การพยากรณ์ปริมาณฝุ่น PM2.5 โดยใช้วิธีวิเคราะห์อนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล กรณีศึกษา: กรุงเทพฯ เขตปทุมวัน

ณภัทร เลาหไพฑูรย์[1] ธนกร สุวรรณโสภณ[2] ทัศภูมิ รันระนา[3] นิตินัย ไพศาลพยัคฆ์[4]*

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม

38 ถนนเพชรเกษม แขวงบางหว้า เขตเจริญกรุง กรุงเทพมหานคร

Email: lonesome_road@windowlive.com

บทคัดย่อ

ฝุ่นละออง PM2.5 เป็นฝุ่นขนาดที่เล็กสามารถเข้าสู่ร่างกายได้โดยตรง ในช่วงปีที่ผ่านมาค่าฝุ่น PM2.5 ได้สูงเกินค่ามาตรฐานที่กรมควบคุมมลพิษและสิ่งแวดล้อมกำหนด ผู้วิจัยจึงได้ทำการวิจัยเพื่อพยากรณ์ปริมาณของฝุ่น PM2.5 โดยใช้เทคนิคเหมืองข้อมูล และใช้ขั้นตอนการทำงานของ CRISP-DM ซึ่งผู้วิจัยได้ศึกษาและเลือกใช้อัลกอริทึม 3 ตัวได้แก่ การถดถอยเชิงเส้น แบบจำลองโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอยมาใช้งาน โดยหลังจากเตรียมข้อมูลได้ใช้โปรแกรม WEKA ในการคำนวนหาผลการพยากรณ์ โดยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอยให้ค่าความแม่นยำสูงสุดที่ 64.76% จึงสามารถนำแบบจำลองไปพยากรณ์ค่าฝุ่นละออง PM2.5 ในเขตอื่นๆเพื่อเฝ้าระวังพร้อมแก้ไขปัญหาต่อไป

คำสำคัญ: *การวิเคราะห์อนุกรมเวลา, ปริมาณฝุ่น PM2.5, เหมืองข้อมูล*

Predict amount of PM2.5 by using Data Mining – Time series techniques Case Study: Pathum Wan District, Bangkok

Napat Laohapaitoon[1], Thanakorn Suwanasophon[2], Tusaphum Runrana[3], Nitinai Phaisanpayak[4]*

Faculty Information Technology Siam University

38 Petchkasem Road, Bang Wa, Charoen Krung, Bangkok

Email: lonesome_road@windowlive.com

Abstract

PM2.5 dust is tiny particles that directly enter the body. In past year PM2.5 is higher than measure that Pollution Control Department set. Researchers study to forecast amount of PM2.5 by using Data Mining technique and by follow working order of CRISP-DM. Researcher using 3 algorithm Linear Regression ,Multilayer Perceptron ,Support Vector Machine, after data preparation use WEKA program to get the result of forecast. Algorithm Support Vector Machine is the highest accurate at 64.76%. After this can bring this model to forecast amount of PM2.5 in other district for surveillance and problem resolution

Keywords: *Time Series Analysis, Amount of Pollution PM2.5, Data Mining*

บทนำ

มลภาวะทางอากาศถือเป็นปัญหาที่สำคัญในช่วงปีที่ผ่านมาหนึ่งในจำนวนนั้นคือ ฝุ่น PM2.5 ที่มีการเพิ่มขึ้นและลดลงในแต่ละช่วงของปี มลภาวะทางอากาศสร้างความอันตรายแก่ประชากรประเทศไทยโดยเฉพาะผู้มีความเสี่ยงสูงและผู้ป่วยเกี่ยวกับโรคทางเดินหายใจ ฝุ่นPM2.5 ซึ่งเป็นอันตรายต่อระบบทางเดินหายใจของสิ่งมีชีวิต ฝุ่นPM2.5 เป็นละอองฝุ่นที่มีขนาดไม่เกิน 2.5 ไมครอน หรือเล็กประมาณ 1 ใน 25 ของเส้นผมของมนุษย์ ซึ่งขนจมูกของมนุษย์ไม่สามารถกรองฝุ่นละอองชนิดนี้ได้ จึงทำให้ฝุ่นละอองชนิดนี้แพร่กระจายเข้าสู่ระบบทางเดินหายใจของมนุษย์ และกระจายตัวเข้าสู่ส่วนต่างๆภายในร่างกายของมนุษย์ ทำให้มีความเสี่ยงที่อาจจะก่อเกิดโรคต่างๆ เช่น โรคหอบหืด ไซนัสอักเสบเรื้อรัง และโรคภูมิแพ้ เป็นต้น (จินตนา ประชุมพันธ์, 2561)

เนื่องจากผู้วิจัยไม่ใช่ผู้รับผิดชอบโดยตรงเกี่ยวกับเรื่องนี้ จึงได้นำผลการดำเนินงานหรือผลการพยากรณ์ไปส่งมอบให้แก่หน่วยงานที่เกี่ยวข้องได้รับผิดชอบต่อไป

