การเรียนรู้ของเครื่องสำหรับการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

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

กชรัตน์ นฤพัฒน์ผจง

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2566

การเรียนรู้ของเครื่องสำหรับการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

กชรัตน์ นฤพัฒน์ผจง

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2566

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

KOJCHARAT NARUPATPAJONG

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2023

สารนิพนธ์

เรื่อง

การเรียนรู้ของเครื่องสำหรับการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

ของ

กชรัตน์ นฤพัฒน์ผจง

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.นภา แซ่เบ๊)

..... ประธาน

(อาจารย์ ดร.นิดา ชาติวัฒนศิริ)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

ชื่อเรื่อง	การเรียนรู้ของเครื่องสำหรับการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)
ผู้วิจัย	กชรัตน์ นฤพัฒน์ผจง
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2566
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. นภา แซ่เบ๊

สถานการณ์ปัจจุบันปัญหาฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5) เป็นปัญหาสำคัญของประเทศไทย งานวิจัยนี้มุ่งศึกษาการนำข้อมูลภาคอุตุนิยมวิทยามาใช้ร่วมกับเทคนิคการเรียนรู้ของเครื่องเพื่อสร้างแบบจำลองเบื้องต้นที่ใช้สำหรับในการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (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 หลักสี่ กรุงเทพมหานคร 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 ทั้งในแง่ของความถูกต้องแม่นยำและความคลาดเคลื่อนที่ต่ำลดลง 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 และ24 ชั่วโมงย้อนหลัง และข้อมูลภาคอุตุนิยมวิทยา

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

Title	MACHINE LEARNING MODEL FOR FORECASTING PARTICULATE MATTER CONCENTRATION (PM 2.5)
Author	KOJCHARAT NARUPATPAJONG
Degree	MASTER OF SCIENCE
Academic Year	2023
Thesis Advisor	Assistant Professor Dr. Napa Sae-Bae

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).

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

กิตติกรรมประกาศ

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

กชรัตน์ นฤพัฒน์ผจง

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ภูมิหลัง

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 (กระทรวงสาธารณสุข, 2566)ซึ่งก่อให้เกิดผลกระทบต่อสุขภาพมนุษย์ได้โดยตรงและกระทบต่อภาพลักษณ์ในฐานะศูนย์กลางการท่องเที่ยวและเศรษฐกิจของเอเชียตะวันออกเฉียงใต้

นอกจากนี้(กรมควบคุมมลพิษ กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม, 2561)ได้กล่าวถึงเรื่อง “การพัดพาและแปรสภาพของมลพิษ (transportation and transformation of pollutants)” ปัจจัยที่ส่งผลต่อการแพร่กระจายของมลพิษ ได้แก่ สภาพอุตุนิยมวิทยาและสภาพแวดล้อม โดยสภาพอุตุนิยมวิทยา ได้แก่ ฤดูมรสุมตะวันออกเฉียงเหนือ ตั้งแต่ช่วงกลางเดือนตุลาคมถึงกลางเดือนกุมภาพันธ์ และฤดูมรสุมตะวันตกเฉียงใต้ ตั้งแต่ช่วงกลางเดือนพฤษภาคมถึงกลางเดือนตุลาคม ซึ่งฤดูมรสุมนี้ ส่งผลต่อสภาพอากาศ รวมถึงความกดอากาศ ทิศทางลมประจำฤดู อุณหภูมิ ปริมาณฝน ความชื้น ปัจจัยเหล่านี้ทำให้ระดับฝุ่นละอองในฤดูมรสุมตะวันออกเฉียงเหนือในพื้นที่กรุงเทพมหานคร มีระดับสูงขึ้น เนื่องจากสภาพอากาศแห้ง และทิศทางลมตะวันออกเฉียงเหนือได้พัดพาฝุ่นละอองจากการเผาชีวมวลพื้นที่เกษตรกรรมในภาคกลางเข้าสู่พื้นที่กรุงเทพมหานคร ประกอบกับความแปรปรวนสภาพอากาศรายวัน ซึ่งหากในช่วงวันดังกล่าวมีอุณหภูมิต่ำ ความกดอากาศสูง ท้องฟ้าปิด สภาพอากาศสงบ ไม่กระจายตัว จะก่อให้เกิดการสะสมของมลพิษระดับมลพิษสูงกว่าปกติจากปัจจัยดังกล่าว ทำให้เกิดงานวิจัยนี้ โดยนำข้อมูลภาคอุตุนิยมวิทยาที่เกิดขึ้นในอดีตมาใช้รวมกับเทคนิคการเรียนรู้ของเครื่องเพื่อทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)ในอนาคต

การเรียนรู้ของเครื่อง (Machine Learning) เป็นเทคนิคกระบวนการที่ใช้คอมพิวเตอร์ในการสร้างแบบจำลอง โดยเรียนรู้จากข้อมูล ตัวแปรต่างๆ เพื่อทำนายผลลัพธ์ต่างๆ ใช้วิธีการทางสถิติและคณิตศาสตร์ เพื่อให้แบบจำลองสามารถทำนายผลลัพธ์ในสิ่งที่ต้องการได้อย่างแม่นยำ

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

1.2 ความมุ่งหมายของงานวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1. เพื่อสร้างแบบจำลองเบื้องต้นที่ใช้หลักการทำงานของการเรียนรู้ของเครื่องในรูปแบบต่างๆ สำหรับนำมาใช้ในการทำนายระดับความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคต โดยอาศัยตัวแปร ในภาคอุตุนิยมวิทยา เช่น อุณหภูมิ ความชื้น ความกดอากาศ และค่าเฉลี่ยของ PM2.5 ในช่วงเวลาต่างๆย้อนหลัง ฯลฯ มาใช้เป็นคุณลักษณะในการสร้างแบบจำลองการทำนายระดับความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคต

2. เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องในรูปแบบต่างๆ

1.3 ความสำคัญของการวิจัย

การวิจัยเกี่ยวกับการทำนายระดับความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคตนี้มีความสำคัญอย่างมาก จากสถานการณ์ในปัจจุบันเกิดปัญหาหมอกควันและฝุ่นละออง

โดยเฉพาะฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน (PM2.5) ด้วยขนาดอนุภาคที่เล็กมากของฝุ่น ทำให้ไม่สามารถมองเห็นด้วยตาเปล่า นอกจากนี้ขนจมูกของมนุษย์ ก็ไม่สามารถกรองฝุ่นนี้ได้ ทำให้ฝุ่นเหล่านี้สามารถซึมผ่านทางเดินหายใจของมนุษย์และเข้าสู่ระบบทางเดินหายใจได้ ก่อให้เกิดผลกระทบต่อสุขภาพมนุษย์ เช่น เช่น โรคหัวใจและหลอดเลือด โรคปอดเรื้อรัง ภาวะหอบหืด มะเร็งปอดแม้จะไม่สูบบุหรี่ (กระทรวงสาธารณสุข, 2566)

การทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) โดยใช้ข้อมูลทางด้านอุตุนิยมวิทยาในอดีต เช่น อุณหภูมิ ความชื้น ความกดอากาศ และค่าเฉลี่ยของ PM2.5 ในช่วงเวลาต่างๆย้อนหลัง เป็นต้น ในการสร้างแบบจำลองเบื้องต้นเพื่อทำนายปริมาณฝุ่น PM2.5 ในอนาคต จะช่วยสร้างความเข้าใจในผลกระทบต่อสุขภาพ ประเมินระดับความเสี่ยงที่เกิดจากฝุ่นละอองขนาดเล็ก ปรับเปลี่ยนพฤติกรรมการใช้พลังงาน การเดินทางได้ให้เหมาะสมกับสภาพอากาศและสิ่งแวดล้อม รวมถึง พัฒนาแนวทางการบริหารจัดการในระยะยาว ในการรับมือปริมาณฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคต

1.4 ขอบเขตของการวิจัย

1.4.1 กลุ่มตัวอย่างประชากรที่ใช้ในการวิจัย

ในการดำเนินงานวิจัยนี้ มีการใช้งานชุดข้อมูล 2 ชุดมารวมกัน โดยนำเข้าข้อมูลจากแหล่งข้อมูลสาธารณะแบบเปิด ผ่านหน้าเว็บไซต์และวิธีการ Web Scraping ซึ่งเป็นกระบวนการในการดึงข้อมูล โดยใช้สคริปต์ดึงข้อมูลจากหน้าเว็บไซต์ รายละเอียดดังนี้

1. ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) จากเว็บไซต์ Berkeley Earth 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 ซึ่งเป็นอันดับที่ 1 ในการให้บริการสภาพอากาศเชิงพาณิชย์ ที่ให้ข้อมูลสภาพอากาศแบบเรียลไทม์ผ่านทางอินเทอร์เน็ต ภารกิจคือทำให้ข้อมูลสภาพอากาศ มีคุณภาพ พร้อมใช้งานสำหรับทุกคนบนโลก ซึ่งเจ้าของคือ The Weather Company ซึ่งเป็นบริษัทย่อยของ IBM โดยข้อมูลภาคอุตุนิยมวิทยา ที่นำมา ผ่านวิธีการ 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

1.5 กรอบแนวคิดในงานวิจัย

กรอบแนวคิดในงานวิจัย การทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ตามที่ผู้วิจัยต้องการศึกษาดังนี้

1. กระบวนการรวบรวมข้อมูล (Data Acquisition)

กระบวนการในการรวบรวมข้อมูลที่เกี่ยวกับปัจจัยต่างๆ ที่นำไปสู่การทำนายระดับค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ได้แก่ ข้อมูลค่าความเข้มของฝุ่นละอองขนาดเล็ก(PM2.5) , ข้อมูลในภาคอุตุนิยมวิทยาในพื้นที่ที่สนใจ ซึ่งกระบวนการในการรวบรวมข้อมูล จะนำเข้าข้อมูลจากแหล่งสาธารณะแบบเปิด 2 ชุดข้อมูลมารวมกัน โดยผ่านหน้าเว็บไซต์และวิธีการ Web Scraping ข้อมูลจากแหล่งสถานีตรวจอากาศ เพื่อใช้เป็นข้อมูลในการสร้างแบบจำลองการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

2. การเตรียมข้อมูล (Data Preparation)

การนำข้อมูลที่ได้มา ทำความสะอาดข้อมูล (Data Cleansing) จัดการกับค่าผิดปกติ เช่น ข้อมูลที่กรอกไม่ครบถ้วน ขาดหาย หรือบางส่วนไม่จำเป็นต่อการวิเคราะห์ อาจจะต้องตัดทิ้ง เพื่อลดระยะเวลาในการประมวลผลของแบบจำลอง รวมถึงการแปลงข้อมูลคุณลักษณะใหม่ ให้อยู่ในรูปแบบที่สามารถนำไปสร้างแบบจำลองได้

3. การสำรวจข้อมูล (Exploratory Data Analysis: EDA)

ขั้นตอนหลังจากที่ได้ข้อมูลที่ทำความสะอาดแล้ว มาวิเคราะห์ค้นหาสิ่งที่แฝงอยู่ในข้อมูลนั้น รวมถึงการวิเคราะห์ทางด้านสถิติ ดูรูปแบบความสัมพันธ์ระหว่างตัวแปร เพื่อ ใช้เป็นคุณลักษณะในการสร้างแบบจำลองการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

4. การสร้างแบบจำลอง (Modelling)

การสร้างแบบจำลองพื้นฐานเพื่อทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคต ในวิจัยนี้มีการเลือกใช้แบบจำลอง 4 แบบ ประกอบด้วย LR (Linear Regression) , SVR (Support Vector Regression), XGBoost (eXtreme Gradient Boosting) และ MLP (Multi-Layer Perceptron) ซึ่งเป็นแบบจำลองที่นิยมและได้รับการยอมรับในปัจจุบัน

5. การประเมินผล (Evaluation)

การประเมินผลแบบจำลองที่สร้างขึ้น โดยใช้ตัวชี้วัดประสิทธิภาพ 4 วิธีได้แก่ ค่า R-squared (R2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) เพื่อประเมินความแม่นยำ ความคลานเคลื่อน ของแบบจำลอง ว่ามีมากน้อยเพียงใด เหมาะที่จะนำไปใช้งานหรือไม่

1.6 สมมุติฐานในการวิจัย

1. ค่า PM2.5 ในช่วง 24 ชั่วโมงย้อนหลัง (Last24hrs_mean), 48 ชั่วโมงย้อนหลัง (Last48hrs_mean) และ 72 ชั่วโมงย้อนหลัง (Last72hrs_mean) จะมีผลต่อความแม่นยำของแบบจำลองการทำนายความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

2. นอกจากตัวแปรด้านอุตุนิยมวิทยาแล้ว การใช้ตัวแปรด้านฤดูกาล จะมีผลต่อความแม่นยำของแบบจำลองการทำนายความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

1.7 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ทราบขั้นตอนในการสร้างชุดข้อมูลจากแหล่งข้อมูลสาธารณะ 2 แหล่งมารวมกัน เพื่อใช้ในการสร้างแบบจำลองการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5

2. สามารถสร้างแบบจำลองเบื้องต้นสำหรับนำมาใช้ในการทำนาย ความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5)

3. ผู้มีส่วนเกี่ยวข้อง หน่วยงานภาครัฐสามารถนำแบบจำลองการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) มาเป็นส่วนหนึ่งในการประกอบการตัดสินใจวางแผน ต่อยอดได้ในหลายๆด้าน เช่น

3.1. การจัดการและควบคุมมลพิษทางอากาศ

วางแผนการจัดการและควบคุมมลพิษทางอากาศได้อย่างมีประสิทธิภาพ การกำหนดนโยบายที่มีมาตรฐาน โดยเฉพาะอย่างยิ่งในสถานที่ ที่มีระดับมลพิษทางอากาศสูง เช่น ในเมืองหรือในโรงงาน โดยการสร้างพื้นที่สีเขียว การปลูกต้นไม้ และการใช้งานพลังงานทดแทน

3.2. การวางแผนพัฒนาทางเศรษฐกิจ

วางแผนการพัฒนาและวางนโยบาย ทางเศรษฐกิจได้อย่างมีประสิทธิภาพ โดยใช้ข้อมูลประกอบในการวิเคราะห์แนวโน้มของระดับมลพิษทางอากาศในเขตต่างๆ วางแผนทิศทางการพัฒนาทางเศรษฐกิจในระยะยาว รวมถึงภาคการท่องเที่ยว ที่ประเทศไทยในฐานะศูนย์กลางการท่องเที่ยวและเศรษฐกิจของเอเชียตะวันออกเฉียงใต้

3.3. การปรับปรุงคุณภาพชีวิตและสิ่งแวดล้อม

การลดระดับมลพิษทางอากาศเป็นสิ่งสำคัญต่อการปรับปรุงคุณภาพชีวิตของสิ่งมีชีวิต เพราะมลพิษก่อให้เกิดปัญหาในด้านสุขภาพ เช่น การเพิ่มความเสี่ยงของโรคหัวใจ โรคเกี่ยวกับทางเดินทางใจ โรคมะเร็ง นอกจากนี้ไม่ใช่เพียงแค่มนุษย์ที่ได้รับผลกระทบ มลพิษยังสะสมในพืชและสัตว์ แหล่งน้ำส่งผลต่อการเจริญเติบโตของพืชและสัตว์ หากรู้ทันเหตุการณ์ก็ช่วยลดความเสี่ยงลงได้

3.4. การเพิ่มความรู้ ต่อยอดการสร้างนวัตกรรม พัฒนาเทคโนโลยี

การวิจัยช่วยให้สามารถสร้างการรับรู้ ตระหนักถึงปัญหา เตรียมการรับมือ รวมถึงได้ความรู้ใหม่ เกิดการเรียนรู้ที่สามารถนำไปใช้ต่อยอดในงานวิจัยอื่น ๆ หรือในการพัฒนาแนวทางการแก้ไขปัญหาต่าง ๆ ในสาขาวิชาชีพต่าง ๆ เช่น เทคโนโลยีทางอากาศ

บทที่ 2
ทบทวนวรรณกรรม

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

1. ฝุ่นละอองขนาดเล็ก PM 2.5
2. แหล่งข้อมูลฝุ่นละอองขนาดเล็ก PM2.5 และสภาพอากาศ
3. แบบจำลองการทำนายฝุ่นละอองขนาดเล็ก PM2.5
4. งานวิจัยที่เกี่ยวข้อง (Literature Review)

2.1 ฝุ่นละอองขนาดเล็ก PM 2.5

PM หรือ Particulate Matters เป็นคำเรียกค่ามาตรฐานของฝุ่นละอองขนาดเล็กที่เป็นอันตรายต่อสุขภาพซึ่งมีอยู่ด้วยกัน 2 ชนิด ได้แก่ PM 10 และ PM 2.5 (สำนักหอสมุดมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, 2565) ฝุ่นละอองขนาดเล็ก PM 2.5 (Particulate Matter 2.5) หรือเรียกว่า ฝุ่นละเอียด (Final PM2.5 Particles) คือ อนุภาคฝุ่นละอองในอากาศที่มีขนาดเส้นผ่าศูนย์กลางไม่เกิน 2.5 ไมครอน มีหน่วยเป็น ไมโครกรัมต่อลูกบาศก์เมตร (μg/m3) ฝุ่นละอองจะแขวนลอยอยู่ในอากาศรวมกับไอน้ำ ควัน และก๊าซต่าง ๆ แต่ถึงจะเป็นเพียงฝุ่นละอองขนาดเล็ก ไม่สามารถมองเห็นได้ด้วยตาเปล่า แต่เมื่ออยู่รวมกัน จะกินพื้นที่ในอากาศมหาศาล ลอยอยู่ในชั้นบรรยากาศปริมาณสูง จนมองเห็นได้ในลักษณะที่คล้ายกับมีหมอกควัน นอกจากนี้ด้วยขนาดของฝุ่นละอองที่มีขนาดเล็ก จนขนจมูกของมนุษย์ ที่ทำหน้าที่กรองฝุ่น ไม่สามารถกรองได้ ก่อให้เกิดผลกระทบต่อสุขภาพมนุษย์ได้โดยตรง ทั้งแบบเฉียบพลัน อาทิ แสบตา คัดจมูก ภูมิแพ้ ฯลฯ และแบบเรื้อรัง สามารถแพร่กระจายเข้าสู่ทางเดินหายใจ กระแสเลือด และเข้าสู่อวัยวะอื่น ๆ ในร่างกายได้ นอกจากนี้ ตัวฝุ่นยังเป็นพาหะนำสารอื่น เข้ามาสู่ร่างกายด้วย เช่น แคดเมียม ปรอท โลหะหนัก และสารก่อมะเร็งอีกด้วย (สำนักหอสมุดมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, 2565)

โดยค่ามาตรฐานของประเทศไทย ที่ประกาศของทางคณะกรรมการสิ่งแวดล้อมแห่งชาติ ได้กำหนดมาตรฐานฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน ในบรรยากาศโดยทั่วไป ดังนี้ “ค่าเฉลี่ยในเวลา 24 ชั่วโมง จะต้องไม่เกิน 50 ไมโครกรัมต่อลูกบาศก์เมตร โดยให้มีผลจนถึงวันที่ 31 พฤษภาคม พ.ศ. 2566 และตั้งแต่วันที่ 1 มิถุนายน พ.ศ. 2566 เป็นต้นไป ให้ค่าเฉลี่ย ในเวลา 24 ชั่วโมง จะต้องไม่เกิน 37.5 ไมโครกรัมต่อลูกบาศก์เมตร” (ราชกิจจานุเบกษา, 2565)

