การทำนายค่าปริมาณฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอน ในเขตจังหวัดเชียงใหม่โดยใช้การเรียนรู้ของเครื่อง

Using Machine Learning Techniques to Predict $PM_{2.5}$ Level in Chiang Mai, Thailand

ประหยัด ปวงจักร์ทา

590510559

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

ปีการศึกษา 2562

การทำนายค่าปริมาณฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอนในเขตจังหวัดเชียงใหม่โดยใช้เทคนิคการเรียนรู้ของเครื่อง

Using Machine Learning Techniques to Predict $PM_{2.5}$ Level in Chiang Mai, Thailand

ประหยัด ปวงจักร์ทา

590510559

การค้นคว้าอิสระนี้ได้รับการพิจารณาอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่
ปีการศึกษา 2562

คณะกรรมการสอบการค้นคว้าอิสระ

..... ประธานกรรมการ
(อาจารย์ ดร.ปราการ อุณจักร)
..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.จักริน ชวชาติ)

วันที่.....เดือน.....พ.ศ.....

กิตติกรรมประกาศ

รายงานการค้นคว้าอิสระเล่มนี้เป็นส่วนหนึ่งของหลักสูตรปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ ในกระบวนวิชา 204499 ซึ่งผู้ค้นคว้าได้จัดทำในหัวข้อเรื่องการทำนายค่าฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอนในเขตจังหวัดเชียงใหม่โดยใช้เทคนิคการเรียนรู้ของเครื่อง

การค้นคว้าอิสระเล่มนี้สำเร็จลุล่วงได้โดยความอนุเคราะห์จากบุคคลหลายท่านขอกราบขอบพระคุณอาจารย์ ดร.ปราการ อุณจักร ซึ่งเป็นอาจารย์ที่ปรึกษา ที่กรุณาให้ความรู้ แนวคิด คำแนะนำ วิธีการในการแก้ปัญหา และคำปรึกษารวมทั้งได้เสียสละเวลาอันมีค่าในการตรวจแก้ไขข้อบกพร่องของเนื้อหาและสำนวนภาษาไทยด้วยความใส่ใจจนทำให้การค้นคว้าอิสระนี้สำเร็จลุล่วง

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.จักริน ชวชาติ ที่กรุณารับเป็นกรรมการสอบการค้นคว้าอิสระนี้รวมทั้งได้ให้คำแนะนำเป็นอย่างดีมาโดยตลอด

ขอขอบพระคุณ อาจารย์ ดร.ว่าน วิริยา อาจารย์ประจำภาควิชาเคมี คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่ได้ให้ความอนุเคราะห์การให้คำปรึกษาตลอดเวลาของการดำเนินงาน

ขอขอบพระคุณกรมอุตุนิยมวิทยาและกรมควบคุมมลพิษ ที่ได้ให้ความอนุเคราะห์ข้อมูลในส่วนของการตรวจวัดคุณภาพอากาศและข้อมูลอื่นๆ ที่เกี่ยวข้องกับการทำวิจัยฉบับนี้

สุดท้ายนี้ขอกราบขอบพระคุณคุณพ่อ คุณแม่ และครอบครัว ผู้เป็นที่รัก ผู้ให้กำลังใจและให้โอกาสการศึกษาอันมีค่ายิ่งรวมถึงทุกความช่วยเหลือ ทุกคำแนะนำ ทุกคำติชม ซึ่งส่งผลให้การทำการค้นคว้าอิสระนี้สำเร็จลุล่วงไปด้วยดี

นายประหยัด ปวงจักร์ทา

590510559

หัวข้อการค้นคว้าอิสระ	การทำนายค่าปริมาณฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอนในเขตจังหวัดเชียงใหม่โดยใช้การเรียนรู้ของเครื่อง
ชื่อเจ้าของโครงงาน	นาย ประหยัด ปวงจักร์ทา **รหัสประจำตัว** 590510559
วิทยาศาสตรบัณฑิต	สาขาวิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.ปราการ อุณจักร

บทคัดย่อ

เนื่องจากปัญหามลพิษทางอากาศในจังหวัดเชียงใหม่โดยเฉพาะปัญหาฝุ่นละอองที่มีขนาดเล็กกว่า 2.5 ไมครอน (Particulate Matters 2.5: PM2.5) มีความหนาแน่นสูงเกินกว่าระดับมาตรฐานที่องค์กรอนามัยโลก (World Health Organization: WHO) ได้กำหนดไว้และความหนาแน่นที่สูงเกินไปส่งผลให้เกิดผลเสียแก่จังหวัดเชียงใหม่ ได้แก่ ปัญหาทางสุขภาพของประชาชน ปัญหาทางเศรษฐกิจที่เกี่ยวกับการท่องเที่ยว โดยปัญหาที่เกิดขึ้นนี้เป็นปัญหาที่เกิดขึ้นมานานหลายปีและมีแนวโน้มที่จะเกิดขึ้นต่อเนื่องในอนาคต ซึ่งในปัจจุบันยังไม่สามารถหาแนวทางในการแก้ไขได้อย่างชัดเจนและไม่สามารถแก้ไขได้ในระยะเวลาอันสั้น จึงทำให้ผู้วิจัยได้เล็งเห็นถึงปัญหาและตระหนักถึงผลกระทบที่จะเกิดขึ้นในอนาคต จากปัญหาดังกล่าวทำให้ผู้วิจัยได้นำข้อมูล มลพิษทางอากาศ ข้อมูลอุตุนิยมวิทยาและข้อมูลจุดความร้อนที่เกิดขึ้น ณ บริเวณจุดที่สนใจในการทำการทดลองคือ จังหวัดเชียงใหม่ มาใช้ในการวิเคราะห์ หาปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลงของค่าฝุ่นละอองที่มีขนาดเล็กกว่า 2.5 ไมครอนและใช้ข้อมูลดังกล่าวในการทำนายค่าฝุ่นละอองที่มีขนาดเล็กกว่า 2.5 ไมครอนในอีกสามชั่วโมงข้างหน้าสำหรับการทำนายนั้นจะใช้แบบจำลองทางคณิตศาสตร์ดังนี้ การถดถอยพหุคูณ (Multiple Linear Regression), แรนดอมฟอร์เรส (Random Forest), เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (Extreme Gradient Boosting) และโครงข่ายประสาทเทียม (Artificial Neural Network) จะทำการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองเพื่อหาแบบจำลองที่ดีที่สุดจากนั้นนำแบบจำลองที่ดีที่สุดไปใช้ในการทำนายค่าฝุ่นละอองที่มีขนาดเล็กกว่า 2.5 ไมครอน โดยได้ทำแบบจำลองที่สามารถทำนายค่าฝุ่นละอองในอีก 3 ชั่วโมงข้างหน้าเพื่อให้ประชาชนและหน่วยงานต่างๆ สามารถเตรียมแผนรับมือได้ทัน

จากการนำข้อมูลที่ผ่านการทำความสะอาดเข้าสู่แต่ละแบบจำลองผลลัพธ์ที่ได้ปรากฏว่าแบบจำลองเอ็กซ์ตรีมเกรเดียนต์บูสติ้งเป็นแบบจำลองที่ให้ประสิทธิภาพสูงสุด โดยใช้ตัววัดประสิทธิภาพดังนี้ รากที่สองของค่าเฉลี่ยความผิดพลาดกำลังสอง (Root Mean Square Error: RMSE) = 6.31769 $\mu g/m^3$, ค่าคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) = 4.1775 $\mu g/m^3$, ค่าสัมบูรณ์ของเปอร์เซ็นต์ความคลาดเคลื่อน (Mean Absolute Percentage Error: MAPE) = 18.63 % และ R^2 = 0.91318

Independent Study Title	Using Machine Learning Techniques to Predict $PM_{2.5}$ Level in Chiang Mai, Thailand
Author	Mr. Prayat Puangjaktha **Student ID** 590510559
Bachelor of Science	Computer Science
Supervisor	Lect. Dr. Prakarn Unachak

Abstract

The particle matters 2.5 (PM2.5) density in Chiang Mai is much higher than the standards set by the World Health Organization (WHO), resulting in many ill effects to Chiang Mai, including the health problems of local population and economic problems related to tourism setback. This problem has occurred for many years and is likely to continue in the future. Currently, there is no clear solution to the problem and thus the problem won't be resolved in a short time. So, the researcher can recognize the problems and realize the effect that will occur in the future. This research, used air pollution data, meteorological data and wildfire hotspots that occur at the study area, which is Chiang Mai for analysis, aim to find the factors that affect the change of PM2.5 and use this information to predict PM2.5 in the next three hours. For prediction models, we use Multiple Linear Regression, Random Forest, Extreme Gradient Boosting and Artificial Neural Network, and will compare the efficiencies among the models. Another goal is to find the best model that can be used to predict the density of PM2.5.

From using cleansed data to train each model, the result shows the Extreme Gradient Boosting is the most effective model by using performance indicators as follows 1. Root Mean Square Error (RMSE) = 6.31769 μg/m^3, Mean Absolute Error (MAE) = 4.1775 μg/m^3 and Mean Absolute Percentage Error (MAPE) = 18.63 %, R^2 = 0.91318.

สารบัญ

หน้า

กิตติกรรมประกาศ.....ก
บทคัดย่อ.....ข
Abstract.....ค
สารบัญ.....ง
สารบัญตาราง.....ฌ
สารบัญรูปภาพ.....ญ
บทที่ 1 บทนำ..... 1
1.1 หลักการและเหตุผล..... 1
1.2 วัตถุประสงค์ของโครงงาน..... 2
1.3 ประโยชน์ที่ได้รับจากการศึกษาเชิงประยุกต์..... 2
1.4 ขอบเขตของโครงงาน 3
1.4.1 ขอบเขตทางสถาปัตยกรรม..... 3
1.4.2 ขอบเขตของระบบงาน 3
1.4.3 ขอบเขตข้อมูล..... 4
1.5 แผนการดำเนินงานและระยะเวลาดำเนินงาน..... 5
บทที่ 2 ทฤษฎี เอกสาร และงานวิจัยที่เกี่ยวข้อง..... 6
2.1 Particulate Matter 2.5 ($PM_{2.5}$)..... 6
2.2 ดัชนีคุณภาพอากาศ (Air Quality Index: AQI)..... 6
2.3 การคำนวณดัชนีคุณภาพอากาศรายวันของสารมลพิษทางอากาศแต่ละประเภท..... 8
2.4 ค่ามาตรฐานฝุ่นละอองขนาดเล็ก 9
2.5 แหล่งที่มาของ $PM_{2.5}$..... 10
2.6 สารมลพิษทางอากาศ 6 ชนิด 10
2.6.1 ฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน ($PM_{2.5}$)..... 10
2.6.2 ฝุ่นละอองขนาดไม่เกิน 10 ไมครอน (PM_{10})..... 10

สารบัญ(ต่อ)

หน้า

2.6.3 ก๊าซโอโซน (O_3) 10

2.6.4 ก๊าซคาร์บอนมอนอกไซด์ (CO) 11

2.6.5 ก๊าซไนโตรเจนไดออกไซด์ (NO_2) 11

2.6.6 ก๊าซซัลเฟอร์ไดออกไซด์ (SO_2) 11

2.7 ปรากฏการณ์อุณหภูมิผกผัน (Temperature Inversion) 11

2.8 ค่าสหสัมพันธ์ (Correlation) 12

2.9 สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) 13

2.10 การเรียนรู้ของเครื่อง (Machine Learning) 14

2.11 โครงข่ายประสาทเทียม (Artificial Neural Network) 14

2.12 การเรียนรู้เชิงลึก (Deep Learning) 15

2.13 Loss Function, Cost Function, Error Function 16

2.14 Underfitting และ Overfitting 17

2.15 Gradient Descent 18

2.15.1 Stochastic Gradient Descent (SGD) 18

2.15.2 Mini Batch Gradient Descent 18

2.15.3 Momentum 18

2.15.4 Adagrad 19

2.15.5 AdaDelta 19

2.15.6 Adaptive Moment Estimation (Adam) 19

2.16 ฟังก์ชันกระตุ้น (Activation Function) 19

2.16.1 Sigmoid Function 20

2.16.2 Hyperbolic Tangent Function (Tanh Function) 20

2.16.3 Rectified Linear Unit (ReLU Function) 20

สารบัญ(ต่อ)

หน้า

2.16.4 Leaky ReLU Function 20

2.17 Batch Size, Iterations, Epoch 21

2.18 การป้องกันไม่ให้เกิด Overfitting 21

2.18.1 Cross-Validation 21

2.18.2 เพิ่มจำนวนข้อมูลฝึกฝนให้มากขึ้น 21

2.18.3 การลบคุณสมบัติที่ไม่จำเป็น (Remove Features) 21

2.18.4 การหยุดฝึกฝนแบบจำลองก่อนการเกิด Overfitting (Early Stopping) 22

2.18.5 Regularization 22

2.18.6 Ensembling 23

2.19 Bootstrap Aggregating (Bagging) และ Boosting 23

2.20 Extreme Gradient Boosting และ Random Forest 24

2.21 Training Set, Validation Set และ Test Set 25

2.22 Python 25

2.23 TensorFlow 26

2.24 VIIRS I-Band 375 m Active Fire Data 26

2.25 Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. 28

2.26 A Bayesian Downscaler Model to Estimate Daily $PM_{2.5}$ levels in the Continental Us 29

2.27 Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC Satellite-Based AOD Data 31

2.28 Estimating Ground-Level PM2.5 in China using Satellite Remote sensing 34

2.29 Estimating ground-level $PM_{2.5}$ fusing Satellite and station observations: A geo-intelligent deep learning approach 36

สารบัญ(ต่อ)

หน้า

2.30 $PM_{2.5}$ Prediction Based on Random Forest,XGBoost, and Deep Learning using Multisource Remote Sensing Data 38

2.31 Estimation of Daily PM10 and PM2.5 Concentation in Italy, 2013-2015 use Spatiotemporal land use Random Forest model 39

บทที่ 3 วิธีการดำเนินงานวิจัย 43

3.1 การหาหรือการเก็บข้อมูลที่เกี่ยวข้องกับ $PM_{2.5}$ (Data Gathering) 43

3.1.1 ข้อมูลจากกรมควบคุมมลพิษ (Pollution Control Department: PCD) 43

3.1.2 ข้อมูลอุตุนิยมวิทยาจากเว็บไซต์ www.Wunderground.com 45

3.1.3 ข้อมูลจุดความร้อนจาก Fire Information of Resource Management System 46

3.2 การสำรวจลักษณะของข้อมูล (Data Visualization) 46

3.3 การทำความสะอาดข้อมูล (Cleaning Data) 47

3.3.1 การกำจัดข้อมูลรบกวน (Noisy Data) 47

3.3.2 การจัดการกับค่าที่หายไป (Missing Value) 47

3.4 การเติมแต่งข้อมูลและการปรับแต่งข้อมูล (Data Engineering) 49

3.4.1 การเลือกคุณสมบัติที่เหมาะสม (Feature Selection) 49

3.4.2 การแปลงหน่วยให้เหมาะสม 49

3.4.3 การลดขนาดข้อมูลจากรายชั่วโมงเป็นรายสามชั่วโมง 49

3.4.4 การเพิ่มคุณสมบัติแก่ชุดข้อมูล 50

3.4.5 การสร้างข้อมูลในอดีต (Lag Features) 51

3.4.6 การรวมข้อมูลจากแหล่งข้อมูลทั้งหมดเข้าด้วยกัน 51

3.5 การฝึกฝนแบบจำลอง (Training Models) 52

3.6 การปรับแต่งพารามิเตอร์ (Hyper Parameter Tuning) 52

3.6.1 Random Forest 52

3.6.2 Extreme Gradient Boosting 53

สารบัญ(ต่อ)

หน้า

3.6.3 Neural Network 53

3.7 การทำนายค่า $PM_{2.5}$ ในอีกสามชั่วโมงข้างหน้า (Prediction Next 3 Hour $PM_{2.5}$ Values) 54

บทที่ 4 ผลการดำเนินงานและผลการวิเคราะห์ข้อมูล 55

4.1 การวิเคราะห์หาสาเหตุของการเกิดปัญหา PM2.5 ด้วยวิธี Data Visualization 55

4.1.1 การวิเคราะห์ $PM_{2.5}$ กับเวลา 55

4.1.2 การวิเคราะห์จากแหล่งข้อมูลกรมควบคุมมลพิษ 62

4.1.3 การวิเคราะห์จากแหล่งข้อมูล Wunderground 63

4.1.4 การวิเคราะห์จากแหล่งข้อมูล FIRMS 66

4.2 การวิเคราะห์หาสาเหตุของการเกิดปัญหา $PM_{2.5}$ ด้วยวิธีการเรียนรู้ของเครื่อง 71

4.3 การหาชุดข้อมูลที่เหมาะสมสำหรับการทำนายผล 74

4.3.1 Multiple Linear Regression 74

4.3.2 Extreme Gradient Boosting 76

4.3.3 Random Forest 78

4.3.4 MultiPerceptron Neural Network (Sklearn) 80

4.3.5 Artificial Neural Network (Keras) 82

4.4 การหาค่าพารามิเตอร์ที่เหมาะสมของแต่ละแบบจำลอง 84

4.4.1 Random Forest 84

4.4.2 Extreme Gradient Boosting 85

4.4.3 Multi-Perceptron Neural Network (Sklearn) 86

4.4.4 Artificial Neural Network (Keras) 87

4.5 การทำนายผลค่าปริมาณฝุ่นละอองที่น้อยกว่า 2.5 ไมครอน หรือ $PM_{2.5}$ ในสามชั่วโมงข้างหน้า 94

สารบัญ(ต่อ)

หน้า

บทที่ 5 สรุปผลงานวิจัย 96

5.1 สรุปผลการทดลอง 96

5.2 ปัญหาและอุปสรรค 98

5.3 ข้อเสนอแนะและแนวทางในการพัฒนาในอนาคต 99

เอกสารอ้างอิง 100

สารบัญตาราง

หน้า

ตารางที่ 1.1 การดำเนินงานและระยะเวลาการดำเนินงาน 5
ตารางที่ 2.1 เกณฑ์ของดัชนีคุณภาพอากาศสำหรับประเทศไทย 7
ตารางที่ 2.2 ค่าความเข้มข้นของสารมลพิษทางอากาศที่เทียบเท่ากับค่าดัชนีคุณภาพอากาศ 9
ตารางที่ 2.3 เปรียบเทียบค่ามาตรฐานของ $PM_{2.5}$ และ PM_{10} ระหว่างประเทศไทยและองค์กรอนามัยโลก 9
ตารางที่ 2.4 การประมาณการปล่อยมลพิษทางอากาศจากแหล่งกำเนิดๆ (ตันต่อปี) 10
ตารางที่ 2.5 การพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ 13
ตารางที่ 2.6 Loss Function สำหรับ Regression Problem 16
ตารางที่ 2.7 แสดงความหมายของข้อมูล FIRMS 27
ตารางที่ 2.8 สรุปตัวแปรที่ใช้ในแต่ละงานวิจัย 41
ตารางที่ 4.1 ผลลัพธ์จากการฝึกฝนแบบจำลอง Multiple Linear Regression 74
ตารางที่ 4.2 ผลลัพธ์การฝึกฝนแบบจำลองด้วย Extreme Gradient Boosting 76
ตารางที่ 4.3 ผลลัพธ์จากการฝึกฝนแบบจำลองแบบ Random Forest 78
ตารางที่ 4.4 ผลลัพธ์จากการฝึกฝนแบบจำลองแบบ MultiPerceptron Neural Network จาก Sklearn 80
ตารางที่ 4.5 ผลลัพธ์การฝึกฝนของแบบจำลอง Artificial Neural Network จาก Keras 82
ตารางที่ 4.6 ผลลัพธ์การปรับแต่งพารามิเตอร์ของแบบจำลอง Random Forest 84
ตารางที่ 4.7 ผลการปรับแต่งพารามิเตอร์ของแบบจำลอง Extreme Gradient Boosting 85
ตารางที่ 4.8 ผลการปรับแต่งพารามิเตอร์ของ Multi-Perceptron Neural Network จาก Sklearn 86
ตารางที่ 4.9 ผลลัพธ์ของการปรับแต่งจำนวนโหนดใน Artificial Neural Network จาก Keras 87
ตารางที่ 4.10 ผลลัพธ์การปรับแต่ง Dropout Rate ของ Artificial Neural Network จาก Keras 88
ตารางที่ 4.11 ผลการปรับแต่ง Regularization Rate ของ Artificial Neural Network จาก Keras 88
ตารางที่ 4.12 ผลการปรับแต่ง Batch Size ของ Artificial Neural Network จาก Keras 89

สารบัญตาราง(ต่อ)

หน้า

ตารางที่ 4.13 ผลการปรับแต่ง Activation Function ของ Artificial Neural Network จาก Keras 90

ตารางที่ 4.14 ผลการปรับแต่ง Optimizer ของ Artificial Neural Network จาก Keras..... 90

ตารางที่ 4.15 ผลการปรับแต่ง Learning Rate ของ Artificial Neural Network จาก Keras..... 91

ตารางที่ 4.16 พารามิเตอร์แต่ละตัวที่ทำให้แบบจำลอง Artificial Neural Network (Keras) มีประสิทธิภาพสูงสุด..... 91

ตารางที่ 4.17 การเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองหลังจากผ่าน การปรับแต่งพารามิเตอร์ 93

ตารางที่ 5.1 การเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองหลังจากผ่าน การปรับแต่งพารามิเตอร์ 97

สารบัญรูปภาพ

หน้า

ภาพที่ 2.1 ปรากฏการณ์อุณหภูมิผกผัน 12
ภาพที่ 2.2 เซลล์ประสาทของสมองมนุษย์ 14
ภาพที่ 2.3 แบบจำลองโครงข่ายประสาทเทียมอย่างง่าย 15
ภาพที่ 2.4 โครงข่ายประสาทเทียมแบบการเรียนรู้เชิงลึก 16
ภาพที่ 2.5 การเกิด Underfitting และ Overfitting 17
ภาพที่ 2.6 ตัวอย่างฟังก์ชันกระตุ้นแต่ละประเภท 20
ภาพที่ 2.7 การวัดประสิทธิภาพของการฝึกฝนแบบจำลอง 22
ภาพที่ 2.8 ไดอะแกรมการทำงานของ Boosting และ Bagging 24
ภาพที่ 2.9 โครงสร้างของโครงข่ายประสาทเทียมสำหรับการประมาณค่า $PM_{2.5}$ 28
ภาพที่ 2.10 พื้นที่ของประเทศสหรัฐอเมริกาที่ถูกแบ่งเป็น 9 เขต 29
ภาพที่ 2.11 โครงสร้าง Geoi-DBN 36
ภาพที่ 3.1 ข้อมูลจากกรมควบคุมมลพิษ ณ ศูนย์ราชการ จังหวัดเชียงใหม่ 45
ภาพที่ 3.2 ภาพรวมของกระบวนการทำงาน 54
ภาพที่ 4.1 ข้อมูล $PM_{2.5}$ ทั้งหมดแสดงในลักษณะของ Time Series ณ โรงเรียนยุพราช จังหวัดเชียงใหม่ 56
ภาพที่ 4.2 ข้อมูล $PM_{2.5}$ ทั้งหมดแสดงในลักษณะของ Time Series ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 56
ภาพที่ 4.3 ข้อมูล $PM_{2.5}$ ในรูปแบบรายวัน ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 57
ภาพที่ 4.4 ข้อมูล $PM_{2.5}$ ในรูปแบบรายวัน ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่ 57
ภาพที่ 4.5 ข้อมูล $PM_{2.5}$ ในแต่ละฤดูกาล ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่ 58
ภาพที่ 4.6 ข้อมูล $PM_{2.5}$ ในแต่ละฤดูกาล ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 58
ภาพที่ 4.7 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในแต่ละปี ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 59
ภาพที่ 4.8 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในปี ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่ 59
ภาพที่ 4.9 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในแต่ละฤดูกาล ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 60
ภาพที่ 4.10 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในแต่ละฤดูกาล ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่ 60

สารบัญรูปภาพ (ต่อ)

หน้า

ภาพที่ 4.11 $PM_{2.5}$ ในแต่ละชั่วโมงของวันในสัปดาห์ ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 61
ภาพที่ 4.12 $PM_{2.5}$ ในแต่ละชั่วโมงของวันในสัปดาห์ ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่ 61
ภาพที่ 4.13 แผนภาพการกระจายของข้อมูลที่เข้าคู่กันระหว่าง $PM_{2.5}$ กับตัวแปรต่างๆ ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 62
ภาพที่ 4.14 แผนภาพการกระจายของข้อมูลที่เข้าคู่กันระหว่าง $PM_{2.5}$ กับตัวแปรต่างๆ ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่ 62
ภาพที่ 4.15 แผนภาพการกระจายของอุณหภูมิ ความไวลม ความชื้นที่เข้าคู่กับ $PM_{2.5}$ 64
ภาพที่ 4.16 แผนภาพการกระจายของความกดอากาศและจุดไอน้ำกลั่นตัวที่เข้าคู่กับ $PM_{2.5}$ 65
ภาพที่ 4.17 เงื่อนไขสภาพอากาศในแต่ละวันกับ $PM_{2.5}$ 65
ภาพที่ 4.18 การกระจายของค่า $PM_{2.5}$ ในวันที่ฝนตก 66
ภาพที่ 4.19 การเปรียบเทียบจำนวนจุดความร้อนในระยะ 144 กิโลเมตรกับ $PM_{2.5}$ 67
ภาพที่ 4.20 การเปรียบเทียบจำนวนจุดความร้อนในระยะ 288 กิโลเมตรกับ $PM_{2.5}$ 67
ภาพที่ 4.21 การเปรียบเทียบจุดความร้อนในระยะ 432 กิโลเมตรกับ $PM_{2.5}$ 68
ภาพที่ 4.22 การเปรียบเทียบจุดความร้อนในระยะมากกว่า 432 กิโลเมตรแต่น้อยกว่า 3000 กิโลเมตร กับ $PM_{2.5}$ 68
ภาพที่ 4.23 การเปรียบเทียบพลังงานที่แผ่จากจุดความร้อนในระยะ 144 กิโลเมตรกับ $PM_{2.5}$ 69
ภาพที่ 4.24 การเปรียบเทียบพลังงานที่แผ่จากจุดความร้อนในระยะ 288 กิโลเมตรกับ $PM_{2.5}$ 69
ภาพที่ 4.25 การเปรียบเทียบพลังงานที่แผ่จากจุดความร้อนในระยะ 432 กิโลเมตรกับ $PM_{2.5}$ 70
ภาพที่ 4.26 การเปรียบเทียบพลังงานที่แผ่ออกจากจุดความร้อนระยะมากกว่า 432 กิโลเมตรแต่น้อยกว่า 3000 กิโลเมตรกับ $PM_{2.5}$ 70
ภาพที่ 4.27 ความสำคัญของตัวแปรอิสระในชุดข้อมูลแบบวนซ้ำ (Cyclical Dataset) จากกรมควบคุมมลพิษ 72
ภาพที่ 4.28 ความสำคัญของตัวแปรอิสระจากชุดข้อมูลแบบดัมมี่ (Dummy Dataset) จากกรมควบคุมมลพิษ 73
ภาพที่ 4.29 การทำนายค่า PM2.5 ณ วันที่ 2020-04-13 ในอีกสามชั่วโมงข้างหน้า ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ 94

สารบัญรูปภาพ (ต่อ)

หน้า

ภาพที่ 4.30 การทำนายค่า $PM_{2.5}$ ณ วันที่ 2020-04-13 ในอีกสามชั่วโมงข้างหน้า
ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่..... 95

บทที่ 1

บทนำ

ในปัจจุบันสิ่งแวดล้อมที่มนุษย์นั้นอาศัยอยู่นั้นได้เปลี่ยนแปลงไปอย่างมากหากเทียบกับ 5 ถึง 10 ปีที่ผ่านมา โดยการเปลี่ยนแปลงที่กล่าวถึงนั้นในจังหวัดเชียงใหม่เองก็มีการเปลี่ยนแปลงเกิดขึ้นเช่นกันซึ่งในงานวิจัยฉบับนี้จะกล่าวถึง การเปลี่ยนแปลงในเรื่องของปริมาณละอองฝุ่นที่น้อยกว่า 2.5 ไมครอน (Particulate Matters 2.5: $PM_{2.5}$) ในอากาศมากเกินกว่าค่ามาตรฐานซึ่งเป็นปัญหาที่เกิดขึ้นมาอย่างยาวนานในจังหวัดเชียงใหม่และเขตภาคเหนือของประเทศไทยรวมถึงปัญหาดังกล่าวยังเป็นปัญหาที่ไม่สามารถหาแนวทางในการแก้ไขได้อย่างชัดเจนและไม่สามารถแก้ไขได้ในระยะเวลาสั้น ทำให้ผู้คนที่อาศัยในพื้นที่บริเวณที่มีปัญหาดังกล่าวนั้นได้รับผลกระทบไม่ว่าจะเป็นทางตรงหรือทางอ้อม ด้วยเหตุผลดังกล่าวจึงได้เกิดงานวิจัยฉบับนี้ขึ้นโดยได้นำเทคนิคการเรียนรู้ของเครื่องเข้ามาช่วยในการวิเคราะห์ถึงปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลงของปริมาณฝุ่นที่น้อยกว่า 2.5 ไมครอนและใช้ในการทำนายค่าปริมาณฝุ่นที่น้อยกว่า 2.5 ไมครอนในอนาคต

1.1 หลักการและเหตุผล

ปัญหาปริมาณฝุ่นที่น้อยกว่า 2.5 ไมครอน หรือ $PM_{2.5}$ สูงกว่ามาตรฐานนั้นเป็นปัญหาเรื้อรังที่อยู่คู่กับประเทศไทยในบริเวณภาคเหนือมาอย่างยาวนานซึ่งสาเหตุที่ทำให้เกิดปัญหานี้มีหลายสาเหตุด้วยกันยกตัวอย่างเช่น การเผาป่า การเผาป่าในที่นี้รวมถึงการเผาป่าของประเทศเพื่อนบ้าน ควันไอเสียจากยานพาหนะที่มีจำนวนเพิ่มมากขึ้น การเกิดอุณหภูมิผกผัน การเพิ่มขึ้นของประชากรซึ่งส่งผลให้เกิดการบริโภคสิ่งต่างๆ เพิ่มมากขึ้นเมื่อความต้องการเพิ่มมากขึ้นส่งผลให้อุตสาหกรรมการผลิตเพิ่มมากขึ้นตามกฎอุปสงค์อุปทานทำให้เกิดโรงงานอุตสากรรมเพิ่มมากขึ้นตามไปด้วย และลักษณะทางภูมิประเทศ เช่น เชียงใหม่มีลักษณะภูมิประเทศที่มีภูเขารอบล้อมทำให้เกิดเป็นแอ่งขึ้นหรือที่ชาวบ้านเรียกกันว่า “ก้นกระทะ” ทำให้หมอกควันต่างๆ ที่อยู่ในจังหวัดเชียงใหม่ไม่สามารถระบายเคลื่อนย้ายไปที่อื่นได้ ส่งผลให้จังหวัดเชียงใหม่ถือเป็นอันดับต้นๆ ที่ต้องเจอกับปัญหา $PM_{2.5}$ สูงเกินกว่ามาตรฐานโดยในเดือนเมษายนปี 2562 $PM_{2.5}$ ที่ถูกวัดได้สูงสุดในจังหวัดเชียงใหม่สูงถึง 500 $\mu g/m^3$ ซึ่งสูงเป็นอันดับ 1 ของโลกโดยผลกระทบจากปัญหาดังกล่าวนั้นส่งผลกระทบต่อประชากรที่อาศัยอยู่ในพื้นที่ที่มี $PM_{2.5}$ สูงเกินกว่าค่ามาตรฐานทั้งทางตรงและทางอ้อม ทั้งนี้ทางตรงที่กล่าวถึงหมายถึงปัญหาเกี่ยวกับสุขภาพของประชากรโดย $PM_{2.5}$ นั้นส่งผลให้ประชากรในพื้นที่มีความเสี่ยงในการป่วยเป็นโรคเกี่ยวกับระบบทางเดินหายใจเพิ่มขึ้น ส่วนผลกระทบทางอ้อมหมายถึงผลกระทบทางด้านเศรษฐกิจไม่ว่าจะเป็นเกี่ยวกับการท่องเที่ยว ฯลฯ เป็นต้น

จากผลกระทบดังกล่าวส่งผลให้มีองค์กรทางการศึกษาได้ให้ความสนใจ เริ่มมีการทำสำรวจและติดตั้งเครื่องตรวจวัดปริมาณ $PM_{2.5}$ ที่เกิดขึ้น ณ เวลานั้นๆ ตามสถานที่ต่างๆ ในจังหวัดเชียงใหม่โดยแสดงข้อมูลผ่านเว็บไซต์ ยกตัวอย่างเช่น www.cmuccdc.org, www.aqmthai.com/public_report.php ทั้งสองเว็บไซต์เป็นแหล่งให้ข้อมูลเกี่ยวกับปริมาณของคุณภาพอากาศที่เกิดขึ้น ณ เวลานั้นๆ โดยมีจุดตรวจวัดที่ถูกติดตั้งไว้ตามสถานที่ต่างๆ เมื่อประชากรทราบข้อมูลก็ทำให้รู้ว่าควรจะปฏิบัติตนเช่นไร เช่น หาก $PM_{2.5}$ มีปริมาณสูงกว่ามาตรฐานก็ควรจะอยู่ในสถานที่ที่มีเครื่องกรองอากาศ หรืองดกิจกรรมกลางแจ้ง เป็นต้น สำหรับปัญหา $PM_{2.5}$ นั้นถือเป็นปัญหาที่เกิดขึ้นทั่วโลกพร้อมทั้งมีนักวิจัยมากมายที่พยายามหาข้อเท็จจริงของปัญหา $PM_{2.5}$ ในพื้นที่ต่างๆ โดยใช้แบบจำลองทางสถิติมากมาย เช่น โครงข่ายประสาทเทียม (Neural Network), การถดถอยพหุคูณ (Multiple Linear Regression), แบบจำลองถดถอยแบบถ่วงน้ำหนักภูมิศาสตร์ (Geographically Weighted Regression) ฯลฯ โดยมีน้อยงานวิจัยที่สามารถบอกได้อย่างแน่ชัดว่าปัญหา $PM_{2.5}$ ได้รับอิทธิพลมาจากปัจจัยใด ซึ่งในแต่ละพื้นที่จะมีปัจจัยที่ส่งผลนั้นแตกต่างกันออกไปด้วยเหตุผลดังกล่าว ผู้วิจัยจึงต้องการวิเคราะห์หาสาเหตุปัจจัยที่ส่งผลกระทบต่อการเกิดปัญหา $PM_{2.5}$ สูงกว่ามาตรฐานและทำการทำนายค่า $PM_{2.5}$ ที่จะเกิดขึ้นในอนาคตระยะสั้น โดยใช้การเรียนรู้ของเครื่องในการแก้ปัญหาและจำกัดขอบเขตของงานในพื้นที่จังหวัดเชียงใหม่เท่านั้น เพื่อให้ประชาชน สถานศึกษาและองค์กรต่างๆ สามารถนำข้อมูลส่วนนี้ไปใช้ในการวางแผนในการทำงานหรือการดำเนินชีวิตประจำวันต่อไป

1.2 วัตถุประสงค์ของโครงงาน

1) เพื่อหาปัจจัยที่ส่งผลกระทบต่อการเกิดปัญหาปริมาณละอองฝุ่นที่น้อยกว่า 2.5 ไมครอนสูงกว่ามาตรฐานในจังหวัดเชียงใหม่

2) เพื่อพัฒนาแบบจำลองที่ใช้ในการทำนายปริมาณละอองฝุ่นที่น้อยกว่า 2.5 ไมครอน ในอีก 3 ชั่วโมงข้างหน้า ให้มีความถูกต้องและแม่นยำมากที่สุด

1.3 ประโยชน์ที่ได้รับจากการศึกษาเชิงประยุกต์

1) เป็นการประยุกต์ใช้ความรู้ในการดึงสารสนเทศจากข้อมูลดาวเทียมที่มีอยู่แล้วมาใช้ให้เกิดประโยชน์
2) เป็นเครื่องมือที่ช่วยคาดการณ์ปริมาณฝุ่นละอองที่น้อยกว่า 2.5 ไมครอน ให้แก่ประชาชนและองค์กรต่างๆ ในระยะเวลาสั้นๆ เพื่อให้สามารถเตรียมการป้องกันได้อย่างทันท่วงที
3) สามารถประมาณค่าปริมาณละอองฝุ่นที่น้อยกว่า 2.5 ไมครอน ได้ในพื้นที่กว้างทำให้สามารถลดอุปกรณ์ตรวจวัดลงได้

1.4 ขอบเขตของโครงงาน

1.4.1 ขอบเขตทางสถาปัตยกรรม

ระบบที่ได้ทำการพัฒนาขึ้นมานี้เป็นระบบสแตนอะโลน (Standalone)

1) ฮาร์ดแวร์ (Hardware) ที่ใช้ในการพัฒนาระบบ ประกอบด้วย
   - คอมพิวเตอร์มีหน่วยประมวลผล (CPU) Intel-Core i7-2600 4 Core
   - หน่วยความจำเข้าถึงแบบสุ่ม (Random Access Memory) 8 Gigabyte
   - จานบันทึกแบบเข็ง (Hard Disk) ขนาดความจุ 200 Gigabyte
2) ซอฟต์แวร์ (Software) ที่ใช้พัฒนาแบบจำลอง ประกอบด้วย
   - ระบบปฏิบัติการไมโครซอฟท์วินโดว์ 10 เอนเทอไพรส์ (Microsoft Windows 10 Enterprise)
   - Python 3.7.4 (ภาษาสำหรับพัฒนาโปรแกรมที่ใช้ในงานวิจัย)
   - TensorFlow 2.1.0 (ไลบรารีสำหรับสร้างแบบจำลองทางสถิติ)
   - Scikit-learn 0.22.2 (ไลบรารีสำหรับสร้างแบบจำลองทางสถิติ)
   - Matplotlib 3.1.3 (ไลบรารีสำหรับการแสดงผลในรูปของกราฟ)
   - Seaborn 0.10.0 (ไลบรารีสำหรับการแสดงผลในรูปของกราฟ)

1.4.2 ขอบเขตของระบบงาน

ลักษณะการทำงานหลักแบ่งออกเป็นดังนี้

1) การดึงข้อมูลจากแหล่งออนไลน์

   ในการทำนายค่าฝุ่น $PM_{2.5}$ นั้นมีปัจจัยที่เกี่ยวข้องอยู่มากมาย เช่น สภาพอากาศ การเผาไหม้ที่เกิดขึ้นในแต่ละพื้นที่ สารมลพิษทางอากาศอื่นๆ จากปัจจัยที่กล่าวมาข้างต้นเป็นเพียงปัจจัยบางส่วนที่ส่งผลต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ เนื่องจากข้อมูลดังกล่าวในปัจจุบันสามารถหาได้ผ่านเว็บไซต์ออนไลน์ โดยผู้วิจัยจะต้องทำการสืบหาข้อมูลเกี่ยวกับปัจจัยต่างๆ ที่เกี่ยวข้องและหาแหล่งข้อมูลเหล่านั้นมาใช้ในการคำนวณทำนายค่า $PM_{2.5}$

2) การเตรียมข้อมูลก่อนนำข้อมูลเข้าสู่แบบจำลอง

   เนื่องจากปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลง $PM_{2.5}$ นั้นมีหลายส่วนด้วยกันทำให้มีข้อมูลที่ต้องใช้จากแหล่งข้อมูลหลายแหล่งข้อมูล ซึ่งข้อมูลจากแต่ละแหล่งข้อมูลนั้นมีมาตรฐานในการเก็บข้อมูลที่แตกต่างกันไป จึงต้องทำการปรับแต่งข้อมูลทั้งหมดให้อยู่ในมาตรฐานเดียวกันพร้อมทั้งเป็นการแก้ไขค่าต่างๆ ที่ไม่เหมาะสมโดยวิธีในการปรับแต่งนั้นก็มีหลายวิธีเช่นเดียวกัน ยกตัวอย่างเช่น การแทนค่าที่ว่างเปล่าด้วยค่าเฉลี่ยของคุณสมบัตินั้นๆ เป็นต้น ก่อนที่จะนำข้อมูลเหล่านี้ไปคำนวณกับตัวแบบจำลองที่ผู้วิจัยได้ออกแบบไว้

3) การฝึกฝนแบบจำลอง

ในการฝึกฝนแบบจำลองนั้นผู้วิจัยต้องทำการเลือกแบบจำลองที่จะใช้ในการฝึกฝนข้อมูล ซึ่งแบบจำลองที่ผู้วิจัยได้เลือกไว้มี 4 แบบด้วยกัน ได้แก่ การถดถอยพหุคูณ (Multiple Linear Regression), แรนดอมฟอร์เรส (Random Forest), เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (Extreme Gradient Boosting) และแบบจำลองโครงข่ายประสาทเทียม (Neural Network) โดยจะต้องทำการแบ่งข้อมูลเป็นสองส่วนในการใช้สำหรับฝึกฝนและทดสอบแบบจำลอง เมื่อทำการฝึกฝนเสร็จแล้วจะต้องทำการเลือกแบบจำลองที่ให้ผลลัพธ์ได้ออกมาดีที่สุด ในขั้นตอนการฝึกฝนนี้จากความแม่นยำที่เกิดขึ้นจะทำให้สามารถวิเคราะห์ความสำคัญของปัจจัยที่ส่งผลกระทบต่อ $PM_{2.5}$ ได้อีกด้วย

4) การทำนายค่าฝุ่น $PM_{2.5}$

สำหรับการทำนายค่า $PM_{2.5}$ ที่เกิดขึ้นในงานวิจัยฉบับนี้นั้นจะเป็นการทำนายค่า $PM_{2.5}$ ที่จะเกิดขึ้นในอนาคตในระยะเวลาอันสั้นหรือกล่าวให้ชัดเจนก็คือจะเป็นการทำนายค่า $PM_{2.5}$ ในอีก 3 ชั่วโมงข้างหน้า โดยจะทำการคำนวณจากปัจจัยต่างๆ ที่ได้ทำการเลือกมาในขั้นตอนที่ 2 พร้อมกับนำค่าจากการทำนายไปเปรียบเทียบกับข้อมูลจริงที่เกิดขึ้น

1.4.3 ขอบเขตข้อมูล

แหล่งข้อมูล มี 3 ส่วนด้วยกันดังนี้ ส่วนแรกเป็นข้อมูลจากกรมควบคุมมลพิษ ส่วนที่สองเป็นข้อมูลสภาพอากาศจากเว็บไซต์ www.Wunderground.com และส่วนสุดท้ายเป็นข้อมูลการตรวจจับจุดความร้อน (Fire Information for Resource Management System : FIRMS) จากดาวเทียมขององค์กรนาซ่า (The National Aeronautics and Space Administration) จากเว็บไซต์ firms.modaps.eosdis.nasa.gov

1) ข้อมูลนำเข้า

ได้แก่ ส่วนของกรมควบคุมมลพิษ ได้แก่ NO, NO_X, NO_2, SO_2, Wind Direction, Temperature, Relative Humidity, $PM_{2.5}$, PM_{10}, Datetime, Season, Wind Speed, Location เป็นต้น ส่วนของ www.Wunderground.com ได้แก่ Dew Point, Pressure, Condition, Wind Speed, Wind Direction, Temperature, Relative Humidity, Pressure, Precipitation, Datetime เป็นต้น และส่วนสุดท้ายข้อมูล FIRMS ได้แก่ Latitude, Longitude, Bright_ti4, Satellite, Instrument, Confidence, Version, Bright_ti5, Frp, Type, Datetime

2) ข้อมูลออก ค่าตัวเลขที่ทำนาย $PM_{2.5}$ ในอีก 3 ชั่วโมงข้างหน้าในหน่วย $\mu g/m^3$

1.5 แผนการดำเนินงานและระยะเวลาดำเนินงาน

ระยะเวลาดำเนินงานเริ่มตั้งแต่เดือนสิงหาคม พ.ศ.2562 ถึง เดือนเมษายน พ.ศ.2563 ซึ่งมีขั้นตอนการดำเนินงานดังตารางที่ 1.1

ตารางที่ 1.1 การดำเนินงานและระยะเวลาการดำเนินงาน

ขั้นตอนการดำเนินงาน	**ระยะเวลาการดำเนินงาน**								
	พ.ศ. 2562					พ.ศ. 2563			
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย
ศึกษาทฤษฎี, ศึกษางานวิจัยเดิม, หาแหล่งข้อมูล									
วิเคราะห์แหล่งปัจจัยที่ใช้ในการทำนายค่าฝุ่น									
Pre-processing									
ออกแบบโมเดล									
พัฒนาโมเดล									
ทดสอบและแก้ไขตัวโมเดล									
สรุปผลและจัดทำเอกสาร									

บทที่ 2
ทฤษฎี เอกสาร และงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงหลักการและทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์ ออกแบบ และพัฒนาแบบจำลองสำหรับการใช้ข้อมูลจากดาวเทียมและค่า $PM_{2.5}$ ในการทำนายด้วยการเรียนรู้ของเครื่อง ซึ่งได้ทำการศึกษาหลักการและทฤษฎีที่เกี่ยวข้อง รวมไปถึงเอกสารและงานวิจัยดังนี้

2.1 Particulate Matter 2.5 ($PM_{2.5}$)

PM ย่อมาจาก Particulate Matter หมายถึงหน่วยวัดขนาดของอนุภาคที่อยู่ในอากาศ และ $PM_{2.5}$ ก็คืออนุภาคที่มีขนาดเล็กกว่า 2.5 ไมโครเมตร และหากนำไปเปรียบเทียบกับขนาดเส้นผมของมนุษย์ที่มีขนาด 50-70 ไมโครเมตร ฝุ่นขนาดเล็กที่กล่าวถึงมีขนาดเล็กกว่าขนาดเส้นผมถึง 20-30 เท่า ปกติในทางเดินหายใจของมนุษย์มีขนพัดโบก (Respiratory Cilia) และมีการหลั่งน้ำมูกเสมหะ เพื่อดักจับไม่ให้สิ่งแปลกปลอมเข้าสู่ทางเดินหายใจ แต่ฝุ่นขนาดเล็กกว่า 2.5 ไมโครเมตรจะสามารถผ่านเข้าไปได้ถึงหลอดลมฝอยขนาดเล็ก (Bronchiole) และหากมีขนาดเล็กกว่า 1 ไมโครเมตร ก็จะสามารถซึมเข้าสู่กระแสเลือดได้ โดยตัวฝุ่นเป็นพาหะนำสารอื่นเข้ามาด้วย เช่น แคดเมียม (Cd) ปรอท (Hg) โลหะหนัก และสารก่อมะเร็งอื่นๆ ดังนั้นหากสภาพอากาศมีค่า $PM_{2.5}$ เกินมาตรฐาน คนกลุ่มแรกที่ได้รับผลกระทบอย่างมากได้แก่ เด็ก ผู้สูงอายุ ผู้ที่มีโรคทางเดินหายใจ เรื้อรังโรคปอด และโรคหัวใจ เนื่องจากกระบวนการป้องกันของร่างกายที่อ่อนแอกว่าคนที่มีสุขภาพแข็งแรงส่วนในคนที่มีสุขภาพแข็งแรงนั้น จากการศึกษาพบว่าหากสัมผัสกับอากาศที่มี $PM_{2.5}$ เกินมาตรฐานในระยะเวลานานมีความสัมพันธ์กับการเพิ่มโอกาสในการเสียชีวิตด้วยโรคปอดและโรคหัวใจ (Cardiopulmonary Mortality) และเพิ่มความเสี่ยงต่อการเป็นมะเร็งปอดด้วย [1]

