การพยากรณ์ปริมาณน้ำฝนระยะสั้นในบริเวณพื้นที่สนามบินสุวรรณภูมิด้วยโครงข่ายระบบประสาทแบบย้อนกลับ

น.ส.รักษ์คณา ภูสีเขียว

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Short Term Precipitation Forecasting using Recurrent Neural Networks, a Case Study of Suvarnabhumi Airport.

Miss Ragkana Phooseekhiwe

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Statistics
Department of Statistics
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2021

หัวข้อวิทยานิพนธ์	การพยากรณ์ปริมาณน้ำฝนระยะสั้นในบริเวณพื้นที่สนามบินสุวรรณภูมิด้วยโครงข่ายระบบประสาทแบบย้อนกลับ
โดย	น.ส.รักษ์คณา ภูสีเขียว
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

..... กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร.วรพจน์ กรีสุระเดช)

รักษ์คณา ภูสีเขียว : การพยากรณ์ปริมาณน้ำฝนระยะสั้นในบริเวณพื้นที่สนามบินสุวรรณภูมิด้วยโครงข่ายระบบประสาทแบบย้อนกลับ. (Short Term Precipitation Forecasting using Recurrent Neural Networks, a Case Study of Suvarnabhumi Airport.) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุรณพีร์ ภูมิวุฒิสาร

ปริมาณน้ำฝนนับเป็นปัจจัยสำคัญอย่างหนึ่งที่มีผลต่อการดำเนินชีวิตของมนุษย์ การพยากรณ์ปริมาณน้ำฝนที่มีความแม่นยำช่วยให้มนุษย์เตรียมพร้อมสำหรับกิจกรรมต่างๆ ที่จะเกิดขึ้นในอนาคตได้ดี อย่างไรก็ตามในบางสถานการณ์ความพร้อมใช้งานของข้อมูลสภาพอากาศมีจำกัด ทำให้การพยากรณ์ปริมาณน้ำฝนอย่างแม่นยำนั้นเป็นเรื่องที่ยาก ปัจจุบันหลายๆ งานวิจัยที่เกี่ยวข้องได้เลือกโครงข่ายประสาทเทียมเชิงลึกเป็นอัลกอริทึมในการฝึกแบบจำลองเพื่อใช้ในการพยากรณ์ แนวคิดหลักคือการสร้างตัวแปรคุณลักษณะ (Feature) ที่เกี่ยวข้องในระดับสถาปัตยกรรม จากหลักการนี้สถาปัตยกรรมโครงข่ายประสาทเทียมเชิงลึกที่เหมาะสมสามารถผสมผสานและจับคู่คุณลักษณะที่เกี่ยวข้องในการพยากรณ์ได้อย่างเหมาะสม ผลที่ตามมางานวิจัยที่มีอยู่ส่วนใหญ่จึงมุ่งเน้นไปที่เทคนิคบางอย่างเพื่อปรับปรุงประสิทธิภาพของแบบจำลองโดยไม่ได้ให้ความสำคัญกับการเพิ่มคุณลักษณะให้กับตัวแบบมากนัก อย่างไรก็ตามเมื่อข้อมูลการฝึกฝนมีจำนวนจำกัดโครงข่ายประสาทเทียมเชิงลึกอาจจะทำงานได้ไม่เต็มประสิทธิภาพมากนัก ทำให้การผสมผสานและจับคู่คุณลักษณะที่เกี่ยวข้องในการพยากรณ์ทำได้ไม่ดีตามไปด้วย สิ่งนี้ทำให้เกิดคำถามงานวิจัยว่าแบบจำลองการพยากรณ์ปริมาณน้ำฝนที่ได้ถูกนำเสนอมีประสิทธิภาพที่ดีเพียงพอหรือไม่ เมื่อไม่ได้มีการเพิ่มคุณสมบัติที่เกี่ยวข้องให้กับแบบจำลอง งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของแบบจำลองต่างๆ ในการพยากรณ์ปริมาณน้ำฝนสะสมในระยะสั้นที่มีและไม่มีการเพิ่มตัวแปรคุณสมบัติที่เกี่ยวข้อง โดยได้แบ่งการทดลองออกเป็น 2 ส่วนเพื่อวัดประสิทธิภาพ คือ 1) การเปรียบเทียบประสิทธิภาพของตัวแบบที่มีการเพิ่มตัวแปรคุณลักษณะที่เกี่ยวข้องว่ามีความถูกต้องแม่นยำดีขึ้นหรือไม่เมื่อเทียบกับแบบจำลองที่ไม่ได้มีการเพิ่มตัวแปรคุณลักษณะในสภาพแวดล้อมที่เทียบเท่ากัน และ 2) การเปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมของแบบจำลองที่สนใจศึกษา ได้แก่ ARIMA ARIMAX RNN LSTM และ GRU ข้อมูลที่นำมาใช้ในงานวิจัยนี้เป็นข้อมูลสภาพอากาศและปริมาณน้ำฝนสะสมที่รวบรวมมาจากพื้นที่สนามบินสุวรรณภูมิ จากผลการศึกษาทั้ง 2 ส่วนพบว่าการเพิ่มตัวแปรคุณลักษณะสามารถเพิ่มประสิทธิภาพการพยากรณ์ให้กับตัวแบบได้ในกรณีที่ข้อมูลที่นำมาฝึกฝนตัวแบบมีจำนวนจำกัด โดยแบบจำลอง GRU ให้ประสิทธิภาพในการพยากรณ์มากที่สุด

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....
ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6280266526 : MAJOR STATISTICS

KEYWORD: Long Short-Term Memory, Predictive analytics, Rainfall prediction, Recurrent Neural Network, Neural Network, Gated Recurrent Unit, Imbalanced data

Ragkana Phooseekhiwe : Short Term Precipitation Forecasting using Recurrent Neural Networks, a Case Study of Suvarnabhumi Airport.. Advisor: Asst. Prof. SURONAPEE PHOOMVUTHISARN

Rainfall is one of the key important factors affecting human life. Accurate precipitation forecasting allows humans to better prepare for various activities that will happen in the future. However, in some situations, the availability of weather data is limited, making it difficult to make accurate forecasting. Currently, much research has chosen deep neural networks as an algorithm to train forecasting models. The main idea is to come up with relevant features at the architecture level. Based on this paradigm, it has been shown that the appropriate deep neural network architecture can flexibly mix and match features that are relevant in predictions. Consequently, most of the existing research then focuses on some of the techniques to improve the performance of the models without paying much attention on the issues of adding relevant features. However, when the training data is limited, deep neural networks might not scale very well, thus making it difficult to mix and match features for predictions. This imposes a research question of how effective the proposed forecasting models might be when not adding relevant features to the models. The aims of this research are to develop and compare the efficiency of various models for short-term precipitation forecasting with and without adding relevant features. The experiment consists of 2 parts: (1) The experiment by exploring whether the models with relevant featured provided can achieve higher accuracy compared to original models in equivalent environments, and (2) The experiment by comparing accuracy between 5 models (ARIMA, ARIMAX, RNN, LSTM and GRU). The weather dataset, which is very limited in quantity, used in this research was collected from Suvarnabhumi Airport. The results show that adding relevant features can enhance the forecasting performance of the model when the weather data is limited. In addition, the GRU model with relevant featured provided is the most effective prediction.

Field of Study:	Statistics	Student's Signature .....
Academic Year:	2021	Advisor's Signature .....

กิตติกรรมประกาศ

การที่วิทยานิพนธ์เล่มนี้เสร็จสมบูรณ์ได้ด้วยดีนั้น เกิดจากความช่วยเหลือและเอาใจใส่จากบุคคลสำคัญหลายท่าน อันเป็นส่วนสำคัญในการจัดทำวิทยานิพนธ์ฉบับนี้ ผู้วิจัยจึงได้จัดทำในส่วนของกิตติกรรมประกาศนี้ขึ้น เพื่อแสดงความขอบพระคุณอย่างสูงมา ณ ที่นี้

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้เป็นแรงบันดาลใจให้ผู้วิจัยมีความสนใจในการศึกษาเกี่ยวกับการพัฒนาแบบจำลองการเรียนรู้เชิงลึก ที่สามารถนำมาช่วยในการพยากรณ์ข้อมูลให้เกิดความแม่นยำมากขึ้น อีกทั้งยังคอยผลักดัน ให้คำปรึกษา ให้ความรู้ที่เกี่ยวข้อง ตลอดจนให้ความช่วยเหลือในการปรับปรุงแก้ไขตัววิทยานิพนธ์ จนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ขอขอบพระคุณ รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์ ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์ และรองศาสตราจารย์ ดร.วรพจน์ กรีสุระเดช กรรมการสอบวิทยานิพนธ์ ที่ให้เกียรติเป็นคณะกรรมการในการสอบวิทยานิพนธ์ให้กับผู้วิจัย อีกทั้งยังให้คำแนะนำ ตรวจสอบและช่วยเหลือในการปรับปรุงแก้ไขตัววิทยานิพนธ์ให้มีความสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณ คุณนพสิทธิ์ ศรีพล เจ้าหน้าที่กลุ่มบริการสารสนเทศอุตุนิยมวิทยา สำนักบริการดิจิทัล กรมอุตุนิยมวิทยา ที่ให้คำปรึกษา และสนับสนุนเกี่ยวกับข้อมูลอุตุนิยมทั้งหมด ซึ่งเป็นส่วนสำคัญที่ต้องนำมาใช้ในการจัดทำวิทยานิพนธ์ฉบับนี้

ขอขอบพระคุณครอบครัวของผู้วิจัย ผู้คอยส่งเสริม และสนับสนุนในทุกๆ ด้าน เพื่อให้ผู้วิจัยได้มีโอกาสในการศึกษาต่อในระดับปริญญาโทดังที่ต้องการ จนกระทั่งสำเร็จการศึกษา และสุดท้ายขอขอบคุณแมว และเพื่อนๆ อันเป็นที่รัก ที่คอยช่วยเหลือ ผลักดัน และให้กำลังใจแก่ผู้วิจัยเสมอมา

รักษ์คณา ภูสีเขียว

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

สภาพอากาศถือเป็นอีกหนึ่งสิ่งที่มีความเกี่ยวพันกับมนุษย์มาตั้งแต่สมัยอดีต เพราะไม่ว่าจะเป็นการใช้ชีวิตประจำวัน การเดินทาง การประกอบอาชีพ (เช่น การทำการเกษตร ประมง เป็นต้น) หรือแม้แต่ใช้ในการตัดสินใจในการบริหารจัดการทรัพยากรระดับประเทศ ต่างก็มีความเกี่ยวข้องกับสภาพอากาศทั้งสิ้น และในบรรดาปรากฏการณ์ที่เกี่ยวข้องกับสภาพอากาศทั้งหมด ปรากฏการณ์ฝนตกก็ถือเป็นหนึ่งในปรากฏการณ์ที่สำคัญมากในระบบภูมิอากาศ เพราะเป็นปรากฏการณ์ที่ส่งผลโดยตรงกับระบบนิเวศ แต่การจะพยากรณ์ปริมาณน้ำฝนอย่างแม่นยำ และมีประสิทธิภาพยังคงเป็นปัญหาสำคัญที่รัฐบาลหรือแม้แต่ภาคธุรกิจยังคงให้ความสนใจ เพราะนอกจากปัจจัยที่จะทำให้เกิดปรากฏการณ์ฝนตกนั้นจะมีความซับซ้อนแล้ว ตัวปรากฏการณ์ฝนตกเองก็เหมือนเป็นปรากฏการณ์แบบสุ่มที่ไม่อาจคาดเดาได้เช่นกัน เนื่องจากลักษณะข้อมูลที่ไม่ได้มีความสัมพันธ์แบบเชิงเส้น (non-linear) และเป็นรูปแบบสุ่ม (stochastic) เพื่อที่จะสามารถคาดการณ์หรือพยากรณ์ปริมาณน้ำฝนที่จะตกลงมาในช่วงต่างๆได้ จึงได้มีการนำเอาศาสตร์ต่างๆ เข้ามาช่วย

ในเริ่มแรก แบบจำลองที่มักถูกนำมาใช้ในการพยากรณ์ข้อมูลปริมาณน้ำฝนก็คือแบบจำลองอาริมา(ARIMA) ดังเช่นในงานวิจัยของ (Wang et al., 2013) ที่มีการนำตัวแบบ SARIMA มาใช้ในการวิเคราะห์และพยากรณ์ กับงานวิจัยของ (Narayanan et al., 2013) ก็ได้ใช้ตัวแบบจำลอง ARIMA ในการพยากรณ์ปริมาณน้ำฝนในแต่ละสถานี พบว่าตัวแบบ ARIMA นั้นไม่ได้ตอบโจทย์สำหรับทุกสถานี เนื่องจากตัวแบบจำลอง ARIMA นั้นเป็นตัวแบบจำลองที่พิจารณาโดยใช้ตัวแปรเพียงตัวแปรเดียว แต่ในความเป็นจริง ปัจจัยหรือตัวแปรที่ส่งผลกับปริมาณน้ำฝนที่เกิดขึ้นในแต่ละช่วงเวลานั้นไม่ได้มีเพียงปัจจัยเดียว เพื่อให้สามารถนำปัจจัยร่วมอื่นๆ มาพิจารณาร่วมกับปริมาณน้ำฝนสะสมได้ ในงานวิจัยของ (Jalalkamali et al., 2015) จึงได้เลือกใช้แบบจำลอง ARIMAX ในการพยากรณ์ปริมาณน้ำฝนเพื่อคาดการณ์ปัญหาภัยแล้ง

จากปัญหาที่ได้กล่าวไปข้างต้น ทำให้ในปัจจุบัน งานวิจัยส่วนใหญ่ที่เกี่ยวข้องกับการพยากรณ์อากาศ หรือการพยากรณ์ปริมาณน้ำฝน ได้เลือกใช้โครงข่ายประสาทเทียมเข้ามาช่วยในการพยากรณ์

อย่างเช่นในงานวิจัยของ (Hung et al., 2009) ที่ได้มีการนำโครงข่ายประสาทเทียม (ANN) เข้ามาช่วยในการวิเคราะห์นั้น ทำให้สามารถเพิ่มตัวแปรหรือคุณลักษณะด้านสภาพอากาศอื่นๆ เข้าไปเพื่อเพิ่มประสิทธิภาพในการพยากรณ์ ทั้งนี้ การสร้างตัวแบบโดยอาศัยโครงข่ายประสาทเทียมนั้นก็มีเทคนิคหรือวิธีการแตกแขนงแยกย่อยลงไปอีกมากมาย แต่ในการจัดการกับข้อมูลภูมิอากาศนั้น การใช้แบบจำลองการเรียนรู้เชิงลึก (Deep Learning) เช่น โครงข่ายระบบประสาทแบบย้อนกลับ (RNNs) ดูจะเป็นตัวเลือกที่เหมาะสมที่สุดเนื่องจากโครงข่ายดังกล่าวนั้นถูกพัฒนามาเพื่อจัดการกับข้อมูลที่มีลักษณะเป็นลำดับ หรืออนุกรมเวลา

อย่างไรก็ตาม ในแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับ ที่ได้กล่าวไปก่อนหน้านี้นั้น ได้มีการพัฒนาและต่อยอด ทำให้เกิดตัวแบบขึ้นมาทั้งหมด 3 ตัวแบบ อันได้แก่ โครงข่ายระบบประสาทแบบย้อนกลับ (RNNs) แบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long short-term memory) และโครงข่ายประตูวกกลับ (Gated Recurrent Unit) ซึ่งต่อมา งานวิจัยส่วนใหญ่ก็ได้พยายาม พัฒนาต่อยอดตัวแบบโครงข่ายระบบประสาทย้อนกลับเพื่อเพิ่มประสิทธิภาพในการพยากรณ์ให้กับข้อมูล ดังเช่นในงานวิจัยของ (Poornima & Pushpalatha, 2019) ได้ทำการพัฒนาแบบจำลอง Intensified LSTM และทำการเปรียบเทียบประสิทธิภาพของแบบจำลองดังกล่าวกับแบบจำลองอาริมาและโครงข่ายระบบประสาทแบบย้อนกลับ ซึ่งพบว่าแบบจำลองที่เสนอนั้นมีความแม่นยำในการพยากรณ์มากกว่าอีก 2 แบบจำลอง ในขณะที่บางงานวิจัยก็ต้องการทราบว่าตัวแบบโครงข่ายระบบประสาทแบบใดเป็นตัวแบบที่มีประสิทธิภาพมากที่สุด จึงมีงานวิจัยที่พยายามเปรียบเทียบประสิทธิภาพของตัวแบบทั้ง 3 กับชุดข้อมูลในรูปแบบที่หลากหลาย ดังเช่น งานวิจัยของ (Yang et al., 2020) ได้ทดลองทำการเปรียบเทียบประสิทธิภาพของแบบจำลองหน่วยความจำระยะสั้นแบบยาวกับโครงข่ายประตูวกกลับ พบว่า นอกจากประสิทธิภาพของแบบจำลองทั้ง 2 จะดีหรือไม่ขึ้นอยู่กับจำนวนข้อมูลแล้วยังขึ้นกับและประเภทของข้อมูลที่นำมาวิเคราะห์ด้วย เช่นเดียวกับในงานวิจัยของ (Aurnhammer & Frank, 2019) และ (Shewalkar, 2019) ได้ทำการศึกษาเปรียบเทียบประสิทธิภาพของโครงข่ายระบบประสาทแบบย้อนกลับกับชุดข้อมูลที่มีความหลากหลาย ซึ่งพบว่า แบบจำลอง GRU และ LSTM นั้น มีความเหมาะสมในการคาดการณ์หรือพยากรณ์กับชุดข้อมูลหลายๆประเภท แต่ทั้งนี้ ขึ้นอยู่กับตัว Short-term memory ด้วยเช่นกัน อย่างไรก็ตาม งานวิจัยที่กล่าวมาข้างต้นมุ่งเน้นในเรื่องการปรับปรุงอัลกอริทึมของตัวแบบจำลอง (เนื่องจากตัวแบบจำลองการเรียนรู้เชิงลึกนั้นมีความสามารถในการผสมผสาน และจับคู่คุณลักษณะที่เกี่ยวข้องกับ

คุณลักษณะที่ต้องการพยากรณ์ รวมไปถึงการสกัดคุณลักษณะที่ไม่มีความสำคัญ หรือส่งผลต่อการพยากรณ์ออกไป) จึงไม่ได้ทำการทดลองเกี่ยวกับการปรับปรุง แก้ไขข้อมูล หรือเพิ่มคุณลักษณะใดๆ (Feature Engineering) เพื่อเพิ่มประสิทธิภาพในการพยากรณ์

งานวิจัยนี้ จึงมุ่งเน้นการพัฒนาและวิเคราะห์เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมล่วงหน้าในระยะสั้นร่วมกับการใช้ข้อมูลหลายตัวแปร โดยตัวแบบที่เหมาะสมจะมาจากผลการศึกษาและเปรียบเทียบใน 2 ประเด็นหลักคือ 1) การเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบที่ใช้ชุดข้อมูลที่แตกต่างกัน ระหว่างชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะพิเศษกับชุดข้อมูลดั้งเดิม 2) เปรียบเทียบประสิทธิภาพในการพยากรณ์ในช่วงเวลาถัดไปของโครงข่ายระบบประสาทแบบย้อนกลับที่สนใจศึกษากับแบบจำลอง ARIMA และแบบจำลอง ARIMAX ซึ่งเป็นแบบจำลองที่เป็นที่นิยมทางสถิติในการพยากรณ์ข้อมูลประเภทอนุกรมเวลา และสำหรับโครงข่ายระบบประสาทแบบย้อนกลับที่สนใจนำมาศึกษาเปรียบเทียบ จะมีทั้งหมด 3 ตัว ได้แก่ โครงข่ายระบบประสาทแบบย้อนกลับ (RNN) แบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long short-term memory) และโครงข่ายประตูวกกลับ (Gated Recurrent Unit)

งานวิจัยนี้ได้ใช้ข้อมูลสภาพอากาศรายชั่วโมงจากสถานีตรวจวัดสภาพอากาศสนามบินนานาชาติสุวรรณภูมิ จังหวัดสมุทรปราการ (ข้อมูลดังกล่าวได้รับการสนับสนุนจากกรมอุตุนิยมวิทยา) ชุดข้อมูลดังกล่าวมีการเก็บสะสมตั้งแต่ปี ค.ศ. 2016 ถึงปี ค.ศ. 2020 โดยเราจะนำชุดข้อมูลนี้มาสร้างโครงข่ายระบบประสาทแบบย้อนกลับ เพื่อพยากรณ์ปริมาณน้ำฝนในอีก 3 ชั่วโมงข้างหน้า และเพื่อเพิ่มประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝน จึงจะทำการเพิ่มคุณลักษณะบางอย่างเข้าไปก่อนที่จะนำข้อมูลเข้าไปวิเคราะห์ในโครงข่ายระบบประสาทแบบย้อนกลับ พร้อมทั้งเปรียบเทียบค่าความแม่นยำในการพยากรณ์ที่ได้จากการพัฒนาตัวแบบต่างๆ

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนาแบบจำลองที่เหมาะสมต่อการพยากรณ์ปริมาณน้ำฝน
2. เพื่อการเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบที่ใช้ชุดข้อมูลที่แตกต่างกัน ระหว่างชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะพิเศษกับชุดข้อมูลดั้งเดิม เพื่อทดสอบว่า การเพิ่มคุณลักษณะทางสถิติให้กับชุดข้อมูลก่อนที่จะนำเข้าแบบจำลองนั้น จะช่วยเพิ่มประสิทธิภาพให้กับผลการพยากรณ์ได้หรือไม่

3. เพื่อเปรียบเทียบค่าความแม่นยำในการพยากรณ์ที่ได้จากการพัฒนาตัวแบบจำลองอนุกรมเวลา(แบบจำลอง ARIMA และแบบจำลอง ARIMAX) แบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) โครงข่ายระบบประสาทแบบย้อนกลับ (RNN) และโครงข่ายประตูวกกลับ (GRU)

1.3 คำจำกัดความที่ใช้ในการวิจัย

1. การพยากรณ์อากาศ (Weather forecast) คือ การคาดหมายสภาวะของลมฟ้าอากาศ รวมทั้งปรากฏการณ์ทางธรรมชาติที่อาจจะเกิดขึ้นในช่วงเวลาข้างหน้า
2. การพยากรณ์ปริมาณน้ำฝน (Rainfall prediction) คือการคาดการณ์ปริมาณน้ำฝนที่จะตกลงมา โดยในการคาดการณ์ดังกล่าวนั้นสามารถบ่งบอกได้ถึงระดับความรุนแรงของฝนที่จะตกด้วย
3. คุณลักษณะ (Feature) คือ ตัวแปร หรือคุณสมบัติของสิ่งที่เราต้องการศึกษาหรือสังเกต ซึ่งเป็นสิ่งที่สามารถวัดได้ด้วยค่าทางสถิติ เพื่อให้สามารถนำไปใช้ประกอบการวิเคราะห์ คำนวณหรือพยากรณ์ทางสถิติได้
4. ชุดข้อมูล (Data set) คือ กลุ่มของคุณลักษณะที่มากกว่าหรือเท่ากับ 1 คุณลักษณะขึ้นไป ที่จะนำเข้าตัวแบบเพื่อสร้างแบบจำลองในการพยากรณ์
5. ฟังก์ชันกระตุ้น (Activation function: $f_{..}$) ทำหน้าที่รวมค่าเชิงตัวเลขจากตัวแปรเข้า (Input)ที่เข้ามาในโหนด แล้วทำการตัดสินใจว่าจะส่งสัญญาณเอาต์พุตออกไปในรูปใด ค่าของเอาต์พุตที่ออกไปจากโหนดจะเรียกว่า Activation โดยที่ฟังก์ชันกระตุ้นอาจเป็นฟังก์ชันเส้นตรงหรือไม่ก็ได้ ฟังก์ชันกระตุ้นมีหลายรูปแบบ ขึ้นอยู่กับลักษณะงาน
6. Bias Unit ($b_{..}$) คือ ตัวเลขที่บวกเข้าไปเพื่อปรับให้ค่าที่คำนวณออกมาถูกต้องมากขึ้น โดยมักจะมีค่าเท่ากับ 1 เสมอ ซึ่งหากมองในมุมมองของสมการเส้นตรง bias จะทำหน้าที่เหมือนกับจุดตัดแกน y ซึ่งสามารถทำให้สมการเส้นตรง สามารถเลื่อนไปในทิศทางไหนก็ได้ โดยที่ความชันเท่าเดิม
7. เมทริกซ์ค่าถ่วงน้ำหนัก (Weight Matrix: $W_{..}$) คือ ค่าน้ำหนักที่มอบให้กับข้อมูลแต่ละตัว โดยน้ำหนัก(Weight) ของแต่ละข้อมูลจะสะท้อนถึงความสำคัญของข้อมูลแต่ละตัวที่นำมาคำนวณในตัวแบบ หมายความว่า หากค่าถ่วงน้ำหนักมีค่าเหมาะสม โครงข่ายประสาทเทียมก็จะมีประสิทธิภาพสูงตามไปด้วย

1.4 ขอบเขตการวิจัย

1.4.1 ขอบเขตเนื้อหา

การพัฒนาและวิเคราะห์เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมล่วงหน้าในระยะสั้นร่วมกับการใช้ข้อมูลหลายตัวแปร โดยตัวแบบที่เหมาะสมจะมาจากผลการศึกษาและเปรียบเทียบใน 2 ประเด็น

1) การเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบที่ใช้ชุดข้อมูลที่แตกต่างกันระหว่างชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะพิเศษกับชุดข้อมูลดั้งเดิม

2) เปรียบเทียบประสิทธิภาพในการพยากรณ์ในช่วงเวลาถัดไปของโครงข่ายระบบประสาทแบบย้อนกลับที่สนใจศึกษากับแบบจำลองอาริมา

ในการทดสอบความถูกต้องของผลการพยากรณ์จะใช้เกณฑ์ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE)

1.4.2 ขอบเขตพื้นที่

ชุดข้อมูลที่นำมาใช้ในงานวิจัยนี้เป็นข้อมูลสถิติสภาพอากาศรายชั่วโมงของพื้นที่สนามบินสุวรรณภูมิ จังหวัดสมุทรปราการ

1.4.3 ขอบเขตเวลา

ข้อมูลสถิติที่รวบรวมมาประกอบไปด้วย อุณหภูมิ ค่าความกดอากาศ และค่าความชื้นสัมพันธ์เป็นข้อมูลรายชั่วโมง และข้อมูลปริมาณน้ำฝนสะสม(มิลลิเมตร) เป็นราย 3 ชั่วโมง จากสถานีเก็บข้อมูลของกรมอุตุนิยมวิทยา ตั้งแต่เดือนมกราคม ปี ค.ศ. 2016 จนถึงเดือนธันวาคม ปี ค.ศ. 2020 เป็นระยะเวลาทั้งสิ้น 4 ปี

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้แบบจำลองที่เหมาะสม และมีประสิทธิภาพในการพยากรณ์แนวโน้มของปริมาณน้ำฝน
2. เพื่อช่วยให้ผู้ที่ต้องการทราบสภาพอากาศล่วงหน้า ได้นำผลที่ได้จากการพยากรณ์ไปใช้ในการตัดสินใจ วางแผน หรือดำเนินการต่างๆ ที่มีส่วนเกี่ยวข้องกับสภาพอากาศได้อย่างมีประสิทธิภาพ
3. เพื่อให้ทราบว่าการเพิ่มคุณลักษณะทางสถิติให้กับชุดข้อมูลก่อนนำเข้าแบบจำลองนั้น ส่งผลให้ประสิทธิภาพในการพยากรณ์ของตัวแบบเพิ่มขึ้นได้จริง
4. เพื่อให้ทดลองว่าแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับแบบใดที่จะมีความเหมาะสมกับการพยากรณ์อากาศด้วยชุดข้อมูลที่มีคุณลักษณะจำกัด

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะแบ่งเนื้อหาออกเป็น 2 ส่วนหลัก ส่วนแรกเป็นส่วนของทฤษฎีที่เกี่ยวข้อง ซึ่งจะมีทั้งหมด 9 หัวข้อ โดย 2 หัวข้อแรกจะเป็นการกล่าวถึงศาสตร์และความรู้ที่เกี่ยวข้องกับการพยากรณ์อากาศโดยทั่วไป อีก 5 หัวข้อจะเป็นการกล่าวถึงความรู้พื้นฐานในเรื่องของระบบโครงข่ายประสาทเทียม ตัวแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับต่างๆ ที่ถูกนำมาใช้ในงานวิจัยนี้ 2 หัวข้อสุดท้ายจะกล่าวถึงการวัดประสิทธิภาพของตัวแบบ และการแก้ปัญหาชุดข้อมูลไม่สมดุล ส่วนที่สองจะเป็นส่วนของงานวิจัยที่เกี่ยวข้อง ซึ่งจะเป็นการศึกษาเกี่ยวกับการพยากรณ์ปริมาณน้ำฝนและการเพิ่มประสิทธิภาพในการพยากรณ์ให้กับแบบจำลองต่างๆ

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การพยากรณ์อากาศ

หมายถึง การคาดหมายสภาพลมฟ้าอากาศในอนาคต การที่จะพยากรณ์อากาศได้ต้องมีองค์ประกอบ 3 ประการ ประการแรกคือความรู้ความเข้าใจในปรากฏการณ์และกระบวนการต่าง ๆ ที่เกิดขึ้นในบรรยากาศ ประการที่ สองคือสภาวะอากาศปัจจุบัน และประการสุดท้ายคือความสามารถที่จะผสมผสานองค์ประกอบทั้งสองข้างต้น เข้าด้วยกันเพื่อคาดหมายการเปลี่ยนแปลงของบรรยากาศที่จะเกิดขึ้นในอนาคต

1. ระยะเวลาของการพยากรณ์อากาศ

การพยากรณ์อากาศอาจเป็นการคาดหมายสำหรับช่วงเวลาไม่กี่ชั่วโมงข้างหน้าจนถึงการคาดหมายสิ่งที่จะเกิดขึ้นในอีกหลายปีจากปัจจุบัน สามารถแบ่งชนิดของการพยากรณ์อากาศตามระยะเวลาที่คาดหมายได้ดังนี้

1) การพยากรณ์ปัจจุบัน (nowcast) หมายถึงการรายงานสภาวะอากาศที่เกิดขึ้นในปัจจุบันและการคาดหมายสภาพลมฟ้าอากาศสำหรับช่วงเวลาไม่เกิน 2 ชั่วโมง
2) การพยากรณ์ระยะสั้นมาก คือการพยากรณ์สำหรับช่วงเวลาไม่เกิน 12 ชั่วโมง
3) การพยากรณ์ระยะสั้น หมายถึง การพยากรณ์สำหรับระยะเวลาเกินกว่า 12 ชั่วโมงขึ้นไปจนถึง 3 วัน

4) การพยากรณ์อากาศระยะปานกลาง คือ การพยากรณ์สำหรับช่วงเวลาที่เกินกว่า 3 วันขึ้น ไปจนถึง 10 วัน
5) การพยากรณ์ระยะยาว คือการพยากรณ์สำหรับช่วงเวลาระหว่าง 10 ถึง 30 วัน โดยปกติมักเป็นการพยากรณ์ว่าค่าเฉลี่ยของตัวแปรทางอุตุนิยมวิทยาในช่วงเวลานั้นจะแตกต่างไปจากค่าเฉลี่ยทางภูมิอากาศอย่างไร
6) การพยากรณ์ระยะนาน คือการพยากรณ์ตั้งแต่ 30 วัน จนถึง 2 ปี ซึ่งยังแบ่งย่อยออกเป็น 3 ชนิด คือ
   - การคาดหมายรายเดือน คือการคาดหมายว่าค่าเฉลี่ยของตัวแปรทางอุตุนิยมวิทยาในช่วงนั้น จะเบี่ยงเบนไปจากค่าเฉลี่ยทางภูมิอากาศอย่างไร
   - การคาดหมายรายสามเดือน คือการคาดหมายค่าว่าเฉลี่ยของตัวแปรทางอุตุนิยมวิทยาในช่วงนั้น จะเบี่ยงเบนไปจากค่าเฉลี่ยทางภูมิอากาศอย่างไร
   - การคาดหมายรายฤดู คือการพยากรณ์ค่าเฉลี่ยของฤดูนั้นว่าจะแตกต่างไปจากค่าเฉลี่ยทางภูมิอากาศอย่างไร
7) การพยากรณ์ภูมิอากาศ คือการพยากรณ์สำหรับช่วงเวลามากกว่า 2 ปีขึ้นไป โดยแบ่งเป็น
   - การพยากรณ์การผันแปรของภูมิอากาศ คือการพยากรณ์ที่เกี่ยวข้องกับการผันแปรไปจากค่าปกติเป็นรายปีจนถึงหลายสิบปี
   - การพยากรณ์ภูมิอากาศคือการพยากรณ์สภาพภูมิอากาศในอนาคต โดยพิจารณาทั้งสาเหตุจากธรรมชาติและจากการกระทำของมนุษย์

2. ข้อมูลและเครื่องมือทางอุตุนิยมวิทยา

ข้อมูลอุตุนิยมวิทยา คือ ค่าทางภูมิอากาศต่างๆ ที่ทำการตรวจวัด รวบรวม อย่างต่อเนื่อง เพื่อนำมาใช้ในการพยากรณ์อากาศ เป็นปัจจัยที่มีความสัมพันธ์โดยตรงกับลักษณะอากาศที่เกิดขึ้นในแต่ละวัน ได้แก่ ความกดอากาศ ลม อุณหภูมิ ความชื้นสัมพัทธ์ เมฆ หยาดน้ำฟ้า รังสีดวงอาทิตย์ การระเหยของน้ำ และทัศนะวิสัย เป็นต้น นอกจากข้อมูลทางอุตุนิยมวิทยาแล้ว ระบบของการตรวจวัดสภาพอากาศก็มีความสำคัญ และจำเป็นอย่างยิ่งต่อการพยากรณ์อากาศให้มีประสิทธิภาพ ดังนั้นก่อนที่จะได้ผลการพยากรณ์อากาศ เราจำเป็นจะต้องมีสถานีตรวจวัดสภาพอากาศทั้งทางพื้นผิวโลก และอากาศชั้นบน เพื่อทำการตรวจวัดค่าทางภูมิอากาศต่างๆ สำหรับเครื่องมือที่ใช้ในระบบการพยากรณ์อากาศสามารถแบ่งออกได้เป็นประเภทใหญ่ๆ ได้ 3 ประเภท ดังนี้