National Geographic (NGThai, 2020)ได้กล่าวถึง “สาเหตุที่ทำให้เกิดฝุ่นละอองขนาดเล็ก (PM2.5)” ส่วนใหญ่มักเกิดจากการเผาไหม้ที่ไม่สมบูรณ์ เช่น การเผาไหม้จากเครื่องยนต์ดีเซล โรงงานอุตสาหกรรม การเผาไหม้ในที่โล่งของภาคการเกษตร การคมนาคม ไอเสียจากรถยนต์ รวมถึงภัยธรรมชาติอย่างไฟป่าและภูเขาไฟระเบิด และยังได้กล่าวถึงอีกหนึ่งสาเหตุของสิ่งที่ทำให้ฝุ่นละอองลอยขึ้นสู่ชั้นบรรยากาศ คือ “สภาวะของปรากฏการณ์อุณหภูมิผกผัน” คือ ช่วงที่ชั้นบรรยากาศมีอุณหภูมิสูงกว่าบริเวณที่อยู่ใกล้เคียงกัน เกิดขึ้นได้ในบริเวณที่มีการระบายความร้อนของพื้นผิวดินหรือทะเล โดยเฉพาะในช่วงเวลาที่มีแสงแดดจัด และไม่มีลมพัดเข้ามา ซึ่งในช่วงเวลากลางคืนหรือฤดูหนาว อุณหภูมิเหนือพื้นดินจะมีความเย็นกว่าอากาศด้านบน เนื่องจากมีการคายความร้อนของพื้นผิวโลก จึงทำให้เกิดปรากฏการณ์อุณหภูมิผกผัน (Inversion Layer) ชั้นบรรยากาศเป็นจึงเปรียบเหมือนโดมครอบพื้นที่ไว้ ทำให้ฝุ่นละอองไม่สามารถขึ้นสู่ด้านบนได้ และในเวลาปกติชั้นบรรยากาศจะไล่อุณหภูมิความร้อนบริเวณพื้นดินขึ้นไปสู่ชั้นบรรยากาศด้านบน ทำให้ฝุ่นต่างๆ ลอยขึ้นสูงและถูกกระแสลมพัดออกไป และสะสมจนกลายเป็นฝุ่นควันฟุ้งกระจายทั่วเมืองในที่สุด

2.2 แหล่งข้อมูลฝุ่นละอองขนาดเล็ก PM2.5 และสภาพอากาศ

1. Berkeley Earth

โครงการวิจัยทางวิทยาศาสตร์ข้อมูลและสิ่งแวดล้อม ที่โปร่งใส เชื่อถือได้ ที่มุ่งเน้นวิเคราะห์และวัดสภาพอากาศ ภูมิอากาศของโลก โครงการนี้มีวัตถุประสงค์หลักในการรวบรวมข้อมูลจากหลายแหล่งทั่วโลก เพื่อศึกษาและทำความเข้าใจการเปลี่ยนแปลงของสภาพอากาศที่มีอยู่ตลอดเวลา รวมถึงให้ข้อมูลมลพิษทางอากาศทั่วโลกแบบโอเพนซอร์สที่ครอบคลุมข้อมูลอุณหภูมิโลก เข้าถึงได้ง่าย ทันเวลา เป็นกลางและตรวจสอบได้ Berkeley Earth เกิดขึ้นโดยคุณ Richard และ คุณ Elizabeth Muller ในต้นปี 2010 พวกเขารวมกลุ่มนักวิทยาศาสตร์เพื่อวิเคราะห์ บันทึกอุณหภูมิพื้นผิวโลกและเผยแพร่ ตั้งแต่ปี 2010-2012 และในปี 2012 กลายเป็นองค์กรไม่แสวงหาผลกำไร อิสระ เป็นต้นมา โดย Berkeley Earth นำเข้าข้อมูลการวัดสภาพอากาศจากหลายแหล่งต่าง ๆ ทั่วโลก ซึ่งเป็นข้อมูลสาธารณะแบบเปิด ถูกรวบรวมไว้บนหน้าเว็บไซต์ โดย Berkeley Earth มุ่งมั่นในการให้ข้อมูลที่มีคุณภาพและเชื่อถือได้ ทั้งในด้านวิศวกรรมข้อมูลและการนำเสนอผลลัพธ์ของการวิจัย ในด้านการเผยแพร่ข้อมูลของ Berkeley Earth จะผ่านพื้นที่ออนไลน์หรือผ่านแหล่งทางวิชาการเพื่อให้นักวิจัยและประชาชนทั่วไปสามารถเข้าถึงข้อมูล ในงานวิจัยนี้ได้นำข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) ที่มีความน่าเชื่อถือ จาก Berkeley Earth มาใช้งาน (Berkeley Earth)

2. Weather Underground

เว็บไซต์และแอพพลิเคชันที่ให้บริการข้อมูลสภาพอากาศและพยากรณ์อากาศ ซึ่งเป็นอันดับที่ 1 ในการให้บริการสภาพอากาศเชิงพาณิชย์ บริการข้อมูลสภาพอากาศแบบเรียลไทม์ผ่านทางอินเทอร์เน็ต ภารกิจคือทำให้ข้อมูลสภาพอากาศ มีคุณภาพ พร้อมใช้งานสำหรับทุกคนบนโลก ซึ่งเจ้าของคือ The Weather Company ซึ่งเป็นบริษัทย่อยของ IBM โดยข้อมูลภาคอุตุนิยมวิทยา เป็นข้อมูลสภาพอากาศที่ครอบคลุมทั่วโลก เป็นข้อมูลสภาพอากาศจริงจากสถานีวัดต่าง ๆ รวมถึงข้อมูลเช่น อุณหภูมิ, ความชื้น, แรงลม, และความดันอากาศ ข้อมูลพยากรณ์ระยะสั้นและระยะยาว ที่ทางเว็บไซต์มีการอัปเดตอยู่ตลอดเวลา ซึ่งเป็นแหล่งข้อมูลที่สำคัญและเชื่อถือได้(Weather Underground)

2.3 แบบจำลองการทำนายฝุ่นละอองขนาดเล็ก PM2.5

ในงานวิจัยนี้ ได้ใช้แบบจำลอง Machine Learning 4 แบบ ดังนี้

1. Linear Regression

Linear Regression คือ อัลกอริทึมหนึ่งใน Supervised Learning ใช้สำหรับการวิเคราะห์และคาดการณ์ ทำนายค่าของตัวแปรตาม (Dependent Variable) จากตัวแปรอิสระ (Independent Variable) โดย Linear Regression จะใช้แบบจำลองการทำนายซึ่งแทนด้วยสมการเชิงเส้น เป็นการวาดเส้นตรง (line) บนกราฟ ที่เชื่อมโยงความสัมพันธ์ระหว่างตัวแปรอิสระ x (Independent Variable) และตัวแปรตาม y (Dependent Variable) โดยใช้ข้อมูล Training Data พยากรณ์ค่า y จากค่า x ที่มีอยู่แล้ว ในการเรียนรู้และปรับค่าพารามิเตอร์ให้เหมาะสม เพื่อทำนายค่าตัวแปรที่ต้องการในข้อมูลอื่น ๆ (Jay Chugh, 2018)

โดยสมการของ Linear Regression มีดังนี้:

$$y = b0 + b1 * x$$

โดยที่:

y = Dependent variable (ตัวแปรตาม)

x = Independent variable (ตัวแปรอิสระ)

b0 = ค่าคงที่ (Intercept)

b1 = ค่า Slope หรือความชันของเส้นตรง (Coefficient)

Linear Regression เป็นแบบจำลอง ที่ง่ายต่อการนำไปใช้งานและอธิบายผลได้เข้าใจง่าย เหมาะสำหรับชุดข้อมูลที่มีคุณลักษณะ Feature ไม่มากนัก และไม่ซับซ้อน ทำงานได้ดีก็ต่อเมื่อข้อมูลมีความสัมพันธ์เชิงเส้นกัน ถ้าข้อมูลเป็นแบบค่าของตัวแปรตามเป็น Binary จะไม่สามารถใช้ได้ รวมถึงข้อมูลที่มีค่า Outlier มาก ทำให้ผลลัพธ์ไม่แน่นอน และมีความคลาดเคลื่อนสูง

2. Support Vector Regression (SVR)

เป็นหนึ่งในอัลกอริทึม Supervised Learning ในกลุ่ม Regression แบบสมการเชิงเส้น โดย SVR (Support Vector Regression) เป็นอัลกอริทึมสำหรับการสร้างแบบจำลองเพื่อทำนายค่าตัวเลข สามารถคาดเดาค่าที่ไม่ต่อเนื่องได้ ใช้ความแตกต่างของค่าของตัวแปรตาม y จากตัวแปรอิสระ x เพื่อสร้างแบบจำลองที่สามารถทำนายค่าตัวแปรตาม y ได้ โดยมีการใช้เทคนิคคล้าย Support Vector Machine (SVM) แต่เปลี่ยนจากการจำแนกเป็นการทำนายค่า โดย SVR จะใช้ Kernel Functions วิธีการหาค่าของเส้นโค้ง (curve) ที่เหมาะสมที่สุดในการทำนายค่าตัวแปรตาม y โดยที่แต่ละ Kernel Function จะเลือกใช้ตามรูปแบบของข้อมูลและวัตถุประสงค์ที่ต้องการ ซึ่งส่งผลให้ SVR สามารถใช้งานกับข้อมูลที่มีความซับซ้อนและขนาดใหญ่ได้ดีกว่า Linear Regression

โดยสมการของ Support Vector Regression (SVR) มีดังนี้:

$$y = w^T x + b = \sum_{i=1}^{n} \alpha_i \, k(x_i, x) + b$$

โดยที่:

y คือ ค่าที่ต้องการทำนาย

w คือ เวกเตอร์น้ำหนัก (Weight Vector)

x คือ เวกเตอร์ข้อมูลเข้า (Input Vector)

b คือ ค่าไบแอส (Bias)

n คือ จำนวนข้อมูล (Data Points)

alpha_i คือ ค่าเวกเตอร์ของ Lagrange Multiplier

k (x_i,x) คือ ฟังก์ชันความสัมพันธ์ระหว่างเวกเตอร์ข้อมูลเข้า

ฟังก์ชันความสัมพันธ์ k ที่ใช้กับ SVR สามารถเลือกได้หลากหลาย โดยที่ Gaussian Kernel (RBF Kernel) ถือว่าเป็นที่นิยมมากที่สุด เช่น ข้อมูลที่ไม่เป็นเชิง โดยมีสมการดังนี้:

$$K(x, y) = exp\left(-gamma * ||x - y||^2\right)$$

โดยที่:

x และ y คือ ข้อมูลตัวอย่างแต่ละตัวที่จะนำมาทำการวิเคราะห์

gamma คือ พารามิเตอร์ที่ควบคุมความชันของคำนวณ

ในการเลือกค่า gamma นั้นจะต้องพิจารณาตามปริมาณข้อมูลที่มีอยู่ เนื่องจากการเลือก gamma ไม่เหมือนกันตามแต่ละกรณีและอาจทำให้ผลลัพธ์ที่ต่างกัน

Support Vector Regression (SVR) เป็นแบบจำลองที่นิยมใช้ในการทำนายข้อมูลที่มีความซับซ้อน ใช้กับตัวแปรตาม ที่เป็น Non-Linear ได้ สามารถใช้งานได้กับชุดข้อมูลที่มีจำนวนตัวอย่างน้อยได้ การใช้ SVR ก็ต้องมีการแยกชุดข้อมูลออกเป็นชุด Train และ Test เพื่อป้องกัน Overfitting และไม่สามารถจัดการ Outlier ได้ดี และการทำงานของ SVR อาจช้ากว่าแบบอื่นๆ เนื่องจากมีการเพิ่ม Kernel Function เข้าไปในการประมวลผล

3. XGBoost (eXtreme Gradient Boosting)

Extreme Gradient Boosting หรือเรียกสั้น ๆ ว่า XGBoost การเรียนรู้ในกลุ่ม Ensemble ถือเป็น แบบจำลองที่ล้ำสมัย การพัฒนาและเปิดตัวครั้งแรกของ XGBoost ที่การแข่งขัน Machine learning ของ Kaggle ในช่วงปี 2015 ในบรรดา 29 โซลูชันที่ชนะนั้น มีถึง 17 โซลูชัน ที่ใช้ XGBoost (Jason Brownlee, 2021)อัลกอริทึมถูกพัฒนาขึ้นมา เพื่อแก้ปัญหาการ Regression Predict หรือ Classification ที่มี Dataset ขนาดใหญ่ โดยพัฒนาขึ้นจากแนวคิดพื้นฐานเดียวกับ Gradient Boosting โดยใช้แบบจำลอง Boosting ที่เป็น Ensemble Learning Algorithm ที่รวมแบบจำลองหลายๆ แบบจำลองเข้าด้วยกัน เพื่อสร้างแบบจำลองที่มีความแม่นยำสูงกว่าแบบจำลองเดียว ซึ่ง XGBoost จะทำการเรียนรู้จากข้อมูลที่มีอยู่แล้ว และทำการ Optimize Hyperparameter เพื่อให้แบบจำลองทำนายได้อย่างแม่นยำ (บัญชา ปะสีละเตสัง, 2564)

XGBoost เป็น Gradient Boosting ที่มีประสิทธิภาพมากขึ้น โดยเพิ่มการประมวลผลแบบขนาดใหญ่ (Parallel Processing) ป้องกันการเกิด Overfitting ด้วย Regularization แต่สามารถเกิด Overfitting หรือ Underfitting ได้ง่าย หากมีการเลือก Hyperparameter ที่ไม่เหมาะสม มีการ

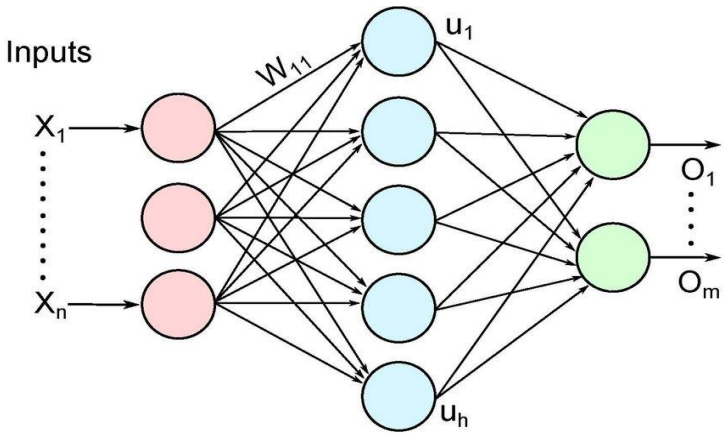
จัดการค่า Missing Values ได้อย่างดี และสามารถประมวลผลข้อมูลหลายมิติได้ เช่น ข้อมูลแบบช่วงเวลา (Time Series) และข้อมูลแบบที่มีการแบ่งกลุ่ม (Clustering)

สำหรับ สมการของ XGBoost ไม่ได้มีสมการเดียวกันที่ใช้ในทุกๆ กรณี เนื่องจากการประมวลผลของ XGBoost เป็นการเพิ่มความซับซ้อนของแบบจำลองโดยใช้ค่าความคลาดเคลื่อน (Loss Function) ที่ต้องการลดลงในแต่ละรอบการเรียนรู้ สามารถกำหนด Loss Function และสามารถกำหนดพารามิเตอร์ที่เหมาะสมเพื่อปรับความซับซ้อนของแบบจำลอง
ดังนั้น สมการของ XGBoost จะไม่เท่ากัน ขึ้นอยู่กับแต่ละปัญหาและการกำหนดพารามิเตอร์ของแต่ละแบบจำลอง (Chen & Guestrin, 2016)

ค่าพารามิเตอร์ ของ XGBoost
n_estimators: จำนวนต้นไม้ที่ต้องการสร้างในแบบจำลอง
max_depth: ความลึกของต้นไม้
eta,learning_rate: อัตราการเรียนรู้ของแบบจำลอง
subsample: สัดส่วนของข้อมูลที่จะถูกสุ่มเพื่อใช้ในการฝึกแบบจำลองในแต่ละรอบ
colsample_bytree: สัดส่วนของคอลัมน์ของข้อมูลที่จะถูกสุ่มเพื่อใช้ในการฝึกแบบจำลองในแต่ละรอบ

4. MLP หรือ Multi-Layer Perceptron (MLP Regressor)

MLP (Multi-Layer Perceptron) เป็นแบบจำลอง Neural Network โครงข่ายเซลล์ประสาท จัดอยู่ในกลุ่ม Supervised Learning สามารถใช้ได้ทั้งงาน Regression และ Classification ข้อมูลที่ไม่เป็นเชิงเส้น (Non-Linear Data) ได้ มีความซับซ้อนมากขึ้นจาก Single-Layer Perceptron โดยที่มีการเชื่อมต่อกันของ Node (หรือ Neuron) ใน Layer ต่างๆ โดยมีหลาย Hidden Layer ที่มีหน้าที่เปลี่ยนแปลง Feature ของข้อมูลเพื่อทำให้แบบจำลองมีความสามารถในการจำแนกและทำนายข้อมูลได้ดียิ่งขึ้น *(บัญชา ปะสีละเตสัง, 2564)*



ภาพประกอบ 1 Multi-Layer Perceptron

ที่มา : (Yamany, Waleed & Fawzy, Mohammed & Tharwat, Alaa & Hassanien, Aboul Ella ,2015)(Yamany, 2015)

MLP จะแบ่งโครงสร้างออกเป็น 3 ชั้น Layer ประกอบด้วย

Input Layer เป็นชั้นที่รับข้อมูลเข้ามาจากภายนอก แต่ละInput จะมีค่าน้ำหนัก Weight ของตัวเอง ค่าสัมประสิทธิ์ (Coefficient) เพื่อบ่งชี้ความสำคัญต่อ Output

Hidden Layer เป็นชั้นกลางระหว่าง Input Layer และ Output Layer ซึ่งมีได้หลายชั้น (Multiple Hidden Layer) และแต่ละชั้นอาจมีจำนวนนิวรอน Neuron ได้มากกว่า 1 อัน เพื่อให้แบบจำลองสามารถเรียนรู้และสกัดคุณลักษณะของข้อมูลได้มากขึ้น

Output Layer เป็นชั้นที่แสดงผลลัพธ์ ซึ่งรับค่ามาจาก Hidden Layer โดยชั้นนี้ อาจมีจำนวนนิวรอน Neuron ได้มากกว่า 1 อันเช่นกัน

วิธีการคำนวณของ MLP ยังคงเดิมเหมือนกัน กับ Single-Layer Perceptron นอกจากนี้มีรายละเอียดที่ต้องกำหนดเพิ่มเติม คือ ค่าน้ำหนัก Weight ของ Input แต่ละค่า, ค่า Bias ที่เติมลงไป ต้องมีค่า Bias สำหรับแต่ละนิวรอน ถ้าเทียบสมการเส้นตรง คือ จุดตัดแกน y, Activation Function ใช้ในการแปลงนิวรอน หลังการคูณ Input เข้ากับน้ำหนักบวกด้วยค่า Bias โดยการกำหนดค่าให้แก่ Activation Function เพื่อได้ Output ให้ตรงตามลักษณะผลลัพธ์ที่ต้องการ ซึ่งค่าที่ได้จาก Function นั้นคือผลการทำนายของ Perceptron