นักวิจัยได้ทราบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอยให้ค่าผลลัพธ์ความแม่นยำที่ 63.76%มีความแม่นยำมากที่สุดใน 3 อัลกอริทึมที่เลือกนำมาทดลองให้การพยากรณ์ค่าฝุ่น PM2.5

ซึ่งได้สร้างผลลัพธ์โมเดลในการทำงานแบบใหม่ให้เป็นทางเลือกในการนำไปใช้กับผู้วิจัยท่านอื่นๆที่สามารถนำไปใช้พัฒนาหรืออ้างอิงต่อไป

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการวิเคราะห์ข้อมูลอนุกรมเวลาของปริมาณฝุ่น PM2.5 ในเขตปทุมวัน

วิธีดำเนินการวิจัย

งานวิจัยครั้งนี้มุ่งเน้นการสร้างแบบจำลองเพื่อพยากรณ์ปริมาณฝุ่น PM2.5 โดยดำเนินการวิจัยตามขั้นตอน 3 ขั้นตอนดังนี้

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1.1. การทำเหมืองข้อมูล (Data Mining) คือการค้นหาข้อมูลจากฐานข้อมูลจำนวนมาก เพื่อนำข้อมูลที่ได้มาวิเคราะห์เพื่อช่วยตัดสินใจ ซึ่งการทำเหมืองข้อมูลต้องอาศัยเทคนิคหรือวิธีการต่างๆ เช่น การค้นหาความสัมพันธ์ของข้อมูล การจำแนกกลุ่ม การแบ่งกลุ่มข้อมูล เป็นต้น เพื่อให้ได้ข้อสรุป

1.2. ข้อมูลอนุกรมเวลา (Time Series) คือชุดข้อมูลที่มีการเก็บรวบรวมตามระยะเวลาต่างๆอย่างต่อเนื่องกัน การจัดเก็บข้อมูลแบบอนุกรมเวลามีวัตถุประสงค์เพื่อทำนายสิ่งที่จะเกิดขึ้นในอนาคต โดยอาศัยข้อมูลจากช่วงเวลาต่างๆในอดีต ในงานวิจัยนี้ได้นำเอา 3 เทคนิคที่มีความนิยมนำมาใช้ในการวิเคราะห์ข้อมูล ได้แก่

1.2.1 การวิเคราะห์การถดถอย (Linear Regression) การวิเคราะห์การถดถอยเป็นวิธีการวิเคราะห์ข้อมูลรูปแบบความสัมพันธ์ระหว่างตัวแปรอิสระ (X) และตัวแปรตาม (Y) เช่น ความดันโลหิต มูลค่าการส่งออกสินค้ากับปริมาณการผลิต ความสัมพันธ์ระหว่างค่าใช้จ่ายเพื่อการบริโภคกับรายได้ (Cai et al., 2006: 2159-2179)

1.2.2 โครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron:MLP) เป็นเทคนิคที่จำลองจากการทำงานของสมองมนุษย์ที่มีโครงสร้างหลายๆชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลอย่างดี โดยโครงข่ายประสาทเทียมประกอบไปด้วย ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นผลลัพธ์ เป็นการสอนให้ AI เกิดการเรียนรู้และจดจำจากข้อมูลที่ป้อนเข้าไป เช่น การสอนให้ AI จดจำการเล่นหมากล้อมในหลายๆรูปแบบ (Ghorbanian et al., 2011: 1095 -1105)

1.2.3 ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย (Support Vector Machine for Regression) เป็นเทคนิคที่ใช้ในการจำแนกกลุ่มข้อมูล โดยประเภทข้อมูลออกเป็น 2 ส่วน โดยแยกออกจากกันซัพพอร์ตเวกเตอร์แมชชีนเป็นการนำข้อมูลปัจจุบันและข้อมูลในอดีตจำนวนหนึ่งมาเรียนรู้สำหรับการพยากรณ์ที่เกิดขึ้นในอนาคต (Shevade et al., 2000: 1188 – 1193)

1.3 Root Mean Square Error (RMSE) เป็นค่าที่ใช้ในการวัดขนาดของความคลาดเคลื่อนของการพยากรณ์ โดยค่าดังกล่าวได้จากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย Mean Square Error (MSE) ซึ่งเป็นการนำผลต่างของค่าจริงและค่าที่ได้จากการพยากรณ์ยกกำลังสอง ถ้าค่าผลต่างมีค่ามากจะส่งผลให้ค่าความคลาดเคลื่อนมีค่าที่สูง จึงมีการนำค่าดังกล่าวมาคำนวณด้วยรากที่สอง (Square Root) เพื่อให้ค่าดังกล่าวมีหน่วยวัดเดียวกับค่าที่ทำการทดลอง สำหรับค่า RMSE ที่ได้จากการทดลองมีค่าน้อย แสดงให้เห็นว่าตัวแบบสำหรับพยากรณ์ สามารถทำนายผลลัพธ์ที่มีความคลาดเคลื่อนที่ต่ำ