1) เครื่องมือตรวจอากาศผิวพื้น ในสถานีตรวจอากาศผิวพื้นแต่ละสถานีจะมีสนามอุตุนิยมวิทยา ซึ่งเป็นที่สำหรับตรวจวัดอากาศผิวพื้น โดยเครื่องมือต่าง ๆ เหล่านี้จะตรวจวัดสภาพทางภูมิอากาศต่างๆ ตามช่วงเวลาที่กำหนดไว้ในแต่ละวัน
2) เครื่องมือตรวจอากาศชั้นบน เนื่องจากการเปลี่ยนแปลงของลักษณะอากาศบางอย่าง มีความเกี่ยวพันกับการเปลี่ยนแปลงของสภาวะชั้นบรรยากาศระดับชั้นบน ดังนั้น จึงมีความจำเป็นต้องตรวจวัดสภาพภูมิอากาศในบรรยากาศด้วย โดยเฉพาะในชั้นโทรโพสเฟียร์ ข้อมูลอุตุนิยมวิทยาที่ตรวจวัดในอากาศชั้นบนนี้ โดยมากจะเป็นข้อมูลหลักทางอุตุนิยมวิทยา เช่น อุณหภูมิ ความกดอากาศ ลมและความชื้นในระดับต่าง ๆ การตรวจวัดข้อมูลเหล่านี้ใช้เครื่องมือหลักที่เรียกว่า Radiosonde
3) เครื่องมือตรวจอากาศพิเศษ เป็นเครื่องมือที่ใช้สำหรับตรวจวัดปรากฏการณ์หรือลักษณะอากาศที่เกิดขึ้นเพื่อช่วยเสริมในการวิเคราะห์พยากรณ์อากาศ เครื่องมือตรวจอากาศพิเศษเหล่านี้มีหลายอย่าง อาทิ เรดาร์ตรวจอากาศ ดาวเทียมอุตุนิยมวิทยา และเครื่องมือสำหรับตรวจวัดความสูงของคลื่น แสดงภาพถ่ายดาวเทียมแบบต่าง ๆ VI, IR ภาพแสดงตรวจฝนด้วย Radar

3. ความผิดพลาดในการพยากรณ์อากาศ

แม้ว่าในปัจจุบันการพยากรณ์อากาศจะก้าวหน้าไปอย่างรวดเร็ว แต่การพยากรณ์อากาศ ให้ถูกต้องสมบูรณ์โดยไม่มีความผิดพลาดนั้น เป็นสิ่งที่ไม่อาจกระทำได้ สาเหตุสำคัญสามประการของความผิดพลาดในการพยากรณ์อากาศได้แก่ ประการแรก ความรู้ความเข้าใจเกี่ยวกับปรากฏการณ์ต่าง ๆ ทางอุตุนิยมวิทยายังไม่สมบูรณ์ ประการที่สอง บรรยากาศเป็นสิ่งที่ต่อเนื่องและมีการเปลี่ยนแปลงอยู่ตลอดเวลา แต่สถานีตรวจอากาศมีจำนวนน้อยและอยู่ห่างกันมาก รวมทั้งทำการตรวจเพียงบางเวลาเท่านั้น เช่น ทุก 3 ชั่วโมง ทำให้ไม่อาจทราบสภาวะที่แท้จริงของบรรยากาศได้เมื่อไม่ทราบสภาวะอากาศที่กำลังเกิดขึ้นอย่างสมบูรณ์ จึงเป็นไปไม่ได้ที่จะพยากรณ์อากาศให้มีรายละเอียดครบถ้วนถูกต้อง ประการสุดท้าย ธรรมชาติของกระบวนการที่เกิดขึ้นในบรรยากาศ มีความละเอียดอ่อนซับซ้อนอย่างยิ่ง ปรากฏการณ์ซึ่งมีขนาดเล็กหรือเกิดขึ้นในระยะสั้น ๆ และไม่อาจตรวจพบได้จาการตรวจอากาศ อาจทำให้เกิดการเปลี่ยนแปลงของสภาพลมฟ้าอากาศเป็นอย่างมากในระยะเวลาต่อมา ซึ่งจะทำให้ผลการพยากรณ์อากาศผิดพลาดไปได้อย่างมาก สาเหตุประการสุดท้ายนี้เป็นข้อจำกัดอย่างยิ่งในการพยากรณ์อากาศ เพราะเป็นเหตุให้การพยากรณ์อากาศจะมีความถูกต้องลดลงตามระยะเวลานั่นคือการพยากรณ์สำหรับช่วงเวลาที่สั้นจะมีความถูกต้องมากกว่าการพยากรณ์

สำหรับช่วงเวลาที่นานกว่า การพยากรณ์อากาศบริเวณเขตร้อนของโลกเช่นประเทศไทย จะยากกว่าการพยากรณ์ในเขตอบอุ่นและเขตหนาวเนื่องจากจากเหตุผลหลัก 3 ประการ

1) ความรู้ความเข้าใจเกี่ยวกับอุตุนิยมวิทยาเขตร้อนยังไม่ก้าวหน้าทัดเทียมกับอุตุนิยมวิทยาในเขตละติจูดสูงเพราะการศึกษาวิจัยเกี่ยวกับอุตุนิยมวิทยาในเขตร้อนมีน้อยกว่ามาก
2) สถานีตรวจอากาศในเขตร้อนมีจำนวนน้อยกว่าในเขตอบอุ่นและเขตหนาว ทำให้ผลการตรวจอากาศมีน้อยกว่า
3) ลมฟ้าอากาศในบริเวณละติจูดสูงส่วนมากเป็นระบบขนาดใหญ่ ซึ่งเกิดจากมวลอากาศที่แตกต่างกันมาพบกัน ทำให้ตรวจพบได้โดยง่าย เช่นฝนที่เกิดจากแนวปะทะอากาศมีความยาวมากกว่า 1,000 กิโลเมตร ในขณะที่ระบบลมฟ้าอากาศในเขตร้อนส่วนมากมีขนาดเล็กเพราะไม่ได้เกิดจากความแตกต่างของมวลอากาศ เช่นฝนที่ตกเป็นบริเวณแคบ ๆ

(กรมอุตุนิยมวิทยา โดย (Sukawat) [ระบบออนไลน์])

2.1.2 สภาพภูมิอากาศ และปริมาณน้ำฝนในประเทศไทย

1. สภาพภูมิอากาศในไทย

ประเทศไทยตั้งอยู่ในเขตร้อนทางทิศตะวันออกเฉียงใต้ของทวีปเอเชียระหว่างละติจูด 5 องศา 37 ลิปดา เหนือ กับ 20 องศา 27 ลิปดา เหนือ และระหว่างลองจิจูด 97 องศา 22 ลิปดา ตะวันออก กับ 105 องศา 37 ลิปดา ตะวันออก มีพื้นที่ทั้งหมดประมาณ 513,115 ตารางกิโลเมตร เนื่องจากประเทศไทยเป็นประเทศเล็ก ลักษณะภูมิประเทศ และลมฟ้าอากาศส่วนใหญ่จึงมีความคล้ายคลึงกัน อาจมีแตกต่างกันบ้างเพียงเล็กน้อย การแบ่งภาคของประเทศไทยในทางอุตุนิยมวิทยา จะพิจารณาจากรูปแบบภูมิอากาศ โดยแบ่งประเทศไทยออกได้เป็น 5 ภาค ได้แก่ ภาคเหนือ ภาคกลาง ภาคตะวันออกเฉียงเหนือ ภาคตะวันตก และภาคใต้

เนื่องจากประเทศไทยเป็นประเทศที่ตั้งอยู่ในเขตร้อนใกล้เส้นศูนย์สูตร ภูมิอากาศส่วนใหญ่ของประเทศจะมีลักษณะเป็นแบบร้อนชื้น หรือภูมิอากาศแบบทุ่งหญ้าสะวันนา ในขณะที่ภาคใต้และทางตะวันออกสุดของภาคตะวันออกจะเป็นเขตภูมิอากาศแบบมรสุมเขตร้อน ทั่วประเทศมีอุณหภูมิเฉลี่ยระหว่าง 19-38 องศาเซลเซียส อากาศจะร้อนที่สุดช่วงกลางเดือนเมษายน หลังจากนั้นภายใต้อิทธิพลของลมมรสุมตะวันตกเฉียงใต้และตะวันออกเฉียงเหนือทำให้ประเทศไทยเข้าสู่ฤดูฝน

และฤดูหนาวตามลำดับ ซึ่งภูมิอากาศในประเทศไทยนั้นสามารถแบ่งออกได้เป็น 3 ฤดู ได้แก่ ฤดูร้อน (ช่วงกลางเดือนกุมภาพันธ์ถึงกลางเดือนพฤษภาคม) ฤดูฝน (ช่วงกลางเดือนพฤษภาคมถึงกลางเดือนตุลาคม) และฤดูหนาว (ช่วงกลางเดือนตุลาคมถึงกลางเดือนกุมภาพันธ์)

2. ปริมาณน้ำฝน

โดยเฉลี่ยแล้ว ประเทศไทยมีปริมาณน้ำฝนสะสมทั่วประเทศรวมตลอดปีเฉลี่ยประมาณ 1,200-1,600 มิลลิเมตรต่อปี อีกทั้งนอกเหนือจากปัจจัยในเรื่องของฤดูกาลแล้ว ปริมาณฝนในแต่ละพื้นที่ยังผันแปรไปตามลักษณะของภูมิประเทศ ดังเช่น ประเทศไทยตอนบน ที่ภูมิอากาศส่วนใหญ่จะมีความแห้งแล้งและมีปริมาณฝนน้อยในฤดูหนาว เมื่อเข้าสู่ฤดูร้อนปริมาณฝนจะเริ่มเพิ่มขึ้นเนื่องจากอิทธิพลของพายุฝนฟ้าคะนอง และเมื่อเข้าสู่ฤดูฝน ปริมาณฝนจะเพิ่มขึ้นอย่างมาก โดยในช่วงเดือนที่มีปริมาณฝนมากที่สุดก็คือช่วงเดือนสิงหาคมถึงกันยายน พื้นที่ที่จะมีปริมาณฝนมากในประเทศไทย ส่วนใหญ่จะอยู่ทางด้านหน้าทิวเขา หรือด้านรับลมมรสุมตะวันตกเฉียงใต้ ได้แก่ พื้นที่ทางด้านตะวันตกและบริเวณภาคตะวันออกของประเทศ ส่วนพื้นที่ที่มีปริมาณฝนน้อย ส่วนใหญ่ก็จะอยู่ด้านหลังทิวเขา ซึ่งก็คือ พื้นที่แถบบริเวณตอนกลางของภาคเหนือและภาคกลาง รวมไปถึงพื้นที่บริเวณด้านตะวันตกของภาคตะวันออกเฉียงเหนือ สำหรับในส่วนของภาคใต้นั้นจะมีฝนตกเกือบตลอดปี ยกเว้นในช่วงฤดูร้อน โดยในช่วงฤดูฝนของภาคใต้นั้น ภาคใต้ฝั่งตะวันตกจะมีปริมาณน้ำฝนสะสมมากกว่าภาคใต้ฝั่งตะวันออก เนื่องจากเป็นด้านที่ต้องรับอิทธิพลจากลมมรสุมตะวันตกเฉียงใต้ แต่ในช่วงฤดูหนาว ภาคใต้ฝั่งตะวันออกจะมีปริมาณน้ำฝนสะสมที่มากกว่าภาคใต้ฝั่งตะวันตก เนื่องจากได้รับอิทธิพลจากลมมรสุมตะวันออกเฉียงเหนือ

1) เครื่องมือตรวจวัดปริมาณน้ำฝน

จากที่กล่าวไว้ในหัวข้อที่ 2.1.1 เกี่ยวกับเครื่องมือตรวจวัดสภาพภูมิอากาศ ว่าสามารถแบ่งออกได้เป็น 3 ประเภท เช่นเดียวกันกับเครื่องมือตรวจวัดปริมาณน้ำฝนก็สามารถแบ่งออกได้เป็น 3 ประเภทเช่นกัน ได้แก่ โทรมาตร (เครื่องมือตรวจวัดปริมาณน้ำฝนภาคพื้นดิน) เรดาร์ตรวจอากาศ และดาวเทียมอุตุนิยมวิทยา ซึ่งข้อมูลปริมาณน้ำฝนที่นำมาใช้ในงานวิจัยนี้ เป็นข้อมูลที่เก็บมาจากสถานีตรวจวัดภาคพื้นดิน หรือระบบโทรมาตร

โทรมาตร เป็น อุปกรณ์ที่สามารถตรวจวัดค่าทางฟิสิกส์ เคมี หรือ ชีวภาพ แล้วจึงทำการส่งค่าที่วัดได้ดังกล่าวไปเก็บยังระบบที่กำหนดไว้ด้วยเงื่อนไขต่างๆ ข้อมูลที่ตรวจวัดได้อาจจะเป็นข้อมูลระดับเสียง อุณหภูมิ ความชื้น ค่าความเป็นกรด ด่าง หรือ ปริมาณ

ออกซิเจนที่ละลายในน้ำ หรือ แม้กระทั่งภาพถ่าย หรือ ข้อมูลที่เกิดขึ้นจากตัวระบบโทรมาตรเอง เช่น สถานะการทำงาน เป็นต้น

2) เกณฑ์การพิจารณาปริมาณน้ำฝน

เกณฑ์การพิจารณาปริมาณน้ำฝนด้วยระบบโทรมาตร ในระยะเวลา 24 ชั่วโมงของแต่ละวันตั้งแต่เวลา 07.00 น. ของวันหนึ่งถึงเวลา 07.00 น.ของวันรุ่งขึ้นตามลักษณะของฝนที่ตกในประเทศที่อยู่ในเขตร้อนย่านมรสุมมีดังนี้

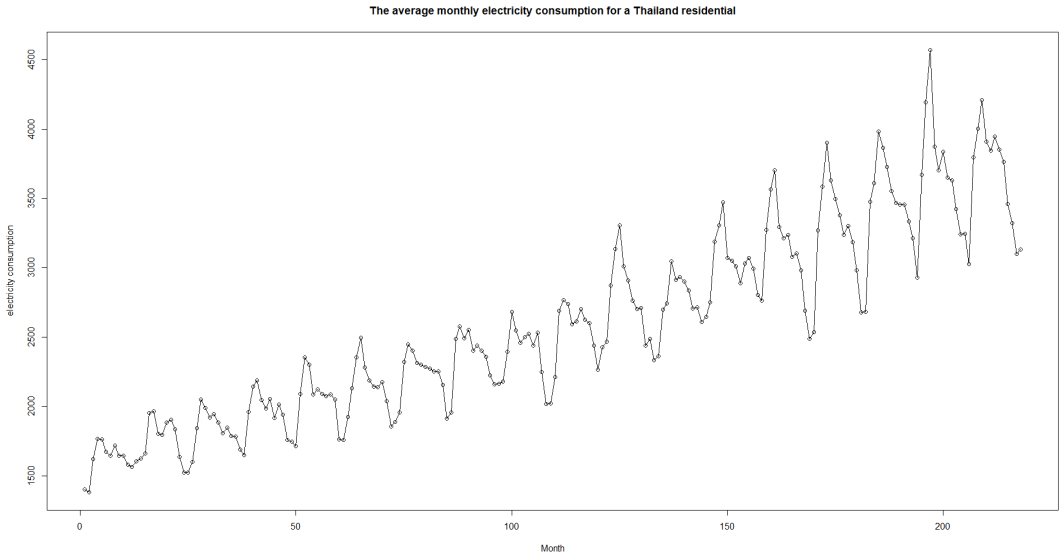
- ปริมาณน้ำฝนสะสมน้อยกว่า 0.1 มิลลิเมตร ไม่สามารถวัดปริมาณน้ำฝนได้
- ปริมาณน้ำฝนสะสมระหว่าง 0.1 - 10.0 มิลลิเมตร ฝนตกเล็กน้อย
- ปริมาณน้ำฝนสะสมระหว่าง 10.1 - 35.0 มิลลิเมตร ฝนตกปานกลาง
- ปริมาณน้ำฝนสะสมระหว่าง 35.1 - 90.0 มิลลิเมตร ฝนตกหนัก
- ปริมาณน้ำฝนสะสมตั้งแต่ 90.1 มิลลิเมตรขึ้นไป ฝนตกหนักมาก

2.1.3 แบบจำลองอนุกรมเวลา (Time series)

1. อนุกรมเวลา

ชุดของข้อมูลหรือข้อมูลเชิงปริมาณที่ทำการเก็บรวบรวมมามากกว่า 1 ช่วงเวลา โดยมีการจัดเก็บตามลำดับเวลาที่เกิดขึ้นอย่างต่อเนื่อง ข้อมูลอนุกรมเวลาอาจอยู่ในลักษณะที่เป็นข้อมูลรายปี รายไตรมาส หรือรายเดือนก็ได้ ทั้งนี้ขึ้นอยู่กับความเหมาะสมในการนำไปใช้ประโยชน์ ส่วนประกอบของอนุกรมเวลาประกอบไปด้วย แนวโน้ม (Trend) ความผันแปรตามฤดูกาล (Seasonal Variation) ความผันแปรตามวัฏจักร (Cycle Variation) และความผันแปรเนื่องจากเหตุการณ์ที่ผิดปกติ (Irregular Variation)

ตัวอย่างของกราฟข้อมูลอนุกรมเวลาที่ข้อมูลมีแนวโน้ม และการแปรผันตามช่วงเวลา สามารถแสดงได้ดังรูปที่ 1



รูปที่ 1 กราฟแสดงปริมาณการใช้พลังงานไฟฟ้าภาคครัวเรือนของไทย(รายเดือน) ตั้งแต่เดือนมกราคม ปี 2543 ถึงเดือนกุมภาพันธ์ ปี 2562

2. Autoregressive Integrated Moving Average (ARIMA)

แบบจำลอง ARIMA เป็นหนึ่งในแบบจำลองที่ใช้ในการวิเคราะห์ข้อมูลอนุกรมเวลาแบบตัวแปรเดี่ยว (Univariate variable) ซึ่งถูกต้องและมีความแม่นยำสูงโดยพิจารณาจากการวัดค่าความคลาดเคลื่อนของการพยากรณ์ ถูกพัฒนาขึ้นโดย George E.P.Box และ Gwilym M.Jenkins เป็นการสร้างตัวแบบจากข้อมูลที่เก็บรวบรวมในอดีตตามลำดับเวลาอย่างต่อเนื่องเพื่อกำหนดรูปแบบของข้อมูลและพยากรณ์ข้อมูลในอนาคต โดยข้อมูลอนุกรมเวลาที่จะนำมาวิเคราะห์ได้นั้นต้องมีความคงที่ (Stationary) และไม่มีแนวโน้ม (Trend) องค์ประกอบของแบบจำลอง ARIMA จะประกอบไปด้วยพารามิเตอร์สำคัญ 3 ตัว ได้แก่ p, d และ q ซึ่งมาจากส่วนสำคัญ 3 ส่วน ของแบบจำลอง ดังนี้

1) ตัวแบบ Autoregressive หรือ ตัวแบบ AR(p)

ตัวแบบ AR(p) เป็นตัวแบบที่ค่าเวลาปัจจุบันของอนุกรมเวลา x_t จะขึ้นอยู่กับค่าของอนุกรมเวลาก่อนหน้า หรือค่าอนุกรมเวลาในอดีต p ค่า คือ $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$ สามารถแสดงเป็นสมการทางคณิตศาสตร์ได้ดังนี้

$$x_t = \phi x_{t-1} + \phi x_{t-2} + \ldots + \phi x_{t-p} + w_t$$

หรือ $$x_t = \sum_{j=1}^{p} \phi x_{t-j} + w_t$$

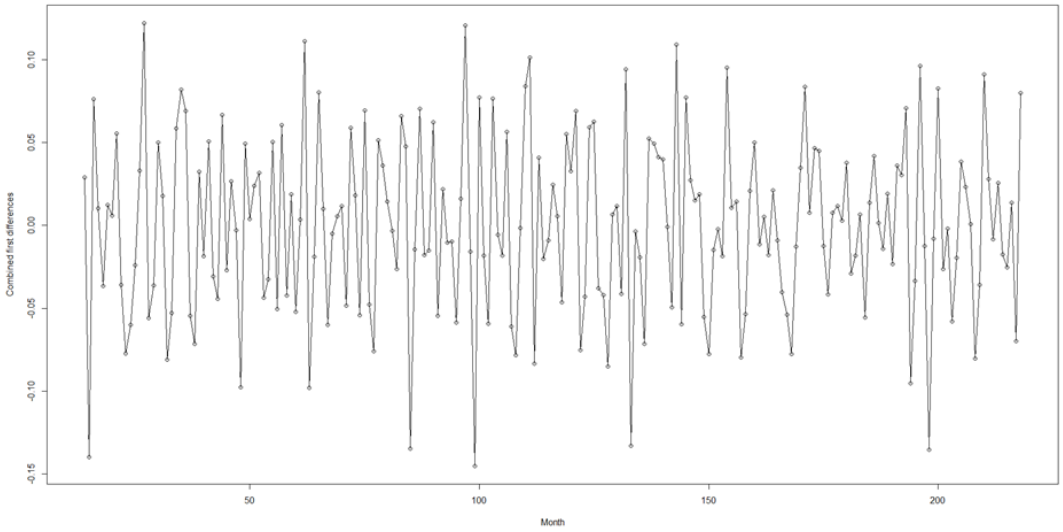
เมื่อ x_t เป็นอนุกรมเวลาแบบคงที่ (Stationary) ที่มีค่าเฉลี่ยเท่ากับ 0

$\phi_1, \phi_2, \ldots, \phi_p$ เป็นค่าคงที่ และ ไม่เท่ากับ 0 ($\phi_p \neq 0$)

w_t เป็น Gaussian white noise ที่มีค่าเฉลี่ยเท่ากับ 0 มีค่าความแปรปรวนเท่ากับ σ^2 และเป็นอิสระกับ x_t (ค่าสหสัมพันธ์ระหว่าง x_t และ w_t เท่ากับ 0)

2) Integrated หรือ I

เป็นการหาผลต่าง (Differencing) อันดับที่ d ของอนุกรมเวลา x_t (The difference of order d: $\nabla^d x_t$) เพื่อแปลงข้อมูลอนุกรมเวลาที่ไม่คงที่ (Nonstationary) ให้เป็นอนุกรมเวลาคงที่ (Stationary) เนื่องจากข้อมูลที่จะถูกนำมาใช้ในแบบจำลอง ARIMA จะต้องมีค่าเฉลี่ยและความแปรปรวนคงที่ ดังนั้น ข้อมูลจึงไม่ควรมีแนวโน้ม (Trend) และไม่ควรแปรผันตามช่วงเวลา (Seasonal)



รูปที่ 2 แบบจำลองอนุกรมเวลาที่มีความนิ่ง

3) ตัวแบบ Moving Average หรือตัวแบบ MA(q)

ตัวแบบ MA(q) เป็นตัวแบบที่ค่าเวลาปัจจุบันของอนุกรมเวลา x_t จะขึ้นอยู่กับปัจจัยอื่น w_t ตั้งแต่อดีตจนถึงปัจจุบันจำนวน q ค่า คือ $w_t, w_{t-2}, \ldots, w_{t-q}$ สามารถแสดงเป็นสมการทางคณิตศาสตร์ได้ดังนี้

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ \ldots \ + \theta_q w_{t-q}$$

หรือ $$x_t = w_t + \sum_{j=1}^{q} \theta_j w_{t-j}$$

เมื่อ x_t เป็นอนุกรมเวลาแบบคงที่ (Stationary) ที่มีค่าเฉลี่ยเท่ากับ 0

$\theta_1, \theta_2, \ \ldots \ , \theta_p$ เป็นค่าคงที่ และ ไม่เท่ากับ 0 ($\theta_p \neq 0$)

w_t เป็น Gaussian white noise ที่มีค่าเฉลี่ยเท่ากับ 0 มีค่าความแปรปรวนเท่ากับ σ^2 และเป็นอิสระกับ x_t (ค่าสหสัมพันธ์ระหว่าง x_t และ w_t เท่ากับ 0)

3. Autoregressive Integrated Moving Average with Exogenous model (ARIMAX)

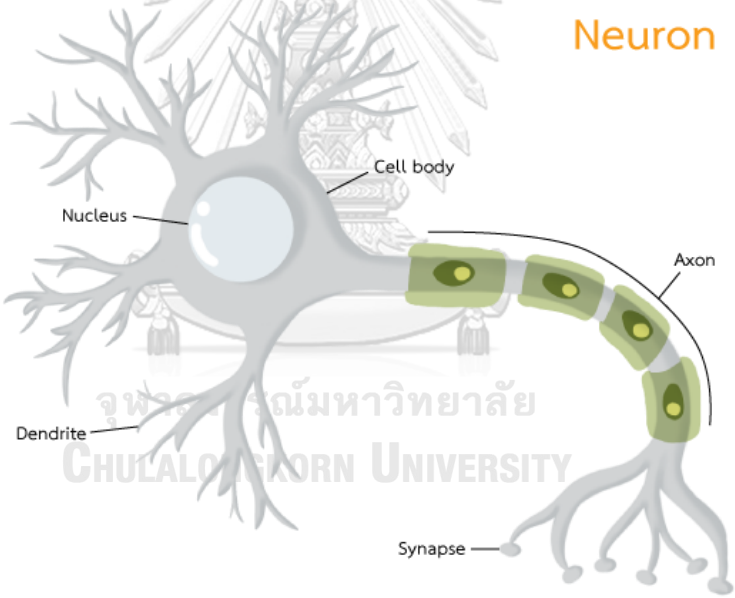
แบบจำลอง ARIMAX เป็นแบบจำลองที่ถูกพัฒนาต่อดยอดมาจากตัวแบบ ARIMA ซึ่งเป็นแบบจำลองที่ใช้ในการวิเคราะห์ข้อมูลอนุกรมเวลาแบบตัวแปรเดี่ยว (Univariate variable) เพื่อให้สามารถเพิ่มตัวแปรอิสระอื่นๆ เข้าไปในตัวแบบ หรือเป็นการทำให้ตัวแบบสามารถวิเคราะห์ข้อมูลอนุกรมเวลากับตัวแปรหลายตัวได้ (Multivariate variables) เพื่อเพิ่มประสิทธิภาพในการพยากรณ์ของตัวแบบให้ดียิ่งขึ้น โดยในส่วนของสมการนั้นจะมีความคล้ายคลึงกับตัวแบบจำลอง ARIMA แต่จะมีการเพิ่มในส่วนของชุดตัวแปรอิสระอื่นๆที่จะเข้ามาในสมการในรูปแบบของเวกเตอร์ (สำหรับ 1 ตัวแปร) หรือเมทริกซ์ (สำหรับหลายตัวแปร)

4. การพยากรณ์อนุกรมเวลาแบบบอกซ์แอนด์เจนกินส์

วิธีบอกซ์แอนด์เจนกินส์เป็นวิธีหนึ่งในการสร้างตัวแบบสำหรับข้อมูลอนุกรมเวลาซึ่งเป็นหนึ่งในวิธีที่ถูกต้องและมีความแม่นยำสูงโดยพิจารณาจากการวัดค่าความคลาดเคลื่อนของการพยากรณ์ ซึ่งสามารถพิจารณาจากค่าความคลาดเคลื่อนน้อยที่สุดของเทคนิคที่ใช้ในการพยากรณ์ โดยงานวิจัยนี้ จะทำการพิจารณาเบื้องต้นว่าอนุกรมเวลานั้นมีลักษณะแบบใด โดยทำการทดสอบองค์ประกอบแนวโน้มและความแปรผันตามฤดูกาลโดยพิจารณาจากกราฟอัตตสหสัมพันธ์ (Auto Correlation Function: ACF) และกราฟอัตต-สหสัมพันธ์บางส่วน (Partial Auto Correlation Function: PACF) ซึ่งเป็นเครื่องมือที่ใช้ในการพิจารณาลำดับหรือจำนวนเทอมของตัวพารามิเตอร์ที่จะพิจารณาย้อนหลัง เพื่อที่จะใช้สร้างตัวแบบอนุกรมเวลาสำหรับอนุกรมเวลาที่มีคุณสมบัติคงที่ (Stationary)

2.1.4 โครงข่ายระบบประสาทเทียม (Artificial Neural Network)

โครงข่ายระบบประสาทเทียม (Artificial Neural Network) เป็นตัวแบบทางคณิตศาสตร์สำหรับคอมพิวเตอร์ในการประมวลผลข้อมูลสารสนเทศด้วยการคำนวณแบบ connectionist โดยมีหลักการเบื้องต้นมาจากการจำลองการทำงานของระบบเซลล์ประสาท (Neuron) ของมนุษย์ โดยที่แต่ละเซลล์ประสาทจะมีหลักการในการสื่อสารกันผ่านการกระตุ้นด้วยศักย์ไฟฟ้า กล่าวคือ กระแสประสาทจะเริ่มต้นด้วยการส่งสัญญาณจากตัวเซลล์ (Cell body) ผ่านแกนประสาทนำออก (Axon) ไปยังจุดประสานประสาท (Synapse) ซึ่งจะเชื่อมต่อกับเซลล์ประสาทอื่นผ่านใยประสาทนำเข้า (Dendrite) การเชื่อมต่อกันของเซลล์ประสาทเป็นสิ่งสำคัญของการทำงานของโครงข่ายประสาท ขนาดและความแข็งแรงของจุดประสานประสาท (Synapse) นั้น จะเปลี่ยนแปลงไปตามประสบการณ์การเรียนรู้ต่างๆ ที่ผ่านมา ณ จุดที่เซลล์ประสาททำการสื่อสารกันนี้ เป็นหลักการสำคัญที่ถูกนำไปประยุกต์ใช้ในโครงข่ายประสาทเทียม (Sanguansat, 2019)



รูปที่ 3 เซลล์ประสาท (Neuron)

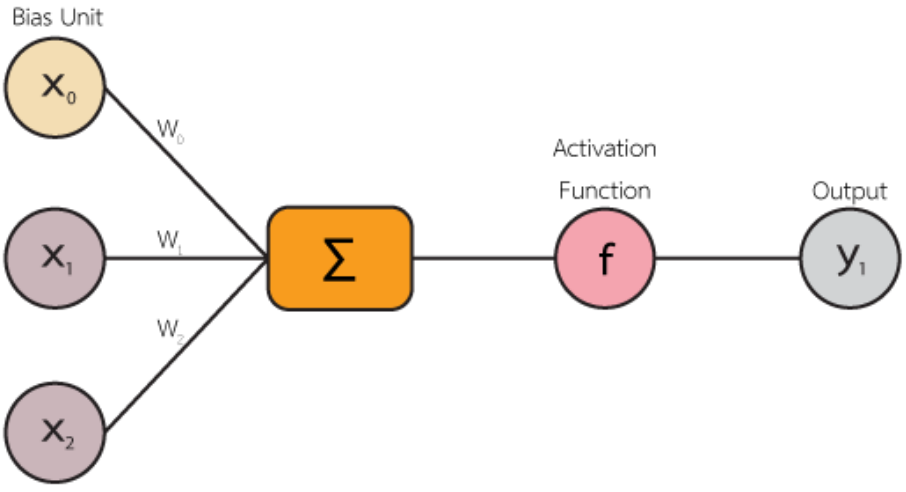
สำหรับโครงข่ายระบบประสาทเทียมนั้น ได้มีการเริ่มต้นขึ้นในปี ค.ศ. 1943 โดย Warren S. McCulloch กับ Walter Pitts ที่ได้เสนอแบบจำลองเซลล์ประสาทในลักษณะของแบบจำลองคณิตศาสตร์ เพื่ออธิบายการทำงานอันซับซ้อนของสมองมนุษย์ และได้มีการพัฒนาต่อยอดมาเรื่อยๆ เพื่อสร้างการเรียนรู้ให้กับคอมพิวเตอร์ทั้งแบบที่มีผู้สอน (Supervised Learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ก่อนที่ในปี ค.ศ. 1958 Frank Rosenblatt จะเสนอวิธี เพอร์เซปตรอน (Perceptron) ขึ้น แต่ตัวของเพอร์เซปตรอนนั้นยังมีข้อจำกัดในการแก้ปัญหาบางประเภท

เช่น ปัญหา XOR (Excusive OR) ซึ่งต่อมา โครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation) ได้ถูกพัฒนาขึ้นเพื่อแก้ปัญหาที่เพอร์เซปตรอนไม่สามารถแก้ได้ เช่น ปัญหา XOR โดยอาศัยโครงข่ายแบบหลายชั้น (Multi-Layer) และวิธีการคำนวณค่าความผิดพลาดระหว่างชั้นดังกล่าว อีกทั้งหลักการนี้ยังเป็นหลักการสำคัญที่ถูกนำไปใช้ในการพัฒนาต่อยอดในการพัฒนาแบบจำลองการเรียนรู้เชิงลึก (Deep Learning) อีกด้วย

ลักษณะการคำนวณในโครงข่ายระบบประสาทเทียมจะมี Node (หรือ Perceptron) ทำหน้าที่คล้าย Neuron ในการรวบรวมข้อมูลนำเข้าแบบขนานให้กับหน่วยประมวลผลย่อยๆ (Layer) และทำงานเป็นขั้นๆ จนได้คำตอบ การเชื่อมต่อนี้เป็นส่วนสำคัญที่สามารถหาความสัมพันธ์หรือรูปแบบที่ซับซ้อนได้ ยิ่งตัวโครงข่ายระบบประสาทเทียมมี layer มากเท่าไหร่ ก็จะยิ่งสามารถทำงานที่มีซับซ้อนได้มากขึ้น