Activation Function ที่นิยมใช้งานในปัจจุบัน เช่น Sigmoid Function หรือ Logistic Function, ReLU (Rectified Linear Unit) Function, Tanh Function (Hyperbolic Tangent

Function), Softmax Function, Leaky ReLU Function, ELU (Exponential Linear Unit) Function เป็นต้น (Kasidis Satangmongkol, 2022) แต่ละ Activation Function จะมีคุณสมบัติและการทำงานที่แตกต่างกันไปตามลักษณะของข้อมูลและงานที่ต้องการจะทำ ดังนั้นการเลือก Activation Function ที่เหมาะสมกับงานและข้อมูลที่ต้องการ เป็นสิ่งสำคัญที่มีผลต่อประสิทธิภาพของ Neural Network

สำหรับ MLP Regressor ใน Scikit-Learn จะใช้พารามิเตอร์หลัก ๆ ดังนี้

ตาราง 1 พารามิเตอร์สำหรับ MLP Regressor ใน Scikit-Learn

พารามิเตอร์	**คำอธิบาย**
hidden_layer_sizes	จำนวน Hidden Layers และจำนวน Nodes ในแต่ละ Hidden Layers
activation	Activation Function ของชั้น Hidden เช่น Sigmoid, ReLU, Tanh
solver	อัลกอริทึมที่ใช้ในการปรับค่า Weight ของแบบจำลอง ‘lbfgs’,’sgd’,’adam’
max_iter	จำนวนการวนซ้ำสูงสุดในการฝึกแบบจำลอง
random_state	ใช้ในการกำหนด seed สุ่มเพื่อเอาไว้ซ้ำกันในการฝึกแบบจำลอง

ที่มา : (Scikit-learn.org)

โดย MLP จะเหมาะสำหรับการแก้ไขปัญหาที่มีความซับซ้อนและความสำคัญของความสัมพันธ์ของข้อมูลสูง เช่น การทำนายราคาหุ้น แต่ต้องระวังปัญหาการตั้งค่า Parameter ที่มีความซับซ้อน อาจทำให้เกิดปัญหา Overfitting ได้ง่าย

2.4 **งานวิจัยที่เกี่ยวข้อง** (literature Review)

การทบทวนวรรณกรรมของงานวิจัยที่เกี่ยวข้องกับการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) โดยงานวิจัยที่เกี่ยวข้องมีรายละเอียดดังต่อไปนี้

1. **บทความวิจัยเรื่อง** Using Machine Learning to Predict Air Quality Index in New Delhi **โดย** Samayan Bhattacharya **และ** Sk Shahnawaz (Samayan Bhattacharya, 2021)

งานวิจัยนี้ได้ทำนายคุณภาพอากาศ ช่วยให้รัฐบาลและองค์กรที่เกี่ยวข้องสามารถพัฒนากลยุทธ์และนโยบาย เพื่อป้องกันประชาชนจากการสัมผัสกับอากาศที่เป็นพิษ โดยใช้แบบจำลอง Support Vector Regression (SVR) เพื่อคาดการณ์ระดับมลพิษต่างๆ เช่น NO2, SO2, PM2.5 และ PM10 และดัชนีคุณภาพอากาศ (AQI) โดยใช้ข้อมูลมลพิษจากคณะกรรมการควบคุมมลพิษและสถานเอกอัครราชทูตสหรัฐอเมริกาประจำกรุงนิวเดลี ที่เก็บเผยแพร่ต่อสาธารณะ มีการใช้เทคนิค Principal Component Analysis (PCA) ร่วมด้วย มีการใช้ 2 Kernel คือ Radial Basis (RBF) และ Polynomial เปรียบเทียบกัน ในบรรดาวิธีที่ทดสอบสำหรับ PM2.5 ฟังก์ชัน Radial Basis (RBF) เป็น Kernel ที่ดีที่สุด ผลการทดลองTraining set PCA SVR-RBF ค่า R2 ความแม่นยำ 0.881 และแบบไม่ใช้ PCA คือ SVR-RBF ค่า R2 ความแม่นยำที่ 0.936

2. **บทความวิจัยเรื่อง** Modelling and Forecasting Temporal PM2.5 Concentration Using Ensemble Machine Learning Methods**โดย** Obuks Augustine Ejohwomu , Olakekan Shamsideen Oshodi , Majeed Oladokun ,Oyegoke Teslim Bukoye , Nwabueze Emekwuru , Adegboyega Sotunbo and Olumide Adenuga *(Ejohwomu et al., 2022)*

ในงานวิจัยนี้ ผู้วิจัยได้สร้างแบบจำลองการทำนายที่เชื่อถือได้ เป็นเครื่องมือที่มีประโยชน์ในการทำความเข้าใจปัจจัยที่สามารถส่งผลต่อความเข้มของปริมาณ PM2.5 ซึ่งข้อมูลนี้สามารถใช้ในการพัฒนากลยุทธ์และนโยบายการลดปริมาณของมลพิษทางอากาศ ผู้วิจัยใช้ข้อมูลคุณภาพอากาศ ที่รวบรวมในเมืองลากอส 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. ตัวแปรทางมาตรวิทยาหรือข้อมูลที่สามารถวัดหรือวิเคราะห์ได้ตามหลักการทางวิทยาศาสตร์นั้น ไม่สามารถทำนายความเข้มข้นของ PM2.5 ในช่วงสูงสุดและต่ำสุดได้อย่างเพียงพอ ชี้ให้เห็นถึงความจำเป็นในการรวบรวมข้อมูลปัจจัยอื่นๆ เช่น จำนวนรถ ประเภทรถ และแหล่งที่มาอื่นๆ ของมลพิษทางอากาศ 2. เรื่องความก้าวหน้าด้านวิทยาศาสตร์ข้อมูล อาจจะมีเครื่องมือที่สามารถใช้ในการสร้างการทำนาย

ความเข้มข้นของ PM2.5 เกิดขึ้น 3. ข้อจำกัดที่สำคัญที่สุดของการศึกษานี้ คือใช้ตัวแปรทางอากาศวิทยา เป็นตัวแปรเพียงอย่างเดียว ในการพัฒนาแบบจำลอง ตัวแปรอื่น ๆ อย่างเช่น ตัวแปรที่เกี่ยวกับปฏิทินและเวลา ถูกสร้างขึ้นจากองค์ประกอบเวลาเท่านั้น ไม่ได้ถูกนำมาพิจารณาในการสร้างแบบจำลอง แต่แบบจำลองที่พัฒนาขึ้นก็มีประสิทธิภาพดีกว่าโมเดลแบบ Naive โดยค่า MASE มีค่าน้อยกว่า 1 ซึ่งแสดงให้เห็นว่าเชื่อถือได้

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 จากค่าเผยให้เห็นว่า XGBoost_All และ RF_All มีค่าที่แม่นยำกว่าโมเดล ARIMA จากผลการวิเคราะห์ตัวแปร เช่น ความชื้นสัมพัทธ์, อุณหภูมิ, Time-Related และ Lag Features ที่รวมอยู่ในโมเดล XGBoost_All และ RF_All เป็นตัวทำนายที่ดีสำหรับความเข้มข้นของ PM2.5

การรวมโมเดล (Ensemble Methods) เป็นการผสมผสานหลายอัลกอริทึมการเรียนรู้ของเครื่อง เพื่อให้ได้ผลทำนายที่ดีกว่า เมื่อใช้อัลกอริทึมแยกแต่ละตัว โดยการวิจัยที่ผ่านมา การใช้ Ensemble นั้นมีผลการทำนายที่ดีกว่าการใช้อัลกอริทึมแต่ละตัวแยก ในการศึกษานี้ เราได้ใช้โมเดลที่ดีสุด 3 อันดับ ได้แก่ (XGBoost_All, RF_All, และ ARIMA) มาผสมกันเพื่อสร้าง Ensemble Model โดยจะมี 3 แบบ ได้แก่ 1 ค่าเฉลี่ย (average), 2 มัธยฐาน (median), และ 3 ค่าน้ำหนัก (weighted) โดยกำหนดน้ำหนัก ตามประสิทธิภาพของแต่ละแบบจำลอง XGBoost_All ดีสุด ให้น้ำหนัก = 3, RF_All น้ำหนัก = 2 และ ARIMA น้ำหนัก = 1 โดยกำหนดน้ำหนักตามผลการทำนายของแต่ละโมเดล เห็นได้ว่าตัวชี้วัดผลการทำนาย (MAE, MASE และ RMSE) Ensemble Model แบบ Ensemble (Weighted) มีค่าต่ำที่สุด 1.57, 0.71, 2.1876 ตามลำดับ

3.บทความวิจัยเรื่อง Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models **โดย** Doreswamy, Harishkumar K S, Yogesh KM, Ibrahim Gad (Doreswamy et al., 2020)

ในงานวิจัยนี้ใช้ชุดข้อมูลจากเครือข่ายการตรวจสอบคุณภาพอากาศของไต้หวัน (TAQMN) ปี 2012 ถึง 2017 มีสถานีทางอากาศ 76 แห่ง โดยใช้แบบจำลองการเรียนรู้ของเครื่องได้แก่ Linear Regression , Lasso, Ridge, Random Forest Regressor, Gradient Boosting Regressor, K Neighbors ,MLP, Decision Tree ตัวชี้วัดที่ใช้ ได้แก่ MAE, MSE, RMSE และ R2 ผลลัพธ์ เมื่อเปรียบเทียบกัน แบบจำลอง Gradient Boosting Regressor ได้ผลลัพธ์ดีที่สุดสำหรับ

การทำนายมลพิษทางอากาศข้อมูลจากTAQMN ค่า R2 ของ Train อยู่ที่ 0.9983 Test อยู่ที่ 0.8891

4.บทความวิจัยเรื่อง Feature extraction and prediction of fine particulate matter (PM2.5)chemical constituents using four machine learning models **โดย** Young Su Lee , Eunhwa Choi , Minjae Park , Hyeri Jo , Manho Park , Eunjung Nam , Dai Gon Kim, Seung-Muk Yi, Jae Young Kim (Lee et al., 2023)

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

5. บทความวิจัยเรื่อง An improved deep learning model for predicting daily PM 2 . 5 concentration **โ ด ย** Fei Xiao, MeiYang, Hong Fan, Guanghui Fan **แ ล ะ** MohammedA.A.Alqaness (Xiao et al., 2020)

ในวิจัยนี้ผู้วิจัยได้กล่าวว่า ในช่วงหลายทศวรรษที่ผ่านมามลพิษทางอากาศได้ทำให้สุขภาพของประชาชนได้รับความเสียหายอย่างมีนัยสำคัญ ดังนั้น การทำนาย PM2.5 อย่างแม่นยำเป็นงานที่สำคัญ โดยได้ชุดข้อมูลจากเมืองปักกิ่ง เทียนจิน และ 11 เมืองของมณฑลเหอเป่ย(BTH) ของประเทศจีน เก็บข้อมูลการช่วงวันที่ 1 มกราคม 2015- 31 ธันวาคม 2017 ข้อมูลถูกแบ่งออกเป็น 3 หมวดใหญ่ ได้แก่ Ground measured , Near-real time analysis, Satellite

products ผู้วิจัย มีการทำ ONE HOT และรวมตารางข้อมูลอุณหภูมิที่รวบรวมจากชุดข้อมูล MOD11A1 และ ECMWF ถูกรวมเข้าด้วยกัน ผู้วิจัยได้กล่าวถึงเรื่องการกระจายตัวของสถานีตรวจวัดมลพิษทางอากาศที่ไม่ได้เท่ากัน ทำให้ความสัมพันธ์พื้นที่และเวลา ระหว่างสถานีในศูนย์กลางกับสถานีรอบๆ มีความแตกต่างกัน ซึ่งแก้ปัญหานี้ โดยศึกษานำเสนอวิธีการแบบจำลองที่ชื่อว่า weighted long short-term memory neural network extended (WLSTME) ซึ่งปรับปรุงปัญหาเรื่องว่าต้องพิจารณาผลของความหนาแน่นของสถานีและเงื่อนไขของลมต่อความสัมพันธ์พื้นที่และเวลา ในขั้นแรกจำนวนของสถานีรอบๆ ที่ใกล้ที่สุด ถูกเลือกเป็นสถานีเพื่อนบ้านกับสถานีในศูนย์กลาง และระยะทาง, ความเข้มของมลพิษทางอากาศและเงื่อนไขของลม ถูกนำเข้าไปในแบบจำลอง Multilayer Perception (MLP) เพื่อสร้างข้อมูล time series ของ PM2.5 ในขั้นตอนถัดไป นำความเข้มของ PM2.5 ในอดีตของสถานีในศูนย์กลางและข้อมูล PM2.5 ที่มีน้ำหนักของสถานีเพื่อนบ้าน นำเข้าไปใน long short-term memory (LSTM) เพื่อที่จะทำประสิทธิภาพสัมพันธ์ทางพื้นที่และเวลาพร้อมๆ กันและดึงข้อมูลทางพื้นที่และเวลาออก และใช้ multilayer perception (MLP) อีกตัวเพื่อรวมข้อมูลทางพื้นที่และเวลาที่ถูกดึงออกมากับข้อมูลอุตสาหกรรมของสถานีในศูนย์กลางเพื่อสร้างการทำนายค่าความเข้มของ PM2.5 ในอนาคตของสถานีในศูนย์กลาง

ผลทดลองที่เปรียบเทียบกับวิธีการที่มีอยู่แล้วสามวิธี แสดงให้เห็นว่าแบบจำลอง WLSTME ที่นำเสนอมีค่า RMSE (40.67) และ MAE (26.10) ที่ต่ำที่สุดและค่า p (0.59) ที่สูงที่สุด มีทดลองเพิ่มเติมแสดงให้เห็นว่าในทุกฤดูกาลและภูมิภาค แบบจำลอง WLSTME ทำงานได้ดีที่สุด ผลลัพธ์นี้ยืนยันว่า WLSTME สามารถปรับปรุงความแม่นยำในการทำนาย PM2.5 อย่างมีนัยสำคัญ

บทที่ 3
วิธีดำเนินการวิจัย

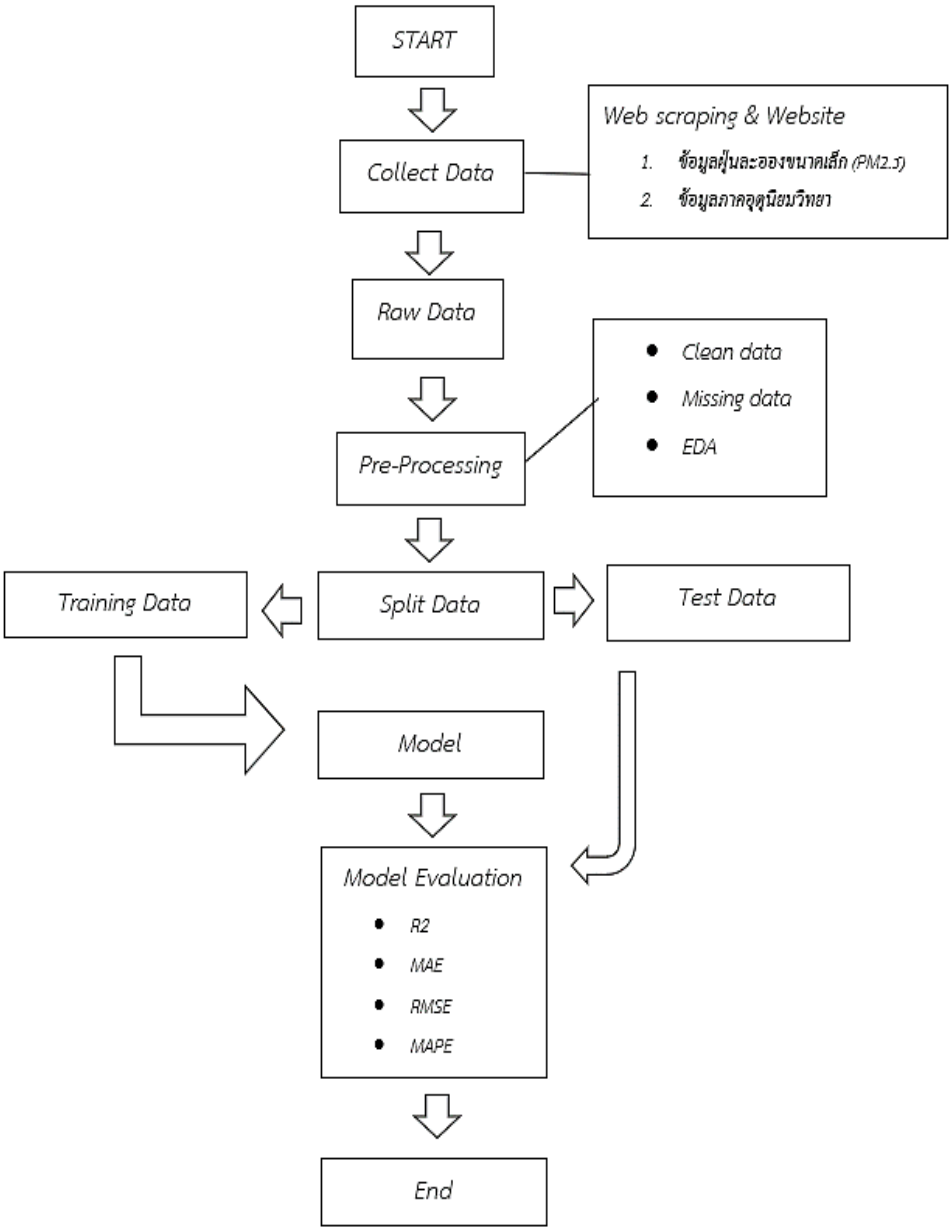
การวิจัยนี้เป็นการสร้างแบบจำลองเพื่อทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) โดยอาศัยค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง และข้อมูลสภาพอากาศมาเป็น ข้อมูลคุณลักษณะที่ใช้ในการทำนาย ซึ่งกระบวนการดำเนินงานมีดังนี้

1. การสร้างชุดข้อมูลค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 และข้อมูลสภาพอากาศประชากร

2. การจัดการข้อมูลและการสำรวจและวิเคราะห์ข้อมูลเบื้องต้น

3. การสร้างแบบจำลองการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 ด้วยเทคนิคการเรียนรู้ของเครื่อง

โดยมีแผนผังกระบวนการสร้างแบบจำลองเพื่อทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 แสดงดังภาพประกอบที่ 2 โดยเริ่มตั้งแต่การเก็บรวบรวมข้อมูล Collect Data จากแหล่งข้อมูลสาธารณะ จากนั้นนำข้อมูล Raw Data ที่ได้มาเตรียมให้อยู่ในรูปแบบที่เหมาะสำหรับการสร้างแบบจำลอง Pre-Processing เช่น การจัดรูปแบบข้อมูล Data Formatting การทำความสะอาดข้อมูล Clean Data การจัดการข้อมูลสูญหาย Missing Data การสำรวจข้อมูล EDA จากนั้นทำการแบ่งชุดข้อมูล Split Data ที่ทำความสะอาดแล้ว โดย Training Data แบ่งสำหรับให้แบบจำลองเรียนรู้ และ Test Data สำหรับในการประเมินผลแบบจำลอง จากนั้นนำ Training Data ที่เตรียม เข้าสู่แบบจำลอง Model ทั้ง 4 แบบ ได้แก่ LR (Linear Regression), SVR (Support Vector Regression) , XGBoost (eXtreme Gradient Boosting) และ MLP (Multi-Layer Perceptron) ขั้นสุดท้าย เป็นการประเมินผลของแบบจำลอง Model Evaluation โดยนำ Test Data มาใช้ในการเปรียบเทียบค่าการทำนาย ซึ่งใช้เครื่องมือวัดผลประสิทธิภาพ ได้แก่ R2 Score, MAE, RMSE, และ MAPE เพื่อประเมินความแม่นยำของแบบจำลอง