โดยองค์การอนามัยโลก หรือ WHO กำหนดให้ฝุ่น $PM_{2.5}$ จัดอยู่ในกลุ่มที่ 1 ของสารก่อมะเร็งประกอบกับรายงานของธนาคารโลก (World Bank) ที่ระบุว่า ประเทศไทยมีผู้เสียชีวิตจากมลพิษทางอากาศมากถึง 50,000 ราย ส่งผลไปถึงระบบเศรษฐกิจรวมไปถึงค่าใช้จ่ายที่รัฐต้องสูญเสียเกี่ยวเนื่องกับค่ารักษาพยาบาลผู้ป่วยจากมลพิษทางอากาศ

2.2 ดัชนีคุณภาพอากาศ (Air Quality Index: AQI)

ดัชนีคุณภาพอากาศ เป็นการรายงานข้อมูลคุณภาพอากาศในรูปแบบที่ง่ายต่อความเข้าใจของประชาชนทั่วไปเพื่อเผยแพร่ประชาสัมพันธ์ให้สาธารณะชนได้รับรู้ข้อมูลถึงสถานการณ์มลพิษทางอากาศในแต่ละพื้นที่ว่าอยู่ในระดับใด มีผลกระทบต่อสุขภาพหรือไม่ โดยที่ดัชนีคุณภาพอากาศเป็นการแสดงในรูปแบบสากลที่ใช้กันอย่างแพร่หลาย เช่น ประเทศสหรัฐอเมริกา ออสเตรเลีย สิงคโปร์ มาเลเซีย และประเทศไทยเป็นต้น

ดัชนีคุณภาพอากาศที่ใช้ในประเทศไทย คำนวณเทียบจากมาตรฐานคุณภาพอากาศในบรรยากาศโดยทั่วไปของสารมลพิษทางอากาศ 5 ชนิด ได้แก่ ก๊าซโอโซน (O_3) เฉลี่ย 1 ชั่วโมง ก๊าซไนโตรเจนไดออกไซด์ (NO_2) เฉลี่ย 1 ชั่วโมง ก๊าซคาร์บอนมอนอกไซด์ (CO) เฉลี่ย 8 ชั่วโมง ก๊าซซัลเฟอร์ไดออกไซด์ (SO_2) เฉลี่ย 24 ชั่วโมง และฝุ่นละอองขนาดเล็กกว่า 10 ไมครอน (PM_{10}) เฉลี่ย 24 ชั่วโมง โดยดัชนีคุณภาพอากาศที่คำนวณได้ของสารมลพิษทางอากาศประเภทใดมีค่าสูงสุด จะใช้เป็นดัชนีคุณภาพอากาศของวันนั้น

ดัชนีคุณภาพอากาศของประเทศไทย ได้แบ่งเป็น 5 ระดับ ตั้งแต่ 0 - มากกว่า 201 ซึ่งแต่ละระดับจะใช้สีเป็นสัญลักษณ์เปรียบเทียบระดับของผลกระทบต่อสุขภาพ ซึ่งดัชนีคุณภาพอากาศ 100 จะมีค่าที่เทียบเท่ากับมาตรฐานคุณภาพอากาศในบรรยากาศโดยทั่วไป หากดัชนีคุณภาพอากาศมีค่าสูงเกินกว่า 100 แสดงว่าค่าความเข้มข้นของมลพิษทางอากาศมีค่าเกินมาตรฐานและคุณภาพอากาศ ณ วันนั้นจะเริ่มมีผลกระทบต่อสุขภาพของประชาชน [2]

ตารางที่ 2.1 เกณฑ์ของดัชนีคุณภาพอากาศสำหรับประเทศไทย

AQI	ความหมาย	สีที่ใช้	แนวทางการป้องกันผลกระทบ
0 - 25	คุณภาพดีมาก	ฟ้า	คุณภาพอากาศดีมาก เหมาะสำหรับทำกิจกรรมกลางแจ้งและท่องเที่ยว
26 - 50	คุณภาพดี	เขียว	คุณภาพอากาศดี สามารถทำกิจกรรมกลางแจ้งและท่องเที่ยวได้ปกติ
51 -100	ปานกลาง	เหลือง	**ประชาชนทั่วไป:** สามารถทำกิจกรรมกลางแจ้งได้ตามปกติ **ผู้ที่ต้องดูแลสุขภาพพิเศษ:** หากมีอาการเบื้องต้นเช่น ไอ หายใจลำบาก ระคายเคืองตา ควรลดระยะเวลาการทำกิจกรรมกลางแจ้ง

ตารางที่ 2.1 เกณฑ์ของดัชนีคุณภาพอากาศสำหรับประเทศไทย (ต่อ)

AQI	ความหมาย	สีที่ใช้	แนวทางการป้องกันผลกระทบ
101-200	เริ่มมีผลกระทบต่อสุขภาพ	ส้ม	**ประชาชนทั่วไป:** ควรเฝ้าระวังสุขภาพ ถ้ามีอาการเบื้องต้น เช่น ไอ หายใจลำบาก ระคายเคืองตา ควรลดระยะเวลาการทำกิจกรรมกลางแจ้ง หรือใช้อุปกรณ์ป้องกันตนเองหากมีความจำเป็น **ผู้ที่ต้องดูแลสุขภาพเป็นพิเศษ:** ควรลดระยะเวลาการทำกิจกรรมกลางแจ้งหรือใช้อุปกรณ์ป้องกันตนเองหากมีความจำเป็น ถ้ามีอาการทางสุขภาพควรปรึกษาแพทย์
มากกว่า 201	มีผลกระทบต่อสุขภาพ	แดง	ทุกคนควรหลีกเลี่ยงกิจกรรมกลางแจ้งหลีกเลี่ยงพื้นที่ที่มีมลพิษทางอากาศสูงหรือใช้อุปกรณ์ป้องกันตนเองหากมีความจำเป็น หากมีอาการทางสุขภาพควรปรึกษาแพทย์

2.3 การคำนวณดัชนีคุณภาพอากาศรายวันของสารมลพิษทางอากาศแต่ละประเภท

การคำนวณจากค่าความเข้มข้นของสารมลพิษทางอากาศแต่ละชนิดจากข้อมูลผลการตรวจวัดคุณภาพอากาศโดยมีระดับของค่าความเข้มข้นของสารมลพิษทางอากาศที่เทียบเท่ากับค่าดัชนีคุณภาพอากาศที่ระดับต่างๆ ดังตารางที่ 2.2 การคำนวณดัชนีคุณภาพอากาศภายในช่วงระดับเป็นสมการเส้นตรงดังนี้ [2]

$$\mathrm{I_i} = \frac{\mathrm{I_{ij+1}} - \mathrm{I_{ij}}}{\mathrm{X_{ij+1}} - \mathrm{X_{ij}}}\left(\mathrm{X_i} - \mathrm{X_{ij}}\right) + \mathrm{I_{ij}}$$

กำหนดให้

$\mathrm{X_i} =$ ความเข้มข้นของสารมลพิษทางอากาศจากผลการตรวจวัด

$\mathrm{X_{ij}} =$ ความเข้มข้นของสารมลพิษทางอากาศที่เป็นค่าต่ำสุดของช่วงพิสัยที่มีค่า $\mathrm{X_i}$ นั้น

$\mathrm{X_{ij+1}} =$ ความเข้มข้นของสารมลพิษทางอากาศที่เป็นค่าสูงสุดของช่วงพิสัยที่มีค่า $\mathrm{X_i}$

$\mathrm{I_i} =$ ค่าดัชนีย่อยคุณภาพอากาศ

$\mathrm{I_{ij}} =$ ค่าดัชนีย่อยคุณภาพอากาศที่เป็นค่าต่ำสุดของช่วงพิสัยที่มีค่า $\mathrm{I_i}$ นั้น

$\mathrm{I_{ij+1}} =$ ค่าดัชนีย่อยคุณภาพอากาศที่เป็นค่าสูงสุดของช่วงพิสัยที่มีค่า I_i นั้น

ตารางที่ 2.2 ค่าความเข้มข้นของสารมลพิษทางอากาศที่เทียบเท่ากับค่าดัชนีคุณภาพอากาศ

AQI	$PM_{2.5}$ (มคก./ลบ.ม.)	PM_{10} (มคก./ลบ.ม.)	O_3 (ppb)	CO (ppm)	NO_2 (ppb)	SO_2 (ppb)
	เฉลี่ย 24 ชั่วโมงต่อเนื่อง		**เฉลี่ย 8 ชั่วโมงต่อเนื่อง**		**เฉลี่ย 1 ชั่วโมง**	
0 - 25	0 - 25	0 - 50	0 - 35	0 - 4.4	0 - 60	0 - 100
26 - 50	26 - 37	51 - 80	36 - 50	4.5 - 6.4	61 - 106	101 - 200
51 - 100	38 - 50	81 - 120	51 - 70	6.5 - 9.0	107 - 170	201 - 300
101 - 200	51 - 90	121 - 180	71 - 120	9.1 - 30.0	171 - 340	301 - 400
มากกว่า 200	91 ขึ้นไป	181 ขึ้นไป	121 ขึ้นไป	30.1 ขึ้นไป	341 ขึ้นไป	401 ขึ้นไป

2.4 ค่ามาตรฐานฝุ่นละอองขนาดเล็ก

มาตรฐานคุณภาพอากาศของประเทศไทยถือว่ามีช่องว่าง เมื่อเทียบกับข้อแนะนำขององค์กรอนามัยโลก ค่ามาตรฐานรายปีของ $PM_{2.5}$ อยู่ที่ 25 ไมโครกรัมต่อลูกบาศก์เมตร สูงกว่าค่ามาตรฐานขององค์การอนามัยโลก 2.5 เท่า ค่าเฉลี่ย 24 ชั่วโมงอยู่ที่ 50 ไมโครกรัมต่อลูกบาศก์เมตรซึ่งสูงกว่า 2 เท่าเมื่อเทียบกับมาตรฐานขององค์การอนามัยโลก

ส่วนค่ามาตรฐานรายปีของ PM_{10} ของประเทศไทยอยู่ที่ 50 ไมโครกรัมต่อลูกบาศก์เมตร เมื่อเทียบกับค่ามาตรฐานขององค์การอนามัยโลกซึ่งอยู่ที่ 20 ไมโครกรัมต่อลูกบาศก์เมตร ในขณะที่มาตรฐานเฉลี่ย 24 ชั่วโมงอยู่ที่ 120 ไมโครกรัมต่อลูกบาศก์เมตร เมื่อเทียบกับค่ามาตรฐานขององค์การอนามัยโลก 50 ไมโครกรัมต่อลูกบาศก์เมตรดังตารางที่ 2.3 [3]

ตารางที่ 2.3 เปรียบเทียบค่ามาตรฐานของ $PM_{2.5}$ และ PM_{10} ระหว่างประเทศไทยและองค์กรอนามัยโลก

ประเทศ, องค์กร	ค่าเฉลี่ย	$PM_{2.5}$ ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)
ไทย	ค่าเฉลี่ยรายปี	25	50
	ค่าเฉลี่ยรายชั่วโมง	50	120
องค์การอนามัยโลก	ค่าเฉลี่ยรายปี	10	20
	ค่าเฉลี่ยรายชั่วโมง	25	50

2.5 แหล่งที่มาของ $PM_{2.5}$

ตารางที่ 2.4 การประมาณการปล่อยมลพิษทางอากาศจากแหล่งกำเนิด (ตันต่อปี)

ประเภท	$PM_{2.5}$	SO_2	NO_x คือ NO_2
การคมนาคมขนส่ง	50,240	14,000	246,000
การผลิตไฟฟ้า	31,793	231,000	227,000
อุตสาหกรรมการผลิต	65,140	212,000	222,000
ที่อยู่อาศัย/ธุรกิจการค้า	28,265	0	31,000
การเผาในที่โล่ง	209,937	5,000	84,346

จากตารางที่ 2.4 แม้ว่าภาคการผลิตไฟฟ้าจะเป็นแหล่งกำเนิด $PM_{2.5}$ เป็นลำดับรองจากการเผาในที่โล่งการคมนาคมขนส่งและอุตสาหกรรมการผลิตแต่การปล่อยซัลเฟอร์ไดออกไซด์และออกไซด์ของไนโตรเจนต่อปีจากภาคการผลิตไฟฟ้านั้นมีสัดส่วนมากที่สุดในบรรดาแหล่งกำเนิดต่างๆ ซึ่งนำไปสู่เกิด $PM_{2.5}$ จากกระบวนการทางเคมีในบรรยากาศที่มีก๊าซซัลเฟอร์ไดออกไซด์และออกไซด์ของไนโตรเจนเป็นสารตั้งต้น [3]

2.6 สารมลพิษทางอากาศ 6 ชนิด [4]

2.6.1 ฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน ($PM_{2.5}$)

หมายถึง ฝุ่นที่มีเส้นผ่านศูนย์กลางไม่เกิน 2.5 ไมครอน มีสาเหตุมาจากการเผาไหม้ เช่น ไฟป่า การเผาวัสดุเกี่ยวกับการเกษตร หรือ การเผาไหม้จากยานพาหนะและกระบวนการอุตสาหกรรม ซึ่งส่งผลกระทบต่อระบบทางเดินหายใจและโรคปอด

2.6.2 ฝุ่นละอองขนาดไม่เกิน 10 ไมครอน (PM_{10})

หมายถึง ฝุ่นที่มีเส้นผ่านศูนย์กลางไม่เกิน 10 ไมครอน มีสาเหตุมาจากการเผาไหม้เชื้อเพลิง การเผาในที่โล่ง กระบวนการอุตสาหกรรม การบด การโม่ หรือ การทำให้เป็นผงจากการก่อสร้าง ส่งผลกระทบต่อระบบทางเดินหายใจเช่นเดียวกับ $PM_{2.5}$

2.6.3 ก๊าซโอโซน (O_3)

มีคุณลักษณะเป็นก๊าซไม่มีสีหรือมีสีฟ้าอ่อน มีกลิ่นที่ฉุน สามารถละลายในน้ำได้เล็กน้อย ก๊าซโอโซนที่เป็นสารมลพิษมีผลกระทบต่อสุขภาพ การระคายเคืองดวงตาและระคายเคืองต่อระบบทางเดินหายใจและเยื่อบุต่างๆ

2.6.4 ก๊าซคาร์บอนมอนอกไซด์ (CO)

เกิดจากการเผาไหม้ที่ไม่สมบูรณ์ของเชื้อเพลิงที่มีคาร์บอนเป็นองค์ประกอบ คุณลักษณะเป็นก๊าซไม่มีสี กลิ่น และรส เมื่อสะสมในร่างกายจะจับตัวกับฮีโมโกลบินในเลือด ส่งผลให้ร่างกายเกิดอาการอ่อนเพลียและหัวใจทำงานหนักขึ้นเนื่องจากการลำเลียงออกซิเจนไปสู่เซลล์ต่างๆ ของร่างกายลดน้อยลง

2.6.5 ก๊าซไนโตรเจนไดออกไซด์ (NO_2)

พบได้ทั่วไปในธรรมชาติและเกิดจากการกระทำของมนุษย์ เช่น การเผาไหม้เชื้อเพลิงต่างๆ อุตสาหกรรมบางชนิด เป็นต้น มีคุณลักษณะเป็นก๊าซที่ไม่มีสีและกลิ่นรวมถึงละลายน้ำได้เล็กน้อยส่งผลกระทบต่อการมองเห็นหรือโรคเกี่ยวกับทางเดินหายใจ

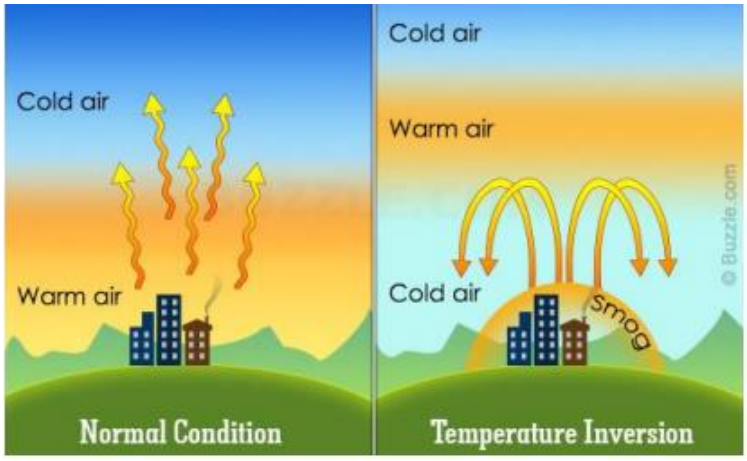
2.6.6 ก๊าซซัลเฟอร์ไดออกไซด์ (SO_2)

ก๊าซชนิดนี้มีผลกระทบโดยตรงต่อสุขภาพทำให้เกิดการระคายเคืองต่อเยื่อบุตา ผิวหนัง และระบบทางเดินหายใจ มีคุณลักษณะเป็นก๊าซไม่มีสีหรืออาจมีสีเหลืองอ่อนๆ มีรสและกลิ่นที่ระดับความเข้มข้นสูงเกิดจากธรรมชาติและการเผาไหม้เชื้อเพลิงที่มีกำมะถันเป็นส่วนประกอบ

2.7 ปรากฏการณ์อุณหภูมิผกผัน (Temperature Inversion)

การที่มีการเรียงสลับของอุณหภูมิของอากาศในชั้นโทรโพสเฟียร์ (Troposphere) ซึ่งโดยปกติอุณหภูมิของอากาศจะลดลงประมาณ 6.4 ถึง 6.5 องศาเซลเซียสต่อความสูงหนึ่งกิโลเมตรแต่เมื่อมีการเกิดปรากฏการณ์อุณหภูมิผกผันจะมีชั้นอากาศอุ่นไปแทรกอยู่ระหว่างกลางอากาศที่เย็นกว่าทั้งด้านล่างด้านบนทำให้ลำดับชั้นของอุณหภูมิของอากาศเกิดความผิดปกติ

ปรากฏการณ์อุณหภูมิผกผันมีบทบาทสำคัญในการขัดขวางการพาความร้อน (Convection) ตามธรรมชาติส่งผลโดยตรงต่อการพัฒนาตัวของเมฆ การเกิดฝน และยังส่งผลให้เกิดลักษณะคล้ายเพดานห้องที่กักฝุ่นละอองหรือแก๊สมลพิษที่ก่อตัวในระดับพื้นผิวโลกไม่ให้ลอยขึ้นสู่เบื้องบน การเกิดปรากฏการณ์อุณหภูมิผกผันมีได้หลายสาเหตุขึ้นอยู่กับลักษณะภูมิประเทศบริเวณนั้นๆ เช่น หุบเขาในต่างประเทศมักเกิดจากการไหลของลมเย็นเข้าไปใต้อากาศอุ่น แต่สำหรับในในประเทศไทยมักเกิดในฤดูหนาวที่ลมนิ่ง หลังจากช่วงบ่ายแสงแดดได้ให้ความร้อนกับอากาศและพื้นดินในเมือง เมื่อเข้าสู่ช่วงค่ำพื้นดินจะเย็นลงอย่างรวดเร็วไม่เหมือนฤดูร้อนที่ยังร้อนอบอ้าวไปถึงกลางคืน เมื่อพื้นดินเย็นลงแต่อากาศสูงขึ้นไปกลับเย็นลงช้ากว่าเนื่องจากลักษณะของเมืองทำให้เกิดโดมความร้อนคลุมเอาไว้ทำให้เกิดปรากฏการณ์อุณหภูมิผกผันขึ้นมาในช่วงกลางดึกคือชั้นความร้อนที่ปิดกั้นการลอยตัวของฝุ่นละออง จึงมักพบสภาพที่ฝุ่นละอองในเมืองลดลงในช่วงบ่าย(เนื่องจากอากาศมีการเรียงลำดับความร้อนได้ถูกต้อง)และไปเพิ่มสูงในช่วงกลางดึก [5] ดังภาพที่ 2.1



ภาพที่ 2.1 ปรากฏการณ์อุณหภูมิผกผัน

ที่มา https://www.rihes.cmu.ac.th

2.8 ค่าสหสัมพันธ์ (Correlation)

เป็นการหาความสัมพันธ์ระหว่างตัวแปรหรือข้อมูลตั้งแต่ 2 ตัวขึ้นไป ในการพิจารณาความความสัมพันธ์ระหว่างตัวแปรว่ามีมากหรือมีน้อยนั้น จะใช้ค่าสัมประสิทธ์สหสัมพันธ์ (Correlation Coefficient) เป็นค่าที่ใช้ในการวัดความสัมพันธ์ซึ่งในแง่สถิติมีวิธีการอยู่หลายวิธีในขั้นตอนการเลือกใช้วิธีใดนั้นขึ้นอยู่กับลักษณะของตัวแปร ในขั้นตอนการตีความหมายจะดูในเรื่องของความสัมพันธ์ ความเกี่ยวพัน ความสอดคล้อง ความแปรผันร่วมกัน แต่ไม่ได้หมายความว่าตัวแปรหนึ่งเป็นสาเหตุและตัวหนึ่งเป็นผล เช่น ศึกษาความสัมพันธ์ระหว่างส่วนสูงกับน้ำหนัก ไม่สามารถบอกได้ว่าส่วนสูงหรือน้ำหนักเป็นเหตุและตัวใดเป็นผล สามารถตีความหมายได้เพียงว่ามีความสัมพันธ์กันหรือไม่และมีความสัมพันธ์มากน้อยเพียงใด ค่าสัมประสิทธิ์สหสัมพันธ์จะใช้สัญลักษณ์ r แทนสัมประสิทธิ์สหสัมพันธ์ของกลุ่มตัวอย่าง

การบอกระดับหรือขนาดของความสัมพันธ์ ใช้ตัวเลขของค่าสัมประสิทธิ์สหสัมพันธ์หากค่าสัมประสิทธิ์สหสัมพันธ์ มีค่าเข้าใกล้ -1 หรือ 1 หมายถึงการมีความสัมพันธ์กันในระดับสูง แต่หากค่าสัมประสิทธิ์สหสัมพันธ์เข้าใกล้ 0 หมายถึงมีความสัมพันธ์กันในระดับน้อยหรือไม่มีเลย

ตารางที่ 2.5 การพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์

ค่า r	ระดับความสัมพันธ์
0.90 - 1.00	มีความสัมพันธ์ระดับสูงมาก
0.70 - 0.90	มีความสัมพันธ์ระดับสูง
0.50 - 0.70	มีความสัมพันธ์ระดับปานกลาง
0.30 - 0.50	มีความสัมพันธ์ระดับต่ำ
0.00 - 0.30	มีความสัมพันธ์ระดับต่ำมาก

สำหรับเครื่องหมาย +, - หน้าสัมประสิทธิ์สหสัมพันธ์จะบอกถึงทิศทางของความสัมพันธ์

หาก r มีเครื่องหมาย + หมายถึง ตัวแปรที่นำมาทดสอบมีความสัมพันธ์ไปในทิศทางเดียวกัน

หาก r มีเครื่องหมาย – หมายถึง ตัวแปรที่นำมาทดสอบมีความสัมพันธ์ในทิศทางตรงกันข้าม ค่าสัมประสิทธิ์สหสัมพันธ์จะใช้ได้เหมาะสมกับข้อมูลที่มีความสัมพันธ์เชิงเส้นเท่านั้น สำหรับการคำนวณหากพบว่า r = 0 การตีความหมายข้อมูลว่าไม่มีความสัมพันธ์กันอาจจะไม่ถูกต้อง เพราะข้อมูลอาจมีความสัมพันธ์กันในลักษณะอื่นที่ไม่ได้เป็นเชิงเส้น [6]

2.9 สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient)

สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันเป็นวิธีทางสถิติที่ใช้ในการหาความสัมพันธ์ระหว่างตัวแปรหรือข้อมูล 2 ชุด โดยที่ตัวแปรหรือข้อมูล 2 ชุดนั้นต้องอยู่ในรูปของข้อมูลในมาตราอันตรภาคหรืออัตราส่วน (Interval or Ration Scale) เช่น การหาความสัมพันธ์ระหว่างภาวะสุขภาพกับการดูแลตนเอง การหาความสัมพันธ์ระหว่างน้ำหนักแรกเกิดของทารกกับอายุของมารดา เป็นต้น โดยมีหลักการคำนวณตามสมการด้านล่าง

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

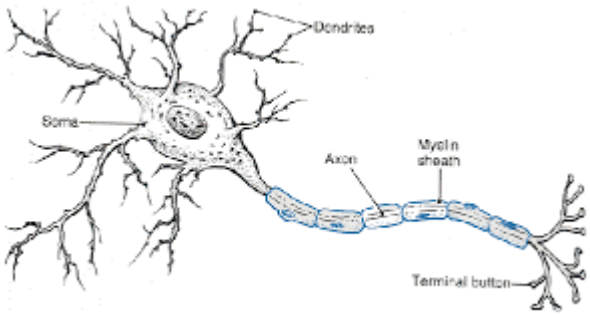
โดยค่า Covariance (cov) เป็นตัวเลขที่บ่งบอกถึงความสัมพันธ์เพียงแต่ค่าของมันไม่สามารถที่จะเทียบเคียงกันระหว่างหลายๆ ตัวแปรได้ เช่น ไม่สามารถบอกได้ว่า cov(x,y) = 20 มีค่ามากกว่า cov(z,y) = 10 เพราะการหาค่า Covariance นั้นจะไม่ได้อยู่ในช่วงขอบเขตเดียวกันจึงทำการหารด้วยส่วนเบี่ยงเบนมาตรฐานของทั้งสองตัวแปรเพื่อปรับเปลี่ยนให้มีขอบเขต -1 ถึง 1 เพื่อให้สามารถนำมาเปรียบเทียบกันได้ [7]

2.10 การเรียนรู้ของเครื่อง (Machine Learning)

เทคนิคการเรียนรู้ส่วนใหญ่เป็นการเรียนรู้เชิงอุปนัย (Inductive Learning) และมีบางเทคนิคเป็นการเรียนรู้เชิงวิเคราะห์ (Analytical Learning) การเรียนรู้เชิงอุปนัยคือการเรียนรู้ที่หากฎเกณฑ์หรือความรู้ที่แฝงอยู่ในชุดตัวอย่างสอน (Training Example Set) เพื่อเรียนรู้ให้ได้ความรู้ใหม่ที่สอดคล้องกับชุดตัวอย่างสอน ส่วนการเรียนเชิงวิเคราะห์เป็นการจัดรูปแบบของความรู้ใหม่เพื่อให้ใช้งานได้อย่างมีประสิทธิภาพมากขึ้นและสามารถทำงานได้เร็วมากขึ้น [8]

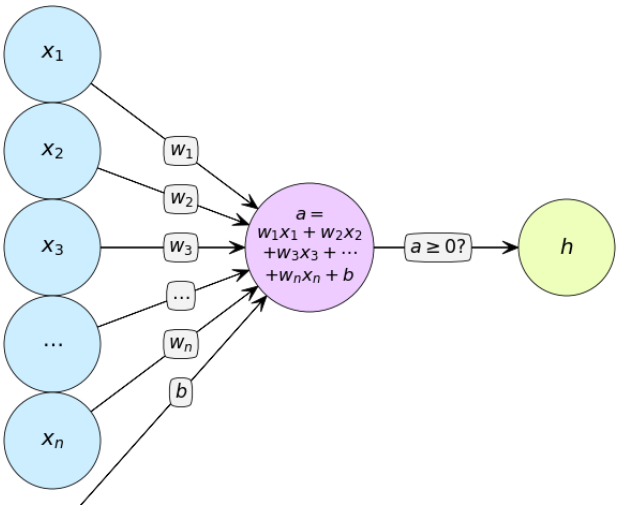
2.11 โครงข่ายประสาทเทียม (Artificial Neural Network)

โครงข่ายประสาทถูกพัฒนามาจากการเลียนแบบการทำงานของสมองมนุษย์ที่มีโครงสร้างแสดงตามภาพที่ 2.2 สมองมนุษย์มีส่วนประกอบไปด้วย เซลล์ประสาท (Neuron) มากมาย โดยเซลล์เหล่านี้ติดต่อสื่อสารระหว่างกันโดยการส่งข้อมูลผ่านสัญญาณไฟฟ้าเคมี (Electrochemical Signal) สัญญาณนี้วิ่งผ่าน จุดประสาท (Synapse) โดยที่สัญญาณที่ผ่านจุดประสานตำแหน่งต่างๆ จะถูกผสมเข้าด้วยกันเป็นหนึ่งเดียวถ้าสัญญาณรวมนี้มีค่าสูงกว่าขีดแบ่งเซลล์ประสาทที่รับสัญญาณนี้จะถูกกระตุ้นให้ส่งผลลัพธ์ไปยังเซลล์ประสาทอื่นผ่านทางแกนประสาท



ภาพที่ 2.2 เซลล์ประสาทของสมองมนุษย์

โครงข่ายประสาทประกอบด้วย เซลล์ประสาทประดิษฐ์ (Artificial Neuron) ที่มีโครงสร้างคล้ายกับเซลล์ประสาทมนุษย์ ข้อมูลเข้าที่ถูกส่งไปยังเซลล์ประสาทแต่ละเซลล์จะถูกให้น้ำหนักและถูกรวมเข้าด้วยกัน ซึ่งถ้าผลรวมที่ได้มีค่าสูงกว่าขีดแบ่ง เซลล์ประสาทจะส่งข้อมูลออกไป ยกตัวอย่างเช่น ข้อมูลออก คือ 1 ในกรณีที่ผลรวมมีค่าเกินขีดแบ่ง นอกเหนือจากนี้ให้ค่าข้อมูลออกเป็น 0 เป็นต้น [9]

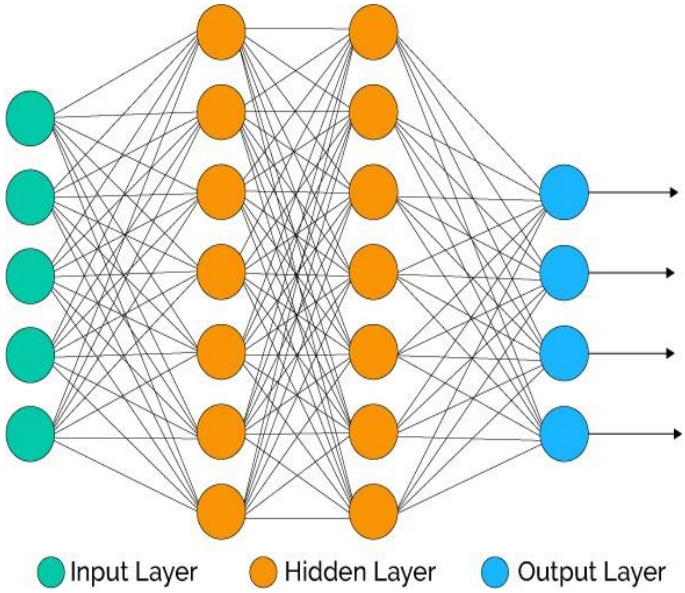


ภาพที่ 2.3 แบบจำลองโครงข่ายประสาทเทียมอย่างง่าย

ที่มา https://phyblas.hinaboshi.com/umaki02

2.12 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึกคือระบบคอมพิวเตอร์ที่พยายามเลียนแบบการทำงานของระบบโครงข่ายประสาทในสมองมนุษย์ ถือเป็นซับเซ็ตของการเรียนรู้ของเครื่อง (Machine Learning) แนวคิดของการเรียนรู้เชิงลึกถูกสร้างขึ้นจากการนำเอาโครงข่ายประสาทเทียม (Artificial Neural Network) หลายชั้นมาต่อกัน โดยชั้นแรกสุดจะทำหน้าที่ในการรับข้อมูล (Input Layer) ชั้นสุดท้ายจะทำหน้าที่ส่งผลลัพธ์การประมวลผลออกมา (Output Layer) ส่วนชั้นระหว่างชั้นแรกสุดและชั้นสุดท้ายจะถูกเรียกว่า Hidden Layer การเรียนรู้เชิงลึก (Deep Learning) มีที่มาจากการใช้ชั้นของโครงข่ายประสาทเทียมหลายอันมาต่อกัน เนื่องจากชั้นเหล่านี้เป็นโครงสร้างที่ถูกจัดเก็บแบบกองซ้อน (Stack) จึงเปรียญได้ว่าชั้นที่จำนวนมากก็จะทำให้มีโครงสร้างที่ลึกมากยิ่งขึ้น แบบจำลองที่ใช้การเรียนรู้เชิงลึกให้ความแม่นยำที่สูงในหลายๆ ปัญหาตั้งแต่การตรวจจับวัตถุไปจนถึงการรู้จำเสียงพูด โดยที่ไม่จำเป็นต้องให้ความรู้พื้นฐานใดๆ กับแบบจำลองไว้ล่วงหน้าเลย เพียงแค่ให้ข้อมูลตัวอย่าง (Input Data) แบบจำลองก็จะทำการเรียนรู้จากแหล่งข้อมูลและทำการสังเคราะห์เป็นองค์ความรู้ออกมาได้อย่างอัตโนมัติ [10]



ภาพที่ 2.4 โครงข่ายประสาทเทียมแบบการเรียนรู้เชิงลึก

ที่มา https://medium.com/mmp-li/deep-learning-แบบฉบับคนสามัญชนทั่วไป-ep-1-neural-network-history-f7789236a9a3

2.13 Loss Function, Cost Function, Error Function

สำหรับทั้งสามชื่อนี้หมายถึงสิ่งเดียวกัน คือ ฟังก์ชันทางคณิตศาสตร์ที่พยายามจับคู่ค่าผลลัพธ์จากข้อมูลนำเข้าหลายๆ ตัวให้ออกมาเป็นตัวเลขจำนวนจริงเพียงหนึ่งตัวเดียว เพื่อใช้ในการบอกประสิทธิภาพในการทำงานนั้นๆ โดยถูกใช้ในการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง โดยต้องการให้ ได้ Loss Function มีค่าน้อยที่สุด โดย Loss Function เหล่านี้ก็มีหลายประเภทด้วยกัน แต่สำหรับงานที่เป็น Regression Problem [9] แล้วส่วนใหญ่จะใช้สมการดังตารางที่ 2.6

ตารางที่ 2.6 Loss Function สำหรับ Regression Problem

ชื่อ	สมการ
Mean Absolute Error (MAE)	$MAE = \frac{1}{N} * \sum \lvert Prediction - Actual \rvert$
Mean Square Error (MSE)	$MSE = \frac{1}{N} * \sum (Prediction - Actual)^2$

ตารางที่ 2.6 Loss Function สำหรับ Regression Problem (ต่อ)

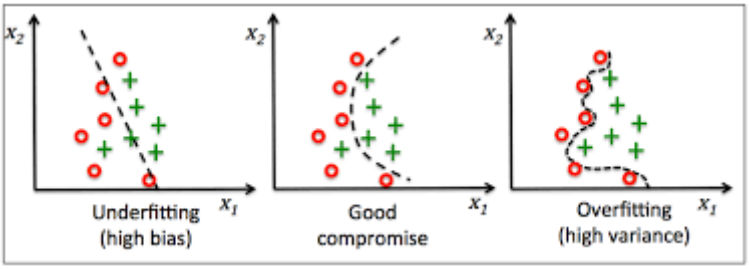
ชื่อ	สมการ
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{N} * \sum (Prediction - Actual)^2}$
R-squared	$R^2 = 1 - \left(\frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}\right)$

MAE, MSE, RMSE มีค่ายิ่งต่ำหมายความว่าแบบจำลองมีประสิทธิภาพในการทำงานที่ดี ในทางปฏิบัติโอกาศที่จะฝึกฝนแบบจำลองได้ Loss Function = 0 นั้นเป็นไปได้ยาก เพราะสามารถนำไปสู่ปัญหา Overfitting และ สำหรับ R^2 จะมีค่าระหว่าง 0-1 ยิ่งเข้าใกล้ 1 แสดงว่าแบบจำลองสามารถทำนายผลได้ดีมาก

2.14 Underfitting และ Overfitting

Underfitting หมายถึง แบบจำลองที่ไม่สามารถทำงานได้ เนื่องจากไม่สามารถหารูปแบบความสัมพันธ์หรือแนวโน้มของข้อมูลได้ อาจเกิดขึ้นมาจากแบบจำลองที่ไม่เหมาะสมกับข้อมูลหรือข้อมูลที่ใช้ฝึกฝนมีจำนวนน้อยเกินไปในกรณีนี้แบบจำลองจะมี Bias สูง เช่น นำข้อมูลฝึกฝนมาทดสอบจะได้ความแม่นยำที่ต่ำ เมื่อนำข้อมูลทดสอบมาทดสอบก็ได้ความแม่นยำที่ต่ำเช่นกัน

Overfitting หมายถึง แบบจำลองตอบสนองต่อการรบกวนจำนวนมาก จนแบบจำลองเริ่มทำการฝึกฝนจากข้อมูลรบกวนทำให้แบบจำลองได้เรียนรู้ลักษณะข้อมูลที่ไม่ถูกต้อง หรือการที่แบบจำลองทำการฝึกฝนข้อมูลฝึกฝนได้ดีจนเกินไป เมื่อนำข้อมูลฝึกฝนมาทดสอบจะได้ความแม่นยำที่สูงมาก แต่เมื่อนำข้อมูลทดสอบมาทำการทดสอบแบบจำลองกลับพบว่าความแม่นยำที่ได้นั้นต่ำลง เนื่องจากข้อมูลที่ใช้ทดสอบเป็นข้อมูลที่แบบจำลองไม่เคยฝึกฝนมาก่อนทำให้ไม่สามารถทำนายได้ถูกต้อง กรณีนี้ข้อมูลมีความ Variance สูง [11]



ภาพที่ 2.5 การเกิด Underfitting และ Overfitting

2.15 Gradient Descent

ในการเรียนรู้ของเครื่อง (Machine Learning) ปกติเมื่อแบ่งกลุ่มข้อมูลสำหรับฝึกฝนและทดสอบเสร็จ จากนั้นทำการฝึกฝนแบบจำลองและทดสอบด้วยข้อมูลทดสอบจะถือว่าเสร็จสิ้นกระบวนการ สามารถนำไปใช้งานได้ทันทีแต่ในการเรียนรู้เชิงลึก (Deep learning) ในระหว่างการฝึกฝนข้อมูลด้วยเซลล์ประสาท (Neuron) สามารถปรับปรุง Error และ Loss ได้อีกด้วย เพื่อให้ได้ค่าที่ดีที่สุดโดยขั้นตอนนี้เรียกว่า Optimization ซึ่งจะทำการแปลงค่า น้ำหนัก (Weight), ค่า (Bias) ที่เชื่อมกับเซลล์ประสาทนั้นๆ [12]

โดย Optimization ก็มีหลายวิธีด้วยกัน วิธีที่นิยมมีดังนี้

2.15.1 Stochastic Gradient Descent (SGD)

เป็นวิธีที่จะมีการเปลี่ยนแปลงค่าพารามิเตอร์ (Parameter) ในทุกๆชุดข้อมูลฝึกฝน มีการอัพเดทค่าครั้งเดียวต่อการฝึกฝน 1 รอบ

$\boldsymbol{\theta=\theta-\eta\cdot\nabla J(\theta;x(i);y(i))}$; โดยที่ { x(i), y(i) } คือชุดข้อมูลฝึกฝน

ในทุกครั้งที่มีการอัพเดท ค่าพารามิเตอร์ที่อัพเดทจะมีความแปรปรวนสูงส่งผลกับค่าของ Loss Function แปรผันไปตาม Different Intensities โดยจะสามารถช่วยให้พบค่าที่น้อยที่สุดได้แต่ สำหรับข้อเสียของวิธีนี้คือยิ่งค่าที่ได้ต่ำมากเท่าไหร่ค่าที่ได้ก็จะมีความแปรปรวนและซับซ้อนมาก

2.15.2 Mini Batch Gradient Descent

วิธีนี้เป็นวิธีที่พัฒนาเพื่อแก้ปัญหาของ Gradient Descent โดยนำข้อดีของ Gradient Descent และ Stochastic Gradient มารวมกันและวิธีนี้จะทำการอัพเดทค่าเป็นชุดโดยแต่ละชุดจะประกอบด้วยข้อมูลจำนวน n ข้อมูล

2.15.3 Momentum

ในการปรับค่าแต่ละครั้ง SGD ทำให้เกิดความแปรปรวนมาก ทำให้มีความยากที่จะเข้าสู่จุดที่ต่ำที่สุดได้จึงเกิดวิธีที่เรียกว่า Momentum ขึ้นคิดค้นเพื่อเร่งความเร็วในการ Optimize ของ SGD ให้ความสำคัญกับการพุ่งไปยังทิศทางที่ใกล้จุดกลางมากที่สุดก่อนทำให้ทิศทางที่ไม่เกี่ยวข้องมีความความสำคัญที่น้อยลงไปด้วยเหตุนี้ทำให้เกิดทิศทางที่ถูกต้องขึ้น โดยมีการเพิ่ม $\boldsymbol{\gamma}$ เข้ามาในการอัพเดททิศทางแต่ละครั้ง

$\boldsymbol{V(t)=\gamma V(t-1)+\eta\nabla J(\theta).}$ อัพเดทค่าพารามิเตอร์ด้วย $\boldsymbol{\theta=\theta-V(t).}$

2.15.4 Adagrad

เป็นวิธีที่สามารถปรับค่า Learning Rate ให้คู่ควรกับพารามิเตอร์ได้ โดยจะมีการอัพเดทจำนวนมากสำหรับพารามิเตอร์ที่มีจำนวนน้อย และอัพเดทน้อยถ้าค่าพารามิเตอร์มีจำนวนมาก จึงทำให้วิธี Adagrad เหมาะกับข้อมูลที่มีการกระจัดกระจาย (Sparse Data) ค่า Learning Rate จะถูกเปลี่ยนทุกครั้งสำหรับพารามิเตอร์ **θ** โดยมีการอ้างอิงจากทิศทางที่ผ่านมาจาก Gradient Descent

2.15.5 AdaDelta

เป็นวิธีที่มีการต่อยอดมาจาก Adagrad ซึ่งวิธีนี้สามารถลบปัญหา Decaying Learning Rate ที่เกิดใน Adagrad ได้ โดยแทนที่จะทำการเก็บสะสมการคำนวณทั้งหมดที่ผ่านไปแล้วของ Gradient Descent วิธี AdaDelta จะกำจัดการสะสมค่าของการคำนวณ Gradient Descent เพื่อแก้ขนาดของน้ำหนัก (Weight) ที่จะเกิดขึ้น แทนที่จะทำการเก็บค่าน้ำหนักที่ได้รับการอัพเดทมาก่อนหน้าที่ยังไม่เหมาะสม เปลี่ยนเป็นการหาผลรวมของ Gradient ทำแบบนี้ซ้ำๆ เพื่อแก้ปัญหา Decaying Learning Rate

2.15.6 Adaptive Moment Estimation (Adam)

เป็นวิธีที่สามารถปรับ Learning Rate ของค่าพารามิเตอร์ในแต่ละครั้งได้และแก้ปัญหา Decaying Learning Rate ของ Gradient Descent ในแต่ละขั้นตอนที่ผ่านมาได้เหมือนกับ AdaDelta พร้อมทั้งสามารถอธิบายการเกิด Decaying Average ที่ผ่านมาได้เหมือนกับวิธี Momentum โดยวิธีนี้เป็นวิธีที่นิยมมากที่สุดเพราะรวมจุดเด่นของแต่วิธีและลบข้อเสียต่างๆ ออกไปทำให้แบบจำลองไม่หยุดฝึกฝนและยังทำงานไวกว่า Gradient Descent และลดปัญหาการแกว่งของค่าพารามิเตอร์

2.16 ฟังก์ชันกระตุ้น (Activation Function)

ในเซลล์ประสาท (Neuron) แต่ละตัวจะได้รับข้อมูลนำเข้า (Input) มากมายแล้วข้อมูลนำเข้าเหล่านั้นมาทำการประมวลผลแล้วส่งเป็นข้อมูลออก (Output) 1 ตัวออกไปที่ Axon เพื่อส่งไปให้เซลล์ประสาทอื่นๆ ต่อไปฟังก์ชันกระตุ้น (Activation Function) คือ ฟังก์ชันที่จะรับผลรวมของการประมวลผลทั้งหมดจากทุกข้อมูลนำเข้าภายใน 1 เซลล์ประสาทแล้วพิจารณาว่าจะส่งต่อข้อมูลออกเป็นค่าเท่าไหร่หรือป็นฟังก์ชันที่ใช้ในการหาข้อมูลออกของเซลล์ประสาทมีชื่อเรียกอีกชื่อว่า Transfer Function สามารถแบ่งประเภทได้เป็น 2 ประเภท ได้แก่ 1.ฟังก์ชันเชิงเส้น 2.ฟังก์ชันแบบไม่เชิงเส้น Activation Function ที่เห็นได้บ่อยมีดังนี้ [12]

2.16.1 Sigmoid Function

เป็นฟังก์ชัน S-Curve เป็นที่นิยมเนื่องจากข้อมูลออกของฟังก์ชันนี้มีค่า 0 ถึง 1 ดังนั้นจึงเหมาะสมมากหากต้องการค่าความน่าจะเป็น (Probability) ของ Output ที่มีค่าความน่าจะเป็นตั้งแต่ 0-1 โดยฟังก์ชันนี้ส่วนใหญ่จะใช้กับการ Binary Classification Problem หรือใช้ใน Logistic Classification