1. เพอร์เซปตรอน (Perceptron)

ขั้นตอนวิธีเพอร์เซปตรอน (Perceptron algorithm) เป็นขั้นตอนวิธีที่ทำให้เครื่องคอมพิวเตอร์สามารถทำการเรียนรู้จากตัวอย่างที่นำเข้าไปได้ โครงสร้างของเพอร์เซปตรอนจะมีลักษณะคล้ายแบบจำลองเซลล์ประสาท คือ จะมี Node ที่ทำหน้าที่คล้าย Neuron ในการรวบรวมข้อมูลนำเข้าแต่ละตัว (ที่ผ่านการถ่วงน้ำหนัก) ผ่านการเชื่อมต่อด้วย Synapse เพื่อนำไปคำนวณหาผลลัพธ์เป็นข้อมูลส่งออก (output) ซึ่งข้อมูลที่ส่งออกไปนั้นก็จะแตกต่างกันไปตามฟังก์ชันกระตุ้น (Activation Function) ที่เลือกใช้ ลักษณะโครงสร้างของเพอร์เซปตรอนสามารถแสดงได้ดังรูปต่อไปนี้



รูปที่ 4 โครงสร้างของเพอร์เซปตรอน

จากรูปที่ 4 ซึ่งแสดงถึงโครงสร้างของเพอร์เซปตรอน เราจะเห็นว่า โครงสร้างคร่าวๆ ของเพอร์เซปตรอนจะแบ่งออกเป็นส่วนสำคัญ 3 ส่วนคือ 1) Input layer 2) Hidden layer และ 3) Output layer ในส่วนของสัญลักษณ์ต่างๆ ในภาพจะสามารถแจกแจง ได้ดังนี้

1) Bias Unit (x_0) เป็นตัวปรับค่า จะมีค่าเท่ากับ 1 เสมอ (และจะมีการใส่ค่า Bias นี้เอาไว้ในทุกๆ layer ยกเว้นใน output layer) ถ้ามองในรูปแบบของสมการเส้นตรง ตัว x_0 นี้จะทำหน้าที่เป็นจุดตัดแกน y ซึ่งสามารถทำให้สมการเส้นตรง สามารถเลื่อนไปตำแหน่งไหนก็ได้ โดยที่ความชันเท่าเดิม ซึ่งก็จะทำให้มีชุดพารามิเตอร์ที่หลากหลายมากขึ้น และสามารถนำไปทำการทดสอบหาชุดพารามิเตอร์ที่เหมาะสมกับชุดข้อมูลได้
2) ค่าถ่วงน้ำหนัก (Weight) คือ ค่าน้ำหนักที่มอบให้กับข้อมูลแต่ละตัว โดยน้ำหนักของแต่ละข้อมูลจะสะท้อนถึงความสำคัญของข้อมูลนั้นๆ ที่นำมาคำนวณในตัวแบบ
3) ฟังก์ชันกระตุ้น (Activation Function) ทำหน้าที่รวมค่าเชิงตัวเลขจากข้อมูลนำเข้า (Input) ที่เข้ามาใน Node แล้วทำการตัดสินใจว่าจะส่งสัญญาณเอาต์พุตออกไปในรูปใด ค่าของเอาต์พุตที่ออกไปจากโหนดจะเรียกว่า Activation โดยที่ฟังก์ชันกระตุ้นอาจเป็นฟังก์ชันเส้นตรงหรือไม่ก็ได้ ฟังก์ชันกระตุ้นมีหลายรูปแบบ ขึ้นอยู่กับลักษณะงาน โดยฟังก์ชันกระตุ้นที่นำมาใช้ในงานวิจัยจะมีดังต่อไปนี้
    - ฟังก์ชันซิกมอยด์ (Sigmoid Function) จะมีความเหมาะสมกับ output layer ที่แก้ปัญหาเกี่ยวกับการแบ่งกลุ่ม (classification) เนื่องจากเอาต์พุตของ Sigmoid Function จะมีค่าระหว่าง 0 ถึง 1 แต่จะไม่เหมาะกับ Hidden layer เพราะเอาต์พุตจาก Node จะถูกบีบและส่งต่อ ทำให้ความถูกต้องในการดึงไปใช้งานของ Node ใน layer ถัดไปมีน้อยลง สมการของฟังก์ชันซิกมอยสามารถแสดงได้ ดังนี้

$$\sigma(x) = \frac{1}{1 - e^{-x}}$$

    - Hyperbolic Tangent Activation Function (Tanh) มีเหมาะสมกับ output layer ที่ต้องการใช้เพื่อแก้ปัญหาการแบ่งกลุ่ม (classification) มากกว่า sigmoid เนื่องจากเอาต์พุตของ Tanh Function จะมีค่าระหว่าง -1 ถึง 1 ทำให้เอาต์พุตที่ออกมามีโอกาสเข้าใกล้ 0 ทำให้ผลลัพธ์ที่ได้จากการคำนวณมีค่าไม่มาก และ

เช่นเดียวกันกับฟังก์ชันซิกมอยด์ คือข้อมูลจะถูกบีบ จึงไม่ควรนำไปใช้ใน Hidden layer สมการของ Tanh Function สามารถแสดงได้ ดังนี้

$$f(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

- Rectified Linear Unit (Relu) เป็นฟังก์ชันกระตุ้นที่เหมาะที่จะใช้ใน Hidden layer เนื่องจากสามารถให้เอาต์พุตที่อยู่ในรูปแบบที่ไม่ใช่เส้นตรง (non-linear) ได้ เอาต์พุตที่ออกมาจะมีค่ามากกว่าหรือเท่ากับ 0 ทำให้เอาต์พุตไม่ถูกบีบ สมการของ Relu สามารถแสดงได้ ดังนี้

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

2. ฟังก์ชันต้นทุน (Cost Function / Loss Function)

เป็นสมการที่เอาไว้ใช้ในการวัดความถูกต้องของตัวแบบจำลองที่เราพัฒนาขึ้น โดยมีชุดพารามิเตอร์ต่างๆ เป็นข้อมูลนำเข้า ซึ่งผลลัพธ์ที่ได้จากจากฟังก์ชันดังกล่าวคือต้นทุนที่แสดงถึงค่าความผิดพลาดของตัวแบบ ซึ่งหมายความว่า ชุดพารามิเตอร์ที่ให้ค่าต้นทุนที่ต่ำที่สุด จะเป็นชุดพารามิเตอร์ที่ดีที่สุดที่ควรเลือกใช้ ทั้งนี้รูปแบบของตัวสมการดังกล่าวจะขึ้นอยู่กับว่า ปัญหาที่ต้องการแก้นั้น เป็นปัญหาแบบใด โดยที่สมการของฟังก์ชันต้นทุนนั้นจะแบ่งออกเป็น 2 ประเภทตามลักษณะของปัญหา ซึ่งสามารถแสดงได้ดังนี้

Regression Problem

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Classification Problem

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})))$$

2.1.5 โครงข่ายระบบประสาทแบบย้อนกลับ (Recurrent Neural Networks: RNN)

โครงข่ายระบบประสาทแบบย้อนกลับ (RNN) คือ โครงข่ายประสาทเทียม (Artificial Neuron Network : ANN) รูปแบบหนึ่ง ที่ออกแบบมาเพื่อแก้ปัญหาเกี่ยวกับข้อมูลประเภทอนุกรมเวลา (Time Series data) หลักการของ RNN คือการปรับรูปแบบของโครงข่ายประสาทเทียมแบบเดิม เพื่อให้มีชั้นซ่อน (Hidden state) ในการจดจำความรู้หรือข้อมูลก่อนหน้า มารวมเข้ากับข้อมูลตัวใหม่ที่เข้ามา (Input data) เพื่อใช้ในการทำนายหรือพยากรณ์ ซึ่งสามารถแสดงเป็นสมการได้ดังนี้

$$h_t = f_h(W_{HH}h_{t-1} + W_{IH}x_t)$$
$$y_t = f_y(W_{OH}h_t)$$

โดยที่

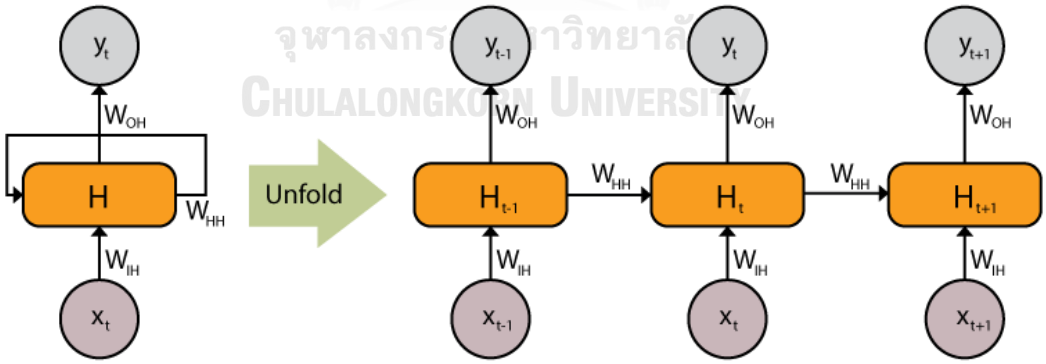
h_t คือ Hidden state ณ เวลา t

y_t คือ Output vector ณ เวลาที่ t

x_t คือ Input vector ณ เวลา t

f_h และ f_y คือ ฟังก์ชันกระตุ้น

W_{HH} , W_{IH} และ W_{OH} คือ เมทริกซ์ค่าถ่วงน้ำหนัก



รูปที่ 5 โครงสร้างของโครงข่ายระบบประสาทแบบย้อนกลับ (Recurrent Neural Networks: RNN)

อย่างไรก็ตาม RNN ก็มีปัญหาเกี่ยวกับการลดลงของเกรเดียนต์ (Vanishing Gradient or Exploding Gradient) ซึ่งปกติแล้ว ในโครงข่ายประสาทเทียมจะทำการปรับเปลี่ยนค่าถ่วงน้ำหนัก (update weights) โดยอาศัยวิธีการแพร่กลับของความผิดพลาด(Backward Propagation of

Errors: Backpropagation) ซึ่งจะทำการคำนวณกราเดียนต์ของ Loss Function เพื่อมาใช้ในการปรับเปลี่ยนค่าถ่วงน้ำหนัก แต่สำหรับ RNN ที่ เอาต์พุตไม่ได้มาจากแค่ช่วงเวลาเพียงช่วงเวลาเดียว (t=t) แต่มาจากหลายๆช่วงเวลา (t=t-1, t-2, ..., t) นั้น ทำให้เกิดการคูณกันของอนุพันธ์หลายๆ ตัว ซึ่งหากกราเดียนต์มีค่าน้อย (น้อยกว่า 1) การคูณกันแบบนี้ก็จะส่งผลให้ค่ากราเดียนต์นั้นลดลงไปเรื่อยๆ ตามความยาวของลำดับข้อมูลที่มี หรืออีกนัยก็คือ ตัวของ RNN นั้น จะมีปัญหากับข้อมูลที่มีจำนวนลำดับมากเกินไป

2.1.6 แบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM)

แบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) เป็นรูปแบบหนึ่งของโครงข่ายระบบประสาทแบบย้อนกลับ (RNN) ถูกพัฒนาขึ้นโดย (Hochreiter & Schmidhuber, 1997) เพื่อแก้ปัญหาการลดลงของเกรเดียนต์ (Vanishing Gradient) เมื่อลำดับของข้อมูลที่รับเข้ามามีจำนวนมากเกินไป โดยการทำงานของ LSTM นั้นจะมีความคล้ายคลึงกับ RNN แต่ในส่วนของชั้นซ่อน (Hidden state) ที่ใช้สำหรับจดจำลำดับของข้อมูลก่อนหน้านั้น LSTM จะมีการเรียนรู้ว่า เมื่อใดควรลืม (Forget) เขียน (Write) หรืออนุญาตให้อ่าน (Read) ซึ่งทำให้สามารถรองรับข้อมูลที่มีลำดับปริมาณมากที่เข้ามาได้ โดยส่วนประกอบต่างๆ ของแบบจำลองหน่วยความจำระยะสั้นแบบยาว มีดังนี้

1) สถานะเซลล์ (Cell state) ทำหน้าที่เป็นตัวเก็บสถานะของเซลล์ความจำ (Memory cell)
2) ประตู (Gate) ทำหน้าที่เป็นตัวควบคุมการไหลของข้อมูลที่เข้ามาในแบบจำลอง ซึ่งประตูแต่ละบานจะมีฟังก์ชันกระตุ้นเป็นของตัวเองเพื่อตัดสินใจว่าควรจะจัดการกับข้อมูลที่จะเข้ามาอย่างไร โดยประตูควบคุมการทำงานนั้น จะประกอบไปด้วย
   - ประตูลืม (Forget gate) ทำหน้าที่เป็นประตูที่จะตัดสินใจว่าควรจะเก็บสถานะเซลล์เดิมเอาไว้ หรือจะลบสถานะเซลล์เดิมออกไป เป็นเสมือนการเตรียมพื้นที่เพื่อรับข้อมูลใหม่ โดยฟังก์ชันกระตุ้นที่ควบคุมการทำงานประตูนี้อยู่คือ ฟังก์ชันซิกมอยด์ (Sigmoid Function: σ) ผลลัพธ์ที่จะได้ออกมาจะมีค่าอยู่ระหว่างค่า 0 ถึง 1 ถ้าผลลัพธ์ที่ได้มีค่าเข้าใกล้ 0 หมายถึง การลบสถานะเซลล์เดิมออกไป แต่ถ้าผลลัพธ์ทีได้มีค่าเข้าใกล้ 1 หมายถึง การเก็บสถานะเซลล์เดิมเอาไว้ สามารถเขียนสมการของประตูลืม (f_t) ได้ดังนี้

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)$$

- ประตูเขียน (Write gate) ในส่วนของประตูเขียนนั้น จะมีการทำงาน 2 ส่วน โดยในส่วนแรก คือการตัดสินใจว่าจะปรับสถานะเซลล์เมื่อมีข้อมูลใหม่เข้ามาหรือไม่ โดยส่วนนี้จะถูกควบคุมด้วยประตูนำเข้า (Input gate) ซึ่งมีฟังก์ชันซิกมอยด์ทำหน้าที่เป็นฟังก์ชันกระตุ้น เพื่อทำการตัดสินใจว่า จะให้มีการปรับเปลี่ยนสถานะเซลล์หรือไม่ ในการคำนวณส่วนนี้จะใช้ข้อมูลนำเข้าปัจจุบันที่เข้ามาพร้อมกับชั้นซ่อน (Hidden state) ก่อนหน้า ถ้าผลลัพธ์ที่ได้มีค่าเข้าใกล้ 0 ก็จะไม่มีการปรับเปลี่ยนสถานะของเซลล์ใดๆ แต่ถ้าผลลัพธ์ที่ได้มีค่าเข้าใกล้ 1 ก็จะทำการปรับเปลี่ยนสถานะของเซลล์ใหม่ให้เป็นสถานะปัจจุบัน โดยสำหรับสมการการคำนวณของประตูนำเข้า (i_t) สามารถแสดงได้ดังนี้

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

หลังจากที่ทำการตัดสินใจไปแล้วในส่วนแรกว่าจะทำการปรับเปลี่ยนค่าสถานะเซลล์ใหม่ให้เป็นปัจจุบัน ก็จะนำไปสู่การทำงานในส่วนที่สองของประตูเขียน นั่นคือการหาค่าสถานะเซลล์ใหม่เพื่อทำการปรับเปลี่ยน โดยในส่วนนี้จะถูกควบคุมด้วยประตูปรับค่านำเข้า (Input moderation gate) ซึ่งมีฟังก์ชันไฮเพอร์โบลิกแทนเจนต์ (Hyperbolic Tangent: tanh) หรือ ฟังก์ชันแทน ทำหน้าที่เป็นฟังก์ชันกระตุ้น โดยผลลัพธ์ที่ได้ออกมานั้น จะเป็นเหมือนกับผู้ท้าชิงในสนาม (Cell state candidate) ก่อนที่จะถูกนำไปปรับเปลี่ยนอีกครั้งตามสถานะที่ถูกคำนวณมาก่อนหน้าเพื่อส่งเป็นข้อมูลขาออกต่อไป โดยสำหรับสมการการคำนวณของประตูปรับค่านำเข้า ($\tilde{C}_t$) สามารถแสดงได้ดังนี้

$$\tilde{C}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

จะเห็นว่าในตอนนี้มีข้อมูลจากทั้ง ประตูลืม ประตูนำเข้า และประตูปรับค่านำเข้า ซึ่งทั้งหมดนี้ก็เพียงพอสำหรับการปรับค่าสถานะเซลล์ (Update cell state) แล้ว โดยสำหรับสมการการคำนวณของสถานะเซลล์ (C_t) สามารถแสดงได้ดังนี้

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

เมื่อพิจารณาจากฝั่งขวามือของสมการจะพบว่า ถ้าประตูลืม (f_t) มีค่าเป็น 0 ค่าสถานะเซลล์ก่อนหน้า (C_{t-1}) ก็จะไม่ถูกนำมาพิจารณาประกอบการปรับเปลี่ยนค่าสถานะเซลล์ปัจจุบัน แต่ถ้าประตูลืมมีค่าเป็น 1 ค่าสถานะเซลล์ก่อนหน้า (C_{t-1}) ก็จะถูกนำมา

พิจารณาประกอบการปรับเปลี่ยนค่าสถานะเซลล์ปัจจุบัน และเมื่อพิจารณาฝั่งซ้ายมือของสมการจะพบว่า ถ้าประตูนำเข้า (i_t) มีค่าเป็น 0 แสดงว่าค่าที่ได้จากการคำนวณในประตูปรับค่านำเข้า ($\widetilde{C_t}$) จะไม่ถูกนำมาใช้ประกอบการพิจารณาในการปรับเปลี่ยนค่าสถานะเซลล์ปัจจุบัน แต่ถ้าประตูนำเข้า (i_t) มีค่าเป็น 1 แสดงว่าค่าที่ได้จากการคำนวณในประตูปรับค่านำเข้า ($\widetilde{C_t}$) จะถูกนำมาใช้ประกอบการพิจารณาในการปรับเปลี่ยนค่าสถานะเซลล์ปัจจุบัน และจากค่าทั้งหมดที่ได้มาก็จะทำให้ได้ผลลัพธ์ออกมาเป็นค่าสถานะเซลล์ในปัจจุบัน (C_t)

- ประตูนำออก (Output gate) ทำหน้าที่เป็นประตูที่จะตัดสินใจว่าจะทำการส่งค่าสถานะชั้นซ่อน (Hidden state: h_t) ออกไปหรือไม่ โดยฟังก์ชันกระตุ้นที่ควบคุมการทำงานประตูนี้อยู่คือ ฟังก์ชันซิกมอยด์ (Sigmoid Function: σ) ในการคำนวณส่วนนี้จะใช้ข้อมูลนำเข้าปัจจุบันที่เข้ามาพร้อมกับชั้นซ่อน (Hidden state) ก่อนหน้า สำหรับสมการการคำนวณของประตูนำออก (o_t) สามารถแสดงได้ดังนี้

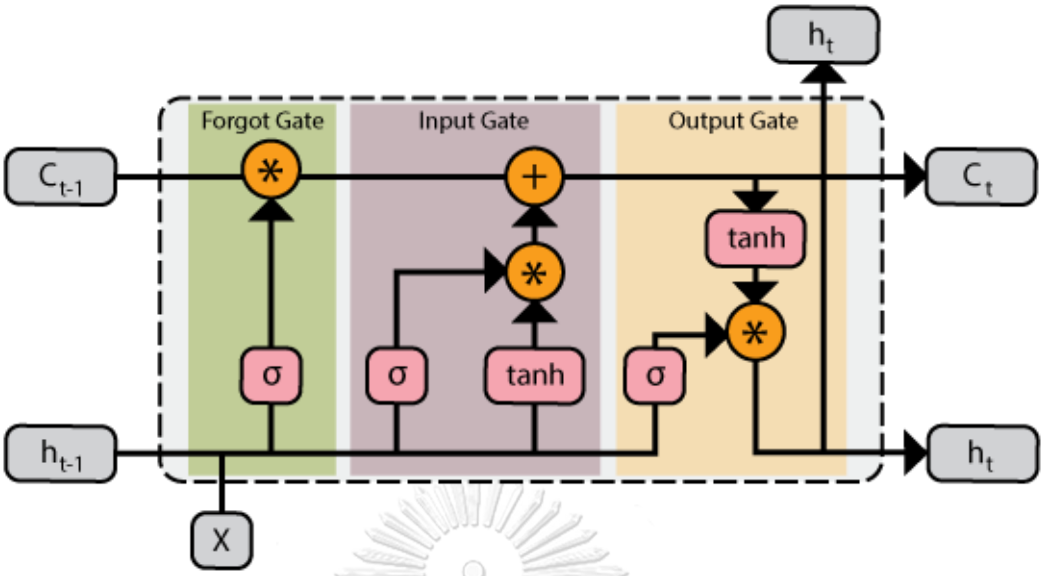
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_0)$$

เนื่องจากผลลัพธ์ที่จะต้องผลิตออกไปคือค่าสถานะชั้นซ่อนที่เวลาปัจจุบัน(h_t) ดังนั้นสมการสำหรับคำนวณค่าดังกล่าว จึงแสดงได้ดังนี้

$$h_t = o_t \odot \tanh(C_t)$$

จากสมการดังกล่าวพบว่า ถ้าประตูนำออก (o_t) มีค่าเป็น 0 ก็จะไม่มีการส่งค่าใดๆ ออกไป แต่ถ้าประตูนำออกมีค่าเป็น 1 ก็จะคำนวณค่า h_t แล้วส่งออกไป โดยที่ค่า h_t ก็คือผลลัพธ์สุดท้ายที่ได้จากสถานะเซลล์และสถานะชั้นซ่อน เพื่อใช้กับหน่วยความจำระยะสั้นแบบยาวในลำดับถัดไป ซึ่งในส่วนประตูนำออกนี้ก็จะเป็นเสมือนประตูอนุญาตให้อ่าน (Read gate) ที่ได้กล่าวถึงไปข้างต้น ที่ทำหน้าที่ตัดสินใจว่าจะอนุญาตให้นำข้อมูลผลลัพธ์สุดท้ายที่ได้ออกไปหรือไม่

สำหรับโครงสร้างภายในของแบบจำลองหน่วยความจำระยะสั้นแบบยาว สามารถแสดงได้ดังรูปที่ 6



รูปที่ 6 แบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM)

2.1.7 โครงข่ายประตูวกกลับ (Gated Recurrent Unit: GRU)

เป็นรูปแบบหนึ่งของโครงข่ายระบบประสาทแบบย้อนกลับ (RNN) ถูกพัฒนาขึ้นโดย (Chung et al., 2014) โดยมีจุดประสงค์เพื่อแก้ปัญหาการลดลงของเกรเดียนต์ (Vanishing Gradient) หรือก็เพื่อแก้ปัญหาการมีหน่วยความจำระยะสั้นของ RNN เช่นเดียวกับ LSTM ซึ่งการทำงานภายในของ GRU จะมีความคล้ายคลึงกับ LSTM เช่นกัน แต่จะลดความซับซ้อนของประตู (Gate) ลงเหลือเพียงแค่ 2 ประตู เพื่อให้เกิดความรวดเร็วในการคำนวณ ได้แก่

- Update Gate (z) เป็นประตูนำเข้าข้อมูลเพื่อนำไปคำนวณในการปรับสถานะของค่าเซลล์ เพื่อส่งเป็นข้อมูลขาออกในการคำนวณในขั้นต่อไป ฟังก์ชันกระตุ้นที่ควบคุมการทำงานประตูนี้อยู่คือ ฟังก์ชันซิกมอยด์ (Sigmoid Function: σ) ผลลัพธ์ที่จะได้ออกมาจะมีค่าอยู่ระหว่างค่า 0 ถึง 1 (หรือก็คือ z_t) โดยสำหรับสมการการคำนวณของประตูปรับค่านี้ สามารถแสดงได้ดังนี้

$$z_t = \sigma\left(W_{xz} x_t + W_{hz} h_{t-1} + b_z \right)$$

- Reset Gate (r) เป็นประตูที่ใช้ในการตัดสินใจว่าจะปรับสถานะเซลล์เมื่อมีข้อมูลใหม่เข้ามาหรือไม่ มีฟังก์ชันซิกมอยด์ทำหน้าที่เป็นฟังก์ชันกระตุ้น เพื่อทำการตัดสินใจว่า จะให้มีการปรับเปลี่ยนสถานะเซลล์หรือไม่ ในการคำนวณส่วนนี้จะใช้ข้อมูลนำเข้าปัจจุบันที่เข้า

มาพร้อมกับชั้นซ่อน (Hidden state) ก่อนหน้า โดยสำหรับสมการการคำนวณของประตูรีเซ็ตค่านี้ สามารถแสดงได้ดังนี้

$$r_t = \sigma\left(W_{xr} x_t + W_{hr} h_{t-1} + b_r \right)$$

หลังจากที่ได้สมการของประตูทั้ง 2 แล้ว สิ่งที่เราต้องการคือการคำนวณเอาต์พุตที่ทำนายได้ในสถานะชั้นซ่อนที่เวลาปัจจุบัน ($\widehat{h_t}$) ซึ่งคำนวณจากผลรวมระหว่างข้อมูลนำเข้าที่เวลาปัจจุบัน (x_t) คูณกับเมทริกซ์ค่าถ่วงน้ำหนักของเอาต์พุต (W_{xh}) กับการทำ Element-wise product ($\odot$) ระหว่าง Reset Gate (r) กับสถานะชั้นซ่อนก่อนหน้า (Hidden state) เพื่อตัดสินใจว่าจะเอาสถานะของชั้นซ่อนก่อนหน้านี้มาเท่าใด ซึ่งผลลัพธ์ของการทำ Element-wise product ($\odot$) จะคูณอยู่กับเมทริกซ์ค่าถ่วงน้ำหนักของ Update Gate ดังนั้น สมการสำหรับคำนวณค่าดังกล่าว จึงแสดงได้ดังนี้

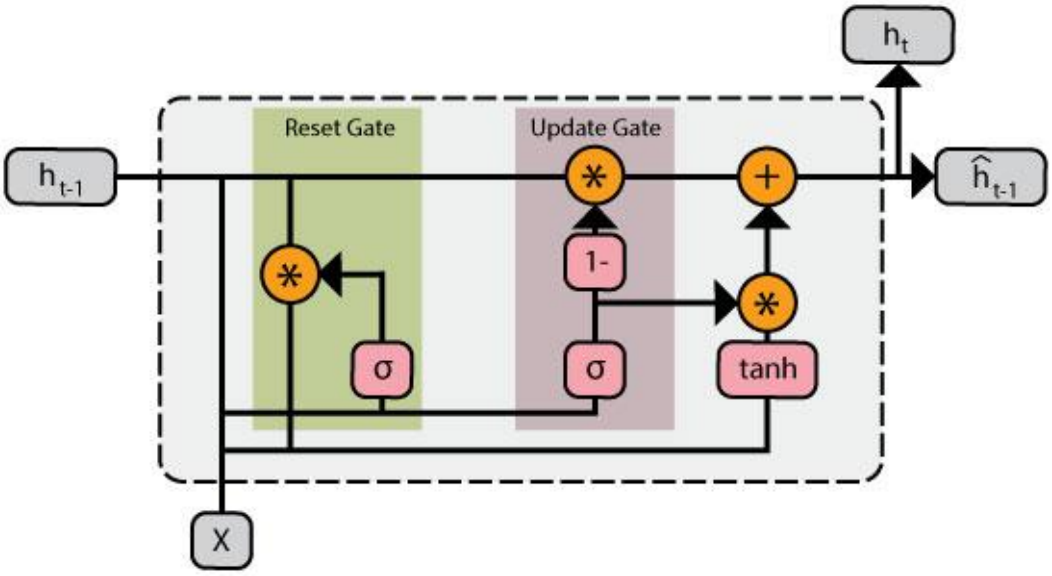
$$\widehat{h_t} = \tanh\left(W_{xh} x_t + W_{rhh}\left(r_t \odot h_{t-1} \right) \right)$$

เนื่องจากฟังก์ชันกระตุ้นที่ใช้คือ ฟังก์ชันไฮเพอร์โบลิกแทนเจนต์ (Hyperbolic Tangent: tanh) หรือ ฟังก์ชันแทน ซึ่งเป็นฟังก์ชั่นไม่เชิงเส้น (nonlinearity function) ค่าที่ได้ออกมาจึงจะอยู่ในช่วง -1 ถึง 1 ซึ่งจะช่วยป้องกันการเกิดปัญหาการลดลงของเกรเดียนต์ เนื่องจากฟังก์ชั่นนี้จะจำกัดค่า (Bounds) ให้อยู่ระหว่างช่วงดังกล่าว จึงทำให้สามารถที่จะคำนวณค่าการแพร่กลับของความผิดพลาด (Backpropagation) ได้โดยที่ค่าที่ได้ออกมาจะไม่เป็นค่าอนันต์

ขั้นตอนสุดท้ายคือการคำนวณสถานะชั้นซ่อนที่เวลาปัจจุบันเพื่อส่งออกไปเป็นผลลัพธ์สุดท้าย หรือก็คือเป็นการตัดสินใจว่าจะทำการปรับเปลี่ยนค่าให้เป็นค่าใด ระหว่างผลลัพธ์ที่ทำนายได้จากสถานะชั้นซ่อนก่อนหน้า ($\widehat{h_t}$) หรือผลลัพธ์ที่ทำนายได้จากสถานะชั้นซ่อนปัจจุบัน (h_t) ในสัดส่วนเท่าใด สมการสำหรับคำนวณค่าดังกล่าว จึงแสดงได้ดังนี้

$$h_t = \left(1 - z_t \right) h_{t-1} + z_t \widehat{h_t}$$

สำหรับโครงสร้างภายในของโครงข่ายประตูวกกลับ สามารถแสดงได้ดังรูปที่ 7



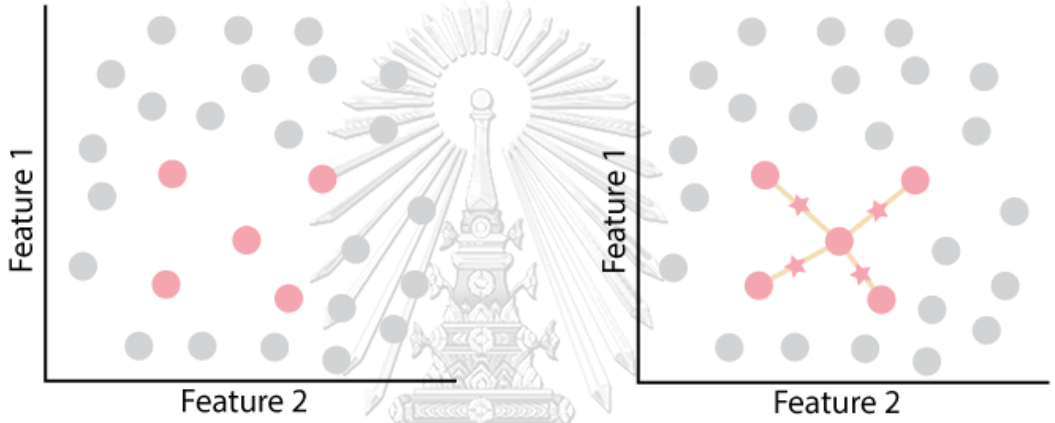
รูปที่ 7 โครงข่ายประตูวกกลับ (GRU)

2.1.8 ปัญหาข้อมูลไม่สมดุล (Imbalanced data)

ปัญหาข้อมูลไม่สมดุล (Imbalanced data) เป็นปัญหาที่เกิดขึ้นในชุดข้อมูลที่มีการแบ่งประเภทของข้อมูลในชุดข้อมูล ออกเป็น 2 กลุ่ม (หรือมากกว่า) แล้วข้อมูลในกลุ่มใดกลุ่มหนึ่งมีจำนวนมากกว่าข้อมูลอีกกลุ่มหนึ่งมาก กล่าวคือ จะมีข้อมูลที่เป็นกลุ่มส่วนมาก (Majority) ซึ่งจะมีจำนวนของข้อมูลในกลุ่มมากหรือเป็นกลุ่มส่วนใหญ่ของชุดข้อมูล ในขณะเดียวที่ข้อมูลในกลุ่มส่วนน้อย (Minority) จะมีจำนวนของข้อมูลในกลุ่มน้อย ซึ่งอาจเกิดขึ้นได้ทั้งจากการที่ลักษณะทางธรรมชาติของข้อมูลมีความแตกต่างของจำนวนในแต่ละกลุ่ม หรืออาจเกิดจากข้อจำกัดในการเก็บข้อมูล

อย่างไรก็ตาม การที่ข้อมูลในชุดข้อมูลหนึ่งๆ มีความไม่สมดุลเกิดขึ้น จะส่งผลต่อการจำแนกข้อมูลในกลุ่มส่วนน้อย เพราะโดยทั่วไปนั้น วิธีการจำแนกประเภท (Classification) ต่างๆ จะมีประสิทธิภาพสูงก็ต่อเมื่อข้อมูลในแต่ละกลุ่มในชุดข้อมูลนั้นมีจำนวนที่ใกล้เคียงกัน ซึ่งหากชุดข้อมูลมีความไม่สมดุล หรือ ข้อมูลที่เป็นกลุ่มส่วนมากมีมากกว่ากลุ่มส่วนน้อยมากจนเกินไป ก็จะส่งผลให้วิธีการจำแนกประเภทที่ใช้โดยทั่วไปนั้นมีโอกาสที่จะทำนายกลุ่มของชุดข้อมูลออกมาเป็นกลุ่มส่วนมาก อันนำไปสู่ปัญหาที่เรียกว่า ปัญหาการแบ่งกลุ่มข้อมูลผิดกลุ่ม (Misclassification) เนื่องจากตัวแบบจะมองว่าการทำนายออกมาเป็นกลุ่มส่วนมากจะมีความแม่นยำมากกว่าเพราะเป็นกลุ่มที่เป็นข้อมูลส่วนใหญ่ของชุดข้อมูล ซึ่งก็จะทำให้ตัวแบบไม่สามารถทำนาย หรือแยกประเภทข้อมูลที่เป็นกลุ่มส่วนน้อยออกมาได้เลย