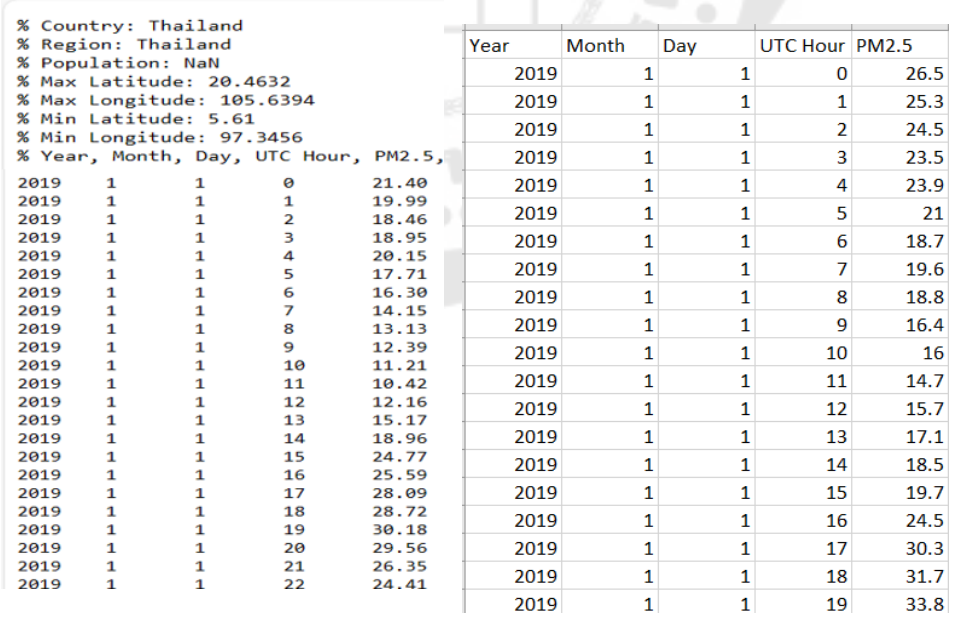


ภาพประกอบ 2 แผนผังกระบวนการสร้างแบบจำลองเพื่อทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5

3.1 การสร้างชุดข้อมูลค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 และข้อมูลสภาพอากาศ

ในการดำเนินงานวิจัยนี้ มีการสร้างชุดข้อมูล 2 ชุดมารวมกัน โดยนำเข้าข้อมูลจากแหล่งข้อมูลสาธารณะแบบเปิด ผ่านหน้าเว็บไซต์และวิธีการ Web Scraping ซึ่งเป็นกระบวนการในการดึงข้อมูล โดยใช้สคริปต์ดึงข้อมูลจากหน้าเว็บไซต์ รายละเอียดดังนี้

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 3 ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5)จากเว็บไซต์ Berkeley Earth

2. ข้อมูลภาคอุตุนิยมวิทยา ผ่านวิธีการ Web Scraping โดยใช้สคริปต์ดึงข้อมูลจากหน้าเว็บไซต์ Weather Underground โดย สคริปต์นำมาจาก ผู้ใช้ GitHub.com ชื่อว่า Karlheinzniebuhr/the-weather-scraper (Karlheinzniebuhr, 2022) 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 ดังภาพประกอบ

```
!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 4 โปรแกรม "unrar" เพื่อแตกไฟล์จากไฟล์ที่มีนามสกุล .rar

จากนั้น ทำการตั้งค่า ในไฟล์ **config.py** กำหนดระยะเวลาที่จะดึงข้อมูล Start Date – End Date ดังภาพประกอบ 5

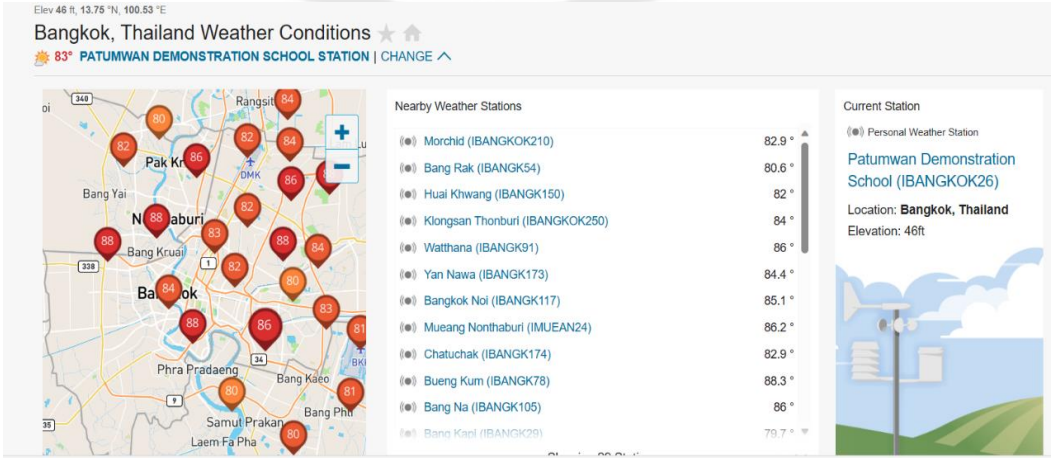
```
from datetime import date

# 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 5 กำหนดระยะเวลาที่จะดึงข้อมูลStart Date – End Date

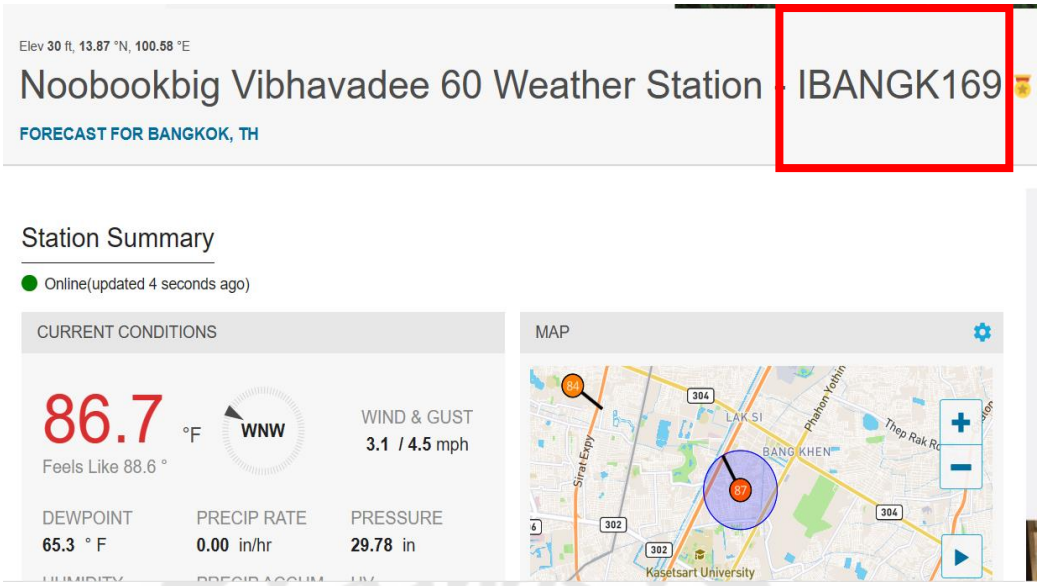
ที่มา (Karlheinzniebuhr, 2022)

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 6 บนเว็บไซต์ Weather Underground



ภาพประกอบ 7 Station ID สถานีตรวจวัด

เมื่อกดเข้าไปยังสถานีตรวจวัดที่เลือก จะปรากฏ Station ID ที่ด้านขวาดังภาพประกอบ 7 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 8 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 9

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 9 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 รวมคอลัมน์วันและเวลานำมาเก็บเป็นไฟล์ CSV

คุณลักษณะและตัวแปรในชุดข้อมูลดิบ Raw Data ที่ได้จากหน้าเว็บไซต์และวิธีการ Web Scraping

1. ตัวแปรอิสระของข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) ทั้งหมด 5 คอลัมน์ ดังนี้

1.1 ข้อมูลเวลา

1.1.1 Year (ปี)

1.1.2 Month (เดือน)

1.1.3 Day (วัน)

1.1.4 UTC Hour (ชั่วโมงแบบ UTC)

1.2 ข้อมูลอุตุนิยมวิทยาและสภาพอากาศ

1.2.1 PM2.5 (ค่าฝุ่นละอองขนาดเล็ก PM2.5)

2. ตัวแปรอิสระของข้อมูลภาคอุตุนิยมวิทยา ทั้งหมด 13 คอลัมน์ ดังนี้

2.1 ข้อมูลเวลา

2.1.1 Date (วัน, เดือน, ปี)

2.1.2 Time (เวลา)

2.2 ข้อมูลอุตุนิยมวิทยาและสภาพอากาศ

2.2.1 Temperature_C (อุณหภูมิ)

2.2.2 Dew_Point_C (จุดน้ำค้าง)

2.2.3 Humidity_% (ความชื้น)

2.2.4 Wind (ทิศทางลม)

2.2.5 Speed_kmh (ความเร็วลม)

2.2.6 Gust_kmh (ลมกระโชก)

2.2.7 Pressure_hPa (ความกดอากาศ)

2.2.8 Precip_Rate_mm (หยาดน้ำฟ้า)

2.2.9 Precip_Accum_mm (หยาดน้ำฟ้าสะสม)

2.2.10 UV (รังสีอัลตราไวโอเลต)

2.2.11 Solar_w/m2 (รังสีดวงอาทิตย์)

3.2 การจัดการข้อมูล สำรวจและวิเคราะห์ข้อมูลเบื้องต้น

ในขั้นตอนนี้เป็นการเตรียมข้อมูลที่ได้มาให้อยู่ในรูปแบบที่เหมาะสม สำหรับการนำเข้าสู่การสร้างแบบจำลอง Pre-Processing ดังนี้

1.การจัดรูปแบบข้อมูล Data Formatting ตรวจสอบการจัดเรียงข้อมูลเรียงลำดับตามเวลา รูปแบบของข้อมูลถูกต้อง การแปลงข้อมูลที่เป็นตัวอักษรเป็นตัวเลข

2.การทำความสะอาดข้อมูล Clean Data การจัดการข้อมูลซ้ำ

3.การจัดการข้อมูลสูญหาย Missing Data ตรวจสอบความครบถ้วน ความถูกต้อง ไม่ขาดหาย เช่น ค่าฝุ่นละอองขนาดเล็ก PM2.5 ไม่ควรเป็นค่าว่าง หรือค่าที่ติดลบ

4.การทำวิศวกรรมคุณลักษณะข้อมูล Feature Engineering สร้างตัวแปรใหม่โดยใช้ข้อมูลที่มีอยู่ การทำ Timestamp ,การแปลงตัวแปรข้อมูลให้มีการกระจายปกติ

3.2.1 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 คอลัมน์ดังนี้ Season_Rainy, Season_Summer, Season_Winter ดังภาพประกอบ 10

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 10 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

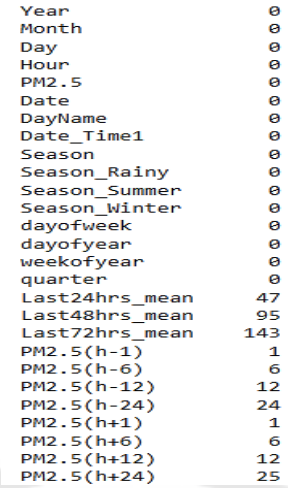
5. สร้างคอลัมน์ใหม่ จากข้อมูลคอลัมน์ “PM 2.5” โดยการ shift (1),(6),(12),(24) ตามลำดับ เพื่อเลื่อนข้อมูลมาใช้ในชั่วโมงย้อนหลัง ซึ่งใช้เป็นค่า PM2.5 ของ 1,6,12,24 ชั่วโมงย้อนหลัง ได้คอลัมน์ ดังนี้ PM2.5(h-1), PM2.5(h-6), PM2.5(h-12), PM2.5(h-24)

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 ที่ไม่ใช่ค่าว่าง มาเติม

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 11



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

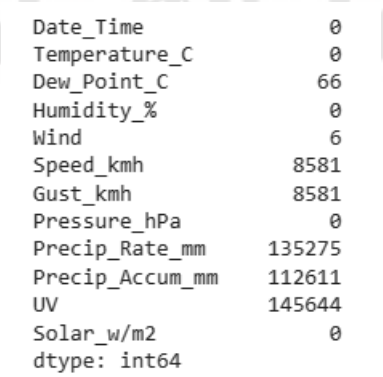
ภาพประกอบ 11แสดงค่าว่างในข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5)

8. ใช้. duplicated() ตรวจสอบข้อมูลซ้ำ ไม่พบข้อมูลซ้ำ

3.2.2 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 ดังภาพประกอบ 12



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 12 แสดงค่าว่างในข้อมูลภาคอุตุนิยมวิทยา

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

อยู่ในแถวก่อนหน้า 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

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
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

ภาพประกอบ 13 การ One-Hot Encoding ในคอลัมน์ “ Wind”

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 เพื่อให้ได้ค่าตัวกลางที่อาจ

ไม่มีอยู่ในข้อมูลต้นฉบับ 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 14 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**

Date_Time	Year	Month	Day	Hour	PM2.5	Season_Rain	Season_Summe	Season_Winter
1/1/2019 7:00	2019	1	1	7	26.5	0	0	1
1/1/2019 8:00	2019	1	1	8	25.3	0	0	1
1/1/2019 9:00	2019	1	1	9	24.5	0	0	1
1/1/2019 10:00	2019	1	1	10	23.5	0	0	1
1/1/2019 11:00	2019	1	1	11	23.9	0	0	1

Date_Time	Last24hrs_	Last48hrs	Last72hrs	PM2.5(h-1)	PM2.5(h-6)	PM2.5(h-12)
1/1/2019 7:00	23.7125	23.5854	24.8917	26.5	26.5	26.5
1/1/2019 8:00	23.7125	23.5854	24.8917	26.5	26.5	26.5
1/1/2019 9:00	23.7125	23.5854	24.8917	25.3	26.5	26.5
1/1/2019 10:00	23.7125	23.5854	24.8917	24.5	26.5	26.5
1/1/2019 11:00	23.7125	23.5854	24.8917	23.5	26.5	26.5

ภาพประกอบ 15 Merge Data Frame โดยใช้ key เป็นตัวเชื่อมที่ คอลัมน์ 'Date_Time'

3.2.3 คุณลักษณะและตัวแปรชุดข้อมูลในการวิจัย

คุณลักษณะและตัวแปรในชุดข้อมูล หลังผ่านการรวม Data Frame และทำความสะอาดข้อมูลเรียบร้อยแล้ว ประกอบด้วยข้อมูลดังตารางที่ 2

ตาราง 2 ข้อมูลคุณลักษณะของชุดข้อมูลในการดำเนินงานวิจัย

ตัวแปร	คำอธิบาย
Date_Time	วันที่ เดือน ปีคริสตศักราช และ เวลาที่มีระยะห่างกัน 1 ชั่วโมง
Year	ปีคริสตศักราช
Month	เดือน
Day	วันที่
Hour	เวลาเป็นชั่วโมง
Season_Rainy	ฤดูฝน
Season_Summer	ฤดูร้อน
Season_Winter	ฤดูหนาว
Last24hrs_mean (μg/m3)	ค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 24 ชั่วโมงย้อนหลัง
Last48hrs_mean (μg/m3)	ค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 48 ชั่วโมงย้อนหลัง
Last72hrs_mean (μg/m3)	ค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 72 ชั่วโมงย้อนหลัง
PM2.5 (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5
PM2.5(h-1) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 1 ชั่วโมงย้อนหลัง
PM2.5(h-6) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 6 ชั่วโมงย้อนหลัง
PM2.5(h-12) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 12 ชั่วโมงย้อนหลัง
PM2.5(h-24) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 24 ชั่วโมงย้อนหลัง
Temperature_C	อุณหภูมิ
Dew_Point_C	จุดน้ำค้าง
Humidity_%	ความชื้น
Speed_kmh (km/h)	ความเร็วลม
Gust_kmh	ลมกระโชก

ตาราง 2 (ต่อ)

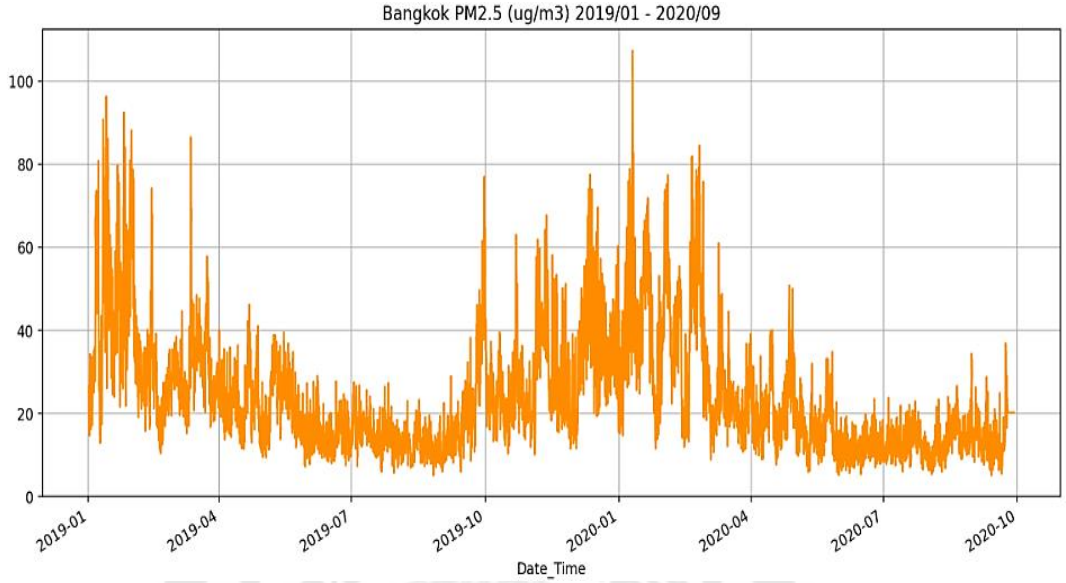
ตัวแปร	คำอธิบาย
Pressure_hPa (hPa)	ความกดอากาศ
Wind_North	ลมทิศทางเหนือ North 0.00°
Wind_NNE	ลมทิศทางเหนือ-ตะวันออกเฉียงเหนือ North-Northeast 22.50°
Wind_NE	ลมทิศทางตะวันออกเฉียงเหนือ Northeast 45.00°
Wind_ENE	ลมทิศทางตะวันออก-ตะวันออกเฉียงเหนือ East-Northeast 67.50°
Wind_East	ลมทิศทางตะวันออก East 90.00°
Wind_ESE	ลมทิศทางตะวันออก-ตะวันออกเฉียงใต้ East-Southeast 112.50°
Wind_SE	ลมทิศทางตะวันออกเฉียงใต้ Southeast 135.00°
Wind_SSE	ลมทิศทางใต้-ตะวันออกเฉียงใต้ South-Southeast 157.50°
Wind_South	ลมทิศทางใต้ South 180.00°
Wind_SSW	ลมทิศทางใต้-ตะวันตกเฉียงใต้ South-Southwest 202.50°
Wind_SW	ลมทิศทางตะวันตกเฉียงใต้ Southwest 225.00°
Wind_WSW	ลมทิศทางตะวันตก- ตะวันตกเฉียงใต้ West-Southwest 247.50°
Wind_West	ลมทิศทางตะวันตก West 270.00°
Wind_WNW	ลมทิศทางตะวันตก-ตะวันตกเฉียงเหนือ West-Northwest 292.50°
Wind_NW	ลมทิศทางตะวันตกเฉียงเหนือ Northwest 315.00°
Wind_NNW	ลมทิศทางเหนือ-ตะวันตกเฉียงเหนือ North-Northwest 337.50°

