การพยากรณ์ราคาผักบุ้งไทย ด้วยวิธีการเรียนรู้เชิงลึก

THAI MORNING GLORY PRICE FORECAST USING DEEP LEARNING

กนกวรรณ แววดี

KANOKWAN WAEODI

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2566

KMITL-2023-SC-M-017-073

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

THAI MORNING GLORY PRICE FORECAST USING DEEP LEARNING

KANOKWAN WAEODI

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIRENMENT FOR THE DEGREE OF MASTER OF SCIENCE

IN DATA SCIENCE AND ANALYTICS

KMITL DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2023

KMITL-2023-SC-M-017-073

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2023

SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อการค้นคว้าอิสระ การพยากรณ์ราคาผักบุ้งไทย ด้วยวิธีการเรียนรู้เชิงลึก
ชื่อนักศึกษา นางสาวกนกวรรณ แววดี
รหัสประจำตัว 64605005
ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์)
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง
พ.ศ. 2566
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ รองศาสตราจารย์ ดร.ละออ บุญเกษม

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อทำนายราคาผักบุ้ง ด้วยวิธีการเรียนรู้เชิงลึก เปรียบเทียบกับวิธีการเรียนรู้เชิงลึกร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก ข้อมูลที่ใช้ฝึกตัวแบบมีจำนวน 2,800 เรคอร์ด เป็นข้อมูลระหว่างวันที่ 1 มกราคม 2559 ถึง 31 สิงหาคม 2566 โดยมีการเก็บข้อมูลจาก 4 แหล่ง คือ กรมการค้าภายใน, เว็ปไซต์ wunderground , สถาบันสารสนเทศทรัพยากรน้ำ และเว็ปไซต์ myhora ทั้งหมด 22 คุณลักษณะ จากนั้นนำคุณลักษณะทั้งหมดมาทำการทดสอบ Multicollinearity และคัดเลือกคุณลักษณะด้วยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Selection) ได้คุณลักษณะที่มีความสัมพันธ์กับราคาผักบุ้งสูงที่สุด 4 คุณลักษณะ คือ ค่าความชื้นในอากาศต่ำสุด, ค่าอุณหภูมิสูงสุด, ค่าความกดอากาศสูงสุด และ ค่าความเร็วลมสูงสุด จากนั้นนำตัวแบบมาปรับค่า Hyperparameters คือ epoch, batch size, window_len ผลลัพธ์ที่ได้คือตัวแบบที่ให้ค่าความเคลื่อน MSE ต่ำที่สุด โดยมีค่าเท่ากับ 0.0008 คือ ตัวแบบที่ 2 ที่ epoch เท่ากับ 300, batch size เท่ากับ 32 และ window_len เท่ากับ 5 และตัวแบบที่ 6 ที่ epoch เท่ากับ 300, batch size เท่ากับ 32 และ window_len เท่ากับ 21 ตัวแบบที่ให้ค่า RMSE ต่ำที่สุด โดยมีค่าเท่ากับ 0.0258 คือ ตัวแบบที่ 2 ที่ epoch เท่ากับ 300, batch size เท่ากับ 32 และ window_len เท่ากับ 5 ตัวแบบที่ให้ค่าความเคลื่อน MAPE ต่ำที่สุด โดยมีค่าเท่ากับ 3.7803 คือ ตัวแบบที่ 1 ที่ epoch เท่ากับ 300, batch size เท่ากับ 32 และ window_len เท่ากับ 5 และตัวแบบที่ให้ค่าความเคลื่อน MAE ต่ำที่สุด โดยมีค่าเท่ากับ 0.0157 คือ ตัวแบบที่ 6 ที่ epoch เท่ากับ 300, batch size เท่ากับ 32 และ window_len เท่ากับ 21

คำสำคัญ : การเรียนรู้เชิงลึก, การคัดเลือกคุณลักษณะ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Independent Study Title	Thai Morning Glory Price Forecast Using Deep Learning
Student Name	Miss Kanokwan Waeodi
Student ID	64605005
Degree	Master of Science (Data Science and Analytics) KMITL Digital Analytics and Intelligence Center
Year	2023
Independent Study Advisor	Assoc.Prof. Dr. Laor Longsome

Abstract

The purpose of this study is to predict Thai Morning Glory prices using deep learning techniques compared to deep learning techniques with feature selection. The data used in training the model has 2,800 records from 1 January 2016 to 31 August 2023. The data were collected from 4 sources; the Department of Internal Trade of Thailand, the Wunderground website, the Hydro-Informatics Institute, and the Myhora website, a total of 22 features. Then all features tested multicollinearity and selected features by stepwise selection method. It resulted from a feature selection of four features that had the highest relationship with Thai Morning Glory prices, which are the lowest air humidity, the highest temperature, the highest air pressure, and the highest wind speed. The next step is taking the model to adjust in Hyperparameters which are epoch, batch size and window_len. In conclusion, the model that gives the lowest value of MSE equal to 0.0008 are model 2 that epoch is equal to 300, batch size is equal to 32, and window_len is equal to 5, and model 6 that epoch is equal to 300, batch size is equal to 32, and window_len is equal to 21. Next, the model that gives the lowest value of RMSE equal to 0.0258 is model 2 that epoch is equal to 300, batch size is equal to 32, and window_len is equal to 5. Following, the model that gives the lowest value of MAPE equal to 3.7803 is model 1 that epoch is equal to 300, batch size is equal to 32 and window_len is equal to 5. Lastly, the model that gives the lowest MAE equal to 0.0157 is model 6 that epoch is equal to 300, batch size is equal to 32, and window_len is equal to 21.

Keywords: deep learning, feature selection

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

งานวิจัยฉบับนี้สำเร็จลุล่วงด้วยดีเนื่องจากความกรุณาและความช่วยเหลืออย่างดียิ่งจาก รศ.ดร.ละออ บุญเกษม อาจารย์ที่ปรึกษาการค้นความอิสระ ท่านได้เสียสละเวลาอันมีค่ายิ่ง ในการให้คำปรึกษาการดำเนินงานวิจัย ตลอดจนได้ตรวจสอบและแก้ไขข้อบกพร่องต่างๆอันเป็นประโยชน์ในการจัดทำงานวิจัยนี้ ตั้งแต่เริ่มดำเนินการจนกระทั่งดำเนินการเสร็จสมบูรณ์ ผู้วิจัยขอกราบขอบพระคุณเป็น อย่างสูงไว้ ณ โอกาสนี้

ขอขอบคุณอาจารย์ ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบังทุกท่าน ที่ให้ความรู้ในการทำงานวิจัยนี้ ตลอดจนเพื่อนๆ คณะวิทยาศาสตร์ สาขาวิทยาการข้อมูลและการวิเคราะห์ที่ให้ความช่วยเหลือในการทำงานวิจัยครั้งนี้

ขอกราบขอบพระคุณพ่อ คุณแม่ ที่ให้คำปรึกษาในเรื่องต่างๆ รวมทั้งเป็นกำลังใจที่ดีเสมอมาในการทำงานวิจัยนี้ให้สมบูรณ์

สุดท้ายนี้ผู้วิจัยหวังว่างานวิจัยฉบับนี้คงเป็นประโยชน์สำหรับหน่วยงานที่เกี่ยวข้อง และผู้ที่สนใจศึกษาต่อไป

นางสาวกนกวรรณ แววดี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ก
บทคัดย่อภาษาอังกฤษ	ข
กิตติกรรมประกาศ	ค
สารบัญ	ง
สารบัญตาราง	ฉ
สารบัญรูป	ช
บทที่ 1 บทนำ	**1**
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของงานวิจัย	1
1.3 ขอบเขตของงานวิจัย	2
1.4 ประโยชน์ที่คาดว่าจะได้รับ	2
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	**3**
2.1 ทฤษฎีที่เกี่ยวข้อง	3
2.1.1 ผักบุ้งไทย	3
2.1.2 ลักษณะทางพฤกษศาสตร์	3
2.1.3 ประโยชน์ผักบุ้งไทย	4
2.1.4 การปลูกผักบุ้งไทย	5
2.1.5 เส้นทางราคาผักและผลไม้ที่มีผลต่อการใช้จ่ายของผู้บริโภค	5
2.1.6 การคัดเลือกคุณลักษณะ (Feature selection)	7
2.1.7 ค่าสัมประสิทธิ์สหสัมพันธ์(Correlation Coefficient)	9
2.1.8 ตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กัน (Multicollinearity)	10
2.1.9 Variance Inflation Factor (VIF)	10
2.1.10 การเข้ารหัสเชิงกลุ่ม	11
2.1.11 การทำให้ข้อมูลเป็นค่ามาตรฐาน (Data Normalization)	14
2.1.12 Long Short-Term Memory (LSTM)	15
2.1.13 วิธีวัดผลความแม่นยำโมเดล	17
2.2 งานวิจัยที่เกี่ยวข้อง	19