วิธีการแก้ปัญหาข้อมูลไม่สมดุลนั้น มีด้วยกันหลายวิธี แต่ในที่นี้จะขอกล่าวถึงแค่วิธีเดียว ซึ่งเป็นวิธีที่ผู้วิจัยได้เลือกใช้ในการนำมาแก้ปัญหาความไม่สมดุลที่เกิดขึ้นในชุดข้อมูล ซึ่งก็คือวิธีการสุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วยการสังเคราะห์ (Synthetic Minority Over-sampling Technique: SMOTE) เป็นวิธีที่ถูกเสนอขึ้นในปี 2002 โดย (Chawla et al., 2002) เป็นการสร้างกลุ่มตัวอย่างส่วนน้อยตัวอย่างใหม่ขึ้นมาอย่างสุ่ม โดยใช้หลักการกำหนดจำนวนเพื่อนบ้านที่อยู่ใกล้ที่สุดจำนวน k ตัว (k-nearest neighbors: KNN) แล้วทำการสุ่มสร้างข้อมูลขึ้นมาบนแนวเส้นที่เชื่อมต่อระหว่างตัวอย่างกลุ่มส่วนน้อยใดๆ วิธีการดังกล่าวสามารถอธิบายได้ด้วยแผนภาพดังต่อไปนี้



รูปที่ 8 โครงสร้าง Synthetic Minority Over-sampling Technique: SMOTE

อย่างไรก็ตาม เทคนิค SMOTE นั้น ถูกพัฒนาขึ้นมาสำหรับตัวแบบประเภท Classification ที่มีการแบ่งกลุ่มของชุดข้อมูลอย่างชัดเจน ไม่ใช่ตัวแบบประเภท Regression แต่ก็ได้มีการศึกษาวิจัยและพัฒนาต่อยอด เพื่อให้ SMOTE นั้น สามารถจัดการกับตัวแบบที่เป็น Regression ได้ จนกระทั่งในปี 2017 เทคนิค SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) โดย (Branco et al., 2017) ก็ได้ถูกเสนอขึ้น หลักการในการทำงานของ SMOGN คือ การกำหนดข้อมูลที่เป็นข้อมูลกลุ่มส่วนน้อย และข้อมูลกลุ่มส่วนมาก โดยจะทำการสุ่มลด (Under sampling) ข้อมูลส่วนมาก และสุ่มเพิ่ม (Over sampling) ข้อมูลส่วนน้อย สำหรับในขั้นตอนการสุ่มเพิ่มข้อมูลส่วนน้อยนั้นจะมีการเพิ่มเทคนิค Gaussian Noise เข้าไปด้วย ซึ่งเทคนิค SMOTE ปกตินั้น จะทำการสุ่มสร้างข้อมูลด้วยหลักการ k-nearest neighbors ซึ่งเป็นการหาข้อมูลใกล้เคียงในกลุ่มเดียวกันที่ใกล้ที่สุด แต่หากข้อมูลในกลุ่มเดียวกันอยู่ไกลออกไป SMOGN ก็จะใช้ Gaussian noise ในการสุ่มสร้างแทน

2.1.9 การประเมินประสิทธิภาพตัวแบบ (Model Assessment)

หลังจากที่ได้ทำการพัฒนาตัวแบบจำลองขึ้นมาแล้ว การวัดประสิทธิภาพของตัวแบบนั้นถือเป็นเรื่องสำคัญ เนื่องจากประสิทธิภาพของตัวแบบจะเป็นตัวบอกความแม่นยำในการพยากรณ์ของตัวแบบ ว่าตัวแบบที่เราพัฒนาขึ้นมานั้น มีความเหมาะสมกับข้อมูลหรือไม่ สามารถพยากรณ์ค่าได้แม่นยำหรือคลาดเคลื่อนจากค่าที่เกิดขึ้นจริงมากน้อยขนาดไหน ซึ่งตัวแบบที่ดีนั้น ไม่ควรที่จะมีความคลาดเคลื่อนในการพยากรณ์ที่สูงมาก ค่าที่ได้ควรมีความใกล้เคียงกับค่าที่เกิดขึ้นจริง หรือหากค่าที่พยากรณ์ได้จากตัวแบบนั้นคลาดเคลื่อนไปจากค่าจริงมาก ก็ควรที่จะกลับไปตรวจสอบขั้นตอนการพัฒนาตัวแบบใหม่อีกครั้งว่า ตัวแบบที่เราได้ทำการพัฒนาขึ้นมานั้น มีปัญหาอะไร มีความ Over-fitting หรือ Under-fitting กับข้อมูลอย่างไร เพื่อที่จะได้ทำการปรับปรุง และเลือกตัวแบบที่มีความเหมาะสมกับรูปแบบของชุดข้อมูลที่สุดมาเป็นตัวแบบที่ใช้ในการพยากรณ์

สำหรับเกณฑ์การวัดประสิทธิภาพของตัวแบบจำลองนั้น ในงานวิจัยนี้ได้เลือกใช้เกณฑ์ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) ซึ่งเหตุผลที่เลือกใช้เกณฑ์ MAE นี้ เนื่องจากเป็นวิธีที่มีความอ่อนไหวกับตัวข้อมูลที่ผิดแปลกไปจากปกติ น้อยกว่า MSE และ RMSE เนื่องจากเป็นการนำค่าความผิดพลาด (Error) มาใส่สัมบูรณ์ (Absolute) ซึ่งหากตัวชุดข้อมูลมีค่าผิดปกติอยู่เยอะ ค่านี้ก็อาจเหมาะสมกว่าในการนำไปใช้ประเมินประสิทธิภาพของโมเดล และตัวข้อมูลปริมาณน้ำฝนนี้ก็มักจะมีค่าผิดปกติ ซึ่งเกิดจากช่วงเวลาที่มีปริมาณน้ำฝนสะสมมากกว่าปกติ หรือก็คือช่วงที่ฝนตกหนัก หรือฝนตกหนักมาก โดยสมการของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) สามารถแสดงได้ดังนี้

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

โดยที่

n คือ จำนวนข้อมูลทั้งหมดของชุดข้อมูล

y_i คือ ข้อมูลจริงที่เกิดขึ้นของชุดข้อมูลตัวที่ i

$\hat{y_i}$ คือ ข้อมูลที่ตัวแบบจำลองทำนายได้ของชุดข้อมูลตัวที่ i

2.2 งานวิจัยที่เกี่ยวข้อง

เนื่องจากตัวข้อมูลสภาพอากาศนั้นเป็นข้อมูลเชิงอนุกรมเวลา กล่าวคือ เป็นข้อมูลที่ได้จากการเก็บข้อมูลตามลำดับเวลาต่อเนื่องกันเป็นช่วงๆ ในอดีต จึงได้มีการทำการศึกษาและปรับปรุงตัวแบบทางสถิติต่างๆ เพื่อให้สามารถพยากรณ์ข้อมูลอนุกรมเวลาให้ออกมาแม่นยำให้มากที่สุด โดยตัวแบบแรกที่ถูกนำมาใช้ในการพยากรณ์ดังกล่าว คือตัวแบบอนุกรมเวลา (Time series) เนื่องจากตัวแบบอนุกรมเวลาถือเป็นตัวแบบตั้งต้นสำหรับการพยากรณ์ข้อมูลอนุกรมเวลา และในงานวิจัยนี้ก็ได้นำตัวแบบนี้เข้ามาพิจารณาเปรียบเทียบร่วมด้วย โดยตัวแบบอนุกรมเวลาตัวแรกที่ถูกนำมาพิจารณาคือตัวแบบ ARIMA

(Wang et al., 2013) ได้ทำการพยากรณ์ปริมาณน้ำฝนสะสมในเมืองโซว่กวง มณฑลซานตง ประเทศจีน ซึ่งได้พบว่าปริมาณน้ำฝนนั้นมีสหสัมพันธ์อัตโนมัติที่โดดเด่น (Strong autocorrelation) จึงได้เลือกใช้ตัวแบบ SARIMA (Seasonal Autoregressive and Moving Average) ในการวิเคราะห์และพยากรณ์ ซึ่งหลังจากการเตรียมข้อมูล การกำหนดพารามิเตอร์ และเลือกตัวแบบที่ดีที่สุดตามวิธีการของการทำตัวแบบอนุกรมเวลาแล้ว จากการทดสอบประสิทธิภาพพบว่า ค่าที่พยากรณ์ได้จากตัวแบบมีค่าใกล้เคียงกับปริมาณน้ำฝนสะสมจริง โดยมีค่าความคลาดเคลื่อนเพียงร้อยละ 27.42 เท่านั้น แต่ในงานวิจัยของ Wang และคณะ ไม่ได้กล่าวถึงการเปรียบเทียบแบบจำลอง SARIMA กับแบบจำลองอื่นๆ

(Narayanan et al., 2013) ได้ทำการพยากรณ์ปริมาณน้ำฝนสะสมในช่วงฤดูก่อนมรสุมตะวันตกเฉียงใต้ (ระหว่างเดือนมีนาคมถึงพฤษภาคม) ในภาคตะวันตกของประเทศอินเดีย โดยใช้ข้อมูลสภาพอากาศรายเดือน เลือกเฉพาะช่วงเดือนมีนาคมถึงพฤษภาคม จากสถานีวัดสภาพอากาศในประเทศอินเดีย 6 สถานี ในช่วงเวลา 60 ปี โดยมีการใช้ Mann Kendall test ในการทดสอบสหสัมพันธ์ระหว่างตัวแปร และใช้ตัวแบบ ARIMA ในการพยากรณ์ พบว่า ผลการพยากรณ์ทั้ง 6 สถานีมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย(RMSE) ที่มากน้อยแตกต่างกันไป หมายความว่า การใช้ตัวแบบ ARIMA ในการพยากรณ์ปริมาณน้ำฝนสะสมอาจจะไม่ได้ตอบโจทย์สำหรับทุกสถานี หรือบางสถานีอาจจะต้องการการจัดการกับตัวแปรรับเข้าก่อนที่จะนำไปพยากรณ์ ทั้งนี้ งานวิจัยของ Narayanan และคณะ ก็ไม่ได้กล่าวถึงการเปรียบเทียบแบบจำลอง ARIMA กับแบบจำลองอื่นเช่นกัน

อย่างไรก็ตาม แบบจำลอง ARIMA ก็ยังมีข้อเสีย นั่นคือ เป็นแบบจำลองที่พิจารณาโดยใช้ตัวแปรเพียงตัวเดียว ซึ่งในความเป็นจริงนั้น ปัจจัยที่ทำให้เกิดฝนหรือส่งผลต่อปริมาณน้ำฝนสะสมนั้น มีได้หลายปัจจัย และเพื่อเพิ่มประสิทธิภาพในการพยากรณ์จึงควรที่จะเอาปัจจัยสภาพอากาศด้านอื่นๆ

มาพิจารณาร่วมด้วย เพื่อให้ตอบโจทย์ดังกล่าว แบบจำลอง ARIMAX ที่เป็นแบบจำลองที่สามารถนำเข้าข้อมูลคุณลักษณะอื่นๆ ที่เกี่ยวข้องกับตัวข้อมูลที่ต้องการจะพยากรณ์ได้ จึงถูกยกขึ้นมาเพื่อทำการเปรียบเทียบกับแบบจำลอง ARIMA

(Wangdi et al., 2010) ได้เปรียบเทียบผลการพยากรณ์การเกิดไข้มาลาเรีย ระหว่างแบบจำลอง ARIMA ที่รับข้อมูลสถิติผู้ป่วยไข้มาลาเรียเพียงตัวแปรเดียว กับตัวแบบ ARIMAX ที่นอกจากจะรับข้อมูลของจำนวนผู้ป่วยแล้ว ยังรับข้อมูลด้านสภาพภูมิอากาศเพิ่มเข้าไปด้วย พบว่า ตัวแบบ ARIMAX ให้ผลการพยากรณ์ที่ดีกว่าตัวแบบ ARIMA

(Jalalkamali et al., 2015) ได้พัฒนาตัวแบบพยากรณ์ปริมาณน้ำฝนเพื่อคาดการณ์การเกิดปัญหาภัยแล้ง โดยในงานวิจัยได้ทำการเปรียบเทียบผลการพยากรณ์ระหว่างตัวแบบ ARIMAX ตัวแบบโครงข่ายประสาทเทียม และตัวแบบ support-vector machines (SVM) พบว่าในการพยากรณ์ระยะสั้น ตัวแบบ ARIMAX เป็นตัวแบบที่มีประสิทธิภาพมากที่สุด

จากงานวิจัยข้างต้นจะเห็นได้ว่า การเพิ่มคุณลักษณะอื่นๆ ที่เกี่ยวข้องกับคุณลักษณะที่เราต้องการจะพยากรณ์นั้นช่วยเพิ่มประสิทธิภาพของการพยากรณ์ได้จริงๆ จึงทำให้ตัวแบบ ARIMAX เป็นตัวเลือกที่ดีกว่าตัวแบบ ARIMA ทั้งนี้ นอกจากการใช้ตัวแบบ ARIMA และ ARIMAX การพยากรณ์โดยใช้แบบจำลองการเรียนรู้เชิงลึก โครงข่ายระบบประสาทแบบย้อนกลับ รวมไปถึงการทำ Feature Engineering ให้กับชุดข้อมูล เพื่อเพิ่มประสิทธิภาพให้กับตัวแบบก็ได้เข้ามามีบทบาทมากขึ้น

(Srachoom, 2007) ได้พัฒนาโครงข่ายประสาทเทียมเพื่อให้สามารถพยากรณ์ปริมาณน้ำฝนสะสมระยะสั้น จากข้อมูลสภาพอากาศย้อนหลังก่อนวันที่ต้องการพยากรณ์ได้ โดยข้อมูลที่ใช้เป็นข้อมูลพื้นที่อำเภอเมือง จังหวัดเชียงใหม่ ได้แก่ ความกดอากาศสูงสุด-ต่ำสุด อุณหภูมิสูงสุด-ต่ำสุด ความชื้นสัมพัทธ์สูงสุด-ต่ำสุด และปริมาณน้ำฝน ย้อนหลัง 9 ปี (พ.ศ. 2541-2549) และเพื่อให้โครงข่ายประสาทได้เรียนรู้แนวโน้มของสภาพอากาศก่อนวันฝนตกได้อย่างชัดเจน จึงได้เพิ่มชุดข้อมูลของสภาพอากาศย้อนหลัง 3 วัน ก่อนวันพยากรณ์ และเพิ่มดัชนีแสดงฤดูกาลเข้าไปด้วย ส่งผลให้มีตัวแปรต้นทั้งหมด 24 ตัว พบว่าโครงข่ายประสาทเทียมที่เหมาะสมมี 1 ชั้นซ่อน มีจำนวน 30 นิวรอน มีความถูกต้องเฉลี่ยร้อยละ 73.36

(Hung et al., 2009) ได้ทำการพยากรณ์ปริมาณน้ำฝนในเขตกรุงเทพมหานครแบบรายงานผลได้ทันการณ์(real time) ด้วยโครงข่ายประสาทเทียม(ANN) โดยทำการเปรียบเทียบระหว่างแบบ

Simple multilayer perceptron และแบบ Generalized feedforward ร่วมกับการปรับจูนพารามิเตอร์ ข้อมูลที่นำมาใช้เป็นข้อมูลที่ทำการเก็บข้อมูลแบบรายชั่วโมงตลอดระยะเวลา 4 ปี จาก 75 สถานีตรวจวัดสภาพอากาศในกรุงเทพมหานคร พบว่า ในการพยาการณ์ปริมาณน้ำฝนในช่วงเวลา 1-6 ชั่วโมงนั้น ยิ่งชั่วโมงเพิ่มขึ้นเท่าไหร่ ค่าความคลาดเคลื่อนของการพยากรณ์ก็จะยิ่งเพิ่มขึ้นเช่นกัน นอกจากนี้ในงานวิจัยนี้ยังได้ทำการหาความสัมพันธ์ระหว่างตัวแปรต้นกับตัวแปรตามเพื่อพิจารณาว่าตัวแปรใดบ้างที่ส่งผลกับปริมาณน้ำฝน พบว่า เมื่อนำข้อมูลสภาพอากาศจากสถานีตรวจวัดใกล้เคียง (ในรัศมี 21 ตารางกิโลเมตร) มาพิจารณาร่วมกับสถานีตรวจวัดที่สนใจจะพยากรณ์แล้วนั้น ส่งผลให้ค่าความคลาดเคลื่อนลดลงถึงร้อยละ 71.19 เมื่อเทียบกับการพิจารณาข้อมูลจากสถานีตรวจวัดที่สนใจเพียงสถานีเดียว

(Hernández et al., 2016) ได้ศึกษาเรื่องการพัฒนาแบบจำลองการเรียนรู้เชิงลึกเพื่อทำนายปริมาณน้ำฝนสะสมรายวัน โดยใช้ข้อมูลสภาพอากาศรายวันจากสถานีตรวจวัดในเมือง Manizales ประเทศ Colombia สิ่งที่น่าสนใจของงานวิจัยนี้คือการสร้างตัวแปรต้นจำนวน 47 ตัว โดยมาจากข้อมูลพื้นฐานที่เก็บรวบรวมมา ได้แก่ อุณหภูมิ ความชื้นสัมพันธ์ ความกดอากาศ ความเข้มของแสงอาทิตย์ ความเร็วลม และทิศทางลม ในส่วนของข้อมูลที่สร้างขึ้นจะมาจากการข้อมูลพื้นฐานแต่ละตัวอีกที โดยการนำมา 1) วัดย้อนหลังไป 3 วัน 2) ค่าเฉลี่ยในช่วง 5 วันย้อนหลัง 3) ความแตกต่างของอุณหภูมิที่วัดได้ในช่วงเวลา 4:00 น. และ 24:00 น. เป็นต้น เนื่องจากตัวแปรต้นที่มีจำนวนมาก จึงได้มีการนำตัวเข้ารหัสอัตโนมัติ (Autoencoder) มาช่วยในการสกัดคุณลักษณะสำคัญและทำการหาความสัมพันธ์ของแต่ละตัวแปร ก่อนที่จะส่งไปต่อไปยัง Multilayer Perceptron (MLP) เพื่อทำหน้าที่ในการพยากรณ์ปริมาณน้ำฝนสะสม

โครงข่ายประสาทเทียมนั้น มีจุดเด่นที่การมีชั้นซ่อนซึ่งจะช่วยจัดการในเรื่องของการผสมผสานและจับคู่คุณลักษณะที่เกี่ยวข้องในการพยากรณ์ได้อย่างเหมาะสม ทำให้ประสิทธิภาพในการพยากรณ์เพิ่มขึ้น ทั้งนี้ นอกจากการทำ Feature Engineering เพื่อเพิ่มประสิทธิภาพให้กับตัวแบบแล้ว งานวิจัยข้างต้นยังกล่าวถึงการจัดการกับข้อมูลเบื้องต้น เพื่อไม่ให้เกิดแนวโน้มการพยากรณ์ผิดกลุ่ม (Misclassification) เนื่องจากข้อมูลที่นำมาใช้นั้นมีทั้งข้อมูลช่วงที่ฝนตก และข้อมูลในช่วงที่ฝนไม่ตก หากนำข้อมูลในช่วงที่ฝนไม่ตกมาทำนายร่วมด้วยก็จะทำให้ไม่สามารถทำนายปริมาณน้ำฝนได้อย่างถูกต้อง โดยงานวิจัย (Srachoom, 2007) ได้แก้ไขปัญหาชุดข้อมูลไม่สมดุลด้วยการเพิ่มดัชนีฤดูกาลด้วยเลขไบนารีขนาด 3 บิต เพื่อให้แบบจำลองสามารถรับรู้กลุ่มประเภทของข้อมูลตามฤดูกาลได้ และงานวิจัย (Hung et al., 2009) ได้ทำการศึกษาเปรียบเทียบโดยการนำเข้าชุดข้อมูลหลายๆ

แบบ พบว่า นอกจากการใช้ข้อมูลสภาพอากาศจากสถานีตรวจวัดใกล้เคียงมาพิจารณาร่วมกับสถานีตรวจวัดที่สนใจจะพยากรณ์แล้ว การใส่ข้อมูลนำเข้าที่จะทำให้การพยากรณ์มีความแม่นยำมากที่สุดก็คือ ใส่ข้อมูลเฉพาะวันที่ฝนตก งานวิจัย (Hernández et al., 2016) จะมีความคล้ายงานวิจัย (Srachoom, 2007) แต่จะแบ่งข้อมูลด้วยการเพิ่มดัชนีรายเดือนด้วยเลขไบนารีขนาด 12 บิต อย่างไรก็ตาม การทำ Feature Engineering ของงานวิจัยที่ยกมาข้างต้นนั้น ไม่ได้มีการเปรียบเทียบ หรืออธิบายว่า การเพิ่มคุณลักษณะพิเศษทางสถิติให้กับชุดข้อมูลนั้น จะช่วยให้ผลการพยากรณ์ดีขึ้นกว่าการใช้ชุดข้อมูลตั้งต้นจริงๆ หรือไม่ ซึ่งคาดว่าอาจจะเป็นเพราะตัวแบบโครงข่ายประสาทเทียมมีความสามารถในการสกัดตัวคุณลักษณะที่ไม่มีความสำคัญออกไปจากตัวแบบได้ จึงทำให้หลายๆ งานวิจัยมุ่งเน้นไปในเรื่องของการพัฒนาเพื่อเพิ่มประสิทธิภาพให้กับตัวแบบมากกว่าที่จะให้ความสำคัญการการเพิ่มคุณลักษณะให้กับตัวแบบ

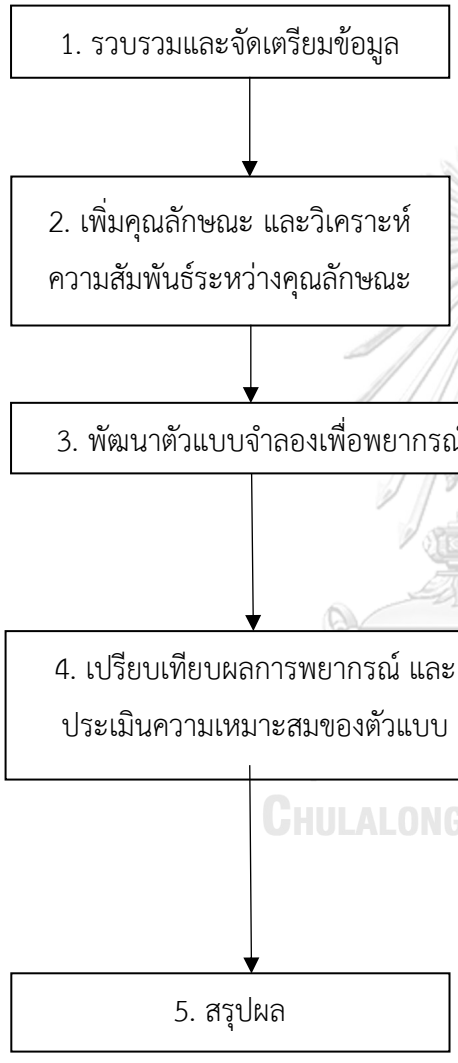
ในส่วนของการพยากรณ์ปริมาณน้ำฝนสะสมที่เกิดขึ้น นอกจากแบบจำลองอนุกรมเวลา และโครงข่ายประสาทเทียมที่ถูกกล่าวถึงในงานวิจัยข้างต้น ในปัจจุบันได้มีการนำความรู้เรื่องแบบจำลองการเรียนรู้เชิงลึกเข้ามาเพื่อช่วยทำนาย โดยแบบจำลองการเรียนรู้เชิงลึกที่นิยมนำมาใช้ทำนายปริมาณน้ำฝนคือ แบบจำลองที่มีความเกี่ยวข้องกับอนุกรมเวลา ได้แก่ RNN LSTM และ GRU โดยงานวิจัย (Salman et al., 2018) ได้ทำการทดลองเพื่อเพิ่มประสิทธิภาพให้กับตัวแบบ LSTM โดยการสร้างตัวแบบ LSTM ผสมผสานกับตัวแบบ ARIMA และทำการเปรียบเทียบตัวแบบ ARIMA LSTM กับตัวแบบผสมผสาน ซึ่งพบว่าการผสมผสานตัวแบบ LSTM เข้ากับตัวแบบ ARIMA นั้นช่วยเพิ่มประสิทธิภาพให้การทำนายได้ เช่นเดียวกับ (Fan et al., 2020) ได้ทำการทดลองพัฒนาแบบจำลอง LSTM เพื่อให้มีความเหมาะสมกับข้อมูลและพยากรณ์ปริมาณน้ำฝน พร้อมนำแบบจำลองที่ได้ไปเปรียบเทียบกับแบบจำลองอื่นๆ เช่น ARIMA พบว่าแบบจำลอง LSTM ที่พัฒนาขึ้นนั้นมีประสิทธิภาพในการพยากรณ์ที่โดดเด่น ต่อมา (Aurnhammer & Frank, 2019) กับ (Yang et al., 2020) ได้ทำการทดลองเพื่อเปรียบเทียบความแม่นยำในการพยากรณ์ระหว่างแบบจำลองโครงข่ายระบบประสาทย้อนกลับด้วยกัน (RNN LSTM และ GRU) พบว่า แบบจำลอง LSTM และ GRU มีความโดดเด่น และสามารถให้ผลการพยากรณ์ที่มีความถูกต้องและรวดเร็ว อย่างไรก็ตาม ในด้านความเหมาะสมในการเลือกใช้ตัวแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับนั้นขึ้นอยู่กับตัว Short-term memory ที่ใส่เข้าไปด้วยเช่นกัน โดยทั้ง 2 งานวิจัยพบว่า ในชุดข้อมูลที่มีขนาดเล็กและถ้าจำนวนของข้อมูลก่อนหน้า (Short-term memory) ที่จะใส่เข้าไปในแบบจำลองไม่ได้ยาวมาก

แบบจำลอง GRU จะมีประสิทธิภาพในการเรียนรู้ที่ดีกว่า แต่สำหรับชุดข้อมูลขนาดใหญ่ ที่มีข้อมูลก่อนหน้าที่จะใส่เข้าไปจำนวนมากนั้น แบบจำลอง LSTM จะมีประสิทธิภาพการเรียนรู้ที่ดีกว่า

จากการศึกษางานวิจัยทั้งงานทางด้านการทำ Feature Engineering และด้านการปรับปรุงตัวแบบโครงข่ายระบบประสาทแบบย้อนกลับพบว่า งานวิจัยทั้ง 2 แบบจะมุ่งเน้นในส่วนของการปรับจูนไฮเปอร์พารามิเตอร์ หรือการพัฒนาต่อยอดตัวแบบเพื่อให้มีความเหมาะสมและสามารถพยากรณ์ชุดข้อมูลให้มีความแม่นยำมากที่สุดเป็นหลักอยู่ ไม่ได้กล่าวถึงการทำ Feature Engineering โดยเฉพาะการเพิ่มคุณลักษณะทางสถิติให้กับชุดข้อมูลก่อนที่จะนำเข้าตัวแบบจำลอง ซึ่งคาดว่าเป็นเพราะความสามารถของแบบจำลองการเรียนรู้เชิงลึกที่มีชั้นซ่อน (Hidden Layer) ที่สามารถทำการสกัดตัวแปรที่ไม่มีความสำคัญกับตัวแบบออกไป จึงทำให้หลายงานวิจัยนั้น ไม่ได้มีการเปรียบเทียบให้เห็นว่าการเพิ่มคุณลักษณะทางสถิติต่างๆ ให้กับชุดข้อมูลนั้น ทำให้ประสิทธิภาพของการพยากรณ์เพิ่มขึ้นได้จริงหรือไม่ จึงนำมาสู่จุดมุ่งหมายในการศึกษาข้อแรก คือการเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบที่ใช้ชุดข้อมูลที่แตกต่างกัน ระหว่างชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะทางสถิติกับชุดข้อมูลดั้งเดิม เพื่อดูว่าการเพิ่มคุณลักษณะนั้นส่งผลที่ดีกับประสิทธิภาพในการพยากรณ์จริงหรือไม่ และสำหรับจุดมุ่งหมายข้อที่สอง คือ เปรียบเทียบประสิทธิภาพในการพยากรณ์ในช่วงเวลาถัดไปของโครงข่ายระบบประสาทแบบย้อนกลับกับแบบจำลองอาริมา ซึ่งจะเป็นการทดลองเพื่อหาแบบจำลองที่มีความเหมาะสมกับประเภทชุดข้อมูลที่นำมาศึกษามากที่สุด ทั้งนี้ ดังที่กล่าวไปแล้วว่าในกลุ่มตัวแบบโครงข่ายระบบประสาทแบบย้อนกลับนั้น แต่ละตัวแบบมีข้อดี และความได้เปรียบกับชุดข้อมูลที่แตกต่างกัน กล่าวคือ บางตัวแบบอาจจะไม่มีความเหมาะสม หรือมีความเหมาะสมน้อยกับชุดข้อมูลที่เรานำมาศึกษา ซึ่งการจะจำเพาะเจาะจงตัวแบบใดตัวแบบนึงเลยนั้น อาจจะทำให้เราไม่สามารถทราบได้ว่าตัวแบบใดเป็นตัวแบบที่มีความเหมาะสมกับชุดข้อมูลของเรา จึงนำมาสู่จุดมุ่งหมายที่ 2 ของงานวิจัยนี้

บทที่ 3
วิธีดำเนินการวิจัย

ในส่วนของลำดับขั้นตอนในการดำเนินการของงานวิจัยชิ้นนี้ สามารถเขียนเป็นแผนผังโดยคร่าวได้ดังต่อไปนี้



ศึกษาลักษณะของข้อมูลที่มี พร้อมทั้งแก้ไขค่าว่าง

เพิ่มคุณลักษณะ(Feature) ให้กับชุดข้อมูล พร้อมทั้งวิเคราะห์ความสัมพันธ์ของคุณลักษณะที่มีปริมาณน้ำฝนสะสม

พัฒนาตัวแบบจำลอง ARIMA ARIMAX Simple-RNN LSTM และ GRU ที่มีความเหมาะสมกับชุดข้อมูล

หลังจากได้ผลการพยากรณ์ออกมาแล้ว จะทำการตรวจสอบความถูกต้องของการพยากรณ์ด้วย MAE และทำการเปรียบเทียบความแม่นยำในการพยากรณ์ด้วยแบบจำลองวิเคราะห์การถดถอยต่างๆ

แสดงผลลัพธ์การพยากรณ์ปริมาณน้ำฝนที่ได้ และสรุปผลการทดลอง

3.1 ศึกษาและจัดเตรียมข้อมูล

งานวิจัยนี้ได้รับการสนับสนุนข้อมูลสภาพอากาศบริเวณสนามบินสุวรรณภูมิย้อนหลัง 5 ปี (ตั้งแต่ปี 2016 จนถึงปี 2020) จากกรมอุตุนิยมวิทยาประเทศไทย โดยชุดข้อมูลที่ได้รับมาจะมีทั้งหมด 4 ชุดข้อมูล ได้แก่ อุณหภูมิรายชั่วโมง (Celsius) ความชื้นสัมพันธ์รายชั่วโมง (Percent) ความกด

อากาศรายชั่วโมง (hPa) และปริมาณน้ำฝนสะสม 3 ชั่วโมง (Millimeter) ซึ่งตัวอย่างของ 1 ในชุดข้อมูลที่ได้รับมาจากกรมอุตุนิยมวิทยา สามารถแสดงได้ดังรูปที่ 9

ปริมาณฝน(มิลลิเมตร)
รายชั่วโมง

ที่	รหัสสถานี-สถานี-จังหวัด	วันที่	เวลาทำการตรวจ								รวม
			0100	0400	0700	1000	1300	1600	1900	2200	
1	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	1/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
2	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	2/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
3	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	3/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
4	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	4/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
5	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	5/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
6	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	6/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
7	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	7/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
8	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	8/1/2016	0.0	0.0	2.5	1.0	T	0.0	0.0	0.0	3.5
9	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	9/1/2016	0.0	T	T	1.4	T	0.0	0.0	0.0	1.4
10	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	10/1/2016	0.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0
11	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	11/1/2016	0.0	3.4	0.5	0.2	0.0	0.0	0.0	0.0	4.1
12	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	12/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
13	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	13/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
14	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	14/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
15	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	15/1/2016	0.0	0.0	T	0.0	0.0	0.0	0.0	0.0	T
16	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	16/1/2016	0.0	0.1	2.0	0.7	T	0.0	0.0	T	2.8
17	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	17/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
18	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	18/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
19	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	19/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
20	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	20/1/2016	0.0	9.7	0.8	0.0	0.0	0.0	0.0	0.0	10.5
21	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	21/1/2016	0.0	0.0	0.0	0.0	4.8	0.0	0.0	0.0	4.8
22	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	22/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
23	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	23/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
24	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	24/1/2016	0.0	0.0	4.0	0.0	T	T	0.0	0.0	4.0
25	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	25/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
26	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	26/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
27	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	27/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
28	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	28/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
29	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	29/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
30	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	30/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
31	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	31/1/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
32	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	1/2/2016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-