ผลลัพธ์ของตัวแปรตาม หรือ Target ได้แก่ ค่าฝุ่นละอองขนาดเล็ก PM2.5 ชั่วโมงล่วงหน้า ดังตารางที่ 3

ตาราง 3 ข้อมูลผลลัพธ์ตัวแปรตาม หรือ Target ของชุดข้อมูลในการดำเนินงานวิจัย

ตัวแปร	คำอธิบาย
PM2.5 (h+1) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 1 ชั่วโมงล่วงหน้า
PM2.5 (h+6) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 6 ชั่วโมงล่วงหน้า
PM2.5(h+12) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 12 ชั่วโมงล่วงหน้า
PM2.5(h+24) (μg/m3)	ค่าฝุ่นละอองขนาดเล็ก PM2.5 24 ชั่วโมงล่วงหน้า

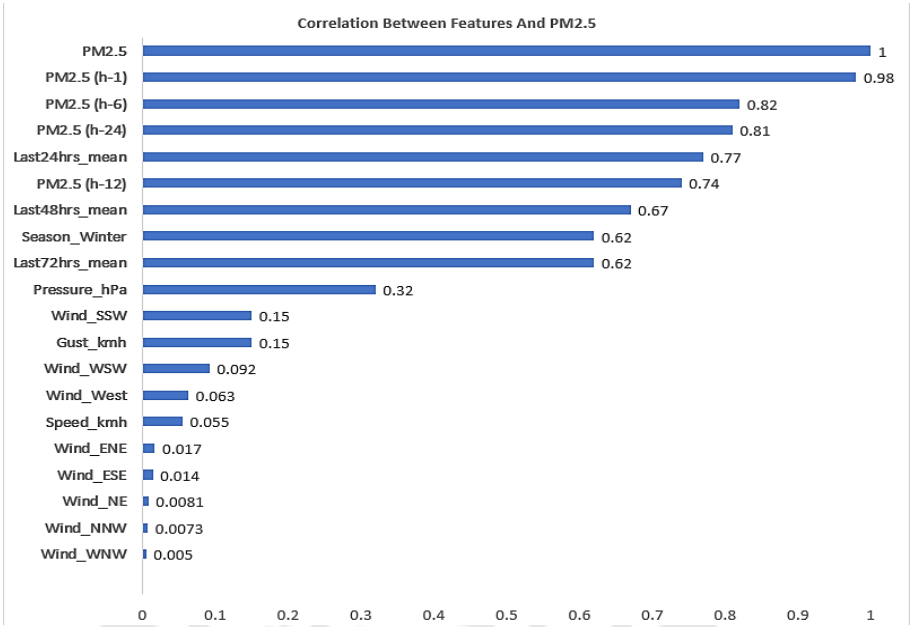
3.2.4 สำรวจและวิเคราะห์ข้อมูลเบื้องต้น



ภาพประกอบ 16 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 เพิ่มสูงขึ้นอีกครั้ง เมื่อเข้าสู่ปลายปีจนถึงต้นปี

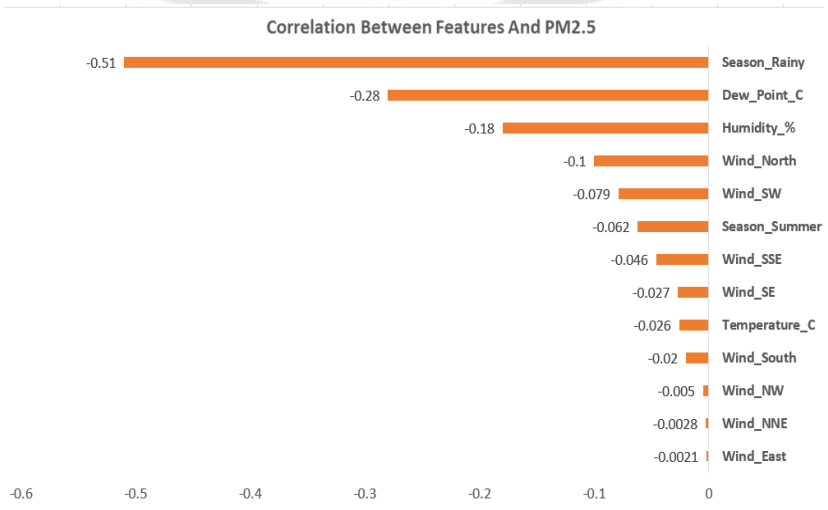
การสำรวจความสัมพันธ์ ระหว่างข้อมูลฝุ่นละอองขนาดเล็ก PM2.5 กับ คุณลักษณะอื่นประกอบด้วย Season_Rainy, Season_Summer, Season_Winter, Last24hrs_mean, Last48hrs_mean, Last72hrs_mean, PM2.5(h-1), PM2.5(h-6), PM2.5(h-12), PM2.5(h-24), Temperature_C (อุณหภูมิ), Dew_Point_C (จุดน้ำค้าง), Humidity_% (ความชื้น), 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 ซึ่งมีทั้งความสัมพันธ์ในทางบวกและทางลบ ดังภาพประกอบ 17-18



ภาพประกอบ 17แสดงความสัมพันธ์ระหว่างชุดข้อมูลฝุ่น 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

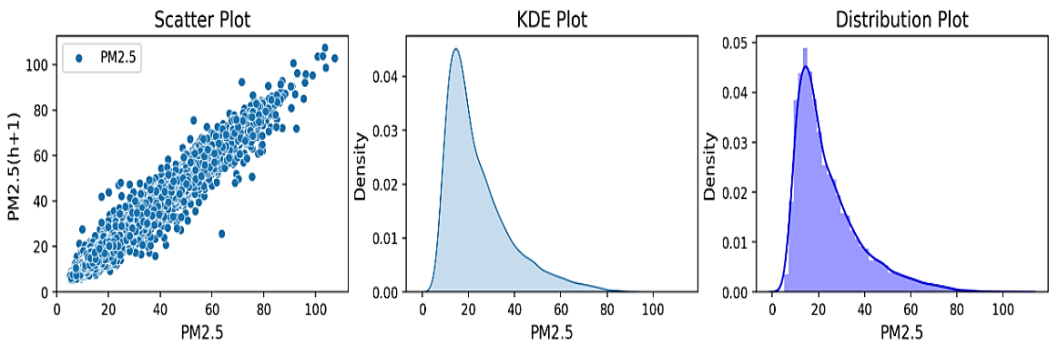


ภาพประกอบ 18 แสดงความสัมพันธ์ระหว่างชุดข้อมูลฝุ่น PM2.5 กับ คุณลักษณะอื่นในทางลบ

เรียงลำดับความสัมพันธ์ทางลบดังนี้

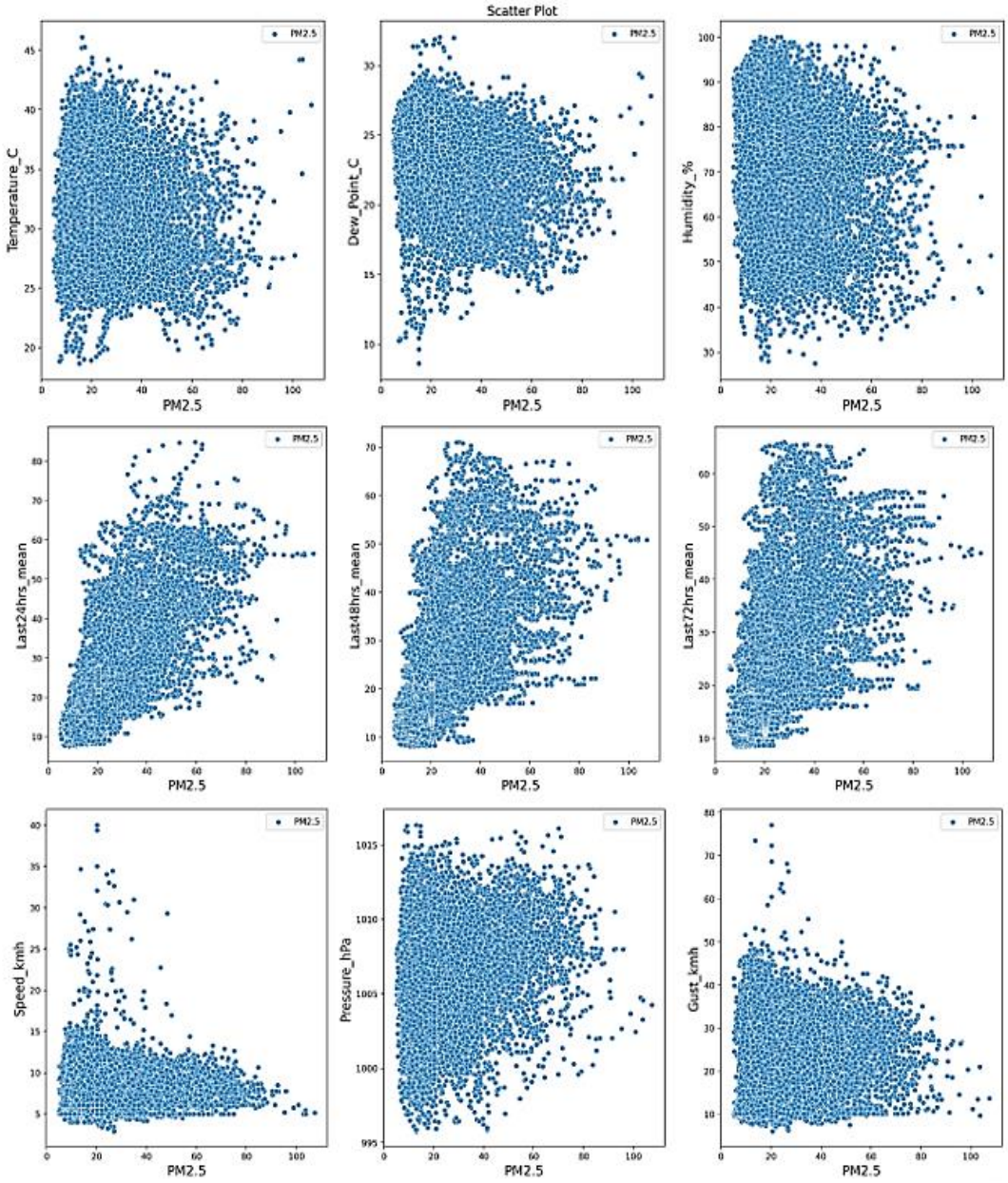
Season_Rainy > Dew_Point_C (จุดน้ำค้าง) > Humidity_% (ความชื้น) > ทิศทางลม Wind_North >ทิศทางลม Wind_SW > Season_Summer > ทิศทางลม Wind_ SSE > ทิศทางลม Wind_SE > ทิศทางลม Wind_South >> ทิศทางลม Wind_NW > ทิศทางลม Wind_NNE > ทิศทางลม Wind_East

สำรวจดูค่าความหนาแน่น Density ของปริมาณข้อมูลฝุ่นละอองขนาดเล็ก PM2.5 มีลักษณะ โค้งเบ้ขวา **(right-skewed)**



ภาพประกอบ 19 ความหนาแน่น Density ของข้อมูลฝุ่นละอองขนาดเล็ก PM2.5

จากการสำรวจการกระจายตัวของชุดข้อมูล พบว่าส่วนใหญ่มีการกระจายตัวที่ไม่เท่ากัน ดังนั้นก่อนนำข้อมูลเข้าสู่แบบจำลองจะต้องมีการปรับระดับของข้อมูลให้อยู่ในระดับเดียวกัน Scale ในวิจัยนี้ได้ใช้ การ Scaling ข้อมูล ใน 2 รูปแบบเปรียบเทียบกัน คือ Standard Scaling และ Min-Max Scaling เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมสำหรับการสร้างและฝึกสอนแบบจำลอง

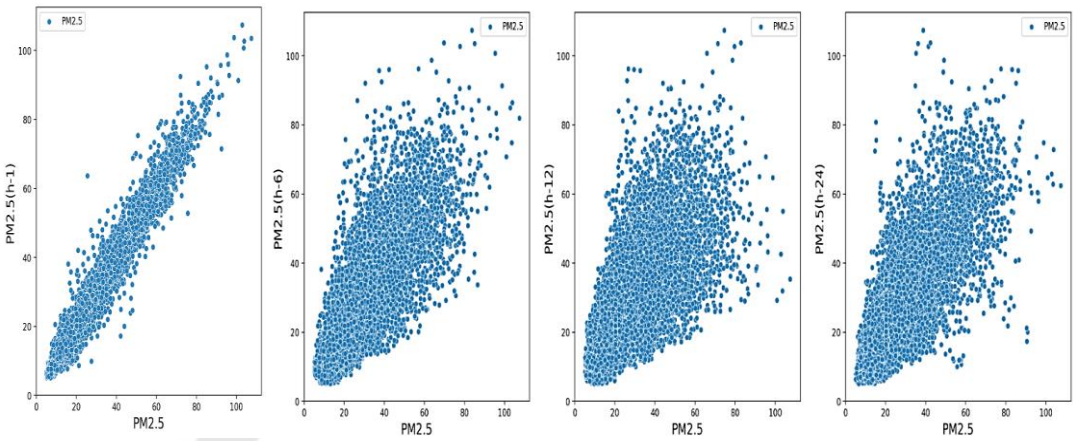


ภาพประกอบ 20 ความสัมพันธ์ระหว่างค่าฝุ่นละออง PM2.5 กับข้อมูลทางอุตุนิยมวิทยา

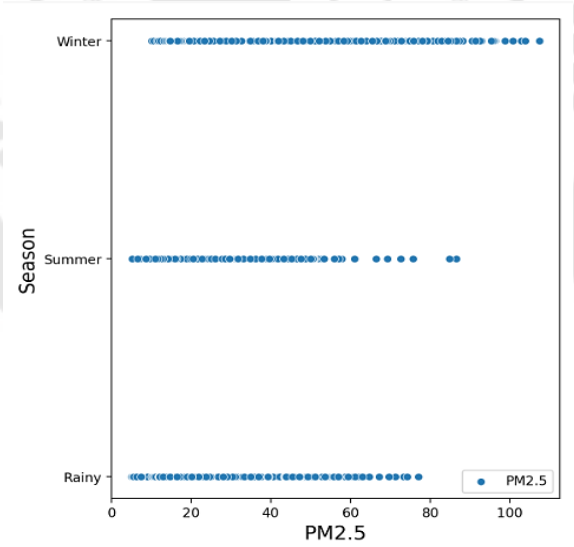
จากภาพประกอบ 20 **แถวแรก** แสดงความสัมพันธ์ระหว่างฝุ่นละออง PM2.5 กับ Temperature_C (อุณหภูมิ) ,Dew_Point_C (จุดน้ำค้าง) , Humidity_% (ความชื้น) ตามลำดับ

แถวสอง แสดงความสัมพันธ์ระหว่างฝุ่นละออง PM2.5 กับ Last24hrs_mean ,Last48hrs_mean ,Last72hrs_mean ตามลำดับ

แถวสาม แสดงความสัมพันธ์ระหว่างฝุ่นละออง PM2.5 กับ Speed_kmh (ความเร็วลม) , Pressure_hPa (ความกดอากาศ), Gust_kmh (ลมกระโชก) ตามลำดับ



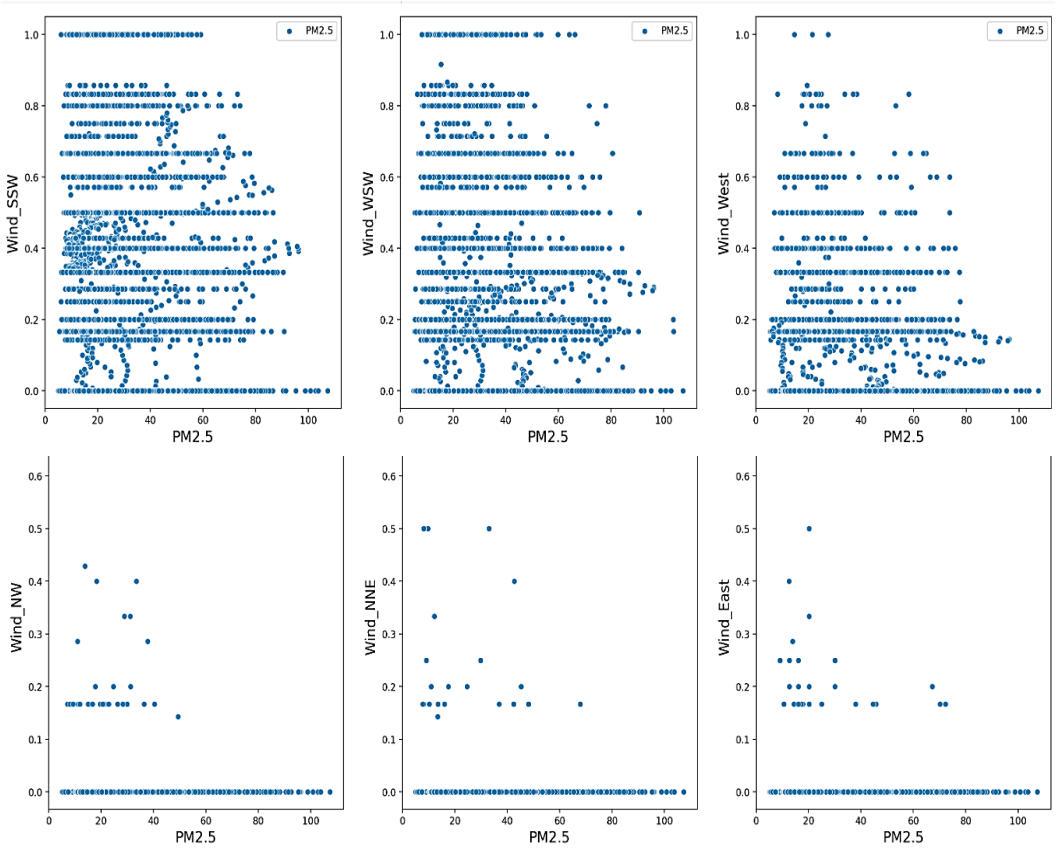
ภาพประกอบ 21ความสัมพันธ์ระหว่างค่าฝุ่นละออง PM2.5 กับข้อมูลฝุ่นละออง PM2.5 ใน 1 6 12 24 ชั่วโมงย้อนหลัง ตามลำดับ