1.4 Mean Absolute Error (MAE) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเป็นค่าเฉลี่ยของความแตกต่างสัมบูรณ์ระหว่างค่าพยากรณ์กับค่าจริง หากค่า MAE มีค่าน้อย แสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง

การประมาณค่าความแม่นยำ (Evaluation Criterion) จากวิธีต่าง ๆ ที่สร้างขึ้น แบบจำลองที่สร้างขึ้นต้องมีความแม่นยำ เข้ากันได้กับข้อมูลที่ใช้ในการสร้างแบบจำลองนั้นสูงสุด (Model Best Fit) แบบจำลองนี้จะถูกนำไปทดสอบกับกลุ่มข้อมูลชุดที่ทราบค่าจริง (Actual Data) ผลจากการพยากรณ์ข้อมูลชุดใหม่ (Predicted Data) จะถูกนำมาคำนวณหาค่าความคลาดเคลื่อนสัมพัทธ์ (Magnitude of Relative Error: MRE)

หากข้อมูลมีจำนวนมากต้องนำ มาหาค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (Mean Magnitude of Relative Error: MMRE) โดยที่ MMRE มีค่าสูงหมายถึง เปอร์เซ็นต์ของความคลาดเคลื่อนสูง ถ้าค่า MMRE = 0 หมายถึง ค่าของการพยากรณ์เท่ากับค่าจริงทุก ๆ ค่า ถ้า MMRE มีค่าน้อย หมายถึง การพยากรณ์ที่ได้มีความแม่นยำสูง

1.5 Time Lag คือ ข้อมูลย้อนหลังของตัวแปรที่เป็นตัวแปรผันร่วมตามเวลานั้น ๆ โดยย้อนกลับไปจากสภาพปัจจุบันซึ่งมีเหตุการณ์เกิดขึ้น

การดำเนินการตามรูปแบบของ CRISP-DM

ทำความเข้าใจกับปัญหา (Business Understanding)

ในปัจจุบันปัญหามลพิษทางอากาศถือเป็นปัญหาที่สำคัญที่สุด ซึ่งมลพิษทางอากาศที่เป็นอันตรายอย่างมากต่อประชากรที่อาศัยอยู่ในกรุงเทพฯคือ ละอองฝุ่น PM2.5 ซึ่งเป็นฝุ่นขนาดเล็กไม่เกิน 2.5 ไมครอน ซึ่งขนจมูกมนุษย์ไม่สามารถกรองได้ ทำให้ฝุ่นแพร่กระจายเข้าสู่ระบบทางเดินหายใจทำให้เกิดโรคที่เกี่ยวข้องกับทางเดินหายใจ

ทำความเข้าใจกับข้อมูล (Data Understanding)

งานวิจัยชิ้นนี้คณะผู้จัดทำได้รวบรวมและเก็บข้อมูลปริมาณฝุ่นละออง PM2.5 ในช่วงเดือนกรกฎาคม พ.ศ. 2561 ถึง เดือน ตุลาคม พ.ศ. 2562 เขตปทุมวัน เป็นจำนวนทั้งหมด 15 เดือน ซึ่งเป็นชุดข้อมูลจากกรมควบคุมมลพิษ ผ่านทางเว็บไซต์ www.bangkokairquality.com โดยข้อมูลที่รวบรวมมาประกอบไปด้วย แรงลม ทิศทางลม อุณหภูมิ ปริมาณน้ำ ความกดอากาศ และ PM2.5

การเตรียมข้อมูล (Data Preparation)

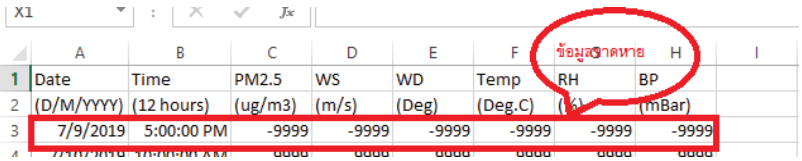
งานวิจัยนี้มุ่งเน้นการสร้างแบบจำลองเพื่อใช้ในการทำนายปริมาณของฝุ่นละออง PM2.5 โดยข้อมูลที่ถูกใช้งานจากฐานข้อมูลจำนวนมากนั่น จำเป็นต้องผ่านการคัดกรองข้อมูล หรือการทำความสะอาดข้อมูล (Data Cleaning) เพื่อคัดเลือกข้อมูลที่ถูกต้องตามวัตถุประสงค์ที่ต้องการและมีความเหมาะสมต่อการนำไปวิเคราะห์ข้อมูล

การคัดเลือกข้อมูล (Select Data)

ในขั้นตอนการเตรียมข้อมูลสำหรับนำไปวิเคราะห์ จากฐานข้อมูลที่มีอยู่เป็นจำนวนมากเพื่อคัดเลือกข้อมูลที่มีความเกี่ยวข้องกันและตรงตามเป้าหมายในการวิจัยครั้งนี้ เช่น แรงลม ทิศทางลม อุณหภูมิ ปริมาณน้ำ ความกดอากาศ และ PM 2.5 โดยข้อมูลที่นำมาวิเคราะห์เริ่มใช้ตั้งแต่กรกฎาคม พ.ศ. 2561 ถึง เดือน ตุลาคม พ.ศ. 2562 เขตปทุมวัน เป็นจำนวนทั้งหมด 15 เดือน