2.16.2 Hyperbolic Tangent Function (Tanh Function)

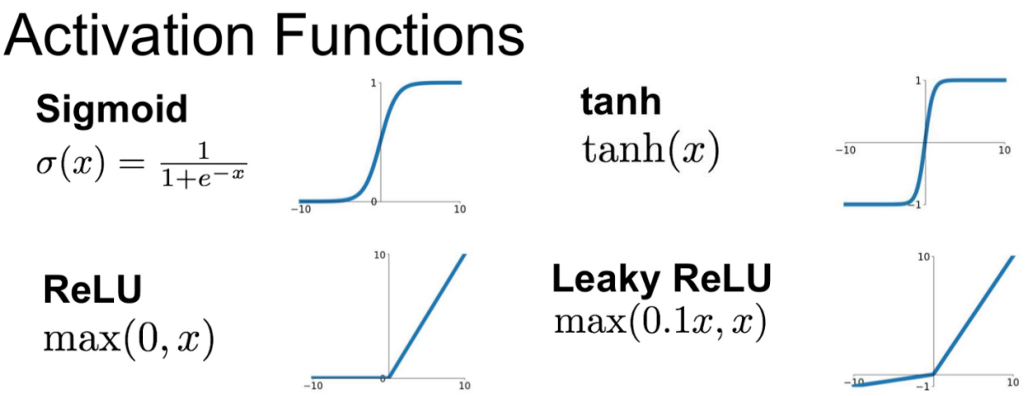
เป็นฟังก์ชัน S-Curve เช่นเดียวกับ Sigmoid Function แต่ข้อมูลออกของวิธีนี้จะเป็น -1 ถึง 1 จุดเด่นของฟังก์ชันนี้คือหากผลลัพธ์ที่ได้ออกมาเป็นค่าติดลบจะมีแนวโน้วที่จะถูกจับคู่ไปยังค่าที่เป็นลบสูงแต่หากได้รับค่าที่เป็น 0 ก็จะมีแนวโน้มไปจับคู่กับค่าที่ใกล้กับ 0 ส่วนใหญ่ถูกนำไปใช้กับงาน Binary Classification Problem

2.16.3 Rectified Linear Unit (ReLU Function)

เป็นฟังก์ชันที่ได้รับความนิยมมาก นิยมใช้ใน Convolutional Neural Network หรือ Deep learning โดยตัวฟังก์ชันนี้จะทำการจับคู่ค่าที่ต่ำกว่า 0 จะได้ข้อมูลออกเป็น 0 ไปโดยปริยายซึ่งหากค่ามากกว่าหรือเท่ากับ 0 จะได้ข้อมูลออกที่มากกว่า 0 ตามดังภาพที่ 2.6 ฟังก์ชันนี้มีข้อเสียคือการจับคู่ค่าที่เป็นลบทั้งหมดเป็น 0 จะทำให้ลดความสามารถของแบบจำลองลงและข้อมูลออกมีขอบเขตตั้งแต่ 0 ถึง บวกอินฟินิตี้ ทำให้ยากต่อการจัดการเนื่องจากไม่มีขอบเขต

2.16.4 Leaky ReLU Function

ฟังก์ชันนี้ถูกคิดค้นเพื่อแก้ปัญหาของ ReLU Function โดยจะทำการเพิ่มขอบเขตของ ข้อมูลออกจาก 0 ถึง บวกอินฟินิตี้ เป็น ลบอินฟินิตี้ ถึง บวกอินฟินิตี้ โดยเมื่อได้รับข้อมูลนำเข้าที่ติดลบจะทำการจับคู่เป็นค่าข้อมูลออกที่ติดลบน้อยๆ แทน



ภาพที่ 2.6 ตัวอย่างฟังก์ชันกระตุ้นแต่ละประเภท

2.17 Batch Size, Iterations, Epoch [13]

Epoch หมายถึง การฝึกฝนแบบจำลองโดยทำการนำข้อมูลเข้าแบบจำลองโดยใช้ข้อมูลในชุดข้อมูลทั้งหมด (Dataset) ที่มีอยู่จนครบทั้งหมดทุกตัวนับเป็น 1 รอบ

Batch Size หมายถึง การแบ่งข้อมูลที่จะใช้ในการฝึกฝนเป็นกลุ่มที่เล็กลง เนื่องจากคอมพิวเตอร์นั้นไม่สามารถนำข้อมูลขนาดใหญ่ เช่น รูปที่มีความละเอียดสูง เป็นต้น ที่มีนำเข้าแบบจำลองเพื่อทำการประมวลผลได้ในครั้งเดียว จึงต้องทำการแบ่งข้อมูลออกเป็นกลุ่มเล็กๆ

Iterations หมายถึง จำนวนรอบของ Batch Size ที่ต้องส่งเข้าไปยังแบบจำลองเพื่อให้ครบ 1 Epoch หรือจำนวนทั้งหมดที่ใช้ Batch size จำนวนเท่านี้เพื่อฝึกฝนแบบจำลองจนครบข้อมูลทั้งหมด

2.18 การป้องกันไม่ให้เกิด Overfitting [14]

2.18.1 Cross-Validation

เป็นวิธีที่ใช้ได้ผลมากในการป้องกันไม่ให้แบบจำลองเกิดการ Overfitting ทำโดยการสร้างหลายๆ ข้อมูลฝึกฝนและข้อมูลทดสอบเรียกว่า K-Fold Cross-Validation ซึ่งจะทำการแบ่งข้อมูลออกมาเป็น K ส่วนแล้วทำการฝึกฝนแบบจำลองด้วย K-1 ส่วน ส่วนที่เหลือเป็นข้อมูลสำหรับใช้ในการทดสอบแบบจำลองแล้วทำการสลับข้อมูลที่ใช้ในการฝึกฝนและทดสอบไปเรื่อยๆ จนหมด

2.18.2 เพิ่มจำนวนข้อมูลฝึกฝนให้มากขึ้น

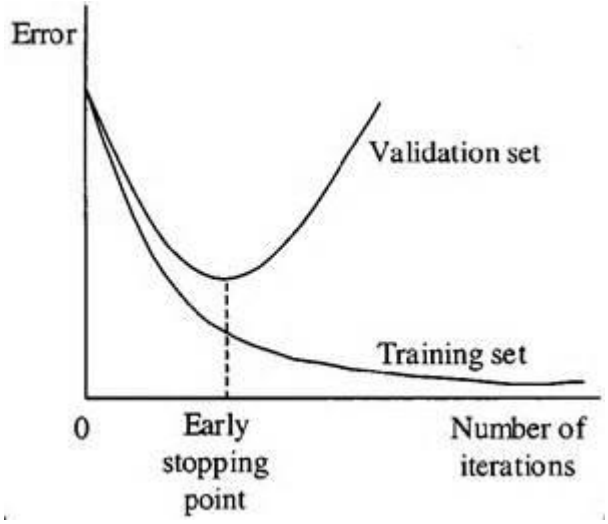
สำหรับวิธีนี้อาจจะใช้ไม่ได้ผลในทุกๆ ครั้ง แต่การฝึกฝนแบบจำลองด้วยข้อมูลที่มากขึ้นจะทำให้แบบจำลองที่พัฒนาสามารถเรียนรู้ตรวจจับรูปแบบของข้อมูลได้ดีขึ้น ในกรณีที่ข้อมูลที่นำมาฝึกฝนนั้นไม่มีข้อมูลรบกวนอยู่ (Noise) หากจะใช้วิธีนี้ควรแน่ใจแล้วว่าข้อมูลที่มีอยู่ผ่านการทำความสะอาดมาแล้วและแม้จะผ่านการทำความสะอาดมาแล้วก็ควรแน่ใจด้วยว่าข้อมูลที่ผ่านการทำความสะอาดมาแล้วนั้นไม่มีข้อมูลรบกวนอยู่ด้วย

2.18.3 การลบคุณสมบัติที่ไม่จำเป็น (Remove Features)

ในแบบจำลองอาจจะมีการเลือกคุณสมบัติในตัวเองอยู่แล้ว (Built-in Feature Selection) แต่ในแบบจำลองบางแบบจำลองที่ไม่มีคุณสมบัตินี้นักวิเคราะห์ก็สามารถทำเองได้ โดยขั้นตอนการหาคุณสมบัติที่ส่งผลต่อประสิทธิภาพของแบบจำลองนั้นมีหลายวิธีขึ้นอยู่กับประสบการณ์ของผู้ดำเนินงานด้วย เช่น การหาความสัมพันธ์ของตัวแปรอิสระกับตัวแปรตาม หากไม่มีความสัมพันธ์กันอาจจะหมายความว่าคุณสมบัตินั้นไม่จำเป็นสำหรับแบบจำลอง การฝึกฝนแบบจำลองด้วยคุณสมบัติที่ไม่จำเป็นส่งผลเสียหลายด้านไม่ว่าจะเป็นการเพิ่มข้อมูลรบกวนในตัวของชุดข้อมูลเป็นสาเหตุของการเกิด Overfitting การเพิ่มเวลาในการฝึกฝนแบบจำลองเนื่องจากมิติที่ใช้มีขนาดเพิ่มมากขึ้น เป็นต้น

2.18.4 การหยุดฝึกฝนแบบจำลองก่อนการเกิด Overfitting (Early Stopping)

เมื่อผู้ดำเนินการทำการฝึกฝนแบบจำลองซ้ำๆ สามารถทำการเช็คได้ว่าในแต่ละรอบที่ทำการฝึกฝนแบบจำลองนั้น ตัวแบบจำลองมีประสิทธิภาพเป็นอย่างไร โดยจะทำการฝึกฝนไปเรื่อยๆ จนกว่าจะไม่ได้ประสิทธิภาพที่ดีขึ้นจากภาพที่ 2.7 จะเห็นได้ว่าเมื่อทำการฝึกฝนข้อมูลไปเรื่อยๆ แล้วทำการวัดประสิทธิภาพของแบบจำลองด้วยข้อมูลตรวจสอบ (Validation Set) เมื่อผ่านจุดๆ หนึ่งไปประสิทธิภาพที่ได้กลับไม่ดีขึ้น นอกจากนี้ยังแย่ขึ้นไปเรื่อยๆ จึงควรทำการหยุดการฝึกฝนแบบจำลองให้อยู่ภายในรอบที่มีประสิทธิภาพมากที่สุด



ภาพที่ 2.7 การวัดประสิทธิภาพของการฝึกฝนแบบจำลอง

ที่มา https://elitedatascience.com/overfitting-in-machine-learning

2.18.5 Regularization

เป็นเทคนิคที่หลากหลายที่บังคับให้แบบจำลองมีความง่ายขึ้น โดยวิธีการนั้นขึ้นอยู่กับประเภทของแบบจำลองที่ใช้ ยกตัวอย่างเช่น การพรูนในต้นไม้ตัดสินใจ (Prune Decision Tree) หรือการ Dropout ในโครงข่ายประสาทเทียมหรือการเพิ่มพารามิเตอร์ใน Cost Function ของ Regression Problem สำหรับโครงข่ายประสาทเทียมส่วนใหญ่มีปัญหา Overfitting ได้ง่ายแต่สามารถแก้ปัญหาได้โดยการใช้ Ensembles คือการสร้างหลายๆ แบบจำลองแล้วนำข้อมูลออกมาทำการเฉลี่ยกันแต่การทำหลายแบบจำลองนั้นสิ้นเปลืองทรัพยากร เวลา และต้องทำการดูแลรักษาหลายแบบจำลอง จึงใช้การ Dropout ในการแก้ปัญหานี้ วิธีการก็คือ การ Dropout จะทำการสุ่มถอดเซลล์ประสาทบางเซลล์ประสาทออกในระหว่างการฝึกฝน ทำให้สามารถใช้แบบจำลองเดียวแต่จำลองเป็นแบบจำลองหลายแบบจำลองได้

2.18.6 Ensembling

เป็นวิธีของการเรียนรู้ของเครื่องแบบหนึ่งโดยเป็นการรวมการทำนายของหลายๆ แบบจำลองเข้าด้วยกัน นอกจากจะช่วยในการป้องกันการเกิด Overfitting แล้วยังสามารถพัฒนาประสิทธิภาพการทำนายได้อีกด้วย การทำ Ensembling หลักๆ แล้วมีสองวิธี ได้แก่

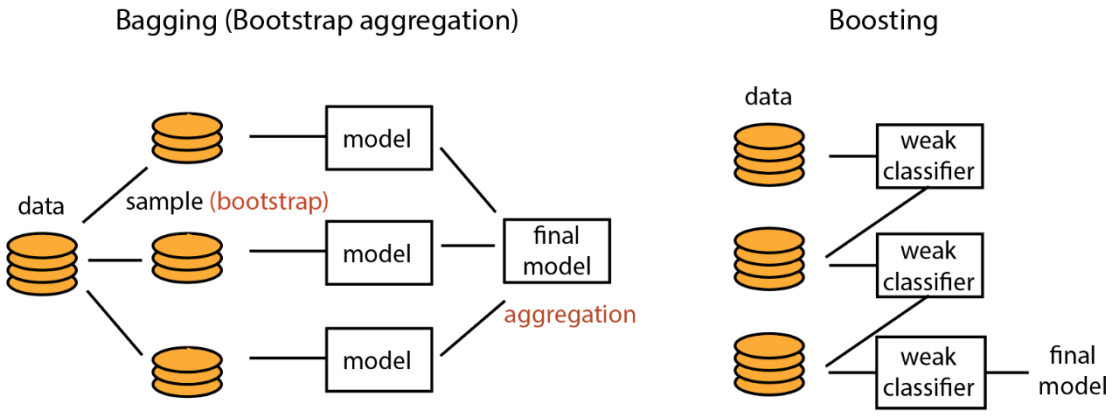
1) Bootstrap Aggregation (Bagging) เป็นการพยายามลดโอกาสการเกิด Overfitting ของแบบจำลองที่มีความซับซ้อนสูง

2) Boosting พยายามที่จะพัฒนาความสามารถในการทำนายให้มีความยืดหยุ่นมากยิ่งขึ้นได้แบบจำลองที่ง่ายขึ้น

2.19 Bootstrap Aggregating (Bagging) และ Boosting [15]

Bagging เป็นการสุ่มข้อมูลจากข้อมูลทั้งหมดแล้วนำ ตัวเรียนรู้ (Learner) หรือแบบจำลองมาเรียนรู้ข้อมูลที่สุ่มออกมาโดยทำการสุ่มข้อมูลหลายๆ ครั้งเพื่อให้ได้แบบจำลองหลายๆ ตัวสำหรับการทำนายแบบจำลองทุกตัวที่สร้างขึ้นมาเพื่อทำนายชุดข้อมูลใหม่ที่เจอการทำนายสามารถมีได้หลายแบบตั้งแต่การใช้ค่าเฉลี่ย ไปจนถึงการโหวต ซึ่งนอกจากการสุ่มข้อมูลแล้ว Bagging ยังสามารถสุ่มคุณสมบัติ (Features) ได้อีกด้วย เนื่องจากการใช้แบบจำลองหลายๆ ตัวมาช่วยในการทำนายทำให้ลดการเกิด Overfitting อีกทางหนึ่งเช่นกัน แบบจำลองที่อยู่ในวิธีการนี้ได้แก่ Random Forest

Boosting เป็นการนำตัวเรียนรู้ที่มีความแม่นยำต่ำ (Weak Learner) มาทำนายข้อมูลที่มีจากนั้นจะให้ตัวเรียนรู้ตัวใหม่มาแก้ไขค่าความผิดพลาดที่เกิดขึ้น โดยผลรวมของตัวเรียนรู้เดิมจะทำให้เกิดตัวเรียนรู้ใหม่ขึ้น เมื่อทำแบบนี้ไปเรื่อยๆ จะทำให้ได้แบบจำลองที่ดีที่สุดจากผลรวมของตัวเรียนรู้ ข้อเสียของวิธี Boosting คือต้องรันหลายครั้งและเป็นลำดับกว่าจะได้แบบจำลองที่ต้องการซึ่งต่างจากวิธี Bagging ที่สามารถสุ่มข้อมูลแล้วฝึกฝนแบบจำลองได้พร้อมกัน แบบจำลองที่จัดอยู่ในประเภทนี้ได้แก่ Extreme Gradient Boosting



ภาพที่ 2.8 ไดอะแกรมการทำงานของ Boosting และ Bagging

ที่มา https://tupleblog.github.io/bagging-boosting/

2.20 Extreme Gradient Boosting และ Random Forest [16]

Random Forest หมายถึง เป็นแบบจำลองที่นำต้นไม้ตัดสินใจ (Decision Tree) แบบ Regression Tree หากเป็นปัญหาแบบ Regression หลายต้นมาฝึกฝนรวมกันบนข้อมูลย่อยของชุดข้อมูลทั้งหมด โดยจะทำการแบ่งข้อมูลออกเป็นข้อมูลย่อยๆ เพื่อสร้างต้นไม้หลายต้นมาทำการเรียนรู้ข้อมูลแต่ละชุดที่ได้แบ่งไว้ นอกจากจะแบ่งชุดข้อมูลได้แล้วยังสามารถแบ่งคุณสมบัติ (Features) ในการฝึกฝนของแต่ละต้นไม้ได้อีกด้วย สำหรับการผลการทำนายจะทำการหาค่าเฉลี่ยของคำตอบที่แต่ละต้นไม้ตอบออกมา วิธีนี้สามารถเพิ่มความแม่นยำในการทำนายผลและป้องกันการเกิด Overfitting ได้ดี

Extreme Gradient Boosting หมายถึง เป็นแบบจำลองที่นำต้นไม้ตัดสินใจแบบ Regression Tree หากเป็นปัญหาแบบ Regression มาฝึกฝนแบบจำลองต่อกันหลายต้น โดยการฝึกฝนนั้นจะมีการฝึกฝนเป็นลำดับต่อกัน แต่ละต้นไม้จะเรียนรู้ค่าความผิดพลาดจากต้นไม้ก่อนหน้าทำให้ความแม่นยำในการทำนายสูงมากขึ้นเรื่อยๆ โดยเฉพาะเมื่อมีการเรียนรู้ของต้นไม้ที่ความลึกที่เหมาะสมและแบบจำลองจะทำการหยุดฝึกฝนเมื่อค่าความผิดพลาดจากต้นไม้ก่อนหน้าหมดลง ทุกการเรียนรู้จะใช้ Negative Gradient ของ Loss Function ที่ได้กำหนดให้กับต้นไม้

ความแตกต่างระหว่าง Random Forest และ Extreme Gradient Boosting ได้แก่ วิธีที่ใช้ในการเรียนรู้แม้ทั้งสองจะเป็นการสร้างแบบจำลองจากการใช้ต้นไม้หลายๆ ต้นเหมือนกัน แต่วิธีที่ใช้ในการเรียนรู้นั้นต่างกัน Extreme Gradient Boosting จะสร้างต้นไม้หลายต้นแล้วทำการฝึกฝนแบบเป็นลำดับเพื่อเรียนรู้ความผิดพลาดจากต้นไม้ต้นก่อนหน้า ซึ่งวิธีการอยู่ในประเภท Boosting ส่วน Random Forest นั้นจะสร้างต้นไม้แล้วทำการเรียนรู้ข้อมูลที่ถูกแบ่งแต่ละส่วนแยกกันจากนั้นนำคำตอบมาเฉลี่ยเพื่อหาผลลัพธ์ที่ดีที่สุดเป็นวิธีการที่อยู่ในประเภท Bagging ดังภาพที่ 2.8

2.21 Training Set, Validation Set และ Test Set

ในขั้นตอนก่อนการนำข้อมูลเข้าไปฝึกฝนแบบจำลองนั้นต้องทำการแบ่งชุดข้อมูลออกเป็นส่วน ส่วนใหญ่จะแบ่งเป็น 2 ส่วนหลักๆ ได้แก่ ชุดข้อมูลสำหรับการฝึกฝน (Training Set) และ ชุดข้อมูลสำหรับทดสอบ (Test Set) ซึ่งอัตราส่วนในการแบ่งข้อมูลนั้นก็ขึ้นอยู่กับจำนวนของข้อมูล หากจำนวนข้อมูลทั้งหมดมีมาก ก็อาจจะใช้อัตราส่วน 60:40 โดย 60 ใช้สำหรับการฝึกฝนแบบจำลองและ 40 ใช้สำหรับทดสอบแบบจำลอง เหตุผลที่ต้องแบ่งข้อมูลออกเป็นสองส่วนเพราะหากทำการฝึกฝนแบบจำลองด้วยข้อมูลทั้งหมดแล้วนั้นจะไม่สามารถประเมินแบบจำลองที่ผ่านการฝึกฝนว่ามีประสิทธิภาพดีแค่ไหน ส่วนใหญ่จะใช้อัตราส่วน 80:20 ทั้งนี้ทั้งนั้นก็ขึ้นอยู่กับจำนวนข้อมูลที่มีเช่นกัน โดยข้อมูลทั้งสองส่วนไม่ควรที่จะน้อยเกินไปเนื่องจากอาจทำให้เกิดการ Overfitting ขึ้นได้ แต่การแบ่งข้อมูล 2 ส่วนอาจเกิดปัญหาขึ้นได้เช่นกันเนื่องจากมีการนำข้อมูลทดสอบมาใช้ซ้ำๆ แล้วทำการปรับพารามิเตอร์ตามข้อมูลทดสอบทำให้เกิด Overfitting ได้เช่นกัน จึงมีแนวคิดอีกวิธีหนึ่งให้ทำการแบ่งข้อมูลออกเป็น 3 ส่วน ได้แก่ ข้อมูลฝึกฝน (Training Set), ข้อมูลตรวจสอบ (Validation Set) และ ข้อมูลทดสอบ (Test set) โดยจะใช้ชุดข้อมูลตรวจสอบในการวัดประสิทธิภาพของแบบจำลอง และใช้ข้อมูลทดสอบในการทดสอบกับแบบจำลองที่ผ่านการปรับพารามิเตอร์เสร็จแล้วเพียงครั้งเดียว

2.22 Python

Python เป็นภาษาที่ใช้ในการเขียนโปรแกรมภาษาหนึ่งซึ่งเป็นภาษาระดับสูงถูกสร้างขึ้นมาจากภาษาซี โดย Guido Van Rossum ในตัวภาษาให้ความสำคัญไปที่การอ่านโค้ด (Code) ได้ง่าย ลดความซับซ้อนของการเขียนโปรแกรมลง มีจุดมุ่งหมายเพื่อให้นักเขียนโปรแกรมเขียนโปรแกรมอย่างชัดเจน ตัวภาษารองรับการเขียนโปรแกรมได้ทั้งขนาดเล็กและขนาดใหญ่

นอกจากนี้ Python เป็นภาษาที่นิยมใช้ในปัจจุบันอย่างกว้างขวางเนื่องจาก สนับสนุนกระบวนทัศน์ของการเขียนโปรแกรมที่หลากหลาย ไม่ว่าจะเป็น การเขียนโปรแกรมเชิงวัตถุ (Object-Oriented Programming), การเขียนโปรแกรมเชิงโครงสร้าง (Structured programming) รวมถึง การเขียนโปรแกรมเชิงฟังก์ชัน (Functional Programming) นอกจากการเขียนกระบวนทัศน์ด้านบนแล้วยังมีการเขียนโปรแกรมอีกมากมายที่ Python รองรับทำให้มีคลังโปรแกรม (Library) มากมายให้สามารถเลือกใช้ได้ตามงานที่เหมาะสม [17]

2.23 TensorFlow

เป็นแฟลตฟอร์มที่เปิดให้ผู้ใช้งานด้านการเรียนรู้ของเครื่องได้ใช้งานฟรี ซึ่งมีความกว้างขวาง, มีระบบการใช้งานที่ยืดหยุ่นสำหรับเครื่องมือต่างๆ รวมถึงมีสมาคม องค์กรต่างๆ มีการให้นักวิจัยได้ทำการสร้างหรือพัฒนาแบบจำลองของตัวเองขึ้นมาซึ่งผู้พัฒนาสามารถสร้างได้อย่างง่ายดาย และสามารถนำไปผนวกเข้ากับแอปพลิเคชันได้อย่างมีประสิทธิภาพ [18]

ข้อดีของ TensorFlow

- ง่ายสำหรับการสร้างแบบจำลอง สามารถสร้างและฝึกฝนโดยใช้ APIs ระดับสูงอย่าง Keras หรืออื่นๆ ในการทำงานได้นอกจากนี้ยังง่ายสำหรับการตรวจสอบข้อผิดพลาดของแบบจำลอง (Debugging)
- สามารถสร้างแบบจำลองได้ทุกที่ เนื่องจากตัว TensorFlow สามารถใช้งานร่วมกับคลาว์ (Cloud) สามารถใช้งานบนเบราว์เซอร์ (Browser) หรือ ทำงานบนอุปกรณ์ส่วนตัวก็ได้ตามความสะดวกของผู้ใช้งาน
- มีความยืดหยุ่นในการเขียนโปรแกรม โดยสามารถเขียนโปรแกรมแบบใหม่ จากคอนเซ็ปเดิมและสามารถเผยแพร่ได้อย่างรวดเร็ว
- ผู้เริ่มต้นทำงานด้านการเรียนรู้ของเครื่องสามารถเรียนรู้วิธีการทำงานได้อย่างรวดเร็ว เนื่องจากตัว TensorFlow มีการใช้งานที่ง่าย และไม่ซับซ้อน

2.24 VIIRS I-Band 375 m Active Fire Data

Visible Infrared Imaging Radiometer Suite (VIIRS) 375 m thermal anomalies/active fire product เป็นผลผลิตสุดท้ายที่ถูกเพิ่มเข้ามาโดย Fire Information of Resource Management System (FIRMS) ซึ่งเป็นองค์กรที่ถูกจัดตั้งขึ้นเพื่อจัดการเกี่ยวกับไฟขององค์กรนาซ่า (The National Aeronautics and Space Administration: NASA) โดยได้รับข้อมูลผ่านตัวรับข้อมูลจากอุปกรณ์ของ VIIRS ที่ถูกติดตั้งอยู่บนดาวเทียม Suomi National Polar-Orbiting Partnership (Suomi NPP) และ ดาวเทียม NOAA-20 สามารถตรวจจับไฟข้อมูลโดยมีความระเอียด 375 เมตร ซึ่งมีความละเอียดสูงกว่า อุปกรณ์ Moderate Resolution Imaging Spectroradiometer (MODIS) สามารถตรวจจับไฟในพื้นที่ที่มีขนาดเล็กได้และยังพัฒนาประสิทธิภาพในการตรวจจับไฟในตอนกลางคืนอีกด้วย [19]

จากข้อมูลให้มามีคุณสมบัติทั้งหมดตามตารางที่ 2.7 ดังนี้

ตารางที่ 2.7 ความหมายของข้อมูล FIRMS

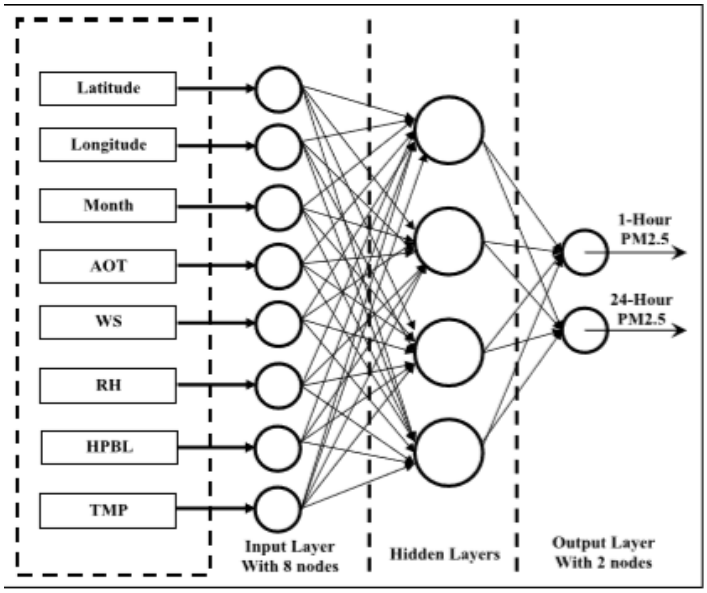
Attribute	Short Description	Long Description
Latitude	Latitude	ศูนย์กลางของค่าไฟในหน่วย pixel
Longitude	Longitude	ศูนย์กลางของค่าไฟในหน่วย pixel
Bright_ti4	Brightness Temperature 1-4	หมวด 1 – 4 VIIRS ที่ใช้สำหรับการตรวจจับอุณหภูมิในหน่วย Kelvin
Scan	Along Scan Pixel Size	ประมาณค่าที่ต่ำสุดเพื่อใช้ในการสะท้อนถึงค่าพิกเซลล์จริง
Track	Along Track Pixel Size	ประมาณค่าที่ต่ำสุดเพื่อใช้ในการสะท้อนถึงค่าพิกเซลล์จริง
Acq_Date	Acquisition Date	วันที่ตรวจพบ
Acq_Time	Acquisition Time	เวลาที่ตรวจพบ
Satellite	Satellite	ดาวเทียมที่ใช้ในการตรวจจับ
Confidence	Confidence	เป็นค่าความมั่นใจของจุดไฟที่ตรวจพบ
Version	Version (Collection and Source)	เวอร์ชันที่บ่งบอกถึงการเก็บข้อมูลหรือแหล่งของข้อมูล
Bright_ti5	Brightness Temperature 1-5	หมวด 1 – 5 VIIRS ที่ใช้สำหรับการตรวจจับอุณหภูมิในหน่วย Kelvin
FRP	Fire Radiative Power	พลังงานที่แผ่ออกมาของไฟที่ตรวจจับได้ในหน่วยเมกะวัตต์ (megawatts)
DayNight	Day or Night	D = กลางวัน, N = กลางคืน

ในส่วนต่อไปผู้วิจัยจะได้กล่าวถึงงานวิจัยที่ผ่านมาที่เกี่ยวข้องกับปัญหา $PM_{2.5}$ เพื่อทำการศึกษาแต่ละงานวิจัยว่ามีแนวทางในการแก้ไขปัญหาอย่างไร และเพื่อนำความรู้จากการศึกษางานวิจัยข้างต้นมาใช้เป็นต้นแบบของงานวิจัย โดยงานวิจัยทั้งหมดที่ได้ทำการศึกษามีหัวข้อดังต่อไป

2.25 Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach.

Pawan Gupta และ Sundar A. Christopher จาก [20] ได้ทำงานวิจัยเกี่ยวกับการประมาณค่า $PM_{2.5}$ โดยงานวิจัยนี้เป็นการนำโครงข่ายประสาทมาใช้ในการลดความไม่แน่นอนของการประมาณค่า $PM_{2.5}$ จากข้อมูลดาวเทียมโดยใช้ข้อมูลค่าความลึกเชิงแสงของฝุ่นละออง (Aerosol Optical Depth, AOD) ที่ 0.55 ไมโครเมตร มีความละเอียดเท่ากับ 10 × 10 กิโลเมตรที่ทำการเก็บสะสมมาจากระบบ MODIS เป็นระยะเวลาทั้งหมด 3 ปี และข้อมูลทางด้านอุตุนิยมวิทยาสำหรับการพัฒนาแบบจำลองในการประมาณ $PM_{2.5}$ ณ ระดับพื้นผิว บนภาคตะวันออกเฉียงใต้ของสหรัฐอเมริกา โดยได้เปรียญเทียบค่าสัมประสิทธิ์การถดถอยที่ได้จาก สหสัมพันธ์อย่างง่าย (R = 0.6) และ การถดถอยพหุคูณ (R = 0.68) โดยผลที่ได้จากการใช้โครงข่ายประสาทเทียมในการประมาณ PM2.5 ค่าที่ได้จากแบบจำลองโครงข่ายประสาทเทียมรายชั่วโมงเปรียบเทียบกับการค่าจริง R = 0.74 ส่วนค่าที่ได้จากแบบจำลองโครงข่ายประสาทเทียมแบบรายวันเปรียบเทียบกับค่าจริง R = 0.78

ในโครงสร้างของแบบจำลองโครงข่ายประสาทเทียมที่งานวิจัยนี้ใช้ประกอบด้วย Input Layer, Hidden Layer และ Output Layer ตัว Input Layer ประกอบด้วยข้อมูลนำเข้า 8 ตัวด้วยกัน ได้แก่ ชื่อ, ละติจูด, ลองติจูด, เดือน, ค่าความลึกเชิงแสงของฝุ่นละออง, ความไวลม, ค่าความชื้น, ค่าความสูงของชั้นบรรยากาศ และ ค่าอุณหภูมิ โดยสามารถนำข้อมูลทั้งหมดนี้มาสร้างเป็นโครงข่ายประสาทเทียมได้ดังนี้



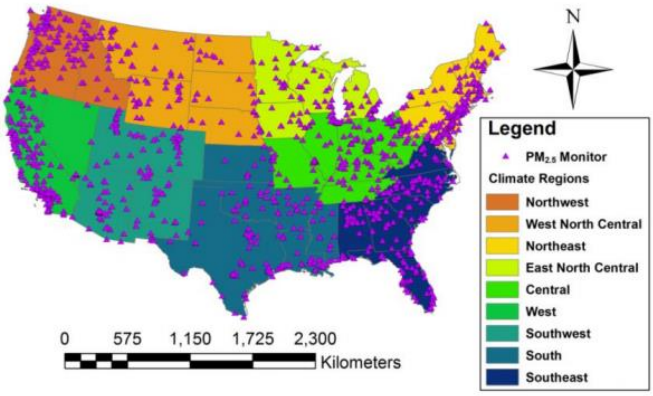
ภาพที่ 2.9 โครงสร้างของโครงข่ายประสาทเทียมสำหรับการประมาณค่า $PM_{2.5}$

2.26 A Bayesian Downscaler Model to Estimate Daily $PM_{2.5}$ levels in the Continental Us

Yikai Wang และ Xuefei Hu จาก [21] ได้ทำงานวิจัยชิ้นนี้เป็นงานวิจัยที่มีพื้นที่ศึกษาบนประเทศสหรัฐอเมริกา โดยใช้เทคนิค Bayesian Downscaling ซึ่งข้อมูลที่ใช้ได้แก่

- ข้อมูล $PM_{2.5}$ เฉลี่ย 24 ชั่วโมง ในปี 2011
- ข้อมูล Level 2 Aqua Modis ที่ความละเอียด 12×12 ตารางกิโลเมตร
- คำนวณค่าเฉลี่ยค่า AOD ด้วยวิธี Deep-Blue และ Dark-target
- ข้อมูลอุตุนิยมวิทยาที่ความละเอียดเชิงพื้นที่ 32 กิโลเมตรและเป็นข้อมูลรายสามชั่วโมง
- ความสูงจากระดับน้ำทะเลที่ความละเอียด 30 เมตร
- ข้อมูลพื้นดินบนอเมริกาเหนือเป็นความละเอียดเชิงพื้นที่ 13 กิโลเมตรและเป็นข้อมูลรายชั่วโมง
- ข้อมูลเปอร์เซ็นต์ของป่าที่ปกคลุมที่ความละเอียดเชิงพื้นที่ 30 m
- ข้อมูลถนน
- กิจกรรมหลักที่ปล่อย $PM_{2.5}$

เพื่อทำให้การคำนวณมีประสิทธิภาพที่เพิ่มมากขึ้นจึงทำการแบ่งพื้นที่ทั้งหมดออกเป็น 9 เขต ได้แก่ ตะวันออกเฉียงเหนือ, ตะวันออกเฉียงใต้, ทิศใต้, เขตโอไฮโอ วัลเลย์ (Ohio Valley), ส่วนบนของมิดเวสต์ (Midwest), ทิศเหนือของร็อกกี้ (Rockies) และ พื้นที่ราบ,ตะวันตกเฉียงใต้, ตะวันตกเฉียงเหนือและตะวันตก



ภาพที่ 2.10 พื้นที่ของประเทศสหรัฐอเมริกาที่ถูกแบ่งเป็น 9 เขต

มีการเพิ่มพื้นที่เก็บข้อมูลสำหรับส่วนที่มีภูมิอากาศทับซ้อนกัน 100 กิโลเมตร นอกจากนี้ในแต่ละพื้นที่ยังมีรูปแบบสภาพอากาศที่หลากหลายตลอดทั้งปี เพื่อความยืดหยุ่นของแบบจำลองจึงแบ่งช่วงเวลาออกเป็น 3 ส่วน ส่วนละ 4 เดือน และใช้ Bayesian Downscaling กับแต่ละช่วงเวลา

โดยงานวิจัยนี้ได้ให้สมการการประมาณค่า $PM_{2.5}$ ได้ดังนี้

$$PM_{2.5}(s,t) = \alpha_0(s,t) + \alpha_1(s,t)AOD(s,t) + \gamma_{reg,tem}(s,t)z(s,t) + \varepsilon(s,t)$$

โดยที่

$\alpha_0(s,t)$, $\alpha_1(s,t)$ คือ Random Intercept และ Slope

$\varepsilon(s,t)$คือ Residual Error ขึ้นอยู่กับ Mean Zero, พื้นที่และเวลา

z(s,t) แสดงความแปรปรวนของตัวแปรที่เกี่ยวข้องกับ $PM_{2.5}$ ที่ $\gamma_{reg,tem}(s,t)$ แสดงสำหรับพื้นที่และเวลาที่ระบุเป็นผลกระทบที่ขึ้นอยู่ระหว่าง z(s,t) และ $PM_{2.5}$ (s,t) รวมถึง ไฟป่า พื้นที่ป่า ความชื้น อุณหภูมิ ความไวลม ความยาวถนนเส้นหลัก ความหนาชั้นบรรยากาศ AOD

Random effect $\alpha_0(s,t)$,$\alpha_1(s,t)$ ใช้ $\alpha_i(s,t) = \beta_i(s) + \beta_i(t)$; i=0,1

$\beta_i(s)$ โครงสร้างที่ซ่อนไว้ใช้ Gaussian process $W_1(s)$ and $W_2(s)$

$$\beta_0 = c_1 W_1(s)$$
$$\beta_1 = c_2 W_1(s) + c_3 W_2(s)$$

$\beta_0(t)$,$\beta_1(t)$ เป็นตัวแปรที่มีเวลาเข้ามาเกี่ยวข้องใช้ First Order Random Walk ในการกำหนดเงื่อนไขการกระจายของวันที่จะนำไปกำหนดแก่วันอื่นๆ

ในการทำ Model fitting ใช้ Markov Chain Monte Carlo (MCMC) และในการวัดประสิทธิภาพใช้ 2 Cross Validation ได้แก่ Fully-Random Cross-Validation (R-CV) และ Spatial Cross-Validation (S-CV) สำหรับ R-CV จะทำการสุ่มแบ่งข้อมูลเป็น 10 ส่วน ใช้ 9 ส่วนในการฝึกฝนโมเดลและใช้ส่วนที่เหลือในการประมาณโมเดล เพื่อเป็นการทดสอบสำหรับการทำนายภาพรวมของโมเดล ส่วน S-CV จะทำการแบ่งข้อมูลตามพื้นที่โดยผลลัพธ์เพื่อประมาณความสามารถในการคาดการณ์เชิงพื้นที่ นอกจากนี้การใช้ MCMC ยังทำให้สามารถทำนายความไม่แน่นอนในเชิงตัวเลขได้ พร้อมกับเปรียบเทียบผลที่ได้กับ $PM_{2.5}$ ที่วัดมาได้ โดย Root Mean Squared Error , Empirical Coverage Probability, Linear Coefficient of Determination R^2 การ Cross-Validation ภาพรวมของ R^2 ทั้งพื้นที่ที่สนใจและช่วงเวลาที่สนใจ คือ 0.7 และความชันได้ 0.98 ในปี 2011 สำหรับ S-CV ที่ทำ Cross-Validation ในพื้นที่ต่างๆภายใต้ R^2 ที่สูงที่สุดสำหรับ Cross-Validation Setting (Complete = R^2 0.78 and Spatial R^2= 0.7) ได้แก่ พื้นที่ตะวันตกเฉียงเหนือ มิดเวสต์ตอนบน โอไฮโอวัลเลย์ ส่วนพื้นที่ทางทิศใต้ได้ R^2 = 0.54 ต่ำกว่า Complete 10-Fold Setting นอกจากนี้ตัว Spatial Cross Validation Setting ทำงานได้ไม่ดีเท่า Complete Cross Validation Setting สำหรับงานวิจัยนี้มีแบบจำลองมีความยืดหยุ่นเนื่องจากมีการแยกพื้นที่ตามสภาพอากาศในแต่ละพื้นที่และในแต่ละช่วงเวลา ในพื้นที่ที่ไม่มีข้อมูลก็ใช้การยืมข้อมูลจากพื้นที่ใกล้เคียงและสามารถจัดการกับความไม่แน่นอนต่างๆได้ดี ส่วนข้อเสีย

คือแบบจำลองมีความแม่นยำน้อยกว่าแบบจำลองอื่นๆ และในบางพื้นที่ที่ได้ทำการแบ่งเพื่อจำแนกสภาพอากาศมีการแบ่งพื้นที่ที่ใหญ่เกินไปทำให้ความเป็นไปได้ในพื้นที่นั้นกว้างไปด้วย

2.27 Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC Satellite-Based AOD Data

งานวิจัยนี้เป็นของ Itai Kloog และ Meytar Sorek-Hamer จาก [22] มีจุดมุ่งหมายเพื่อตรวจสอบการใช้งาน MODIS-Based MAIAC Data ในประเทศอิสราเอลและสำรวจความน่าเชื่อถือของการทำนาย $PM_{2.5}$ และ PM_{10} ในการใช้ความละเอียดสูงที่ระดับ 1 กิโลเมตร เนื่องจากประเทศอิสราเอลมีความแตกต่างกันทางด้านภูมิศาสตร์ และพื้นที่ภูมิอากาศ เช่น มีพื้นที่การเกษตรขนาดใหญ่, พื้นที่ป่าขนาดใหญ่, ส่วนที่เป็นน้ำ, ภูเขา และที่ราบชายฝั่ง ข้อมูลที่ใช้ได้แก่

- MODIS Aerosol Product ได้แก่ Spectral AOD โดยใช้ Product Level 2 ที่สามารถเข้าถึงข้อมูลได้ทั้งโลก และใช้วิธีการ Deep-Blue ในการคำนวณค่า AOD ที่ความละเอียด 10 กิโลเมตร แต่ว่าในงานวิจัยนี้จะใช้วิธี MAIAC ด้วยเป็นวิธีการใหม่ในการคำนวณค่า AOD ที่ถูกพัฒนาขึ้นโดย Lyapustin and Colleagues โดยให้ความละเอียดที่ 1 กิโลเมตร วิธี MAIAC จะใช้งานได้ดีกว่าในพื้นที่ที่มีความสว่าง โดยใช้ข้อมูล จากดาวเทียมที่มีชื่อว่า Aqua จากองค์กร Nes Ziona AERONET มีค่าสหสัมพันธ์ (Correlation) ระหว่าง MAIAC AOD และ AERONET AOD เท่ากับ 0.85 ซึ่งถูกคำนวณจากข้อมูลรายชั่วโมง งานวิจัยนี้มีการเติม Flag ให้กับข้อมูลโดยใช้ QA คือ Quality Assurance และ UN คือ Uncertainly เพื่อใช้การกรองค่าที่เป็นปัญหาซึ่งเกิดจากพื้นที่ที่มีความสว่าง ได้แก่ พื้นผิวที่เป็นหิมะ เป็นต้น

- มีการเก็บข้อมูล $PM_{2.5}$ และ PM_{10} ตั้งแต่ปี 2003-2013 จาก TEOM โดยมีความแม่นยำ ±5% เนื่องจากพื้นที่บางพื้นที่ไม่ได้มีการเก็บข้อมูลที่ต่อเนื่อง

- ข้อมูลที่ใช้ในการทำนายเชิงพื้นที่ ได้แก่ ความหนาแน่นของประชากร (Population Density), ความสูงจากระดับน้ำทะเล (Elevation), ความหนาแน่นของการจราจร (Traffic Density), ระยะทางจากถนนหลัก (Distance from Major Road), ระยะทางจากชายฝั่งทะเล (Distance from Shoreline), เปอร์เซ็นต์ของพื้นที่ที่เป็นที่เปิดโล่ง ปัจจัยที่ถูกเลือกมาทั้งหมดนี้เพราะว่าสามารถนำมาทำการคำนวณหาการปล่อยของอนุภาคต่างๆได้

ข้อมูลที่ใช้ในการทำนายเชิงเวลาที่ ได้แก่

1.ข้อมูลที่เกี่ยวกับอุตุนิยมวิทยาคือ อุณหภูมิของอากาศ ความชื้น และปริมาณฝนตก

2.การจำแนกฝุ่นเพื่อที่จะรู้ว่ามีกิจกรรมใดที่ฝุ่นส่งผลกระทบบ้าง โดยจะทำการเติมหมายเลขให้ถ้าได้รับผลกระทบโดยฝุ่น จะได้หมายเลย 1 แต่ถ้าไม่จะได้หมายเลข 0

3.NDVI ใช้ Monthly MODIS NDVI Product ที่ความละเอียดเชิงพื้นที่ 1 กิโลเมตร

4. PBL เป็นค่ารายวันของความสูงของชั้นบรรยากาศ (Planetary Boundary Layer) ใช้ความละเอียดเชิงพื้นที่ที่ 50 กิโลเมตร

งานวิจัยนี้นำข้อมูลทั้งหมดไปสร้างในรูปแบบของระบบกริดเช่นกัน โดยในช่องที่ MAIAC AOD หายไป จะไม่ใช้การสุ่ม ปัญหาที่เป็นไปได้ ได้แก่ พายุฝุ่น หิมะ หรือ เมฆที่ปกคลุมมากเกินไป และการตรวจสอบความถูกต้อง (Calibration Process) จะใช้เทคนิค Inverse Probability Weighting (IPW) เพื่อหลีกเลี่ยงการเกิด Bias ในแบบจำลอง Regression

IPW= $\frac{1}{P}$; p คือความน่าจะเป็นของ AOD ที่สามารถนำมาใช้งานได้ในแต่ละวัน

กระบวนการทำงานมีสามขั้นตอนด้วยกัน ได้แก่

ขั้นตอนแรก จะทำการ Fit Daily Calibration โดยการใช้ ข้อมูลจากในจุดภาพ (Pixel)

$PM_{2.5}$ = AOD + Other Spatio-Temporal Predictors ใช้ Linear Regression

ขั้นตอนที่สอง จะทำการ ใช้ Calibration Model ในการ Fit เพื่อทำนาย $PM_{2.5}$ ในแต่ละกริดโดยใช้ AOD ที่สามารถเข้าถึงได้เท่านั้น

$$PM_{ij} = (\alpha + u_j + g_{j(reg)}) + (\beta_1 + v_j + h_{j(reg)})AOD_{ij} + \sum_{m=1}^{6} X_{1mi} + \sum_{m=1}^{6} X_{2mj} + \varepsilon_{ij}$$

$$(u_j v_j k_j) \sim N[(0,0), \sigma]$$

$$(g_{j(reg)} h_{j(reg)}) \sim N[(0,0), \sigma_{reg}]$$

แบบจำลองด้านบนเป็นแบบจำลองที่ใช้ในการคำนวณเพื่อทำนายค่า $PM_{2.5}$ และ PM_{10} โดย

PM_{ij} เป็นค่า $PM_{2.5}$, PM_{10} ที่พื้นที่ i และวันที่ j

α เป็นค่าคงที่ที่ตั้งค่าไว้

u_j เป็นค่าสุ่มสกัดกั้น (Random Intercept)