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
บทที่ 3 วิธีการดำเนินงานวิจัย	**21**
3.1 ชุดข้อมูล (Dataset)	21
3.2 การจัดเตรียมข้อมูลเพื่อการประมวลผล	23
3.2.1 ETL	24
3.2.2 การสำรวจข้อมูล (data exploration)	24
3.2.3 การจัดการกับข้อมูลสูญหาย	24
3.2.3 การทำให้ข้อมูลเป็นค่ามาตรฐาน (Data Normalization)	25
3.2.4 การเข้ารหัสข้อมูล (One-Hot encoding)	25
3.2.5 การคัดเลือกคุณลักษณะ (Feature selection)	26
3.3 การสร้างตัวแบบการพยากรณ์ (Modelling)	26
3.3.1 อัลกอริทึมที่ใช้	26
3.3.2 การแบ่งข้อมูลสำหรับการฝึกและทดสอบ	26
3.3.3 การสร้างตัวแบบการเรียนรู้เชิงลึก (Deep Learning Model)	26
3.3.4 การหาค่า Hyperparameter ที่เหมาะสม	27
3.4 การวัดประสิทธิภาพของตัวแบบ (Model Evaluation)	27
บทที่ 4 ผลการวิจัยการอภิปรายผล	**28**
4.1 การคัดเลือกคุณลักษณะ	28
4.1.1 การทดสอบ Multicollinearity	28
4.1.2 การเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน	33
4.2 ตัวแบบการพยากรณ์	33
4.3 การปรับค่า Hyperparameters	35
4.4 ผลการศึกษาและการวัดประสิทธิภาพ	37
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ	**38**
5.1 สรุปผลการวิจัย	38
5.2 ข้อเสนอแนะ	39
เอกสารอ้างอิง	40
ประวัติผู้เขียน	41

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่ **หน้า**

ตารางที่ 2.1 แสดงตัวอย่างข้อมูลการทำ One Hot Encoding 11
ตารางที่ 2.2 แสดงตัวอย่างข้อมูลการทำ Ordinal Encoding แบบที่ 1 13
ตารางที่ 2.3 แสดงตัวอย่างข้อมูลการทำ Ordinal Encoding แบบที่ 2 13
ตารางที่ 2.3 เปรียบเทียบงานวิจัยที่ได้ศึกษามา 20
ตารางที่ 3.1 แสดงคุณลักษณะของชุดข้อมูลทั้งหมด 22
ตารางที่ 3.2 แสดงวิธีการจัดการกับข้อมูลสูญหาย 25
ตารางที่ 3.3 แสดงตัวอย่างการทำข้อมูลให้เป็นมาตรฐาน 25
ตารางที่ 3.4 แสดงโครงสร้างตัวแบบของการเรียนรู้เชิงลึก 26
ตารางที่ 3.5 แสดงการกำหนดค่า Hyperparameters ที่ใช้ในตัวแบบ 27
ตารางที่ 4.1 แสดงประสิทธิภาพของตัวแบบการพยากรณ์ 35
ตารางที่ 4.2 แสดงประสิทธิภาพของตัวแบบจากการทดลองกำหนดค่า epoch 36
ตารางที่ 4.3 แสดงประสิทธิภาพของตัวแบบจากการทดลองกำหนดค่า batch size 36
ตารางที่ 4.4 แสดงประสิทธิภาพของตัวแบบจากการทดลองกำหนดค่า window 36
ตารางที่ 4.5 แสดงการเปรียบเทียบประสิทธิภาพของตัวแบบ 37

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่	หน้า
รูปที่ 2.1 แสดงกราฟราคาปุ๋ยเคมีขายปลีกท้องถิ่น (บาท/ตัน)	6
รูปที่ 2.2 แสดงโครงสร้างของ LSTM	16
รูปที่ 3.1 แสดงขั้นตอนหลักในการดำเนินงานวิจัย	21
รูปที่ 3.2 แสดงขั้นตอนการจัดเตรียมข้อมูลเพื่อการประมวลผล	23
รูปที่ 3.3 แสดงตัวอย่างการเข้ารหัสข้อมูลด้วยวิธี One-Hot Encoding ของคุณลักษณะ weekend คุณลักษณะเดิม (ซ้าย) คุณลักษณะที่สร้างขึ้นใหม่ (ขวา)	25
รูปที่ 4.1 ผลลัพธ์จากการทำ Scatter matrix	28
รูปที่ 4.2 ผลลัพธ์จากการทำ Correlation matrix	29
รูปที่ 4.3 ผลลัพธ์จากการทำ Variance Inflation Factor (VIF)	29
รูปที่ 4.4 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 1	30
รูปที่ 4.5 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 2	30
รูปที่ 4.6 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 3	31
รูปที่ 4.7 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 4	31
รูปที่ 4.8 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 5	31
รูปที่ 4.9 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 6	32
รูปที่ 4.10 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 7	32
รูปที่ 4.11 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 8	32
รูปที่ 4.12 ผลลัพธ์จากการคัดเลือกตัวแปรด้วยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน	33
รูปที่ 4.13 แสดงค่าสูญเสียของการฝึก 0 - 100 รอบ (LSTM)	34
รูปที่ 4.14 แสดงค่าสูญเสียของการฝึก 0 - 100 รอบ (LSTM+ปัจจัยภายนอก)	34
รูปที่ 4.15 แสดงค่าสูญเสียของการฝึก 0 - 300 รอบ (ด้านซ้าย) เปรียบเทียบกับ 0 – 100 รอบ (ด้านขวา)	35

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ประเทศไทยมีพื้นที่เกษตรกรรม 153,184,527 ไร่ หรือร้อยละ 47.77 ของพื้นที่ประเทศไทย แบ่งออกเป็น 4 ประเภท คือ กสิกรรม ปศุสัตว์ การประมง และด้านป่าไม้ ส่งผลให้ประเทศไทยมีจำนวนประชากรในภาคเกษตรไม่น้อยกว่าสามสิบล้านคน ดังนั้นการทำเกษตรกรรมจึงมีความสำคัญ เพราะส่งผลให้ประชากรส่วนใหญ่ของประเทศมีรายได้สูงขึ้น และมีความเป็นอยู่ที่ดีขึ้น

เกษตรกรรมประเภทกสิกรรม คือ การเพาะปลูกพืช เช่น การทำนา การทำสวน การทำไร่ เป็นต้น ซึ่งปัญหาหลักของเกษตรกรรมประเภทนี้ คือ เมื่อพืชผักไม่คุ้มกับราคาและเกษตรกรรู้ราคาตลาดไม่ชัดเจน การคาดการณ์ราคาพืชผลทางการเกษตรต่างๆ อย่างทันท่วงทีและเชื่อถือได้มีความจำเป็นอย่างยิ่งและจำเป็น สำหรับวันนี้การพยากรณ์ราคาสินค้าเกษตรยังคงทำได้ยากเนื่องจากได้รับอิทธิพลจากหลายปัจจัย ความไม่แน่นอนของราคา การผลิต และการบริโภคในอนาคต ทำให้กลยุทธ์การตลาดเกษตรและการวางแผนการลงทุนทำได้ยาก ความเน่าเสียง่าย การเปลี่ยนแปลงราคา และธรรมชาติตามฤดูกาลของผักมีผลกระทบอย่างมากต่อราคาผัก

ผักบุ้งไทย ถือเป็นเกษตรกรรมประเภทกสิกรรม มีลักษณะเป็นไม้ล้มลุกมีข้อโปร่งเลื้อยแผ่ไปตามพื้นหรือผิวน้ำนิยมปลูกทั้งในดินและในน้ำ การขยายพันธุ์ เพาะเมล็ด หรือใช้ลำต้นแก่มีรากติด การเก็บเกี่ยวได้ 2 วิธีคือการถอนออกมาทั้งราก และการตัดยอดไปเพียงอย่างเดียว ซึ่งถ้าเป็นวิธีการตัดยอดไปนั้นเราสามารถบำรุงใหม่ด้วยปุ๋ยคอก และหมั่นรดน้ำ ประมาณ 15 วันผักบุ้งก็จะแตกยอดมาอีกครั้ง ซึ่งต้นทุนในการปลูกผักบุ้งไทยนั้น จะมีทั้งค่าปุ๋ย ค่าแรงงาน ค่าขนส่ง ค่าอุปกรณ์ต่างๆในการแพค จะดียิ่งขึ้นถ้าเราสามารถคาดการณ์ราคาผักได้ เพื่อวางแผนการลงทุน

ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะพยากรณ์ราคาผักบุ้ง และศึกษาปัจจัยที่เกี่ยวข้องกับราคาผักบุ้ง เพื่อให้เกษตรกรสามารถวางแผนการลงทุนได้อย่างเหมาะสม

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาปัจจัยภายนอกที่มีผลต่อราคาผักบุ้งไทย
2. เพื่อหาตัวแบบการพยากรณ์ที่เหมาะสมของราคาผักบุ้งไทย
3. เพื่อพยากรณ์ราคาผักบุ้งไทย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.3 ขอบเขตของงานวิจัย