การทำความสะอาดข้อมูล (Clean Data)

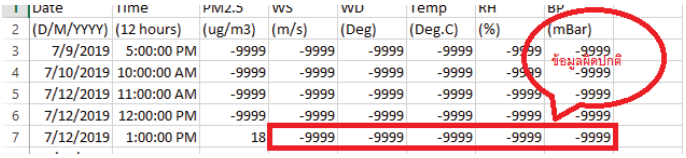
หลังจากทำการคัดกรองข้อมูลเพื่อที่จะนำไปใช้วิเคราะห์แล้ว ทางผู้วิจัยได้ทำความสะอาดข้อมูลที่มีความผิดปกติ หรือข้อมูลที่สูญหาย เช่น ข้อมูลปริมาณฝุ่นละออง PM2.5 มีค่าผิดปกติแม้ในวันที่เก็บจะเป็นช่วงที่มีฝนตก ดังภาพที่ 1



	A	B	C	D	E	F	G	H	I
1	Date	Time	PM2.5	WS	WD	Temp	RH	BP	
2	(D/M/YYYY)	(12 hours)	(ug/m3)	(m/s)	(Deg)	(Deg.C)	(%)	(mBar)	
3	7/9/2019	5:00:00 PM	-9999	-9999	-9999	-9999	-9999	-9999	

ภาพที่ 3.3 ข้อมูลที่สูญหาย (Missing Value)

จากภาพที่ 1 เป็นข้อมูลสูญหาย (Missing Value) กล่าวคือข้อมูลที่มีการจัดเก็บคุณภาพอากาศที่มีการจัดเก็บเป็นรายชั่วโมง พบว่ามีการมีความขาดหายไป ตัวอย่างเช่น บางเดือนข้อมูลที่มีการจัดเก็บก็เริ่มเก็บตั้งแต่วันที่ 7 ซึ่งข้อมูลก่อนหน้านั่นมีการขาดหายไป 6 วัน เป็นต้น

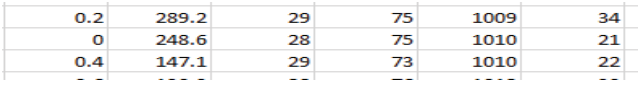


	Date	Time	PM2.5	WS	WD	Temp	RH	BP
2	(D/M/YYYY)	(12 hours)	(ug/m3)	(m/s)	(Deg)	(Deg.C)	(%)	(mBar)
3	7/9/2019	5:00:00 PM	-9999	-9999	-9999	-9999	-9999	-9999
4	7/10/2019	10:00:00 AM	-9999	-9999	-9999	-9999	-9999	-9999
5	7/12/2019	11:00:00 AM	-9999	-9999	-9999	-9999	-9999	-9999
6	7/12/2019	12:00:00 PM	-9999	-9999	-9999	-9999	-9999	-9999
7	7/12/2019	1:00:00 PM	18	-9999	-9999	-9999	-9999	-9999

ภาพที่ 3.4 ข้อมูลที่มีค่าผิดปกติ (Outliers)

จากภาพที่ 2 เป็นข้อมูลที่มีค่าผิดปกติ (Outliers) คือจำนวนเลขที่วัดได้มีค่าสูงผิดปกติเกินความเป็นจริง ดังตัวอย่างภาพที่ 2 จะสังเกตเห็นว่าค่าต่างๆ ที่เก็บมามีต่อเลขผิดปกติอยู่เป็นจำนวนมาก-9999

โครงสร้างข้อมูลใหม่ (Construct Data)



0.2	289.2	29	75	1009	34
0	248.6	28	75	1010	21
0.4	147.1	29	73	1010	22

ภาพที่ 3.5 ข้อมูลที่มีการแก้ไขแล้ว

รูปแบบของข้อมูล (Format Data)

ผู้วิจัยได้นำไฟล์ Format “.CSV” การโปรแกรม Microsoft Excel มาแปลงเป็นไฟล์ Format “.ARFF” เพื่อใช้ในการแสดงผลบนโปรแกรม Weka

การพัฒนาแบบจำลอง (Modeling)

หลังจากรวบรวมข้อมูลปริมาณฝุ่น PM2.5 แล้ว จึงนำข้อมูลที่ได้ผ่านการคัดเลือกนำมาทดสอบกับโปรแกรม Weka (Weikato Environment For Knowledge Analysis) เป็นเครื่องมือที่ใช้ในการวิจัย สำหรับการสร้างแบบจำลองการพยากรณ์ ผู้วิจัยเลือกใช้โปรแกรม Weka version 3.9.3 เพื่อ วิเคราะห์ข้อมูลและสร้างแบบจำลองการพยากรณ์ปริมาณฝุ่น PM2.5 โดยใช้วิธีการวิเคราะห์อนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล (Time Series Data Mining Techniques) โดยใช้เทคนิค