AOD_{ij} เป็นค่า AOD ที่พื้นที่ i และวันที่ j

β_1, v_j เป็นค่าที่ตั้งไว้และเป็นค่าของการสุ่มความชันตามขึ้นอยู่กับวันที่ระบุ

X_{1mi} คือ ข้อมูลที่ใช้ในการทำนายเชิงพื้นที่ (m-th Spatial Predictor)

X_{2mj} คือ ข้อมูลที่ใช้ในการทำนายเชิงเวลา (m-th Temporal Predictor)

$g_{j(reg)}$ คือ Daily Random Intercept

$h_{j(reg)}$ คือ ความชันของค่า AOD ในแต่ละพื้นที่

σ เป็น เมทริกซ์แนวทะแยงขนาด 3 $\times$ 3 ที่ประกอบด้วย $\sigma^2_U, \sigma^2_V, \sigma^2_K$

σ_{reg} เป็น เมทริกซ์แนวทะแยงขนาด 2 $\times$ 2 ที่ประกอบด้วย σ^2_g, σ^2_h

ขั้นตอนสุดท้าย จะเป็นการประมาณค่า $PM_{2.5}$ ในช่องที่ AOD ไม่สามารถอ่านค่าได้โดยจะทำการใช้ Spatial Smoothingของกริดเพื่อนบ้านมาคำนวณ

$$PredPM_{ij} = (\alpha + u_j) + (\beta_1 + v_j)MPM_{ij} + s(X_i, Y_i)_{k(j)} + \varepsilon_{ij}$$
$$(u_j v_j) \sim N[(0,0), \Omega_\beta]$$

$PredPM_{ij}$ คือ $PM_{2.5}$, PM_{10} ในกริดที่ i และ วันที่ j

MPM_{ij} คือ ค่าเฉลี่ยของ PM ในระยะ 30 กิโลเมตร ของพื้นที่ i และวันที่ j

α เป็นค่าคงที่ที่ตั้งค่าไว้

u_j เป็นค่าสุ่มสกัดกั้น(Random Intercept)

β_1, v_j เป็นค่าที่ตั้งไว้และเป็นค่าของการสุ่มความชันตามขึ้นอยู่กับวันที่ระบุ

X_i, Y_i คือ ละติจูด และ ลองติจูด ของจุดศูนย์กลางของกริดที่ i

$s(X_i, Y_i)_{k(j)}$ คือ Smooth Function ของตำแหน่งที่ระบุไปจนถึง Bi-Monthly Period k(j)ใน วันที่ j

โมเดลประมาณความแม่นยำโดยใช้ Square Root of The Mean Squared Prediction Errors (RMSPE) เมื่อทำการ Cross-Validation แล้วให้ผลลัพธ์ดังนี้ R^2 = 0.79, Slop = 0.99, RMSPE = 25.10 μg/m^3 สำหรับ PM_{10} ส่วน $PM_{2.5}$ ผลลัพธ์ดังนี้ R^2 = 0.72, Slop = 0.99, RMSPE = 8.52 μg/m^3 สำหรับข้อจำกัดตัวโมเดลมีความลำเอียงเล็กน้อย(Bias) เนื่องจากมีค่าความชันถึง 0.99 และสำหรับความละเอียดที่ 1 กิโลเมตรที่ใช้บางครั้งอาจจะยังไม่สามารถจับผลกระทบที่เกิดขึ้นในแหล่งชุมชนได้ เช่น ถนนไฮเวย์ที่มีความละเอียดน้อยกว่า 1 กิโลเมตร

2.28 Estimating Ground-Level PM2.5 in China using Satellite Remote sensing

งานวิจัยนี้ถูกพัฒนาโดย Zongwei Ma และ Xuefei Hu จาก [23] ซึ่งพัฒนาแบบจำลองทางสถิติชื่อว่า National Scale Geographically Weighted Regression (GWR) สำหรับการประมาณค่า $PM_{2.5}$ รายแบบรายวันในประเทศจีนใช้ข้อมูล AOD จากดาวเทียมเป็นส่วนประกอบหลัก โดยในงานวิจัยนี้ใช้ข้อมูลจากดาวเทียม Terra และ Aqua

ข้อมูลที่ใช้ได้แก่

- เป็นข้อมูลเฉลี่ยรายวันของ $PM_{2.5}$ ที่ถูกวัดได้ ตั้งแต่วันที่ 22 ธันวาคม 2012 จนถึง 30 พฤศจิกายน 2013

- AOD Retrieval and Calibration ใช้ MODIS Level 2 Product โดยงายวิจัยนี้บอกว่าข้อมูลนี้จะมีผลประมาณ 1-2 วัน ใช้ MODIS Retrieves AOD ที่ความละเอียด 10 กิโลเมตร โดยข้อมูลก่อนปี 2013 ใช้สำหรับการ Calibration ส่วนข้อมูลหลัง 22 ธันวาคม 2013 ใช้สำหรับการสร้างโมเดล

สำหรับ AOD ที่ใช้ Parameter ชื่อว่า Image_Optical_Depth_Land_and_Ocean ที่ 550 นาโนเมตร ใช้ข้อมูล MISR level 2 Aerosol Product โดยมีความละเอียดของค่า AOD เชิงพื้นที่ 17.6 กิโลเมตร ข้อมูลนี้มีผลประมาณ 2- 9 วัน ข้อมูลที่เหมาะสมอยู่ที่ 558 นาโนเมตร ที่มีค่า Chi-Square น้อยที่สุด การเทียบข้อมูล AOD เพื่อการใช้งานข้อมูลที่ดีขึ้นจึงทำการรวมข้อมูล Aqua MODIS, Terra MODIS and MISR AOD เข้าด้วยกันเนื่องจากข้อมูลทั้งสองมีช่วงค่าที่แตกต่างกัน จึงต้องทำการ Calibration แยกกันโดยใช้ข้อมูลจาก Aerosol Robotic Network (Aeronet) ที่ AOD 550 นาโนเมตร ในการ Validate ข้อมูล MODIS และ MISR

- ข้อมูลที่เกี่ยวกับอุตุนิยมวิทยา ได้จาก GEOS-5 โดยใช้ข้อมูลรายชั่วโมงหรือ 3ชั่วโมง ค่าที่ใช้ได้แก่ ค่าเฉลี่ยของขอบเขตชั้นความสูง (Planetary Boundary Layer Height (PBLH, m)), อุณหภูมิที่ความสูง 2 เมตร, ความไวลม (WS, m/s) ที่ความสูง 10 เมตร ค่าความชื้นในขอบเขตชั้นความสูง (RH_PBLH, %) และ ความดันของพื้นที่ (PS, hPa) ระหว่าง 10 และ 11 โมงตอนเช้า

- ข้อมูลส่วนที่เป็น พื้นที่ปกคลุมของพื้นดิน และ ข้อมูลประชากร ใช้ข้อมูลรายเดือนของค่า NDVI ที่ความละเอียด 0.25 × 0.25 และ ข้อมูลความหนาแน่นของประชากรแบบรายเดือน

ข้อมูลทั้งหมดจะรวมอยู่ในกริด โดยใช้ความละเอียดที่ 50 กิโลเมตร และข้อมูล AOD ถูกรวมให้เหลือค่า AOD เดียวเป็นข้อมูลนำเข้า งานวิจัยนี้ได้พัฒนาโมเดล GWR โดยตัวโมเดลนี้สามารถสร้างพื้นที่ต่อเนื่องของตัวแปรที่รับเข้ามาโดยทำการวัดตัวแปรที่รับเข้าในแต่ละพื้นที่ที่สังเกตเพื่อกำหนดความหลากหลายเชิงพื้นที่

โดยสามารถแสดงเป็นสมการได้ดังนี้

$$PM_{2.5st} = \beta_{0,st} + \beta_{1,st}AOD_{st} + \beta_{2,st}PBLH_{st} + \beta_{3,st}T2M_{st} + \beta_{4,st}WS_{st} + \beta_{5,st}RH_PBLH_{st} + \beta_{6,st}PS_{st} + \beta_{7,st}POP_s + \beta_{8,st}NDVI_{sm} + \varepsilon_{st}$$

$PM_{2.5st}$ คือ ค่า $PM_{2.5}$ รายวันที่กริดช่องที่ s วันที่ t

$\beta_{0,st}$ คือ กำหนดให้เป็น location specific intercept on day

$\beta_{1,st}$-$\beta_{8,st}$ คือ Location specific slope on day

AOD_{st} คือ ค่า AOD ที่ทำการรวมข้อมูลมาแล้วในกริดช่องที่ s และวันที่ t

$PBLH_{st}$, $T2M_{st}$, WS_{st},$RH_{PBLH_{st}}$,PS_{st} คือ ข้อมูลที่เกี่ยวข้องกับอุตุนิยมวิทยาที่กริดช่องที่s และวันที่ t

POP_s คือ ประชากรรวมของกริดช่องที่ s

$NDVI_{sm}$ คือ ค่า NDVI ของกริดช่องที่ s และ เดือนที่ m

ε_{st} คือ ค่าความผิดพลาดที่จะเกิดขึ้นในกริดช่องที่ s และวันที่ t

ผลลัพธ์ Cross-validation ของโมเดลแบบเต็มรูปแบบ R^2 = 0.64 และ RMSE = 32.98 ไมโครกรัมต่อลูกบาศก์เมตร MPE = 8.28 ไมโครกรัมต่อลูกบาศก์เมตร ในงานวิจัยนี้ ค่า $PM_{2.5}$ มีค่าตั้งแต่ 1- 795 ไมโครกรัมต่อลูกบาศเมตร Cross-Validation Prediction Error กำหนดโดยการใช้ RMSE / Mean Ground $PM_{2.5}$ ของ Full Model อยู่ที่ 51.3% ตัวโมเดลค่อนข้างจะให้ผลลัพธ์ที่น้อยกว่าความเป็นจริงเมื่อค่า $PM_{2.5}$ มีค่ามากกว่า 60 ไมโครกรัมต่อลูกบาศก์เมตร

ข้อเสีย

- ยังมีบางพื้นที่มีความผิดพลาดเกิดขึ้น โดยเฉพาะในฤดูหนาว
- คำแนะนำควรใช้โมเดลที่มีความยืดหยุ่นแบบไม่ใช่ปัญหาเชิงเส้น (Non-Linear) ในการแก้ปัญหา
- ความแม่นยำยังถือว่ามีน้อยเมื่อเทียบกับแบบจำลองประเภทอื่นๆ

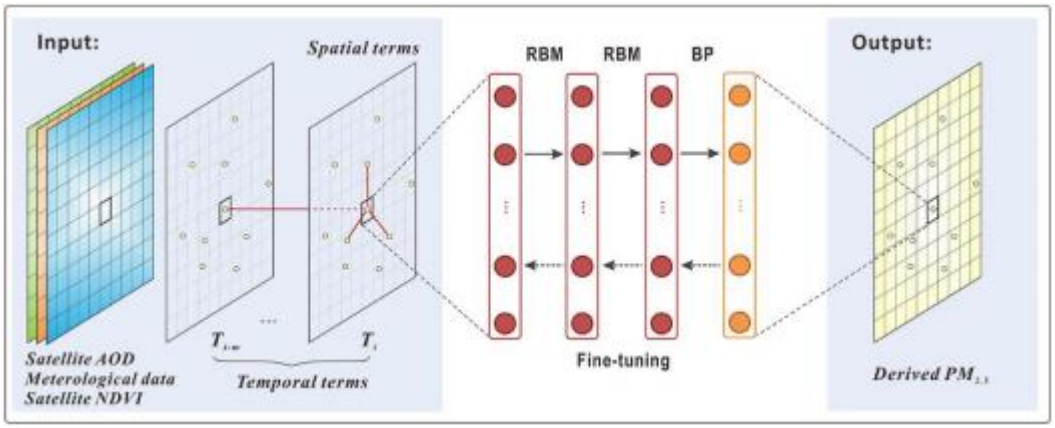
2.29 Estimating ground-level $PM_{2.5}$ fusing Satellite and station observations: A geo-intelligent deep learning approach

งานวิจัยชิ้นนี้ก็เป็นการนำความรู้ทางลักษณะภูมิศาสตร์ของ Tongwen Li และ Huanfeng Shen จาก [24] โดยนำมาใช้ร่วมการเรียนเชิงลึก (Deep Learning) ที่เรียกว่า Geoi-DBN เข้ามาใช้ในการประมาณค่า $PM_{2.5}$ โดยการนำข้อมูลระหว่างข้อมูลจากดาวเทียมและข้อมูลที่วัดได้ที่สถานีมาใช้รวมกันโดยมีพื้นที่ที่ให้ความสนใจในประเทศจีน โดยใช้ช่วงเวลาตั้งแต่ 1 มกราคม 2015 ไปจนถึง 31 ธันวาคม 2015

ข้อมูลที่ใช้แบ่งเป็นสี่ส่วนด้วยกันดังนี้

- ข้อมูลรายชั่วโมงของค่า $PM_{2.5}$ สำหรับปี 2015 โดยใช้ค่าเฉลี่ยรายชั่วโมงในการหาค่าเฉลี่ยรายวัน
- MODIS AOD โดยค่า AOD นี้ใช้จากดาวเทียมทั้งสองโมเดลได้แก่ Terra และ Aqua MODIS level 2 โดยใช้ความละเอียดที่ 10 กิโลเมตร ซึ่งได้จากการคำนวณของการใช้ Deep-Blue และ Dark Target อัลกอลิทึม โดยค่าเหล่านี้จะถูกนำมาใช้ในการประมาณค่า $PM_{2.5}$ แบบรายวัน
- ข้อมูลอุตุนิยมวิทยา จะใช้ค่าความชื้น (RH, %), อุณหภูมิที่ความสูง 2 เมตร (TMP, K), ความไวลมที่ความสูง 10 เมตร(WS, m/s), ความดันพื้นผิว (PS, Pa), ความสูงของชั้นบรรยากาศ (Planetary Boundary Layer Height (PBL, m))
- MODIS Normalized Difference Vegetation Index (NDVI) ข้อมูลส่วนนี้ในความละเอียด 1 กิโลเมคร สามารถเข้าถึงได้ทุกๆ 16 วัน

จากข้อมูลทั้งหมดจะต้องนำมารวมกันโดยจะประกอบด้วย ข้อมูลที่เป็นข้อมูลเกี่ยวกับเวลาและข้อมูลเชิงพื้นที่โดยรวมให้อยู่ในระบบกริด ที่ 0.1 องศา และปรับมาตรฐานของข้อมูลให้อยู่ในมาตรฐานเดียวกัน โครงสร้างของโมเดลจะประกอบไปด้วย Multiple Restricted Boltzmann Machine (RBN) Layers และ Back-Propagation (BP) Layers ดังภาพที่ 2.11



ภาพที่ 2.11 โครงสร้าง Geoi-DBN

ข้อมูลนำเข้าที่จะใช้ ได้แก่ ค่า AOD, ข้อมูลอุตุนิยมวิทยา, ค่า NDVI และข้อมูลที่ในรูปแบบเวลาและพื้นที่ เพราะว่า จุดที่ใกล้กับกริดที่ n ของการวัด $PM_{2.5}$ และ การสังเกตค่า $PM_{2.5}$ ในอดีตในจุดเดียวกันนั้นก็เป็นข้อมูลที่สำคัญสำหรับการประมาณค่า $PM_{2.5}$เช่นกัน

$$S - PM_{2.5} = \frac{\sum_{i=1}^{n} ws_i \, PM_{2.5,i}}{\sum_{i=1}^{n} ws_i} \qquad ws_i = \frac{1}{ds_i^2}$$

$$T - PM_{2.5} = \frac{\sum_{j=1}^{n} wt_j \, PM_{2.5,j}}{\sum_{j=1}^{n} ws_j} \qquad wt_j = \frac{1}{dt_j^2}$$

$$DIS = \min\left(\frac{1}{ds_i}\right) \; i = 1,2,3, \ldots\ldots, n$$

ค่า ds, dt หมายถึง ระยะทางเชิงพื้นที่ และ ระยะทางด้านเวลา

ค่า m = 3 , n = 10 และ DIS เป็นค่าที่ใช้เพื่อสะท้อนความหลากหลายของความไม่สม่ำเสมอของการกระจายสถานี

Geoi – DBN ใช้ 2 Hidden Layer และใช้จำนวนโหนดในแต่ละ Hidden layer คือ 15 โหนด และมี 1 Output Layer ที่เป็น BP ซึ่งมีโหนดเดียว โดยมีความสัมพันธ์ของ $PM_{2.5}$ ดังนี้

$$PM_{2.5} = f(AOD, RH, WS, TMP, PBL, PS, NDVI, S - PM_{2.5}, T - PM_{2,5}, DIS)$$

การทำงานแบ่งออกเป็น 3 ขั้นตอน

Pre-Training ใช้ข้อมูลที่เก็บรวบรวมมา RBMS จะทำการเทรนชั้นต่อชั้นโดยใช้วิธี Unsupervised Learning เพื่อทำการแยกลักษณะที่สำคัญที่เกี่ยวข้องกับ $PM_{2.5}$ แล้วส่งข้อมูลไปให้ RBM Layer ชั้นถัดไป ดังนั้น

Fine-Tuning Weight เริ่มต้นของ Geoi-DBN จะถูกสร้างขึ้นแล้วทำการคำนวณ $PM_{2.5}$ เปรียบเทียบกับ ค่า $PM_{2.5}$ดั้งเดิมที่วัดได้ แล้วคำนวณความผิดพลาดที่เกิดขึ้น จากนั้นจะมีการส่งค่าความผิดพลาดนั้นกลับไปยังโมเดลแล้วทำการปรับ Weight โดยการทำ Back Propagation

Prediction ขั้นตอนนี้เป็นการประมาณประสิทธิภาพของ Geoi-DBN Model โดยการป้อนข้อมูลนำเข้าที่บันทึกแล้วทำนายค่า $PM_{2.5}$ ของตำแหน่งเหล่านั้นโดยไม่มี ข้อมูลจากสถานีภาคพื้นดิน ในการประมาณประสิทธิภาพของโมเดลใช้ 10 Fold Cross-Validation ใช้ Correlation Coefficient (R), Root-Mean Square Error (RMSE, ไมโครกรัมต่อลูกบาศก์เมตร), Mean Prediction Error (MPE, ไมโครกรัมต่อลูกบาศก์เมตร) และ Relative Prediction Error (RPE , RMSE/Mean $PM_{2.5}$)

ผลลัพธ์ที่ได้ R = 0.94, RMSE = 13.68 ไมโครกรัมต่อลูกบาศก์เมตร MPE = 9.03 ไมโครกรัมต่อลูกบาศก์เมตร RPE = 24.90%

2.30 $PM_{2.5}$ Prediction Based on Random Forest,XGBoost, and Deep Learning using Multisource Remote Sensing Data

งานวิจัยนี้เป็นงานวิจัยจาก [25] เขียนโดย Mehdi Zamani Joharestani และคณะโดยมีจุดมุ่งหมายเพื่อหาแบบจำลองที่มีประสิทธิภาพสูงสุดในการทำนาย $PM_{2.5}$ และบอกคุณสมบัติที่มีผลต่อการทำนาย $PM_{2.5}$ โดยใช้การเรียนรู้ของเครื่องสามแบบด้วยกันในการทำนายค่า $PM_{2.5}$ ในเขตตัวเมือง Tehran's ประเทศอิตาลี ได้แก่ 1. Random Forest 2.Extreme Gradient Boost 3.Deep Learning ในการทำนายค่า $PM_{2.5}$ โดยอาศัยปัจจัยที่แบ่งออกเป็นสามส่วนด้วยกันได้แก่ ข้อมูลจากดาวเทียม ซึ่งได้แก่ ค่าความลึกเชิงแสงของฝุ่นละออง (Aerosol Optical Depth) ที่ใช้ข้อมูลจากดาวเทียม Aqua ความละเอียดขนาด 3 กิโลเมตร และความละเอียด 10 กิโลเมตร, ข้อมูลสภาพอากาศและสารมลพิษต่างๆ จากสถานีวัดภาคพื้นดิน ได้แก่ อุณหภูมิ (Temperature), ค่าอุณหภูมิสูงสุด (Max Temperature), ค่าอุณหภูมิต่ำสุด (Min Temperature) ทั้งสามตัวแปรนี้ใช้หน่วยองศาเซลเซียส, ค่าความชื้น (Relative Humidity) หน่วยคือเปอร์เซ็นต์, ค่าปริมาณฝนตกรายวัน (Daily Rainfall) หน่วยคือมิลลิเมตร, ความสามารถในการมองเห็น (Visibility) หน่วยคือกิโลเมตร, ความไวลม (Wind Speed), (Sustained Wind Speed) สองตัวแปรนี้ใช้หน่วยคือเมตรต่อวินาที, ความกดอากาศ (Air Pressure) หน่วยคือปาสคาล, จุดไอน้ำกลั่นตัว (Dew Point) หน่วยคือองศาเซลเซียส, ค่า $PM_{2.5}$ รายวันในหน่วยไมโครกรัมต่อลูกบาศก์เมตร , ข้อมูลเชิงพื้นที่ได้แก่ ละติจูด ลองจิจูด ของสถานีที่ใช้วัดข้อมูล, จำนวนประชากร, ความสูงจากระดับน้ำทะเล, ความชื้นเฉลี่ยรายปี, ช่วงของค่าอุณหภูมิที่เกิดขึ้นในพื้นที่นี้ เมื่อเลือกคุณสมบัติหรือปัจจัยที่ส่งผลกระทบต่อการเกิด $PM_{2.5}$ ขั้นตอนต่อไปได้แก่ การเตรียมข้อมูลแบ่งเป็นขั้นตอนย่อยๆ ได้ดังนี้

1.การรวมข้อมูล (Data Integrated) เนื่องจากแหล่งข้อมูลที่จะนำมาใช้ในแบบจำลองมีมาจากหลายแหล่งด้วยกันทำให้ต้องนำข้อมูลเหล่านั้นรวบกันให้เป็นเพียงข้อมูลเดียวที่จะนำเข้าไปใช้ในแบบนจำลอง

2.การเติมค่าข้อมูลที่หายไป โดยใช้การ Interpolate ในการประมาณค่า $PM_{2.5}$ ที่ขาดหายแล้วเติมค่าที่หายไปเหล่านั้น ส่วนค่า AOD ที่ขนาดความละเอียด 3 กิโลเมตรมีอัตราของข้อมูลที่หายไป 94.09 เปอร์เซ็นต์ และขนาดความละเอียด 10 กิโลเมตร มีอัตราของข้อมูลที่หายไปได้แก่ 63.43 เปอร์เซ็นต์ จึงได้มีการทดลองโดยนำข้อมูลที่ไม่รวมค่า AOD และ ข้อมูลที่รวมค่า AOD ไปให้แบบจำลองเรียนรู้และทดลอง แล้วทำการเปรียบเทียบผลลัพธ์

3.การเพิ่มคุณสมบัติด้านเวลา ได้แก่ Day of Year, Season, Weekday และการเติมคุณสมบัติในอดีตเช่น $PM_{2.5}$ ในวันที่ผ่านมา ($PM_{2.5}$_lag1) ปริมาณฝนตกในวันที่ผ่านมา (rainfall_lag1)

4.คำนวณค่าทางสถิติ เช่น ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation), ค่ามากสุด (Max), ค่าน้อยสุด (Min), ของทุกคุณสมบัติและของทุกการแบ่งข้อมูลออกเป็นควอไทล์ที่ 25, 50 และ 75

5.ทำการปรับขนาดของข้อมูล (Normalize Data) โดยให้แต่ละคุณสมบัติมีค่าได้ไม่เกิน [-1,1]

6.แบ่งข้อมูลสำหรับการฝึกฝนแบบจำลอง 70 เปอร์เซ็นต์ และข้อมูลสำหรับทดสอบแบบจำลอง 30 เปอร์เซ็นต์

ถัดจากกระบวนการเตรียมข้อมูลแล้วก็เป็นการฝึกฝนแบบจำลองและทดสอบแบบจำลองโดยแบบจำลองที่งานวิจัยนี้ได้เลือกมาใช้ได้แก่ Random Forest, Extreme Gradient Boost, Deep Learning โดยการเรียนรู้เชิงลึกนี้ได้ใช้ Long Short Term Memory ผลลัพธ์ปรากฏว่าแบบจำลองที่มีประสิทธิภาพสูงสุดได้แก่ Extreme Gradient Boost ซึ่งเป็นแบบจำลองที่เรียนรู้โดยข้อมูลที่ไม่มีค่า AOD ที่ความละเอียด 3 กิโลเมตร โดยมี R^2 = 0.81, Mean Absolute Error = 9.93 ไมโครกรัมต่อลูกบาศก์เมตร, Root Mean Square Error = 13.58 ไมโครกรัมต่อลูกบาศก์เมตร และยังให้การสรุปอีกว่า ค่า AOD ที่มีความละเอียด 3 กิโลเมตรไม่ได้เข้ามาช่วยในการทำนายค่า $PM_{2.5}$ เนื่องจาก AOD ที่มีความละเอียด 3 กิโลเมตรนั้นมีค่าข้อมูลที่หายไปเป็นจำนวนมากทำให้เมื่อรวมกับข้อมูลที่เหลือแล้วทำให้ขนาดของข้อมูลเล็กลงทำให้ฝึกฝนได้น้อยลงจึงมีความแม่นยำที่ลดลง

2.31 Estimation of Daily PM10 and PM2.5 Concentration in Italy, 2013-2015 use Spatiotemporal land use Random Forest model

งานวิจัยนี้จาก [26] เป็นงานวิจัยที่มีจุดมุ่งหมายเพื่อประมาณค่า $PM_{2.5}$, PM_{10} และ $PM_{2.5-10}$ (หมายถึงค่า Particular matter ที่มีขนาดระหว่าง 2.5 ไมครอน จนถึง 10 ไมครอน) โดยใช้ความละเอียดเชิงพื้นที่ 1 ตารางกิโลเมตร ต่อ 1 กริดเซลล์สำหรับปี 2013-2015 และใช้ Random Forest ในการประมาณค่าฝุ่นที่เกิดขึ้น และงานวิจัยนี้มีพื้นที่ที่ใช้ในการทดลองได้แก่ ประเทศอิตาลี โดยจะทำการแบ่งประเทศอิตาลีทั้งหมดให้ประกอบด้วย 1 ตารางกิโลเมตรกริดเซลล์แล้วทำการประมาณค่าฝุ่นที่เกิดขึ้นในระดับประเทศ โดยข้อมูลที่ใช้ได้แก่

1.Particulate Matter เป็นข้อมูลเฉลี่ย 24 ชั่วโมงของ PM_{10}, $PM_{2.5}$ (2006-2015) จากสถานีตรวจวัดต่างๆในประเทศอิตาลี นอกจากนี้การหาค่า $PM_{2.5-10}$ นั้นหาได้จากการเอาค่า PM_{10} – $PM_{2.5}$ แต่ด้วยเนื่องจากค่า $PM_{2.5}$ ในช่วงก่อนปี 2013 นั้นมีข้อจำกัด ทำให้สำหรับการทำ $PM_{2.5}$ นั้นใช้แค่ปี 2013-2015

2.Aerosol Optical Depth หรือค่า AOD นี้เป็นค่าจากการวัดการหายไปของแสงจากสิ่งต่างๆที่มันไปสะท้อนหรือถูกดูดซับเข้าไป โดยเมื่อไม่นานมานี้องค์กร NASA ได้มีการพัฒนาวิธีที่จะใช้ในการคำนวณเพื่อหาค่า AOD ออกมาโดยวิธีนั้นมีชื่อว่า MAIAC โดยให้ผลลัพธ์เป็น AOD ที่มีขนาด 1 ตารางกิโลเมตร แต่ว่าค่า AOD ที่ได้นั้นก็ยังมีข้อเสียอยู่นั้นคือไม่สามารถได้ค่า AOD ในพื้นที่ที่มีขนาดใหญ่เนื่องจากปัญหาการปกคลุมของเมฆ หรือ การปกคลุมของน้ำ, หิมะ และอื่นๆ ที่มีผลกระทบต่อการสะท้อนหรือการดูดซับด้วยแสง ทำให้ค่า AOD จาก MAIAC ประกอบด้วยค่าที่หายไปเป็นจำนวนมาก

3.Meteorological Parameter เป็นข้อมูลเกี่ยวกับสภาพอากาศทั่วไปได้แก่ ค่าเฉลี่ยของอุณหภูมิในอากาศ ความไวของลม ทิศทางของลม ความชื้น ความสูงของชั้นบรรยากาศ ความกดอากาศที่ระดับน้ำทะเล การตกตะกอน ดัชนีพืชพรรณ

4.Spatial Data เป็นข้อมูลที่เกี่ยวข้องกับพื้นที่ ได้แก่ ประชากร ลักษณะสภาพภูมิอากาศของพื้นที่ สารมลพิษที่ถูกปลดปล่อยจากสาเหตุต่างๆ เช่น PM_{10}, SO_2, NO_2, CO และ NH_3 ค่าเฉลี่ยของระดับความสูงจากน้ำทะเล ความหนาแน่นของถนน ระยะทางจากจุดกึ่งกลางของช่องที่กำหนดไปจนถึงถนนหลักหรือจุดสำคัญต่างๆของเมืองเช่น สนามบิน หรืออาจไม่เกี่ยวข้องกับสภาวะเศรษฐกิจ แต่มีผลกระทบต่อปัญหามลพิษ ได้แก่ ทะเลสาบ หรือ ทะเล การจัดแบ่งพื้นที่ว่าพื้นที่นี้เป็นพื้นที่ทางการเกษตร มีต้นไม้น้อย เป็นทุ่งหญ้า หรือเป็นเขตเลี้ยงสัตว์ เป็นต้น

เมื่อได้ข้อมูลทั้งหมดแล้วสำหรับวิธีการดำเนินการถูกแบ่งออกเป็น 5 ขั้นตอน ได้แก่

1.การทำนายค่า $PM_{2.5}$ PM_{10} และ $PM_{2.5-10}$ รวมถึงในเขตพื้นที่ที่ไม่มีค่า $PM_{2.5}$ ในแต่ละปีก็ใช้ค่า PM_{10} ในการประมาณแทนก่อน

2.การเติมค่าที่หายไปของค่า AOD โดยค่าที่หายไปนั้นจะใช้การประมาณจาก Atmospheric Ensemble Model โดยใช้บริการที่มีชื่อว่า CAMS ในการดาวน์โหลดข้อมูลซึ่งค่าที่ได้จาก CAMS นั้นเป็นค่า AOD ที่ถูกทำนาย มี 6 ประเภท ได้แก่ Tropospheric aerosols, sea salt, dust, organic, black carbon, sulfates เป็นข้อมูลราย 3 ชั่วโมงที่ความยาวคลื่นแตกต่างกัน ได้แก่ 469 นาโนเมตร, 550 นาโนเมตร, 670 นาโนเมตร, 865 นาโนเมตร, 1240 นาโนเมตร

3.Calibration หาความสัมพันธ์ของค่า PM และข้อมูลจากดาวเทียมต่างๆ รวมถึงข้อมูลที่ถูกวัดจากพื้นดินต่างๆด้วย

4.ทำนาย ค่า $PM_{2.5}$ PM_{10} $PM_{2.5-10}$ โดยใช้ข้อมูลที่ผ่านการ Calibration มาแล้วในแต่ละ Grid Cell

5.ปรับผลการทำนายให้มีประสิทธิภาพที่ดีมากขึ้น โดยการปรับขนาดของพื้นที่ที่ใช้คำนวณให้ลดลง โดยจะคำนวณพื้นที่รอบ Station ระยะ 150 เมตร เพื่อปรับให้ผลลัพธ์ดีขึ้นจากขั้นตอนที่ 3 การวัดประสิทธิภาพใช้ Mean Cross Validation R^2 = 0.75, 0.80 เป็นของ PM_{10} และ $PM_{2.5}$ ตามลำดับในขั้นตอนที่ 3 R^2 ในขั้นตอนที่ 5 เท่ากับ 0.84 และ 0.86 ซึ่งถือว่าประสิทธิภาพพัฒนาขึ้นจากขั้นตอนที่สาม ในส่วนที่ประมาณได้แย่นั้นได้แก่ช่วงฤดูร้อนและตอนใต้ของอิตาลี

ตารางที่ 2.8 สรุปตัวแปรที่ใช้ในแต่ละงานวิจัย

Input / Paper	Elevation	NDVI	Major Road Length	Traffic Density	Distance from Shoreline	Rainfall	Actual $PM_{2.5}$	Forest Area	Latitude, Longitude	Wildfire	Aerosol Optical Depth MODIS	Aerosol Optical Depth MISR	Pressure
งานวิจัยจากหัวข้อ 2.25							✓		✓		✓		
งานวิจัยจากหัวข้อ 2.26	✓		✓				✓	✓	✓	✓	✓		
งานวิจัยจากหัวข้อ 2.27	✓	✓		✓	✓	✓	✓		✓		✓		
งานวิจัยจากหัวข้อ 2.28		✓					✓		✓		✓	✓	✓
งานวิจัยจากหัวข้อ 2.29		✓					✓		✓		✓		✓

ตารางที่ 2.8 สรุปตัวแปรที่ใช้ในแต่ละงานวิจัย (ต่อ)

Wind Speed	Relative Humidity	High Planetary Boundary Layer	Population Density	Temperature	Time	Technique	Performance
✓	✓	✓		✓	✓	Neural Network	R = 0.74 สำหรับรายชั่วโมง R = 0.78 สำหรับรายวัน
✓	✓	✓		✓	✓	Bayesian Downscaling, Markov Chan Monte Carlo	R^2 = 0.7 ในพื้นที่ที่ทำงานได้ดี R^2 <= 0.54 ในพื้นที่ที่ทำงานได้แย่
	✓	✓	✓	✓	✓	Linear Regression	R^2 = 0.72 RMSPE = 8.52 $\mu g/m^3$
✓	✓	✓		✓	✓	National Scale Geographically Weight Regression	R^2 = 0.64 RMSE = 32.98 $\mu g/m^3$ MPE = 8.28 $\mu g/m^3$
✓	✓	✓		✓	✓	Deep Learning Geoi-DBN	R = 0.94 MPE = 9.03 $\mu g/m^3$ RPE = 24.90 %

บทที่ 3
วิธีการดำเนินงานวิจัย

งานวิจัยฉบับนี้มีวัตถุประสงค์อยู่สองส่วน ได้แก่ 1.เพื่อหาปัจจัยที่สงผลกระทบต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ และ 2.เพื่อทำนายค่า $PM_{2.5}$ ในอีกสามชั่วโมงข้างหน้า โดยใช้เทคนิคการเรียนรู้ของเครื่องในการแก้ปัญหา ซึ่งได้มีการวางแผนดำเนินการดังนี้

1) การหาหรือเก็บข้อมูลที่เกี่ยวข้องกับ $PM_{2.5}$ (Data Gathering)

2) การสำรวจลักษณะของข้อมูล (Data Visualization)

3) การทำความสะอาดข้อมูล (Cleaning Data)

4) การเติมแต่งข้อมูลและปรับแต่งข้อมูล (Data Engineering)

5) การฝึกฝนแบบจำลอง (Training Models)

6) การปรับแต่งพารามิเตอร์ (Hyper Parameters Tuning)

7) การทำนายค่า $PM_{2.5}$ ในอีกสามชั่วโมงข้างหน้า (Prediction Next 3 Hour $PM_{2.5}$ Values)

3.1 การค้นหาหรือการเก็บข้อมูลที่เกี่ยวข้องกับ $PM_{2.5}$ (Data Gathering)

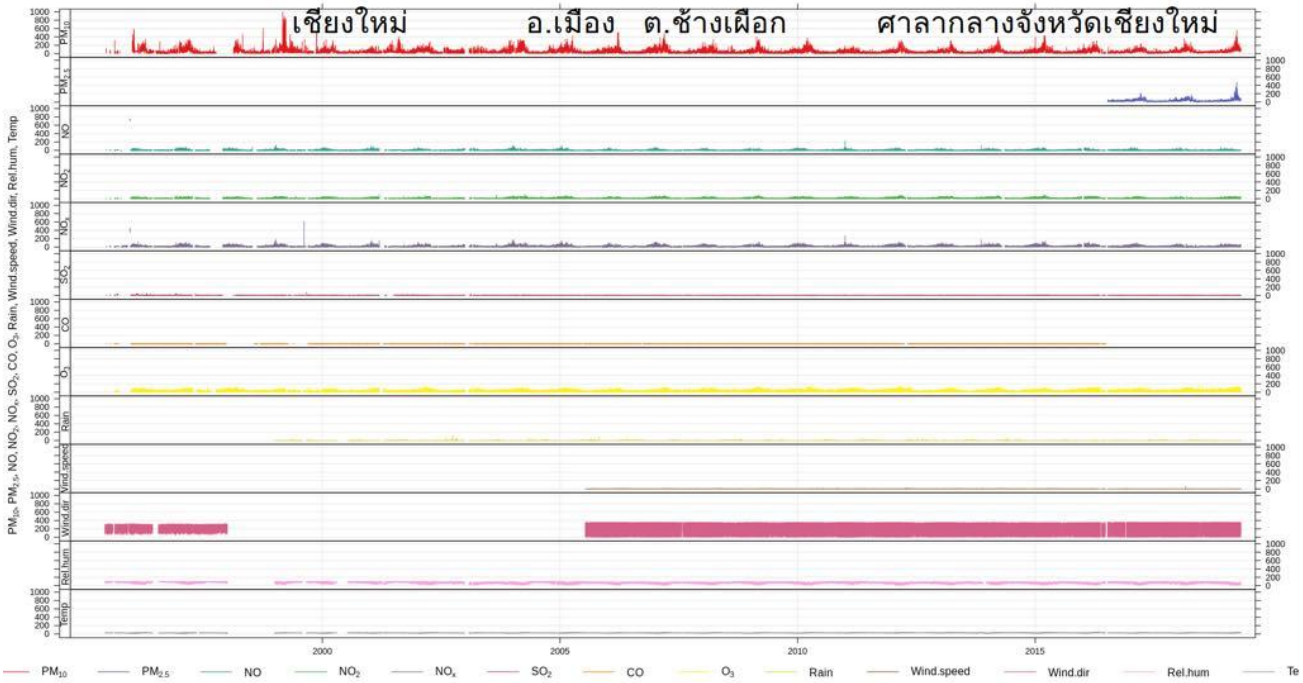
เนื่องจากการวิเคราะห์ปัญหา $PM_{2.5}$ และการทำนายโดยใช้วิธีการเรียนรู้ของเครื่องนั้นจำเป็นต้องใช้ข้อมูลที่เกี่ยวข้องจำนวนมากดังนั้นผู้วิจัยจึงต้องหาข้อมูลที่ไม่ล้าสมัยมากจนเกินไป เนื่องจากข้อมูลในปัจจุบันอาจจะมีความเปลี่ยนแปลงเพราะสิ่งแวดล้อม ณ ปัจจุบันนั้นแตกต่างจากในอดีตและจากการสอบถามผู้เชี่ยวชาญรวมกับการสืบค้นข้อมูลด้วยต้นเองพบว่า ข้อมูลที่มีความเกี่ยวข้องกับปัญหา $PM_{2.5}$ ส่วนใหญ่นั้นได้แก่ ข้อมูลไฟ ข้อมูลอุตุนิยมวิทยา ค่าความลึกเชิงแสงของฝุ่นละออง (Aerosol Optical Depth: AOD) ข้อมูลเชิงพื้นที่และข้อมูลเกี่ยวกับมลพิษทางอากาศ โดยในงานวิจัยอื่นๆ ที่เกี่ยวข้องกับ $PM_{2.5}$ จะใช้ข้อมูลดังกล่าวในการวิเคราะห์และทำนายผล จากปัจจัยที่ได้กล่าวข้างต้นในงานวิจัยฉบับนี้ผู้วิจัยจะใช้ข้อมูลจากแหล่งข้อมูล 3 แหล่งข้อมูลด้วยกันได้แก่

3.1.1 ข้อมูลจากกรมควบคุมมลพิษ (Pollution Control Department: PCD)

กรมควบคุมมลพิษมีการเก็บข้อมูลทางด้านมลพิษทั่วประเทศไทย แต่เนื่องจากงานวิจัยฉบับนี้มีพื้นที่ที่สนใจในการทำการทดลองได้แก่ จังหวัดเชียงใหม่ ดังนั้นจึงจะเลือกใช้ข้อมูลเพียงจุดรับข้อมูลที่อยู่ในเขตจังหวัดเชียงใหม่เท่านั้นโดยสถานีที่ตรวจรับข้อมูลในจังหวัดเชียงใหม่ที่อยู่ในเครือข่ายของกรมควบคุมมลพิษมีอยู่ 2 สถานี ได้แก่ โรงเรียนยุพราชและศูนย์ราชการ โดยมีรหัสประจำสถานีได้แก่ 35t และ 36t ตามลำดับ

เนื่องจากปัญหา $PM_{2.5}$ เริ่มมีการตื่นตระหนกไม่กี่ปีที่ผ่านมาดังนั้นอาจจะมีข้อมูล $PM_{2.5}$ ไม่มากนักและจากการตรวจสอบข้อมูลพบว่าแต่ละสถานีเริ่มเก็บข้อมูลค่า $PM_{2.5}$ ในช่วงเวลาที่แตกต่างกัน จึงต้องเลือกช่วงเวลาที่ทั้งสองสถานีมีข้อมูลที่ตรงกัน จากสถานีศูนย์ราชการพบว่ามีการเก็บข้อมูล $PM_{2.5}$ ตั้งแต่ปี 2016-ปัจจุบันแต่โรงเรียนยุพราชมีการเก็บข้อมูลตั้งแต่ปี 2015-ปัจจุบันดังนั้นจึงจะใช้ข้อมูลตั้งแต่ปี 2016 เป็นต้นมาแต่เนื่องจากในปี 2019 ข้อมูลที่ผู้วิจัยสามารถเข้าถึงได้มีเพียง ค่า $PM_{2.5}$ ซึ่งขาดข้อมูลส่วนอื่นๆ ไปทำให้ผู้วิจัยเลือกใช้ข้อมูลตั้งแต่ 2016 จนถึง 2018 ในการวิเคราะห์และทำนายผลของ $PM_{2.5}$ โดยข้อมูลที่ผู้วิจัยจะใช้ในการวิเคราะห์และทำนายผลของกรมควบคุมมลพิษนั้นมีปัจจัยทั้งหมดดังนี้

- ข้อมูลอุณหภูมิ (Temperature)
- ข้อมูลความชื้น (Relative Humidity)
- ข้อมูลทิศทางลม (Wind Direction)
- ข้อมูลความไวของลม (Wind Speed)
- ข้อมูลปริมาณฝน (Rain)
- ข้อมูลก๊าซโอโซน (O_3)
- ข้อมูลก๊าซคาร์บอนมอนออกไซด์ (CO)
- ข้อมูลก๊าซซัลเฟอไดออกไซด์ (SO_2)
- ข้อมูลออกไซด์ของไนโตรเจน (NO_x)
- ข้อมูลไนโตรเจนออกไซด์ (NO_2)
- ข้อมูล $PM_{2.5}$
- ข้อมูล PM_{10}
- ข้อมูลไนโตรเจนออกไซด์ (NO)



ภาพที่ 3.1 ข้อมูลจากกรมควบคุมมลพิษ ณ ศูนย์ราชการ จังหวัดเชียงใหม่

3.1.2 ข้อมูลอุตุนิยมวิทยาจากเว็บไซต์ www.Wunderground.com

เว็บไซต์นี้เป็นเว็บไซต์ที่ให้ข้อมูลอุตุนิยมวิทยาจากสถานีเก็บข้อมูลทั่วโลก ผู้วิจัยจะต้องทำการดาวน์โหลดข้อมูลที่ต้องการผ่านหน้าเว็บโดยใช้เทคนิคที่เรียกว่า Web Scraping โดยได้ทำการเลือกพื้นที่คือ จังหวัดเชียงใหม่ สถานีที่ทำการรับข้อมูลนั้นตั้งอยู่ที่ สนามบินเชียงใหม่ (Chiang Mai International Airport) ซึ่งจะต้องทำการเลือกช่วงเวลาให้สอดคล้องกับข้อมูลจากกรมควบคุมมลพิษคือ ปี 2016 – 2018 โดยข้อมูลที่มีให้ได้แก่

- ข้อมูลเวลา (Time)
- ข้อมูลอุณหภูมิ (Temperature)
- ข้อมูลจุดไอน้ำกลั่นตัว (Dew Point)
- ข้อมูลความชื้น (Relative Humidity)
- ข้อมูลทิศทางลม (Wind Direction)
- ข้อมูลความไวลม (Wind Speed)
- ข้อมูลความกดอากาศ (Pressure)
- ข้อมูลหยาดน้ำฟ้า (Precipitation)
- ข้อมูลเงื่อนไขสถาพอากาศในวันนั้น (Condition)

จะเห็นว่าข้อมูลบางส่วนนั้นทับซ้อนกันกับข้อมูลจากกรมควบคุมมลพิษดังนั้นผู้วิจัยจึงต้องทำการตัดข้อมูลที่ทับซ้อนกันเช่น อุณหภูมิ ความชื้น ทิศทางลม ความไวลม ออกจากชุดข้อมูลก่อนการฝึกฝนแบบจำลองแต่ยังใช้ในการวิเคราะห์บางส่วนเพื่อให้ลดความคลาดเคลื่อนในการทำนายผลลงเนื่องจากสถานีรับข้อมูลของกรมควบคุมมลพิษและเว็บไซต์ Wunderground นั้นตั้งอยู่ในจุดที่แตกต่างกันการรวมข้อมูลเข้าด้วยกันอาจส่งผลถึงความแม่นยำในการทำนายผลได้

3.1.3 ข้อมูลจุดความร้อนจาก Fire Information of Resource Management System

สามารถดาวน์โหลดได้จากเว็บไซต์ firms.modaps.eosdis.nasa.gov และนอกจากแหล่งข้อมูลด้านบนแล้วยังมีข้อมูลไฟหรือข้อมูลจุดความร้อนที่ยังส่งผลกระทบต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ จากตารางที่ 2.4 ในบทที่ 2 จะเห็นว่าการเผาไหม้นั้นเป็นสาเหตุอันดับหนึ่งที่ทำให้เกิดค่า $PM_{2.5}$ ดังนั้นข้อมูลส่วนนี้จึงมีความสำคัญเช่นกัน ซึ่งข้อมูลส่วนนี้จะใช้ข้อมูลจากดาวเทียมที่ใช้ในการตรวจจับความผิดปกติของอุณหภูมิและไฟที่เกิดขึ้นผ่านอุปกรณ์ VIIRS ที่ความละเอียด 375 เมตรโดยได้กล่าวไว้แล้วในบทที่สองหัวข้อ 2.24 ข้อมูลที่ใช้ในการวิเคราะห์และทำนายผลได้แก่

- Latitude
- Longitude
- Acq_Date (Acquisition Date)
- Acq_Time (Acquisition Time)
- FRP (Fire Radiation Power)