1. การสร้างตัวแบบพยากรณ์ราคาเฉลี่ยผักบุ้งจะเป็นการวิเคราะห์ความเปลี่ยนแปลงราคาเป็นรายวัน ตั้งแต่เดือน มกราคม พ.ศ. 2559 – สิงหาคม พ.ศ. 2566
2. ข้อมูลราคาผักบุ้งไทย เป็นข้อมูลที่เกิดจากการเก็บสถิติจากแหล่งข้อมูลต่าง ๆ ที่กรมการค้าภายในได้เลือกไว้ให้เป็นกลุ่มตัวอย่างที่ใช้ในการศึกษางาน เป็นราคาสินค้าเกษตรในพื้นที่กรุงเทพมหานคร
3. ข้อมูลราคาผักบุ้งที่ได้เป็นราคาในช่วงวันจันทร์ – อาทิตย์
4. ข้อมูลราคาผักบุ้งและข้อมูลปัจจัยนอก เป็นข้อมูลเฉพาะในจังหวัดกรุงเทพมหานคร

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ผลลัพธ์ที่ได้จากการพยากรณ์สามารถนำไปช่วยตัดสินใจเพื่อวางแผนการลงทุนในการผลิตเพื่อให้ไม่เกิดการขาดทุน
2. ทำให้ทราบถึงปัจจัยภายนอกที่ส่งผลต่อราคาผักบุ้งไทย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง เพื่อเป็นแนวทางในการศึกษางานวิจัย การพยากรณ์ราคาผักบุ้งไทย ด้วยวิธีการเรียนรู้เชิงลึก (Thai Morning Glory Price Forecast Deep Learning And External Actor Groups) โดยได้ทำการศึกษาเอกสาร ทฤษฎี งานวิจัย หรือวรรณกรรมที่เกี่ยวข้อง ซึ่งมีสาระสำคัญดังต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ผักบุ้งไทย

ผักบุ้งไทย เป็นพันธุ์ผักบุ้งท้องถิ่น และเป็นพันธุ์ผักบุ้งดั้งเดิมของไทย ซึ่งเป็นที่นิยมนำมารับประทาน และใช้ประกอบอาหารไม่แพ้พันธุ์ผักบุ้งจีน เนื่องจาก มีลำต้น และใบใหญ่ ให้รสกรอบหวาน เนื้อลำต้น และใบไม่เหนียว

ผักบุ้งไทย แบ่งเป็น 2 พันธุ์ คือ พันธุ์ก้านแดง และพันธุ์ก้านเขียวขาว ซึ่งทั้ง 2 พันธุ์ มีลักษณะที่แตกต่างกันอย่างชัดเจนที่สีของลำต้น รวมถึงมีขนาดใบและลำต้นที่ต่างกัน โดยพันธุ์เขียวขาวจะมีลักษณะลำต้น และใบใหญ่กว่า และให้เนื้อที่กรอบมากกว่า จึงเป็นที่ต้องการของตลาด และมีราคาสูงมากกว่าพันธุ์ก้านแดง เนื่องจากพันธุ์ก้านแดงมีลำต้น และใบเล็ก เนื้อลำต้น และใบค่อนข้างเหนียว ซึ่งพันธุ์นี้จะนิยมเก็บมารับประทานเฉพาะส่วนยอดอ่อนเท่านั้น

2.1.2 ลักษณะทางพฤกษศาสตร์

1) พันธุ์ก้านแดง

ลำต้น : ลำต้นพันธุ์ก้านแดงที่อยู่บนบกจะตั้งตรง แต่ไม่สูงมาก สูงประมาณ 5-15 ซม. หากสูงมากกว่านี้ลำต้นจะโน้มลงพร้อมเลื้อย ขนาดลำต้นเล็กประมาณ 0.3-0.8 เซนติเมตร แตกกิ่งได้มาก เป็นพันธุ์ที่มีลำต้นเลื้อยได้ยาวที่สุดในบรรดาผักบุ้งทุกพันธุ์ อาจเลื้อยยาวได้ถึง 10 เมตร โดยเฉพาะเมื่อเวลาที่มีน้ำท่วม และสามารถแตกรากตามข้อได้ดี พันธุ์นี้ลำต้นจะมีสีเขียวเข้ม เขียวอมแดง หรือแดงม่วง มียอดยาวได้มากกว่า 50 เซนติเมตร ลำต้นมีเนื้อค่อนข้างเหนียว แต่ยอดอ่อนให้ความกรอบได้ดี

ใบ : ออกเป็นเดี่ยว สีเขียวเข้ม แทงออกบริเวณข้อตรงข้ามกันเป็นคู่ ใบค่อนข้างแหลมยาว ฐานใบมน ปลายใบแหลม ใบบริเวณยอดจะสั้น ยาว 2-4 เซนติเมตร กว้าง 1-2 เซนติเมตร ใบส่วนโคนจะยาวกว่า 5-10 เซนติเมตร กว้าง 3-5 เซนติเมตร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ดอก : ดอกพันธุ์ก้านแดงจะออกเป็นช่อ กลีบเลี้ยงดอกมีสีเขียว ส่วนกลีบดอกมีสีขาวอมม่วง ดอกมีขนาดเล็กกว่าพันธุ์ก้านเขียวขาว ส่วนอื่นมีลักษณะคล้ายกัน

ผล และเมล็ด : ผลมีลักษณะเป็นแคปซูล ภายในมีเมล็ด 2-4 เมล็ด เมล็ดมีรูปร่างคล้ายสามเหลี่ยม มีสีดำ ส่วนฐานเมล็ดมนใหญ่

2) พันธุ์ก้านเขียวขาว

ลำต้น : ลำต้นบนบกมีลักษณะตั้งตรง พร้อมเลื้อยหากต้นยาวมาก แต่หากอยู่ในแหล่งน้ำจะเป็นเถาเลื้อยลอยน้ำ ลำต้นมีลักษณะอวบใหญ่ ประมาณ 1-2 ซม. ซึ่งจะใหญ่กว่าผักบุ้งไทยพันธุ์ก้านแดง และใหญ่กว่าพันธุ์ผักบุ้งจีน มีความสูงของลำต้นปานกลาง สูงประมาณ 20-30 ซม. สูงกว่าผักบุ้งไทยพันธุ์ก้านแดง แต่จะต่ำกว่าผักบุ้งจีน ลำต้นมีสีเขียวอมขาวเล็กน้อย ลำต้น ทั้งส่วนโคน และยอดอ่อนจะให้เนื้อกรอบได้ดีมาก มีความกรอบมากกว่าผักบุ้งจีน และผักบุ้งพันธุ์ก้านแดง

ใบ : เป็นใบเดี่ยว ออกแบบสลับตรงข้ามกัน ก้านใบใหญ่ยาว ฐานใบใหญ่ มนเป็นรูปหัวใจ ก้านใบ มีสีเขียวอมขาว ใบมีสีเขียวสด ใบยาว 10 – 15 เซนติเมตร กว้าง 5 – 10 เซนติเมตร

ดอก : ดอกพันธุ์ก้านเขียวอมขาวจะออกดอกเป็นสีขาวทั่วดอก มีกลีบเลี้ยงสีเขียว ดอกมีขนาดใหญ่ มีรูปทรงกรวย แทงออกตามซอกใบ ดอกออกเป็นช่อ แต่ละช่อมีดอกย่อย 2-7 ดอก มีเกสรตัวผู้มี 5 อัน ความยาวไม่เท่ากัน

ผล และเมล็ด : ผล และเมล็ดมีลักษณะคล้ายพันธุ์ก้านแดง แต่มีขนาดใหญ่กว่าทั้งฝัก และเมล็ด

2.1.3 ประโยชน์ผักบุ้งไทย

1. ผักบุ้งไทยนิยมนำมาประกอบอาหาร ได้แก่ ผัดผักบุ้ง ลวกผักบุ้ง ใส่ก๋วยเตี๋ยว เป็นต้น

2. ผักบุ้งไทย นิยมเก็บมาเลี้ยงสัตว์หลายชนิด ได้แก่ เลี้ยงกระต่าย เต่า หมู เป็ด ไก่ ห่าน เป็นต้น การเลี้ยงจะใช้วิธีการสับเป็นชิ้นเล็กผสมกับอาหารพวกรำ และข้าวนึ่ง แต่สัตว์ที่กินเฉพาะพืช เช่น เต่า กระต่าย จะให้ทั้งต้น

3. ผักบุ้งไทย นิยมใช้เป็นอาหารปลา เช่น บ่อเลี้ยงปลานิล ปลาตะเพียน เป็นต้น รวมถึงใช้ปลูกหรือปล่อยตามบ่อเลี้ยงปลา เพื่อให้เป็นแหล่งพักอาศัยของปลาวัยอ่อน หรือเป็นแหล่งวางไข่สำหรับปลาบางชนิด