รูปที่ 9 ตัวอย่างชุดข้อมูลปริมาณน้ำฝนที่ได้รับมาจากกรมอุตุนิยมวิทยา

จะเห็นว่าชุดข้อมูลที่ได้รับมานั้น ยังไม่อยู่ในสภาพที่สามารถจะนำไปใช้ในการคำนวณ หรือนำเข้าแบบจำลองได้ ซึ่งการจะนำข้อมูลทั้งหมดที่ได้รับมานั้นไปใช้งานต่อ จำเป็นที่จะต้องทำการทำความสะอาดข้อมูล (Data cleaning) ใหม่เสียก่อน แล้วจึงนำข้อมูลที่ผ่านการทำความสะอาดแล้วไปเข้าสู่ขั้นตอนการแปลงข้อมูล (Data transformation) ซึ่งขั้นตอนนี้อาจจะไม่จำเป็นสำหรับข้อมูลปริมาณน้ำฝน แต่ขั้นตอนนี้จำเป็นสำหรับข้อมูลความกดอากาศ ความชื้นสัมพันธ์ และอุณหภูมิ ที่จำเป็นจะต้องปรับความถี่ของช่วงเวลาของข้อมูลทั้ง 4 ให้ตรงกัน และสุดท้ายคือการเชื่อมโยงข้อมูล (Combining data) เนื่องจากข้อมูลทั้ง 4 ไม่ได้อยู่ในตารางเดียวกันตั้งแต่แรก จึงจำเป็นจะต้องเชื่อมโยงข้อมูลทั้งหมดเข้าด้วยกัน โดยให้สัมพันธ์กับทั้งวันและช่วงเวลาที่ได้ทำการเก็บข้อมูลมา

ทั้งนี้ รายละเอียดในการจัดการข้อมูลโดยละเอียด สามารถอธิบายแยกย่อยได้ดังนี้

3.1.1 การจัดการค่าว่างและสัญลักษณ์เฉพาะ

สำหรับข้อมูลสภาพอากาศที่มีการเก็บเป็นรายชั่วโมงนั้น การเกิดขึ้นของค่าว่างนั้นเกิดได้จากหลายปัจจัย ในชุดข้อมูลที่เรานำมาใช้นั้นถือว่ามีค่าว่างอยู่น้อยมากๆ อีกทั้งการเก็บข้อมูลยังเป็นรายชั่วโมงจึงทำให้สามารถเติมค่าว่างได้ด้วยการหาค่าเฉลี่ยระหว่างข้อมูลก่อนหน้าและข้อมูลหลังจากค่าว่าง อย่างไรก็ตาม ในกรณีของปริมาณน้ำฝนสะสมนั้น จะมีสัญลักษณ์พิเศษเพิ่มขึ้นมาคือ สัญลักษณ์ T ซึ่งมีความหมายว่า ปริมาณน้ำฝนวัดค่าไม่ได้หรือน้อยกว่า 0.1 มิลลิเมตร ซึ่งเราได้แทนค่า T ดังกล่าวด้วยค่า 0 เนื่องจากในทางอุตุนิยมวิทยานั้น ไม่สามารถที่จะวัดปริมาณน้ำฝนที่มีค่าน้อยกว่า 0.1 มิลลิเมตรได้ เราจึงจะให้ความหมายค่า T นี้ว่าเป็นช่วงเวลาที่ฝนไม่ตก

3.1.2 การปรับหน่วยเวลาของข้อมูล

เนื่องจากคุณลักษณะแต่ละตัวมีความถี่ในการเก็บข้อมูลที่ไม่เท่ากัน และคุณลักษณะที่เราสนใจจะพยากรณ์ก็คือปริมาณน้ำฝนสะสมซึ่งคุณลักษณะดังกล่าวถูกเก็บข้อมูลที่ความถี่ทุกๆ 3 ชั่วโมง ซึ่งแตกต่างจากอีก 3 คุณลักษณะ คือ ความกดอากาศ อุณหภูมิ และความชื้นสัมพันธ์ที่มีการเก็บข้อมูลในทุกๆชั่วโมง จึงจำเป็นต้องทำให้ทุกคุณลักษณะอยู่ในความถี่เดียวกัน นั่นคือเป็นข้อมูลราย 3 ชั่วโมง ซึ่งในการปรับแก้นั้นได้ทำการหาค่าเฉลี่ยในทุกๆ 3 ชั่วโมงของแต่ละวัน เท่ากับว่า ในงานวิจัยฉบับนี้จะใช้ความถี่ 1 ช่วงเวลาเท่ากับ 3 ชั่วโมง

ความชื้นสัมพัทธ์(เปอร์เซ็นต์)
รายชั่วโมง

ที่	รหัสสถานี-สถานี-จังหวัด	วันที่	เวลาทำการตรวจ																						เฉลี่ย
			0100	0200	0300	0400	0500	0600	0700	0800	0900	1000	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000	2100	2200	2300	2400	
1	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	1/1/2016	57	55	57	59	60	62	64	60	56	53	53	50	47	43	45	43	45	49	53	54	55	59	62	66	54
2	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	2/1/2016	69	73	74	70	69	72	76	67	62	56	52	48	43	40	43	39	42	48	50	52	52	53	53	59	57
3	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	3/1/2016	50	54	66	72	71	73	73	70	61	47	43	43	40	38	36	36	36	44	51	50	50	55	59	61	53
4	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	4/1/2016	62	73	79	76	73	76	75	65	48	40	34	32	30	31	33	31	38	43	52	56	66	68	76	73	55
5	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	5/1/2016	71	71	74	79	80	75	84	83	69	37	35	34	33	31	34	32	34	42	48	51	68	72	78	72	58
6	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	6/1/2016	75	79	83	81	79	76	76	69	62	62	61	52	44	43	39	38	42	49	57	65	74	77	77	81	64
7	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	7/1/2016	81	84	85	85	85	92	94	92	80	72	66	56	49	48	49	45	44	57	67	71	73	75	77	79	71
8	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	8/1/2016	81	82	83	83	84	90	80	81	81	81	82	78	74	70	60	59	56	60	65	65	68	71	71	72	74
9	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	9/1/2016	79	81	81	78	80	81	81	78	76	80	77	62	54	53	52	49	49	62	68	76	73	73	74	76	71
10	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	10/1/2016	77	76	81	79	81	83	83	83	80	63	54	47	41	40	42	43	48	58	67	72	77	75	79	80	67
11	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	11/1/2016	77	79	78	81	81	82	84	93	83	74	66	62	52	50	43	51	50	52	56	57	66	68	68	77	68
12	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	12/1/2016	82	79	81	80	81	80	81	80	75	68	57	56	52	47	38	33	35	45	57	59	61	67	71	75	64
13	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	13/1/2016	77	78	79	80	82	82	83	84	80	71	63	57	46	48	54	53	58	63	66	70	73	80	79	77	70
14	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	14/1/2016	79	81	84	82	83	82	82	83	76	65	57	53	51	51	50	53	56	64	70	72	73	68	68	73	69
15	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	15/1/2016	76	80	83	81	79	83	82	79	72	74	59	54	50	48	49	53	58	63	70	68	68	71	71	71	68
16	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	16/1/2016	71	75	80	82	81	76	80	86	79	78	73	68	61	55	53	58	67	66	71	75	72	72	75	75	72
17	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	17/1/2016	76	73	76	79	80	80	80	76	69	64	57	57	55	54	50	49	57	65	67	74	75	73	74	77	68
18	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	18/1/2016	77	79	80	84	85	87	89	83	78	73	67	60	57	56	48	49	57	64	70	73	78	78	75	78	72
19	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	19/1/2016	78	81	79	79	81	83	83	81	73	69	63	61	59	56	56	56	62	67	72	68	67	68	71	73	70
20	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	20/1/2016	73	82	84	86	86	86	85	83	78	69	62	57	56	58	60	59	63	68	72	75	76	76	77	79	73
21	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	21/1/2016	77	77	77	78	82	82	82	82	80	73	67	69	63	64	60	57	60	64	73	76	77	76	76	77	73
22	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	22/1/2016	79	80	81	82	83	84	80	76	67	59	53	52	51	52	52	53	53	57	70	77	75	77	79	80	69
23	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	23/1/2016	81	82	82	82	83	83	89	78	72	64	59	59	57	57	54	56	58	67	72	75	78	78	78	78	72
24	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	24/1/2016	78	79	78	80	75	79	81	78	70	64	60	61	61	60	57	57	57	60	63	62	63	62	61	60	67
25	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	25/1/2016	61	58	60	59	60	61	57	58	57	55	52	53	55	53	53	52	49	54	56	57	58	59	61	61	57
26	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	26/1/2016	60	59	58	58	57	62	61	60	53	49	47	45	44	44	44	42	42	46	48	49	51	60	65	68	53
27	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	27/1/2016	70	71	71	75	74	74	75	73	63	60	53	51	47	44	43	46	46	49	54	55	59	59	58	63	60
28	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	28/1/2016	64	67	69	71	75	75	75	69	66	60	55	50	46	44	45	41	44	48	54	59	60	62	71	66	60
29	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	29/1/2016	66	69	70	70	71	73	71	64	56	48	46	45	44	42	44	43	44	50	59	65	69	72	73	77	60
30	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	30/1/2016	78	80	81	93	86	86	86	86	78	69	65	57	59	55	53	53	53	60	68	74	78	79	80	81	72
31	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	31/1/2016	81	82	83	85	86	87	86	87	91	74	63	58	50	52	51	54	58	68	77	76	78	79	78	80	74
32	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	1/2/2016	81	82	82	83	80	80	82	82	77	70	65	56	58	50	51	47	50	62	66	69	72	75	76	77	70
33	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	2/2/2016	78	79	71	78	77	78	81	81	75	74	71	66	54	55	54	51	52	54	60	59	57	58	57	59	66
34	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	3/2/2016	59	59	63	63	64	65	68	60	55	52	46	42	43	41	39	38	40	42	47	57	55	55	59	59	53
35	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	4/2/2016	59	60	63	63	71	70	74	65	54	48	47	44	43	41	40	37	39	42	50	50	53	60	57	60	54
36	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	5/2/2016	64	88	71	67	62	65	67	64	52	50	42	42	35	33	32	30	32	36	49	55	46	45	46	51	51
37	429601-สนามบินสุวรรณภูมิ จ.สมุทรปราการ	6/2/2016	55	60	55	60	61	60	55	53	51	46	45	42	39	35	31	32	33	35	37	41	38	39	40	43	45

รูปที่ 10 ตัวอย่างชุดข้อมูลความชื้นสัมพันธ์ที่ได้รับมาจากกรมอุตุนิยมวิทยา

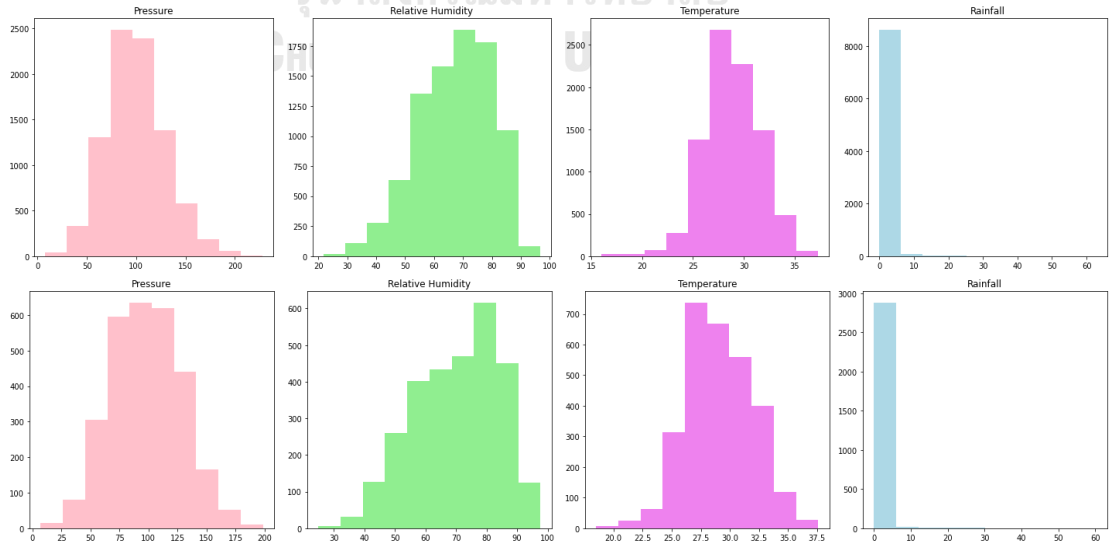
หลังจากจัดการชุดข้อมูลตามขั้นตอน และวิธีการที่ได้กล่าวไปข้างต้นนี้ ตอนนี้เราก็จะมีชุดข้อมูลเพียงชุดเดียวที่จะนำไปใช้ในการศึกษา และทำวิจัยในขั้นต่อไป

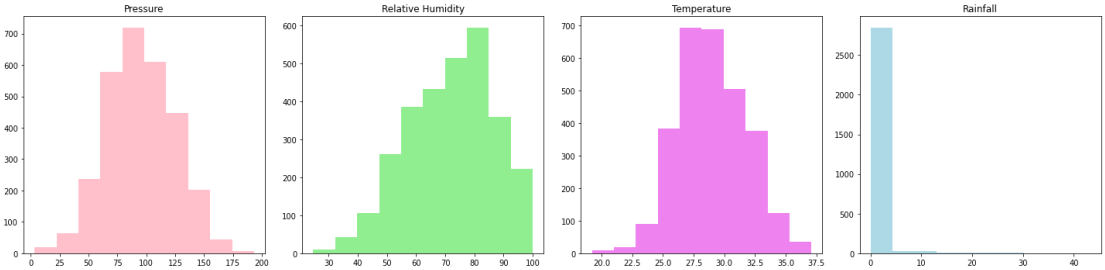
	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	25%	50%	75%	ค่าสูงสุด	ค่าต่ำสุด
อุณหภูมิ	28.92	2.78	27.0	28.8	28.8	37.6	16.0
ความชื้นสัมพันธ์	68.66	13.46	59.0	7.0.	70.0	100.0	21.7
ความกดอากาศ	98.36	29.96	77.3	97.0	97.0	227.7	3.3
ปริมาณน้ำฝนสะสม	0.34	2.23	0.0	0.0	0.0	62.8	0.0

ตาราง 1 คุณลักษณะสรุปโดยคร่าวของชุดข้อมูล

3.1.3 การแบ่งกลุ่มให้กับชุดข้อมูล

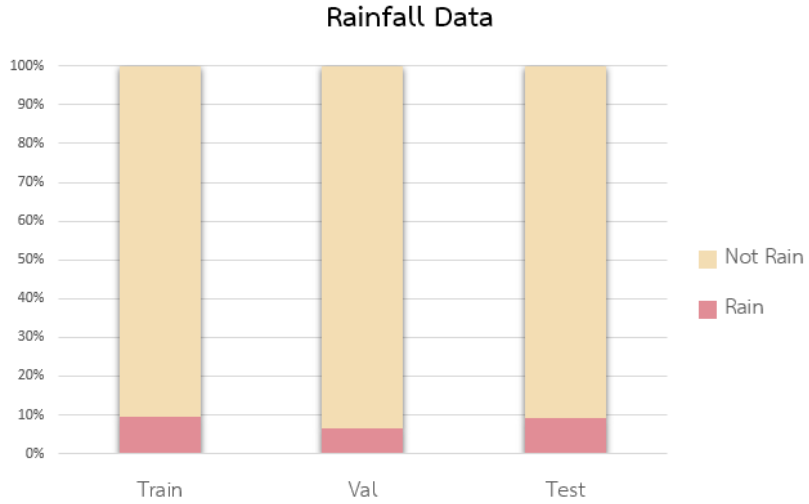
ขั้นตอนในการจัดการกับข้อมูลขั้นต่อมาที่เราจะดำเนินการก็คือการแบ่งกลุ่มให้กับชุดข้อมูลเพื่อทำการสร้างแบบจำลอง และการทดสอบ ซึ่งก่อนที่จะทำการแบ่งชุดข้อมูลออกเป็นกลุ่มๆ นั้นจำเป็นจะต้องทำการดูรูปแบบของชุดข้อมูลที่มีอยู่ก่อน เพื่อประกอบการพิจารณาในการสร้างตัวแบบ โดยหลังจากที่ปรับความถี่ของช่วงเวลาของข้อมูลทั้ง 4 คุณลักษณะให้ตรงกันแล้วจะทำให้เรามีข้อมูลทั้งหมด 14,616 จุด ซึ่งเราจะทำการแบ่งข้อมูลออกเป็น 3 ส่วน ดังนี้ 1) ข้อมูลตั้งแต่ปี 2016 ถึงปี 2018 เป็นกลุ่มฝึกสอน (Training data set) 2) ข้อมูลปี 2019 เป็นกลุ่มตรวจสอบ (Validation data set) 3) ข้อมูลปี 2020 เป็นกลุ่มทดสอบ (Test data set) หลังจากแบ่งชุดข้อมูลออกเป็น 3 กลุ่มแล้วจึงนำข้อมูลดังกล่าวไปทำการปรับปรุงโครงสร้างข้อมูล (Normalization) ก่อนที่จะนำไปเข้าแบบจำลองต่อไป





รูปที่ 11 การกระจายตัวในแต่ละคุณลักษณะของทั้ง 3 ชุดข้อมูล

จากรูปที่ 11 แสดงการกระจายตัวในแต่ละคุณลักษณะจากแต่ละกลุ่มที่เราได้ทำการแบ่งไว้ พบว่า การกระจายตัวของข้อมูลในกลุ่มฝึกสอน ตรวจสอบ และทดสอบนั้น มีความใกล้เคียงกัน อีกทั้งเมื่อเราพิจารณาแต่ละคุณลักษณะจะพบว่า ข้อมูลอุณหภูมิและความกดอากาศมีลักษณะใกล้เคียงกับการกระจายตัวแบบปกติ (Normal Distribution) ในขณะที่ข้อมูลความชื้นสัมพันธ์มีลักษณะเบ้ซ้าย (Left skewed) แต่สำหรับข้อมูลปริมาณน้ำฝนนั้น เราจะเห็นได้อย่างชัดเจนเลยว่าตัวข้อมูลนั้นมีความไม่สมดุลเลย เนื่องจากปริมาณชั่วโมงที่ฝนไม่ตกนั้นมีมากกว่าปริมาณชั่วโมงที่ฝนตก และเมื่อมาพิจารณาอย่างละเอียดก็พบว่า ปริมาณชั่วโมงที่ฝนตกนั้น (ปริมาณน้ำฝนสะสมมากกว่า 0.1 มิลลิเมตร) มีเพียงร้อยละ 10 จากข้อมูลทั้งหมด รายละเอียดสามารถแสดงได้ดังรูปที่ 12



รูปที่ 12 กราฟแสดงสัดส่วนปริมาณน้ำฝนในแต่ละชุดข้อมูล

3.1.4 การแก้ปัญหาชุดข้อมูลไม่สมดุล

จากรูปที่ 12 จะเห็นแล้วว่าสัดส่วนของจำนวนชั่วโมงที่มีฝนตกกับชั่วโมงที่ไม่มีฝนตกนั้นมีความแตกต่างกันมาก ซึ่งก่อให้เกิดปัญหาชุดข้อมูลไม่สมดุล (Imbalance Data) หากปล่อยไว้ก็อาจจะเกิดปัญหาที่ว่าตัวแบบจำลองไม่สามารถพยากรณ์ปริมาณน้ำฝนสะสมในชั่วโมงที่มีฝนตกได้ เนื่องจากการทำนายว่า ฝนไม่ตกนั้นทำให้มีโอกาสที่จะทำนายได้ถูกต้องมากกว่า หรือหากทำนาย

ชั่วโมงที่มีปริมาณน้ำฝนได้ ก็อาจจะไม่สามารถทำนายชั่วโมงที่มีปริมาณน้ำฝนสะสมในปริมาณมากได้ เพื่อป้องกันไม่ให้เกิดปัญหาดังกล่าว ในงานวิจัยนี้จึงได้เลือกใช้ Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) ซึ่งเป็นเทคนิคการทำ Over-sampling สำหรับการพัฒนาตัวแบบประเภท Regression โดยเฉพาะ เพื่อสุ่มสร้างตัวอย่างชั่วโมงที่มีฝนตก และสุ่มลดตัวอย่างชั่วโมงที่ไม่มีฝนตกเพื่อให้แบบจำลองสามารถทำนายปริมาณน้ำฝนสะสมในชั่วโมงที่มีฝนตกได้อย่างแม่นยำมากขึ้น โดยเงื่อนไขในการกำหนดกลุ่มข้อมูลส่วนน้อยคือ 1) หากปริมาณน้ำฝนสะสมในช่วง 3 ชั่วโมงมีปริมาณน้อยกว่า 0.1 มิลลิเมตร จะถือว่าเป็นข้อมูลกลุ่มส่วนมาก 2) หากปริมาณน้ำฝนสะสมในช่วง 3 ชั่วโมงมีปริมาณมากกว่าหรือเท่ากับ 0.1 มิลลิลิตร จะถือว่าเป็นข้อมูลกลุ่มส่วนน้อย

3.2 เพิ่มคุณลักษณะ และวิเคราะห์ความสัมพันธ์ระหว่างคุณลักษณะ

3.2.1 การเพิ่มข้อมูลคุณลักษณะทางสถิติ

ในการทำ Feature Engineering โดยการเพิ่มคุณลักษณะทางสถิตินั้น ในงานวิจัยนี้ได้ทำการศึกษาการเพิ่มคุณลักษณะทางสถิติให้กับชุดข้อมูลประเภทสภาพภูมิอากาศ โดยได้นำเทคนิคการเพิ่มคุณลักษณะเกี่ยวกับฤดูกาลด้วยเลขไบนารี 3 บิตมาจากงานของ (Srachoom, 2007) และได้ศึกษาการเพิ่มตัวคุณลักษณะอื่นด้วยวิธีทางสถิติจากงานวิจัยของ (Hung et al., 2009) และงานวิจัยของ (Hernández et al., 2016) อีกทั้งยังได้ศึกษาเพิ่มเติมเกี่ยวกับคุณลักษณะทางสภาพอากาศแต่ละตัวโดยละเอียดเพื่อสร้างชุดข้อมูลรับเข้าที่เหมาะสม เพื่อให้แบบจำลองสามารถเรียนรู้แนวโน้มของสภาพอากาศก่อนฝนจะตกได้ดียิ่งขึ้น

ค่าคุณลักษณะทางสถิติที่เพิ่มเข้ามาให้กับชุดข้อมูลก่อนที่จะนำเข้าแบบจำลอง เป็นการเพิ่มความสัมพันธ์ที่เกี่ยวข้องกันในเชิงเวลาให้กับแต่ละคุณลักษณะ อีกทั้งยังเพิ่มคุณลักษณะที่บ่งบอกฤดูกาล เช่น ช่วงเดือนพฤศจิกายนถึงเดือนกุมภาพันธ์จะเป็นฤดูหนาว ช่วงเดือนมีนาคมถึงเดือนมิถุนายนเป็นฤดูร้อน และช่วงเดือนกรกฎาคมถึงเดือนตุลาคมเป็นฤดูฝน ซึ่งทำให้คุณลักษณะที่เราจะนำเข้าแบบจำลองมีทั้งหมด 23 คุณลักษณะ ซึ่งคุณลักษณะที่เราทำการเพิ่มเข้ามา มีดังนี้

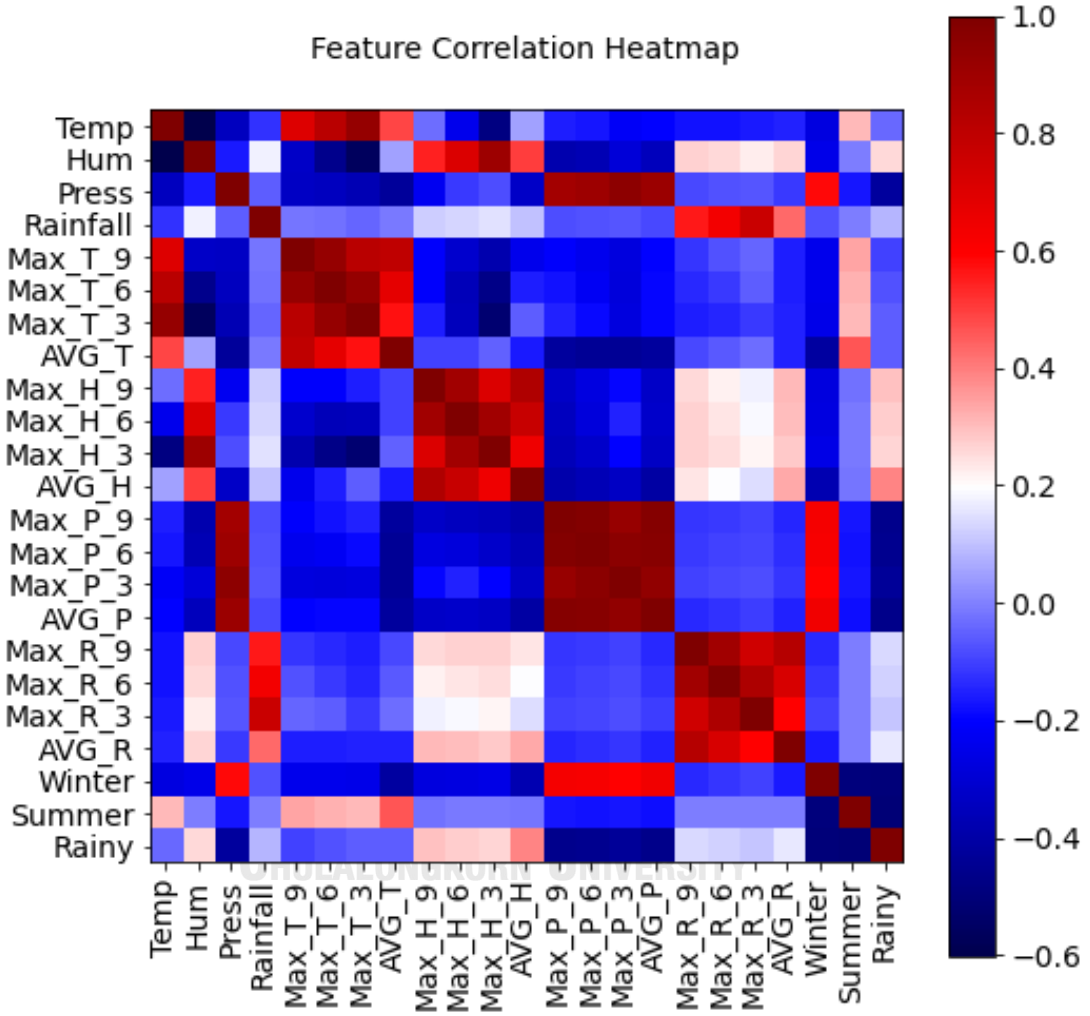
ลำดับ	คำอธิบาย	หน่วยวัด
1	ความชื้นสัมพันธ์สูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 9 ชั่วโมงก่อนหน้า (t-3)	Percent
2	ความชื้นสัมพันธ์สูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 6 ชั่วโมงก่อนหน้า (t-2)	
3	ความชื้นสัมพันธ์สูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน(t)กับ 3 ชั่วโมงก่อน	

	หน้า (t-1)	
4	ค่าเฉลี่ยของความชื้นสัมพันธ์ ณ เวลาปัจจุบัน (t) กับ 15 ชั่วโมงก่อนหน้า (t-5)	
5	ความกดอากาศสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 9 ชั่วโมงก่อนหน้า (t-3)	hPa
6	ความกดอากาศสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 6 ชั่วโมงก่อนหน้า (t-2)	
7	ความกดอากาศสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 3 ชั่วโมงก่อนหน้า (t-1)	
8	ค่าเฉลี่ยของความกดอากาศ ณ เวลาปัจจุบัน (t) กับ 15 ชั่วโมงก่อนหน้า (t-5)	
9	อุณหภูมิสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 9 ชั่วโมงก่อนหน้า (t-3)	Celsius
10	อุณหภูมิสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 9 ชั่วโมงก่อนหน้า (t-3)	
11	อุณหภูมิสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 9 ชั่วโมงก่อนหน้า (t-3)	
12	ค่าเฉลี่ยของอุณหภูมิ ณ เวลาปัจจุบัน (t) กับ 15 ชั่วโมงก่อนหน้า (t-5)	
13	ปริมาณน้ำฝนสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 9 ชั่วโมงก่อนหน้า (t-3)	Millimeter
14	ปริมาณน้ำฝนสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 6 ชั่วโมงก่อนหน้า (t-2)	
15	ปริมาณน้ำฝนสูงสุด เปรียบเทียบ ณ ช่วงเวลาปัจจุบัน (t) กับ 3 ชั่วโมงก่อนหน้า(t-1)	
16	ค่าเฉลี่ยของปริมาณน้ำฝนสะสม ณ เวลาปัจจุบัน (t) กับ 15 ชั่วโมงก่อนหน้า (t-5)	
17-19	คุณลักษณะบ่งบอกฤดูกาล ได้แก่ ฤดูหนาว ฤดูร้อน และฤดูฝน	เลขไบนารี 0,1

ตาราง 2 ตารางแสดงคุณลักษณะทั้งหมด

3.2.2 วิเคราะห์ความสัมพันธ์ระหว่างคุณลักษณะ

ในการเพิ่มประสิทธิภาพการพยากรณ์ปริมาณน้ำฝนในช่วงเวลาถัดไปให้มีความแม่นยำมากยิ่งขึ้น งานวิจัยฉบับนี้จึงได้ทำการเพิ่มคุณลักษณะบางอย่างตามที่ได้กล่าวถึงไปในหัวข้อที่แล้วเข้ามาด้วย หลังจากนั้นจึงได้ทำการวิเคราะห์หาค่าสหสัมพันธ์ (Correlation) ระหว่างคุณลักษณะ ซึ่งได้ผลลัพธ์ดังรูปต่อไปนี้



รูปที่ 13 แผนภาพแสดงค่าสหสัมพันธ์ระหว่างแต่ละคุณลักษณะ

อย่างไรก็ตาม การนำคุณลักษณะทั้งหมด 23 คุณลักษณะเข้าสู่แบบจำลองอาจจะไม่ส่งผลให้ประสิทธิภาพของแบบจำลองเพิ่มมากขึ้นกว่าเดิม เนื่องจากบางคุณลักษณะก็ไม่ได้มีความสัมพันธ์กับตัวคุณลักษณะปริมาณน้ำฝนเลย ดังนั้น เพื่อเป็นการเปรียบเทียบให้เห็นผลว่าการเพิ่มคุณลักษณะนั้นสามารถช่วยเพิ่มประสิทธิภาพให้กับแบบจำลองได้จริง จึงได้ทำการแบ่งชุดข้อมูลก่อนจะนำเข้าแบบจำลองออกเป็น 3 ชุด โดยชุดแรกจะเป็นชุดข้อมูลดั้งต้น ชุดที่สองจะเป็นชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะทางสถิติดังที่ได้กล่าวไปในหัวข้อที่แล้ว และสำหรับชุดสุดท้ายจะเป็นชุดที่ข้อมูลที่เพิ่ม

เฉพาะข้อมูลเกี่ยวกับปริมาณน้ำฝนเท่านั้น ซึ่งเป็นเทคนิคที่ถูกใช้ในงานวิจัยของ (Manokij, 2019) รายละเอียดสำหรับแต่ละชุดข้อมูลมีดังนี้

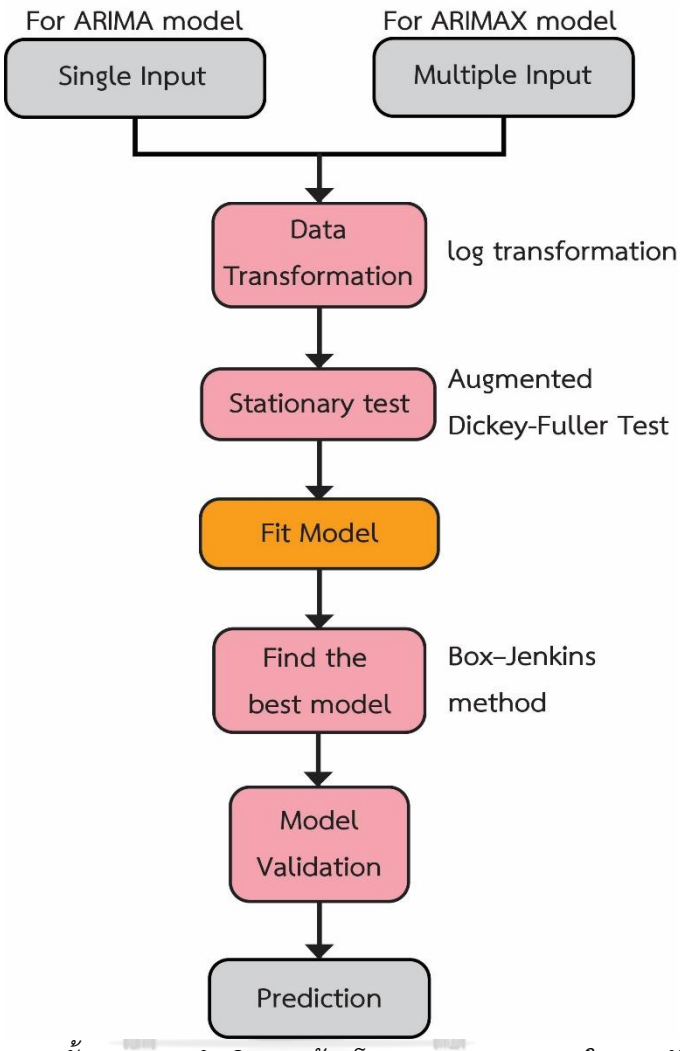
1) ชุดข้อมูลที่มีเพียง 4 คุณลักษณะดั้งเดิม ได้แก่ ปริมาณน้ำฝนสะสม อุณหภูมิ ความกดอากาศ และความชื้นสัมพันธ์
2) ชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะทั้งสิ้น 19 คุณลักษณะ และจะรวมกับคุณลักษณะดั้งเดิม ทำให้มีคุณลักษณะรวมทั้งสิ้น 23 คุณลักษณะ
3) ชุดข้อมูลที่เพิ่มเฉพาะคุณลักษณะที่มีความสัมพันธ์กับปริมาณน้ำฝนสะสมเท่านั้น ซึ่งจะมีเพียง 7 คุณลักษณะ ซึ่งทำให้ข้อมูลชุดที่ 3 จะมีคุณลักษณะทั้งหมด 11 คุณลักษณะ โดยคุณลักษณะที่เพิ่มเข้ามาคือ ปริมาณน้ำฝนสูงสุด เปรียบเทียบช่วงเวลาปัจจุบัน กับ 9 6 และ 3 ชั่วโมงก่อนหน้า ค่าเฉลี่ยในช่วง 15 ชั่วโมง และฤดูกาล