เหมืองข้อมูล 3 เทคนิคคือ การถดถอยเชิงเส้น แบบจำลองโครงข่ายประสาทเทียมแบบเปอร์เซ็ปตรอนหลายชั้น และ ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย โดยผลลัพธ์ของการประมวลผลข้อมูลอนุกรมเวลาจะอยู่ในรูปแบบจำลองของการพยากรณ์ปริมาณฝุ่นละออง PM 2.5 ผู้วิจัยแบ่งข้อมูลออกเป็น 2 ส่วน คือ
ชุดข้อมูลเรียนรู้ (Training Data Set) และชุดข้อมูลทดสอบ (Test Data Set) ดังภาพที่ 3.6

Train Data			Test Data		
Lagged 1 เดือน	Lagged 2 เดือน	Lagged 3 เดือน	Lagged 1 เดือน	Lagged 2 เดือน	Lagged 3 เดือน
มกราคม 2562		มกราคม 2562		มกราคม 2562	
	กุมภาพันธ์ 2562				กุมภาพันธ์ 2562
	มีนาคม 2562	มีนาคม 2562			
		เมษายน 2562	เมษายน 2562		
พฤษภาคม 2562		พฤษภาคม 2562		พฤษภาคม 2562	
					มิถุนายน 2562
	กรกฎาคม 2562	กรกฎาคม 2562			
	สิงหาคม 2562	สิงหาคม 2562	สิงหาคม 2562		
กันยายน 2562		กันยายน 2562		กันยายน 2562	
					ตุลาคม 2562
	พฤศจิกายน 2562	พฤศจิกายน 2562			
	ธันวาคม 2562	ธันวาคม 2562	ธันวาคม 2562		

ภาพที่ 3.6 ชุดข้อมูลสำหรับสอนระบบและทดสอบ

วิเคราะห์โดยใช้การวัดรากของความเคลื่อนที่กำลังสอง และค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ เพื่อแสดงการเปรียบเทียบประสิทธิภาพของแบบจำลองเพื่อใช้ประมาณค่าปริมาณการฝุ่นละออง PM 2.5 ทางคณะผู้วิจัยได้เลือกใช้วิธีทดสอบแบบ 10 Fold – cross validation โดยใช้ข้อมูลแบ่งเป็น 3 Lag ได้แก่

แบบ 1 เดือน กันยายน พ.ศ.2561 ,ธันวาคม พ.ศ.2561 ,เมษายน พ.ศ.2562

แบบ 2 เดือน ตุลาคม พ.ศ.2561 ถึง พฤศจิกายน พ.ศ.2561 ,กุมภาพันธ์ พ.ศ.2562 ถึง มีนาคม พ.ศ.2562 ,มิถุนายน พ.ศ.2562 ถึง กรกฎาคม พ.ศ.2562

แบบ 3 เดือน ตุลาคม พ.ศ.2562 ถึง ธันวาคม พ.ศ.2562 ,กุมภาพันธ์ พ.ศ.2562 ถึง เมษายน พ.ศ.2562 ,กรกฎาคม พ.ศ.2562 ถึง กันยายน พ.ศ.2562

<u>ชุดข้อมูลทดสอบ (Testing Data Set)</u>

ใช้วิธีการประมาณการความแม่นยำในการพยากรณ์ด้วยค่าความคลาดเคลื่อนสัมพัทธ์ สำหรับทดสอบประสิทธิภาพของแบบจำลองโดยแยกแต่ละเดือนในช่วงของชุดข้อมูลทดสอบและใช้ในการวัดประสิทธิภาพโดยรวมของแบบจำลองการพยากรณ์ โดยใช้ข้อมูลทดสอบ 1 เดือนกับข้อมูล Lag ทั้ง 3 แบบ ได้แก่

แบบ 1 เดือน สิงหาคม พ.ศ.2561 ,พฤศจิกายน พ.ศ.2561 ,มีนาคม พ.ศ.2562 ตามลำดับ

แบบ 2 เดือน ธันวาคม พ.ศ.2561 ,เมษายน พ.ศ.2562 ,สิงหาคม พ.ศ.2562 ตามลำดับ

แบบ 3 เดือน มกราคม พ.ศ.2562 ,พฤษภาคม พ.ศ.2562 ,ตุลาคม พ.ศ.2562 ตามลำดับ

<u>การประเมินแบบจำลอง (Evaluation)</u>

หลังจากที่ได้แบบจำลองจากโปรแกรม Weka ที่ได้ผ่านการทดสอบกับชุดข้อมูลสำหรับสอนระบบ ทางผู้วิจัยได้ทำการประเมินผลตัวแบบจำลองทั้ง 3 เทคนิค ที่ได้ทำกับชุดข้อมูลที่แบ่งออกเป็น Lag ทั้ง 3 กลุ่มที่สุ่มแบบคละฤดูการณ์ คือ Lag