4. ใช้สำหรับเป็นพืชพรรณไม้น้ำประดับตู้ปลา

5. ผักบุ้งไทยพันธุ์ก้านแดงมีลำต้นเลื้อยยาว เกษตรกรมักปล่อยให้เลื้อยาวเพื่อคลุมหน้าดินในบางจุด เช่น ริมขอบสระ แต่ในบางครั้ง หากมีการแพร่กระจายมากมักจะมีข้อเสียที่เป็นพืชรุกรานชนิดหนึ่ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.4 การปลูกผักบุ้งไทย

ผักบุ้งไทยตามธรรมชาติสามารถขยบายพันธุ์ และเติบโตได้ด้วยการแตกเหง้าใหม่ และการเจริญจากเมล็ด โดยพันธุ์ก้านแดง เป็นพันธุ์ที่พบแพร่กระจายทั่วไปตามหัวไร่ปลายนา และเป็นพันธุ์ที่พบมากบนบก เป็นพันธุ์ที่เกษตรไม่นิยมปลูก แต่สามารถเลี้ยง และดูแลตามแหล่งธรรมชาติทั่วไปเพื่อนำมาประกอบอาหาร และจำหน่าย

พันธุ์ก้านเขียวขาว เป็นพันธุ์ที่พบได้น้อยกว่าพันธุ์ก้านแดง ส่วนมากพบในแหล่งน้ำ ตามแม่น้ำ ลำคลอง สระน้ำ บึง มากกว่าที่จะพบบนบก และเป็นพันธุ์ที่เกษตรกรเลี้ยงเพื่อนำมารับประทาน และเพื่อจำหน่ายมากกว่าพันธุ์ก้านแดง ซึ่งทั่วไปนิยมเลี้ยงในแหล่งน้ำต่างๆ

การปลูก และขยายพันธุ์ผักบุ้งไทยทั้ง 2 พันธุ์ สามารถทำได้ด้วยการขยายลำต้นใหม่ ด้วยการนำเหง้าหรือลำต้นไปปล่อยเลี้ยงตามแหล่งน้ำ ซึ่งผักบุ้งจะแตกเหง้าใหม่เองตามธรรมชาติ ส่วนการปลูกหรือการขยายพันธุ์บนบกไม่เป็นที่นิยม เพราะผักบุ้งไทยสามารถเก็บหาได้ทั่วไปในทุกที่ และลำต้น และใบที่เติบโตบนบกจะค่อนข้างเล็ก และให้เนื้อเหนียวมากกว่าการปลูกขยายพันธุ์ในน้ำ [2]

2.1.5 เส้นทางราคาผักและผลไม้ที่มีผลต่อการใช้จ่ายของผู้บริโภค

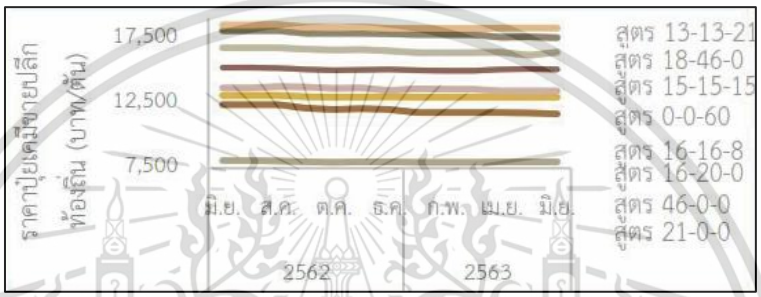
เมื่อพิจารณาถึงปัจจัยที่ทำให้ราคาผักและผลไม้เปลี่ยนแปลง จะพบว่าราคาสามารถเปลี่ยนแปลงไปได้จากปัจจัยทั้งด้านอุปทานและอุปสงค์เช่นเดียวกับสินค้าทั่วไปอื่น ๆ โดยปัจจัยด้านอุปทาน (Supply) จะส่งผลต่อปริมาณและคุณภาพของผลผลิต อันเป็นผลมาจากหลายปัจจัย ทั้งที่ควบคุมไม่ได้และควบคุมได้ยาก ได้แก่

1. สภาพดิน เป็นปัจจัยพื้นฐานในการเจริญเติบโตของพืชผักผลไม้ โดยความสมบูรณ์ของดิน (Soil Fertility) เป็นเครื่องบ่งชี้ความสามารถในการให้ผลผลิต พิจารณาจากธาตุอาหารในปริมาณที่เหมาะสมและสมดุล จากข้อมูลของกรมพัฒนาที่ดิน กระทรวงเกษตรและสหกรณ์ ที่ประเมินความสมบูรณ์ของดินจากข้อมูลปี 2547-2552 และข้อมูลเนื้อที่ใช้ประโยชน์ทางการเกษตร ปี 2562 ของสำนักงานเศรษฐกิจการเกษตร พบว่า ความสมบูรณ์ของทรัพยากรดินในประเทศไทย สอดคล้องกับพื้นที่การเพาะปลูกสวนผักโดยสัดส่วนเนื้อที่การใช้ประโยชน์ทางการเกษตรของสวนผักของภาคกลางมีสัดส่วนมากที่สุดในประเทศ (ร้อยละ 36.22) ตามด้วยภาคเหนือ (ร้อยละ 31.94 ภาคตะวันออกเฉียงเหนือ (ร้อยละ 22.69) และภาคใต้ (ร้อยละ 9.15) ตามลำดับ ขณะที่สวนไม้ผลหรือแหล่งผลิตผลไม้ ส่วนใหญ่กลับพบมากในภาคใต้ ร้อยละ 53.08 ตามด้วยภาคกลางร้อยละ 20.08 ภาคตะวันออกเฉียงเหนือ และภาคเหนือร้อยละ 15.98 และ 10.86 ตามลำดับ ซึ่งจะเห็นได้ว่าไม่ได้สอดคล้องกับคุณภาพดิน

2. น้ำ เป็นวัตถุดิบสำคัญต่อการเพาะปลูกผักผลไม้ และเป็นปัจจัยหนึ่งซึ่งเกษตรกรควบคุมได้ยาก หากปริมาณน้ำมีมากหรือน้อยเกินไป จะส่งผลให้ผักและผลไม้เสียหาย ผลผลิตออกสู่ตลาดไม่เพียงพอและกระทบต่อราคาสินค้าที่เพิ่มสูงขึ้นในที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. ธาตุอาหารหรือปุ๋ย เปรียบเสมือนอาหารเสริมสำหรับผักและผลไม้ที่ช่วยในการเจริญเติบโตให้มีผลผลิตออกสู่ตลาดได้ตามต้องการ ซึ่งผักผลไม้แต่ละชนิด จะมีสูตรและอัตราการใช้ปุ๋ยเคมีที่เหมาะสมแตกต่างกัน เกษตรกรต้องเลือกสูตรปุ๋ยเคมีให้เหมาะสมและคุ้มค่ากับผลผลิตมากที่สุด เนื่องจากปุ๋ยเคมีถือว่าเป็นต้นทุนที่มีมูลค่าค่อนข้างสูงในการผลิต อย่างไรก็ตาม ราคาปุ๋ยเคมีขายปลีกในท้องถิ่น (อ้างอิงจากข้อมูลตารางราคาปุ๋ยเคมีสูตรที่สำคัญ ณ ระดับราคาขายส่งกรุงเทพฯ และราคาขายปลีกท้องถิ่นรายเดือน ปี 2562 - 2563 ของสำนักงานเศรษฐกิจการเกษตร) ราคาค่อนข้างคงที่ จึงอาจสรุปได้ว่าราคาผักและผลไม้ที่เปลี่ยนแปลงไปไม่ได้เกิดจากการเปลี่ยนแปลงของราคาปุ๋ยเคมีโดยตรง แสดงดังรูปที่ 2.1



รูปที่ 2.1 แสดงกราฟราคาปุ๋ยเคมีขายปลีกท้องถิ่น (บาท/ตัน)

4. อากาศ แสงสว่าง และอุณหภูมิ เป็นปัจจัยที่ควบคุมไม่ได้หรือหากต้องการควบคุมต้องมีการลงทุนมูลค่ามาก เช่น การสร้างโรงเรือนควบคุมอุณหภูมิ เป็นต้น ดังนั้นเกษตรกรในประเทศส่วนใหญ่จึงเลือกที่จะเพาะปลูกผักและผลไม้ให้ตรงกับสภาพอากาศและฤดูกาล เพื่อให้ผลผลิตมีปริมาณเหมาะสม คุณภาพสูง และขายได้ราคามากที่สุด รวมทั้งยังสามารถลดค่าใช้จ่ายในการใช้ปุ๋ยเคมีและยาฆ่าแมลงของเกษตรกรได้ อย่างไรก็ตาม สภาพอากาศที่ร้อนจัดในช่วงฤดูร้อนยังคงเป็นปัจจัยสำคัญ ให้ผักบางชนิด (โดยเฉพาะผักโบ เช่น ผักชีและขึ้นฉ่าย) ปรับราคาสูงขึ้นเนื่องจากอากาศที่ร้อนจัดทำให้อายุของผักใบบนชั้นวางสินค้าสั้นลง รวมถึงการเก็บรักษาเป็นไปได้ยาก สำหรับในมุมของผู้บริโภคที่ต้องการ ผักที่สดใหม่ มีความเสี่ยงจากยาฆ่าแมลงปนเปื้อนต่ำ หาซื้อได้ง่าย และราคาถูก