3.2 การสำรวจลักษณะของข้อมูล (Data Visualization)

ในขั้นตอนนี้เป็นการสำรวจลักษณะของข้อมูลที่มีเพื่อเสริมสร้างความเข้าใจให้แก่ผู้วิจัยว่าลักษณะของข้อมูลมีรูปแบบ (Pattern) ไปในทิศทางใด สามารถใช้เป็นขั้นตอนในการเลือกคุณสมบัติที่จะใช้ในการฝึกฝนแบบจำลองได้ โดยวิธีการสำรวจข้อมูลมีหลายแบบด้วยกันไม่ว่าจะเป็นการตรวจสอบค่าเฉลี่ย (Mean), ข้อมูลในแต่ละควอไทล์ (Quartile) มีลักษณะแบบไหน, ลักษณะของการกระจายของแต่ละคุณสมบัติเป็นอย่างไร และการตรวจสอบอื่นๆ อีกมากมาย ซึ่งส่วนใหญ่เพื่อให้ง่ายต่อความเข้าใจของผู้อื่นและผู้วิจัยเองการสำรวจลักษณะของข้อมูลควรทำในรูปแบบของกราฟ (Graph) ในลักษณะต่างๆ ซึ่งกราฟในแต่ละแบบก็จะมีลักษณะที่บ่งบอกข้อมูลที่แตกต่างกันไป ดังนั้นเพื่อให้การวิเคราะห์มีความถูกต้อง, ชัดเจนและเพื่อให้สามารถตอบได้ว่าปัจจัยใดหรือคุณสมบัติใดในข้อมูลที่เลือกมานั้นส่งผลกระทบต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ ตามจุดประสงค์ข้อที่ 1 จึงควรเลือกใช้กราฟและวิธีการวิเคราะห์ให้เหมาะสม

3.3 การทำความสะอาดข้อมูล (Cleaning Data)

การทำความสะอาดข้อมูลถือเป็นขั้นตอนหนึ่งในการเตรียมข้อมูลก่อนที่จะนำข้อมูลที่เลือกมานั้นไปใช้งานในการฝึกฝนแบบจำลองหรือนำไปเติมแต่งในขั้นตอนต่อไปเพราะเนื่องจากชุดข้อมูลที่มีบางครั้งมีความไม่สมบูรณ์อยู่หรือข้อมูลที่มีนั้นผิดเพี้ยนไปจากลักษณะที่ควรจะเป็น ถือเป็นขั้นตอนที่มีความสำคัญเป็นอย่างมากเนื่องจากหากการทำความสะอาดข้อมูลนั้นไม่ถูกต้องหรือทำได้ไม่มีประสิทธิภาพมากพอนั้นอาจจะทำให้แบบจำลองที่ฝึกฝนด้วยข้อมูลชุดนี้เกิด Overfitting ได้ เนื่องจากยังมีข้อมูลบางส่วนที่เป็นข้อมูลรบกวนทำให้แบบจำลองเข้าใจรูปข้อมูลในลักษณะที่ไม่ถูกต้อง สำหรับการทำความสะอาดข้อมูลนั้นจะเป็นการแก้ไขข้อมูลให้มีลักษณะที่ถูกต้องตามที่ข้อมูลควรจะเป็นโดยมีขั้นตอนดังนี้

3.3.1 การกำจัดข้อมูลรบกวน (Noisy Data)

เป็นการลบข้อมูลที่มีลักษณะผิดปกติโดยสามารถสังเกตข้อมูลที่ผิดลักษณะนี้จากขั้นตอนการสำรวจลักษณะข้อมูลได้ เช่น การตรวจสอบโดยใช้ Quartile Plot หรือ การสังเกตจากลักษณะการกระจายของข้อมูลหากข้อมูลส่วนใหญ่กระจุกอยู่ที่ค่าๆ หนึ่งแต่กลับมีข้อมูลบางส่วนที่แตกต่างออกมาข้อมูลส่วนนั้นอาจจะเป็นข้อมูลรบกวนได้ หรืออาจจะสังเกตจากความเป็นจริงเช่น ในความเป็นจริงค่า $PM_{2.5}$ ไม่มีทางเป็นค่า 0 หรือ ติดลบดังนั้นหากในข้อมูลพบค่าที่น้อยกว่า 0 หรือเท่ากับ 0 ก็ให้กำจัดข้อมูลส่วนนั้นออกจากชุดข้อมูล ในอีกกรณีเช่น อุณหภูมิติดลบหรือสูงเกินกว่า 40 องศาเซลเซียสซึ่งไม่มีทางเกิดขึ้นในประเทศไทยดังนั้นจึงให้ลบค่าที่อยู่นอกขอบเขตนี้ออกจากชุดข้อมูลเช่นกัน เป็นต้น

3.3.2 การจัดการกับค่าที่หายไป (Missing Value)

สำหรับค่าที่หายไปผู้วิจัยได้ทำการออกแบบการจัดการกับข้อมูลประเภทดังกล่าวเป็น 4 วิธีด้วยกันซึ่งแต่ละวิธีก็จะมีข้อดีข้อเสียแตกต่างกันไปตามแต่ลักษณะดังนี้

1) การจัดการกับข้อมูลที่หายไปด้วยการลบออกจากชุดข้อมูล เป็นวิธีที่ง่ายและทำให้ได้ชุดข้อมูลที่ไม่มีการเติมแต่งใดๆ เป็นข้อมูลที่เกิดขึ้นจริง ณ เวลานั้นๆ แต่มีข้อเสียคือจะทำให้ได้ชุดข้อมูลที่มีขนาดเล็กลง การฝึกฝนแบบจำลองด้วยข้อมูลขนาดเล็กอาจทำให้เกิดปัญหา Underfitting ได้

2) การจัดการกับข้อมูลที่หายไปด้วยวิธีการแทนแต่ละคุณสมบัติด้วยค่าเฉลี่ย เนื่องจากปัญหา $PM_{2.5}$ หรือปัญหาสภาวะแวดล้อมนั้นเป็นปัญหาที่เกิดการเปลี่ยนแปลงตลอดเวลา ดังนั้นค่าเฉลี่ยที่คิดจากข้อมูลเป็นเวลาทั้งปีนั้นอาจจะไม่ใช่ค่ากลางอย่างแท้จริง ผู้วิจัยจึงทำการแทน ด้วยค่าเฉลี่ยรายเดือนของแต่ละคุณสมบัติ แต่ในชุดข้อมูลก็จะมีคุณสมบัติบางคุณสมบัติที่ไม่สามารถนำไปคำนวณได้ ได้แก่ ทิศทางลม (Wind direction) ค่าที่แสดงในคุณสมบัตินี้เป็นการแสดงของทิศทางดังนั้นจึงไม่สามารถนำไปคำนวณเพื่อหาค่าเฉลี่ยได้ให้ทำการละเว้นคุณสมบัตินี้ไว้ในการแทน หลังจากแทนข้อมูลแล้วหากยังมีส่วนที่ยังหายไปอยู่ให้ทำการเติมข้อมูลด้วยข้อมูลถัด

จากนั้นหรือที่เรียกว่า Back Fill โดยผู้วิจัยจะให้ขอบเขตของการเติมคือ 8 ชั่วโมงหากทำการเติมด้วยข้อมูลถัดไปแล้วยังมีค่าที่หายไปอยู่ให้ทำการลบข้อมูลที่หายไปนั้นออกจากชุดข้อมูล

3) การจัดการกับข้อมูลที่หายไปด้วยการแทนจากชุดข้อมูลอีกชุดข้อมูล วิธีการนี้เป็นวิธีที่ใช้ข้อมูลจากอีกชุดข้อมูลที่คุณสมบัติมีค่าที่ใกล้เคียงกันในการแทนข้อมูล โดยผู้วิจัยได้เลือกชุดข้อมูลจาก Berkeleyearth ซึ่งชุดข้อมูลนี้ได้ทำการเก็บข้อมูล $PM_{2.5}$ ตั้งแต่ปี 2016 – ปัจจุบันไว้ ดังนั้นวิธีนี้จะทำการแทนข้อมูลได้แค่คุณสมบัติ $PM_{2.5}$ เท่านั้นสำหรับการแทนจะทำการแทนคุณสมบัติ $PM_{2.5}$ ของกรมควบคุมมลพิษที่หายไปด้วยข้อมูล ณ วันเวลาที่สอดคล้องกันด้วยข้อมูลจาก Berkeleyearth หาก ณ วันเวลานั้นของ Berkeleyearth ไม่มีข้อมูลเช่นเดียวกันก็ให้ละเว้นไว้เมื่อแทนข้อมูลเสร็จแล้วหากยังมีข้อมูลที่หายไปหลงเหลืออยู่ให้ทำการเติมด้วยข้อมูลถัดไปโดยมีขอบเขตเวลาคือ 8 ชั่วโมงหากเติมแล้วยังเหลือข้อมูลที่หายไปอยู่ให้ทำการลบทิ้งจากชุดข้อมูล

4) การจัดการกับข้อมูลที่หายไปด้วยการใช้ฟังก์ชันไทม์จากไลบรารีแพนด้า (Time Built-in Function from Pandas) วิธีนี้จะเป็นการแทนทุกคุณสมบัติด้วยการใช้ฟังก์ชันไทม์ มีลักษณะการเติมที่คล้ายแบบการเติมแบบเชิงเส้นเพียงแต่จะนำเวลามาคิดบวกเข้ากับลักษณะของข้อมูลเข้าไปด้วยทำให้ตรงกับลักษณะของปัญหาเชิงสิ่งแวดล้อมที่มีเวลาเข้ามากำกับแต่วิธีนี้ผู้วิจัยจะไม่ใช้กับคุณสมบัติ ทิศทางลม (Wind direction) เนื่องจากเป็นตัวเลขที่แสดงถึงค่าทิศทางที่เกิดขึ้นไม่สามารถนำไปคำนวณทางคณิตศาสตร์ได้เช่นเดียวกับ 2 วิธีด้านบนเมื่อแทนข้อมูลเสร็จแล้วหากยังมีข้อมูลที่หายไปหลงเหลืออยู่ให้ทำการเติมด้วยข้อมูลถัดไปโดยมีขอบเขตเวลาคือ 8 ชั่วโมงหากเติมแล้วยังเหลือข้อมูลที่หายไปอยู่ให้ทำการลบทิ้งจากชุดข้อมูล

ทั้ง 4 วิธีเป็นวิธีที่ผู้วิจัยได้ออกแบบโดยเมื่อนำไปใช้กับชุดข้อมูลที่มีแล้วตามปกติจะทำการเลือกมา 1 วิธีแล้วนำชุดข้อมูลที่ทำความสะอาดแล้วไปฝึกฝนแบบจำลองหรือเติมแต่งในขั้นต่อไปต่อ แต่ในงานวิจัยฉบับนี้ต้องการหาว่าการจัดการกับข้อมูลที่หายไปแบบไหนให้ผลลัพธ์ได้ดีกว่ากัน ดังนั้นผู้วิจัยจึงทำการเลือกทั้ง 4 วิธีมาใช้กับชุดข้อมูลที่มีโดยจะได้ผลลัพธ์ออกมาเป็น 4 ชุดข้อมูลที่ผ่านการทำความสะอาดในแต่ละวิธี

3.4 การเติมแต่งข้อมูลและการปรับแต่งข้อมูล (Data Engineering)

ขั้นตอนนี้เป็นการเสริมเติมแต่งชุดข้อมูลที่มีเพื่อให้มีข้อมูลที่ครบถ้วนมากขึ้นพร้อมทั้งยังเป็นขั้นตอนในการเลือกคุณสมบัติที่จำเป็นเบื้องต้นสำหรับการฝึกฝนแบบจำลองอีกด้วย โดยมีวิธีการดังนี้

3.4.1 การเลือกคุณสมบัติที่เหมาะสม (Feature Selection)

สำหรับวิธีนี้ในแบบจำลองบางตัวอาจมีวิธีการเลือกคุณสมบัติของตัวแบบจำลองเองในตัวแต่ก็สามารถคัดกลองคุณสมบัติที่ส่งผลกระทบหรือเกี่ยวข้องด้วยตนเองก่อนได้ เนื่องจากการใส่คุณสมบัติที่ไม่เกี่ยวข้องกับปัญหานั้นนอกจากจะทำให้เวลาในการฝึกฝนเพิ่มขึ้นแล้ว ยังอาจส่งผลให้เกิดปัญหา Overfitting ได้อีกด้วย ในขั้นตอนการเลือกสามารถตัดสินใจได้จากขั้นตอนที่ 3.2 โดยสังเกตจากกราฟต่างๆ ว่าคุณสมบัติไหนที่เกี่ยวข้องกับปัญหา

3.4.2 การแปลงหน่วยให้เหมาะสม

การแปลงหน่วยให้เหมาะสมนั้นขึ้นอยู่กับปัญหาและผู้วิจัยเองว่าต้องการใช้งานคุณสมบัตินั้นๆ ในหน่วยอะไร ตัวอย่างเช่น การใช้คุณสมบัติอุณหภูมิสามารถใช้หน่วยเป็นฟาเรนไฮน์ (Fahrenheit) หรือ เซสเซียส (Celsius) เป็นต้น ซึ่งในการแปลงหน่วยของแต่ละคุณสมบัติก็มีวิธีการคำนวณที่แตกต่างกันสามารถหาการแปลงหน่วยของแต่ละคุณสมบัติจากแหล่งสืบค้นข้อมูลต่างๆ ได้เช่น www.google.com เป็นต้น โดยในขั้นตอนนี้ผู้วิจัยจะทำการแปลงหน่วยของความไวลมจากเมตรต่อชั่วโมงให้กลายเป็นกิโลเมตรต่อชั่วโมงและเนื่องจากอุณหภูมิมีหน่วยเป็นเซลเซียสอยู่แล้วจึงไม่จำเป็นต้องเปลี่ยนแปลงอะไร

3.4.3 การลดขนาดข้อมูลจากรายชั่วโมงเป็นรายสามชั่วโมง

ขั้นตอนนี้เป็นขั้นตอนสำหรับการลดขนาดของข้อมูลเพื่อให้สอดคล้องกับความต้องการในการทำนายตามจุดประสงค์ คือ ผู้วิจัยต้องการทำนายค่า $PM_{2.5}$ ที่จะเกิดขึ้นของ 3 ชั่วโมงข้างหน้าดังนั้นจึงต้องใช้ข้อมูลเป็นรายสามชั่วโมงตาม สำหรับวิธีการลดขนาดข้อมูลนั้นผู้วิจัยจะใช้วิธีการหาค่าเฉลี่ยของแต่ละคุณสมบัติรวมด้วยกัน 4 ชั่วโมงจากข้อมูลรายชั่วโมงให้กลายเป็นข้อมูลรายสามชั่วโมง เช่น ค่า $PM_{2.5}$ ณ วันที่ 4 เมษายน ตั้งแต่เวลา 12.00 – 15.00 จะถูกนำมาหาค่าเฉลี่ยแล้วกลายเป็นข้อมูลค่า $PM_{2.5}$ ณ วันที่ 4 เมษายนของเวลา 15.00 ทำเช่นนี้ไปจนครบในชุดข้อมูลจะได้ข้อมูลที่เป็นรายสามชั่วโมงสำหรับการฝึกฝนแบบจำลอง

3.4.4 การเพิ่มคุณสมบัติแก่ชุดข้อมูล

ในขั้นตอนการเพิ่มคุณสมบัตินี้ประกอบด้วยส่วนย่อยๆ หลายส่วนด้วยกันแบ่งได้ดังนี้

1) การเพิ่มคุณสมบัติเกี่ยวกับเวลา

โดยคุณสมบัติเวลาที่ต้องการเพิ่มนั้น ได้แก่ วันในสัปดาห์ (วันจันทร์–วันอาทิตย์), วันในเดือน (1-31), วันสุดสัปดาห์ (เสาร์-อาทิตย์), ชั่วโมง (0-23), เดือนและฤดูกาล โดยฤดูกาลนั้นมีหลักการในการแบ่งดังนี้ 1.ฤดูร้อน เริ่มจากกลางเดือนกุมภาพันธ์–กลางเดือนพฤษภาคม 2.ฤดูหนาว เริ่มจาก กลางเดือนตุลาคม-กลางเดือนกุมภาพันธ์ 3.ฤดูฝน เริ่มจาก กลางเดือนพฤษภาคม-กลางเดือนตุลาคม โดยจะเพิ่มช่วงคาบเกี่ยวระหว่างฤดูกาลเป็นระยะเวลา 10 วัน

2) การปรับแต่งคุณสมบัติทิศทางลม

โดยทิศทางลมค่าที่แสดงเป็นตัวเลข 0-360 เป็นค่าที่บ่งบอกถึงทิศทางของลมที่พัดผ่านมายังตัวรับข้อมูล (Sensor) ณ เวลาขณะนั้นโดยจะทำการสร้างคุณสมบัติใหม่โดยค่าที่จะแทนในคุณสมบัติใหม่จะเป็นชื่อทิศทางแทนตัวเลข ซึ่งจะทำการจับคู่ตัวเลขที่ในคุณสมบัติเดิมเข้ากับทิศทางลม

3) การปรับแต่งคุณสมบัติที่มีลักษณะวนซ้ำ หรือ คุณสมบัติที่มีค่าข้อมูลเป็นแบบประเภท ส่วนแรกได้แก่คุณสมบัติที่มีลักษณะวนซ้ำ (Cyclical Variable) ข้อมูลที่มีลักษณะวนซ้ำในชุดข้อมูลได้แก่ ชั่วโมง ฤดูกาล ทิศทางลม วันในสัปดาห์ วันในเดือน เป็นต้น โดยคุณสมบัติเหล่านี้จะถูกนำไปสร้างคุณสมบัติใหม่ 2 คุณสมบัติโดยผ่านการแปลงข้อมูลด้วยสมการหาเส้นรอบรูปของวงกลมเส้นรอบรูป = $2\pi r$ โดยจะนำสมการเส้นรอบรูปนี้ไปทำการดัดแปรงโดยนำไปหารกับค่าสูงสุดของคุณสมบัติที่นำมาทำการแปลงแล้วนำฟังก์ชันโคไซน์และฟังก์ชันไซน์ ดังนี้ Sin, Cos($2\pi r/\max(r)$) ซึ่งค่า r จะถูกแทนด้วยข้อมูลที่อยู่ในคุณสมบัตินั้นๆ เมื่อนำคุณสมบัติใหม่ทั้งสองคือค่าที่ผ่านฟังก์ชันไซน์และค่าที่ผ่านฟังก์ชันโคไซน์ทั้งสองไปสร้างกราฟคู่กันจะได้กราฟวงกลมขนาดเท่ากับค่าสูงสุดของแต่ละคุณสมบัติที่ผ่านการแปลง เหตุผลที่ต้องทำเช่นนี้ ยกตัวอย่างให้เห็นภาพเช่น ชั่วโมงมีค่าข้อมูลที่เกิดขึ้นทั้งหมดคือ 0 – 23 โดยช่วงข้อมูลที่ 0 – 23 จะเพิ่มขึ้นทีละหนึ่งหรือทีละสามหลังจากผ่านการปรับขนาดแล้วค่าจะเพิ่มขึ้นตามความต่างนั้นจนถึงค่าสุดท้ายได้แก่ 23 แล้วข้อมูลจะทำการวนกลับไปเป็น 0 ใหม่กลับพบว่าความต่างของข้อมูล ณ ช่วงเวลา 23 กับ 0 นั้นไม่ใช่ 1 หรือ 3 ชั่วโมง ด้วยเหตุนี้การแปลงข้อมูลนี้จะช่วยให้ความต่างของ 23 กับ 0 นั้นเท่ากับส่วนอื่นๆ ของข้อมูล

ส่วนที่สองได้แก่คุณสมบัติที่มีลักษณะประเภท (Category) คุณสมบัติเหล่านี้ไม่สามารถนำมาคำนวณเชิงคณิตศาสตร์ได้เพราะส่วนมากจะอยู่ในรูปของตัวอักษร (String) ดังนั้นแบบจำลองจึงไม่สามารถทำการคำนวณได้เพราะแบบจำลองนั้นสามารถเข้าใจได้เฉพาะตัวเลขเท่านั้นจึงต้องทำการแปลงให้เป็นตัวเลขซึ่งวิธีนี้จะทำการสร้างขึ้นมาเป็นตัวแปรแบบดัมมี่ (Dummy Variable) โดยวิธีดัมมี่นั้นจะทำการแปลงข้อมูลที่อยู่ในคุณสมบัติ

แบบประเภททั้งหมดนำไปสร้างเป็นคุณสมบัติใหม่แล้วแทนค่าด้วย 0-1 โดยคุณสมบัติที่ มีลักษณะแบบประเภทได้แก่ ทิศทางลม ชั่วโมง เดือน วันในเดือน วันในสัปดาห์ ฤดูกาล

3.4.5 การสร้างข้อมูลในอดีต (Lag Features)

เนื่องจากจุดประสงค์เป็นการทำนายข้อมูลในอนาคตดังนั้นจึงจำเป็นต้องใช้ข้อมูลในอดีต เพื่อที่จะนำข้อมูลในอดีตนั้นไปทำนายสิ่งที่จะเกิดในอนาคต ทำให้ผู้วิจัยต้องทำการสร้างข้อมูลที่เป็น Lag ขึ้นมาโดยวิธีการสร้างก็คือจะทำการขยับข้อมูลทั้งหมดทุกคุณสมบัติไปยังชั่วโมงถัดไปหรือเวลาถัดไปจำนวน 1 ขั้นตอน เช่น ข้อมูล ณ เวลา ที่ 12.00 PM จะถูกขยับไปเป็น 15.00 PM เป็นต้น

3.4.6 การรวมข้อมูลจากแหล่งข้อมูลทั้งหมดเข้าด้วยกัน

ในงานวิจัยนี้มีการใช้ข้อมูลจากหลายแหล่งข้อมูลด้วยกัน ดังนั้นในการรวมข้อมูลผู้วิจัยจะทำการรวมข้อมูลด้วยการเข้าคู่กันทั้งหมด จากแหล่งข้อมูลที่มีได้แก่ 1.กรมควบคุมมลพิษ (Pollution Control Department: PCD) 2.Wunderground 3.FIRMS จะทำให้ได้การเข้าคู่กันทั้งหมด 7 แบบ ได้แก่

- PCD (4×2 = 8 ชุดข้อมูล)
- Wunderground (1×4 = 4 ชุดข้อมูล)
- FIRMS (1×4 = 4 ชุดข้อมูล)
- Wunderground – FIMRS (1×4 =4 ชุดข้อมูล)
- PCD – Wunderground (4×2 = 8 ชุดข้อมูล)
- PCD – FIRMS (4×2 = 8 ชุดข้อมูล)
- PCD – Wunderground – FIRMS (4×2 = 8 ชุดข้อมูล)

โดยในแต่ละแบบจะถูกแบ่งย่อยได้อีกคือ การเติมแต่งข้อมูลด้วยวิธี Cyclical Variable และ Dummy Variable โดยแหล่งข้อมูลที่มีการเติมแต่งด้วยสองวิธีด้านบนนั้นมีเพียงแค่ PCD เท่านั้น โดยข้อมูล PCD นั้นสามารถแบ่งย่อยตามวิธีการจัดการกับข้อมูลที่หายไปได้อีก ทำให้มีข้อมูลที่จะถูกนำไปฝึกฝนทั้งหมด (8×4) + (4×3) = 44 ชุดข้อมูลเพื่อทำการหาว่าการแทนที่ข้อมูลที่หายไปแบบไหนจะให้ผลลัพธ์ที่ดีที่สุด, การดัดแปลงข้อมูลแบบ Cyclical Variable และ Dummy Variable แบบไหนให้ผลลัพธ์ดีที่สุด, สุดท้ายนี้เพื่อหาว่าแหล่งข้อมูลส่วนไหนมีความสำคัญต่อปัญหาที่ต้องการแก้ไขมากที่สุด

3.5 การฝึกฝนแบบจำลอง (Training Models)

ขั้นตอนนี้เป็นการฝึกฝนแบบจำลองแต่ละประเภทเพื่อหาชุดข้อมูลที่ดีที่สุดที่จะถูกนำไปใช้ในการทำนายค่า $PM_{2.5}$ ที่จะเกิดขึ้นในอีก 3 ชั่วโมงข้างหน้าโดยแบบจำลองที่เลือกใช้ มีทั้งหมด 4 แบบจำลองด้วยกัน ได้แก่ Multiple Linear Regression, Random Forest, Extreme Gradient Boosting และ Neural Network สำหรับ Neural Network จะถูกแบ่งออกเป็นสองส่วนเพราะแบ่งเป็นส่วนของไลบรารี Sklearn และ Keras โดยจะทำการฝึกฝนแบบจำลองด้วยพารามิเตอร์ที่ถูกตั้งเป็นพื้นฐานโดยไม่มีการปรับแต่งใดๆ สำหรับการแบ่งข้อมูลที่จะใช้ในการเรียนรู้และทดสอบนั้นจะถูกแบ่งเป็นอัตราส่วน 70:30 โดยส่วนที่ใช้เรียนรู้คือ 70% และส่วนที่ใช้สำหรับทดสอบคือ 30%

3.6 การปรับแต่งพารามิเตอร์ (Hyper Parameter Tuning)

การปรับแต่งพารามิเตอร์เป็นขั้นตอนหลังจากที่หาชุดข้อมูลที่ให้ผลลัพธ์ได้ดีที่สุดด้วยแบบจำลองพื้นฐานแล้ว ขั้นตอนต่อมาจะทำการปรับแต่งพารามิเตอร์เหล่านั้นเพื่อให้ได้ประสิทธิภาพของแต่ละแบบจำลองที่ดีขึ้น ซึ่งในแต่ละแบบจำลองก็จะมีพารามิเตอร์ที่ต้องปรับแต่งที่แตกต่างกัน สำหรับการปรับจูนพารามิเตอร์จะใช้ไลบรารี RandomSearchCV, GridSearchCV ของ Sklearn เข้ามาช่วยในการหาพารามิเตอร์ที่จะทำให้แบบจำลองมีประสิทธิภาพดีที่สุด เมื่อทำการปรับจูนพารามิเตอร์ของแต่ละแบบจำลองเสร็จแล้วจะได้แบบจำลองที่มีประสิทธิภาพมากที่สุดของแบบจำลองแต่ละแบบให้ทำการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองว่าแบบจำลองไหนให้ผลลัพธ์ได้ดีที่สุดโดยการวัดผลให้ใช้ RMSE, MAE, R^2, MAPE เป็นต้น

การปรับพารามิเตอร์ของแต่ละแบบจำลองสามารถจำแนกได้ดังนี้

3.6.1 Random Forest

พารามิเตอร์ที่จะทำการปรับแต่งมีดังนี้

n_estimator = จำนวนต้นไม้ทั้งหมดที่ใช้ในการฝึกฝน

max_depth = ความลึกของต้นไม้ จะทำให้จำกัดจำนวนของ Node ในต้นไม้ได้

min_sample_split = จำนวนน้อยที่สุดที่ข้อมูลย่อย (Sample) ที่ต้องการเพื่อแบ่ง Node

min_sample_leaf = จำนวนน้อยที่สุดของข้อมูลย่อย (Sample) ณ จุดใบ (Leaf Node)

max_feature = จำนวนคุณสมบัติ (Feature) ในการคิดเพื่อหาการแบ่งที่ดีที่สุด

bootstrap = เป็นตัวกำหนดในการแบ่งข้อมูลย่อย หากปิดข้อมูลทั้งหมดจะถูกใช้ในการสร้างแต่ละต้นไม้

3.6.2 Extreme Gradient Boosting

พารามิเตอร์ที่จะทำการปรับแต่งมีดังนี้

สำหรับ Extreme Gradient Boosting นั้นจะทำการปรับแต่ง min_sample_splie, min_sample_leaf, max_feature, max_depth ซึ่งมีความหมายเหมือนกับ Random Forest ตามหัวข้อที่ 3.6.1 แต่มีพารามิเตอร์ 2 ตัวที่มีความแตกต่างได้แก่

n_estimator = จำนวนของแต่ละต้นไม้ในแต่ละสเตจ

learning_rate = อัตราการเรียนรู้ของแต่ละต้นไม้ หากใช้ค่าที่เหมาะสมจะทำให้ได้ต้นไม้ที่มีประสิทธิภาพสูงตาม

3.6.3 Neural Network

พารามิเตอร์ที่จะทำการปรับแต่งมีดังนี้ คำอธิบายแต่ละคุณสมบัติได้กล่าวไว้ในบทที่ 2

1) Sklearn

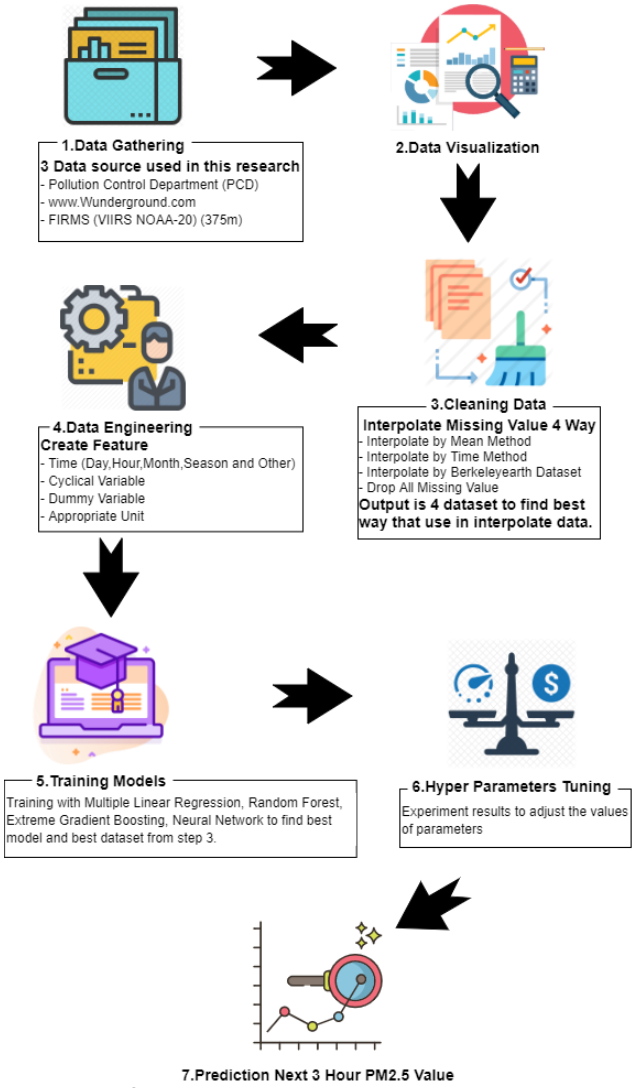
- Activation Function
- Batch Size
- Hidden Layers
- Optimize
- Learning rate
- Alpha

2) Keras

- Activation Function
- Batch Size
- Optimizer
- จำนวนเซลล์ประสาท
- Regularization Rate
- Dropout Rate
- Learning Rate

3.7 การทำนายค่า $PM_{2.5}$ ในอีกสามชั่วโมงข้างหน้า (Prediction Next 3 Hour $PM_{2.5}$ Values)

หลังจากวัดประสิทธิภาพของแต่ละแบบจำลองแล้วจะทำให้ได้แบบจำลองที่ได้มีประสิทธิภาพมากที่สุดโดยจะเลือกใช้แบบจำลองนั้นในการทำนายค่า $PM_{2.5}$ ที่จะเกิดขึ้นในอนาคต ในขั้นตอนนี้ผู้วิจัยจะต้องทำการดาวน์โหลดข้อมูล ณ เวลาปัจจุบันจากในเว็บไซต์ของทางกรมควบคุมมลพิษ, FIMRS, Wunderground เพื่อนำข้อมูลที่เกิดขึ้นใหม่ เข้าสู่แบบจำลองแล้วทำนายผลออกมา โดยแบบจำลองที่นำมาทำนายต้องฝึกฝนด้วยชุดข้อมูลที่ให้ประสิทธิภาพสูงสุดเช่นกัน



ภาพที่ 3.2 ภาพรวมของกระบวนการทำงาน

บทที่ 4
ผลการดำเนินงานและผลการวิเคราะห์ข้อมูล

การทำนายค่าปริมาณฝุ่นที่น้อยกว่า 2.5 ไมครอนและการวิเคราะห์หาปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลงของค่าฝุ่นละอองที่น้อยกว่า 2.5 ไมครอน ซึ่งผลการวิจัยดังกล่าวจะถูกนำเสนอออกเป็นหัวข้อย่อยดังต่อไปนี้

1) การวิเคราะห์หาสาเหตุของการเกิดปัญหา $PM_{2.5}$ ด้วยวิธี Data Visualization

2) การวิเคราะห์หาสาเหตุของการเกิดปัญหา $PM_{2.5}$ ด้วยวิธีการเรียนรู้ของเครื่อง

3) การหาชุดข้อมูลที่เหมาะสมสำหรับการทำนายผล

4) การหาค่าพารามิเตอร์ที่เหมาะสมของแต่ละแบบจำลอง

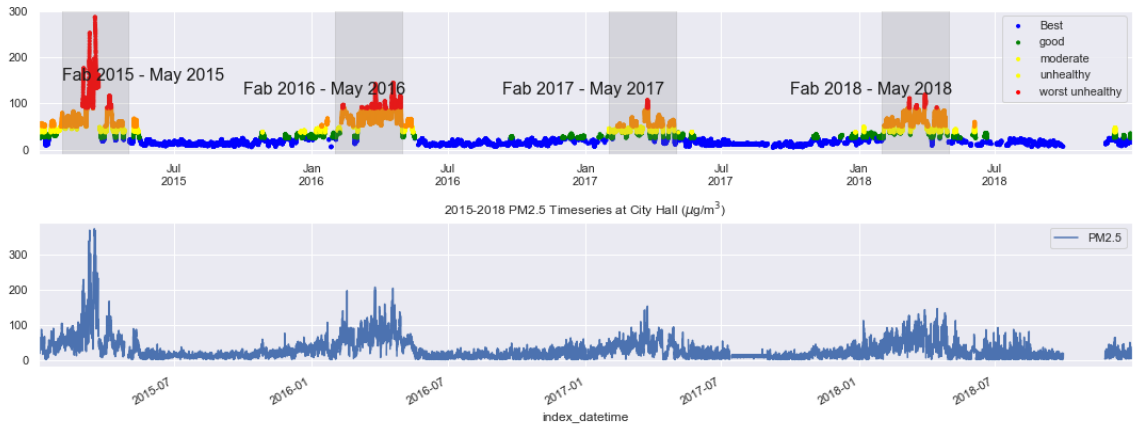
5) การทำนายผลค่าปริมาณฝุ่นละอองที่น้อยกว่า 2.5 ไมครอนหรือ $PM_{2.5}$ ในสามชั่วโมงข้างหน้า

4.1 การวิเคราะห์หาสาเหตุของการเกิดปัญหา PM2.5 ด้วยวิธี Data Visualization

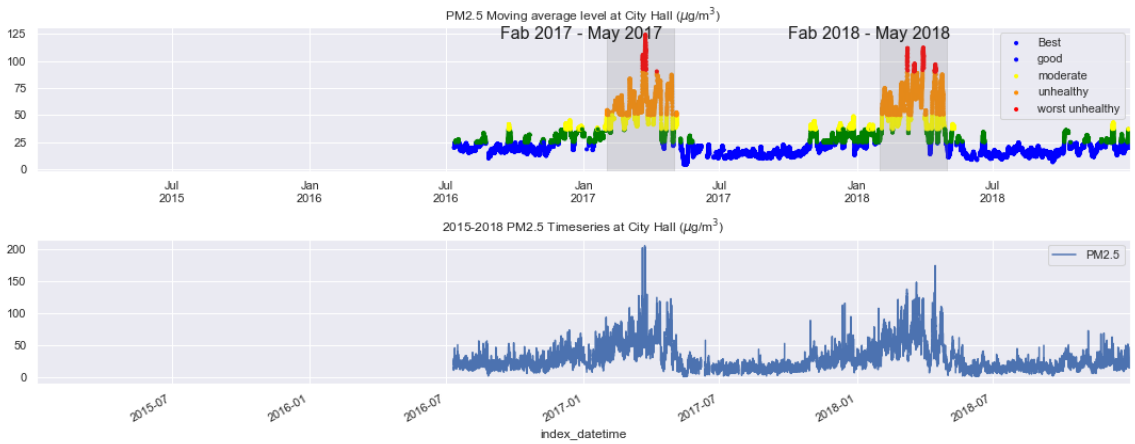
ขั้นตอนนี้เป็นการหาปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ โดยปัจจัยที่กล่าวถึงนั้นจะทำการเลือกจากชุดข้อมูลที่มี แล้วทำการวิเคราะห์หาความสัมพันธ์กับ $PM_{2.5}$ ว่ามีความสัมพันธ์กันหรือไม่ โดยสามารถแบ่งเป็นหัวข้อย่อยได้ดังนี้

4.1.1 การวิเคราะห์ความสัมพันธ์ระหว่างค่า $PM_{2.5}$ กับเวลา

เป็นขั้นตอนการตรวจสอบว่าปัจจัยในเรื่องของเวลามีความสัมพันธ์กับปัญหา $PM_{2.5}$ มากน้อยเพียงใดเริ่มจากภาพที่ 4.1 และภาพที่ 4.2 เป็นข้อมูล $PM_{2.5}$ ทั้งหมดที่ผู้วิจัยสามารถเข้าถึงได้โดยมีข้อมูล 2 สถานีด้วยกันได้แก่ โรงเรียนยุพราช และ ศูนย์ราชการจังหวัดเชียงใหม่ เนื่องจากทั้งสองจุดเป็นข้อมูลที่อยู่ในจังหวัดเชียงใหม่เช่นเดียวกันดังนั้นลักษณะของข้อมูลควรจะมีแนวโน้มไปในทิศทางเดียวกันซึ่งก็สอดคล้องกับกราฟด้านล่างโดยในจุดศูนย์ราชการนั้นจะมีข้อมูล $PM_{2.5}$ ตั้งแต่ปี 2016 – 2018 ส่วนจุดโรงเรียนยุพราชมีข้อมูลตั้งแต่ปี 2015 – 2018 จากการสำรวจข้อมูลพบว่า $PM_{2.5}$ ที่สูงขึ้นนั้นจะเกิดขึ้นในช่วงเดือนกุมภาพันธ์ – เดือนพฤษภาคมแต่จะเริ่มมีค่าสูงขึ้นตั้งแต่เดือนมกราคมโดยจะมีรูปแบบการเกิดเป็นเช่นนี้มาสามปีทำให้ผู้วิจัยคิดว่าในปีต่อๆ ไปนั้นปัญหานี้ก็จะมีค่าที่สูงขึ้นในช่วงเดือนดังกล่าวเช่นเดียวกัน จากกราฟด้านล่างข้อมูลด้านบนเป็นข้อมูลที่ผ่านการหาค่าเฉลี่ยต่อเนื่อง 24 ชั่วโมง (Moving Average) มาแล้วเพื่อให้สามารถนำไปคำนวณค่าคุณภาพอากาศ (Air Quality Index) และสามารถแปลงเปลี่ยนให้เป็นสีเพื่อให้ง่ายต่อการเข้าใจของผู้อ่าน โดยสีแต่ละสีนั้นจะบ่งบอกถึงความอันตรายของค่าที่เกิดขึ้นรายละเอียดสามารถตรวจสอบได้ที่บทที่ 2 ในหัวข้อที่ 2.2 ที่กล่าวไว้ข้างต้น

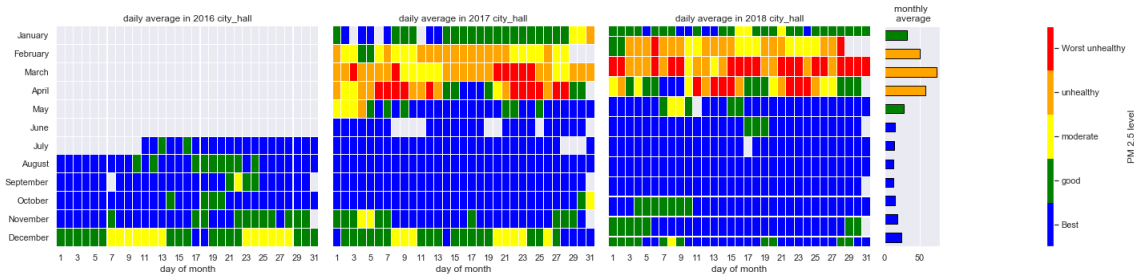


ภาพที่ 4.1 ข้อมูล $PM_{2.5}$ ทั้งหมดแสดงในลักษณะของ Time Series ณ โรงเรียนยุพราช จังหวัดเชียงใหม่

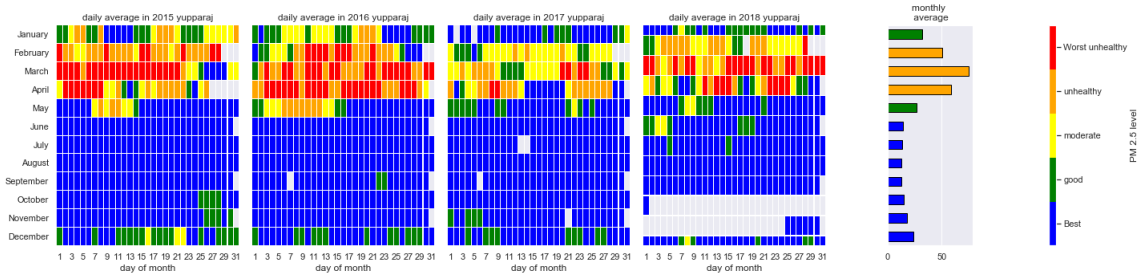


ภาพที่ 4.2 ข้อมูล $PM_{2.5}$ ทั้งหมดแสดงในลักษณะของ Time Series ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่

หากต้องการทำการวิเคราะห์เชิงลึกเพื่อให้เห็นค่าข้อมูลในแต่ละวันกราฟในภาพที่ 4.1 และภาพที่ 4.2 นั้นอาจจะไม่เหมาะสม จึงทำการแสดงเป็นตารางกราฟเพื่อให้ผู้ชมสังเกตเห็นค่า $PM_{2.5}$ ที่ผ่านการแปลงให้เป็นค่าเฉลี่ยต่อเนื่องดังภาพที่ 4.3 และ ภาพที่ 4.4

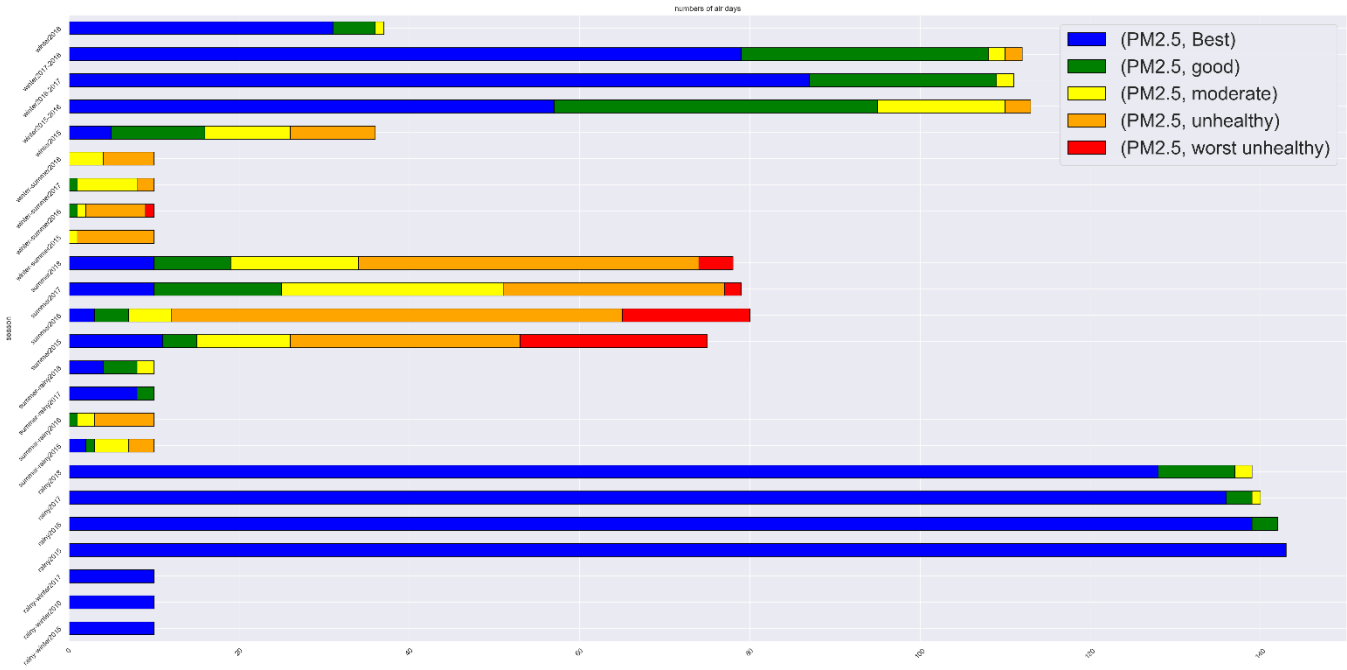


ภาพที่ 4.3 ข้อมูล $PM_{2.5}$ ในรูปแบบรายวัน ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่

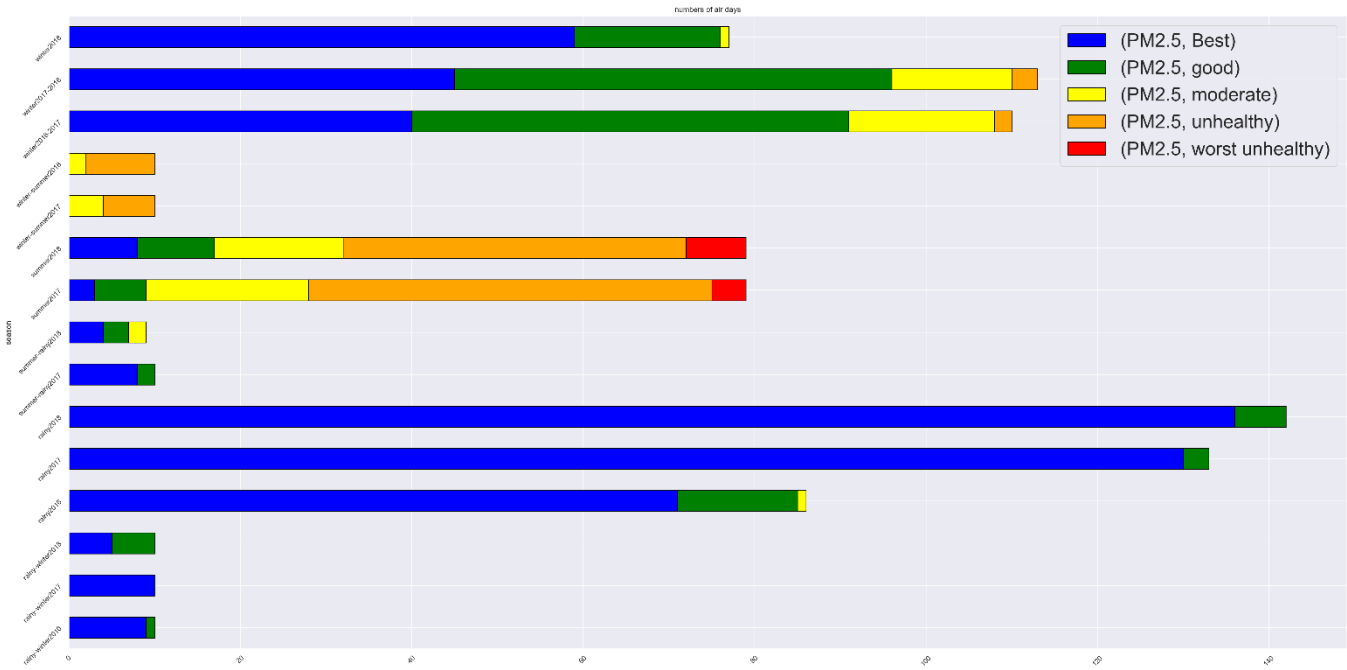


ภาพที่ 4.4 ข้อมูล $PM_{2.5}$ ในรูปแบบรายวัน ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่