ภาพประกอบ 22 แสดงความสัมพันธ์ระหว่างฝุ่นละออง PM2.5 กับ ข้อมูลฤดูกาล

จากภาพประกอบ 22 แสดงความสัมพันธ์ระหว่างฝุ่นละออง PM2.5 กับ ข้อมูลฤดูกาล จะเห็นได้ว่า ในช่วงฤดูหนาว ค่าของฝุ่นละอองขนาดเล็ก PM2.5 มีความเข้มข้นที่ค่อนข้างสูงกว่า ฤดูร้อนและฤดูฝน ซึ่งสอดคล้องกับข้อมูลสภาพอุตุนิยมวิทยาที่ว่า ในฤดูมรสุมตะวันออกเฉียงเหนือ (กลางเดือนตุลาคมถึงกลางเดือนกุมภาพันธ์) ซึ่งตรงกับช่วงฤดูหนาว ในช่วงฤดูมรสุมนี้ ฝุ่นละออง

มีระดับสูงเนื่องจากสภาพอากาศแห้ง ทำให้ค่าของฝุ่นละอองขนาดเล็ก PM2.5 มีความเข้มข้นที่ค่อนข้างสูง



ภาพประกอบ 23 แสดงความสัมพันธ์ระหว่างฝุ่นละออง PM2.5 กับข้อมูลทางอุตุนิยมวิทยา (ทิศทางลม)

จากภาพประกอบ 15 ความสัมพันธ์ระหว่างฝุ่นละออง PM2.5 กับข้อมูลทางอุตุนิยมวิทยา ทิศทางลม จากภาพแสดงความสัมพันธ์ของทิศทางของลมที่มีต่อค่าฝุ่นละอองขนาดเล็ก PM2.5 ทั้งในทางบวกและทางลบ 3 อันดับแรก ดังนี้

ความสัมพันธ์ในทิศทางบวก

ได้แก่ ทิศทางลม Wind_SSW > ทิศทางลม Wind_WSW > ทิศทางลม Wind_West

ความสัมพันธ์ในทิศทางลบ

ทิศทางลม Wind_NW > ทิศทางลม Wind_NNE > ทิศทางลม Wind_East

ซึ่งสอดคล้องกับข้อมูลของกรมควบคุมมลพิษ,2561 เกี่ยวกับสภาพอุตุนิยมวิทยาและสภาพแวดล้อมซึ่งส่งผลต่อการแพร่กระจายของมลพิษ ทิศทางของลม ได้แก่ ฤดูมรสุมตะวันออกเฉียงเหนือ และฤดูมรสุมตะวันตกเฉียงใต้

3.3 การสร้างแบบจำลอง

หลังจากเก็บรวบรวมข้อมูล ทำความสะอาดข้อมูล จัดการกับข้อมูลที่สูญหายและสำรวจข้อมูล เพื่อทำความเข้าใจในรูปแบบและความสัมพันธ์ของข้อมูล รวมถึงมีการสร้างตัวแปรใหม่ เช่น การสร้างค่าเฉลี่ยของ PM2.5 ในช่วง 24 ชั่วโมงย้อนหลัง หรือ ค่า PM2.5 ในชั่วโมงย้อนหลัง หลังจากรวบรวมข้อมูลเรียบร้อย ทำการแบ่งชุดข้อมูลที่ทำความสะอาดแล้ว โดยแบ่ง Training Data สำหรับให้แบบจำลองเรียนรู้ และ Test Data สำหรับในการประเมินผลแบบจำลอง จากนั้นนำ Training Data ที่เตรียมไว้ เข้าสู่แบบจำลอง ซึ่งในการวิจัยนี้ 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 โดยแต่ละแบบนั้น ครอบคลุมการทำนายค่า 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%

หลังจากแบ่งชุดข้อมูลเรียบร้อย ถัดมาทำการ Scaling ข้อมูล ใน 2 รูปแบบคือ Standard Scaling และ Min-Max Scaling และนำเข้าสู่แบบจำลองทั้ง 4 แบบ ได้แก่ LR (Linear Regression), SVR (Support Vector Regression) , XGBoost (eXtreme Gradient Boosting) และ MLP (Multi-Layer Perceptron)

ในงานวิจัยนี้ ได้ทดลองสร้างแบบจำลองเบื้องต้น โดยใช้ ค่าพารามิเตอร์เริ่มต้น จาก scikit-learn ดังนี้ (scikit-learn)

ตาราง 4 พารามิเตอร์ LR (Linear Regression) ใน scikit-learn

พารามิเตอร์	**ชนิด**	**ค่าเริ่มต้น**
copy_X	bool	default=True
n_jobs	int	default=None
positive	bool	default=None

ตาราง 5 พารามิเตอร์ SVR (Support Vector Regression) ใน scikit-learn

พารามิเตอร์	**ชนิด**	**ค่าเริ่มต้น**
kernel{‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’} or callable		default=’rbf’
degree	int	default=3
positive	bool	default=None
gamma{‘scale’, ‘auto’}	float,	default=’scale’
coef0	float	default=0.0
tol	float	default=1e-3
c	float	default=1.0
Epsilon	float	default=0.1
Shrinking	bool	default=True
cache_size	float	default=200
verbose	bool	default=False
max_iter	int	default=-1

ตาราง 6 พารามิเตอร์XGBoost (eXtreme Gradient Boosting) ใน xgboost

พารามิเตอร์	ชนิด	ค่าเริ่มต้น
learning_rate	float	default=0.3
n_estimators	int	default=100
max_depth	int	default=6
min_child_weight	int	default=1
subsample	float	default=1
colsample_bytree	float	default=1
gamma	float	default=0
reg_alpha	float	default=0
reg_lambda	float	default=1
scale_pos_weight	float	default=1
objective	string	default='reg:squarederror' (สำหรับ regression)
booster	string	default='gbtree'
random_state	int	default=None
n_jobs	int	default=1

ตาราง 7 พารามิเตอร์ MLP (Multi-Layer Perceptron)ใน scikit-learn

พารามิเตอร์	ชนิด	ค่าเริ่มต้น
hidden_layer_sizesarray-like of shape(n_layers - 2,)		default=(100,)
activation{‘identity’, ‘logistic’, ‘tanh’, ‘relu’},		default=’relu’
solver{‘lbfgs’, ‘sgd’, ‘adam’},		default=’adam’
alpha	float	default=0.0001
batch_size	int	default=auto

ตาราง 7(ต่อ)

พารามิเตอร์	ชนิด	ค่าเริ่มต้น
learning_rate{‘constant’, ‘invscaling’, ‘adaptive’}		default=’constant’
learning_rate_init	float	default=0.001
power_t	float	default=0.5
max_iter	int	default=200
shuffle	bool	default=True
random_state	int	RandomState instance, default=None
tol	float	default=1e-4
verbose	bool	default=None
warm_start	bool	default=False
momentum	float	default=False
nesterovs_momentum	bool	default=0.9
early_stopping	bool	default=True
validation_fraction	float	default= 0.1
Beta_1	float	default= 0.9
Beta_2	float	default= 0.999
epsilon	float	default= 1e-8
n_iter_no_change	int	default= 10
max_fun	int	default= 15000

หลังจากฝึกแบบจำลองเสร็จสิ้น นำข้อมูล Test Data มาทดสอบประสิทธิภาพของแบบจำลอง โดยการใช้เครื่องมือวัดผลประสิทธิภาพเช่น R2 Score, MAE, RMSE, และ MAPE ซึ่งช่วยให้เข้าใจถึงประสิทธิภาพและความแม่นยำของแบบจำลอง ผลการประเมิน นำไปใช้ประโยชน์ในการปรับปรุงและพัฒนาแบบจำลองในอนาคต อาจจะเป็นการปรับค่าพารามิเตอร์ เพื่อให้ได้ประสิทธิภาพที่ดีที่สุด

3.3.1. ตัวชี้วัดประสิทธิภาพของแบบจำลอง

การประเมินผล (Evaluation) ของแบบจำลอง เป็นขั้นตอนที่สำคัญในการตรวจสอบความแม่นยำของแบบจำลอง โดยจะใช้ข้อมูลที่ไม่เคยใช้ ในการให้แบบจำลองเรียนรู้ เพื่อดูว่าแบบจำลองสามารถทำนายผลได้แม่นยำมากน้อยเพียงใด

การประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องสำหรับทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 ในแบบจำลอง Regression มีวิธีการประเมินผลหลักๆ 4 วิธี(กอบเกียรติ สระอุบล, 2563)

1.R-squared (R2)

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

$$R^2 = 1 - \frac{Rss^2}{Tss^2} = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

RSS (Residual Sum of Square)

ค่าผลรวมของผลต่างระหว่างค่าจริงกับค่าทำนาย

$$RSS = \Sigma(yi - \hat{y}i)^2$$

ยิ่งค่า RSS ยิ่งน้อยเท่าไหร่ แสดงว่า ทำนายค่าได้ดีมากขึ้น

TSS (Total Sum of Square)

ค่าผลรวมของผลต่างระหว่างค่าจริงกับค่าเฉลี่ยของข้อมูล Target Variable (y)

$$TSS = \Sigma(yi - \bar{y})^2$$

2. Mean Absolute Error (MAE)

ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย มีสูตรการคำนวณ ดังนี้

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

โดยที่

MAE คือ ค่าเฉลี่ยความคลาดเคลื่อนที่มาจากการทำนาย

y_i คือ ค่าจริง (Actual Value) ของตัวอย่างที่ i

$\hat{y}_i$ คือ ค่าทำนาย (Predicted Value) ของตัวอย่างที่ i

N คือ จำนวนของ Sample ทั้งหมด

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

3. Root Mean Squared Error (RMSE)

รากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง มีสูตรการคำนวณ ดังนี้

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

โดยที่

RMSE คือ รากที่สองของค่าเฉลี่ยของคลาดเคลื่อนระหว่างค่าจริง (y) กับค่าที่ทำนายยกกำลังสอง ($\hat{\mathbf{y}}$) สำหรับข้อมูล N ตัว

y_i คือ ค่าจริง (actual value) ของตัวอย่างที่ i

$\hat{y}_i$ คือ ค่าทำนาย (predicted value) ของตัวอย่างที่ i

N คือ จำนวนของ Sample ทั้งหมด

Σ คือ ผลรวม

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

4. Mean Absolute Percentage Error (MAPE)

ค่าร้อยละของความคลาดเคลื่อนเฉลี่ยทั้งหมดระหว่างค่าทำนายและค่าจริง (Actual values) มีสูตรการคำนวณ ดังนี้

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{yi}\right| \times 100$$

โดยที่

N คือ จำนวนข้อมูล

y_i คือ ค่าจริง (actual value) ของตัวอย่างที่ i

$\hat{y}_i$ คือ ค่าทำนาย (predicted value) ของตัวอย่างที่ i

MAPE ใช้ในงานที่ต้องการวัดประสิทธิภาพของการทำนาย ซึ่งบอกถึงความคลาดเคลื่อนในรูปแบบร้อยละ หรือ เปอร์เซ็นต์

หลังจากการทดลองได้เสร็จสิ้นลงแล้ว ในถัดไป บทที่4 จะกล่าวถึงผลของการดำเนินการวิจัย ที่ผู้วิจัยได้ทำการศึกษาและทดลองตามกระบวนการขั้นตอนที่ออกแบบไว้ข้างต้น ในบทที่3

บทที่ 4
ผลการศึกษา

ในบทนี้จะกล่าวถึงผลของการดำเนินการวิจัย ที่ผู้วิจัยได้ทำการศึกษาและทดลองตามกระบวนการขั้นตอนที่ออกแบบไว้ ในการวิจัยเกี่ยวกับการใช้วิธีการเรียนรู้ของเครื่องสำหรับทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 ผู้วิจัยได้ดำเนินการวิจัยโดยศึกษาตามขบวนการและขั้นตอนต่างๆ ตลอดจนการประเมินประสิทธิภาพของแบบจำลองตามระยะเวลาในการทำนายล่วงหน้า ดังนี้

1. ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 1 ชั่วโมงล่วงหน้า
2. ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 6 ชั่วโมงล่วงหน้า
3. ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 12 ชั่วโมงล่วงหน้า
4. ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 24 ชั่วโมงล่วงหน้า

โดยในแต่ละการทดลองจะมีการสร้างแบบจำลองใน 4 รูปแบบ เพื่อทำการศึกษาเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลอง

คุณลักษณะพื้นฐานที่ใช้ในการสร้างแบบจำลอง รวมทั้งหมด 27 คอลัมน์ ดังนี้

1. ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) ทั้งหมด 5 คอลัมน์ ดังนี้
   - PM2.5 ค่าฝุ่นละอองขนาดเล็ก (PM2.5)
   - PM2.5(h-1) ค่าฝุ่นละอองขนาดเล็ก PM2.5 1 ชั่วโมงย้อนหลัง
   - PM2.5(h-6) ค่าฝุ่นละอองขนาดเล็ก PM2.5 6 ชั่วโมงย้อนหลัง
   - PM2.5(h-12) ค่าฝุ่นละอองขนาดเล็ก PM2.5 12 ชั่วโมงย้อนหลัง
   - PM2.5(h-24) ค่าฝุ่นละอองขนาดเล็ก PM2.5 24 ชั่วโมงย้อนหลัง
2. ข้อมูลอุตุนิยมวิทยาและสภาพอากาศ ทั้งหมด 22 คอลัมน์ ดังนี้
   - Temperature_C อุณหภูมิ
   - Dew_Point_C จุดน้ำค้าง
   - Humidity_% ความชื้น
   - Speed_kmh ความเร็วลม
   - Gust_kmh ลมกระโชก
   - Pressure_hPa ความกดอากาศ

- Wind_North ลมทิศทางเหนือ North 0.00°
- Wind_NNE ลมทิศทางเหนือ-ตะวันออกเฉียงเหนือ North-Northeast 22.50°
- Wind_NE ลมทิศทางตะวันออกเฉียงเหนือ Northeast 45.00°
- Wind_ENE ลมทิศทางตะวันออก-ตะวันออกเฉียงเหนือ East-Northeast 67.50°
- Wind_East ลมทิศทางตะวันออก East 90.00
- Wind_ESE ลมทิศทางตะวันออก-ตะวันออกเฉียงใต้ East-Southeast 112.50°
- Wind_SE ลมทิศทางตะวันออกเฉียงใต้ Southeast 135.00°
- Wind_SSE ลมทิศทางใต้-ตะวันออกเฉียงใต้ South-Southeast 157.50°
- Wind_South ลมทิศทางใต้ South 180.00°
- Wind_SSW ลมทิศทางใต้-ตะวันตกเฉียงใต้ South-Southwest 202.50°
- Wind_SW ลมทิศทางตะวันตกเฉียงใต้ Southwest 225.00°
- Wind_WSW ลมทิศทางตะวันตก- ตะวันตกเฉียงใต้ West-Southwest 247.50°
- Wind_West ลมทิศทางตะวันตก West 270.00°
- Wind_WNW ลมทิศทางตะวันตก-ตะวันตกเฉียงเหนือ West-Northwest 292.50°
- Wind_NW ลมทิศทางตะวันตกเฉียงเหนือ Northwest 315.00°
- Wind_NNW ลมทิศทางเหนือ-ตะวันตกเฉียงเหนือ North-Northwest 337.50°

ตาราง 8 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 9 คุณลักษณะเพิ่มเติมที่ใช้ในการสร้างแบบจำลอง

คุณลักษณะเพิ่มเติมที่ใช้ในการสร้างแบบจำลอง	**ชื่อคุณลักษณะเพิ่มเติม**	**จำนวนคุณลักษณะทั้งหมดที่ใช้ในการสร้างแบบจำลอง**
Without Mean_PM and Season แบบจำลองโดยใช้คุณลักษณะพื้นฐาน	คุณลักษณะพื้นฐาน	27 คุณลักษณะ
Mean_PM Without Season แบบจำลองโดยการเพิ่มคุณลักษณะค่าเฉลี่ยPM2.5 ย้อนหลัง	Last24hrs_mean Last48hrs_mean Last72hrs_mean	30 คุณลักษณะ
Only Season แบบจำลองโดยการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Season)	Season_Rainy Season_Summer Season_Winter	30 คุณลักษณะ

ตาราง 9 (ต่อ)

คุณลักษณะเพิ่มเติมที่ใช้ในการสร้างแบบจำลอง	ชื่อคุณลักษณะเพิ่มเติม	จำนวนคุณลักษณะทั้งหมดที่ใช้ในการสร้างแบบจำลอง
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

4.1 การเปรียบเทียบ Scaling Method

ในการทดลองนี้ เปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 จากการทดลองด้วย Scaling Method 2 แบบ ได้แก่ Standard Scaling และ Min-Max Scaling ทดสอบโดยใช้แบบจำลอง 4 รูปแบบคือ Linear Regression, Support Vector Regression , Multi-Layer Perceptron , XGBoost ดังนี้

4.1.1 Linear Regression

ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ค่าฝุ่นละอองขนาดเล็ก PM2.5 จากการทดลอง Scaling Method พบว่า ประสิทธิภาพของทั้ง 2 Scaling Method ในแบบจำลอง Linear Regression มีค่าเท่ากัน ทั้งแบบ Standard และ Minmax ดังตาราง 10 ในหน้าถัดไป

ตาราง 10 เปรียบเทียบประสิทธิภาพ Scaling Method ของแบบจำลอง Linear Regression

แบบจำลอง	Scaling	**การประเมิน**	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
LR+1	**Standard**	R2	0.9722	0.972	**0.9722**	0.972
	Standard	MAE	1.6835	1.6929	1.6832	1.6904
	Standard	RMSE	2.4523	2.4616	2.4492	2.4575
	Standard	MAPE (%)	9.0376	9.0408	9.0302	9.0381
	Minmax	R2	0.9722	0.972	**0.9722**	0.9721
	Minmax	MAE	1.6835	1.6929	1.6832	1.6891
	Minmax	RMSE	2.4523	2.4616	2.4492	2.4557
	Minmax	MAPE (%)	9.0376	9.0408	9.0302	9.051
LR+6	**Standard**	R2	0.7655	0.7664	**0.7735**	0.7713
	Standard	MAE	5.0893	5.0061	4.9834	4.9630
	Standard	RMSE	7.1146	7.1015	6.9925	7.0270
	Standard	MAPE (%)	26.8894	25.6720	26.0489	25.4842
	Minmax	R2	0.7655	0.7664	**0.7735**	0.7702
	Minmax	MAE	5.0893	5.0061	4.9834	4.9592
	Minmax	RMSE	7.1146	7.1015	6.9925	7.0270
	Minmax	MAPE (%)	26.8894	25.6720	26.0489	25.2962

4.1.2 Support Vector Regression

ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ค่าฝุ่นละอองขนาดเล็ก PM2.5 จากการทดลอง Scaling Method ของแบบจำลอง Support Vector Regression พบว่า ประสิทธิภาพของ Minmax Scaling ทำได้ดีกว่าStandard Scaling ในทุกกรณี ดังตาราง 11

