b8%a8%e0%b8%b6%e0%b8%81%e0%b8%a9%e0%b8%b2%e0%b9%81%e0%b8%ab%e0%b8%a5%e0%b9%88%e0%b8%87%e0%b8%81%e0%b8%b3%e0%b9%80%e0%b8%99%e0%b8%b4.

[3] Ejohwomu, O. A., Shamsideen Oshodi, O., Oladokun, M., Bukoye, O. T., Emekwuru, N., Sotunbo, A., & Adenuga, O. (2022). Modelling and Forecasting Temporal PM2.5 Concentration Using Ensemble Machine Learning Methods. Buildings, 12(1), 46. Retrieved from https://www.mdpi.com/2075-5309/12/1/46.

[4] Lee, Y. S., Choi, E., Park, M., Jo, H., Park, M., Nam, E., Kim, D. G., Yi, S.-M., & Kim, J. Y. (2023). Feature extraction and prediction of fine particulate matter (PM2.5) chemical constituents using four machine learning models. Expert Systems with Applications, 221, 119696. Retrieved from https://doi.org/https://doi.org/10.1016/j.eswa.2023.119696.

[5] Berkeley Earth. Download Hourly Data. Retrieved October 5, 2022, from https://data.berkeleyearth.org/air-quality/local/Thailand/Thailand.txt.

[6] Weather Underground. Weather Data. Retrieved October 5, 2022, from https://www.wunderground.com/weather/th/bangkok/IBANGK169.

[7] Karlheinzniebuhr. (2022). the-weather-scraper. Retrieved October 4, 2022, from https://github.com/Karlheinzniebuhr/the-weather-scraper.

[8] กอบเกียรติ สระอุบล. (2563). เรียนรู้ Data Science และ AI Machine learning ด้วย Python. มีเดีย เนทเวิร์ค.