เมื่อทำการปรับรูปแบบการมองในช่วงเวลาที่เป็นช่วงวันโดยทำการเปลี่ยนแปลงเป็นฤดูกาล จะเห็นได้ว่าสำหรับจังหวัดเชียงใหม่จากข้อมูลที่มีนั้นสรุปได้ว่าระดับของ $PM_{2.5}$ นั้นจะเพิ่มสูงเกินกว่าระดับมาตรฐานในฤดูร้อนและช่วงคาบเกี่ยวระหว่างฤดูหนาวและฤดูร้อน และจะมีระดับปกติในช่วงฤดูฝน ส่วนฤดูกาลที่เหลือจะมีความผสมกันระหว่างวันที่ปกติและวันที่มีระดับที่สูงซึ่งก็จะไม่สูงมากเท่ากับฤดูร้อนและช่วงคาบเกี่ยวระหว่างฤดูร้อนและฤดูหนาวดังภาพที่ 4.3 และภาพที่ 4.4 โดยทั้งสองจุดรับข้อมูลนั้นให้ข้อมูลที่สอดคล้องกันคือ ฤดูร้อนและช่วงคาบเกี่ยวระหว่างฤดูหนาวกับฤดูร้อนมีระดับ $PM_{2.5}$ ที่สูงเกินกว่าระดับมาตรฐานจำนวนมาก

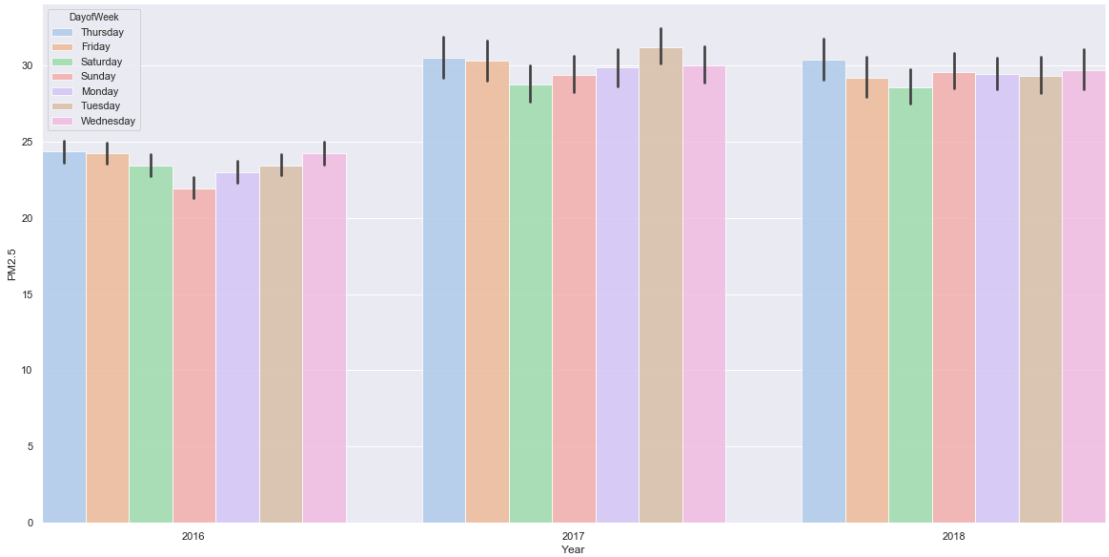


ภาพที่ 4.5 ข้อมูล $PM_{2.5}$ ในแต่ละฤดูกาล ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่

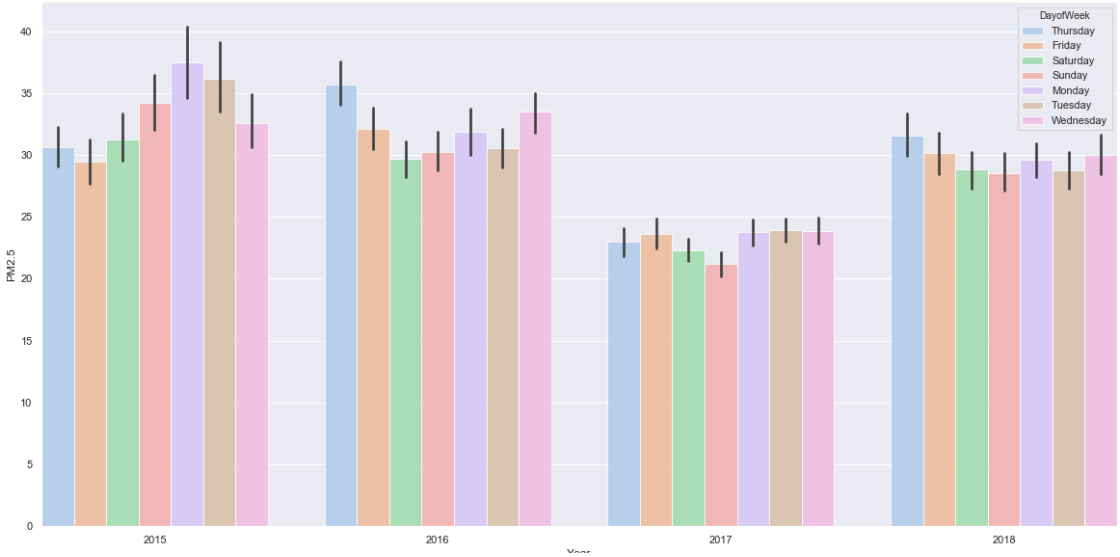


ภาพที่ 4.6 ข้อมูล $PM_{2.5}$ ในแต่ละฤดูกาล ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่

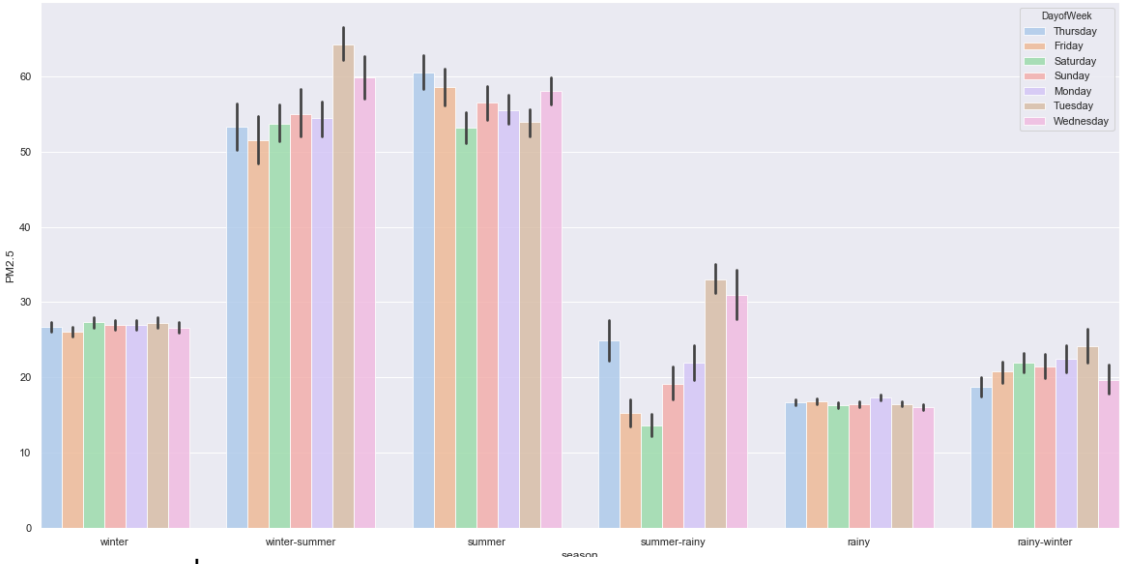
ต่อไปเป็นการวิเคราะห์วันในสัปดาห์ (จันทร์-อาทิตย์) โดยวิธีการวิเคราะห์นั้นก็เหมือนกับด้านบนโดยจะทำการวิเคราะห์ในแต่ละปีหลังจากนั้นจะทำการวิเคราะห์ในแต่ละฤดูกาลจากข้อมูลที่มีไม่สามารถบอกได้ว่าวันในสัปดาห์นั้นมีความสัมพันธ์กับปัญหา $PM_{2.5}$ หรือไม่เพราะจากข้อมูลที่มีและการสังเกตจากภาพที่ 4.5 จะเห็นได้ว่าไม่สามารถบอกได้อย่างชัดเจนว่าวันไหนที่จะมีค่า $PM_{2.5}$ สูงที่สุดหรือต่ำที่สุดในแต่ละปีนั้นวันที่มีค่าสูงหรือต่ำกันนั้นจะสลับกันไปเรื่อยๆ ไม่มีรูปแบบที่แน่นอนดังนั้นจึงสรุปไม่ได้ว่ามีความสัมพันธ์หรือไม่ หากมองในแง่ของฤดูกาลก็ไม่สามารถหาความสัมพันธ์ได้อีกเช่นเดียวกันเหตุผลเช่นเดียวกับที่กล่าวไว้ด้านบนคือ ไม่สามารถหารูปแบบที่ตายตัวของการเกิด $PM_{2.5}$ ในแต่ละวันได้บางวันมีค่าสูง บางวันมีค่าต่ำในแต่ละฤดูกาลที่ไม่ซ้ำกันทำให้ไม่สามารถบอกได้ว่าวันในสัปดาห์มีความสัมพันธ์กับ $PM_{2.5}$ ในรูปแบบใด



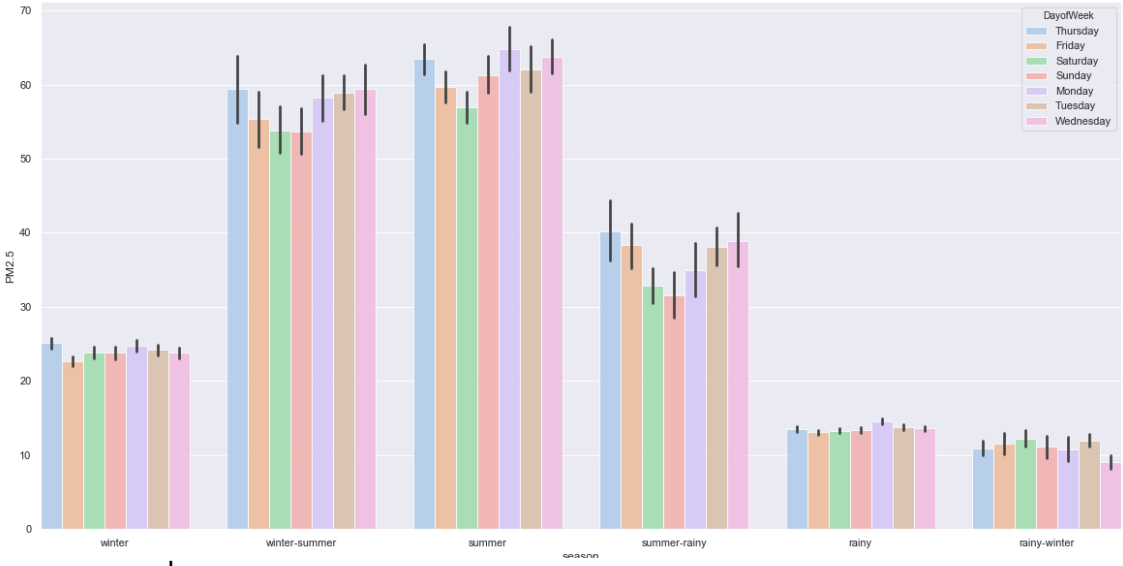
ภาพที่ 4.7 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในแต่ละปี ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่



ภาพที่ 4.8 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในปี ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่

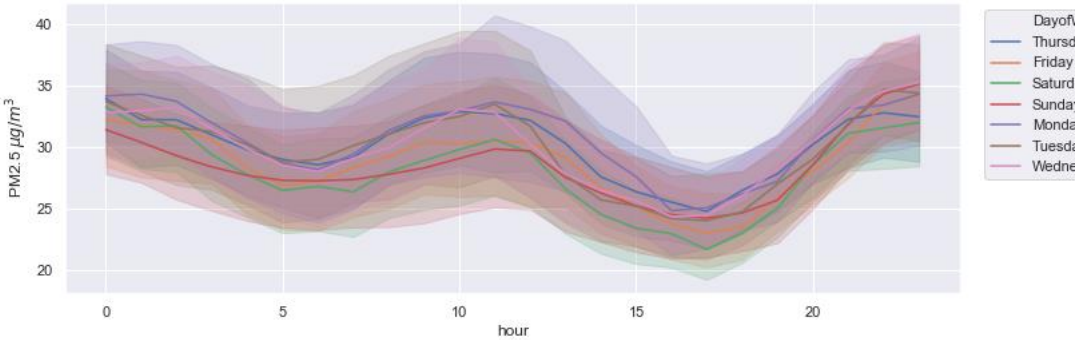


ภาพที่ 4.9 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในแต่ละฤดูกาล ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่

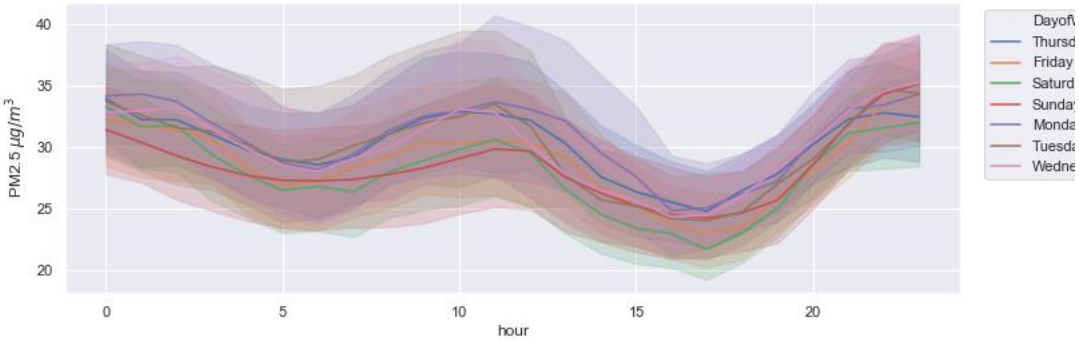


ภาพที่ 4.10 ข้อมูล $PM_{2.5}$ ของวันในสัปดาห์ในแต่ละฤดูกาล ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่

ถัดจากวิเคราะห์วันในสัปดาห์เป็นการวิเคราะห์ค่า $PM_{2.5}$ ในแต่ละชั่วโมงของวันในสัปดาห์ว่ามีลักษณะอย่างไรจากข้อมูลภาพที่ 4.11 และภาพที่ 4.12 สองสถานีให้ข้อมูลที่สอดคล้องกันคือ ระดับของ $PM_{2.5}$ จะเพิ่มสูงขึ้นในช่วงเวลาตี 5– 11 โมงเช้าในช่วงเวลากลางวัน หลังจากนั้นระดับจะค่อยๆ ลดลงจนถึงช่วงเย็น คือ 11.00 – 16.00 จากนั้นจากการวิเคราะห์ของผู้วิจัยคิดว่าในช่วงเย็นอากาศเหนือผิวดินจะมีความเย็นมากกว่าอากาศด้านบน จะทำให้เกิดปรากฏการณ์อุณหภูมิผกผัน (Temperature Inversion) รวมกับปริมาณการจราจรที่เพิ่มมากขึ้นดังนั้นช่วงเย็นจนถึงตอนกลางคืนประมาณ 17.00 – 24.00 ระดับของค่า $PM_{2.5}$ จะค่อยๆ เพิ่มสูงขึ้นจนถึงช่วงสูงสุดประมาณ 23.00 หรือ 24.00 เมื่อเริ่มวันใหม่ ณ ช่วงเวลา 0.00 – 5.00 ในตอนเช้าระดับของค่า $PM_{2.5}$ จะค่อยๆ ลดลงในระดับหนึ่งจากนั้นจะเกิดการวนซ้ำของรูปแบบนี้ไปเรื่อยๆ ในแต่ละวัน

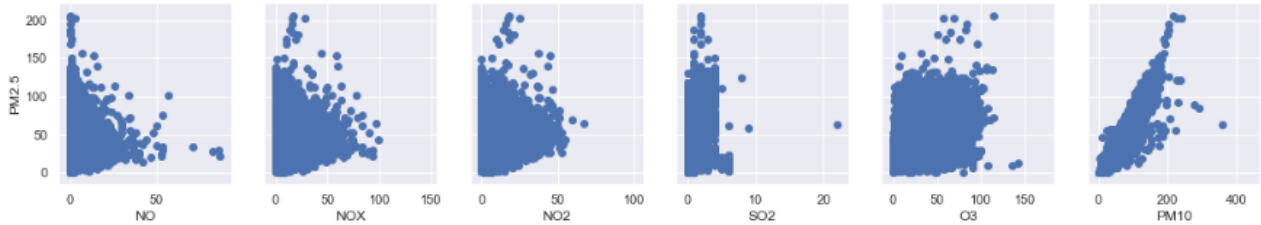


ภาพที่ 4.11 $PM_{2.5}$ ในแต่ละชั่วโมงของวันในสัปดาห์ ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่

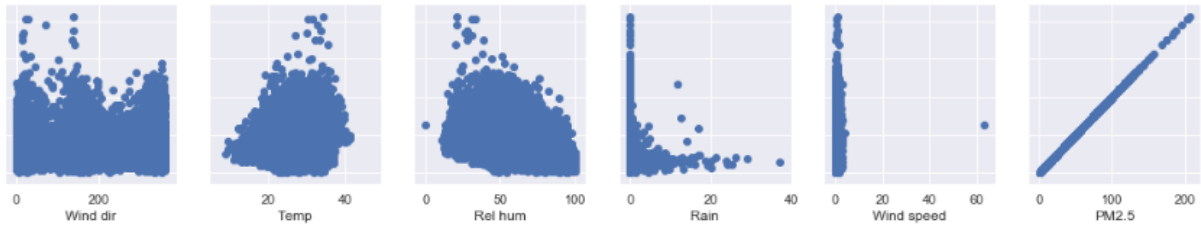


ภาพที่ 4.12 $PM_{2.5}$ ในแต่ละชั่วโมงของวันในสัปดาห์ ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่

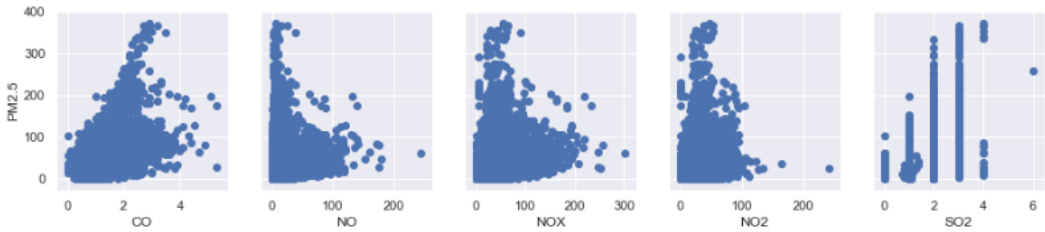
4.1.2 การวิเคราะห์จากแหล่งข้อมูลกรมควบคุมมลพิษ



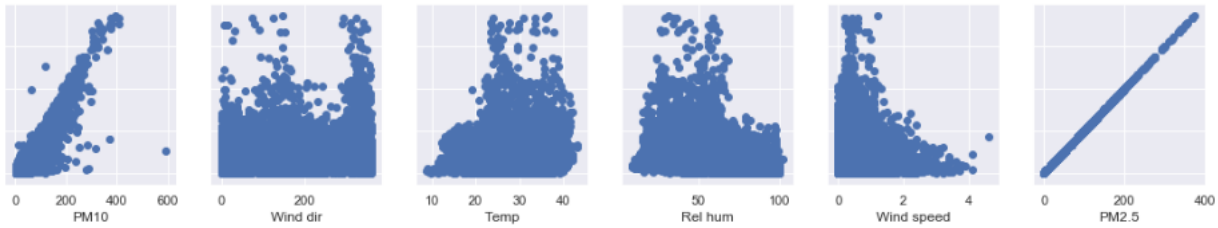
ภาพที่ 4.13 แผนภาพการกระจายของข้อมูลที่เข้าคู่กันระหว่าง $PM_{2.5}$ กับตัวแปรต่างๆ ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่



ภาพที่ 4.13 แผนภาพการกระจายของข้อมูลที่เข้าคู่กันระหว่าง $PM_{2.5}$ กับตัวแปรต่างๆ ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่ (ต่อ)



ภาพที่ 4.14 แผนภาพการกระจายของข้อมูลที่เข้าคู่กันระหว่าง $PM_{2.5}$ กับตัวแปรต่างๆ ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่



ภาพที่ 4.14 แผนภาพการกระจายของข้อมูลที่เข้าคู่กันระหว่าง PM2.5 กับตัวแปรต่างๆ ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่ (ต่อ)

จากภาพที่ 4.13 และ ภาพที่ 4.14 แสดงให้เห็นถึงการกระจายของข้อมูลที่เข้าคู่กับ $PM_{2.5}$ ในสองสถานีจากกราฟด้านบนนอกจากจะแสดงให้เห็นถึงการกระจายแล้วยังแสดงให้เห็นถึงข้อมูลที่มีความผิดปกติอยู่และยังแสดงให้เห็นถึงความสัมพันธ์ของแต่ละคุณสมบัติกับคุณสมบัติ $PM_{2.5}$

จากข้อมูลทั้งสองสถานีพบว่าคุณสมบัติที่แสดงให้เห็นถึงความสัมพันธ์เชิงเส้นกับคุณสมบัติ $PM_{2.5}$ นั้นได้แก่ PM_{10} เนื่องจากลักษณะกราฟแสดงให้ให้ถึง Strong Positive Correlation ส่วนคุณสมบัติอื่นๆ ไม่สามารถบอกได้ว่ามีความสัมพันธ์แบบไหนกับ $PM_{2.5}$ เนื่องจากลักษณะกราฟมีความกำกวมสำหรับคุณสมบัติ ความไวลม ณ จุดโรงเรียนยุพราชมีความสัมพันธ์ในเชิงตรงกันข้ามกับ $PM_{2.5}$ คือ เมื่อค่าความไวลมสูง $PM_{2.5}$ จะมีปริมาณน้อย ในทางกลับกันเมื่อค่าความไวลมต่ำ $PM_{2.5}$ จะมีค่าที่สูงขึ้นแต่ในจุดศูนย์ราชการความไวลมกลับไม่สอดคล้องกับจุดโรงเรียนยุพราชเพราะมีการกระจายของข้อมูลที่อยู่บริเวณใกล้ๆ ค่า 0 ซึ่งเป็นค่าที่น้อยมากผู้วิจัยจึงตั้งสันนิษฐานว่า ณ จุดศูนย์ราชการตัวรับข้อมูลนั้นตั้งอยู่ในจุดที่รับความไวลมได้ไม่ดีพอ ทำให้ค่าที่ได้มีค่าน้อยมาก ส่วนค่าปริมาณน้ำฝนมีการกระจายที่อยู่บริเวณใกล้ค่า 0 เช่นกันเป็นเพราะวันส่วนมากในจังหวัดเชียงใหม่ไม่มีฝนตกทำให้ไม่สามารถวัดปริมาณน้ำฝนที่ตกลงมาได้ค่าที่ได้จึงกระจายที่ 0 มากแต่ในวันที่ฝนตกค่า $PM_{2.5}$ นั้นกลับลดลงเป็นความสัมพันธ์เช่นเดียวกับค่าความไวลม

4.1.3 การวิเคราะห์จากแหล่งข้อมูล Wunderground

ในแหล่งข้อมูล Wunderground นี้จะใช้ข้อมูลได้แก่ อุณหภูมิ (Temperature), จุดไอน้ำกลั่นตัว (Dew Point), ความกดอากาศ (Pressure), ความชื้น (Humidity Relative) และความไวลม (Wind Speed), สภาพเงื่อไขในแต่ละวัน (Condition) โดยจะทำการนำคุณสมบัติเหล่านี้ไปทำการสร้างแผนภาพการกระจายโดยการนำไปเข้าคู่กับคุณสมบัติ $PM_{2.5}$ เพื่อหาความสัมพันธ์ของแต่ละคุณสมบัติ โดยทั้งสองสถานีให้ความสัมพันธ์แบบเดียวกันดังนี้

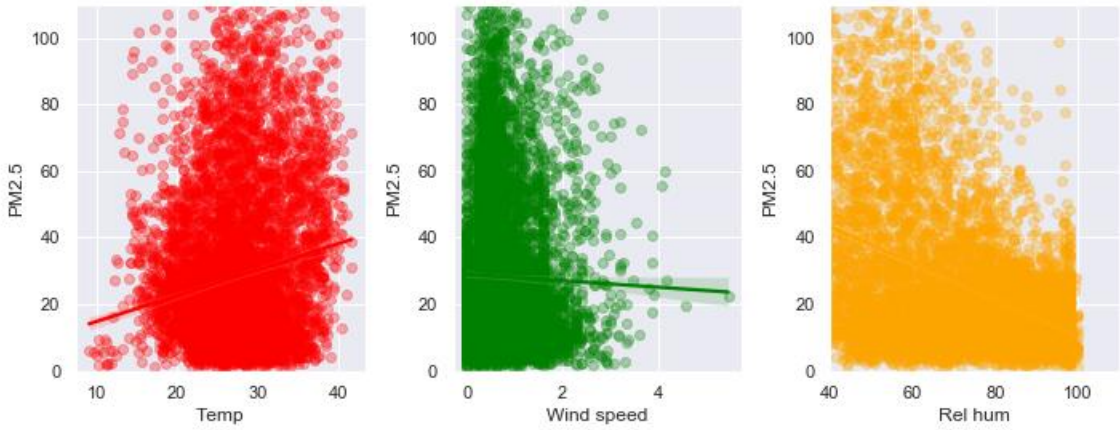
- อุณหภูมิ (Temperature) สำหรับอุณหภูมินั้นจากข้อมูลที่มีแสดงความสัมพันธ์หากดูจากเส้นตรงในภาพที่ 4.15 คือ เมื่ออุณหภูมิสูงขึ้นจะมีแนวโน้มที่จะทำให้ค่า $PM_{2.5}$ นั้นสูงขึ้นตามแต่สำหรับคุณสมบัตินี้ผู้วิจัยคิดว่าไม่ค่อยสมเหตุสมผลเนื่องจากในวันที่อุณหภูมิปกติประมาณ 20 – 30 องศาเซลเซียส ค่า $PM_{2.5}$ ก็เพิ่มสูงขึ้นได้และสังเกตจากกราฟค่า $PM_{2.5}$ ในวันที่อุณหภูมิปกติมีค่า $PM_{2.5}$ สูงเป็นจำนวนมากแต่อาจจะมีแนวโน้มที่เมื่ออุณหภูมิสูงขึ้นอาจจะส่งผลให้เกิดความร้อนซึ่งเป็นสาเหตุของการเกิดไฟป่าหรือการเผาไหม้อื่นๆ ส่งผลให้เกิดค่า $PM_{2.5}$ เพิ่มสูงขึ้น

- ค่าความไวลม (Wind Speed) ค่าความไวลมนั้นหากดูตามเส้นตรงจะพบว่ามีลักษณะเอียงลงเล็กน้อยทำให้สามารถสรุปความสัมพันธ์ได้ว่า ค่าความไวลมมีความสัมพันธ์เชิงตรงข้ามกับปัญหา $PM_{2.5}$ คือเมื่อค่าความไวลมมีค่าสูงขึ้นมีแนวโน้มที่จะทำให้ $PM_{2.5}$ นั้นลดต่ำลง เช่นเดียวกันในทางตรงกันข้ามเมื่อค่าความไวลมมีค่าต่ำลม $PM_{2.5}$ ก็จะสูงขึ้น

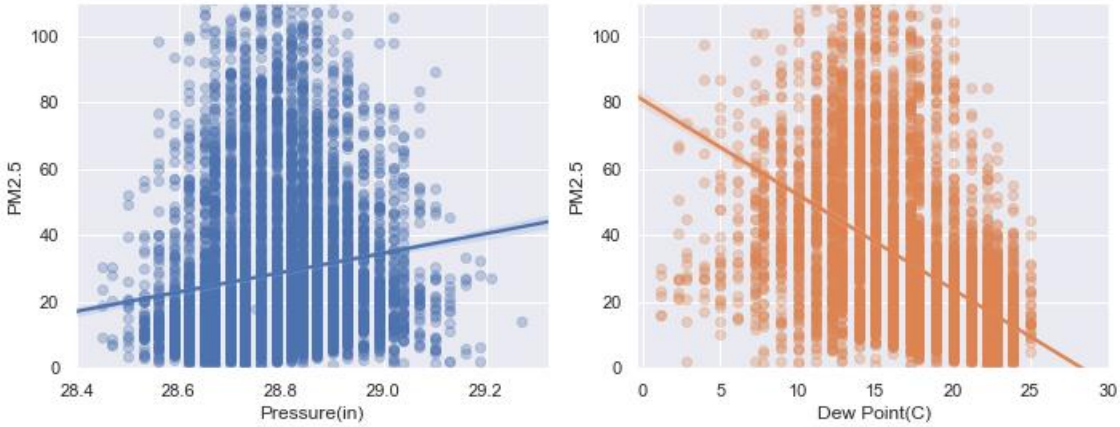
- ค่าความชื้น (Relative Humidity) ค่าความชื้นนั้นเมื่อสังเกตจากลักษณะของเส้นตรงแล้วจะเห็นว่ามีลักษณะเอียงลงมาก ดังนั้นค่าความชื้นนี้จะมีความสัมพันธ์ที่ตรงกันข้ามกับ $PM_{2.5}$ เช่นเดียวกับค่าความไวลมแต่จะมีความชัดเจนมากกว่าเนื่องจากลักษณะของการเอียงลงนั้นมีความเอียงลงมากกว่า

- ค่าจุดไอน้ำกลั่นตัว (Dew Point) สำหรับค่าจุดไอน้ำกลั่นตัวนี้จากลักษณะเส้นตรงจะเห็นได้อย่างชัดเจนเลยว่ามีลักษณะเอียงลงคล้ายๆ กับค่าความชื้นจึงทำให้ทั้งสองมีคุณสมบัติเดียวกันคือมีความสัมพันธ์ในเชิงตรงข้ามกับ $PM_{2.5}$ อาจจะเป็นเพราะความชื้นที่อยู่ในอากาศ เมื่อจุดที่ไอน้ำในอากาศมีอุณหภูมิที่ทำให้เกิดการเปลี่ยนแปลงสถานะจากก๊าซเป็นของเหลวสูงขึ้น ทำให้ไอน้ำในอากาศเปลี่ยนตัวเป็นน้ำหรือของเหลวได้ยากขึ้นทำให้ความชื้นในอากาศมีมากขึ้น ด้วยเหตุนี้อาจส่งผลต่อค่า $PM_{2.5}$

- ค่าความกดอากาศ (Pressure) จากลักษณะของเส้นตรงที่มีลักษณะเอียงขึ้นดังนั้นค่าความกดอากาศจึงมีความสัมพันธ์กับ $PM_{2.5}$ ในเชิงทิศทางเดียวกันเมื่อค่าความกดอากาศเพิ่มขึ้น $PM_{2.5}$ มีแนวโน้มที่จะเพิ่มสูงขึ้นตามแต่อาจจะเพิ่มไม่มากนักเนื่องจากลักษณะเส้นตรงไม่มีความชันเท่าที่ควร

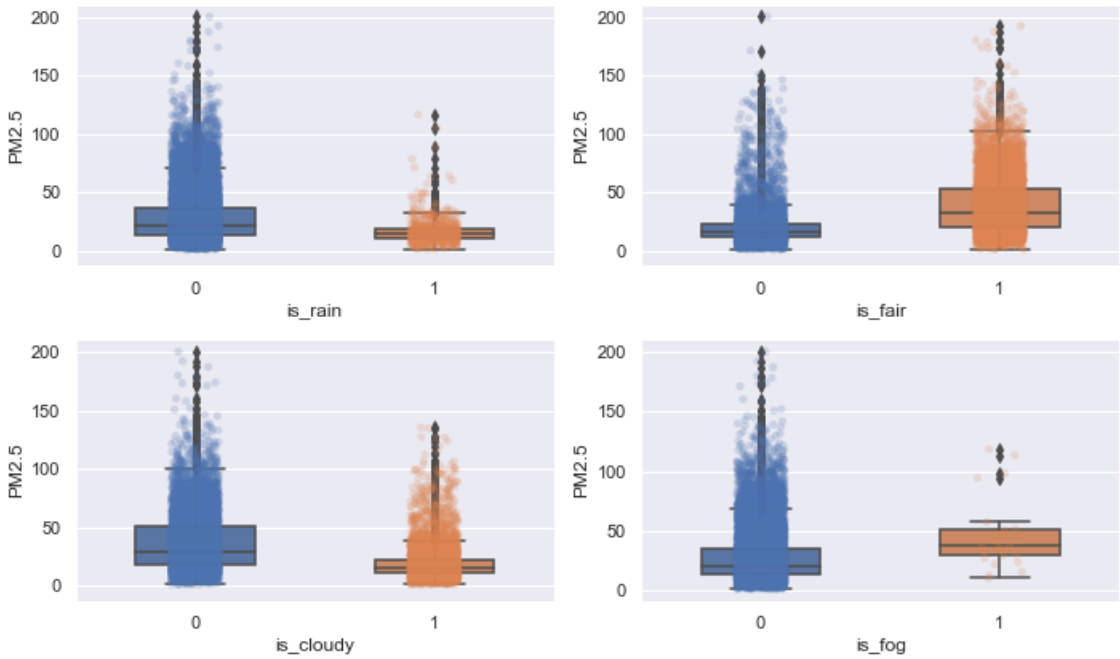


ภาพที่ 4.15 แผนภาพการกระจายของอุณหภูมิ ความไวลม ความชื้นที่เข้าคู่กับ $PM_{2.5}$



ภาพที่ 4.16 แผนภาพการกระจายของความกดอากาศและจุดไอน้ำกลั่นตัวที่เข้าคู่กับ $PM_{2.5}$

ขั้นตอนต่อไปสำหรับข้อมูลจาก Wunderground เป็นการวิเคราะห์เงื่อนไขในแต่วัน โดยเงื่อนไขที่ได้เลือกมาทำการวิเคราะห์กับปัญหา $PM_{2.5}$ นั้นได้แก่ เงื่อนไขฝนตก (is_rain), เงื่อนไขอากาศดี (is_fair), เงื่อนไขเมฆ (is_cloudy) และเงื่อนไขหมอก (is_fog) การวิเคราะห์ทำได้โดยการวิเคราะห์จากภาพที่ 4.17

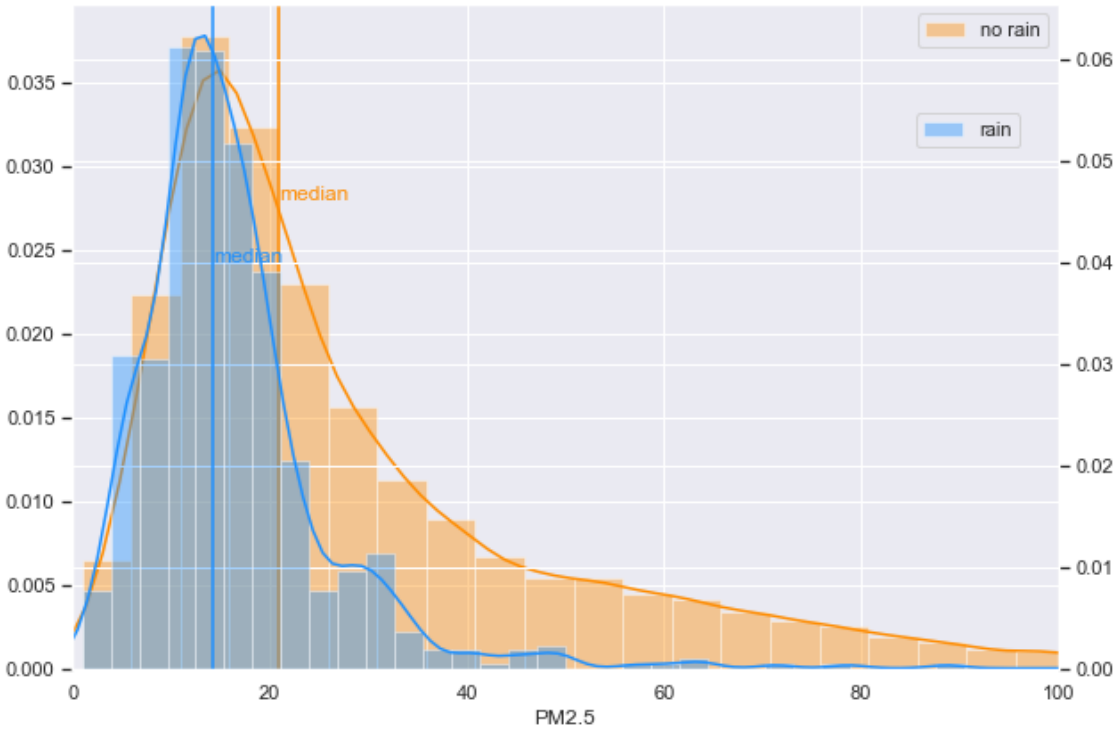


ภาพที่ 4.17 เงื่อนไขสภาพอากาศในแต่ละวันกับ $PM_{2.5}$

จากภาพที่ 4.17 สามารถวิเคราะห์ได้ว่าในวันที่ฝนตกนั้นค่า $PM_{2.5}$ มีระดับที่ลดลงอย่างชัดเจนส่วนในวันที่ไม่มีฝนตกค่า $PM_{2.5}$ กลับสูงขึ้นส่วนอีกคุณสมบัติหนึ่งที่มีลักษณะเดียวกันกับฝนตกได้แก่วันที่มีเมฆแต่ในคุณสมบัตินี้นั้นอาจจะไม่ชัดเจนเท่ากับวันที่ฝนตกเนื่องจากระดับของ

$PM_{2.5}$ รวมกับการกระจายมีค่าสูงกว่าวันที่ฝนตกแต่ก็ยังน้อยกว่าวันที่ไม่มีเมฆจึงตัดสินใจว่าน่าจะมีความสัมพันธ์เช่นเดียวกับวันที่ฝนตก ส่วนวันที่อากาศดีนั้นจากข้อมูลจะมีแนวโน้มที่ค่า $PM_{2.5}$ นั้นสูงกว่าวันที่สภาพอากาศไม่ดี (ซึ่งอาจหมายถึงเป็นวันที่มีลมมาก หรือ ฝนตก) แต่ก็ยังมีความไม่ชัดเจนปะปนอยู่เช่นเดียวกันส่วนคุณสมบัติหมอกในวันที่มีหมอกจากข้อมูลจะมีค่า $PM_{2.5}$ ที่ต่ำกว่าวันที่ไม่มีหมอกแต่เนื่องจากในวันที่มีหมอกมีข้อมูลน้อยจึงทำให้ผู้วิจัยไม่สามารถสรุปความสัมพันธ์ของคุณสมบัตินี้ได้

ภาพที่ 4.18 แสดงให้เห็นว่าในวันที่ฝนตกนั้นมีค่า $PM_{2.5}$ ที่น้อยกว่าวันที่ฝนไม่ตก โดยสังเกตจากค่ากลางของในแต่ละวัน ในวันที่ฝนตกจะมีค่ากลางที่น้อยกว่าวันที่ฝนไม่ตก

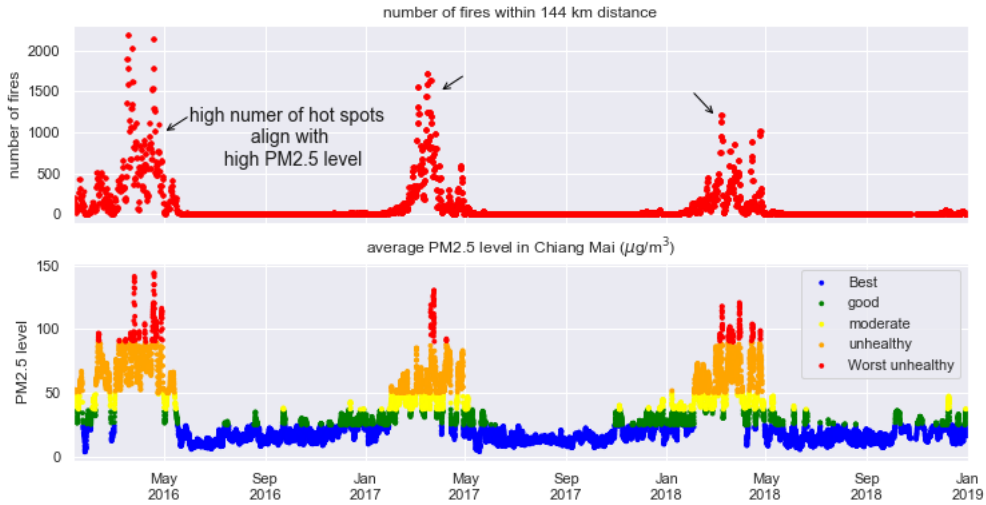


ภาพที่ 4.18 การกระจายของค่า $PM_{2.5}$ ในวันที่ฝนตก

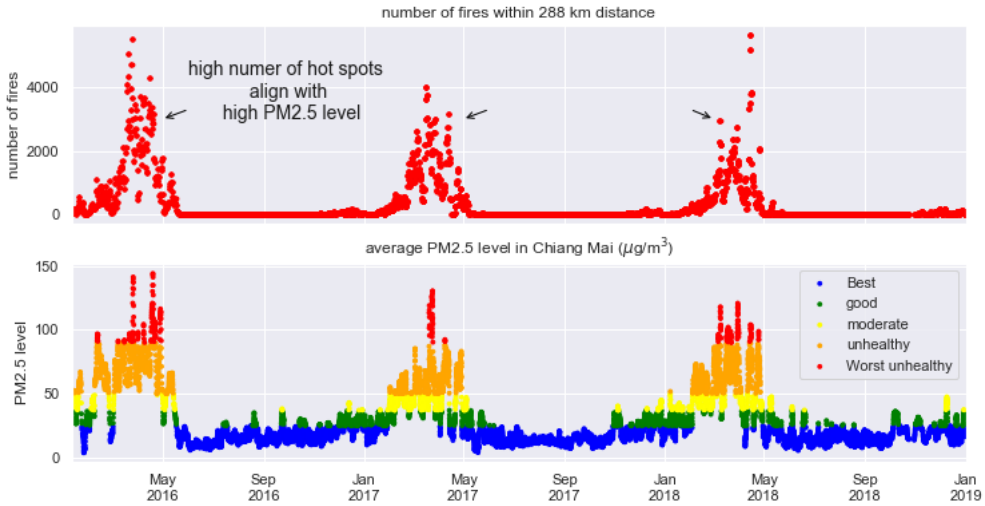
4.1.4 การวิเคราะห์จากแหล่งข้อมูล FIRMS

ข้อมูลส่วนนี้เกี่ยวกับจุดความร้อนที่เกิดขึ้น ณ เวลานั้นๆ ที่ถูกตรวจจับโดยอุปกรณ์ที่ติดตั้งอยู่บนดาวเทียมในความละเอียด 375 เมตร ปัจจัยไฟหรือจุดความร้อนนี้เป็นปัจจัยที่ต้องอาศัยช่วงเวลาระยะหนึ่งจึงจะส่งผลต่อการเกิดปัญหา $PM_{2.5}$ ดังนั้นผู้วิจัยจึงต้องปรับแต่งระยะเวลาให้มีความเหมาะสมโดยจากการคำนวณผู้วิจัยได้เลือกใช้จุดความร้อนที่เกิดในระยะ 144 กิโลเมตร, 288 กิโลเมตร, 432 กิโลเมตร และ มากกว่า 432 กิโลเมตรแต่น้อยกว่า 3000 กิโลเมตรในการวิเคราะห์ โดยแต่ละระยะทางที่เลือกมานั้นคำนวณมาจากค่าความไวลมจาก แหล่งข้อมูล Wunderground ค่าเฉลี่ยของความไวลมในแต่ละปีคือ 6 กิโลเมตรต่อชั่วโมง ดังนั้นใน 1 วัน