ตาราง 11เปรียบเทียบประสิทธิภาพ Scaling Method แบบจำลอง Support Vector Regression

แบบจำลอง	Scaling	**การประเมิน**	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
SVR+1	**Standard**	R2	0.8894	0.8903	0.8952	**0.8961**
	Standard	MAE	2.9289	2.9386	2.8714	2.891
	Standard	RMSE	4.8871	4.8692	4.7573	4.7381
	Standard	MAPE (%)	15.6835	15.5568	15.3876	15.3216
	Minmax	R2	0.931	0.9316	**0.941**	0.9384
	Minmax	MAE	2.5526	2.5338	2.4346	2.4744
	Minmax	RMSE	3.8621	3.8454	3.5850	3.6489
	Minmax	MAPE (%)	14.2835	13.9404	13.3705	13.493
SVR+6	**Standard**	R2	0.6857	0.6921	0.7013	**0.7015**
	Standard	MAE	5.5671	5.4661	5.3951	5.382
	Standard	RMSE	8.2374	8.1527	8.0304	8.0276
	Standard	MAPE (%)	28.4051	27.3001	27.3698	27.1172
	Minmax	R2	0.7301	0.7312	**0.7451**	0.7423
	Minmax	MAE	5.2463	5.2170	5.1296	5.1187
	Minmax	RMSE	7.6330	7.6180	7.4175	7.4587
	Minmax	MAPE (%)	26.8390	26.3808	26.4585	25.9228

4.1.3 Multi-Layer Perceptron

ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ค่าฝุ่นละอองขนาดเล็ก PM2.5 จากการทดลอง Scaling Method ของแบบจำลอง Multi-Layer Perceptron พบว่า ประสิทธิภาพของ Minmax Scaling ทำได้ดีกว่า Standard Scaling ในทุกกรณี ดังตาราง 12

ตาราง 12 เปรียบเทียบประสิทธิภาพ Scaling Method ของแบบจำลอง Multi-Layer Perceptron

แบบจำลอง	Scaling	**การประเมิน**	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
MLP+1	**Standard**	R2	0.9666	0.9661	0.9664	**0.9679**
	Standard	MAE	1.8831	1.9481	1.8865	1.8524
	Standard	RMSE	2.6845	2.7072	2.6951	2.6337
	Standard	MAPE (%)	10.4955	11.16	10.4432	10.0525
	Minmax	R2	0.9708	0.9711	**0.9717**	0.9713
	Minmax	MAE	1.7437	1.7345	1.7110	1.7384
	Minmax	RMSE	2.5101	2.4984	2.4709	2.491
	Minmax	MAPE (%)	9.1794	9.1034	9.0232	9.2409
MLP+6	**Standard**	R2	0.7350	0.7218	**0.7367**	0.7183
	Standard	MAE	5.3059	5.3260	5.1779	5.2804
	Standard	RMSE	7.5631	7.7497	7.5397	7.7984
	Standard	MAPE (%)	27.2241	26.7868	25.7757	26.4037
	Minmax	R2	0.7501	0.7435	**0.7594**	0.7526
	Minmax	MAE	5.1294	5.1847	4.9916	5.0364
	Minmax	RMSE	7.3452	7.4416	7.2074	7.3080
	Minmax	MAPE (%)	24.7874	24.4267	24.2252	24.2278

4.1.4 XGBoost

ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ค่าฝุ่นละอองขนาดเล็ก PM2.5 จากการทดลอง Scaling Method ของแบบจำลอง XGBoost พบว่า ประสิทธิภาพของทั้ง 2 Scaling Method ในแบบจำลอง XGBoost มีค่าเท่ากัน ทั้งในแบบ Standard และ Minmax ดังตาราง 13

ตาราง 13 เปรียบเทียบประสิทธิภาพ Scaling Method ของแบบจำลอง XGBoost

แบบจำลอง	Scaling	**การประเมิน**	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
XGB+1	**Standard**	R2	0.9607	0.9577	0.9608	0.9571
	Standard	MAE	2.0122	2.0778	2.0218	2.0854
	Standard	RMSE	2.913	3.0219	2.9105	3.044
	Standard	MAPE (%)	10.7124	11.4128	10.8274	11.2491
	Minmax	R2	0.9607	0.9577	0.9608	0.9571
	Minmax	MAE	2.0122	2.0781	2.0218	2.0865
	Minmax	RMSE	2.913	3.0221	2.9105	3.0451
	Minmax	MAPE (%)	10.7124	11.4137	10.8274	11.2541
XGB+6	**Standard**	R2	0.6827	0.6991	0.6964	0.7029
	Standard	MAE	5.9327	5.5050	5.6648	5.5024
	Standard	RMSE	8.2759	8.0600	8.0953	8.0085
	Standard	MAPE (%)	31.6585	27.5401	29.3036	27.8301
	Minmax	R2	0.6827	0.6992	0.6964	0.7029
	Minmax	MAE	5.9327	5.5039	5.6648	5.5025
	Minmax	RMSE	8.2759	8.0583	8.0953	8.0085
	Minmax	MAPE (%)	31.6585	27.5374	29.3036	27.8301

ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ค่าฝุ่นละอองขนาดเล็ก PM2.5 จากการทดลอง Scaling Method ในทุกๆแบบจำลอง ผลลัพธ์ที่ได้ การใช้ Min-Max Scaling ในทุกๆแบบจำลอง มีประสิทธิภาพที่ดีกว่าเมื่อเทียบกับการใช้ Standard Scaling เนื่องจากมีข้อได้เปรียบต่อ Standard Scaling ในกรณีของข้อมูลที่มีการกระจายตัวไม่แบบปกติหรือไม่เท่ากัน ซึ่ง Standard Scaling ทำงานดีกับข้อมูลที่มีการกระจายตัวแบบปกติและกราฟที่ได้เป็นรูประฆังคว่ำ (Bell curve) แต่ในการวิจัยนี้ข้อมูลมีการกระจายตัวไม่แบบปกติ หรือกราฟไม่เป็นรูประฆังคว่ำ ดังนั้นการใช้ Min-Max Scaling จึงมีประสิทธิภาพมากกว่าสำหรับชุดข้อมูลนี้

4.2 การเปรียบเทียบผลการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 ล่วงหน้าที่ช่วงเวลาต่าง ๆ

ผลเปรียบเทียบการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 1,6,12,24 ชั่วโมงล่วงหน้า ตามลำดับ

4.2.1 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 14 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
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

ตาราง 14 (ต่อ)

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
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

4.2.2 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 6 ชั่วโมงล่วงหน้า

จากผลลัพธ์ประสิทธิภาพของแบบจำลอง 6 ชั่วโมงล่วงหน้า จะเห็นได้ว่าแบบจำลอง Linear 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 15 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 15 (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

4.2.3 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 12 ชั่วโมงล่วงหน้า

จากผลลัพธ์ประสิทธิภาพของแบบจำลอง 12 ชั่วโมงล่วงหน้า จะเห็นได้ว่าแบบจำลอง Linear Regression (LR+12) โดยการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Only Season) ให้ประสิทธิภาพที่ดีที่สุดในการทำนายฝุ่นละอองขนาดเล็ก PM2.5 โดยมีค่า R2 Score สูงที่สุดและค่า MAE MSE RMSE และ MAPE ต่ำที่สุด ด้วยค่า R2 Score 0.7368, MAE: 5.1076, RMSE: 7.5377, MAPE (%): 26.3422 เรียงอันดับ แบบจำลองได้ดังนี้ แบบจำลอง LR+12 ให้ผลลัพธ์สูงที่สุด รองลงมาเป็น MLP+12, SVR+12 และ XBG+12 ตามลำดับ

ตาราง 16 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 12 ชั่วโมงล่วงหน้า

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
LR+12	Training	R2	0.6519	0.6596	0.6625	0.6653
	Training	MAE	5.6593	5.6133	5.5664	5.5529
	Training	RMSE	8.1914	8.1007	8.0665	8.0329
	Training	MAPE (%)	23.7951	23.3704	23.2710	23.1227
	Test	R2	0.7525	0.7476	0.7623	0.7541
	Test	MAE	5.0909	5.0676	4.9645	5.0145
	Test	RMSE	7.3089	7.3808	7.1637	7.2860
	Test	MAPE (%)	26.6627	25.6283	25.6810	25.4297
SVR+12	Training	R2	0.6547	0.6656	0.6709	0.6761
	Training	MAE	5.4919	5.3975	5.3416	5.2856
	Training	RMSE	8.1591	8.0291	7.9652	7.9015
	Training	MAPE (%)	21.8973	21.5260	21.4348	21.1468
	Test	R2	0.7168	0.7144	0.7368	0.7348
	Test	MAE	5.2828	5.3049	5.1076	5.0696
	Test	RMSE	61.1259	61.6556	56.8171	57.2529
	Test	MAPE (%)	7.8183	7.8521	7.5377	7.5666
MLP+12	Training	R2	0.6724	0.6850	0.6962	0.7036
	Training	MAE	5.4956	5.3909	5.2605	5.2411
	Training	R2	0.6724	0.6850	0.6962	0.7036
	Training	RMSE	7.9468	7.7922	7.6524	7.5595
	Training	MAPE (%)	22.8596	22.0153	21.6425	21.6441
	Test	R2	0.7513	0.7383	0.7556	0.7398
	Test	MAE	4.9848	5.0251	4.8944	5.0174
	Test	RMSE	7.3269	7.5163	7.2636	7.4940
	Test	MAPE (%)	24.3505	23.6528	23.8601	24.4238

ตาราง 16 (ต่อ)

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
XGB+12	Training	R2	0.9531	0.9733	0.9545	0.9744
	Training	MAE	2.2010	1.6886	2.1681	1.6578
	Training	RMSE	3.0057	2.2674	2.9616	2.2203
	Training	MAPE (%)	10.1551	7.9407	9.9863	7.8249
	Test	R2	0.6250	0.6621	0.6626	0.6898
	Test	MAE	6.2164	5.7977	5.8026	5.6152
	Test	RMSE	8.9973	8.5407	8.5343	8.1830
	Test	MAPE (%)	32.0086	28.4651	29.0184	28.0269

4.2.4 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 เป็นเวลา 24 ชั่วโมงล่วงหน้า

จากผลลัพธ์ประสิทธิภาพของแบบจำลอง 12 ชั่วโมงล่วงหน้า จะเห็นได้ว่าแบบจำลอง Linear Regression (LR+12) โดยการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Only Season) ให้ประสิทธิภาพที่ดีที่สุดในการทำนายฝุ่นละอองขนาดเล็ก PM2.5 โดยมีค่า R2 Score สูงที่สุดและค่า MAE MSE RMSE และ MAPE ต่ำที่สุด ด้วยค่า R2 Score 0.7461, MAE: 5.1030 , RMSE: 7.4033, MAPE (%): 26.0673 เรียงอันดับ แบบจำลองได้ดังนี้ แบบจำลอง LR+24 ให้ผลลัพธ์สูงที่สุด รองลงมาเป็น MLP+24, SVR+24 และ XBG+24 ตามลำดับ

ตาราง 17 ผลการทำนายระดับค่าฝุ่นละอองขนาดเล็ก PM2.5 ใน 24 ชั่วโมงล่วงหน้า

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
LR+24	Training	R2	0.6134	0.6254	0.6296	0.6358
	Training	MAE	5.9774	5.9324	5.8323	5.8195
	Training	RMSE	8.6328	8.4976	8.4498	8.3796
	Training	MAPE (%)	25.1692	24.8079	24.4090	24.3064
	Test	R2	0.7295	0.7187	0.7461	0.7355
	Test	MAE	5.3019	5.3390	5.1030	5.1813
	Test	RMSE	7.6406	7.7918	7.4033	7.5555
	Test	MAPE (%)	27.5277	26.5933	26.0673	25.9800
SVR+24	Training	R2	0.6203	0.6324	0.6348	0.6430
	Training	MAE	5.7638	5.6630	5.5936	5.5171
	Training	RMSE	8.5559	8.4185	8.3905	8.2965
	Training	MAPE (%)	23.0514	22.5969	22.3450	21.9777
	Test	R2	0.6994	0.6971	0.7254	0.7216
	Test	MAE	5.4047	5.4260	5.1855	5.1845
	Test	RMSE	8.0542	8.0861	7.6985	7.7516
	Test	MAPE (%)	27.6027	27.4255	26.7576	26.3560
MLP+24	Training	R2	0.6339	0.6509	0.6569	0.6731
	Training	MAE	5.7664	5.7010	5.5859	5.5133
	Training	RMSE	8.4008	8.2035	8.1325	7.9383
	Training	MAPE (%)	23.5544	23.3709	23.0497	22.7946
	Test	R2	0.6455	0.6275	0.6512	0.6244
	Test	MAE	6.0571	6.2158	5.9691	6.1744
	Test	RMSE	8.7466	8.9667	8.6766	9.0040
	Test	MAPE (%)	30.9229	31.2909	30.6313	30.7718

ตาราง 17 (ต่อ)

แบบจำลอง	ชุดข้อมูล	การประเมิน	Without Mean_PM and Season	Mean_PM Without Season	Only Season	Mean_PM and Season
XGB+24	Training	R2	0.9462	0.9697	0.9532	0.9744
	Training	MAE	2.3571	1.7751	2.2034	1.6511
	Training	RMSE	3.2207	2.4187	3.0042	2.2220
	Training	MAPE (%)	10.8807	8.3128	10.2909	7.7660
	Test	R2	0.6455	0.6275	0.6512	0.6244
	Test	MAE	6.0571	6.2158	5.9691	6.1744
	Test	RMSE	8.7466	8.9667	8.6766	9.0040
	Test	MAPE (%)	30.9229	31.2909	30.6313	30.7718

ผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ทั้งหมด ใน 4 ช่วงเวลา ได้แก่ +1, +6 ,+12 ,+24 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

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยการเรียนรู้ของเครื่องสำหรับการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก PM2.5 ผู้วิจัยได้วัดประสิทธิภาพของแบบจำลองในแต่ขั้นตอนต่างๆ เพื่อนำมาเปรียบเทียบและสรุปผลการวิจัย โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

สถานการณ์ในปัจจุบัน พบปัญหาหมอกควันและฝุ่นละอองขนาดเล็ก โดยเฉพาะฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน (PM2.5) ถือเป็นปัญหาสำคัญของประเทศไทย เนื่องจากสถานการณ์ PM2.5 เกินค่ามาตรฐานในทุกปี เพื่อเข้าใจแนวโน้มของสถานการณ์ปริมาณ PM2.5 ทำให้ผู้เกี่ยวข้อง สามารถใช้ในการวางแผนการจัดการที่เหมาะสม ช่วยลดผลกระทบต่อสุขภาพของมนุษย์ ที่อยู่ในพื้นที่ที่มีความเสี่ยงต่อฝุ่นละอองขนาดเล็ก (PM2.5) สร้างความเข้าใจในผลกระทบต่อสุขภาพ ประเมินระดับความเสี่ยงที่เกิด รวมถึงพัฒนาแนวทางการบริหารจัดการในระยะยาว ในการรับมือปริมาณฝุ่นละอองขนาดเล็ก (PM2.5) ในอนาคตได้

ในงานวิจัยนี้เป็นการศึกษาโดยนำข้อมูลภาคอุตุนิยมวิทยาที่เกิดขึ้นในอดีตมาใช้ร่วมกับเทคนิคการเรียนรู้ของเครื่อง เพื่อสร้างแบบจำลองเบื้องต้นที่ใช้สำหรับในการทำนายค่าความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ล่วงหน้า ในการดำเนินงานวิจัยนี้ มีการสร้างชุดข้อมูล 2 ชุดมารวมกัน โดยนำเข้าข้อมูลจากแหล่งข้อมูลสาธารณะแบบเปิด ผ่านหน้าเว็บไซต์และวิธีการ Web Scraping ซึ่งเป็นกระบวนการในการดึงข้อมูล โดยใช้สคริปต์ดึงข้อมูลจากหน้าเว็บไซต์ ดังนี้ 1. ข้อมูลฝุ่นละอองขนาดเล็ก (PM2.5) จากเว็บไซต์ Berkeley Earth 2. ข้อมูลภาคอุตุนิยมวิทยา จากเว็บไซต์ Weather Underground นำมาจากสถานี IKRUNGTH3 บริเวณ วิภาวดี 60 หลักสี่ กรุงเทพมหานคร อยู่ในช่วงวันที่ 1 มกราคม - 31 ธันวาคม 2562 และช่วง 1 มกราคม – 31 ธันวาคม 2563 ภายในชุดข้อมูลมีตัวแปรที่สามารถส่งผลต่อค่า PM2.5 ได้แก่ อุณหภูมิ, จุดน้ำค้าง , ความชื้น, ทิศทางลม, ความเร็วลม ,ลมกระโชก และ ความกดอากาศ จากนั้นนำข้อมูล Raw Data ที่ได้มาเตรียมให้อยู่ในรูปแบบที่เหมาะสำหรับการสร้างแบบจำลอง Pre-Processing เช่น การทำการ One-Hot Encoding ในการจัดการกับคุณลักษณะที่ไม่มีลำดับ , สร้างคอลัมน์ใหม่

จากข้อมูลคอลัมน์ “PM2.5” ด้วย โดยการ shift ข้อมูล เพื่อใช้เป็นข้อมูล Target ที่จะวิเคราะห์ PM 2.5 ในชั่วโมงล่วงหน้า, จัดการข้อมูลสูญหาย โดย ในที่นี้ใช้ 2 วิธี คือ “ .fillna (method='bfill') และ .fillna (method='ffill')” ,ทำการ Resample ข้อมูล เป็นช่วงเวลารายชั่วโมง H เนื่องจากข้อมูลดิบในภาคอุตุนิยมวิทยา ที่ได้จากทางสถานี มีระยะเวลาในการเก็บค่าที่ไม่เท่ากันคนละความถี่ จึงต้องมีการแปลง Index ก่อนนำมารวม Data Frame หลังจากรวบรวมข้อมูลเรียบร้อย ทำการแบ่งชุดข้อมูลที่ทำความสะอาดแล้ว 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%

ในการวิจัยนี้ ได้ใช้ แบบจำลอง ทั้งหมด 4 แบบ ที่ได้รับความนิยมและยอมรับในปัจจุบัน ได้แก่ LR (Linear Regression), SVR (Support Vector Regression) , XGBoost (eXtreme Gradient Boosting) และ MLP (Multi-Layer Perceptron) จากนั้นเริ่มทดลองสร้างแบบจำลองเบื้องต้นที่ใช้ในการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 โดยใช้ ค่าพารามิเตอร์เริ่มต้น จาก scikit-learn ที่ครอบคลุมการทำนายค่า PM2.5 หรือใช้ Target PM2.5 ของชั่วโมงที่ต่างกัน ในช่วงเวลา +1 ชั่วโมง, +6 ชั่วโมง, +12 ชั่วโมง, และ +24 ชั่วโมงล่วงหน้า ตามลำดับ แต่ละช่วงเวลาจะแบ่งเป็น 4 แบบย่อย ซึ่งเกิดจากการเพิ่มเติมคุณลักษณะที่ต่างกัน 4 กรณีเพื่อทดสอบสมมุติฐานที่เกี่ยวข้อง ดังนี้ 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 การสร้างแบบจำลองโดยการเพิ่มคุณลักษณะค่าเฉลี่ย PM2.5 ย้อนหลัง และข้อมูลฤดูกาล(Season) ในวิจัยนี้ทำการ Scaling ข้อมูล ใน 2 รูปแบบคือ Standard Scaling และ Min-Max Scaling เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมสำหรับการสร้างและฝึกสอนแบบจำลอง หลังจากฝึกแบบจำลองเสร็จสิ้น นำข้อมูล Test Data มาทดสอบประสิทธิภาพของแบบจำลอง โดยการใช้เครื่องมือวัดผลประสิทธิภาพสำหรับแบบ Regression เช่น R2 Score, MAE, RMSE, และ MAPE