3.3 พัฒนาตัวแบบจำลองเพื่อพยากรณ์

หลังจากที่จัดการกับชุดข้อมูลเสร็จสิ้นแล้ว ขั้นต่อไปคือการพัฒนาตัวแบบจำลองเพื่อพยากรณ์ปริมาณน้ำฝนสะสมในช่วงเวลาถัดไป ซึ่งแบบจำลองที่เราจะนำมาใช้ในการพยากรณ์เปรียบเทียบประสิทธิภาพในงานวิจัยนี้จะแบ่งออกเป็น 2 กลุ่ม ได้แก่

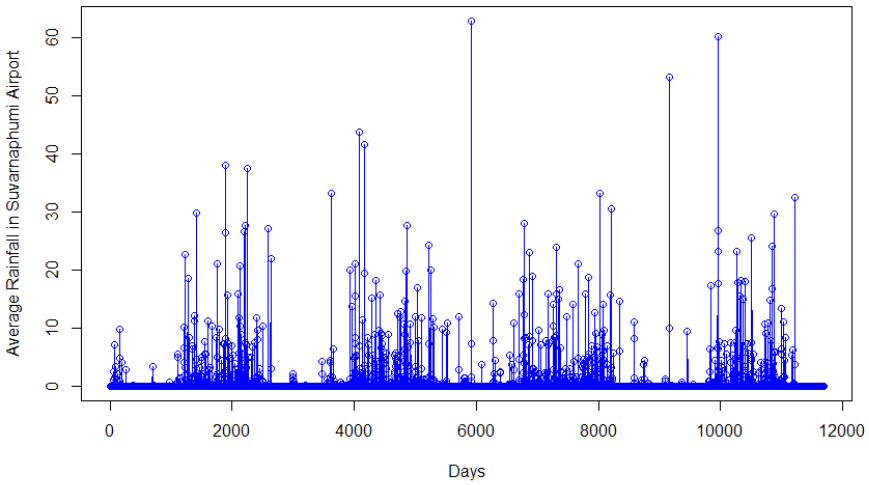
3.3.1 แบบจำลองกลุ่มอนุกรมเวลา

ในกลุ่มนี้จะประกอบไปด้วย 2 แบบจำลอง ได้แก่ แบบจำลอง ARIMA และแบบจำลอง ARIMAX ทั้ง 2 แบบจำลองจะมีขั้นตอนการพัฒนาตัวแบบและการเขียนโปรแกรมที่ค่อนข้างคล้ายกัน ต่างกันแค่เพียงตัวแบบ ARIMA จะมีข้อมูลนำเข้าเพียงข้อมูลเดียว นั่นคือข้อมูลปริมาณน้ำฝน แต่สำหรับตัวแบบ ARIMAX นั้น จะสามารถนำเข้าข้อมูลคุณลักษณะอื่นๆ ที่นอกเหนือจากปริมาณน้ำฝนเข้าแบบจำลองได้ โดยทำให้ข้อมูลคุณลักษณะอื่นๆ อยู่ในรูปแบบของเมทริกซ์ ซึ่งนั่นทำให้เราสามารถทำการทดลองชุดข้อมูลทั้ง 3 ชุดที่เรามีกับแบบจำลอง ARIMAX ได้ ซึ่งในการพัฒนาตัวแบบจำลองทั้ง 2 ตัวแบบนี้ ได้ทำการพัฒนาโดยใช้โปรแกรม RStudio Desktop Version 2022.02.0+443 สำหรับขั้นตอนการทำงานของโปรแกรมโดยคร่าว สามารถแสดงได้ดังรูปต่อไปนี้



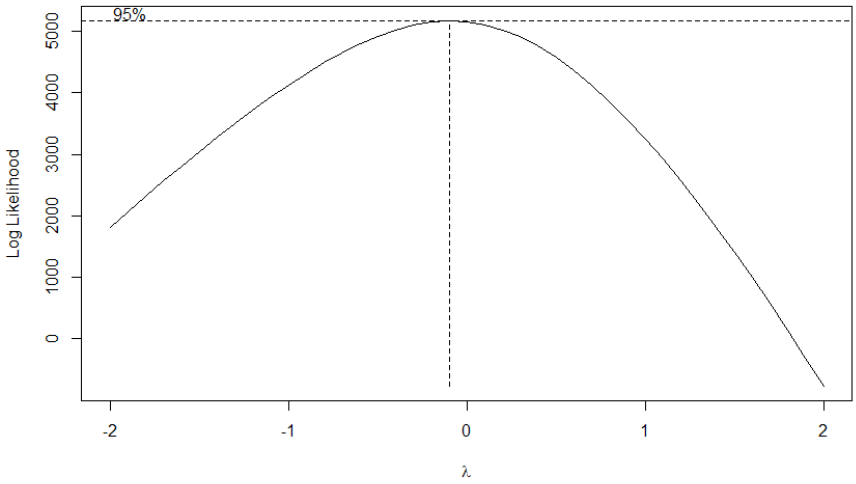
รูปที่ 14 แผนภาพแสดงขั้นตอนการดำเนินการด้วยโปรแกรม R Studio ในการพัฒนาตัวแบบอนุกรมเวลา ARIMA และ ARIMAX

จากรูปที่ 14 จะเห็นว่า ขั้นตอนการสร้างตัวแบบอนุกรมเวลาในโปรแกรม R studio นั้น จะเริ่มต้นจากการแปลงข้อมูลที่มีให้อยู่ในรูปแบบของข้อมูลอนุกรมเวลาเสียก่อน



รูปที่ 15 ตัวอย่างกราฟ Time plot ของข้อมูลปริมาณน้ำฝนสะสม

จากนั้นจึงตามด้วยการตรวจสอบตัวข้อมูลว่าเกิดปัญหาใดบ้าง ตัวอย่างเช่นรูปที่ 15 ที่แสดงข้อมูลปริมาณน้ำฝนสะสมในรูปแบบของ Time plot ซึ่งจะพบว่า ข้อมูลดังกล่าวประสบปัญหาเรื่องการกระจายตัวของค่าความแปรปรวนที่ไม่คงที่ (Constant variance violate) อีกทั้งยังมีข้อมูลปริมาณน้ำฝนสะสมบางช่วงที่สูงมากกว่าตัวอื่น (Error) ซึ่งปัญหาดังกล่าวนี้สามารถแก้ไขได้โดยการเปลี่ยนแปลงข้อมูล (Data Transformations) โดยการเปลี่ยนแปลงข้อมูลนั้น จะต้องเลือกวิธีให้เหมาะสมกับปัญหาของข้อมูล ซึ่งในที่นี้ได้เลือกใช้วิธี Box-Cox Transformations ซึ่งได้ผลสรุปออกมาว่า ควรทำการเปลี่ยนแปลงข้อมูลด้วยวิธี logarithm



รูปที่ 16 การใช้ Box-Cox Transformation เพื่อหาวิธีที่เหมาะสมในการแปลงข้อมูล

หลังจากได้ทำการแปลงข้อมูลด้วยวิธี logarithm แล้ว ขั้นตอนต่อไปจึงจะทำการทดสอบคุณสมบัติความนิ่ง (Stationary) ของข้อมูล ด้วยวิธี Augmented Dickey and Fuller (ADF) เพื่อเตรียมพร้อมครั้งสุดท้ายก่อนที่จะนำข้อมูลไปทำการสร้างตัวแบบ โดยที่เกณฑ์ในการตรวจสอบจะใช้

ค่า p-value กับระดับนัยสำคัญทางสถิติที่ α = 0.05 ซึ่งจากรูปที่ 3.9 จะพบว่า ค่า p-value = 0.01 ซึ่งน้อยกว่า α = 0.05 กล่าวคือ ข้อมูลปริมาณน้ำฝนสะสมนี้มีคุณสมบัติความนิ่งแล้ว สุดท้ายเราจะนำข้อมูลดังกล่าวไปทำการสร้างตัวแบบ ARIMA พร้อมทั้งคัดเลือกตัวแบบที่เหมาะสมที่สุดที่จะนำมาใช้ในการพยากรณ์ด้วยวิธี Box-Jenkins Method ซึ่งจะเป็นการพิจารณาร่วมระหว่าง ค่าเกณฑ์สารสนเทศอะกะอิเกะ (Akaike information criterion : AIC) และค่าเกณฑ์สารสนเทศเบส์เซียน (Bayesian information criterion : BIC)

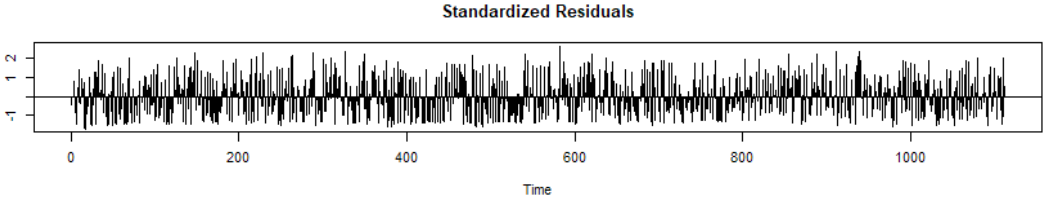
```
        Augmented Dickey-Fuller Test

data:  sa_log
Dickey-Fuller = -4.6344, Lag order = 16, p-value = 0.01
alternative hypothesis: stationary
```

รูปที่ 17 การใช้สถิติทดสอบ ADF เพื่อทดสอบคุณสมบัติความนิ่ง (Stationary) ของข้อมูล

หลังจากทำการสร้างและคัดเลือกตัวแบบด้วยวิธี Box-Jenkins Method ผู้วิจัยได้ทำการตัดสินใจเลือกตัวแบบ ARIMA(0,1,2) มาใช้ในการพยากรณ์

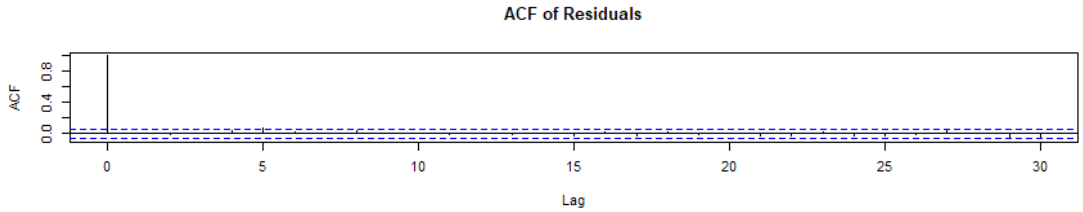
หลังจากที่ได้ตัวแบบแล้ว จึงจะนำตัวแบบดังกล่าวไปประเมินความเหมาะสมของตัวแบบ (Model Validation) ก่อนที่จะนำไปพยากรณ์ ซึ่งวิธีที่ใช้ทดสอบจะมีดังต่อไปนี้

- Shapiro-Wilk normality test เพื่อทดสอบคุณสมบัติการแจกแจงปกติในการวิเคราะห์ค่าความคลาดเคลื่อน (Residual analysis) พบว่า ค่า p-value มีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 ส่งผลให้ไม่เกิดการปฏิเสธสมมติฐานว่าง กล่าวคือ ค่าความคลาดเคลื่อนนั้นมีการแจกแจงแบบปกติที่ระดับนัยสำคัญ 0.05
- คุณสมบัติ White noise ของค่าคลาดเคลื่อน เป็นการทดสอบค่าเฉลี่ยของค่าความคลาดเคลื่อนว่ามีค่าเท่ากับศูนย์หรือไม่ และความแปรปรวนมีความคงที่หรือไม่ ซึ่งค่ามาตรฐานของความคลาดเคลื่อน (Standardized residuals) ทั้งหมดจะต้องอยู่ในช่วงระหว่างค่า -3 ถึง 3 และมีการแกว่งของจุดที่พล็อตรอบจุดศูนย์ ซึ่งจากรูปที่ 3.10 พบว่ากราฟที่ได้ ไม่มีค่าใดที่มีค่าเกินช่วง -3 ถึง 3 เลย



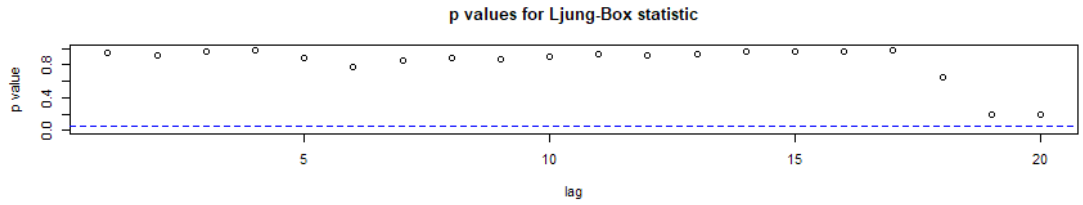
รูปที่ 18 กราฟทดสอบคุณสมบัติ White noise ของค่าคลาดเคลื่อน

- คุณสมบัติความเป็นอิสระต่อกัน (Independence) ของค่าความคลาดเคลื่อน เป็นการทดสอบความเป็นอิสระต่อกันของค่าความคลาดเคลื่อน ซึ่งจะพิจารณาจากกราฟของ ACF of residuals โดยลักษณะของกราฟจะต้องมีค่าอยู่ในช่วงที่ยอมรับได้ (Acceptable range) ทั้งนี้ เมื่อสังเกตในรูปที่ 19 พบว่ากราฟของ ACF of residuals มีค่าอยู่ภายใต้ขอบเขตที่กำหนดและอยู่ในช่วงที่ยอมรับได้ (แนวเส้นประสีน้ำเงิน) ทำให้ได้ข้อสรุปว่าค่าคลาดเคลื่อนที่ได้จากตัวแบบ ARIMA(0,1,2) มีความเป็นอิสระต่อกัน



รูปที่ 19 กราฟของ ACF of residuals

- คุณสมบัติความเป็นลักษณะสุ่ม (Randomness) ของค่าความคลาดเคลื่อน การทดสอบสุดท้ายจะเป็นการทดสอบความเป็นลักษณะสุ่มของค่าความคลาดเคลื่อนจากแนวคิดของการทดสอบ Ljung-Box-Pierce Q-Statistics โดยเป็นการสังเกตค่า P-value ซึ่งต้องมีค่ามากกว่าค่านัยสำคัญ 0.05 และจากผลลัพธ์ในรูปที่ 20 พบว่าตำแหน่งของจุดทุกจุดมีค่า P-value มากกว่า 0.05 ส่งผลให้เกิดการไม่ปฏิเสธสมมติฐานว่างของการทดสอบ กล่าวคือ ค่าความคลาดเคลื่อนจากตัวแบบดังกล่าวมีความเป็นลักษณะสุ่มที่ระดับนัยสำคัญ 0.05



รูปที่ 20 การทดสอบ Ljung-Box-Pierce Q-Statistics

จากผลการทดสอบความเหมาะสมของตัวแบบ ARIMA ที่จะใช้ในการพยากรณ์ปริมาณน้ำฝนที่ได้ทำการแสดงตัวอย่างการดำเนินการไปข้างต้น ได้ข้อสรุปว่า ตัวแบบ ARIMA(0,1,2) มีความเหมาะสมในการนำไปใช้พยากรณ์ปริมาณน้ำฝนสะสม กรณีศึกษาพื้นที่สนามบินสุวรรณภูมิ

สำหรับตัวแบบอนุกรมเวลา ARIMAX นั้นก็จะมีวิธีการดำเนินการที่คล้ายคลึงกัน เพียงแต่เราจะเปลี่ยนรูปแบบการนำเข้าข้อมูลอนุกรมเวลาให้อยู่ในรูปแบบของเมทริกซ์ เพื่อให้สามารถนำข้อมูลคุณสมบัติอื่นๆ ที่มีความเกี่ยวข้องกับการพยากรณ์ปริมาณน้ำฝนเข้าไปในแบบจำลองเพื่อทำการสร้าง

แบบจำลอง ARIMAX ดังที่กล่าวไปข้างต้นว่า แบบจำลอง ARIMAX นั้นสามารถนำเข้าคุณสมบัติอื่นๆ ไปเป็นตัวแปรต้น นอกเหนือจากตัวแปรตาม (ซึ่งก็คือปริมาณน้ำฝนสะสม) ในการคัดเลือกตัวแบบที่ดีที่สุดสำหรับแต่ละชุดข้อมูลก็ได้ใช้วิธี Box-Jenkins Method เช่นเดียวกับตอนที่สร้างตัวแบบจำลอง ARIMA ก่อนที่จะนำตัวแบบที่ได้ไปทำการประเมินความเหมาะสมของตัวแบบ ซึ่งได้ผลลัพธ์ออกมาดังต่อไปนี้

- ตัวแบบที่เหมาะสมที่สุดสำหรับชุดข้อมูลชุดที่ 1 คือ ARIMAX (1,1,3)
- ตัวแบบที่เหมาะสมที่สุดสำหรับชุดข้อมูลชุดที่ 2 คือ ARIMAX (4,1,3)
- ตัวแบบที่เหมาะสมที่สุดสำหรับชุดข้อมูลชุดที่ 3 คือ ARIMAX (1,1,3)

3.3.2 แบบจำลองกลุ่มโครงข่ายระบบประสาทแบบย้อนกลับ

ในกลุ่มนี้จะประกอบไปด้วย 3 แบบจำลอง ได้แก่ แบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) โครงข่ายระบบประสาทแบบย้อนกลับ (RNN) และโครงข่ายประตูวกกลับ (GRU) ซึ่งหลังจากที่เราจัดการกับข้อมูลเบื้องต้นเสร็จแล้ว (ในหัวข้อ 3.1 และ 3.2) เราจะทำการการปรับปรุงโครงสร้างข้อมูล (Normalization) ซึ่งจากรูปที่ 11 เราจะเห็นแล้วว่าชุดข้อมูลในแต่ละคุณลักษณะมีการกระจายตัวที่แตกต่างกัน ดังนั้น จึงจำเป็นต้องปรับปรุงโครงสร้างของข้อมูลเพื่อลดความซับซ้อนและหลีกเลี่ยงความผิดปกติของข้อมูลก่อนที่จะนำเข้าไปในแบบจำลอง โดยจะทำการปรับปรุงโครงสร้างข้อมูลด้วยวิธี Z-Score Normalization หรือ Standardization สามารถแสดงสมการได้ดังนี้

$$x' = \frac{x - \mu}{\sigma}$$

โดยที่

x คือ ค่าของข้อมูลที่ต้องการปรับ

μ คือ ค่าเฉลี่ยของชุดข้อมูล

σ คือ ค่าเบี่ยงเบนมาตรฐานของชุดข้อมูล

x' คือ ค่าของข้อมูลที่ผ่านการปรับแล้ว

	Temp	Hum	Press	Rainfall
0	-1.281799	-0.425714	2.586319	-0.154162
1	-1.354227	-0.796870	2.309166	-0.154162
2	-1.897437	-0.402023	2.676478	-0.154162
3	-1.462869	-0.852149	3.421119	-0.154162
4	-0.304022	-1.349656	3.130609	-0.154162
5	0.456472	-1.847164	2.028673	-0.154162
6	-0.050524	-1.428626	1.851695	-0.154162
7	-0.847231	-0.875840	2.509518	-0.154162
8	-1.173157	-0.536271	2.329201	-0.154162
9	-2.150935	0.411363	2.229025	-0.154162

รูปที่ 21 ตัวอย่างข้อมูลที่ผ่านการทำ Standardization

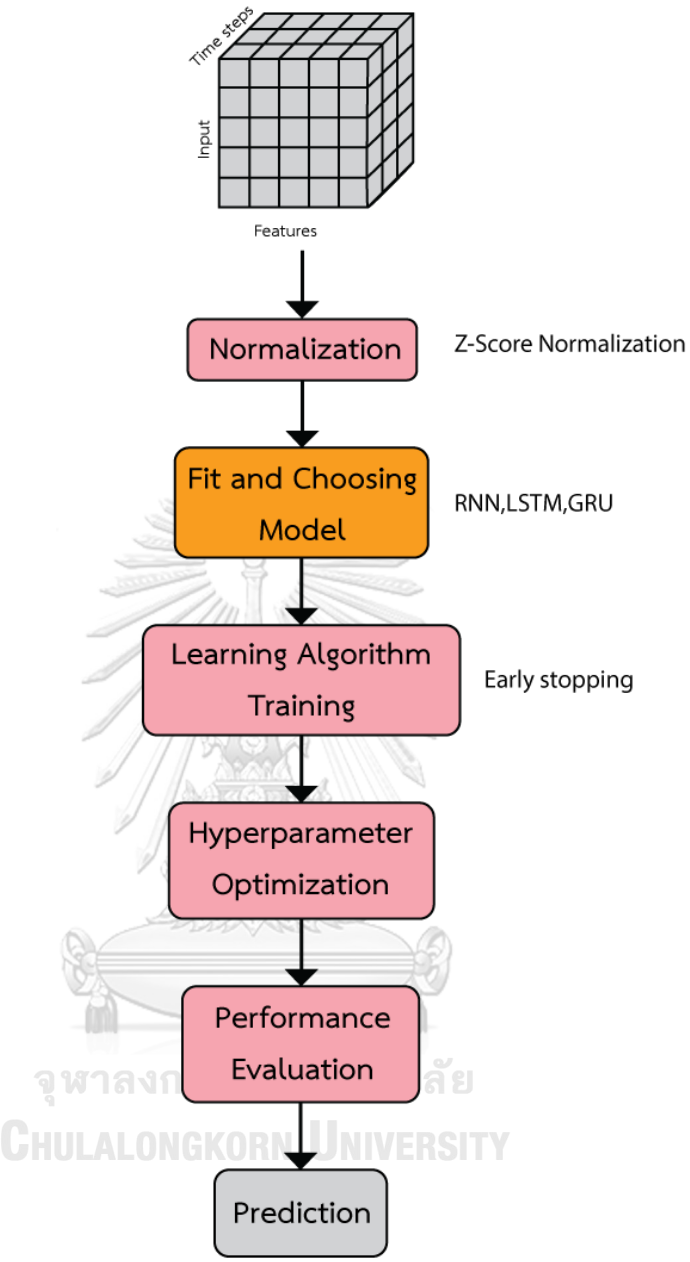
หลังจากนั้นจึงนำข้อมูลทั้งหมดเข้าแบบจำลองเพื่อทำการฝึกสอนโดยจะทำการแปลงข้อมูลนำเข้าให้อยู่ในรูปแบบอาร์เรย์ 3 มิติ คือ [Input, Timestep, Features] โดยที่ Input คือจำนวนข้อมูลที่เราจะนำเข้าแบบจำลอง Timestep คือ จำนวนช่วงเวลาก่อนหน้า (t+n) ซึ่งในที่นี้เราได้ใช้ทั้งหมด 80 ช่วงเวลา (t+80) หรือ 240 ชั่วโมงก่อนหน้า (10 วัน) เพื่อทำนายปริมาณน้ำฝนสะสมในอีกช่วงเวลาถัดไป

พารามิเตอร์ที่ใช้ในการฝึกสอน ได้ทำการทดสอบทั้งหมด 500 Epochs และมีการใช้ Adaptive Moment Estimation (ADAM) เป็นตัว Optimizer เพื่อเพิ่มประสิทธิภาพในการทำนายผลของแบบจำลอง ในการวัดประสิทธิภาพของตัวแบบ ได้ทำการเลือก ฟังก์ชันสูญเสีย (Loss function) คือ Mean Absolute Error (MAE) ซึ่งจะมีความอ่อนไหวกับชุดข้อมูลที่มีค่าผิดปกติ (Error) น้อยกว่า Mean Square Error (MSE) และ Root Mean Square Error (RMSE) เนื่องจากเป็นการนำค่าความคลาดเคลื่อน (Error) มาใส่ สัมบูรณ์ (Absolute) เหตุผลที่เลือก MAE เนื่องจากข้อมูลปริมาณน้ำฝนสะสมนั้นจะมีความเอนเอียง เนื่องจากช่วงเวลาที่ฝนตกมีน้อยกว่าช่วยเวลาที่ฝนไม่ตก (ดังที่แสดงในรูปที่ 12) อีกทั้ง ปริมาณน้ำฝนสะสมในบางช่วงก็มีปริมาณที่สูงมากกว่าปกติ ซึ่งอาจจะเกิดจากอิทธิพลของพายุ หรือปัจจัยอื่นๆ และข้อมูลเหล่านี้เองที่กลายเป็นค่าผิดปกติ (Error) ในชุดข้อมูล การเลือกใช้ MAE ในการประเมินประสิทธิภาพของตัวแบบจึงมีความเหมาะสมกับชุดข้อมูลนี้มากกว่าค่าอื่นๆ ทั้งนี้ สมการของ Mean Absolute Error (MAE) สามารถแสดงได้ดังนี้

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

สำหรับเทคนิคการฝึกสอน ได้เลือกใช้เทคนิค Early stopping ซึ่งเป็นเทคนิคในการสร้างความสมดุลระหว่างค่าความเอนเอียง (Bias) และค่าความแปรปรวน (Variation) เพื่อไม่ให้แบบจำลองประสบปัญหา Overfitting (แบบจำลองอิงกับข้อมูลในชุดข้อมูลฝึกสอนมากเกินไป คือมีค่าความถูกต้องในการทำนายข้อมูลฝึกสอนสูง แต่ค่าความถูกต้องในการทำนายข้อมูลชุดทดสอบต่ำ) โดยการสังเกตค่า Validation loss ในทุกๆ Epochs ด้วยเงื่อนไขว่า หากค่าดังกล่าวไม่ลดลงติดต่อกัน 10 Epochs ให้หยุดการฝึกสอน

สำหรับแบบจำลองในกลุ่มโครงข่ายระบบประสาทแบบย้อนกลับได้ทำการพัฒนาโดยใช้ภาษา Python ร่วมกับโปรแกรม Google Colab โดยจะใช้ TensorFlow Library เพื่อเรียกฟังก์ชันที่ใช้ในการสร้างแบบจำลองการเรียนรู้เชิงลึกที่สนใจจะศึกษา ขั้นตอนการทำงานของโปรแกรมโดยคร่าวสามารถแสดงได้ดังรูปต่อไปนี้



รูปที่ 22 แผนภาพแสดงกระบวนการทำงานของแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับ

ทั้งนี้ สำหรับเนื้อหาเกี่ยวกับผลการทดลอง หรือการปรับจูนไฮเปอร์พารามิเตอร์ ซึ่งเกี่ยวข้องกับหัวข้อการเปรียบเทียบผลการพยากรณ์ และประเมินความเหมาะสมของตัวแบบเราจะอภิปรายหัวข้อดังกล่าวโดยละเอียดในบทที่ 4 และ หัวข้อสรุปผลจะอภิปรายในบทที่ 5

บทที่ 4
ผลการวิจัย

เนื้อหาในบทนี้จะกล่าวถึงผลการทดลองหรือผลการพยากรณ์ที่ได้จากตัวแบบที่เลือก รวมไปถึงการปรับจูนค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมกับตัวแบบโครงงข่ายระบบประสาทย้อนกลับแต่ละตัวเพื่อให้การพยากรณ์มีความแม่นยำ และประสิทธิภาพสูงสุด ก่อนที่จะทำการวิเคราะห์เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมล่วงหน้าในระยะสั้นร่วมกับการใช้ข้อมูลหลายตัวแปร ใน 2 ประเด็นหลักคือ

1) การเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบที่ใช้ชุดข้อมูลที่แตกต่างกันระหว่างชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะพิเศษกับชุดข้อมูลดั้งเดิม

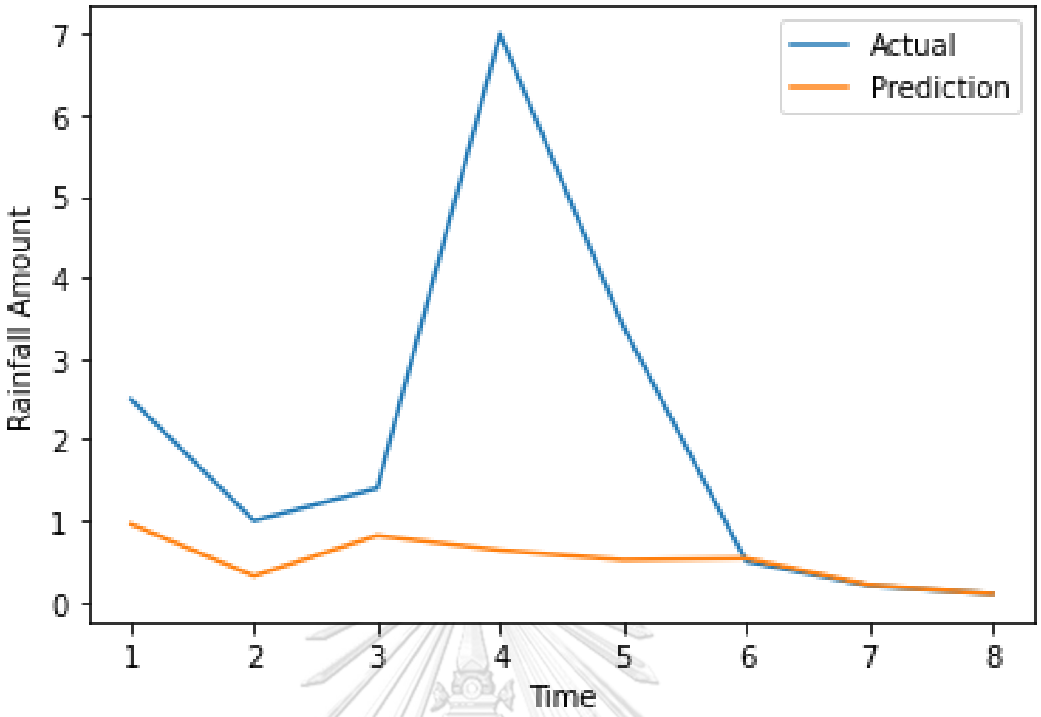
2) เปรียบเทียบประสิทธิภาพในการพยากรณ์ในช่วงเวลาถัดไปของโครงข่ายระบบประสาทแบบย้อนกลับที่สนใจศึกษากับแบบจำลองอาริมา

4.1 ผลการทดลองที่ได้จากตัวแบบอาริมา

ตามที่ได้อภิปรายขั้นตอนและวิธีการในการพัฒนาตัวแบบอนุกรมเวลาอาริมาไปในบทที่ 3 ในส่วนนี้จะทำการอธิบายผลการพยากรณ์ที่ได้จากตัวแบบที่ทำการพัฒนาขึ้นมา ซึ่งได้แก่ ผลการพยากรณ์ปริมาณน้ำฝนสะสมในอีก 1 ช่วงเวลาถัดไป (t+1) และผลการพยากรณ์ปริมาณน้ำในสะสมในอีก 8 ช่วงเวลาถัดไป (t+8)

4.1.1 ผลการพยากรณ์ด้วยตัวแบบ ARIMA

แบบจำลอง ARIMA ถือเป็นแบบจำลองเบื้องต้นพื้นฐานในการพยากรณ์ข้อมูลอนุกรมเวลา โดยแบบจำลอง ARIMA เป็นการพิจารณาโดยใช้ข้อมูลปริมาณน้ำฝนเพียงอย่างเดียวในการนำเข้าแบบจำลองเพื่อทำนายปริมาณน้ำฝนสะสมในช่วงเวลาถัดไป ในงานวิจัยนี้ได้ทำการคัดเลือกตัวแบบ ARIMA ที่ดีที่สุดด้วยวิธีบอกซ์แอนด์เจนกินส์ ซึ่งตัวแบบที่ให้ค่าความคลาดเคลื่อนน้อยที่สุด คือตัวแบบ ARIMA(0,1,2) ซึ่งเมื่อได้ตัวแบบแล้วก็จะนำตัวแบบดังกล่าวไปพยากรณ์ปริมาณน้ำฝนสะสมในอีก 8 ช่วงเวลาถัดไป โดยจะแสดงผลการเปรียบเทียบความแม่นยำในการพยากรณ์ระหว่าง ปริมาณน้ำฝนสะสมที่เกิดขึ้นจริง กับ ปริมาณน้ำฝนสะสมที่ได้จากการพยากรณ์ด้วยตัวแบบ ด้วยกราฟดังต่อไปนี้



รูปที่ 23 กราฟเปรียบเทียบระหว่างปริมาณน้ำฝนสะสมที่เกิดขึ้นจริงกับปริมาณน้ำฝนสะสมที่ได้จากการพยากรณ์ด้วยตัวแบบ ARIMA(1,1,3)

จากรูปที่ 23 จะเห็นว่า เนื่องจากชั่วโมงที่ฝนตกมีน้อย อีกทั้งปริมาณน้ำฝนสะสมในแต่ละชั่วโมงก็ไม่ได้มีค่าสูงมาก ทำให้ตัวแบบ ARIMA ไม่สามารถพยากรณ์ชั่วโมงที่มีปริมาณน้ำฝนสะสมในปริมาณมากได้ ค่าที่ได้จึงเหมือนเป็นค่าเฉลี่ยปกติของปริมาณน้ำฝนในช่วงเวลาปกติ เพราะเมื่อเราพิจารณาในช่วงเวลาที่ 7 และ 8 ที่มีปริมาณน้ำฝนสะสมอยู่ในเกณฑ์ปกติ จะพบว่าแบบจำลองสามารถทำนายออกมาใกล้เคียงกับปริมาณน้ำฝนสะสมจริงในช่วงเวลานั้น

ต่อมาเราจะทำการพิจารณา Mean Absolute Error (MAE) ที่ได้จากผลการทดลองทั้ง 2 ช่วงเวลา ซึ่งสามารถแจกแจงได้ดังนี้