แบบ 1 เดือน 2 เดือน และ 3 เดือน จากนั้นนำผลลัพธ์ที่ได้จาก Lag ทั้ง 3 กลุ่ม มาวิเคราะห์โดยใช้การวัดรากของความเคลื่อนที่กำลังสอง (Root Mean Square Error: RMSE) และค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ (Mean Absolute Error: MAE) เพื่อหาว่า กลุ่มข้อมูล Lag ไหน เหมาะจะนำไปใช้กับชุดข้อมูลสำหรับทดสอบ ซึ่งจากผลการทดสอบพบว่า Lag แบบ 2 เดือน มีความเหมาะสมสำหรับการนำไปใช้กับชุดข้อมูลทดสอบ

การนำไปใช้ (Deployment)

หลังจากรวบรวมข้อมูลปริมาณฝุ่น PM2.5 แล้ว จึงนำข้อมูลที่ได้ผ่านการคัดเลือกนำมาทดสอบกับโปรแกรม Weka เป็นเครื่องมือที่ใช้ในการวิจัย สำหรับการสร้างแบบจำลองการพยากรณ์ ผู้วิจัยเลือกใช้โปรแกรม Weka เพื่อวิเคราะห์ข้อมูลและสร้างแบบจำลองการพยากรณ์ปริมาณฝุ่น PM2.5 โดยใช้วิธีการวิเคราะห์อนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล

ผลการวิจัย

ผลการวิจัยพบว่า การวัดประสิทธิภาพแบบจำลองเพื่อใช้ในการพยากรณ์ปริมาณฝุ่นละออง PM2.5 โดยใช้วิธีการทำเหมืองข้อมูลแบบอนุกรมเวลา ผู้ทำวิจัยได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายกับชุดข้อมูล จากชุดข้อมูลที่ผู้วิจัยเลือกปริมาณฝุ่น PM2.5 ซึ่งเป็นชุดข้อมูลสำหรับการสร้างแบบจำลองการพยากรณ์ ผู้วิจัยนำข้อมูล แบ่งเป็น ชุดข้อมูลย้อนหลัง 1 เดือน 2 เดือน 3 เดือน นำมาสร้างแบบจำลองในการทำนายปริมาณฝุ่นละออง PM2.5 เพื่อวิเคราะห์ประสิทธิภาพที่ได้จากเทคนิควิธีการเหมืองข้อมูลทั้ง 3 วิธี โดยการเปรียบเทียบความคลาดเคลื่อนด้วยวิธีการ ค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ (MAE) และ ค่าของรากของความเคลื่อนที่กำลังสอง (RMSE) จากผลการทดลองสามารถแสดงได้ดังตารางที่ 2

ตารางที่ 1 การเปรียบเทียบประสิทธิภาพแบบจำลองการทำนายโดยใช้ชุดข้อมูลย้อนหลัง (Lagged)

Lagged (เดือน)	Time Series Data Mining Techniques (5 Factor)					
	Linear Regression		Multilayer Perceptron		SMOreg	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
ม.ค.62	19.148	24.084	19.8987	24.6497	18.9907	24.4035
พ.ค.61	9.7961	12.1295	10.5016	13.1119	9.7045	12.1883
ก.ย.61	11.9087	16.2741	11.8951	15.6565	11.4411	16.6576
ค่าเฉลี่ย Lag 1 เดือน	13.6176	17.49586667	14.09847	17.80603333	13.37877	17.7498
ก.พ.-มี.ค.62	81.0947	595.4902	79.022	595.9234	43.9418	595.7633
ก.ค.-ส.ค.62	5.2224	6.7157	6.0259	7.6528	5.2051	6.7254
พ.ย.-ธ.ค.62	10.064	12.6199	14.3469	17.7018	10.0045	12.7885
ค่าเฉลี่ย Lag 2 เดือน	32.12703	204.9419333	33.1316	207.0926667	19.71713	205.0924
มีค-พค 62	54.6992	483.9526	52.1587	484.5339	32.9051	484.8111
กค-กย 62	7.0889	10.3798	8.457	11.2078	6.8614	10.6142
พย-มค 62	13.5968	17.9843	14.8892	19.3407	13.4226	18.2638
ค่าเฉลี่ย Lag 3 เดือน	25.1283	170.7722333	25.1683	171.6941333	17.7297	171.2297