สรุปได้ว่า การทดลอง Scaling Method ในทุกๆแบบจำลอง ผลลัพธ์ที่ได้ การใช้ Min-Max Scaling ในทุกๆ แบบจำลอง มีประสิทธิภาพที่ดีกว่าเมื่อเทียบกับการใช้ Standard Scaling และผลเปรียบเทียบประสิทธิภาพของแบบจำลอง ทั้งหมด ใน 4 ช่วงเวลา ได้แก่ +1, +6 ,+12 ,+24 ชั่วโมงล่วงหน้า พบว่าแบบจำลอง LR- Linear Regression ผลลัพธ์ที่ดีที่สุด ทั้งในแง่ของความถูกต้องแม่นยำ และความคลาดเคลื่อนที่ต่ำลดลง รองลงมาเป็น แบบจำลอง MLP Multi-Layer Perceptron , SVR- Support Vector Regression และ แบบจำลอง XGBoost ตามลำดับ แบบจำลองทุกแบบ ผลลัพธ์จะลดลงตามช่วงเวลาที่มากขึ้น 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 ที่ไม่ใช้ข้อมูลอุตุนิยมวิทยาฤดูกาลเข้ามาเกี่ยวข้อง

5.2. อภิปรายผลการวิจัย

ในการศึกษาและสร้างแบบจำลองเบื้องต้นที่ใช้สำหรับทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 โดยใช้แบบจำลองการเรียนรู้ของเครื่องทั้งหมด 4 แบบ ได้แก่ LR (Linear Regression), SVR (Support Vector Regression) , XGBoost (eXtreme Gradient Boosting) และ MLP (Multi-Layer Perceptron) ในการประเมินประสิทธิภาพจำลองสำหรับแบบ Regression ได้ใช้ตัวชี้วัดประสิทธิภาพ ได้แก่ R2 Score, MAE, RMSE, และ MAPE

จากการทดลอง Scaling Method ในทุกๆแบบจำลอง ผลลัพธ์ที่ได้ การใช้ Min-Max Scaling ในทุกๆ แบบจำลอง มีประสิทธิภาพที่ดีกว่าเมื่อเทียบกับการใช้ Standard Scaling เนื่องจากมีข้อได้เปรียบต่อ Standard Scaling ในกรณีของข้อมูลที่มีการกระจายตัวไม่แบบปกติหรือไม่เท่ากัน ซึ่ง Standard Scaling ทำงานดีกับข้อมูลที่มีการกระจายตัวแบบปกติและกราฟที่ได้เป็นรูประฆังคว่ำ (Bell curve) แต่ในการวิจัยนี้ข้อมูลมีการกระจายตัวไม่แบบปกติ หรือกราฟไม่เป็นรูประฆังคว่ำ ดังนั้นการใช้ Min-Max Scaling จึงมีประสิทธิภาพมากกว่าสำหรับชุดข้อมูลนี้

ซึ่งในการจัดการ การกระจายของข้อมูลเพื่อให้ง่ายต่อการวิเคราะห์และประมวลผล อาจต้องปรับ Variable Distribution ให้เป็นรูปแบบปกติก่อน ซึ่งมีหลายวิธี เช่น Z-Score Standardization เพื่อปรับ Variable Distribution ให้มี mean เป็น 0 และ standard deviation เป็น 1 รวมถึงการลด Outliers ที่อาจทำให้ Variable Distribution ไม่ปกติ.ก่อนทำการ Scaling

จากการทดลองแบบจำลองทั้งหมด 4 แบบ ในตารางผลของแบบจำลอง พบว่าแบบจำลอง LR- Linear Regression และแบบจำลอง MLP- Multi-Layer Perceptron นั้นอยู่ในกลุ่มของได้ผลลัพธ์ที่ดีที่สุด ปัจจัยที่ทำให้แบบจำลอง LR- Linear Regression ได้ผลลัพธ์ที่ดีที่สุดมีได้หลายปัจจัย เช่น Linear Relationship มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรต้นและตัวแปรตาม จากในชุดข้อมูล ความสัมพันธ์ของข้อมูลส่วนใหญ่ มีลักษณะเป็น Linear ส่งผลให้ตัวแบบจำลอง LR- Linear Regression สามารถเรียนรู้และทำนายผลได้แม่นยำที่สุด ส่วนแบบจำลอง MLP- Multi-Layer Perceptron เป็น แบบจำลอง Neural Network โครงข่ายเซลล์ประสาท ทำให้สามารถเรียนรู้ข้อมูลที่ซับซ้อนมากขึ้นได้ ทำให้ทำนายผลออกมาได้อย่างมีประสิทธิภาพ

ส่วนในกลุ่มผลลัพธ์น้อยทั้ง 4 ช่วงเวลา พบว่า แบบจำลอง SVR- Support Vector Regression และ แบบจำลอง XGBoost เนื่องจากในวิจัยนี้ได้สร้างแบบจำลองเบื้องต้นที่ใช้ค่าพารามิเตอร์เริ่มต้น จาก scikit-learn ไม่ได้ตั้งค่า ซึ่งแบบจำลองทั้งสองนั้น สามารถกำหนดค่าพารามิเตอร์ได้หลากหลายรูปแบบ ให้เหมาะสมกับข้อมูล เช่นค่า Kernel Functions ซึ่งค่า

Default คือแบบ 'rbf' แต่ข้อมูลมีลักษณะ Linear ทำให้การเรียนรู้อาจจะไม่ได้เหมาะสมกับข้อมูลที่มี ส่งผลให้แบบจำลองเรียนรู้และทำนายได้อย่างไม่ค่อยมีประสิทธิภาพเท่าที่ควร รวมถึงมีข้อสังเกตจากผลการทดลอง ปัจจัยที่อาจทำให้ผลลัพธ์ของ Linear Regression (LR) สูงกว่า XGBoost ที่เป็นแบบจำลองที่จับความซับซ้อนได้ดี ได้แก่ 1. เกิด Overfitting ของแบบจำลอง XGBoost จากตารางผลการทดลองค่าการทำนายที่วัดได้จากชุดข้อมูลสำหรับฝึกแบบจำลอง (Training set)เมื่อเปรียบเทียบกับชุดข้อมูลทดสอบ (Test set) ผลลัพธ์ในแบบจำลอง XGBoost ให้ค่าที่สูงกว่า LR ในชุดข้อมูลTraining set แต่ผลลัพธ์ลดลงในชุดTest set ซึ่งสลับกันกับแบบจำลอง LR ให้ค่าความแม่นยำที่สูงกว่าในชุด Test set แสดงให้เห็นถึงการเกิด Overfitting หรือ Underfitting ของแบบจำลอง ซึ่ง XGBoost เกิดได้ง่ายหากมีการเลือก Hyperparameter ที่ไม่เหมาะสม 2. Data Size ขนาดของข้อมูล ซึ่ง LR มักทำได้ดีกับข้อมูลที่ไม่ใหญ่มาก ซึ่งในชุดข้อมูลนี้มีขนาดไม่ใหญ่มาก 3. Data Complexity ซึ่งแบบจำลอง XGBoost เป็นโมเดลที่ซับซ้อนจับความซับซ้อนในข้อมูลได้ดี แต่ถ้าความซับซ้อนน้อยและมีความเชิงเส้นมากในข้อมูล แบบจำลอง LR- Linear Regression อาจนำเสนอประสิทธิภาพที่ดีกว่า และจากแบบจำลองทั้ง 4 ช่วงเวลา ผลลัพธ์ลดลงตามจำนวนชั่วโมงที่มากขึ้น เนื่องจากห่างจากจุดข้อมูลที่เป็น PM 2.5 เวลาจริง

จาก 4 กรณีการทดลอง ซึ่งเกิดจากการเพิ่มเติมคุณลักษณะที่ต่างกัน 4 กรณีเพื่อทดสอบสมมุติฐานที่เกี่ยวข้อง โดยมีการเพิ่มตัวแปรด้านค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลังและตัวแปรด้านฤดูกาล (Season) ในข้อมูลนำเข้า Input Data เพื่อใช้ในการสร้างแบบจำลอง พบว่าแบบจำลอง Only Season ที่มีการเพิ่มคุณลักษณะของข้อมูลฤดูกาล (Season) ให้ผลลัพธ์ที่ดีที่สุด สังเกตได้ว่าการเพิ่มคุณลักษณะแบบที่ใช้ Only Season และ Mean_PM and Season ซึ่งเป็นแบบจำลองที่มีฤดูกาลเข้ามาเกี่ยวข้อง จะอยู่ในกลุ่มที่ให้ผลลัพธ์สูงกว่า แบบ Mean_PM Without Season และ Without Mean_PM and Season ที่ไม่ใช้ข้อมูลอุตุนิยมวิทยา ฤดูกาลเข้ามาเกี่ยวข้อง แสดงว่า ตัวแปรด้านฤดูกาล (Season) มีองค์ประกอบที่สำคัญในการทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 ในชุดข้อมูลนี้ ตอบสมมุติฐานที่ว่า ตัวแปรด้านค่าเฉลี่ยฝุ่นละอองขนาดเล็ก PM2.5 ย้อนหลัง มีผลต่อความแม่นยำของแบบจำลองเพียงเล็กน้อย เมื่อเทียบกับการใช้ตัวแปรด้านฤดูกาล จะมีผลต่อความแม่นยำของแบบจำลองการทำนายความเข้มของฝุ่นละอองขนาดเล็ก (PM2.5) ที่สูงกว่า

สุดท้ายนี้ ในงานวิจัยอื่นๆ แต่ละงานวิจัยมีความแตกต่างกันทั้งชุดข้อมูล และแบบจำลองที่ใช้ รวมถึงคุณภาพเซนเซอร์ของแต่ละสถานีตรวจวัดที่มีความแตกต่าง จากความแตกต่างดังกล่าวทำให้ ไม่สามารถสร้างการเปรียบเทียบกันได้แบบตรงไปตรงมาได้ เพราะข้อมูลคนละชุดกัน ซึ่งข้อมูลที่ได้มาส่วนใหญ่มากจากการเก็บข้อมูลจากภายในหน่วยงานหรือองค์กรเกี่ยวกับอุตุนิยมวิทยาของประเทศนั้นๆ ซึ่งเป็นข้อมูลปิด ต่างจากงานวิจัยนี้ ที่การได้มาของข้อมูลไม่ใช่ข้อมูลสำเร็จรูป เป็นการสร้างชุดข้อมูลขึ้นมาจากแหล่งข้อมูลสาธารณะ 2 แหล่ง มารวมกัน เพื่อใช้สร้างแบบจำลองเบื้องต้น ในบางงานวิจัยใช้ชุดข้อมูลแบบปิด ทำให้ไม่สามารถเข้าไปใช้ข้อมูลเพื่อพัฒนาแบบจำลองต่อได้ แต่ในงานวิจัยนี้ใช้ข้อมูลสาธารณะ ซึ่งประโยชน์ของชุดข้อมูลสาธารณะนั้น ทำให้สามารถเข้าถึงได้ สามารถสร้างการเปรียบเทียบได้ รวมถึงให้นักพัฒนาที่สนใจ สามารถนำไปพัฒนาแบบจำลองต่อได้

การทดลองนี้ เป็นการสร้างแบบจำลองในแบบเบื้องต้น โดยเป็นการหาข้อมูลจากแหล่งข้อมูลสาธารณะที่เข้าถึงได้ แล้วนำมาใช้ทำนายค่าฝุ่นละอองขนาดเล็ก PM2.5 ยังไม่ได้มีการปรับค่าพารามิเตอร์ ซึ่งการใช้พารามิเตอร์เริ่มต้นจาก scikit-learn อาจมีผลต่อผลลัพธ์ เห็นถึงความสำคัญในการปรับแต่งพารามิเตอร์เพื่อให้มีประสิทธิภาพที่สูงขึ้น ในงานวิจัยนี้ มีเพียงทดลองเบื้องต้นในการปรับค่าพารามิเตอร์ของแบบจำลอง SVR โดยใช้งาน Grid Search ค้นหาพารามิเตอร์ที่เหมาะสมที่สุดให้แบบจำลอง โดยกำหนดพารามิเตอร์ที่จะค้นหาใน SVR ดังนี้ kernel : ['linear', 'rbf'] และ ค่า 'C': [0.1, 1, 10] หลังใช้งาน Grid Search พารามิเตอร์ Model: SVR (C=0.1, kernel='linear') ได้ผลลัพธ์ที่ดีที่สุด ด้วยค่า R2 Score 0.9722, MAE: 1.6702 MSE: 6.0393 RMSE: 2.4575, MAPE (%): 8.9454 จะเห็นได้ว่าผลลัพท์ดีขึ้น ซึ่งเป็นแนวทางในอนาคตที่ว่า แบบจำลองควรมีการปรับจูน Model (Model Tuning) ค้นหาพารามิเตอร์ที่เหมาะสมที่สุด ช่วยทำให้ได้ผลของการประเมินแบบจำลองที่สูงขึ้นได้

5.3. ข้อเสนอแนะ

1.ในงานวิจัยนี้ ในขั้นตอนของการเก็บรวบรวมข้อมูลข้อมูลสาธารณะ พบว่ามีข้อมูลขาดหาย เนื่องจากการปิดเปิดเซ็นเซอร์วัดของสถานี ทำให้ข้อมูลบางส่วนขาดหาย จึงได้จัดการโดยการ Fill ซึ่งการ Drop ทิ้งทั้งแถว อาจทำให้ข้อมูลที่จำเป็นต้องใช้บางส่วนนั้นถูกลบ

2.เก็บรวบรวมข้อมูลใช้ข้อมูลเพียงแค่ปีเดียว หากมีการใช้ข้อมูลเพิ่มในหลายๆปี อาจช่วยเพิ่มประสิทธิภาพในการทำนายให้ดีมากขึ้นได้

3.ในงานวิจัยนี้ใช้งานเพียงการเรียนรู้ของเครื่อง ยังมีอีกหลายแบบจำลอง อาจจะสามารถเรียนรู้ข้อมูลจำนวนมาก ซับซ้อนมากขึ้น และอาจทำนายค่าได้แม่นยำกว่า เช่น การใช้แบบจำลองที่เกี่ยวข้องข้อมูลอนุกรมเวลา Time Series โดยตรง หรือ Deep Learning

4. ในอนาคต อาจมีการพิจารณา จัดเก็บรวบรวมคุณลักษณะของข้อมูลเพิ่มเติม ที่อาจมีผลต่อคุณภาพของการทำนาย หรือข้อมูลที่เกี่ยวข้องกับฝุ่นละอองขนาดเล็ก เช่น ตัวแปรที่เกี่ยวข้องกับอากาศ ข้อมูลการก่อสารพิษทางอากาศ เช่น CO (คาร์บอนมอนอกไซด์), NO (ไนตริกออกไซด์), NO2 (ไนโตรเจนไดออกไซด์) ,ตัวแปรคมนาคม , สถานที่ที่ส่งผลต่อคุณภาพอากาศ ซึ่งอาจจะช่วยเพิ่มประสิทธิภาพของแบบจำลองได้

5.ในวิจัยนี้ ศึกษาเพียงบริเวณสถานีตรวจวัดคุณภาพอากาศเดียว หากลองใช้ข้อมูลจากบริเวณอื่นร่วมด้วย อาจทำให้ได้ผลลัพธ์ที่แม่นยำมากขึ้น ลองสร้างแบบจำลองเฉพาะแต่ละภูมิภาค เพราะแต่ละพื้นที่มีสภาพภูมิประเทศรวมสภาพภูมิอากาศที่แตกต่างกันออกไป

6.ในงานวิจัยนี้ ใช้พารามิเตอร์เริ่มต้นจาก scikit-learn อาจมีผลต่อผลลัพธ์การทำนาย หากมีการปรับจูน Model (Model Tuning) ค้นหาพารามิเตอร์ที่เหมาะสมที่สุด อาจช่วยทำให้ได้ผลของการประเมินแบบจำลองที่สูงขึ้นได้

7. สำหรับผู้วิจัยและนักพัฒนาที่สนใจ พัฒนาแบบจำลองต่อ สามารถตามลิ้งค์ที่ระบุในอ้างอิงได้ โดยคุณลักษณะที่มีการสำรวจมาเบื้องต้น เป็นไปตามรายงานฉบับนี้

บรรณานุกรม

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

Chen & Guestrin. (2016). A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. Retrieved 13-17 August 2016, from Retrieved from https://www.scirp.org/reference/referencespapers.aspx?referenceid=2962367

Doreswamy, K S, H., Km, Y., & Gad, I. (2020). Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. *Procedia Computer Science*, 171, 2057-2066. https://doi.org/https://doi.org/10.1016/j.procs.2020.04.221

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

Jason Brownlee. (2021). *Extreme Gradient Boosting (XGBoost) Ensemble in Python*. Retrieved from https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/

Jay Chugh. (2018). *Types of Machine Learning and Top* 10 *Algorithms Everyone Should Know*. Retrieved from https://blogs.oracle.com/ai-and-datascience/post/types-of-machine-learning-and-top-10-algorithms-everyone-should-know

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

Kasidis Satangmongkol. (2022). เข้าใจการทำงานพื้นฐานของ *Neurons* ใน *Neural Networks*. DataRockie.

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. https://doi.org/https://doi.org/10.1016/j.eswa.2023.119696

NGThai. (2020). PM 2.5: มลพิษทางอากาศในไทย เหตุใดจึงยังไม่สิ้นสุด. *National Geographic*.

Samayan Bhattacharya, S. S. (2021). Using Machine Learning to Predict Air Quality Index in New Delhi. *arxiv*.

scikit-learn. *scikit-learn Machine Learning in Python*. Retrieved from https://scikit-learn.org/stable/index.html

Scikit-learn.org. *Sklearn.neural_network.MLPRegressor*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html#sklearn.neural_network.MLPRegressor

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

Xiao, F., Yang, M., Fan, H., Fan, G., & Al-qaness, M. A. A. (2020). An improved deep learning model for predicting daily PM2.5 concentration. *Scientific Reports*, 10(1), 20988. https://doi.org/10.1038/s41598-020-77757-w

Yamany, W. (2015). Moth-flame optimization for training Multi-Layer Perceptrons. *Conference:* 2015 11*th International Computer Engineering Conference*. https://doi.org/10.1109/ICENCO.2015.7416360

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

กระทรวงสาธารณสุข, ก. (2566). รายงานสถานการณ์และผลการดำเนินงานด้านการแพทย์และสาธารณสุข กรณี ฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน ปี 2565. กองประเมินผลกระทบต่อ

สุขภาพ กรมอนามัย กระทรวงสาธารณสุข. สืบค้นจาก https://hia.anamai.moph.go.th/th/handbook/3912#wow-book/

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

บัญชา ปะสีละเตสัง. (2564). สร้างการเรียนรู้สำหรับ *AI* ด้วย *Python Machine Learning*. ซีเอ็ดยูเคชั่น.

ราชกิจจานุเบกษา. (2565). กำหนดมาตรฐานฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน ในบรรยากาศโดยทั่วไป (เล่ม 139 ตอนพิเศษ 163 ง หน้า 21).

สำนักหอสมุดมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี. (2565). *PM* 10 และ *PM* 2.5 คืออะไร. สืบค้นจาก https://www.lib.kmutt.ac.th/pm-10-%e0%b9%81%e0%b8%a5%e0%b8%b0-pm-2-5-%e0%b8%84%e0%b8%b7%e0%b8%ad%e0%b8%ad%e0%b8%b0%e0%b9%84%e0%b8%a3/

ประวัติผู้เขียน