นอกจากนี้ พื้นที่ที่มีสภาพอากาศ แตกต่างกันก็จะส่งผลต่อช่วงเวลาที่ผลผลิตออกสู่ตลาด เช่น ทุเรียนที่ปลูกในภาคกลางจะออกสู่ตลาดก่อนทุเรียนที่ปลูกในภาคใต้ และเมื่อพ้นฤดูกาลแล้ว หากมีการวางจำหน่ายจะมีราคาสูงกว่าปกติมาก นอกจากนี้ยังมีปัจจัยอื่นที่มีความสำคัญต่อการเปลี่ยนแปลงราคาเช่นกัน อาทิ ราคาน้ำมัน เนื่องจากเป็นต้นทุนทางตรงในการขนส่ง และเป็นต้นทุนทางอ้อมของวัตถุดิบที่ใช้เพาะปลูก หรืออัตราแลกเปลี่ยน ที่เป็นปัจจัยในการกำหนดราคาปุ๋ยเคมีที่มีการนำเข้าจากต่างประเทศ เป็นต้น[5]

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.6 การคัดเลือกคุณลักษณะ (Feature selection)

1) การเลือกตัวแปรโดยวิธีนำตัวแปรเข้าทั้งหมด (Enter Regression)

เป็นวิธีการเอาตัวแปรอิสระทุกตัวทั้งตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญ ทางสถิติและไม่มีนัยสำคัญทางสถิติเข้าไปวิเคราะห์ในสมการถดถอย ซึ่งการคัดเลือกตัวแปรอิสระด้วยวิธีพิจารณา ทุกตัวแบบการถดถอยสรุปเป็นขั้นตอนดังนี้

1.1 จากตัวแปรอิสระทั้งหมด k ตัว ที่คาดว่ามีความสามารถในการทำนายค่าของตัวแปรตาม (Y) นำตัว แปรอิสระทีละ 1 ตัวแปรสร้างตัวแบบการถดถอยด้วยวิธีกำลังสองน้อยที่สุดแล้วเพิ่มตัวแปรอิสระเข้าไปในตัวแบบ การถดถอยทีละ 1 ตัวแปร ได้ตัวแบบการถดถอยทั้งหมด 2^k - 1 ตัวแบบ

1.2 คำนวณเกณฑ์ต่างๆ ของแต่ละตัวแบบการถดถอย ได้แก่ ค่าความคลาดเคลื่อนมาตรฐาน (S) โดย S = $\sqrt{MSE}$, ค่าสัมประสิทธิ์การกำหนดพหุคูณ (R^2) , ค่าสัมประสิทธิ์การกำหนดพหุคูณปรับแก้ (R^2_{adj}) และ ค่า Mallow's C_p และสามารถเขียนตารางหรือกราฟแสดงค่าเกณฑ์ต่างๆ ของแต่ละตัวแบบการถดถอยเพื่อ สะดวกแก่การเปรียบเทียบและการคัดเลือก

1.3 จากตารางหรือกราฟแสดงค่าเกณฑ์ในขั้นตอนที่ 2 พิจารณาว่ามีตัวแบบการถดถอยใดบ้างที่อยู่ในข่าย ของการถูกเลือก อาจพิจารณาจากเกณฑ์เดียวหรือใช้หลายเกณฑ์ประกอบกัน ตัวแบบที่เลือกอาจมีตัวแบบเดียว หรือหลายตัวแบบขึ้นอยู่กับความเหมาะสมในการใช้งาน เช่น เมื่อใช้เกณฑ์ของค่า (R^2) จะเลือกตัวแบบที่ให้ค่า R^2 สูงสุด หากใช้เกณฑ์ค่า (S) จะเลือกตัวแบบที่ให้ค่า S ต่ำสุดหรือใกล้เคียงต่ำสุด เป็นต้น จุดบกพร่องของ วิธีการนี้ คือ ทำให้ได้ตัวแบบที่จะพิจารณาค่อนข้างหลากหลายเป็นการคัดเลือกตัวแปรเข้าสู่สมการถดถอยที่ไม่เหมาะสม

2) การเลือกตัวแปรโดยวิธีเพิ่มตัวแปร (Forward Selection)

เป็นวิธีการคัดเลือกตัวแปรอิสระเข้าสู่สมการทีละตัวตามลำดับความสัมพันธ์กับ Y โดยตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดจะถูกคัดเลือกเข้าก่อน โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ อย่างง่าย (Simple correlation coefficient) สูงสุด เมื่อตัวแปรถูกคัดเข้าสมการแล้วจะทำการทดสอบว่าตัวแปร อิสระนั้นสามารถทำนายตัวแปรตามได้อย่างมีนัยสำคัญทางสถิติหรือไม่ ซึ่งจะสามารถคัดตัวแปรอิสระนี้เข้าไปใน ตัวแบบการถดถอยได้ก็ต่อเมื่อผลการทดสอบสมมุติฐานด้วยค่าสถิติ F หรือค่าสถิติ t พบว่ามีนัยสำคัญทางสถิติ หมายถึง ตัวแปรอิสระนี้มีความสำคัญในการทำนายค่าของ Y แต่หากผลสรุปพบว่าไม่มีนัยสำคัญทางสถิติ หมายถึงตัวแปรอิสระนี้ไม่มีความสำคัญในการทำนายค่าของ Y จากนั้นจะทำการคัดเลือกตัวแปรอิสระที่มี ความสัมพันธ์กับตัวแปรตามอันดับถัดไปเข้าสมการแล้วทำการทดสอบว่าตัวแปรอิสระที่อยู่ในสมการสามารถ ร่วมกันทำนายตัวแปรตามได้อย่างมีนัยสำคัญทางสถิติหรือไม่ ทำเช่นนี้ไปจนกว่าจะไม่มีตัวแปรอิสระใดเข้าไปใน สมการได้อีกจึงหยุดการคัดเลือกตัวแปรอิสระ ถือว่าสมการที่ได้นั้นมีความเหมาะสม จุดบกพร่องของวิธีนี้คือไม่ได้ ตรวจสอบผลกระทบที่เกิดจากการเพิ่มตัวแปรพยากรณ์ตัวใหม่เข้าไปในตัวแบบต่อตัวแปรพยากรณ์ที่เข้าไปในตัว แบบก่อนหน้านี้แล้ว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3) การเลือกตัวแปรโดยวิธีลดตัวแปร (Backward Elimination)

เป็นวิธีการคัดเลือกตัวแปรอิสระออกจากสมการทีละตัวแปร โดยเริ่มจากการสร้างสมการถดถอยที่นำตัวแปรอิสระทุกตัวเข้าสู่สมการ แล้วจึงคัดเลือกตัวแปรอิสระออกทีละตัว โดยตัวแปรอิสระที่มี ความสัมพันธ์กับตัวแปรตามน้อยที่สุดจะถูกคัดออก ตัวแปรอิสระตัวแรกที่ถูกพิจารณาคัดออกจะมีความสำคัญต่อ การทำนายค่าของ Y น้อยที่สุด โดยมีค่าสถิติ Partial F ต่ำที่สุด หรือมีค่าสถิติ t แบบไม่คิดเครื่องหมาย (t_0) ต่ำที่สุด แต่จะสามารถนำตัวแปรอิสระแรกออกจากตัวแบบการถดถอยได้ต้องผ่านการทดสอบสมมุติฐานว่า ตัวแปรอิสระนี้ไม่มีความสำคัญต่อการทำนายค่าของ Y คือ มีค่า $F_0 < F_{out}$ หรือมีค่า $|t_0| < t_{out}$ หรือมี p value $> p_{out}$ เมื่อค่า F ที่ใช้คัดตัวแปรอิสระออก (F_{out}) คือค่าวิกฤต $F_{\alpha,(1,n-p)}$ ค่า t ที่ใช้คัดตัวแปร อิสระออก (t_{out}) คือค่าวิกฤต , $t_{\frac{\alpha}{2},(n-p)}$ โดยที่ p = k+1 คือ จำนวนพารามิเตอร์ในตัวแบบการถดถอยเริ่มต้นที่มี ตัวแปรอิสระครบเท่ากับ k ตัว และค่า p value ที่ใช้คัดตัวแปรอิสระออก (p_{out}) คือ ระดับนัยสำคัญ จะดำเนินการลดตัวแปรอิสระในตัวแบบการถดถอย ต่อไปทำการทดสอบว่าตัวแปรที่ยังคงอยู่สามารถร่วมกัน ทำนายตัวแปรตามได้อย่างมีนัยสำคัญทางสถิติหรือไม่ ถ้าไม่ได้ก็จะคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัว แปรตามน้อยอันดับถัดมาออกจากสมการ แล้วดูว่าสมการที่เหลือตัวแปรอิสระอยู่มีนัยสำคัญทางสถิติหรือไม่ ถ้ามีนัยสำคัญทางสถิติก็จะหยุดการคัดออก แต่ถ้าไม่มีนัยสำคัญทางสถิติก็จะทำการคัดเลือกตัวแปรอิสระออกต่อไป เรื่อยๆจนกว่าจะไม่มีตัวแปรอิสระที่ถูกคัดออกอีกการคัดเลือกจะสิ้นสุดเมื่อตัวแปรอิสระที่เหลืออยู่ในสมการทุกตัว แปรมีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติ จุดบกพร่องของวิธีนี้ คือ ไม่ทราบว่าตัวแปรอิสระที่อยู่ 5 ในสมการแต่ละตัวสามารถทำนายตัวแปรตามได้มากหรือไม่บอกได้เพียงว่าตัวแปรอิสระในตัวแบบนั้นสามารถ ร่วมกันทำนายตัวแปรตามได้

4) การเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Selection)

เป็นวิธีการคัดเลือกผสมผสานระหว่างวิธีการคัดเลือกตัวแปรอิสระทั้งแบบการเพิ่มตัวแปรและการ ลดตัวแปรเข้าด้วยกัน วิธีการถดถอยแบบขั้นตอนสามารถทำได้ ดังนี้

4.1 เหมือนกับวิธีคัดเลือกตัวแปรโดยเพิ่มตัวแปร คือคัดเลือกตัวแปรอิสระ 1 ตัวเข้าไว้ในตัวแบบการ ถดถอย โดยเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม Y สูงสุด หรือที่ทำให้ตัวแบบการถดถอยมีค่า SSE ต่ำสุดไว้ในตัวแบบการถดถอยเป็นตัวแปรแรก จากนั้นคำนวณค่าสถิติ Partial F หรือค่าสถิติ t เพื่อทดสอบ สมมุติฐานว่าตัวแปรอิสระมีความสำคัญต่อการทำนายค่าของ Y หรือไม่ โดยถ้า $F_0 \le F_{IN} = F_{\alpha,(1,n-2)}$ หรือ $|t_0| < t_{IN} = t_{\frac{\alpha}{2},(n-p)}$ หรือ p value $\ge p_{IN}$ เมื่อ p_{IN} คือระดับนัยสำคัญสำหรับการคัดตัวแปรอิสระเข้า แสดง ว่าตัวแปรอิสระนั้นไม่มีความสำคัญต่อการทำนายค่าของ Y จะสิ้นสุดการคัดเลือกตัวแปรอิสระเข้าไว้ในตัวแบบ การถดถอย แต่ถ้า $F_0 > F_{IN}$ หรือ $|t_0| > t_{IN}$ หรือ p value $< p_{IN}$ แสดงว่าตัวแปรอิสระนั้นมีความสำคัญ ต่อการทำนายค่าของ Y ซึ่งจะคัดเข้าสู่ตัวแบบการถดถอยเป็นตัวแรก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2 ใช้วิธี Blackward elimination คัดตัวแปรอิสระนั้นออกจากตัวแบบการถดถอย หากเป็นไปได้ โดย จะสามารถคัดตัวแปรอิสระที่อยู่ในตัวแบบออกได้ หาก $F_0 \leq F_{out} = F_{\alpha,(1,n-2)}$ หรือ $|t_0| < t_{out} = t_{\frac{\alpha}{2},(n-2)}$ หรือ p-value > p_{out} เมื่อ p_{out} คือระดับนัยสำคัญสำหรับการคัดตัวแปรอิสระออก การดำเนินการคัดเลือกตัวแปร อิสระจะสิ้นสุดลงและสรุปว่าไม่มีตัวแปรอิสระใดเลยที่มีความสำคัญต่อการทำนายค่าของตัวแปรตาม Y แต่ถ้า $F_0 \geq F_{out} = F_{\alpha,(1,n-2)}$ หรือ $|t_0| \geq t_{out} = t_{\frac{\alpha}{2},(n-2)}$หรือ p-value ≤ p_{out} แสดงว่าตัวแปรอิสระนั้นมีความสำคัญ ต่อการทำนายจะไม่สามารถคัดออกจากตัวแบบการถดถอยได้

4.3 พิจารณาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรอิสระกับ Y โดยจะเลือกตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนกับตัวแปรตามสูงที่สุด โดย ถ้า $F_0 > F_{IN}$ หรือ $|t_0| > t_{IN}$ หรือ p-value < p_{IN} แสดงว่าตัวแปรอิสระที่ถูกพิจารณาตัวต่อมามีความสำคัญต่อการทำนายค่าของ Y เมื่อมี ตัวแปรอิสระก่อนหน้าอยู่ในตัวแบบการถดถอยก่อนแล้ว จะคัดตัวแปรอิสระที่ถูกพิจารณาเข้ามาอยู่ในตัวแบบเป็น ตัวต่อมา

4.4 หลังจากนั้นจึงทำการทดสอบความมีนัยสำคัญของตัวแปรที่เข้ามาเป็นตัวสุดท้ายก่อนโดยพิจารณา จากค่าสถิติเอฟบางส่วน ถ้ามีค่าน้อยกว่าเกณฑ์ที่กำหนดไว้จะต้องนำตัวแปรนั้นออกจากตัวแบบ แต่ถ้าค่าสถิติเอฟ บางส่วนมีค่ามากกว่าเกณฑ์ที่กำหนดไว้ ก็จะย้อนกลับไปทดสอบตัวแปรอิสระตัวก่อนหน้าที่เข้ามาอยู่ในตัวแบบ โดยพิจารณาจากค่าสถิติเอฟบางส่วน เช่นเดียวกัน ทำซ้ำในขั้นตอนที่สามและสี่ต่อไปเรื่อยๆ จนกระทั่งไม่สามารถ 6 นำตัวแปรอิสระใดออกจากตัวแบบได้ หรือไม่สามารถนำตัวแปรอิสระใดเข้าสู่ตัวแบบได้อีก ข้อดีของวิธีการนี้ คือ สามารถแก้จุดบกพร่องของวิธีการเลือกตัวแปรโดยวิธีเพิ่มตัวแปร (Forward Selection) และวิธีการเลือกตัว แปรโดยวิธีลดตัวแปร (Backward Elimination) ได้ [8]

2.1.7 ค่าสัมประสิทธิ์สหสัมพันธ์

Correlation Coefficient ค่าสัมประสิทธิ์สหสัมพันธ์ หรืออีกชื่อคือ Pearson Correlation เป็นค่าที่บ่งชี้ถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัวโดยจะแทนด้วยสัญลักษณ์ “r” พูดง่าย ๆ คือเป็นค่าที่บ่งบอกถึงความสัมพันธ์ของตัวแปร 2 โดยที่ค่าสัมประสิทธิ์สหสัมพันธ์จะมีค่าอยู่ระหว่าง -1.0 จนถึง +1.0 โดยหากพบว่าค่า r เข้าใกล้ -1.0 หมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันในเชิงตรงกันข้าม แต่หากค่า r มีค่าเข้าใกล้ +1.0 หมายความว่าตัวแปรทั้งสองมีความสัมพันธ์ไปในทิศทางเดียวกัน แต่ถ้าตัวแปรทั้งสองมีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0 หมายความว่าตัวแปรทั้งสองไม่มีความสัมพันธ์กัน โดยสูตรในการคำนวณหาค่าจะเป็นดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$r = \frac{\sum(x_i - \bar{x})\,(Y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(Y_i - \bar{Y})^2}}$$

โดยกำหนดให้

r	คือ	ค่าสัมประสิทธิ์สหสัมพันธ์
x_i	คือ	ค่าตัวแปร x ณ ชุดข้อมูลที่ i
$\bar{x}$	คือ	ค่าเฉลี่ยของตัวแปร x
Y_i	คือ	ค่าตัวแปร y ณ ชุดข้อมูลที่ i
$\bar{Y}$	คือ	ค่าเฉลี่ยของตัวแปร y

เมื่อเราคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ได้แล้วหลายคนมักจะมีความเข้าใจว่า ค่าที่ได้เป็นตัวบ่งบอกถึงตัวแปร x มีความสัมพันธ์เป็นเหตุเป็นผลกับ y ในลักษณะของการหาค่า y จากตัวแปร x ซึ่งในความเป็นจริงนั้นเป็นความเข้าใจที่ไม่ถูกต้อง ตัวอย่างเช่น กำหนดให้ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร x และ y เท่ากับ 0.9 ไม่ได้หมายความว่า x จะมีค่าเป็น 0.9 เท่าของ y แต่เป็นการบ่งบอกว่าเมื่อนำค่า x และ y มาพลอตกราฟก็จะพบว่าจุดได้เรียงกันเกือบจะเป็นเส้นตรงนั่นเองเพราะค่าที่ได้เข้าใกล้ 1 แต่ในทางกลับกัน ถ้าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเข้าใกล้ 0 กราฟที่ได้จะกระจัดกระจายและเรียงกันไม่เป็นเส้นตรง โดยถ้าเราต้องการหาค่าตัวแปร y จาก x จำเป็นจะต้องนำค่า x และ y ไปหา Linear Regression ต่อนั่นเอง[9]