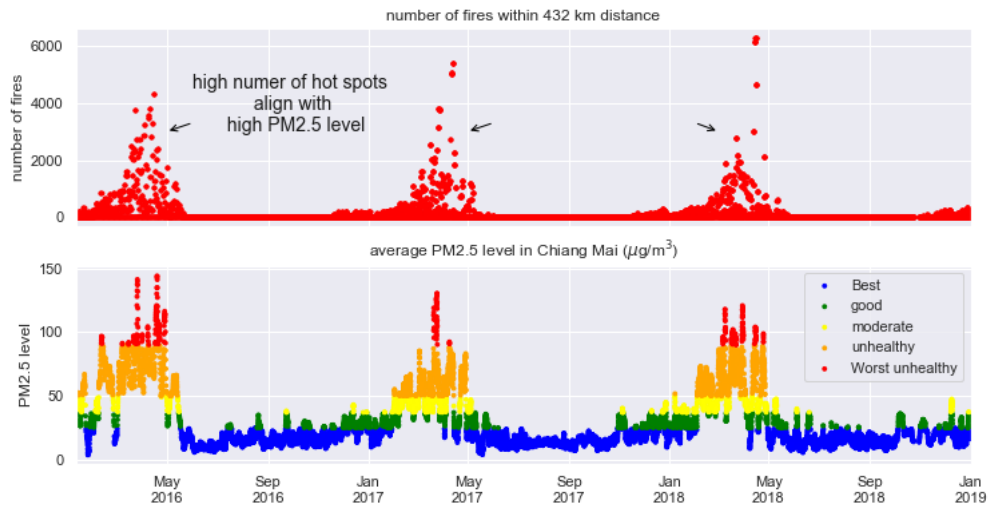
ลมสามารถพัดพา $PM_{2.5}$ ในระยะ 144 กิโลเมตรเข้ามาหาจังหวัดเชียงใหม่ได้ ระยะทาง 288 กิโลเมตรคือเวลา 2 วัน ส่วนระยะทาง 432 กิโลเมตรใช้เวลาประมาณ 3 วัน แต่มากกว่า 432 กิโลเมตรนั้นใช้เวลาหลายวันมากๆ ขึ้นอยู่กับระยะทางผู้วิจัยเลือกใช้ระยะทางไม่เกิน 1500 ระยะทางครึ่งหนึ่งจาก 3000 เพราะจุดความร้อนไม่อาจจะไม่เกิด ณ จุดที่ 3000 กิโลเมตรและเนื่องจากคาดเดาไม่ได้จึงใช้ระยะทางครึ่งหนึ่งในการคำนวณเวลาซึ่งใช้ระยะเวลาในการเดินทางประมาณ 11 วัน จากข้อมูลจะเห็นได้ว่าหลังจากการปรับช่วงเวลาให้เหมาะสมแล้วแม้จะมีความคลาดเคลื่อนเล็กน้อยในเรื่องของการจัดช่วงเวลาแต่ก็ถือว่ายังสามารถนำมาใช้ได้ จะเห็นได้ว่าช่วงที่จำนวนของจุดความร้อนที่ตรวจจับได้เพิ่มขึ้นและพลังงานที่แผ่ออกมาจากจุดความร้อนนั้นจะเพิ่มสูงขึ้นในช่วงที่ค่า $PM_{2.5}$ นั้นเพิ่มสูงขึ้นพอดีดังนั้นจึงถือว่าปัจจัยจุดความร้อนก็เป็นอีกหนึ่งปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ ในจังหวัดเชียงใหม่



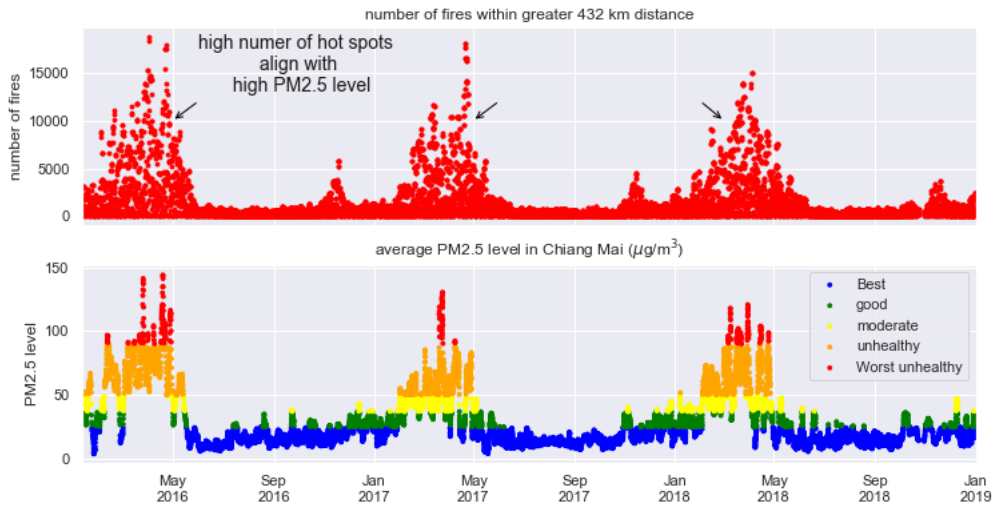
ภาพที่ 4.19 การเปรียบเทียบจำนวนจุดความร้อนในระยะ 144 กิโลเมตรกับ $PM_{2.5}$



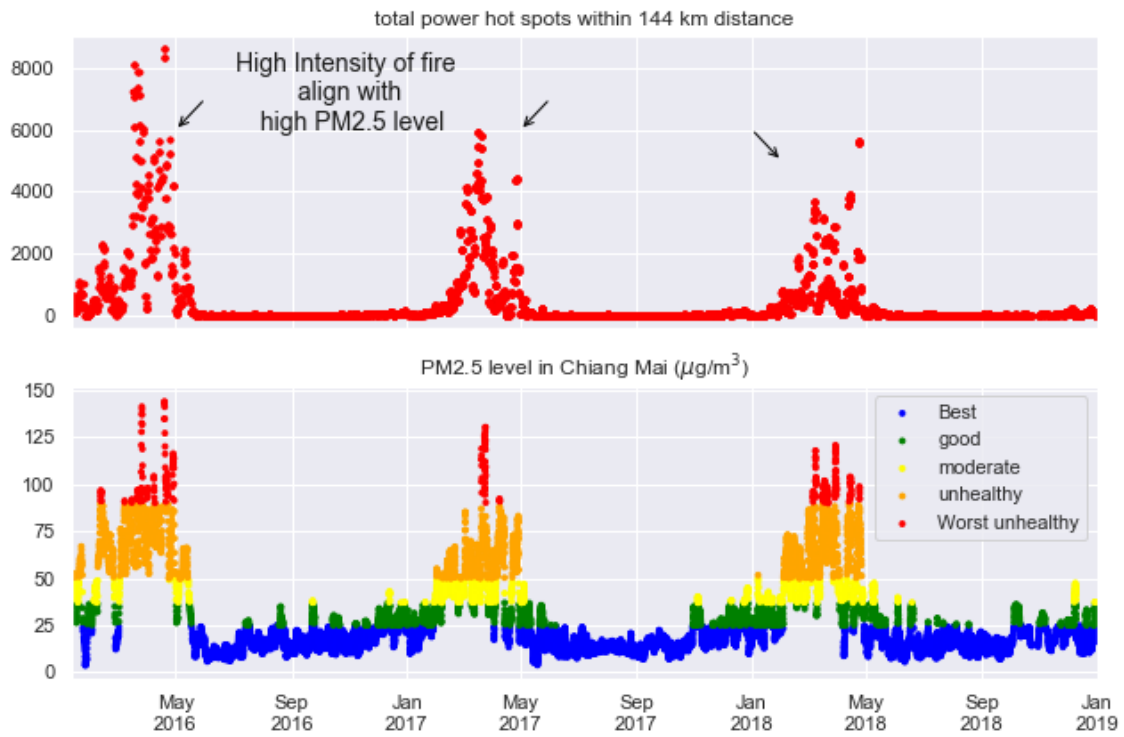
ภาพที่ 4.20 การเปรียบเทียบจำนวนจุดความร้อนในระยะ 288 กิโลเมตรกับ $PM_{2.5}$



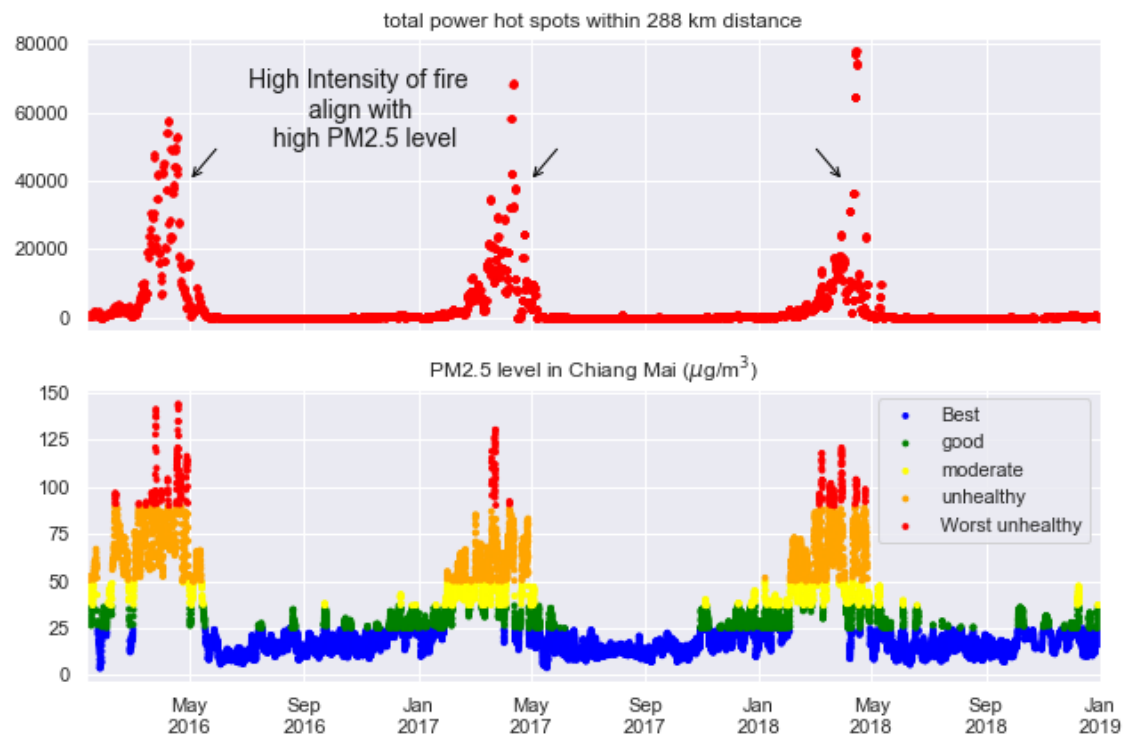
ภาพที่ 4.21 การเปรียบเทียบจุดความร้อนในระยะ 432 กิโลเมตรกับ $PM_{2.5}$



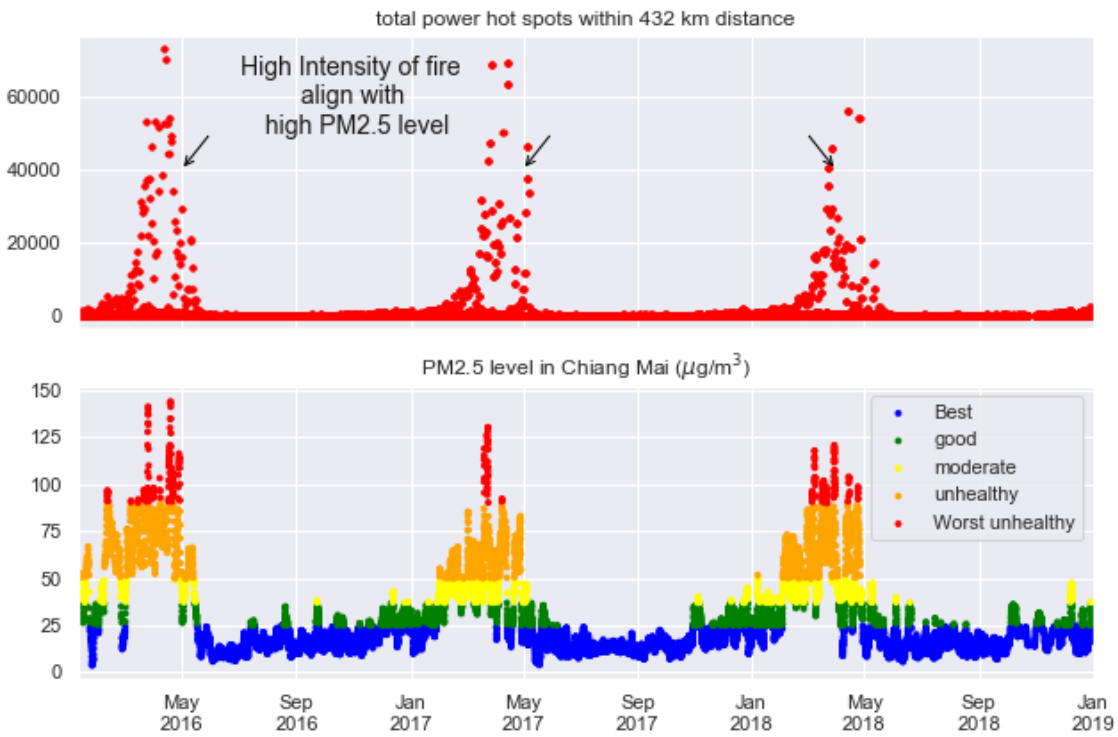
ภาพที่ 4.22 การเปรียบเทียบจุดความร้อนในระยะมากกว่า 432 กิโลเมตรแต่น้อยกว่า 3000 กิโลเมตร กับ $PM_{2.5}$



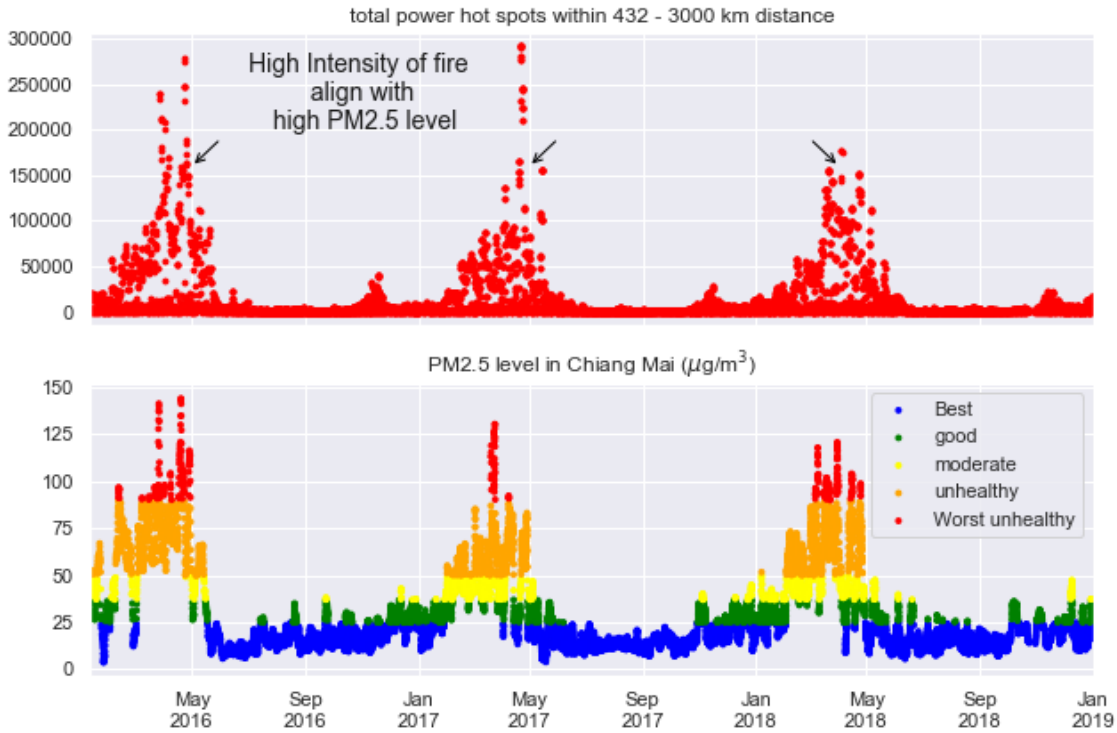
ภาพที่ 4.23 การเปรียบเทียบพลังงานที่แผ่จากจุดความร้อนในระยะ 144 กิโลเมตรกับ $PM_{2.5}$



ภาพที่ 4.24 การเปรียบเทียบพลังงานที่แผ่จากจุดความร้อนในระยะ 288 กิโลเมตรกับ $PM_{2.5}$



ภาพที่ 4.25 การเปรียบเทียบพลังงานที่แผ่จากจุดความร้อนในระยะ 432 กิโลเมตรกับ $PM_{2.5}$



ภาพที่ 4.26 การเปรียบเทียบพลังงานที่แผ่ออกจากจุดความร้อนระยะมากกว่า 432 กิโลเมตร แต่น้อยกว่า 3000 กิโลเมตรกับ $PM_{2.5}$

4.2 การวิเคราะห์หาสาเหตุของการเกิดปัญหา $PM_{2.5}$ ด้วยวิธีการเรียนรู้ของเครื่อง

สำหรับการวิเคราะห์หาสาเหตุของการเกิดปัญหา $PM_{2.5}$ โดยการนำใช้เทคนิคการเรียนรู้ของเครื่องเข้ามาช่วยนั้นจะอาศัยให้แบบจำลองนั้นช่วยเลือกคุณสมบัติที่ส่งผลต่อการทำนายค่า $PM_{2.5}$ ให้เองโดยอัติโนมัติ สำหรับแบบจำลองที่สามารถทำแบบนี้ได้นั้น ได้แก่ Random Forest, Extreme Gradient Boosting โดยทั้งสองแบบจำลองนี้ถือเป็นแบบจำลองที่ประกอบด้วย Decision Tree จำนวนมากที่ช่วยกันเรียนรู้ด้วยคุณสมบัติของ Decision Tree นั้นส่งผลให้แบบจำลองทั้งสองสามารถบอกได้ว่าคุณสมบัติใดที่ส่งผลต่อการทำนายค่า $PM_{2.5}$ ดังนั้นผู้วิจัยจึงจะใช้ข้อดีนี้ในการช่วยวิเคราะห์ถึงสาเหตุของปัจจัยที่ส่งผลกระทบต่อค่า $PM_{2.5}$ โดยผลลัพธ์ที่ได้มีดังนี้

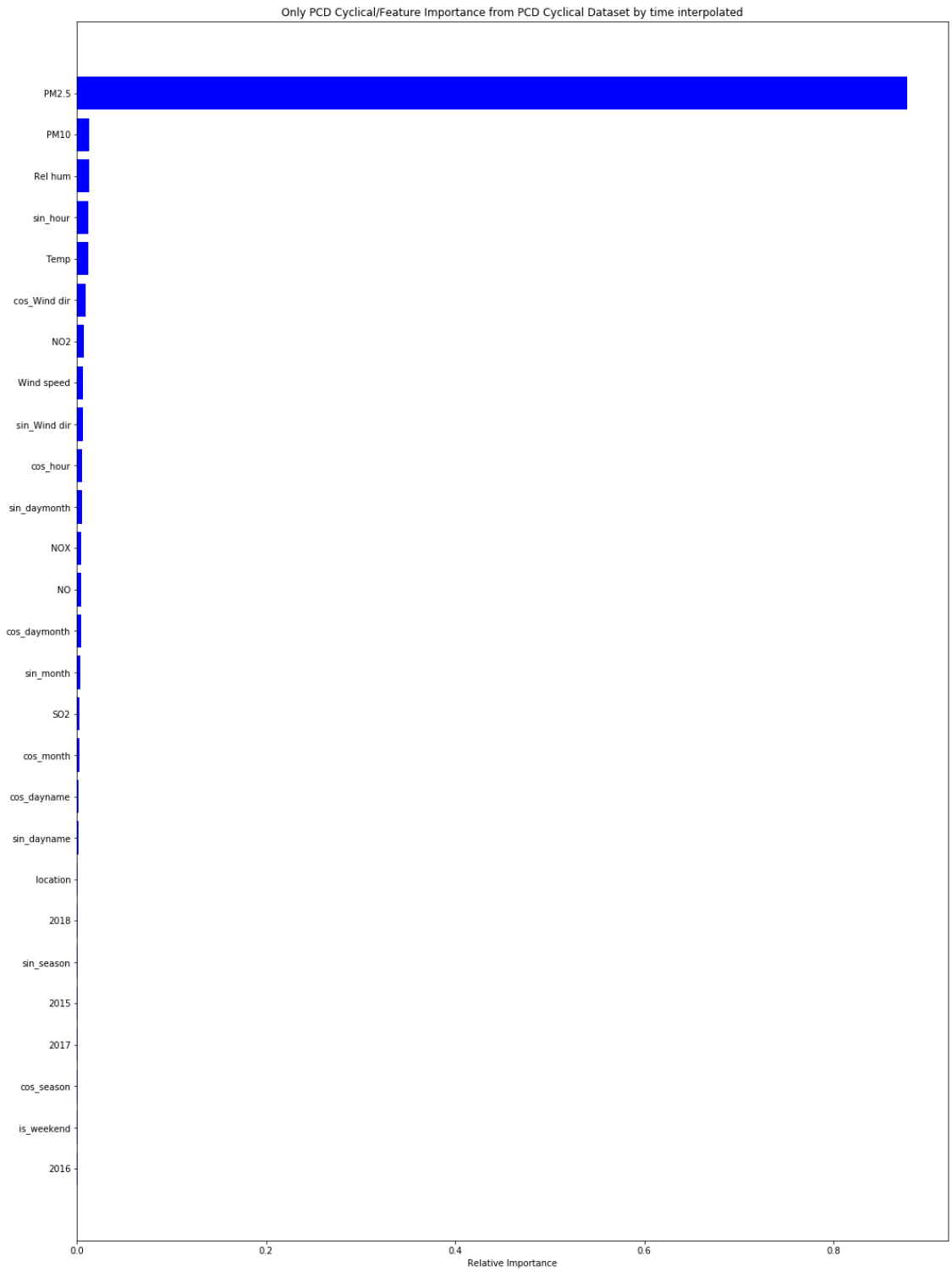
- แบบจำลอง Random Forest

จากคุณสมบัติที่กล่าวมาข้างต้นทำให้ผู้วิจัยสามารถแสดงกราฟความสำคัญของคุณสมบัติหรือตัวแปรต่างๆ ในข้อมูลได้ โดยเมื่อทำการฝึกฝนแบบจำลองแล้วทำการแสดงกราฟความสำคัญของตัวแปรอิสระที่อยู่ในชุดข้อมูลโดยชุดข้อมูลที่มีแบ่งเป็นสองส่วนด้วยกันได้แก่ ชุดข้อมูลที่ใช้วิธีข้อมูลวนซ้ำ (Cyclical Dataset) และชุดข้อมูลแบบดัมมี่ (Dummy Dataset) พบว่าคุณสมบัติที่ส่งผลต่อแบบจำลองนั้นมีดังนี้

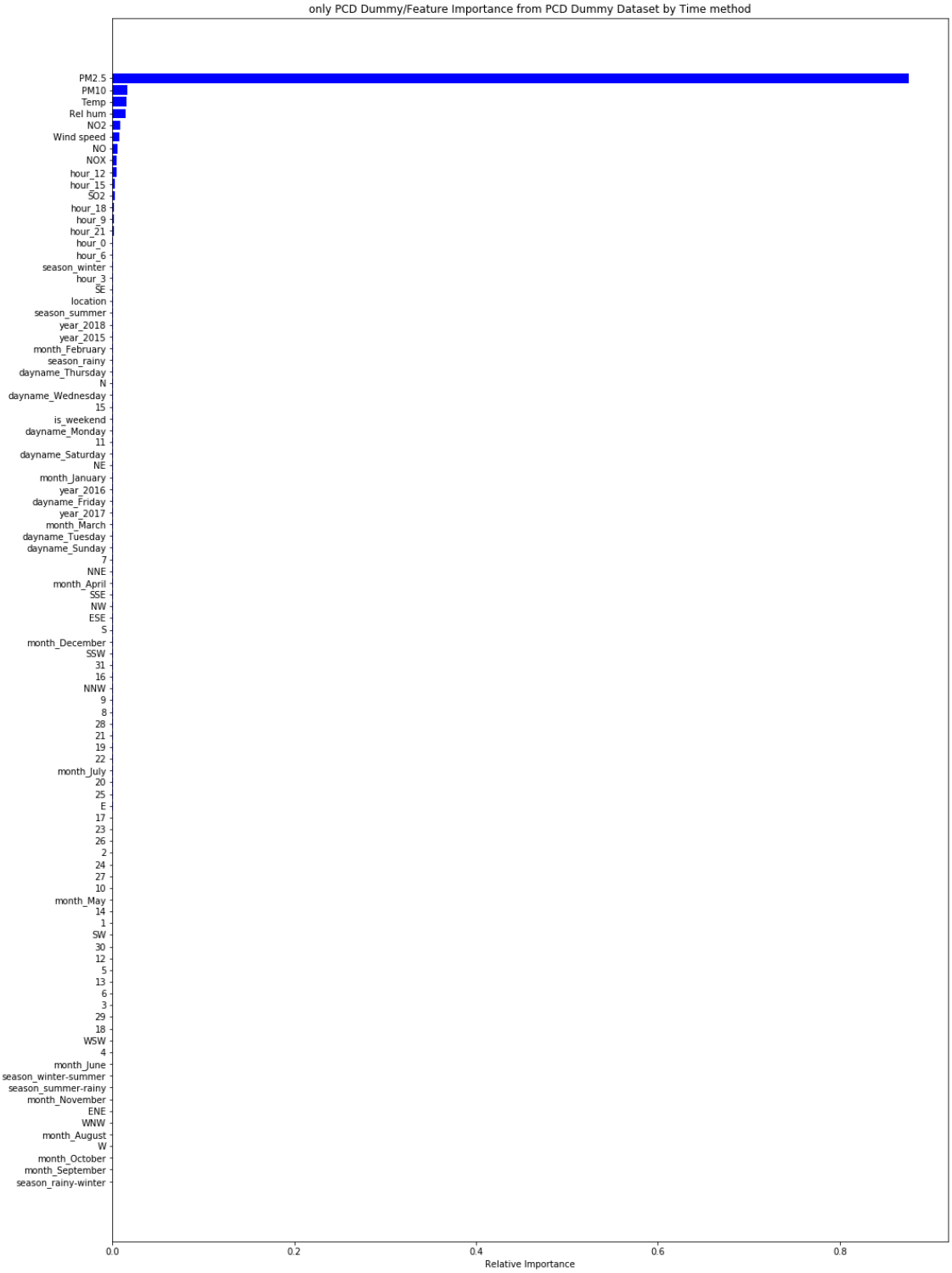
Cyclical Dataset ได้แก่ $PM_{2.5}$, PM_{10}, Relative Humidity, Sin_hour, Temperature, Cos_Wind direction, NO_2, Wind Speed, Sin_Wind direction, Cos_hour, Sin_daymonth, NO_x, NO, Cos_daymonth, Sin_month, SO_2, Cos_month, Cos_dayname, Sin_dayname

Dummy Dataset ได้แก่ $PM_{2.5}$, PM_{10}, Temperature, Relative Humidity, NO_2, Wind Speed, NO, NO_x, hour_12, hour_15, SO_2, hour_18, hour_9, hour_21

สำหรับข้อมูลทั้งสองเป็นชุดข้อมูลที่ถูกปรับช่วงเวลาแล้วหรือข้อมูลที่ใช้ฝึกฝนจะเป็นข้อมูลในอดีตหรือ Lag 1 ขั้นตอน เช่นข้อมูล ณ เวลา 12.00 ที่ใช้ฝึกฝนแบบจำลองคือข้อมูลที่เกิดขึ้น ณ เวลา 9.00 เพราะเป็นการใช้ข้อมูลในอดีตในการทำนายอนาคต จากกราฟทั้งสองจะเห็นว่า $PM_{2.5}$, PM_{10} ในอดีตมีความสัมพันธ์กับค่า $PM_{2.5}$ ในปัจจุบันมากที่สุดตามลำดับ รองลงมาหลังจากนั้นจะเป็นข้อมูลเกี่ยวกับสภาพอากาศและสารมลพิษในอากาศที่มีผลกระทบต่อการทำนายแบบจำลอง ส่วนหลังจากนั้นจะเป็นเกี่ยวกับเวลา ดังภาพที่ 4.27 และ ภาพที่ 4.28 ตามลำดับ



ภาพที่ 4.27 ความสำคัญของตัวแปรอิสระในชุดข้อมูลแบบวนซ้ำ (Cyclical Dataset) จากกรมควบคุมมลพิษ



ภาพที่ 4.28 ความสำคัญของตัวแปรอิสระจากชุดข้อมูลแบบดัมมี่ (Dummy Dataset) จากกรมควบคุมมลพิษ

4.3 การหาชุดข้อมูลที่เหมาะสมสำหรับการทำนายผล

ขั้นตอนนี้เป็นการหาชุดข้อมูลที่ทำให้แบบจำลองมีประสิทธิภาพสูงที่สุดโดยชุดข้อมูลที่จะนำมาฝึกฝนแบบจำลองเพื่อทำการหานั้นมีมากถึง 44 ชุดข้อมูลโดยจะนำไปฝึกฝนให้แต่ละแบบจำลองที่ได้เลือกมาได้แก่ Multiple Linear Regression, Random Forest, Extreme Gradient Boosting, Neural Network (Keras, Sklearn) โดยผลที่ได้วัดโดยข้อมูลทดสอบอัตราส่วน 30 เปอร์เซ็นต์ดังนี้

4.3.1 Multiple Linear Regression

ตารางที่ 4.1 ผลลัพธ์จากการฝึกฝนแบบจำลอง Multiple Linear Regression

Dataset \ Loss Function		Root Mean Square Error	Mean Absolute Error	R^2
PCD Cyclical	Mean	8.736672	5.330405	0.887462
	Time	8.898943	5.416548	0.887758
	Drop Missing Value	9.004609	5.617683	0.877650
	Berkeley Earth	8.677780	5.470895	0.878779
PCD Dummy	Mean	8.685677	5.273939	0.888772
	Time	8.877493	5.378876	0.888298
	Drop Missing Value	8.983390	5.557908	0.878226
	Berkeley Earth	8.636752	5.443873	0.879922
Wunderground	Mean	16.644006	11.186098	0.398948
	Time	17.020814	11.549886	0.398344
	Drop Missing Value	17.366248	11.882481	0.378970
	Berkeley Earth	17.313051	11.774419	0.402207
FIRMS	Mean	14.185629	9.842219	0.544493
	Time	14.487767	10.071243	0.543474
	Drop Missing Value	14.696540	10.263384	0.574134
	Berkeley Earth	14.796022	10.426791	0.563313
PCD Wunderground Cycilcal	Mean	7.484652	5.050960	0.878454
	Time	7.650105	5.132436	0.878459
	Drop Missing Value	8.041046	5.349618	0.866855

ตารางที่ 4.1 ผลลัพธ์จากการฝึกฝนแบบจำลอง Multiple Linear Regression (ต่อ)

Dataset \ Loss Function		Root Mean Square Error	Mean Absolute Error	R^2
	Berkeley Earth	7.869511	5.289309	0.876491
PCD Wunderground Dummy	Mean	7.427363	5.012961	0.880308
	Time	7.546929	5.068088	0.881715
	Drop Missing Value	7.955840	5.282306	0.869662
	Berkeley Earth	7.802811	5.244430	0.878576
PCD FIRMS Cyclical	Mean	7.785382	5.044095	0.862799
	Time	7.463168	4.911497	0.878854
	Drop Missing Value	7.978975	5.229274	0.874473
	Berkeley Earth	8.032178	5.241059	0.871309
PCD FIRMS Dummy	Mean	7.749413	5.017332	0.864064
	Time	7.430127	4.887715	0.879924
	Drop Missing Value	7.943408	5.211927	0.875590
	Berkeley Earth	8.000331	5.213449	0.872328
PCD Wunderground FIRMS Cyclical	Mean	7.614001	4.972134	0.877562
	Time	7.883079	5.250689	0.877618
	Drop Missing Value	7.703972	5.088705	0.879333
	Berkeley Earth	7.709414	5.102752	0.886326
PCD Wunderground FIRMS Dummy	Mean	7.575890	4.958835	0.878785
	Time	7.821126	5.211643	0.879534
	Drop Missing Value	7.669318	5.058312	0.880416
	Berkeley Earth	7.692033	5.087515	0.886838
Wunderground FIRMS	Mean	13.517704	9.028499	0.614082
	Time	13.886042	9.520745	0.620265
	Drop Missing Value	13.857276	9.447894	0.609596
	Berkeley Earth	13.890786	9.435599	0.630960

จากตารางที่ 4.1 พบว่าจากการฝึกฝนแบบจำลองด้วย Multiple Linear Regression พบว่าชุดข้อมูลที่ทำให้แบบจำลองนี้มีประสิทธิภาพสูงสุดได้แก่ ชุดข้อมูลที่จับคู่กันระหว่าง PCD กับ FIRMS ที่ทำการเติมค่าที่หายไปด้วยวิธี Time Function และแก้ข้อมูลที่มีลักษณะวนซ้ำเป็นแบบวงกลมและชุดข้อมูลที่เข้าคู่กันระหว่าง PCD กับ FIRMS ที่ทำการเติมค่าที่หายไปด้วย Time Function และแก้ไขข้อมูลที่มีลักษณะแบบประเภทให้กลายเป็นดัมมี่ โดยทั้งสองชุดข้อมูลนี้ให้ประสิทธิภาพแทบจะเหมือนกันต่างกันเพียงเล็กน้อยดังนั้นผู้วิจัยจึงให้ทั้งสองชุดข้อมูลมีประสิทธิภาพเท่ากัน

4.3.2 Extreme Gradient Boosting

ตารางที่ 4.2 ผลลัพธ์การฝึกฝนแบบจำลองด้วย Extreme Gradient Boosting

Dataset \ Performance Measure		Root Mean Square Error (μg/m^3)	Mean Absolute Error (μg/m^3)	R^2
PCD Cyclical	Mean	7.858620	4.737447	0.908946
	Time	7.459942	4.707408	0.921123
	Drop Missing Value	8.067576	4.939212	0.901789
	Berkeley Earth	7.753405	4.907260	0.903229
PCD Dummy	Mean	7.783761	4.681420	0.910673
	Time	7.355479	4.656890	0.923317
	Drop Missing Value	7.863469	4.914729	0.906695
	Berkeley Earth	7.745159	4.885706	0.903434
Wunderground	Mean	14.785194	9.603695	0.525703
	Time	15.114124	9.938091	0.525590
	Drop Missing Value	15.373492	10.204003	0.513318
	Berkeley Earth	15.261181	10.118397	0.535506
FIRMS	Mean	10.580136	7.366220	0.746615
	Time	10.948846	7.606499	0.739265
	Drop Missing Value	10.920613	7.625059	0.764854
	Berkeley Earth	11.140604	7.743360	0.752430
PCD Wunderground Cycilcal	Mean	6.880583	4.505789	0.897282
	Time	6.989896	4.612923	0.898532
	Drop Missing Value	7.458061	4.824755	0.885461
	Berkeley Earth	7.442761	4.903492	0.889523

ตารางที่ 4.2 ผลลัพธ์การฝึกฝนแบบจำลองด้วย Extreme Gradient Boosting (ต่อ)

Dataset	Loss Function	Root Mean Square Error	Mean Absolute Error	R^2
PCD Wunderground Dummy	Mean	6.935117	4.460789	0.895647
	Time	7.039450	4.589469	0.897088
	Drop Missing Value	7.436442	4.789021	0.886124
	Berkeley Earth	7.360664	4.851035	0.891947
PCD FIRMS Cyclical	Mean	7.107424	4.562444	0.885654
	Time	6.759564	4.482543	0.900620
	Drop Missing Value	7.266628	4.759627	0.895886
	Berkeley Earth	7.509111	4.851822	0.887525
PCD FIRMS Dummy	Mean	7.160400	4.568619	0.883943
	Time	6.781838	4.454778	0.899964
	Drop Missing Value	7.285891	4.763431	0.895333
	Berkeley Earth	7.494745	4.829680	0.887955
PCD Wunderground FIRMS Cyclical	Mean	6.915115	4.469559	0.899008
	Time	7.320804	4.795675	0.894454
	Drop Missing Value	7.198124	4.677305	0.894659
	Berkeley Earth	7.226229	4.762614	0.900128
PCD Wunderground FIRMS Dummy	Mean	6.892797	4.458451	0.899659
	Time	7.323053	4.740145	0.894389
	Drop Missing Value	7.091423	4.596913	0.897759
	Berkeley Earth	7.227290	4.717898	0.900099
Wunderground FIRMS	Mean	6.909593	4.470162	0.899169
	Time	7.324293	4.801357	0.894353
	Drop Missing Value	7.217764	4.684145	0.894083
	Berkeley Earth	7.239435	4.772075	0.899763

จากตารางที่ 4.2 พบว่าจากการฝึกฝนแบบจำลองด้วยพารามิเตอร์พื้นฐานแล้วชุดข้อมูลที่ทำให้ Extreme Gradient Boosting มีประสิทธิภาพสูงสุดได้แก่ ชุดข้อมูลที่รวมกันระหว่าง PCD กับ FIRMS โดยมีวิธีการจัดการกับข้อมูลที่หายไปแบบ Time Function และแก้ข้อมูลที่มีลักษณะแบบวนซ้ำให้เป็นแบบวงกลม

4.3.3 Random Forest

ตารางที่ 4.3 ผลลัพธ์จากการฝึกฝนแบบจำลองแบบ Random Forest

Dataset \ Loss Function		Root Mean Square Error (μg/m^3)	Mean Absolute Error (μg/m^3)	R^2
PCD Cyclical	Mean	7.600287	4.610847	0.903881
	Time	7.385274	4.865031	0.914672
	Drop Missing Value	8.335725	5.211304	0.895152
	Berkeley Earth	8.135817	5.184106	0.893447
PCD Dummy	Mean	8.074259	4.901045	0.903881
	Time	7.759024	4.865031	0.914672
	Drop Missing Value	8.335725	5.211304	0.895152
	Berkeley Earth	8.135817	5.184106	0.893447
Wunderground	Mean	14.4400718	9.239773	0.547546
	Time	14.873997	9.6632279	0.540545
	Drop Missing Value	15.365492	10.029891	0.513824
	Berkeley Earth	14.900569	9.7595568	0.557198
FIRMS	Mean	8.568236	5.925119	0.833819
	Time	8.748686	5.980016	0.833525
	Drop Missing Value	8.888592	6.128673	0.844221
	Berkeley Earth	8.994247	6.177030	0.838635
PCD Wunderground Cycilcal	Mean	6.894927	4.443323	8.896853
	Time	7.045041	4.604460	0.896925
	Drop Missing Value	7.511814	4.847522	0.883804
	Berkeley Earth	7.331903	4.816212	0.892790
PCD Wunderground Dummy	Mean	7.186997	4.661737	0.887929
	Time	7.273006	4.774392	0.890146
	Drop Missing Value	7.809323	5.030185	0.874418

ตารางที่ 4.3 ผลลัพธ์จากการฝึกฝนแบบจำลองแบบ Random Forest (ต่อ)

Dataset	Loss Function	Root Mean Square Error	Mean Absolute Error	R^2
	Berkeley Earth	7.581204	4.978399	0.885375
PCD FIRMS Cyclical	Mean	7.181264	4.537944	0.8832665
	Time	6.742339	4.423149	0.901126
	Drop Missing Value	7.294779	4.759454	0.895078
	Berkeley Earth	7.533405	4.809215	0.886796
PCD FIRMS Dummy	Mean	7.249853	4.592831	0.881025
	Time	6.872445	4.515215	0.897273
	Drop Missing Value	7.554705	4.848565	0.887468
	Berkeley Earth	7.598193	4.865223	0.884840
PCD Wunderground FIRMS Cyclical	Mean	7.033335	4.476938	0.895525
	Time	7.544006	4.907551	0.887920
	Drop Missing Value	7.225820	4.688202	0.893847
	Berkeley Earth	7.406228	4.814343	0.895091
PCD Wunderground FIRMS Dummy	Mean	7.018597	4.459143	0.895962
	Time	7.502479	4.875438	0.889151
	Drop Missing Value	7.191068	4.631895	0.894865
	Berkeley Earth	7.355388	4.776186	0.896526
Wunderground FIRMS	Mean	6.9078823	4.3869.6	0.899221
	Time	7.418111	4.833030	0.891630
	Drop Missing Value	6.997479	4.531870	0.900450
	Berkeley Earth	7.248907	4.718860	0.899500

จากตารางที่ 4.3 พบว่าชุดข้อมูลที่ทำให้แบบจำลอง Random Forest มีประสิทธิภาพดีที่สุด ได้แก่ ชุดข้อมูลที่รวมกันระหว่าง PCD กับ FIRMS โดยผ่านการจัดการกับค่าที่หายไปโดยวิธี Time Function และแก้ไขข้อมูลแบบวนซ้ำเปลี่ยนให้เป็นวงกลม

4.3.4 MultiPerceptron Neural Network (Sklearn)

ตารางที่ 4.4 ผลลัพธ์จากการฝึกฝนแบบจำลองแบบ MultiPerceptron Neural Network จาก Sklearn

Dataset \ Performance Measure		Root Mean Square Error (μg/m^3)	Mean Absolute Error (μg/m^3)	R^2
PCD Cyclical	Mean	8.039761	5.119851	0.904700
	Time	7.920640	5.136005	0.911080
	Drop Missing Value	8.470378	5.452231	0.891737
	Berkeley Earth	8.127597	5.325174	0.893663
PCD Dummy	Mean	8.40019	5.264506	0.895968
	Time	8.060765	5.025405	0.907906
	Drop Missing Value	8.321325	5.191068	0.895514
	Berkeley Earth	8.045669	5.069853	0.895796
Wunderground	Mean	15.446549	10.216523	0.482322
	Time	15.978685	10.380595	0.469763
	Drop Missing Value	16.262794	10.745562	0.455384
	Berkeley Earth	15.991103	10.892256	0.490011
FIRMS	Mean	11.427502	8.000449	0.704403
	Time	11.769187	8.155911	0.698731
	Drop Missing Value	12.085765	8.336430	0.712001
	Berkeley Earth	12.073462	8.257918	0.709234
PCD Wunderground Cycilcal	Mean	7.258232	4.794725	0.885697
	Time	7.427174	4.903518	0.885440
	Drop Missing Value	7.700000	5.505735	0.877910
	Berkeley Earth	7.517289	5.007352	0.887299
PCD Wunderground Dummy	Mean	7.041269	4.607829	0.892428
	Time	7.156455	4.758451	0.893639
	Drop Missing Value	7.526980	4.833396	0.883335
	Berkeley Earth	7.471394	4.916410	0.888671

ตารางที่ 4.4 ผลลัพธ์จากการฝึกฝนแบบจำลองแบบ MultiPerceptron Neural Network จาก Sklearn (ต่อ)

Dataset \ Loss Function		Root Mean Square Error	Mean Absolute Error	R^2
PCD FIRMS Cyclical	Mean	7.251988	4.742556	0.880954
	Time	6.906299	4.620497	0.896258
	Drop Missing Value	7.375063	4.869594	0.892755
	Berkeley Earth	7.657896	4.962942	0.883023
PCD FIRMS Dummy	Mean	7.295615	4.773544	0.879518
	Time	6.954788	4.680077	0.894797
	Drop Missing Value	7.444890	4.952931	0.890715
	Berkeley Earth	7.554781	4.969468	0.886152
PCD Wunderground FIRMS Cyclical	Mean	7.057542	4.620764	0.894805
	Time	7.556203	5.031313	0.887557
	Drop Missing Value	7.201559	4.786168	0.894558
	Berkeley Earth	7.522267	5.003354	0.891777
PCD Wunderground FIRMS Dummy	Mean	6.944396	4.556342	0.898151
	Time	7.337347	4.942164	0.893976
	Drop Missing Value	7.218681	4.838955	0.894057
	Berkeley Earth	7.335112	4.935408	0.897096
Wunderground FIRMS	Mean	7.057542	4.620764	0.894805
	Time	7.556203	5.031313	0.887557
	Drop Missing Value	7.201559	4.786168	0.894558
	Berkeley Earth	7.522267	5.003354	0.891777

จากตารางที่ 4.4 พบว่าชุดข้อมูลที่ทำให้แบบจำลอง Neural Network ด้วยไลบรารี Sklearn นั้นมีประสิทธิภาพดีที่สุดได้แก่ ชุดข้อมูลที่เกิดจากการรวมกันระหว่าง PCD กับ FIRMS โดยมีวิธีการจัดการกับข้อมูลที่หายไปด้วยวิธี Time Function และแก้ไขข้อมูลที่มีการวนซ้ำให้เปลี่ยนเป็นวงกลมเช่นเดียวกับแบบจำลอง Random Forest, Extreme Gradient Boosting, Multiple Linear Regression

4.3.5 Artificial Neural Network (Keras)

ตารางที่ 4.5 ผลลัพธ์การฝึกฝนของแบบจำลอง Artificial Neural Network จาก Keras

Dataset \ Performance Measure		Root Mean Square Error (μg/m^3)	Mean Absolute Error (μg/m^3)	R^2
PCD Cyclical	Mean	8.426096	5.105252	0.889063
	Time	8.364292	5.242438	0.900840
	Drop Missing Value	8.575919	5.578307	0.889022
	Berkeley Earth	8.397325	5.341504	0.886487
PCD Dummy	Mean	8.674322	5.105252	0.889063
	Time	8.711368	5.142746	0.892440
	Drop Missing Value	8.506623	5.182050	0.890808
	Berkeley Earth	8.140460	5.111430	0.893326
Wunderground	Mean	15.760367	10.137364	0.461074
	Time	16.182086	10.319760	0.456178
	Drop Missing Value	16.856170	11.454796	0.414916
	Berkeley Earth	16.152495	10.764031	0.479665
FIRMS	Mean	11.673703	8.119578	0.691529
	Time	12.075888	8.406604	0.682824
	Drop Missing Value	12.477997	8.681726	0.693004
	Berkeley Earth	12.211041	8.301987	0.702569
PCD Wunderground Cycilcal	Mean	7.272897	4.821108	0.885234
	Time	7.407087	4.951480	0.886058
	Drop Missing Value	7.987108	5.215199	0.868635
	Berkeley Earth	7.844296	5.185939	0.877281
PCD Wunderground Dummy	Mean	7.152379	4.706396	0.889006
	Time	7.191315	4.757616	0.892600
	Drop Missing Value	7.652236	4.932495	0.879420
	Berkeley Earth	7.600461	4.992203	0.884792

ตารางที่ 4.5 ผลลัพธ์การฝึกฝนของแบบจำลอง Artificial Neural Network จาก Keras

Dataset	Loss Function	Root Mean Square Error	Mean Absolute Error	R^2
PCD FIRMS Cyclical	Mean	7.305870	4.770795	0.879179
	Time	7.171045	4.769513	0.888152
	Drop Missing Value	7.479314	5.020535	0.889702
	Berkeley Earth	7.772563	5.089478	0.879494
PCD FIRMS Dummy	Mean	7.259089	4.736523	0.880722
	Time	6.897149	4.600044	0.896533
	Drop Missing Value	7.494493	4.944111	0.889254
	Berkeley Earth	7.556056	4.892260	0.886114
PCD Wunderground FIRMS Cyclical	Mean	7.127861	4.680444	0.892698
	Time	7.509967	5.017021	0.888929
	Drop Missing Value	7.292077	4.833313	0.891891
	Berkeley Earth	7.403205	4.969062	0.895176
PCD Wunderground FIRMS Dummy	Mean	7.053301	4.622677	0.894931
	Time	7.318379	4.873645	0.894524
	Drop Missing Value	7.248974	4.803460	0.893165
	Berkeley Earth	7.307462	4.841185	0.897870
Wunderground FIRMS	Mean	11.772336	7.977550	0.707305
	Time	12.193217	8.447852	0.707207
	Drop Missing Value	11.918477	8.327889	0.711198
	Berkeley Earth	12.100100	8.292650	0.719974