จากข้อมูลตารางที่ 1 ผลการพยากรณ์ของแต่ละเทคนิควิธีเหมืองข้อมูลกับการใช้ชุดข้อมูล โดยมีจำนวนเดือนย้อนหลัง (Lagged) ที่แตกต่างกัน โดย MAE และ RMSE เป็นดัชนีชี้วัดประสิทธิภาพ จากการทดลองแสดงให้เห็นว่า เมื่อสร้างแบบจำลองโครงข่ายประสาทเทียม แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย และแบบจำลองด้วยเทคนิคการถดถอยเชิงเส้น ผู้วิจัยพิจารณาแล้วพบว่า เมื่อสร้างข้อมูล ย้อนหลังในเทคนิคเหมืองข้อมูลทั้ง 3 แบบ ด้วยชุดข้อมูลย้อนหลัง1 เดือน 2เดือน 3 เดือน มีประสิทธิภาพสูงที่สุดดังนี้ แบบจำลองโครงข่ายประสาทเทียม 1 เดือนมีค่า MAE เท่ากับ 14.0985 ค่า RMSE เท่ากับ 17.806033 แบบจำลองซัพพอร์ต
เวกเตอร์แมชชีนสำหรับการถดถอย 1 เดือน มีค่า MAE เท่ากับ 13.3788 ค่า RMSE เท่ากับ 17.7498 และแบบจำลองด้วยเทคนิคการถดถอยเชิงเส้น 1 เดือน มีค่า MAE เท่ากับ 13.6176 ค่า RMSE เท่ากับ 17.4958667
สามารถคำนวณได้ดังจากสมการดังนี้

$$MAE = \frac{1}{N}\sum_{i}^{n} = 1(T_i - F_i)$$

T_i คือ ค่าจริง
F_i คือ ค่าพยากรณ์
N คือ จำนวนข้อมูลในชุดข้อมูล

$$RMSE = \sqrt{\frac{\sum_{i}^{n} = 1(Y_i - \hat{Y}_i)^2}{N}}$$

Y_i คือ ค่าที่ได้จากการแบบจำลองพยากรณ์
$\hat{Y}_i$ คือ ค่าจริงที่ใช้ในการสร้างแบบจำลอง
N คือ จำนวนข้อมูลในชุดข้อมูล

จากผลการทดลอง ผู้วิจัยเลือกชุดข้อมูลย้อนหลังที่มีประสิทธิภาพมากที่สุดคือแบบจำลองที่สร้างจากชุดข้อมูลแบบ 1 เดือน นำมาใช้ในการทดลองเพื่อทำนายปริมาณฝุ่น PM2.5 ด้วยชุดข้อมูลทดสอบ โดยแยกเป็นรายชั่วโมงต่อวันจำนวน ด้วยวิธีวิเคราะห์แบบอนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล โดยการใช้อัลกอริทึม 3 แบบ แล้วคำนวณหาค่าความคลาดเคลื่อนสัมพัทธ์ (MRE) และค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (MMRE) โดยได้ผลการเปรียบเทียบประสิทธิภาพแบบจำลองการทำนายในแต่ละเดือน แสดงดังตารางที่ 3

ตารางที่ 3 การเปรียบเทียบประสิทธิภาพแบบจำลองการทำนายในแต่ละเดือน

ปี 2562		Time Series Data Mining Techniques 5 Factors					
		Linear Regression		Multilayer Perceptron		SMOreg	
ช่วงเดือนทดสอบ	Actual	Predict	MRE	Predict	MRE	Predict	MRE
Model ม.ก. :Test ม.ย.	17.15671642	63.2926695	2.689089914	74.30198532	3.330781224	62.92240177	2.667508411
Model ม.ก. :Test ส.ค.	20.62634409	43.84947997	1.125896852	43.49489997	1.108706215	46.75241181	1.266635891
Model ม.ก. :Test ธ.ค.	45.77897574	49.53950277	0.082145285	64.22339426	0.402901512	46.46220982	0.014924626
Model พ.ค. :Test ม.ย.	17.15671642	27.63289804	0.610616937	28.24340048	0.646200811	27.28498787	0.590338571
Model พ.ค. :Test ส.ค.	20.62634409	27.53057651	0.334728849	25.50819846	0.236680546	27.88243452	0.35178752
Model พ.ค. :Test ธ.ค.	45.77897574	23.97089573	0.476377631	64.22339426	0.402901512	23.35997737	0.489722586
Model ก.ย. :Test ม.ย.	17.15671642	51.88371574	2.024105223	62.05464804	2.616930334	39.23591737	1.286912974
Model ก.ย. :Test ส.ค.	20.62634409	26.83292773	0.300905658	28.48653789	0.381075472	24.29942867	0.178077344
Model ก.ย. :Test ธ.ค.	45.77897574	27.23617197	0.405050648	28.72352783	0.372560714	40.83873694	0.107915014
MMRE		0.691964158		0.776058231		0.500792335	
ACCURACY(%)		63.67%		32.75%		64.76%	

$Actual_i$ คือ ค่าจริง

$predicted_i$ คือ ค่าพยากรณ์

$$MRE_i = \frac{Actual_i - predicted_i}{Actual_i}$$

$Actual_i$ คือ ค่าจริง

$predicted_i$ คือ ค่าพยากรณ์

$$MMRE = \frac{1}{n}\sum_{i}^{n}=1\frac{Actual_i - predicted_i}{Actual_i} \times 100$$