2.1.8 ตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กัน (Multicollinearity)

เงื่อนไขข้อหนึ่งในการวิเคราะห์ความถดถอยเชิงพหุ กำหนดให้ตัวแปรอิสระแต่ละตัวต้องเป็นอิสระต่อกัน โดยการตรวจสอบเงื่อนไขนี้จะตรวจสอบโดยใช้ค่าสถิติ คือ Tolerance และ ค่า Variance Inflation Factor (VIF) ถ้าหากค่า Tolerance ของตัวแปร เข้าใกล้ 1 แสดงว่าตัวแปรเป็นอิสระจากกัน แต่ถ้าค่าใกล้ ศูนย์แสดงว่าเกิดปัญหา Multicollinearity และ ค่า Variance Inflation Factor หากมีค่าใกล้ 10 มากแสดงว่าระดับความสัมพันธ์ของตัวแปรอิสระในสมการการวิเคราะห์ความถดถอยพหุเชิงเส้นมีมาก นั้นคือ เกิดปัญหา Multicollinearity

2.1.9 Variance Inflation Factor (VIF)

เพื่อให้เกิดการสรุปผลที่แน่นอนมากกว่า Correlation matrix จึงควรมี Statistic 1 ค่า เพื่อเทียบกับ Threshold ว่าเกิด Multicollinearity ขึ้นหรือไม่ วิธีการที่นิยมใช้เป็นสากลคือ Variance Inflation Factor (VIF)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

VIF เป็นการรัน Regression ระหว่าง Independence variables ด้วยกันเอง โดยให้ 1 ตัวแปรเป็น Dependence variable ในขณะตัวแปรที่เหลือเป็น Independence variables สลับกันไปเรื่อย ๆ จนทุกตัวแปรเป็น Dependence variable ในแต่ละรอบการรันสามารถหาค่า VIF ได้จาก R^2 จากสูตรตามด้านล่าง Threshold ที่นิยมใช้สำหรับ VIF อาจมีตั้งแต่ 2 ไปจนถึง 10 โดย Threshold ที่มากขึ้นหมายถึงการ Relax model มากขึ้นด้วยเช่นกัน Blog ตอนนี้ขอใช้ Threshold กลาง ๆ ที่ 5 เพื่อ Detect multicollinearity ใน Independence variables โดยสูตรในการคำนวณหาค่าจะเป็นดังนี้ [10]

$$VIF_i = \frac{1}{1 - R_i^2}$$

2.1.10 การเข้ารหัสเชิงกลุ่ม

โมเดล Machine Learning เป็นซอฟต์แวร์ที่อยู่ในคอมพิวเตอร์ ไม่ว่าอย่างไรคอมพิวเตอร์ก็ต้องรับข้อมูลที่เป็นตัวเลขเท่านั้น ข้อมูลทุกอย่างไม่ว่าจะเป็น รูปภาพ, เสียง, วิดีโอ, ตัวหนังสือ, ตัวเลข , ข้อมูล Category จะต้องถูกแปลงเป็นตัวเลขในรูปแบบที่ต้องการ เรียกว่า Encoding [5]

1) การเข้ารหัสแบบวัน-ฮอท One Hot Encoding

การเข้ารหัสแบบวัน-ฮอท เป็นวิธีการเข้ารหัสสำหรับข้อมูลนามบัญญัติ โดยใช้การแทนข้อมูลแบบทวิภาค (Binary) แทนแต่ละค่าที่เป็นไปได้ของตัวแปร จากตัวแปร 1 ตัวแปร จะถูกขยายจำนวนตัวแปรเท่ากับจำนวนค่าที่เป็นไปได้ของตัวแปรนั้น ค่าที่แทนลงในตัวแปรที่ขยายมานี้ จะมีค่าเป็น 0 ทั้งหมด ยกเว้นในตัวแปรที่สัมพันธ์กับค่าตัวแปรตั้งต้น จะมีค่าเป็น 1 เพื่อให้เข้าใจมากยิ่งขึ้นจะขอยกตัวอย่างดังต่อไปนี้

ตัวอย่าง จากตัวแปร Sex ของข้อมูลตัวละครจากหนังสือการ์ตูน Marvel มีค่าที่เป็นไปได้ 4 ค่า คือ Agender Female Genderfluid และ Male สามารถทำการ เข้ารหัสแบบวัน-ฮอท โดยขยายตัวแปรออกเป็น 4 ตัวแปร ได้ดังนี้

ตารางที่ 2.1 แสดงตัวอย่างข้อมูลการทำ One Hot Encoding

Sex	isAgender	isFemale	isGenderfluid	isMale
Agender	1	0	0	0
Female	0	1	0	0
Genderfluid	0	0	1	0
Male	0	0	0	1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวแปรใหม่ แต่ละ ตัวแปร ที่ ขยาย ออก ออกมานั้น มีความสัมพันธ์ กับตัวแปรตั้งต้น โดยค่าที่เป็นไปได้ 1 ค่าจะถูกสร้างตัวแปรใหม่ 1 ตัวแปรขึ้นมาแทน และตัวแปรที่ขยายออกมานี้จะมีค่าที่เป็นไปได้ คือ 0 หรือ 1 เท่านั้น และจะต้องมีเพียงตัวแปรเดียวเท่านั้นที่สามารถมีค่าเท่ากับ 1 ได้ โดย

1.1) ถ้าตัวแปรใหม่ isAgender มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Agender

1.2) ถ้าตัวแปรใหม่ isFemale มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่ตัวแปร Sex เดิมเป็น Female

1.3) ถ้าตัวแปรใหม่ isGenderfuid มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัว แปร Sex เดิมเป็น Genderfluid

1.4) ถ้าตัวแปรใหม่ isMale มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Male

ดังนั้น ข้อมูล Spider-Man มีค่าตัวแปร Sex เท่ากับ Male เมื่อทำการเข้ารหัสแล้ว ตัวแปร isMale ซึ่งขยายมาจากตัวแปร Sex เดิม จะมีค่าเท่ากับ 1 ส่วนตัวแปร isAgender isFemale และ isGenderfluid จะมีค่าเท่ากับ 0

ข้อดีของ One Hot Encoding

1. เป็นรูปแบบที่เข้าใจง่าย การแปลงข้อมูลเป็นรูปแบบของเวกเตอร์ที่มีความเป็นระเบียบและเข้าใจง่าย เนื่องจากแต่ละค่าข้อมูลจะถูกแปลงเป็นเวกเตอร์ที่มีมิติเท่ากับจำนวนข้อมูลแต่ละค่า และมีค่าเป็น 0 หรือ 1 เท่านั้น
2. เหมาะสำหรับข้อมูลที่เป็นการจัดกลุ่มหรือหมวดหมู่: One-Hot Encoding มักถูกนำมาใช้กับข้อมูลที่มีการจัดกลุ่มหรือหมวดหมู่ เช่น สี, ประเทศ, หรือประเภทสินค้า เป็นต้น
3. เหมาะสำหรับการใช้กับการเรียนรู้ของเครื่อง (Machine Learning): สามารถนำเข้าข้อมูลที่ถูก One-Hot Encoded ไปใช้กับโมเดลการเรียนรู้ของเครื่องได้โดยง่าย โดยไม่ต้องมีการปรับเปลี่ยนโครงสร้างของโมเดลมาก

ข้อเสียของ One Hot Encoding

1. เปลือง Memory ที่เก็บค่า 0 เป็นส่วนใหญ่ เรียกว่า Sparse Matrix
2. ความหมายของลำดับข้อมูลแบบ Ordinal จะหายไป เนื่องจากทุก Category แตกต่างกันเท่ากันหมด
3. ข้อมูลจะอยู่ได้แต่ 1 Category เท่านั้น ถ้ามากกว่า 1 จะเรียกว่า Multi-Hot Encoding
4. ถ้าข้อมูลมี Value หลากหลายมาก เช่น มีสีเสื้อ 10,000 สี จะทำให้มีปัญหา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. ถ้าข้อมูลมีการเพิ่ม Category ใหม่ ๆ เล็ก ๆ ยิบย่อยอยู่ตลอด จะทำให้มีปัญหา เช่น เพิ่มสีเสื้อใหม่ [4]

2) การเข้ารหัสเชิงลำดับ (Ordinal Encoding)

การเข้ารหัสเชิงลำดับ เป็นวิธีการเข้ารหัสสำหรับข้อมูลเชิงอันดับ โดยการให้ค่าตัวเลขตามลำดับของค่าข้อมูล แทนการใช้ข้อความ ดังนั้น เมื่อข้อมูลถูกเข้ารหัสแล้วความหมายเชิงลำดับของข้อมูลจะยังคงถูกสงวนไว้ ศึกษาวิธีการเข้ารหัสเชิงลำดับได้ ดังตัวอย่างต่อไปนี้