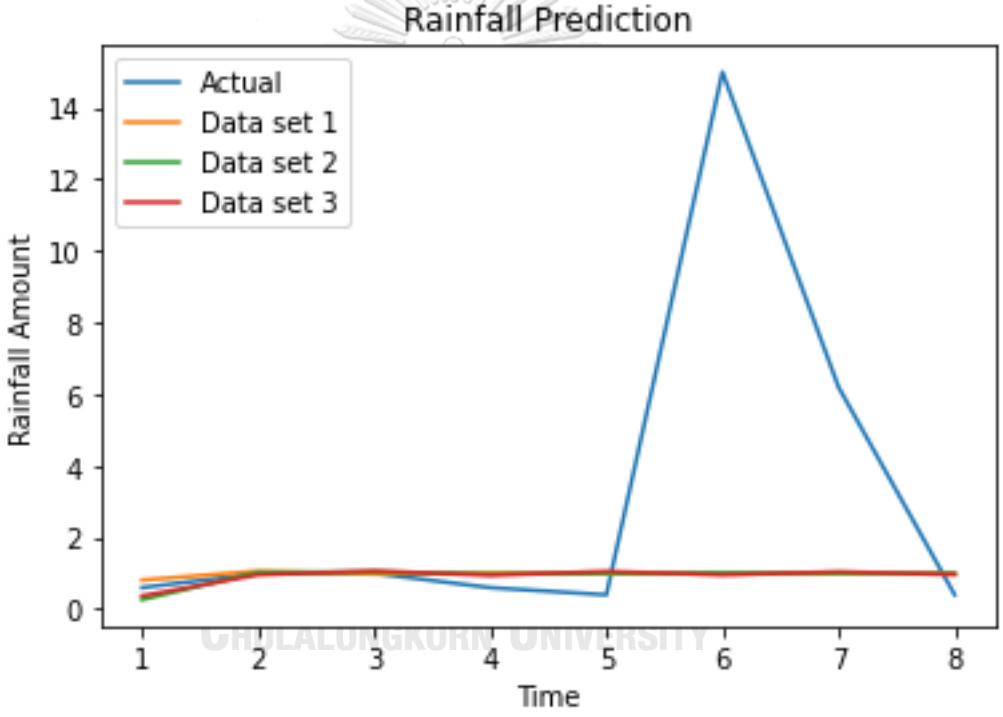
ตัวแบบที่ได้	Mean Absolute Error (MAE)	
	t+1	t+8
ARIMA (1,1,3)	2.2082	6.9857

ตาราง 3 แสดงค่า MAE ที่ได้จากการพยากรณ์ด้วยตัวแบบ ARIMA

จากตาราง 3 จะพบว่า ค่า MAE ในช่วงเวลาที่ t+1 มีค่าความคลาดเคลื่อนที่สูงกว่าค่า MAE ในช่วงเวลาที่ t+8 ทั้งนี้ เนื่องจากในช่วงเวลาจริงๆ ที่ t+1 มีปริมาณน้ำฝนสะสมในปริมาณมากระดับหนึ่ง แต่แบบจำลอง ARIMA ไม่สามารถทำนายค่าได้ (ดังกราฟที่แสดงไปในรูปที่ 23) ส่งผลให้ค่าความคลาดเคลื่อนในช่วงเวลาแรกมีค่าสูงกว่าค่าความคลาดเคลื่อนในอีก 8 ช่วงเวลาถัดไป

4.1.2 ผลการพยากรณ์ที่ได้จากตัวแบบ ARIMAX

แบบจำลอง ARIMAX เป็นแบบจำลองอนุกรมเวลาที่พัฒนาต่อยอดมาจากแบบจำลอง ARIMA เพื่อแก้ไขปัญหาที่ไม่สามารถนำข้อมูลคุณลักษณะอื่นๆ ที่เกี่ยวข้องกับคุณลักษณะที่ต้องการพยากรณ์เข้าไปในแบบจำลองได้ เป็นการพิจารณาโดยใช้ข้อมูลคุณลักษณะอื่นๆ จากแต่ละชุดข้อมูลร่วมกับข้อมูลปริมาณน้ำฝนในการนำเข้าแบบจำลองเพื่อทำนายปริมาณน้ำฝนสะสมในช่วงเวลาถัดไป ในงานวิจัยนี้ได้ทำการคัดเลือกตัวแบบ ARIMAX ที่ดีที่สุดด้วยวิธีบอกซ์แอนด์เจนกินส์ ซึ่งเมื่อได้ตัวแบบแล้วก็จะนำตัวแบบดังกล่าวไปพยากรณ์ปริมาณน้ำฝนสะสมในอีก 8 ช่วงเวลาถัดไป โดยจะแสดงผลการพยากรณ์ดังกล่าวระหว่าง ปริมาณน้ำฝนสะสมที่เกิดขึ้นจริง กับ ปริมาณน้ำฝนสะสมที่ได้จากการพยากรณ์ด้วยตัวแบบ ด้วยกราฟสำหรับข้อมูลแต่ละชุดดังต่อไปนี้



รูปที่ 24 กราฟเปรียบเทียบระหว่างปริมาณน้ำฝนสะสมที่เกิดขึ้นจริงกับปริมาณน้ำฝนสะสมที่ได้จากการพยากรณ์ด้วยตัวแบบ ARIMAX ทั้ง 3 ตัวแบบ

จากรูปที่ 24 จะเห็นว่า ตัวแบบ ARIMAX ทั้ง 3 ตัวประสบปัญหาเดียวกันกับตัวแบบ ARIMA นั่นคือ ไม่สามารถตรวจวัดหรือพยากรณ์ปริมาณน้ำฝนสะสมที่มีค่าสูงได้ แต่สำหรับค่าที่ต่ำๆ ที่เป็นค่าเฉลี่ยของปริมาณน้ำฝนสะสมส่วนมากยังคงพอทำได้ดีอยู่

ต่อมาจะพิจารณาค่า Mean Absolute Error (MAE) ที่ได้จากผลการทดลองทั้ง 2 ช่วงเวลา เพื่อประกอบการพิจารณาร่วมว่าตัวแบบในชุดข้อมูลใด จะให้ค่าความคลาดเคลื่อนน้อยที่สุด ดังนี้

ชุดข้อมูล	ตัวแบบ ARIMAX	Mean Absolute Error (MAE)	
		t+1	t+8
ข้อมูลชุดที่ 1	ARIMAX (1,1,3)	1.1053	2.3084
ข้อมูลชุดที่ 2	ARIMAX (4,1,3)	0.3284	1.2904
ข้อมูลชุดที่ 3	ARIMAX (1,1,3)	0.2568	1.2549

ตาราง 4 แสดงค่า MAE ที่ได้จากการพยากรณ์ด้วยตัวแบบ ARIMAX

จากรูปที่ 24 และตารางแสดงค่า MAE ตารางที่ 4 พบว่า ในบรรดาตัวแบบ ARIMAX ทั้ง 3 ตัวแบบที่ได้ทำการพัฒนาขึ้นโดยใช้ชุดข้อมูลที่แตกต่างกันเพื่อวัดประสิทธิภาพในการพยากรณ์ของตัวแบบ พบว่า ชุดข้อมูลชุดที่ 3 ที่มีการเพิ่มค่าคุณลักษณะทางสถิติที่เกี่ยวข้องกับปริมาณน้ำฝน ลงไปในการช่วยพยากรณ์ ให้ผลการพยากรณ์ที่มีความใกล้เคียงค่าปริมาณน้ำฝนสะสมที่เกิดขึ้นจริงมากที่สุด รองมาคือตัวแบบของชุดข้อมูลชุดที่ 2 ที่เพิ่มคุณลักษณะทางสถิติทั้งหมด 23 คุณลักษณะ และสุดท้ายคือตัวแบบของชุดข้อมูลดั้งเดิมที่มีประสิทธิภาพน้อยกว่าอีก 2 ชุดที่ผ่านการเพิ่มคุณลักษณะพิเศษ

4.2 ผลการทดลองที่ได้จากแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับ (RNNs)

สำหรับขั้นตอนในการพัฒนาแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับนั้น เบื้องต้นได้ทำการอธิบายไว้ในบทที่ 3 แล้ว ดังนั้นในหัวข้อนี้จึงจะกล่าวถึงการปรับจูนค่าไฮเปอร์พารามิเตอร์ (Hyper-Parameter Tuning) และผลการพยากรณ์ที่ได้จากแต่ละแบบจำลอง ดังนี้

4.2.1 การพยากรณ์ด้วยโครงข่ายระบบประสาทแบบย้อนกลับ (Recurrent Neural Network : RNN)

โครงข่ายระบบประสาทแบบย้อนกลับ (RNN) เป็นแบบจำลองการเรียนรู้เชิงลึกตัวแรกๆ ที่ออกแบบมาเพื่อแก้ปัญหาเกี่ยวกับข้อมูลประเภทอนุกรมเวลา โดยใช้การข้อมูลที่ช่วงเวลาก่อนหน้า มารวมเข้ากับข้อมูลตัวใหม่(Input data) เพื่อใช้ในการทำนายหรือพยากรณ์ ซึ่งหลังจากทำการเตรียมข้อมูล และพัฒนาตัวแบบ RNN ด้วยวิธีการที่อธิบายเอาไว้ในบทที่ 3 ก่อนจะนำมาทำการปรับจูน พบว่า การปรับจูนไฮเปอร์พารามิเตอร์ที่มีความเหมาะสม และให้ประสิทธิภาพในการพยากรณ์ที่ดีที่สุด สามารถชี้แจงได้ดังตารางต่อไปนี้

แบบจำลอง	ชั้นของแบบจำลอง (Layer)	ไฮเปอร์พารามิเตอร์ (Hyper-Parameter)	ฟังก์ชันกระตุ้น (Activation Function)
RNN	SimpleRNN layer 1	32	Tanh
	SimpleRNN layer 2	64	
	Dense	32	
	Reshape	[output, number of features]	

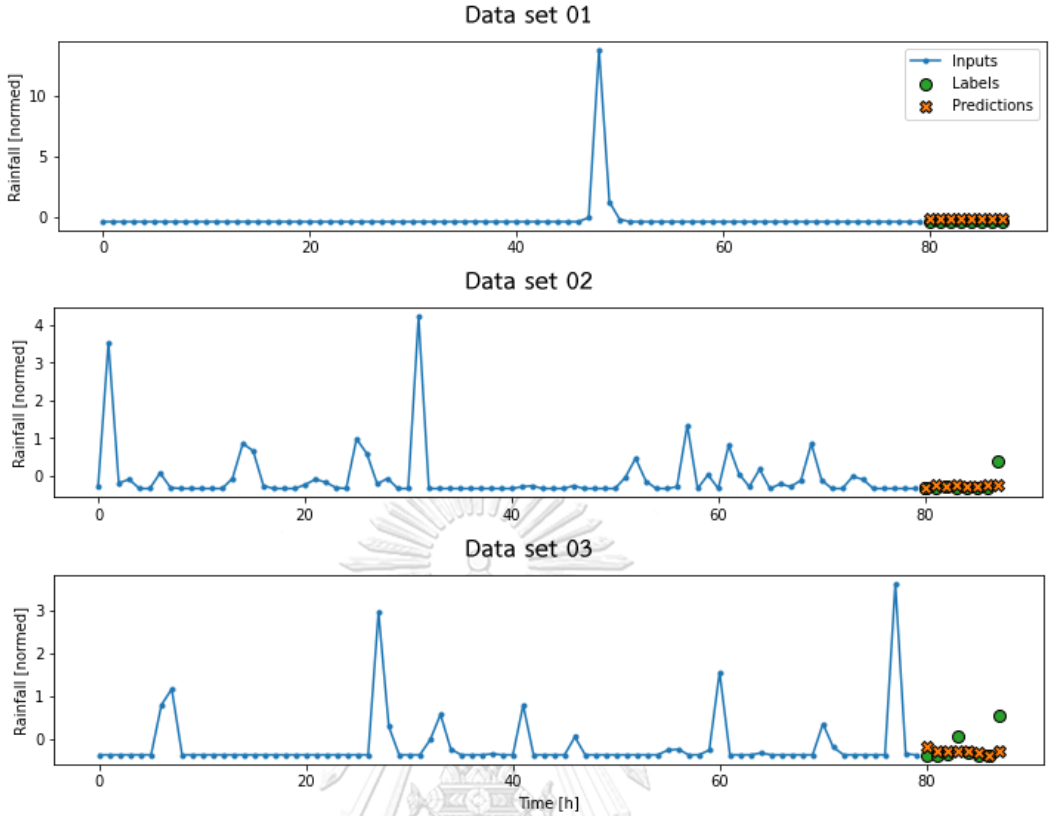
ตาราง 5 แสดงการปรับจูนไฮเปอร์พารามิเตอร์ที่ชั้น (Layer) ต่างๆ ของตัวแบบ RNN

หลังจากที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ตามที่ได้ชี้แจงไปในตาราง 5 แล้ว ขั้นตอนต่อไปคือการนำแบบจำลองที่ได้ไปทำการทดสอบประสิทธิภาพในการพยากรณ์กับกลุ่มข้อมูลตรวจสอบ (Validation data set) และกลุ่มข้อมูลทดสอบ (Test data set) ซึ่งค่า Mean Absolute Error (MAE) ที่ได้จากผลการทดลองทั้ง 2 ช่วงเวลา มีดังนี้

แบบจำลอง	ชุดข้อมูล	Mean Absolute Error (MAE)	
		t+1	t+8
RNN	ข้อมูลชุดที่ 1	0.2829	0.6149
	ข้อมูลชุดที่ 2	0.2655	0.5101
	ข้อมูลชุดที่ 3	0.2221	0.3778

ตาราง 6 แสดงค่า MAE ที่ได้จากการพยากรณ์ด้วยตัวแบบ RNN

จากตาราง 6 จะพบว่า สำหรับการพยากรณ์ปริมาณน้ำฝนสะสมด้วยตัวแบบ RNN ในอีก 1 และ 8 ช่วงเวลาถัดไป ชุดข้อมูลที่ให้ประสิทธิภาพในการพยากรณ์ที่ดีที่สุด คือ ข้อมูลชุดที่ 3 ที่เป็นชุดข้อมูลที่เพิ่มคุณลักษณะทางสถิติที่เกี่ยวข้องกับปริมาณน้ำฝนเท่านั้น ชุดข้อมูลที่มีประสิทธิภาพรองลงมาคือข้อมูลชุดที่ 2 ซึ่งจะเห็นได้ว่า ชุดข้อมูลชุดที่ 1 ซึ่งเป็นชุดข้อมูลดั้งเดิมนั้น ไม่ว่าจะเป็นการพยากรณ์ใน 1 ช่วงเวลาถัดไป หรือ 8 ช่วงเวลาถัดไปก็ยังคงให้ค่า MAE ที่สูงกว่าของทั้งชุดข้อมูลที่ 2 และ 3 อยู่ จึงสามารถสรุปได้คร่าวๆ ว่า ในส่วนของแบบจำลอง RNN นั้น การเพิ่มคุณลักษณะทางสถิติให้กับชุดข้อมูล ส่งผลให้ประสิทธิภาพในการพยากรณ์ของตัวแบบเพิ่มขึ้น ทั้งนี้ ผลการพยากรณ์โดยคร่าวของแบบจำลอง RNN ที่ได้พัฒนาขึ้นดังกล่าว สามารถแสดงได้ดังนี้



รูปที่ 25 ผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยตัวแบบ RNN กับข้อมูลทั้ง 3 ชุด

4.2.2 การพยากรณ์ด้วยตัวแบบแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long short-term memory : LSTM)

แบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) เป็นแบบจำลองการเรียนรู้เชิงลึกที่ถูกพัฒนาต่อยอดมาจากแบบจำลอง RNN เพื่อแก้ปัญหาการลดลงของเกรเดียนต์ (Vanishing Gradient) เมื่อลำดับของข้อมูลที่รับเข้ามามีจำนวนมากเกินไป ในการพัฒนาตัวแบบ LSTM ให้มีความเหมาะสมกับข้อมูล พบว่า การปรับจูนไฮเปอร์พารามิเตอร์ที่มีความเหมาะสม และให้ประสิทธิภาพในการพยากรณ์ที่ดีที่สุด สามารถชี้แจงได้ดังตารางต่อไปนี้

แบบจำลอง	ชั้นของแบบจำลอง (Layer)	ไฮเปอร์พารามิเตอร์ (Hyper-Parameter)	ฟังก์ชันกระตุ้น (Activation Function)
LSTM	LSTM layer 1	64	Tanh
	LSTM layer 2	256	
	Dense	32	
	Reshape	[output, number of features]	

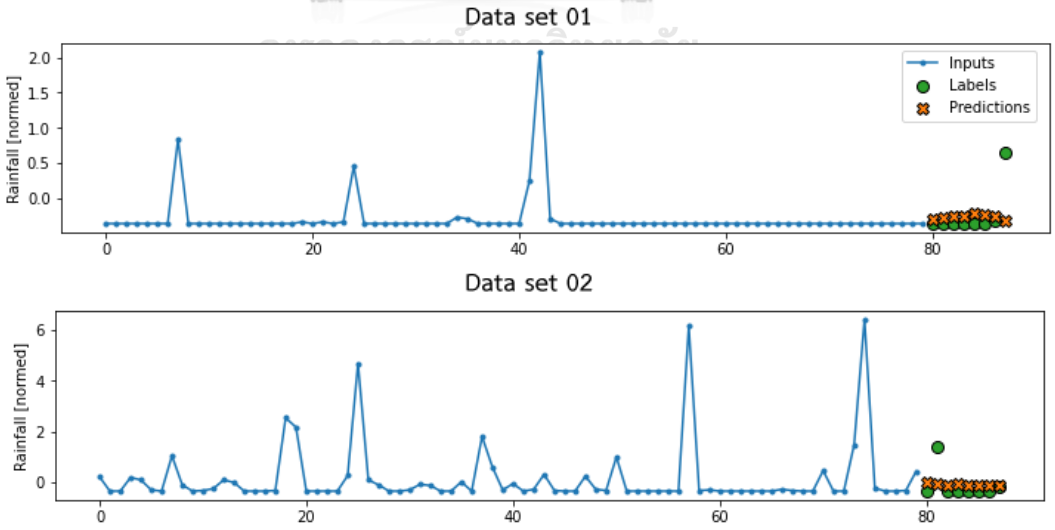
ตาราง 7 แสดงการปรับจูนไฮเปอร์พารามิเตอร์ที่ชั้น (Layer) ต่างๆ ของตัวแบบ LSTM

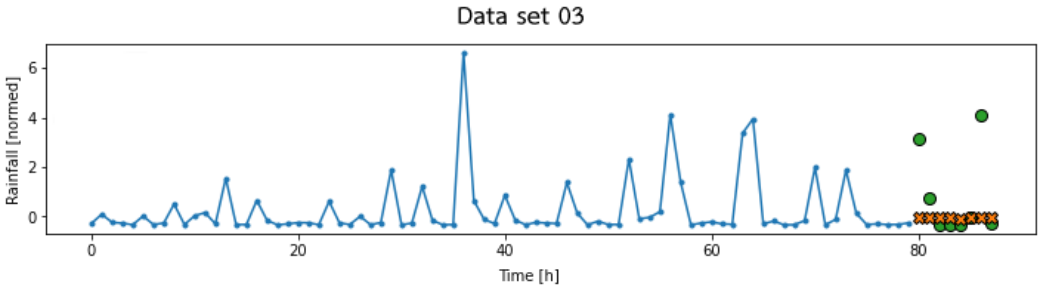
หลังจากที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ตามที่ได้ชี้แจงไปในตาราง 7 แล้ว ขั้นตอนต่อไปคือการนำแบบจำลองที่ได้ไปทำการทดสอบประสิทธิภาพในการพยากรณ์กับกลุ่มข้อมูลตรวจสอบ (Validation data set) และกลุ่มข้อมูลทดสอบ (Test data set) ซึ่งค่า Mean Absolute Error (MAE) ที่ได้จากผลการทดลองทั้ง 2 ช่วงเวลา มีดังนี้

แบบจำลอง	ชุดข้อมูล	Mean Absolute Error (MAE)	
		t+1	t+8
LSTM	ข้อมูลชุดที่ 1	0.2330	0.6022
	ข้อมูลชุดที่ 2	0.3535	0.5703
	ข้อมูลชุดที่ 3	0.2353	0.3834

ตาราง 8 แสดงค่า MAE ที่ได้จากการพยากรณ์ด้วยตัวแบบ LSTM

จากตาราง 8 จะพบว่า สำหรับการพยากรณ์ปริมาณน้ำฝนสะสมในอีก 1 ช่วงเวลาถัดไป ชุดข้อมูลที่ให้ประสิทธิภาพในการพยากรณ์ที่ดีที่สุดคือ ชุดข้อมูลชุดที่ 1 รองลงมาคือชุดข้อมูลชุดที่ 2 ที่ให้ประสิทธิภาพในการพยากรณ์ที่ไม่แตกต่างกันมากนัก แต่สำหรับในอีก 8 ช่วงเวลาถัดไปนั้น พบว่าประสิทธิภาพในการพยากรณ์ของชุดข้อมูลชุดที่ 1 ลดลงอย่างมาก และชุดข้อมูลที่ให้ผลการพยากรณ์ดีที่สุดในอีก 8 ช่วงเวลาต่อมาก็คือ ชุดข้อมูลชุดที่ 3 รองลงมาคือชุดข้อมูลชุดที่ 2 กล่าวคือ สำหรับตัวแบบ LSTM นั้น ในช่วงเวลาแรก การเพิ่มคุณลักษณะทางสถิติจะยังไม่สามารถให้ผลลัพธ์ที่ดีได้ แต่เมื่อเวลาผ่านไป การเพิ่มคุณลักษณะทางสถิติจะช่วยให้ประสิทธิภาพในการพยากรณ์ของตัวแบบดีขึ้น สำหรับผลการพยากรณ์โดยคร่าวของแบบจำลอง LSTM สามารถแสดงได้ดังนี้





รูปที่ 26 ผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยตัวแบบ LSTM กับข้อมูลทั้ง 3 ชุด

4.2.3 การพยากรณ์ด้วยโครงข่ายประตูวกกลับ (Gated Recurrent Unit : GRU)

โครงข่ายประตูวกกลับ (GRU) เป็นแบบจำลองการเรียนรู้เชิงลึกที่ถูกพัฒนาขึ้นเพื่อแก้ปัญหาการลดลงของเกรเดียนต์ (Vanishing Gradient) เช่นเดียวกับแบบจำลอง LSTM แต่จะแตกต่างจากแบบจำลอง LSTM ตรงที่แบบจำลอง GRU จะลดความซับซ้อนของประตู (Gate) ลง เพื่อให้ตัวแบบจำลองสามารถประมวลผลได้รวดเร็วขึ้น ในการพัฒนาตัวแบบ GRU ให้มีความเหมาะสมกับข้อมูล พบว่า การปรับจูนไฮเปอร์พารามิเตอร์ที่มีความเหมาะสม และให้ประสิทธิภาพในการพยากรณ์ที่ดีที่สุด สามารถชี้แจงได้ดังตารางต่อไปนี้

แบบจำลอง	ชั้นของแบบจำลอง (Layer)	ไฮเปอร์พารามิเตอร์ (Hyper-Parameter)	ฟังก์ชันกระตุ้น (Activation Function)
GRU	GRU layer 1	32	Tanh
	GRU layer 2	64	
	Dense	1	
	Reshape	[output, number of features]	

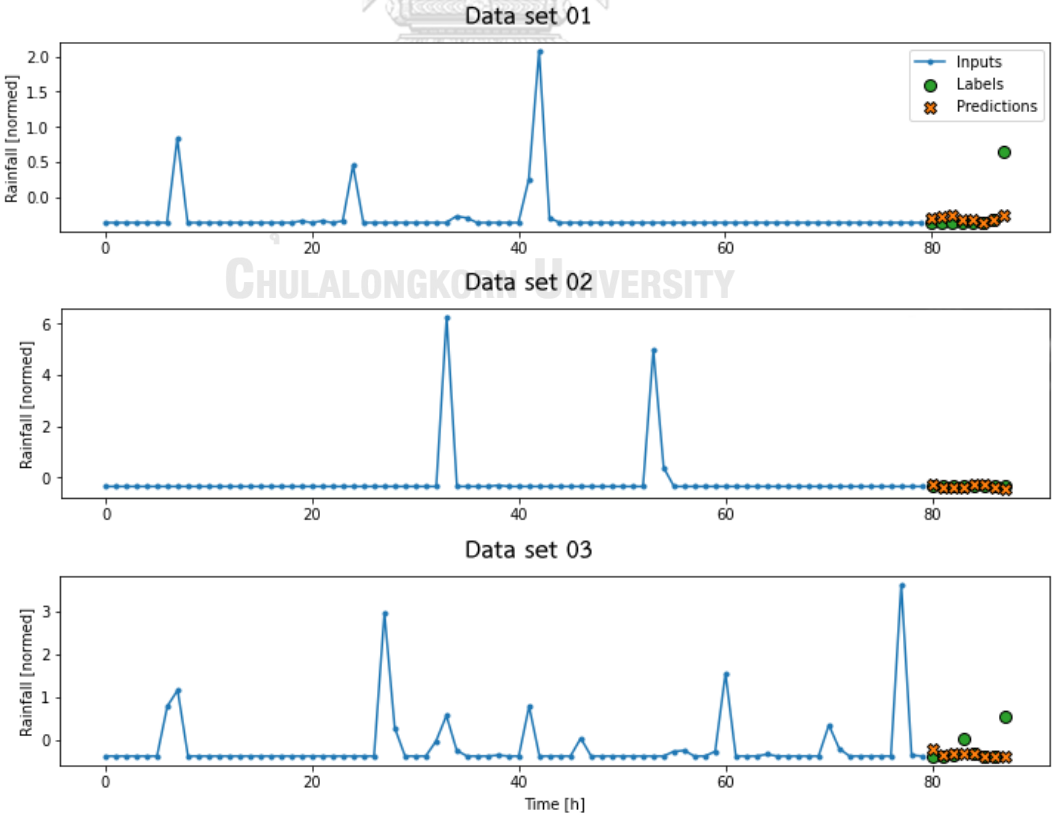
ตาราง 9 แสดงการปรับจูนไฮเปอร์พารามิเตอร์ที่ชั้น (Layer) ต่างๆ ของตัวแบบ GRU

หลังจากที่ทำการปรับจูนไฮเปอร์พารามิเตอร์ตามที่ได้ชี้แจงไปในตาราง 9 แล้ว ขั้นตอนต่อไปคือการนำแบบจำลองที่ได้ไปทำการทดสอบประสิทธิภาพในการพยากรณ์กับชุดข้อมูลตรวจสอบ (Validation data set) และชุดข้อมูลทดสอบ (Test data set) ซึ่งค่า Mean Absolute Error (MAE) ที่ได้จากผลการทดลองทั้ง 2 ช่วงเวลา มีดังนี้

แบบจำลอง	ชุดข้อมูล	Mean Absolute Error (MAE)	
		t+1	t+8
GRU	ข้อมูลชุดที่ 1	0.1767	0.6018
	ข้อมูลชุดที่ 2	0.2900	0.5247
	ข้อมูลชุดที่ 3	0.2164	0.3752

ตาราง 10 แสดงค่า MAE ที่ได้จากการพยากรณ์ด้วยตัวแบบ GRU

จากตาราง 10 จะพบว่า สำหรับการพยากรณ์ปริมาณน้ำฝนสะสมในอีก 1 ช่วงเวลาถัดไปด้วยตัวแบบ GRU พบว่า ชุดข้อมูลที่ให้ประสิทธิภาพในการพยากรณ์ที่ดีที่สุด คือ ข้อมูลชุดที่ 1 ชุดข้อมูลที่มีประสิทธิภาพรองลงมาคือข้อมูลชุดที่ 3 แต่เมื่อเป็นการพยากรณ์ในอีก 8 ช่วงเวลาถัดไป ชุดข้อมูลชุดที่ 1 กลับมีประสิทธิภาพที่ลดลงเป็นอย่างมาก ในขณะที่ชุดข้อมูลที่ให้ประสิทธิภาพในการพยากรณ์ในอีก 8 ช่วงเวลาถัดมาได้ดีที่สุดกลับเป็นชุดข้อมูลชุดที่ 3 และชุดข้อมูลชุดที่ 2 ตามลำดับ จะเห็นได้ว่า ในการใช้ตัวแบบจำลอง GRU ในการพยากรณ์ปริมาณน้ำฝนสะสมนั้น การเพิ่มคุณลักษณะทางสถิติจะยังคงไม่เป็นผลลัพธ์ที่ดีในด้านการเพิ่มประสิทธิภาพในการพยากรณ์ แต่เมื่อเวลาผ่านไป การเพิ่มคุณลักษณะทางสถิติจะช่วยให้ผลการพยากรณ์มีประสิทธิภาพมากยิ่งขึ้น ทั้งนี้ ผลการพยากรณ์โดยคร่าวของแบบจำลอง GRU ที่ได้พัฒนาขึ้นดังกล่าว สามารถแสดงได้ดังนี้



รูปที่ 27 ผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยตัวแบบ GRU กับข้อมูลทั้ง 3 ชุด

4.3 ผลการเปรียบเทียบประสิทธิภาพในการพยากรณ์

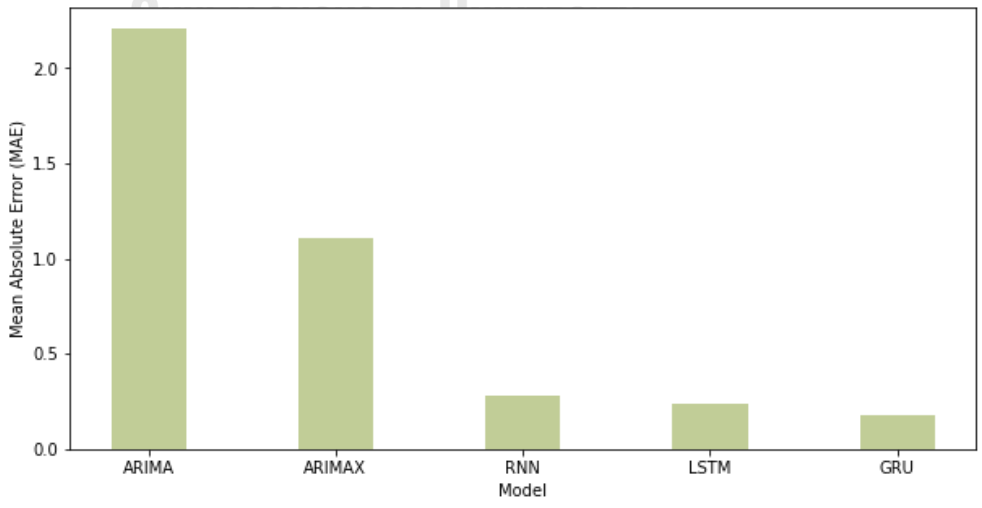
เนื่องจากในงานวิจัยฉบับนี้มีจุดมุ่งหมายในการวิเคราะห์เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมล่วงหน้าในระยะสั้นร่วมกับการใช้ข้อมูลหลายตัวแปร โดยจะทำการเปรียบเทียบใน 2 ประเด็นหลัก ซึ่งประเด็นแรก การเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบที่ใช้ชุดข้อมูลที่แตกต่างกัน ระหว่างชุดข้อมูลที่ผ่านการเพิ่มคุณลักษณะพิเศษกับชุดข้อมูลดั้งเดิม ได้ทำการเปรียบเทียบ อธิบายประสิทธิภาพ และแสดงผลการพยากรณ์ไปแล้วในหัวข้อที่ 4.1 และ 4.2 ดังนั้นในหัวข้อนี้ จะทำการเปรียบเทียบประสิทธิภาพในการพยากรณ์ในช่วงเวลาถัดไปของโครงข่ายระบบประสาทแบบย้อนกลับที่สนใจศึกษากับแบบจำลองอาริมา ซึ่งเป็นประเด็นสุดท้ายที่จะนำมาเปรียบเทียบในงานวิจัยฉบับนี้

ดังที่ ได้อธิบายผลการทดลองไปแล้วในหัวข้อที่ 4.1 และ 4.2 จะเห็นว่า แบบจำลอง ARIMA นั้นให้ผลการพยากรณ์ที่ไม่ค่อยดีนัก เมื่อเทียบกับแบบจำลอง ARIMAX และแบบจำลองโครงข่ายระบบประสาทย้อนกลับ เพื่อให้เห็นภาพการเปรียบเทียบประสิทธิภาพในการพยากรณ์ของแต่ละแบบจำลองกับแต่ละชุดข้อมูล ในหัวข้อนี้จะทำการแบ่งการเปรียบเทียบประสิทธิภาพในการพยากรณ์ตามชุดข้อมูลดังนี้

4.3.1 เปรียบเทียบประสิทธิภาพในการพยากรณ์ด้วยชุดข้อมูลชุดที่ 1

ในการศึกษาประสิทธิภาพในการพยากรณ์ด้วยข้อมูลชุดที่ 1 ของแต่ละแบบจำลองจะแบ่งออกเป็น 2 กรณีศึกษา ได้แก่

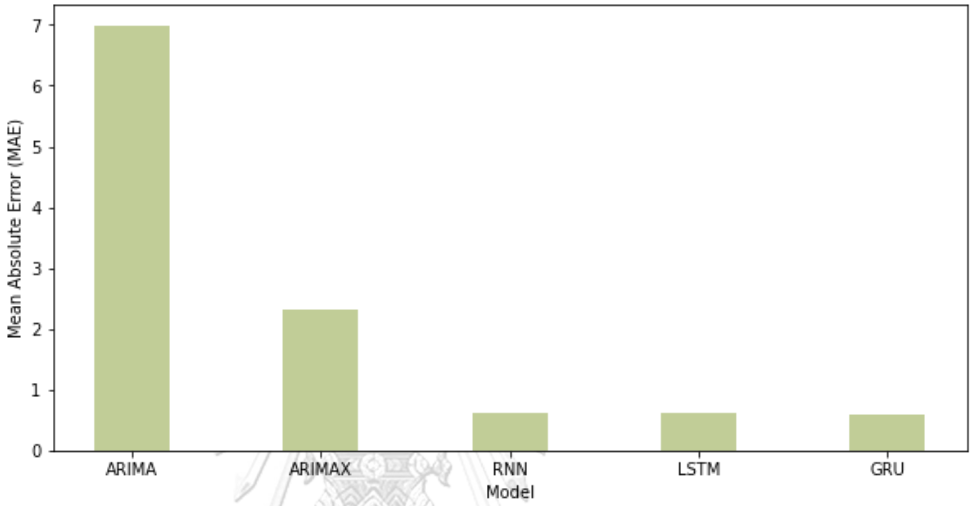
1) เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนด้วยข้อมูลชุดที่ 1 ในอีก 1 ช่วงเวลาถัดไปของแบบจำลองที่สนใจศึกษา