จากตารางที่ 4.5 พบว่าชุดข้อมูลที่ทำให้แบบจำลอง Neural Network ด้วยไลบรารี Keras นั้นมีประสิทธิภาพดีที่สุดได้แก่ ชุดข้อมูลที่เกิดจากการรวมกันระหว่าง PCD กับ FIRMS โดยมีวิธีการจัดการกับข้อมูลที่หายไปด้วยวิธี Time Function และแก้ไขข้อมูลที่มีลักษณะเป็นแบบประเภท (Category) ด้วยวิธีการทำแบบดัมมี่ (Dummy Variable)

4.4 การหาค่าพารามิเตอร์ที่เหมาะสมของแต่ละแบบจำลอง

หลังจากผ่านขั้นตอนการเลือกชุดข้อมูลที่เหมาะสมที่สุดมาแล้วนั้นจะทำให้ได้ชุดข้อมูลที่ทำให้แบบจำลองมีประสิทธิภาพสูงสุดสำหรับพารามิเตอร์ตั้งต้นของแบบจำลอง ดังนั้นจึงทำการปรับแต่งพารามิเตอร์เพื่อทำให้ประสิทธิภาพของแบบจำลองสูงยิ่งขึ้นโดยจะอาศัยไลบรารี RandomSearchCV และ GridSearchCV ของ Sklearn ในการช่วยหาพารามิเตอร์ที่ทำให้แบบจำลองมีประสิทธิภาพสูงที่สุด โดย RandomSearchCV จะทำการสุ่มพารามิเตอร์ที่ผู้วิจัยกำหนดเข้าสู่แบบจำลอง วิธีนี้ให้ผลลัพธ์ที่ใกล้เคียงกับ GridSearchCV บางครั้งอาจดีกว่า โดยที่ใช้เวลาในการหาสั้นกว่าเนื่องจากเป็นการสุ่มไม่ได้หาครบทุกรูปแบบ ส่วน GridSearchCV จะทำการหาทุกรูปแบบของพารามิเตอร์ที่ผู้วิจัยตั้งไว้เนื่องจากต้องหาทุกรูปแบบทำให้สิ้นเปลืองทรัพยากรที่ใช้เป็นอย่างมากบวกกับใช้เวลาในการหานาน

ผลลัพธ์จากการปรับจูนพารามิเตอร์มีดังนี้

4.4.1 Random Forest

ตารางที่ 4.6 ผลลัพธ์การปรับแต่งพารามิเตอร์ของแบบจำลอง Random Forest

	n_estimators	min_samples_split	min_sample_leaf
Random Search	1400	5	2
Grid Search	600	2	1
Default	100	2	1

ตารางที่ 4.6 ผลลัพธ์การปรับแต่งพารามิเตอร์ของแบบจำลอง Random Forest (ต่อ)

	max_features	max_depth	bootstrap
Random Search	sqrt	None	False
Grid Search	sqrt	None	False
Default	auto	None	True

หลังจากการปรับแต่งพารามิเตอร์ประสิทธิภาพที่ได้ คือ

แบบจำลองพื้นฐาน

MAE = 4.4231 μg/m^3 , Accuracy = 80.37%, RMSE = 6.74233 μg/m^3, R^2 =0.9011

แบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์ด้วย RandomSearchCV

MAE = 4.3174 μg/m^3, Accuracy = 80.64%, RMSE = 6.49488 μg/m^3, R^2 = 0.9082

การใช้ RandomSearchCV ทำให้ความแม่นยำ (Accuracy) ดีขึ้นเมื่อเทียบกับแบบจำลองพื้นฐาน โดยคำนวณจาก $\left(\frac{80.64 - 80.37}{80.37}\right)\times 100 = 0.34\ \%$

แบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์ด้วย GridSearchCV

MAE = 4.3146 μg/m^3, Accuracy = 80.56, RMSE = 6.47417, R^2 = 0.9088

การใช้ GridSearchCV ทำให้ความแม่นยำ (Accuracy) ดีขึ้นเมื่อเทียบกับแบบจำลองพื้นฐานโดยคำนวณจาก $\left(\frac{80.56 - 80.37}{80.37}\right)\times 100 = 0.23\ \%$

4.4.2 Extreme Gradient Boosting

ตารางที่ 4.7 ผลการปรับแต่งพารามิเตอร์ของแบบจำลอง Extreme Gradient Boosting

	n_estimators	min_samples_split	min_sample_leaf
Random Search	1200	10	2
Grid Search	1400	10	2
Default	100	2	1

ตารางที่ 4.7 ผลการปรับแต่งพารามิเตอร์ของแบบจำลอง Extreme Gradient Boosting (ต่อ)

	max_features	max_depth	Learning rate
Random Search	sqrt	10	0.01
Grid Search	sqrt	10	0.01
Default	None	3	0.1

หลังจากการปรับแต่งพารามิเตอร์ประสิทธิภาพที่ได้ คือ

แบบจำลองพื้นฐาน

MAE = 4.4825 μg/m^3 , Accuracy = 80.21%, RMSE = 6.7564 μg/m^3, R^2 = 0.90007

แบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์ด้วย RandomSearchCV

MAE = 4.1810 μg/m^3, Accuracy = 81.36%, RMSE = 6.32127 μg/m^3, R^2 = 0.913089

การใช้ RandomSearchCV ทำให้ความแม่นยำ (Accuracy) ดีขึ้นเมื่อเทียบกับแบบจำลองพื้นฐาน โดยคำนวณจาก $\left(\frac{81.36 - 80.21}{80.21}\right)\times 100 = 1.43\ \%$

แบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์ด้วย GridSearchCV

MAE = 4.1775 μg/m^3, Accuracy = 81.37%, RMSE = 6.31769, R^2 = 0.91318

การใช้ GridSearch ทำให้ความแม่นยำ (Accuracy) ดีขึ้นเมื่อเทียบกับแบบจำลองพื้นฐานโดยคำนวณจาก $\left(\frac{81.37 - 80.21}{80.21}\right)\times 100 = 1.45\ \%$

4.4.3 Multi-Perceptron Neural Network (Sklearn)

ตารางที่ 4.8 ผลการปรับแต่งพารามิเตอร์ของ Multi-Perceptron Neural Network จาก Sklearn

	activation	alpha	batch_size
Random Search	relu	0.0003	32
Default	relu	0.0001	auto

ตารางที่ 4.8 ผลการปรับแต่งพารามิเตอร์ของ Multi-Perceptron Neural Network จาก Sklearn (ต่อ)

	hidden_layers	Optimize	Learning rate
Random Search	(32,16)	sgd	0.001
Default	(100,)	adam	0.001

เนื่องจากวิธีนี้การใช้ GridSearchCV นั้นเปลืองทรัพยากรในการทำงานรวมกับการใช้เวลานานกว่าจะได้ผลลัพธ์และจากการดูประสิทธิภาพจาก Extreme Gradient Boosting และ Random Forest แล้วนั้นการใช้ RandomSearchCV นั้นให้ประสิทธิภาพที่ใกล้เคียงกับ GridSearchCV แบบจำลอง Multi-Perceptron จึงใช้ RandomSearchCV ในการปรับพารามิเตอร์เพียงอย่างเดียวหลังจากการปรับแต่งพารามิเตอร์ประสิทธิภาพที่ได้ คือ

แบบจำลองพื้นฐาน

MAE = 4.6025 μg/m^3 , Accuracy = 79.25%, RMSE = 6.9063 μg/m^3, R^2 = 0.8962

แบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์ด้วย RandomSearchCV

MAE = 4.6970 μg/m^3, Accuracy = 79.34%, RMSE = 6.9985 μg/m^3, R^2 = 0.8934

การใช้ RandomSearchCV ทำให้ความแม่นยำ (Accuracy) ดีขึ้นเมื่อเทียบกับแบบจำลองพื้นฐาน โดยคำนวณจาก $\left(\frac{79.34 - 79.25}{79.25}\right)\times 100 = 0.11\ \%$

4.4.4 Artificial Neural Network (Keras)

สำหรับ Neural Network โดยใช้ไลบรารีผู้วิจัยไม่สามารถใช้ RandomSearchCV และ GridSearchCV ดังนั้นจึงต้องทำการปรับพารามิเตอร์เองโดนการปรับพารามิเตอร์นั้นผู้วิจัยได้ทำการทดลองเพิ่มชั้นของ Hidden Layers เข้าไปกลับพบว่าผลลัพธ์ที่ได้ไม่ได้ทำให้ประสิทธิภาพดีขึ้นดังนั้น จึงไม่ทำการเพิ่ม Hidden Layers แต่จะทำการปรับจำนวนโหนดใน Neural Network แทน สำหรับการปรับแต่งพารามิเตอร์ของ Artificial Neural Network นั้นมีหัวข้อย่อยดังนี้

- จำนวนโหนด (Number of Nodes)

ตารางที่ 4.9 ผลลัพธ์ของการปรับแต่งจำนวนโหนดใน Artificial Neural Network จาก Keras

Performance Measure / Number of Node	Mean Absolute Error (μg/m^3)	Root Mean Square Error (μg/m^3)	Accuracy (%) = 100 – (MAPE×100)
20	4.5783	6.8972	79.67
25	4.7369	7.0543	77.82
30	4.6649	6.8510	78.13
35	4.4914	6.7709	80.17
40	4.5811	6.9760	79.53
50	4.5553	6.8337	79.51
60	4.5164	6.7501	79.51
70	4.6164	6.9719	79.53

- การปรับ Dropout Rate

ตารางที่ 4.10 ผลลัพธ์การปรับแต่ง Dropout Rate ของ Artificial Neural Network จาก Keras

Performance Measure / Dropout Rate	Mean Absolute Error ($\mu g/m^3$)	Root Mean Square Error ($\mu g/m^3$)	Accuracy (%) = 100 – (MAPE×100)
0.0	4.6515	6.9700	79.75
0.1	4.5105	6.8584	80.16
0.2	4.5496	6.8584	79.96
0.3	4.6205	6.9166	79.00
0.4	4.7134	7.0633	78.66
0.5	4.9152	7.3107	77.53
0.6	5.0346	7.5155	77.06
0.7	5.2905	7.7039	74.58
0.8	5.8845	8.5210	71.82
0.9	7.5615	10.8691	63.57

- การปรับแต่ง Regularization Rate

ตารางที่ 4.11 ผลการปรับแต่ง Regularization Rate ของ Artificial Neural Network จาก Keras

Performance Measure / Regularization Rate	Mean Absolute Error ($\mu g/m^3$)	Root Mean Square Error ($\mu g/m^3$)	Accuracy (%) = 100 – (MAPE×100)
0.0	4.6315	6.9300	78.86
0.0001	4.5361	6.8716	80.27
0.0002	4.5500	6.8151	79.25

ตารางที่ 4.11 ผลการปรับแต่ง Regularization Rate ของ Artificial Neural Network จาก Keras (ต่อ)

Performance Measure / Regularization Rate	Mean Absolute Error ($\mu g/m^3$)	Root Mean Square Error ($\mu g/m^3$)	Accuracy (%) = 100 – (MAPE×100)
0.0003	4.4588	6.8066	80.46
0.0004	4.5053	6.7604	79.67
0.0005	4.5076	6.8056	80.27
0.0006	4.6123	6.8764	79.06
0.0007	4.5382	6.9601	80.84
0.0008	4.6231	6.8268	78.38
0.0009	4.5529	6.9020	79.79

- การปรับแต่ง Batch Size

ตารางที่ 4.12 ผลการปรับแต่ง Batch Size ของ Artificial Neural Network จาก Keras

Performance Measure / Batch Size	Mean Absolute Error ($\mu g/m^3$)	Root Mean Square Error ($\mu g/m^3$)	Accuracy (%) = 100 – (MAPE×100)
20	4.8207	6.9741	76.36
32	4.5419	6.7828	79.16
64	4.6177	6.9883	79.57

- การปรับแต่ง Activation Function

ตารางที่ 4.13 ผลการปรับแต่ง Activation Function ของ Artificial Neural Network จาก Keras

Performance Measure / Activation Function	Mean Absolute Error (μg/m^3)	Root Mean Square Error (μg/m^3)	Accuracy (%) = 100 – (MAPE×100)
Tanh	4.6356	7.0300	79.92
Relu	4.5825	6.8604	79.28

- การปรับแต่ง Optimizer

ตารางที่ 4.14 ผลการปรับแต่ง Optimizer ของ Artificial Neural Network จาก Keras

Performance Measure / Optimizer	Mean Absolute Error (μg/m^3)	Root Mean Square Error (μg/m^3)	Accuracy (%) = 100 – (MAPE×100)
SGD	4.7383	7.2875	79.24
RMSprop	4.6119	6.9819	79.11
Adagrad	5.8174	8.6341	73.89
Adadelta	9.5843	13.6026	56.32
Adam	4.6161	7.0093	79.92
Adamax	4.6049	7.0435	79.85
Nadam	4.5593	6.9537	80.08

- การปรับแต่ง Learning Rate

ตารางที่ 4.15 ผลการปรับแต่ง Learning Rate ของ Artificial Neural Network จาก Keras

Performance Measure / Learning Rate	Mean Absolute Error (μg/m^3)	Root Mean Square Error (μg/m^3)	Accuracy (%) = 100 – (MAPE×100)
0.1	6.4761	9.2539	67.95
0.01	4.8314	7.1290	77.76
0.001	4.6721	7.0623	79.45
0.0001	4.7171	7.2860	79.46

เมื่อผ่านการทดลองปรับแต่งพารามิเตอร์ตามที่ผู้วิจัยกำหนดผู้วิจัยจะทำการเลือกพารามิเตอร์ที่ทำให้ผลลัพธ์ของแบบจำลองนั้นมีความคลาดเคลื่อนที่น้อยที่สุด โดยจะพิจารณาจากค่า MAE, RMSE, และ Accuracy โดยที่หาก Accuracy มีมากกว่าแต่ MAE, RMSE ของพารามิเตอร์อีกค่านั้นน้อยจะเลือกพารามิเตอร์ที่มีค่า MAE, RMSE น้อยกว่านั้น จากการทดลองทุกพารามิเตอร์ที่ทำให้แบบจำลอง Artificial Neural Network มีประสิทธิภาพสูงสุดมีดังตารางที่ 4.16 จากนั้นจะทำการฝึกฝนแบบจำลองด้วยพารามิเตอร์เหล่านี้เพื่อหาแบบจำลองที่มีประสิทธิภาพมากที่สุด

ตารางที่ 4.16 พารามิเตอร์แต่ละตัวที่ทำให้แบบจำลอง Artificial Neural Network (Keras) มีประสิทธิภาพสูงสุด

	activation	L2 Regularization	batch_size	hidden_layers
Default	relu	0.0001	32	(20,)
Tuned	tanh	0.0001	64	(35,)

ตารางที่ 4.16 พารามิเตอร์แต่ละตัวที่ทำให้แบบจำลอง Artificial Neural Network (Keras) มีประสิทธิภาพสูงสุด (ต่อ)

	Optimize	Learning rate	dropout	epoch
Default	Adam	0.01	0.2	100
Tuned	Nadam	0.001	0.2	100

หลังจากการปรับแต่งพารามิเตอร์ประสิทธิภาพที่ได้ คือ

แบบจำลองพื้นฐาน

MAE = 4.6622 μg/m^3, Accuracy = 79.43%, RMSE = 7.0879 μg/m^3

แบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์

MAE = 4.6263 μg/m^3, Accuracy = 79.58%, RMSE = 7.0587, R^2 = 0.8681

หลังจากการปรับแต่งพารามิเตอร์ความแม่นยำ (Accuracy) สูงขึ้นเมื่อเทียบกับแบบจำลองพื้นฐานโดยคำนวณจาก $\left(\frac{79.58 - 79.43}{79.43}\right) \times 100 = 0.19\ \%$

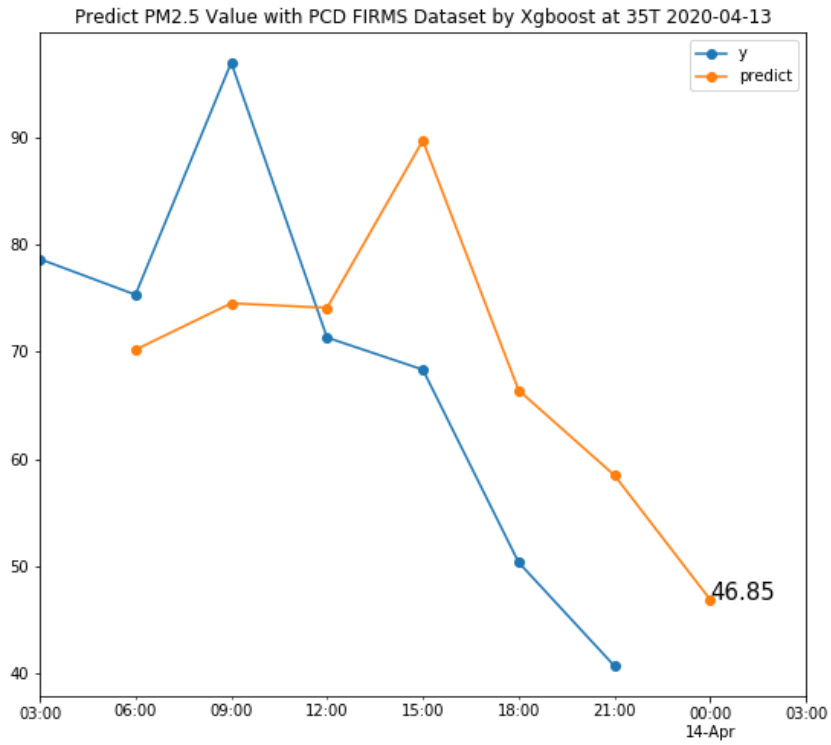
หลังจากขั้นตอนปรับแต่งพารามิเตอร์ของแบบจำลองแต่ละชนิดแล้วเมื่อนำผลลัพธ์ของแบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์มาเปรียบเทียบกันจะเห็นว่าแบบจำลอง Extreme Gradient Boosting มีประสิทธิภาพสูงสุดดังตารางที่ 4.17 ดังนั้นจึงจะใช้แบบจำลอง Extreme Gradient Boosting เป็นแบบจำลองในการทำนาย $PM_{2.5}$ ที่จะเกิดขึ้นในอีกสามชั่วโมงข้างหน้าบวกกับฝึกฝนด้วยชุดข้อมูลที่ทำให้แบบจำลองมีประสิทธิภาพสูงสุดได้แก่ ชุดข้อมูลที่รวมข้อมูลระหว่าง PCD และFIRMS ซึ่งจัดการข้อมูลที่หายไปด้วยวิธี Time Function และจัดการข้อมูลประเภทวนซ้ำ (Cyclical Variable) ให้แปลงให้เป็นวงกลมด้วยฟังก์ชันโคไซน์และไซน์

ตารางที่ 4.17 การเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองหลังจากผ่านการปรับแต่งพารามิเตอร์

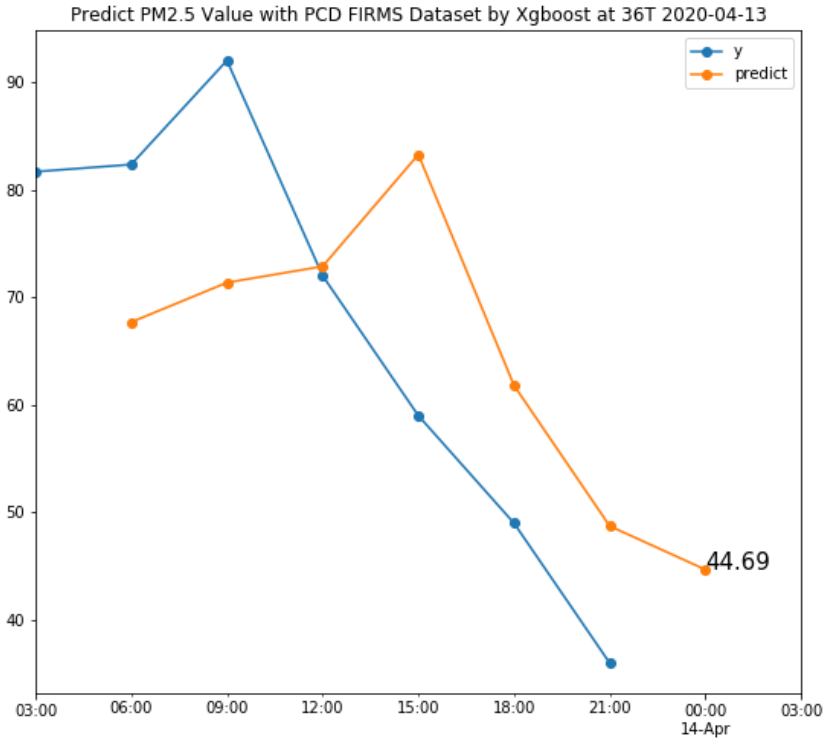
Model	Root Mean Square Error ($\mu g/m^3$)	Mean Absolute Error ($\mu g/m^3$)	Accuracy (%)	R^2
Random Forest	6.49488	4.3174	80.64	0.9082
Extreme Gradient Boosting	6.31769	4.1775	81.37	0.91318
Multiperceptron (Sklearn)	6.9985	4.6970	79.34	0.8934
Artificial Neural Network (Keras)	7.0587	4.6263	79.58	0.8681
Multiple Linear Regression	7.4631	4.911497	78.60	0.8788

4.5 การทำนายผลค่าปริมาณฝุ่นละอองที่น้อยกว่า 2.5 ไมครอน หรือ $PM_{2.5}$ ในสามชั่วโมงข้างหน้า

สำหรับขั้นตอนนี้เป็นขั้นตอนสำหรับการนำแบบจำลองที่ผ่านการเรียนรู้ในขั้นตอนที่ผ่านมานั้นไปใช้งานจริงกับข้อมูลปัจจุบัน โดยผู้วิจัยได้การดึงข้อมูลจากกรมควบคุมมลพิษและข้อมูลจุดความร้อนจาก FIRMS ณ เวลาปัจจุบัน เมื่อดาวน์โหลดข้อมูลเสร็จจะต้องทำการแปลงข้อมูลให้เหมือนกับที่ข้อมูลฝึกฝนถูกแปลงไป โดยต้องทำการแปลงข้อมูลที่มีการวนซ้ำให้เป็นวงกลมของโคไซน์และไซน์ เติมเวลาต่างๆ ในบทที่ 3 เมื่อทำการเติมทุกอย่างแล้วก็นำข้อมูลเหล่านั้นเข้าสู่แบบจำลอง คือ Extreme Gradient Boosting ที่ใช้พารามิเตอร์เดียวกับที่ทำการปรับจูนด้วย RandomSearchCV โดยผลลัพธ์ที่ได้เป็นดังภาพที่ 4.29 และภาพที่ 4.30



ภาพที่ 4.29 การทำนายค่า PM2.5 ณ วันที่ 2020-04-13 ในอีกสามชั่วโมงข้างหน้า ณ จุดศูนย์ราชการ จังหวัดเชียงใหม่



ภาพที่ 4.30 การทำนายค่า $PM_{2.5}$ ณ วันที่ 2020-04-13 ในอีกสามชั่วโมงข้างหน้า ณ จุดโรงเรียนยุพราช จังหวัดเชียงใหม่

บทที่ 5
สรุปผลการวิจัย

ในบทนี้จะกล่าวถึงบทสรุปของปัญหา สิ่งที่สามารถนำไปต่อยอดในงานวิจัยถัดไป ปัญหาและอุปสรรคที่เจอในการทำงานวิจัยรวมทั้งข้อเสนอแนะเพิ่มเติมในการทำงานวิจัย งานจากการทำงานวิจัยฉบับนี้สามารถสรุปผลและประเมินผลได้ดังต่อไปนี้

5.1 สรุปผลการทดลอง

สำหรับงานวิจัยฉบับนี้นั้นมีจุดประสงค์สองข้อคือ 1. เพื่อหาปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลงค่าฝุ่นละอองขนาดเล็กว่า 2.5 ไมครอนและ 2. การทำนายฝุ่นละอองขนาดเล็กกว่า 2.5 ไมครอนในอีกสามชั่วโมงข้างหน้า จากการทดลองพบว่าปัจจัยที่ส่งผลกระทบต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ ที่ชัดเจนที่สุดคือ ปัจจัยไฟหรือจุดความร้อนที่ดาวเทียมนั้นตรวจพบ โดยปัจจัยที่กล่าวถึงนี้เป็นปัญหาหลักที่ส่งผลให้ค่า $PM_{2.5}$ นั้นเพิ่มสูงขึ้นบวกกับสภาพทางภูมิศาสตร์ที่เป็นแอ่งกระทะทำให้ $PM_{2.5}$ นั้นล่องลอยอยู่ในจังหวัดเชียงใหม่เป็นเวลานาน ปัจจัยดังกล่าวนั้นรวมถึงไฟที่เกิดจากกิจกรรมของมนุษย์ เช่น การเผาป่า การเผาทางการเกษตรของเกษตกรในฤดูกาลเผา เป็นต้น ส่วนไฟที่เกิดเองโดยธรรมชาติ ก็เกิดจากไฟป่าสาเหตุอาจมาจากความแห้งที่ขาดฝนมานานเนื่องจากปัญหา $PM_{2.5}$ ส่วนใหญ่นั้นจะเริ่มเกิดตั้งแต่ช่วงคาบเกี่ยวของฤดูหนาวและฤดูร้อนต่อเนื่องไปจนถึงฤดูร้อนเสร็จสิ้นตามที่ได้อธิบายไว้ในบทที่ 4 จะเห็นรูปแบบการเกิด $PM_{2.5}$ ที่เกิดซ้ำทุกปีโดยปัญหา $PM_{2.5}$ นั้นส่งผลเสียทั้งทางด้านสุขภาพของประชาชนที่อาศัยในจังหวัดเชียงใหม่และทั้งทางด้านเศรษฐกิจที่ส่งผลเสียให้การท่องเที่ยวลดลง วิวทิวทัศน์ที่เคยสวยงามก็จางหาย ต่อไปเป็นปัจจัยที่เมื่อ $PM_{2.5}$ สูงขึ้นก็จะพบว่า PM_{10} นั้นก็มีค่าสูงเช่นเดียวกัน แต่ในที่นี้ก็ไม่สามารถบอกได้ว่าปัจจัยใดเป็นเหตุและเป็นผลแต่ทั้งสองปัจจัยจะสูงตามกันและกันและจะลดต่ำลงตามกันและกันเช่นเดียวกัน สำหรับปัจจัยอื่นๆ ที่ส่งผลกระทบต่อการเปลี่ยนแปลงของค่า $PM_{2.5}$ ได้แก่ ปัจจัยทางสภาพอากาศ คือ ฝน ความไวลม ความชื้น จุดกลั่นไอน้ำ โดยปัจจัยทั้ง 4 นี้เป็นปัจจัยที่มีความสัมพันธ์ในทิศทางตรงกันข้ามกับ $PM_{2.5}$ เช่น ระดับของ $PM_{2.5}$ นั้นจะต่ำลงในวันที่ฝนตกหรือต่ำลงในช่วงที่มีความไวลมสูง ความชื้นและจุดกลั่นไอน้ำก็เช่นเดียวกันแต่ความไวลมและความชื้นนั้นต้องขึ้นอยู่กับสถานที่ตั้งของตัวรับข้อมูลด้วยหากจุดรับข้อมูลนั้นไม่ได้มาตรฐานก็อาจทำให้ลักษณะหรือความสัมพันธ์ที่ผู้วิจัยได้วิเคราะห์ออกมาอาจมีความคลาดเคลื่อนและผิดเพี้ยนไปได้ สำหรับปัจจัยที่ส่งผลกระทบต่อการทำนายของ $PM_{2.5}$ นั้นนอกจากปัจจัยที่ได้กล่าวไปข้างต้นแล้ว สำหรับการทำนายปัจจัยที่มีความสำคัญที่สุดได้แก่ ค่า $PM_{2.5}$ ในอดีต เพราะค่าที่เกิดขึ้นในอดีตจะส่งผลอย่างมากต่อการทำนายเนื่องจากค่าถัดไปที่จะเกิดขึ้นนั้นมีความใกล้เคียงกันหากมีการทำนายในอีกหนึ่งชั่วโมงข้างหน้าข้อมูล $PM_{2.5}$ ณ ปัจจุบันกับชั่วโมงข้างหน้านั้นแทบจะไม่แตกต่างกันมากนักแต่เหตุผลที่ทำการทำนายในรายสามชั่วโมงเนื่องจากการทำนายในรายชั่วโมงนั้นไม่สามารถเตรียมการป้องกันได้ทันเวลาเนื่องจากระยะเวลาที่

สั้นเกินไป จึงทำการเพิ่มระยะเวลาเป็นการทำนายในรายสามชั่วโมงดังนั้นความสัมพันธ์ของ $PM_{2.5}$ ในอดีตกับ $PM_{2.5}$ ในปัจจุบันจะมีค่าที่แตกต่างกันมากขึ้นเนื่องจากขอบเขตของเวลาที่แตกต่างกันมากขึ้นแต่แม้จะแตกต่างกันมากขึ้น ก็ยังถือว่ามีความสำคัญมากอยู่ดี ปัจจัยอื่นๆ ที่ส่งผลได้แก่ เวลา เนื่องจากแบบจำลองจะสังเกตเห็นถึงช่วงเวลาที่ค่า $PM_{2.5}$ นั้นเริ่มสูงขึ้นได้แก่ เดือนมกราคม - เดือนพฤษภาคม เป็นต้น หรือ ฤดูกาลโดยฤดูกาลที่มีระดับที่อันตรายได้แก่ ฤดูร้อนและช่วงคาบเกี่ยวระหว่างฤดูหนาวและฤดูร้อน ปัจจัยนอกเหนือจากนี้เป็นปัจจัยที่มีความกำกวมที่ผู้วิจัยไม่สามารถหาความสัมพันธ์กับปัญหา $PM_{2.5}$ ได้

ตารางที่ 5.1 การเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองหลังจากผ่านการปรับแต่งพารามิเตอร์

Model	RMSE (μg/m^3)	MAE (μg/m^3)	Accuracy (%)	R^2
Random Forest	6.49488	4.3174	80.64	0.9082
Extreme Gradient Boosting	6.31769	4.1775	81.37	0.91318
Multiperceptron (Sklearn)	6.9985	4.6970	79.34	0.8934
Artificial Neural Network (Keras)	7.0587	4.6263	79.58	0.8681
Multiple Linear Regression	7.4631	4.911497	78.60	0.8788

สำหรับการทำนายค่า $PM_{2.5}$ ในอนาคตได้ทำการทดลองฝึกฝนแบบจำลองด้วยชุดข้อมูลหลายชุดข้อมูลเพื่อจะหาชุดข้อมูลที่ทำให้แต่ละแบบจำลองมีประสิทธิภาพสูงสุด โดยชุดข้อมูลที่ทำการฝึกฝนมีทั้งหมดด้วยกัน 44 ชุดข้อมูล โดยแต่ละชุดข้อมูลจะเป็นข้อมูลที่ถูกปรับแต่งและจัดการกับข้อมูลที่ขาดหายไปต่างกันโดยวิธีในการแทนข้อมูลที่ขาดหายไปที่ดีที่สุดสำหรับงานวิจัยนี้ได้แก่ การแทนข้อมูลที่ขาดหายไปด้วย Time Function เนื่องจากในแต่ละแบบจำลองชุดข้อมูลนี้มีค่าคลาดเคลื่อนที่น้อยที่สุด ยกตัวอย่างเช่น แบบจำลอง Random Forest การแทนด้วยวิธี Time Function มีค่า RMSE = 6.742 μg/m^3 เทียบกับการแทนด้วยค่าเฉลี่ย มีค่า RMSE = 7.181 μg/m^3 ซึ่งจะเห็นว่า Time Function ทำให้ได้ประสิทธิภาพที่ดีกว่าเหตุผลเป็นเพราะการแทนด้วยวิธีนี้นั้นจะคิดวันเวลาของค่าที่เกิดขึ้นเทียบกับข้อมูลทั้งหมดซึ่งเหมาะกับข้อมูลที่เป็นลักษณะของ Time Series จึงสมเหตุสมผลที่วิธีนี้จะสามารถทำให้แบบจำลองแต่ละแบบนั้นมีประสิทธิภาพสูงที่สุด ต่อไปเป็นการจัดการกับข้อมูลประเภทวนซ้ำและข้อมูลที่มีลักษณะของประเภท โดยทั้ง

สองส่วนนี้จะมีตัวแปรอิสระที่อยู่ในหมวดทั้งสองนี้อยู่ เช่น ทิศทางลมที่มีค่า 0 – 360 แล้วจะเกิดการวนซ้ำ และสามารถแปลงเป็นทิศทางและจัดเป็นข้อมูลที่สามารถแบ่งเป็นหมวดหมู่ได้ (Category) ซึ่งปกติแล้วข้อมูลประเภทนี้ไม่สามารถนำเข้าแบบจำลองได้เนื่องจากลักษณะของข้อมูลไม่เป็นตัวเลขดังนั้นจึงต้องทำการแปลงให้เป็นตัวเลข โดยสามารถแปลงได้สองวิธีได้แก่ การแปลงให้อยู่ในรูปของวงกลมโดยใช้ฟังก์ชันโคไซน์และไซน์ หรือแปลงให้เป็นดัมมี่ จากผลการทดลองส่วนใหญ่แล้วการแปลงข้อมูลให้เป็นแบบวงกลมจะให้ผลลัพธ์ที่ดีกว่าแต่ทั้งนี้ก็ขึ้นอยู่กับข้อมูลด้วยเช่นกัน การแปลงในรูปของดัมมี่นั้นมีข้อเสียคือเป็นการเพิ่มมิติของข้อมูลซึ่งเมื่อมิติของข้อมูลเพิ่มขึ้นเวลาในการฝึกฝนแบบจำลองก็เพิ่มขึ้นตาม การหารูปแบบของข้อมูลในแบบจำลองจะมีความซับซ้อนมากยิ่งขึ้นทำให้ยากต่อการตีความหมาย ต้องใช้แบบจำลองที่มีความซับซ้อนในการแก้ปัญหา

หลังจากฝึกฝนแบบจำลองด้วยชุดข้อมูลที่ทำให้แบบจำลองพื้นฐานมีประสิทธิภาพสูงสุดและผ่านการปรับแต่งพารามิเตอร์ของแต่ละแบบจำลอง จากการทดลองผลลัพธ์ดังตารางที่ 5.1 พบว่าแบบจำลอง Extreme Gradient Boosting นั้นมีประสิทธิภาพสูงสุด โดยมีค่า RMSE = 6.31769 $\mu g/m^3$, MAE = 4.1775 $\mu g/m^3$, มีความแม่นยำ = 81.37%, และมี r^2 = 0.91318

5.2 ปัญหาและอุปสรรค

จากแหล่งข้อมูลที่มีผู้วิจัยต้องตรวจเช็คข้อมูลอย่างละเอียดถี่ถ้วนเนื่องจากข้อมูลเหล่านี้อาจจะมีความคลาดเคลื่อนจากความเป็นจริงแล้วส่งผลให้เกิดความผิดพลาดในการทำนาย นอกจากนี้ชุดข้อมูลที่ผู้วิจัยสามารถเข้าถึงได้นั้นมีช่วงข้อมูลบางช่วงที่ข้อมูลขาดหายไปทำให้ผู้วิจัยต้องหาแนวทางในการแก้ไขปัญหาให้ดีที่สุด เช่น เติมข้อมูลที่ขาดหายไปด้วยค่าเฉลี่ยของแต่ละตัวแปรอิสระในแต่ละเดือน หรือ ลบข้อมูลในช่วงเวลานั้นออกจากชุดข้อมูล เป็นต้น ซึ่งขั้นตอนนี้ถือเป็นขั้นตอนที่สำคัญเนื่องจากหากจัดการได้ไม่มีประสิทธิภาพจะส่งผลต่อการทำนายได้ หากทำการเติมข้อมูลด้วยค่าบางอย่างที่ไม่สอดคล้องกับความเป็นจริงอาจเป็นการเติมข้อมูลรบกวนเข้าไปในชุดข้อมูลได้

ในช่วงแรกของการเริ่มงานวิจัยผู้วิจัยตั้งเป้าจะใช้แบบจำลองทางสถิติที่ใช้ในการแก้ปัญหาที่มีช่วงเวลาเข้ามาเกี่ยวข้องในการทำนายค่า $PM_{2.5}$ หลังจากได้ทำการทดลองกลับพบว่าคอมพิวเตอร์ที่ใช้ในการทำงานนั้นไม่มีประสิทธิภาพมากพอในการใช้แบบจำลองนั้น เนื่องจากมีการคำนวณขนาดใหญ่ทำให้ทรัพยากรที่มีในเครื่องคอมพิวเตอร์นั้นไม่เพียงพอ ส่งผลให้ไม่สามารถใช้แบบจำลองดังกล่าวในการทำการทดลองได้ ปัญหาอีกส่วนหนึ่งได้แก่ การขาดข้อมูล ณ ปัจจุบัน เนื่องจากผู้วิจัยไม่มีข้อมูลในปี 2019 ทำให้การทำนายนั้นมีประสิทธิภาพที่ลดน้อยลงเนื่องจากขาดข้อมูลปีล่าสุดไป

การนำแหล่งข้อมูลหลายแหล่งข้อมูลมารวมกัน เนื่องจากข้อมูลหลายแหล่งข้อมูลนั้นไม่ได้ถูกเก็บข้อมูลมาจากอุปกรณ์ตัวเดียวกันรวมทั้งอุปกรณ์แต่ละตัวไม่สามารถที่จะวัดค่าที่อ่านได้อย่างถูกต้อง 100% เมื่อนำข้อมูลเหล่านั้นมารวมกันส่งผลให้ข้อมูลมีความขัดแย้งกันอาจส่งผลให้ประสิทธิภาพในการทำนายลดลง

ปัญหา $PM_{2.5}$ นั้นเป็นปัญหาที่ละเอียดอ่อนสามารถเกิดได้จากหลายปัจจัยทำให้ในการทำนายค่า $PM_{2.5}$ นั้นมีหลายตัวแปรที่เกี่ยวข้องผู้วิจัยจะต้องทำการเลือกตัวแปรอิสระเหล่านั้นและทำการวิเคราะห์อย่างถี่ถ้วนเพื่อที่จะหาความสัมพันธ์ แต่ในความเป็นจริงตัวแปรที่ถูกเลือกมาบางตัวไม่สามารถอธิบายหรือไม่มีความสัมพันธ์กับค่า $PM_{2.5}$ อย่างชัดเจนทำให้บางครั้งตัวแปรอิสระที่ถูกเลือกมานั้นเป็นสาเหตุที่ทำให้ประสิทธิภาพในการทำนายนั้นลดลง นอกจากนี้หากต้องการลดขนาดข้อมูลโดยการแปลงข้อมูลเป็นรายวันหรือรายเดือนจะยิ่งทำให้ข้อมูลที่มีน้อยลงดังนั้นจึงต้องทำการเลือกช่วงเวลาของข้อมูลให้ดีเพื่อให้มีจำนวนชุดข้อมูลที่สามารถฝึกฝนแบบจำลองได้อย่างเหมาะสมและมีมากพอให้ทดสอบประสิทธิภาพด้วยเช่นกัน

5.3 ข้อเสนอแนะและแนวทางในการพัฒนาในอนาคต

การฝึกฝนแบบจำลองด้วยชุดข้อมูลที่มีมิติมากนั้นจะทำให้เวลาในการฝึกฝนเพิ่มสูงขึ้นมาก ดังนั้นการใช้เทคนิคการลดมิติข้อมูลจะช่วยในเรื่องของการประหยัดเวลาในการฝึกฝนแบบจำลองลงได้ พร้อมทั้งการนำข้อมูลจากจังหวัดใกล้เคียงมาช่วยวิเคราะห์ถึงแนวโน้มการเกิดปัญหา $PM_{2.5}$ และช่วยในการเติมข้อมูลที่ขาดหายไปของ 2 สถานีที่ตั้งอยู่ในจังหวัดเชียงใหม่จะสามารถทำให้มีข้อมูลที่ช่วยในการฝึกฝนแบบจำลองเพิ่มมากขึ้นโดยต้องดูความสัมพันธ์ของค่าข้อมูลที่เกิดขึ้นด้วย การเพิ่มตัวเลือกของการปรับแต่งพารามิเตอร์ก็ถือเป็นอีกส่วนประกอบหนึ่งที่จะทำให้การพัฒนาแบบจำลองมีประสิทธิภาพมากขึ้น สำหรับแนวทางในการพัฒนาในอนาคตผู้วิจัยมีความขาดหวังที่จะใช้ตัวแปรอิสระที่ถูกตรวจจับได้จากดาวเทียมที่มีชื่อว่า ค่าความลึกเชิงแสงของฝุ่นละออง (Aerosol Optical Depth) เนื่องจากมีงานวิจัยหลายงานวิจัยที่ได้ทำการศึกษาว่าตัวแปรอิสระดังกล่าวนั้นมีความสัมพันธ์กับค่า $PM_{2.5}$ ที่เกิดขึ้นนอกจากนี้ยังเป็นข้อมูลจากดาวเทียมหากสามารถใช้ตัวแปรอิสระนี้ในการทำนายค่า $PM_{2.5}$ ได้อย่างมีประสิทธิภาพจะทำให้สามารถวัดค่า $PM_{2.5}$ ที่จะเกิดขึ้นได้โดยไม่จำเป็นต้องใช้ข้อมูลจากภาคพื้นดิน ทำให้สามารถเข้าถึงพื้นที่ที่อุปกรณ์รับข้อมูลไม่สามารถเข้าถึงได้

เอกสารอ้างอิง

[1] “PM2.5 คืออะไร” , [Online]. https://www.phyathai.com/article_detail/2884/th/PM2.5,ฝุ่นละออง. [accessed 10 September 2019]

[2] “ข้อมูลดัชนีคุณภาพอากาศ”, [Online] http://air4thai.pcd.go.th/webV2/aqi_info.php [accessed 6 April 2020]

[3] “มลพิษฝุ่นละอองขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5) ของเมืองในประเทศไทย ช่วงเดือนมกราคม-มิถุนายน พ.ศ. 2560”. [Online]. https://www.greenpeace.or.th/s/right-to-clean-air/PM2.5-in-Thailand_Jan-Jun2017.pdf. [accessed 6 April 2020]

[4] “ดัชนีคุณภาพอากาศ (Air Quality Index: AQI) ในพื้นที่กรุงเทพมหานครและปริมณฑล: กรณีฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน” [Online]. https://library2.parliament.go.th/ebook/content-issue/2562/hi2562-004.pdf. [accessed 6 April 2020]

[5] “ปรากฏการณ์อุณหภูมิผกผัน (Temperature inversion)”. [Online] https://stem.in.th/temperature-inversion/. [accessed 11 September 2019]

[6] “สหสัมพันธ์ (Correlation)”. [Online] http://intraserver.nurse.cmu.ac.th/mis/download/course/lec_567730_lesson_07.pdf.[accessed 7 April 2020]

[7] “Correlation Coefficient”. [Online] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#Definition [accessed 13 November 2019]

[8] บุญเสริม กิจศิริกุล. “ปัญญาประดิษฐ์ (ARTIFICIAL INTELLLIGENCE)”. จุฬาลงกรณ์มหาวิทยาลัย มีนาคม 2548

[9] ชุมพล บุญคุ้มพรภัทร. “PATTERN REGCOGNITION”. มหาวิทยาลัยเชียงใหม่ มิถุนายน 2557

[10] “Deep Learning”. [Online] https://www.thaiprogrammer.org/2018/12/deep-learning-คืออะไร/ .[accessed 13 November 2019]

[11] “Overfitting และ Underfitting”. [Online] https://the-ai-midnight.blogspot.com/2018/12/overfitting-underfitting.html. [accessed 7 April 2020]

[12] “Optimization & Activation Function”. [Online] https://medium.com/mmp-li/deep-learning-แบบฉบับสามัญชน-ep-2-optimization-activation-function-เรียนกันสบายๆสไตล์ชิลๆ-9feb5a87e3b2. [accessed 7 April 2020]

[13] “Epoch vs Batch Size vs Iterations”. [Online] https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9. [accessed 7 April 2020]

[14] “Overfitting in Machine Learning: What It Is and How to Prevent It”. [Online] https://elitedatascience.com/overfitting-in-machine-learning. [Accessed 7 April 2020]

[15] “Bagging หรือ Boosting คืออะไร ทำงานอย่างไร?”. [Online] https://tupleblog.github.io/bagging-boosting/. [Accessed 7 April 2020]

[16] “รู้จัก Decision Tree, Random Forest และ XGBoost”. [Online] https://medium.com/@witchapongdaroontham/รู้จัก -decision-tree-random-forrest-และ -xgboost-part-1-cb49c4ac1315. [Accessed 7 April 2020]

[17] “python (Programming language)”, [Online]. https://en.wikipedia.org/wiki/Python_(programming_language. [accessed 10 September 2019]

[18] “TensorFlow”, [Online]. https://www.tensorflow.org/. [accessed 11 September 2019]

[19] “VIIRS I-Band 375 m Active Fire Data”. [Online] https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/viirs-i-band-active-fire-data. [accessed 11 September 2019]

[20] Pawan Gupta and Sundar A. Christopher, “Particulate matter air quality assessment using integrated surface, satellite, and meteorological products”, JOURNAL OF GEOPHYSICAL RESEARCH, vol. 114, D20205, Doi:10.1029/2008JD011497, 2009

[21] Yikai Wang and Xuefei Hu, “A Bayesian Downscaler Model to Estimate Daily PM2.5 levels in the Continental Us”,International Journal of Environmental Research and Public Health, Doi: 10.3390/ijerph15091999,2018

[22] Itai Kloog and Meytar Sorek-Hamer, “Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC Satellite-based AOD data”, AtmosEnviron, Doi:10.1016/j.atmosenv.2015.10.004,2015

[23] Zongwei Ma and Xuefei Hu,” Estimating Ground-Level PM2.5) in China using Satellite Remote sensing”, Environment Science & Technology, vol.48,7436-7444, Doi: dx.doi.org/10.1021/es5009399,2014

[24] Tongwen Li and Huanfeng Shen, “Estimating ground-level PM2.5 by fusing Satellite and station observations: A geo-intelligent deep learning approach”, Geophysical Research Letters, Doi:10.1002/2017GL075710,2017

[25] Mehdi Zamani Joharestani, Chunxiang Cao, Xiliang Ni, Barjeece Bashir, and Somayeh Talebiesfandarani, “PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning using Multisource Remote Sensing Data”, Atmosphere, Doi:https://doi.org/10.3390/atmos10070373, 2019

[26] Massimo Stafoggia, Tom Bellander, Simone Bucci, Marina Davoli et al., “Estimation of Daily PM10 and PM2.5 Concentation in Italy, 2013-2015 use Spatiotemporal Iand use Random Forest model”, vol 124, Environment International, Doi: https://doi.org/10.1016/j.envint.2019.01.016, 2019