สรุปผลการวิจัย

จากการทดลองแสดงให้เห็นว่า ซัพพอร์ตเวกเตอร์แมชชีนสำหรับถดถอยเหมาะที่สุดในการสร้างแบบจำลองการพยากรณ์ โดย มีค่า MMRE เท่ากับ 0.500792335 ค่าความแม่นยำที่ 64.76% การจำลองด้วยเทคนิคการถดถอยเชิงเส้น มีค่า MMRE เท่ากับ 0.691964158 ค่าความแม่นยำที่ 63.67% และแบบจำลองโครงข่ายประสาทเทียมให้ค่า MMRE เท่ากับ 0.776058231 และความแม่นยำที่ 32.75%

อภิปรายผลการวิจัย

จากวัตถุประสงค์ในการวิจัยที่ต้องการหาอัลกอริทึมที่เหมาะสมที่สุดในการสร้างแบบจำลอง ซึ่งผลที่ได้รับคือ ซัพพอร์ตเวกเตอร์แมชชีน มีความคลาดเคลื่อนน้อยที่สุด และมีความแม่นยำสูงที่สุด

ข้อเสนอแนะ

ผู้วิจัยจึงสรุปเป็นข้อเสนอแนะเพื่อเป็นแนวทางในการวางแผนการให้กับหน่วยงานที่เกี่ยวข้องนำมาแก้ไขปรับใช้ตามสถานการณ์อื่นๆ เพื่อเป็นแนวทางในการป้องกันหรือลดปริมาณการเกิดฝุ่น PM2.5 โดยการณรงค์ลดการใช้รถส่วนตัวในการเดินทาง เพื่อช่วยลดปริมาณฝุ่น PM2.5 ซึ่งเป็นมลพิษที่สำคัญตัวหนึ่งที่เกิดจากท่อไอเสียของรถยนต์ (กัมปนาท เทียนน้อย, 2555: 124-139) นอกจากนี้ทางผู้นำชุมชนควรสร้างความตระหนักและปลูกจิตใต้สำนึกให้กับชุมชน ในเรื่องการเผาป่าหรือการเพาะปลูกโดยใช้วิธีการเผาเป็นต้น (เรไร ลำเจียก, 2558: 27)ผู้วิจัยจึงได้นำเสนอข้อเสนอแนะเหล่านี้ไว้เป็นอีกแนวทางหนึ่งในการแก้ไขปัญหา

เอกสารอ้างอิง

จินตนา ประชุมพันธ์. (2561). **PM2.5 ฝุ่นละอองขนาดเล็กในอากาศ กับวิกฤตสุขภาพที่คนไทยจะต้องแลก.** ค้นเมื่อวันที่ 20 พฤษภาคม 2562 จาก https://thestandard.co/pm-2-5-environmental-nano-pollutants/

พิสุทธิ์ เพียรมนกุล. (2562). **สาเหตุที่แท้ของ PM 2.5 เรียนรู้ เข้าใจ ป้องกัน ไม่ตื่นตระหนก.** ค้นเมื่อวันที่ 21 พฤษภาคม 2562 จาก https://www.khaosod.co.th/pr-news/news_2171737

วิษณุ อรรถวานิช. (2562). **ต้นทุนของสังคมไทยจากมลภาวะทางอากาศและมาตรการณ์รับมือ.** ค้นเมื่อวันที่ 20 พฤษภาคม 2562 จาก https://www.prachachat.net/columns/news-291113

วีระยุทธ พิมพาภรณ์, ปรียาภรณ์ พูลทอง และ บุษกร แก้ววิเชียร. (2559). **การพยากรณ์ปริมาณน้ำไหลเข้าอ่างเก็บน้ำโดยใช้วิธีวิเคราะห์อนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล.** อาจารย์ประจำ หลักสูตรวิทยาศาสตร์บัณฑิต ภาควิชาวิทยาการคอมพิวเตอร์ และสารสนเทศ คณะวิทยาศาสตร์ ศรีราชามหาวิทยาลัยเกษตรศาสตร์ ศรีราชา.

Amnuaysak Thoonsiri (2562**). ควันขาวผสมควันดำ กรรมของคนกรุงเทพฯ.** ค้นเมื่อวันที่ 20 พฤษภาคม 2562 จาก https://www.thairath.co.th/news/local/1488207"1488207

Cai, T.T., Hall, P., & others. (2006). **Prediction in functional linear regression. The Annals of Statistics**. 34(5), 2159-2179.

Ghorbanian, J., Ahmadi, M., & Soltani, R. (2011). **Design predictive tool and optimization of journal bearing using neural network model and multi-objective geneticalgorithm. Scientia Iranica**. 18(5): 1095 -1105.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). **Improvements to the SMO algorithm for SVM regressioan. IEEE Transactions on Neural Networks.** 11(5): 1188 – 1193.

Smola, A. J., & Schölkopf, B. (2004). **A tutorial on support vector regression. Statistics and Computing**.

Thanyaporn Bunthong (2561). **ฝุ่น: เหตุใดสถานการณ์ฝุ่นละอองขนาดเล็ก จึงพุ่งสูงขึ้นมาอีกครั้ง.** ค้นเมื่อวันที่ 20 พฤษภาคม 2562 จาก https://www.bbc.com/thai/thailand-46643980"46643980