รูปที่ 28 เปรียบเทียบผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยข้อมูลชุดที่ 1 ในอีก 1 ช่วงเวลาถัดไป

จากรูปที่ 28 พบว่า ในการพยากรณ์ปริมาณน้ำฝนสะสมด้วยข้อมูลชุดที่ 1 ในอีก 1 ช่วงเวลาถัดไปนั้น แบบจำลองที่มีประสิทธิภาพในการพยากรณ์มากที่สุด คือแบบจำลอง GRU รองมาคือแบบจำลอง LSTM และ RNN ตามลำดับ และแบบจำลองที่มีประสิทธิภาพในการพยากรณ์น้อยที่สุด คือแบบจำลอง ARIMA

2) เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนด้วยข้อมูลชุดที่ 1 ในอีก 8 ช่วงเวลาถัดไปของแบบจำลองที่สนใจศึกษา

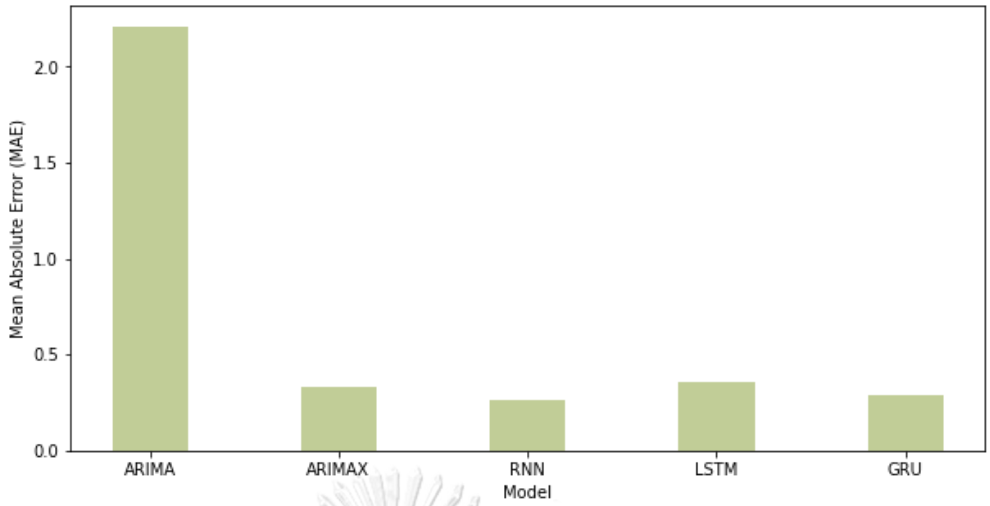


รูปที่ 29 เปรียบเทียบผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยข้อมูลชุดที่ 1 ในอีก 8 ช่วงเวลาถัดไป

จากรูปที่ 29 พบว่า ในการพยากรณ์ปริมาณน้ำฝนสะสมด้วยข้อมูลชุดที่ 1 ในอีก 8 ช่วงเวลาถัดไปนั้น แบบจำลองที่มีประสิทธิภาพในการพยากรณ์มากที่สุด คือแบบจำลอง GRU รองมาคือแบบจำลอง LSTM และ RNN ตามลำดับ และแบบจำลองที่มีประสิทธิภาพในการพยากรณ์น้อยที่สุด คือแบบจำลอง ARIMA

4.3.2 เปรียบเทียบประสิทธิภาพในการพยากรณ์ด้วยชุดข้อมูลชุด 2

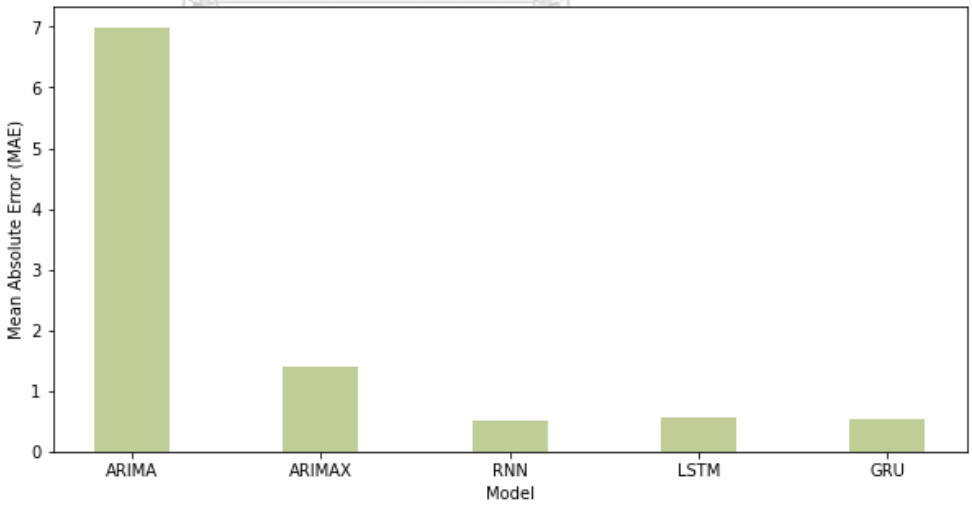
1) เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนด้วยข้อมูลชุดที่ 2 ในอีก 1 ช่วงเวลาถัดไปของแบบจำลองที่สนใจศึกษา



รูปที่ 30 เปรียบเทียบผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยข้อมูลชุดที่ 2 ในอีก 1 ช่วงเวลาถัดไป

จากรูปที่ 30 พบว่า ในการพยากรณ์ปริมาณน้ำฝนสะสมด้วยข้อมูลชุดที่ 2 ในอีก 1 ช่วงเวลาถัดไปนั้น แบบจำลองที่มีประสิทธิภาพในการพยากรณ์มากที่สุด คือ แบบจำลอง RNN รองมาคือแบบจำลอง GRU และ ARIMAX ตามลำดับ และแบบจำลองที่มีประสิทธิภาพในการพยากรณ์น้อยที่สุด คือแบบจำลอง ARIMA

2) เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนด้วยข้อมูลชุดที่ 2 ในอีก 8 ช่วงเวลาถัดไปของแบบจำลองที่สนใจศึกษา



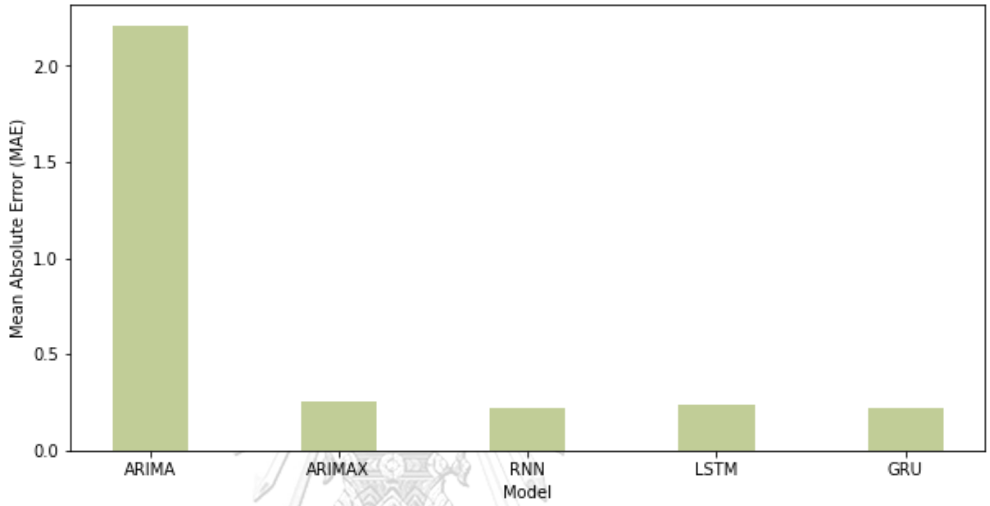
รูปที่ 31 เปรียบเทียบผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยข้อมูลชุดที่ 2 ในอีก 8 ช่วงเวลาถัดไป

จากรูปที่ 31 พบว่า ในการพยากรณ์ปริมาณน้ำฝนสะสมด้วยข้อมูลชุดที่ 2 ในอีก 8 ช่วงเวลาถัดไปนั้น แบบจำลองที่มีประสิทธิภาพในการพยากรณ์มากที่สุด คือแบบจำลอง

RNN รองมาคือแบบจำลอง GRU และ LSTM ตามลำดับ และแบบจำลองที่มีประสิทธิภาพในการพยากรณ์น้อยที่สุด คือแบบจำลอง ARIMA

4.3.3 เปรียบเทียบประสิทธิภาพในการพยากรณ์ด้วยชุดข้อมูลชุด 3

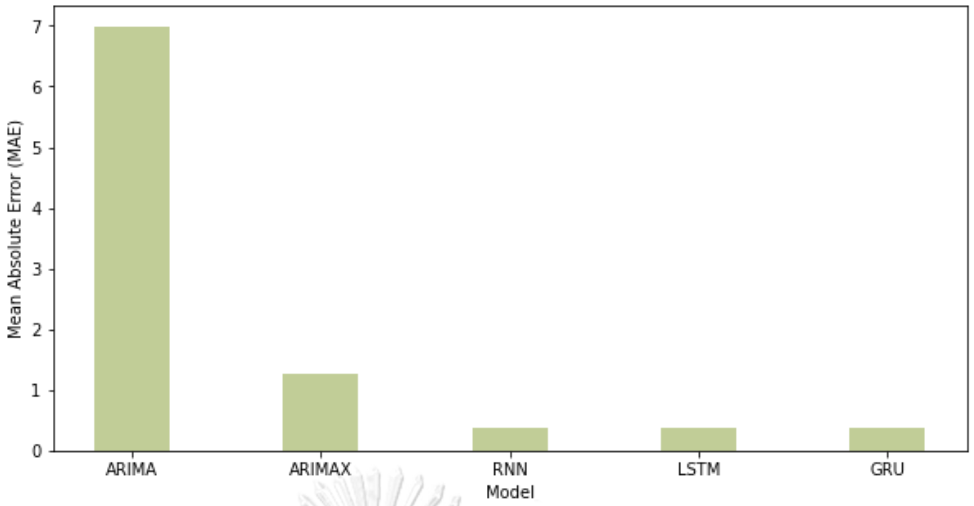
1) เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนด้วยข้อมูลชุดที่ 3 ในอีก 1 ช่วงเวลาถัดไปของแบบจำลองที่สนใจศึกษา



รูปที่ 32 เปรียบเทียบผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยข้อมูลชุดที่ 3 ในอีก 1 ช่วงเวลาถัดไป

จากรูปที่ 32 พบว่า ในการพยากรณ์ปริมาณน้ำฝนสะสมด้วยข้อมูลชุดที่ 3 ในอีก 1 ช่วงเวลาถัดไปนั้น แบบจำลองที่มีประสิทธิภาพในการพยากรณ์มากที่สุด คือ แบบจำลอง GRU รองมาคือแบบจำลอง RNN และ LSTM ตามลำดับ และแบบจำลองที่มีประสิทธิภาพในการพยากรณ์น้อยที่สุด คือแบบจำลอง ARIMA

2) เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนด้วยข้อมูลชุดที่ 3 ในอีก 8 ช่วงเวลาถัดไปของแบบจำลองที่สนใจศึกษา



รูปที่ 33 เปรียบเทียบผลการพยากรณ์ข้อมูลปริมาณน้ำฝนด้วยข้อมูลชุดที่ 3 ในอีก 8 ช่วงเวลา

จากรูปที่ 33 พบว่า ในการพยากรณ์ปริมาณน้ำฝนสะสมด้วยข้อมูลชุดที่ 3 ในอีก 8 ช่วงเวลาถัดไปนั้น แบบจำลองที่มีประสิทธิภาพในการพยากรณ์มากที่สุด คือ แบบจำลอง GRU รองมาคือแบบจำลอง RNN และ LSTM ตามลำดับ และแบบจำลองที่มีประสิทธิภาพในการพยากรณ์น้อยที่สุด คือแบบจำลอง ARIMA

บทที่ 5
สรุปผลการวิจัย

ในบทที่ 5 นี้จะเป็นบทสุดท้ายของวิทยานิพนธ์ฉบับนี้ ซึ่งจะเป็นบทที่พูดถึงใน 2 หัวข้อใหญ่ๆ คือ 1) การสรุปผล จะทำการสรุปผลการพัฒนาตัวแบบ ผลการพยากรณ์ และผลการเปรียบเทียบที่ได้ และ 2) ข้อเสนอแนะ จะเป็นการแนะนำแนวทางในการเพิ่มประสิทธิภาพในการพยากรณ์ สำหรับผู้ที่สนใจจะนำงานวิจัยนี้ไปศึกษา ปรับใช้ หรือต่อยอดในอนาคต

5.1 สรุปผล

งานวิจัยนี้มุ่งเน้นการพัฒนาและวิเคราะห์เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมล่วงหน้าในระยะสั้นร่วมกับการใช้ข้อมูลหลายตัวแปร ข้อมูลที่นำมาใช้ในงานวิจัยนี้เป็นข้อมูลสภาพอากาศและปริมาณน้ำฝนสะสมในพื้นที่สนามบินสุวรรณภูมิที่ได้รับการสนับสนุนจากกรมอุตุนิยมวิทยา ประเทศไทย

ในการพัฒนาตัวแบบ ผู้วิจัยได้เตรียมชุดข้อมูลนำเข้าไว้ 3 ชุด คือ 1) ชุดข้อมูลดั้งเดิม (รวม 4 คุณลักษณะ) 2) ชุดข้อมูลที่เพิ่มคุณลักษณะทางสถิติ (รวม 23 คุณลักษณะ) 3) ชุดข้อมูลที่เพิ่มเฉพาะคุณลักษณะทางสถิติที่เกี่ยวข้องกับปริมาณน้ำฝนสะสม (รวม 11 คุณลักษณะ) เนื่องจากเมื่อลองพิจารณาข้อมูลที่อยู่ในชุดข้อมูลแล้วพบว่า จำนวนชั่วโมงที่มีฝนตกนั้น มีน้อยกว่าจำนวนที่ฝนไม่ตกมาก โดยจำนวนชั่วโมงที่มีฝนตกนั้น คิดเป็นร้อยละ 10 จากจำนวนชั่วโมงที่ไม่มีฝนตก ซึ่งก่อให้เกิดปัญหาชุดข้อมูลไม่สมดุลตามมา จึงได้จัดการกับปัญหาดังกล่าวด้วยการเลือกใช้วิธี Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) เพื่อให้มีจำนวนชั่วโมงที่มีฝนตกมากพอที่จะทำให้แบบจำลองสามารถทำนายปริมาณน้ำฝนสะสมได้ จากนั้นจึงได้ทำการพัฒนาตัวแบบจำลองที่เกี่ยวข้องกับการจัดการข้อมูลที่อยู่ในรูปของอนุกรมเวลาขึ้น ได้แก่

1. ตัวแบบ ARIMA เป็นตัวแบบที่รับข้อมูลนำเข้าเฉพาะข้อมูลปริมาณน้ำฝน จากการพัฒนาและคัดเลือกตัวแบบที่ดีที่สุดด้วยวิธีวิธีบอกซ์แอนด์เจนกินส์ พบว่าตัวแบบที่ให้ค่าความคลาดเคลื่อนน้อยที่สุด คือตัวแบบ ARIMA (0,1,2)
2. ตัวแบบ ARIMAX เป็นตัวแบบที่รับข้อมูลคุณลักษณะอื่นๆ มาพิจารณาร่วมกับข้อมูลปริมาณน้ำฝนที่เราต้องการจะพยากรณ์ จากการพัฒนาและคัดเลือกตัวแบบที่ดีที่สุดด้วยวิธีวิธีบอกซ์แอนด์เจนกินส์ จะทำให้เราได้ตัวแบบ ARIMAX ที่ดีที่สุดสำหรับแต่ละชุดข้อมูล ได้แก่
    - ข้อมูลชุดที่ 1 ตัวแบบที่ดีที่สุด คือ ตัวแบบ ARIMAX (1,1,3)
    - ข้อมูลชุดที่ 2 ตัวแบบที่ดีที่สุด คือ ตัวแบบ ARIMAX (4,1,3)
    - ข้อมูลชุดที่ 3 ตัวแบบที่ดีที่สุด คือ ตัวแบบ ARIMAX (1,1,3)

3. แบบจำลองการเรียนรู้เชิงลึก เป็นตัวแบบที่รับข้อมูลนำเข้าหลายตัวแปร โดยในงานวิจัยนี้ได้เลือกใช้ตัวแบบจำลองการเรียนรู้เชิงลึก 3 ตัว มาเพื่อใช้ในการเปรียบเทียบ ได้แก่ โครงข่ายระบบประสาทแบบย้อนกลับ (RNN) แบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) และโครงข่ายประตูวกกลับ (GRU)

ในส่วนการทดลอง งานวิจัยนี้ได้แบ่งการทดลองออกเป็น 2 ส่วน ดังนี้

1) การทดลองโดยการเปรียบเทียบประสิทธิภาพระหว่างการพัฒนาตัวแบบโดยการเพิ่มคุณลักษณะทางสถิติ

จากการทดลองในเบื้องต้นพบว่า การเพิ่มเฉพาะคุณลักษณะทางสถิติให้กับชุดข้อมูลก่อนที่จะนำไปทำการพัฒนาตัวแบบนั้น ส่งผลให้ประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมในช่วงเวลาไกลๆ เพิ่มขึ้น (ในอีก 8 ช่วงเวลาถัดไป) โดยเฉพาะในชุดข้อมูลชุดที่ 3 ที่มีการเพิ่มเฉพาะคุณลักษณะทางสถิติที่มีความสัมพันธ์กับปริมาณน้ำฝนสะสมให้กับชุดข้อมูลก่อนที่จะนำเข้าแบบจำลองส่งผลให้ Mean Absolute Error (MAE) ลดลงมากกว่าการใช้เพียงคุณลักษณะตั้งต้นเพียง 4 คุณลักษณะ แต่สำหรับการพยากรณ์ในระยะสั้น (1 ช่วงเวลาถัดไป) การเพิ่มคุณลักษณะทางสถิติให้กับชุดข้อมูลในแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับยังไม่สามารถเพิ่มประสิทธิภาพในการพยากรณ์ให้ดีขึ้นได้มากนัก ยกเว้นสำหรับแบบจำลอง ARIMAX ซึ่งเห็นได้อย่างชัดเจนว่า ในการพยากรณ์ระยะสั้นนั้น การเพิ่มคุณลักษณะทางสถิติให้กับชุดข้อมูล ส่งผลให้ประสิทธิภาพในการพยากรณ์ดีขึ้นเกือบเทียบเท่ากับโครงข่ายระบบประสาทแบบย้อนกลับ แต่เมื่อช่วงเวลาไกลออกไป ประสิทธิภาพในการพยากรณ์ของตัวแบบ ARIMAX กลับลดลงอย่างมาก

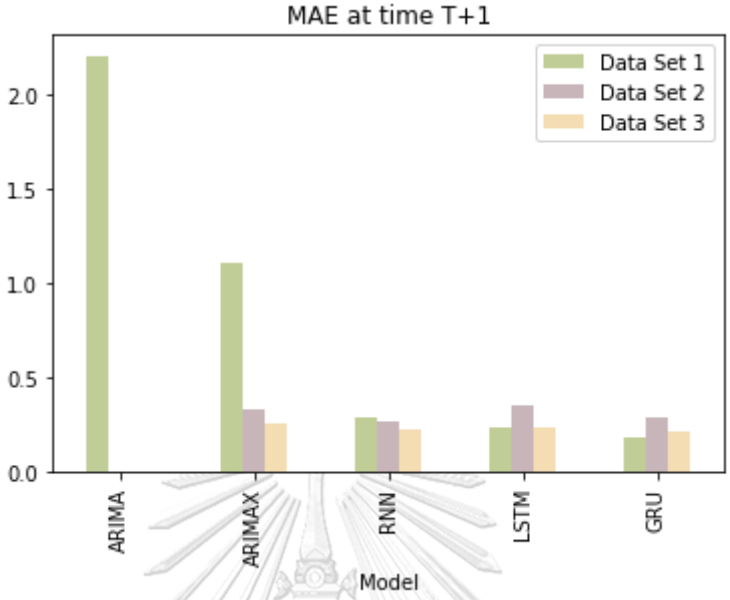
อย่างไรก็ตาม สำหรับข้อมูลชุดที่ 2 หากสามารถพัฒนาตัวแบบให้มีความซับซ้อนได้มากกว่านี้ อาจจะส่งผลให้การสกัดคุณลักษณะในแต่ละชั้นของแบบจำลองโครงข่ายระบบประสาทแบบย้อนกลับทำได้ดีขึ้น และอาจจะส่งผลให้ประสิทธิภาพของการพยากรณ์เพิ่มขึ้น อีกทั้ง ถึงแม้จะพัฒนาตัวแบบจำลองจนสามารถพยากรณ์ชั่วโมงที่มีฝนตกได้ แต่เมื่อพิจารณาจากผลของการพยากรณ์แล้วก็พบว่ายังคงมีปัญหาในเรื่องที่ไม่สามารถทำนายชั่วโมงที่มีปริมาณน้ำฝนสะสมมากกว่าปกติได้ ซึ่งเป็นส่วนที่ต้องพัฒนาต่อไป

2) การทดลองโดยการเปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมในช่วงเวลาถัดไปของแต่ละแบบจำลอง

การทดลองเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบแต่ละตัวในช่วงเวลาที่แตกต่างกัน งานวิจัยได้แบ่งการทดลองส่วนนี้ออกเป็น 2 ส่วนเพื่อวัดประสิทธิภาพ คือ

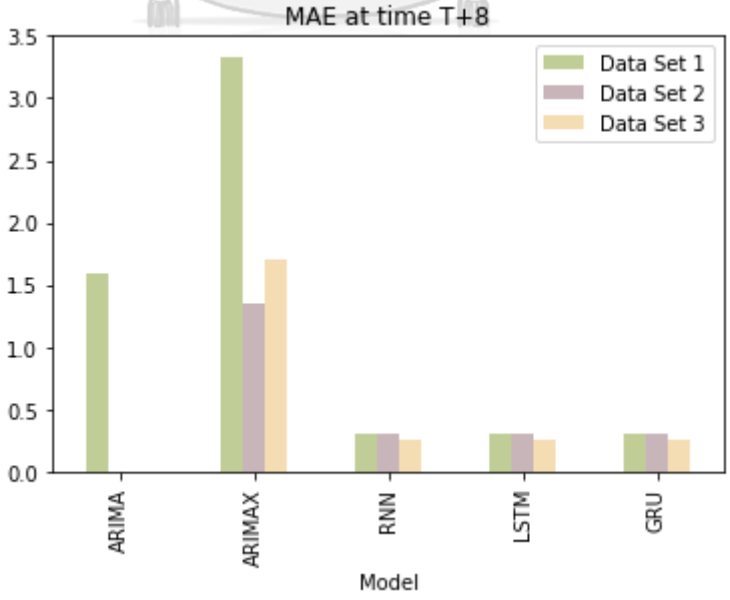
2.1 เปรียบเทียบประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมใน 1 ช่วงเวลาถัดไป (3 ชั่วโมงถัดไป) พบว่า แบบจำลองในกลุ่มโครงข่ายระบบประสาทแบบย้อนกลับให้ประสิทธิภาพในการพยากรณ์ที่ดีกว่าแบบจำลองในกลุ่มอาริมา โดยในการพยากรณ์ปริมาณน้ำฝนสะสมในอีก 1 ช่วงเวลาถัดไป

นั้น แบบจำลอง GRU ให้ประสิทธิภาพในการพยากรณ์ได้ดีที่สุด ซึ่งผลการทดลองและเปรียบเทียบประสิทธิภาพนั้น สามารถแสดงได้ดังแผนภาพต่อไปนี้



รูปที่ 34 กราฟแสดงการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองในการพยากรณ์ 1 ช่วงเวลาถัดไป

2.2 การเปรียบเทียบประสิทธิภาพในการพยากรณ์ใน 8 ช่วงเวลาถัดไป (24 ชั่วโมง) ในการพยากรณ์ในอีก 8 ช่วงเวลาถัดไปนั้น พบว่า แบบจำลอง GRU ให้ประสิทธิภาพในการพยากรณ์ได้ดีที่สุด ซึ่งผลการทดลองและเปรียบเทียบประสิทธิภาพนั้น สามารถแสดงได้ดังแผนภาพต่อไปนี้



รูปที่ 35 กราฟแสดงการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองในการพยากรณ์ 3 ช่วงเวลาถัดไป

นอกจากนี้ เมื่อเปรียบเทียบกับแบบจำลองโครงข่ายประสาทแบบย้อนกลับตัวอื่นๆ แบบจำลอง GRU ยังเป็นแบบจำลองที่มีการประมวลผลที่ไวที่สุด เนื่องจากมีจำนวนประตูที่น้อย ดังนั้น สำหรับการเลือกแบบจำลองโครงข่ายระบบประสาทย้อนกลับเพื่อทำนายปริมาณน้ำฝนสะสม แบบจำลอง GRU จึงเป็นตัวเลือกที่น่าสนใจ

สุดท้ายนี้ เมื่อดูภาพรวมของผลการทดลองเปรียบเทียบทั้ง 2 หัวข้อ จะพบว่า โครงข่ายระบบประสาทแบบย้อนกลับนั้น เป็นตัวแบบที่มีประสิทธิภาพในการพยากรณ์ปริมาณน้ำฝนสะสมที่เกิดขึ้นมากที่สุด และการทำ Feature Engineering โดยการเพิ่มคุณลักษณะทางสถิติ ก็ช่วยให้ผลการพยากรณ์มีความแม่นยำมากขึ้น โดยเฉพาะอย่างยิ่งกับข้อมูลที่มีจำนวนคุณลักษณะน้อย ซึ่งการมีจำนวนคุณลักษณะที่น้อยนั้น อาจส่งผลให้ตัวโครงข่ายระบบประสาทแบบย้อนกลับไม่สามารถให้ประสิทธิภาพในการพยากรณ์ข้อมูลที่สนใจได้ดีเท่าที่ควร ดังนั้น นอกเหนือจากการพัฒนาตัวแบบและการปรับจูนไฮเปอร์พารามิเตอร์เพื่อเพิ่มประสิทธิภาพในการพยากรณ์ให้กับแบบจำลอง ควรให้ความสนใจในเรื่องของการทำ Feature Engineering ให้กับชุดข้อมูลก่อนที่จะนำเข้าแบบจำลองด้วย เพื่อให้แบบจำลองสามารถเรียนรู้ และพยากรณ์ผลลัพธ์ได้อย่างแม่นยำมากขึ้น

5.2 ข้อจำกัดในงานวิจัย

ในการพัฒนาตัวแบบจำลองเพื่อทำนายปริมาณน้ำฝนสะสมในพื้นที่สนามบินนานาชาติสุวรรณภูมิมีข้อจำกัดดังต่อไปนี้

1. ข้อมูลสภาพอากาศในไทยส่วนมาเป็นข้อมูลที่ไม่ได้เผยแพร่แบบสาธารณะ ทำให้ข้อมูลสภาพอากาศที่นำมาใช้สร้างตัวแบบมีน้อย และจำกัด
2. จำนวนสถานีตรวจวัดสภาพอากาศในประเทศไทยมีจำนวนไม่มาก (เฉลี่ยจังหวัดละ 1 แห่ง) จึงไม่สามารถนำข้อมูลสภาพอากาศจากสถานีใกล้เคียงมาพิจารณาร่วมได้
3. ด้านสมรรถภาพของคอมพิวเตอร์ เนื่องจากทรัพยากรด้านคอมพิวเตอร์ที่จำกัดของผู้วิจัย จึงไม่สามารถพัฒนาตัวแบบที่มีความซับซ้อน หรือนำข้อมูลช่วงเวลาในอดีตใส่เข้าไปมากกว่านี้ได้

5.3 แนวทางการวิจัยในอนาคต

ทั้งนี้ เพื่อให้แบบจำลองมีประสิทธิภาพ และความแม่นยำในการพยากรณ์มากยิ่งขึ้น ในอนาคตควรจะมีการเพิ่มเติมในส่วนต่างๆ ดังนี้

1. ด้านข้อมูล ควรนำข้อมูลพื้นที่ หรือข้อมูลเกี่ยวกับสภาพอากาศด้านอื่นๆ เช่น ปริมาณแสงแดด ทิศทางลม ความหนาแน่นของเมฆ ฯ มาพิจารณาร่วมด้วย เนื่องจากทั้งลักษณะทางภูมิศาสตร์ และปัจจัยด้านสภาพอากาศอื่นๆ ก็มีผลต่อปริมาณน้ำฝนสะสม

2. การปรับเพิ่มความยาวของช่วงเวลาในอดีต (Timestep) ที่จะนำเข้าแบบจำลองให้มีความยาวเพิ่มขึ้นมากกว่า 80 ช่วงเวลา เพื่อเปรียบเทียบประสิทธิภาพในการพยากรณ์ของแต่ละแบบจำลองกับแต่ละชุดข้อมูลเพื่อดูว่าการเพิ่มคุณลักษณะให้ชุดข้อมูลนั้นจะส่งผลต่อประสิทธิภาพการพยากรณ์จริงๆ หรือไม่ และแบบจำลองใดที่เหมาะสมกับข้อมูลชุดนี้มากที่สุด
3. การลองใช้ Optimization Algorithm ตัวอื่นๆ ในการเพิ่มประสิทธิภาพ และเปรียบเทียบผลการพยากรณ์
4. การใช้การเรียนรู้แบบกลุ่ม (Ensemble Learning Method) เข้ามาเป็นตัวช่วยในการพัฒนาตัวแบบให้สามารถพยากรณ์ได้อย่างแม่นยำ
5. สำหรับ Library TensorFlow ที่ผู้วิจัยเลือกใช้ในงานวิจัยนี้ ยังมีประเภทของชั้นของแบบจำลอง (Layer) ในหมวดหมู่ของโครงข่ายระบบประสาทแบบย้อนกลับที่พัฒนาต่อยอดจากแบบพื้นฐานไปอีกหลายตัว ซึ่งในอนาคตสำหรับงานวิจัยต่อๆ ไปสามารถทดลองกับชั้นของแบบจำลองอื่นๆ ที่นอกเหนือจากที่ใช้ในงานวิจัยนี้ได้

บรรณานุกรม

Aurnhammer, C., & Frank, S. L. (2019). Comparing gated and simple recurrent neural network architectures as models of human sentence processing.

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: a pre-processing approach for imbalanced regression. First international workshop on learning with imbalanced domains: Theory and applications,

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:*1412.3555.

Fan, H., Jiang, M., Xu, L., Zhu, H., Cheng, J., & Jiang, J. (2020). Comparison of long short term memory networks and the hydrological model in runoff simulation. *Water*, 12(1), 175.

Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016). Rainfall prediction: A deep learning approach. International Conference on Hybrid Artificial Intelligence Systems,

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hung, N. Q., Babel, M. S., Weesakul, S., & Tripathi, N. (2009). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. *Hydrology and Earth System Sciences*, 13(8), 1413-1425.

Jalalkamali, A., Moradi, M., & Moradi, N. (2015). Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized Precipitation Index. *International journal of environmental science and technology*, 12(4), 1201-1210.

Manokij, F. (2019). *Thailand's Precipitation Forecasting Using Deep Learning Approach* Chulalongkorn University]. Bangkok.

Narayanan, P., Basistha, A., Sarkar, S., & Sachdeva, K. (2013). Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India. *Comptes Rendus Geoscience*, 345(1), 22-27. https://doi.org/10.1016/j.crte.2012.12.001

Poornima, S., & Pushpalatha, M. (2019). Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units. *Atmosphere*, 10(11), 668.

Salman, A. G., Heryadi, Y., Abdurahman, E., & Suparta, W. (2018). Weather Forecasting Using Merged Long Short-Term Memory Model (LSTM) and Autoregressive Integrated Moving Average (ARIMA) Model. *J. Comput. Sci.*, 14(7), 930-938.

Sanguansat, P. (2019). *Artificial Intelligence with Machine Learning*. IDC Premier Limited.

Shewalkar, A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, 9(4), 235--245.

Srachoom, C. (2007). *Application of Artificial Neural Network for Weather Forecast* Chiang Mai University]. Chiang Mai.

Sukawat, D. *Weather forecast Knowledge*. Retrieved 7 April from https://www.tmd.go.th/info/info.php?FileID=1

Wang, S. W., Feng, J., & Liu, G. (2013). Application of seasonal time series model in the precipitation forecast. *Mathematical and Computer Modelling*, 58(3-4), 677-683. https://doi.org/10.1016/j.mcm.2011.10.034

Wangdi, K., Singhasivanon, P., Silawan, T., Lawpoolsri, S., White, N. J., & Kaewkungwal, J. (2010). Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. *Malaria Journal*, 9(1), 1-9.

Yang, S., Yu, X., & Zhou, Y. (2020). LSTM and GRU neural network performance comparison study: Taking Yelp review dataset as an example. 2020 International workshop on electronic communication and artificial intelligence (IWECAI),

ประวัติผู้เขียน

ชื่อ-สกุล	รักษ์คณา ภูสีเขียว
วัน เดือน ปี เกิด	9 ธันวาคม 2536
สถานที่เกิด	กรุงเทพมหานคร
ที่อยู่ปัจจุบัน	7/1 ซอยพหลโยธิน 54/1 แยก 8-4-2 แขวงคลองถนน เขตสายไหม กรุงเทพมหานคร 10220