ตัวอย่าง 2.4 จากตัวแปร Align ของข้อมูลตัวละครจากหนังสือการ์ตูน Marvel มีค่าที่เป็นไปได้ 3 ค่า คือ Good Neutral และ Bad โดย กำหนดให้ลำดับของข้อมูลเป็น Good > Neutral > Bad สามารถทำการเข้าเชิงลำดับ ได้ดังนี้

ตารางที่ 2.2 แสดงตัวอย่างข้อมูลการทำ Ordinal Encoding แบบที่ 1

Align	Encoded Align
Bad	1
Neutral	2
Good	3

สังเกตได้ว่า การเข้ารหัสเชิงอันดับ ไม่ทำให้เกิดการเพิ่มตัวแปรในชุดข้อมูล เพียงแต่เปลี่ยนการแทนค่าข้อมูลเชิงอันดับ จากข้อความเป็นตัวเลขเท่านั้น โดยการให้ค่าตัวเลขสัมพันธ์กับลำดับของค่าข้อมูล

อย่างไรก็ตาม การให้ค่าตัวเลขไม่ได้จำกัดให้เริ่มต้นจากค่ 1 เท่านั้น สามารถเลือกใช้ค่าตัวเลขใดก็ได้ แต่ยังคงต้องรักษาลำดับของค่าข้อมูลให้คงอยู่ ตัวอย่างเช่น

ตารางที่ 2.3 แสดงตัวอย่างข้อมูลการทำ Ordinal Encoding แบบที่ 2

Align	Encoded Align
Bad	-1
Neutral	0
Good	1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อดีของ Ordinal Encoding

1. รักษาความหมายของข้อมูลเรื่องลำดับ การเข้ารหัสเชิงลำดับช่วยรักษาความหมายเกี่ยวกับลำดับหรือตำแหน่งของข้อมูล เหมาะสำหรับข้อมูลที่มีลำดับหรือเรียงลำดับอย่างชัดเจน เช่น ระดับการศึกษา (ประถมศึกษา, มัธยมศึกษา, ปริญญาตรี)

2. ใช้งานได้ง่ายและรวดเร็ว Ordinal Encoding เป็นวิธีการที่ง่ายและรวดเร็วในการแปลงข้อมูลเป็นตัวเลขโดยให้ความหมายที่เกี่ยวข้องกับตำแหน่งหรือลำดับของข้อมูล

3. เหมาะสำหรับโมเดลบางประเภท: สามารถนำข้อมูลที่ถูกเข้ารหัสเชิงลำดับไปใช้กับบางโมเดลที่ต้องการข้อมูลที่มีลำดับ เช่น ตัวแบบทางสถิติหรือสมการทางคณิตศาสตร์ที่ต้องการค่าตัวเลข

ข้อเสียของ Ordinal Encoding

1. ไม่เหมาะสำหรับข้อมูลที่ไม่มีลำดับหรือไม่เกี่ยวข้องกับตำแหน่ง: การเข้ารหัสเชิงลำดับไม่เหมาะสำหรับข้อมูลที่ไม่มีความสัมพันธ์เชิงลำดับ เช่น สี (แดง, เขียว, ฟ้า) หรือประเภทสินค้าที่ไม่มีการจัดลำดับ

2. ไม่รักษาความห่างทางตรรกศาสตร์: การกำหนดตัวเลขในการเข้ารหัสเชิงลำดับอาจไม่รักษาความห่างทางตรรกศาสตร์ (ordinality) ระหว่างข้อมูล ซึ่งอาจทำให้โมเดลมองเห็นความสัมพันธ์ระหว่างข้อมูลที่ไม่ถูกต้อง

3. มีความไม่แน่นอนในการกำหนดค่า: การกำหนดค่าตัวเลขในการเข้ารหัสเชิงลำดับอาจมีความไม่แน่นอน ซึ่งอาจส่งผลให้โมเดลมองเห็นความสัมพันธ์ที่ผิดพลาด[6]

2.1.11 การทำให้ข้อมูลเป็นค่ามาตรฐาน (Data Normalization)

การทำ Feature Scaling คือ วิธีการปรับช่วงขอบเขตของข้อมูลชนิดตัวเลข Cardinal แต่ละ Feature (Field) ให้อยู่ในช่วงเดียวกัน ที่เหมาะกับการนำไปประมวลผลต่อ เข้าสูตรคำนวณได้ง่าย เช่น ช่วง [0, 1] หรือ [-1, 1] ได้ผลลัพธ์อยู่ในช่วงที่กำหนด เรียกว่า Data Normalization นิยมทำในขั้นตอน Preprocessing จัดเตรียมข้อมูล ก่อนป้อนให้โมเดลใช้เทรน [6]

1) Rescaling (Min-Max Normalization)

Rescaling หรือ Min-Max Normalization เป็นวิธีที่ง่ายที่สุด ที่จะปรับช่วงข้อมูล ให้เป็นอยู่ในช่วง [0, 1] ด้วยการนำ Feature / Column นั้น ๆ ลบด้วยค่าที่น้อยที่สุด (Min) ของมัน แล้วหารด้วยช่วงของข้อมูลนั้น (Max – Min)

$$x' = \frac{\mathrm{x} - \min(\mathrm{x})}{\max(\mathrm{x}) - \min(\mathrm{x})}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2) Mean Normalization

Mean Normalization คล้ายกับ Rescaling ด้านบน แตกต่างกันที่ใช้ Mean แทน Min ทำให้ช่วงของ Output [-0.5, 0.5] มีทั้งบวกและลบ Balance กัน ตรงเลข 0 (ขยับ Mean มาตรง 0)

$$x' = \frac{\mathrm{x} - \mathrm{average(x)}}{\mathrm{max(x)} - \mathrm{min\ (x)}}$$

เรามักจะเห็น Mean Normalization อีกแบบ ที่หารด้วย Standard Deviation ของมัน ทำให้ Output ที่ออกมามี Standard Deviation เป็น 1 เรียกว่า Standardization

3) Standardization (Z-Score Normalization)

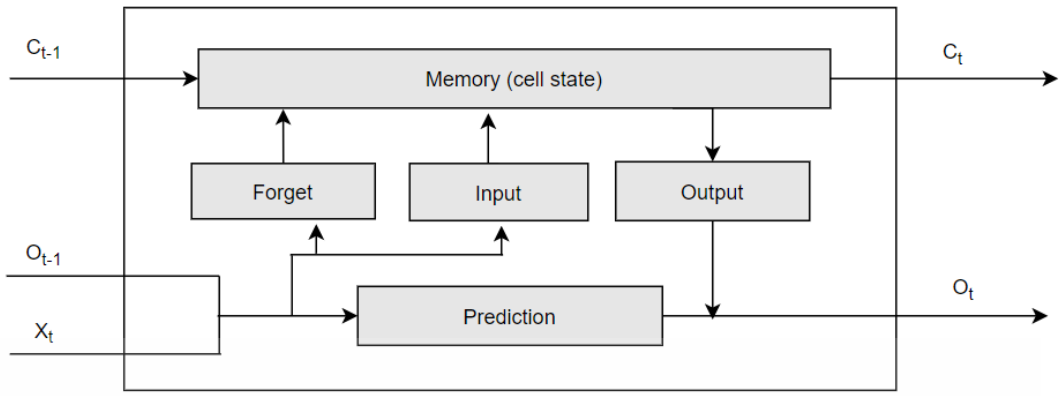
Standardization หรือ Z-Score Normalization คือ การนำข้อมูล Feature / Column มาปรับให้ Mean = 0 และ Standard Deviation = 1 (Unit Variance) [3]

$$x' = \frac{\mathrm{x} - \bar{x}}{\sigma}$$

2.1.12 Long Short-Term Memory (LSTM)

LSTM ย่อจาก Long Short-Term Memory ถือเป็นประเภทหนึ่งของสถาปัตยกรรมแบบ Recurrent Neural Network (RNN) อยู่ในกลุ่มของ Deep Learning ถูกออกแบบให้จดจำ Patterns ในช่วงเวลานานๆ มีประสิทธิภาพสำหรับปัญหาการทำนายที่เป็น Sequential เนื่องจากสามารถเก็บข้อมูลก่อนหน้าและนำมาร่วมใช้ในการประมวลผลได้ สามารถแก้ปัญหา Long-term Dependency ได้ โดย RNN แบบดั้งเดิมจะเผชิญกับ ความท้าทายในเรื่อง Long-range Dependency และมีปัญหา Vanishing Gradient LSTM ถูกออกแบบมาให้จดจำ Long-term Information โดยใช้ Gating Mechanisms ที่ออกแบบมาเฉพาะ มีความยืดหยุ่นที่ดี LSTM สามารใช้ Model ได้ทั้ง Long-term และ Short-term Temporal Sequences [1]

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.2 แสดงโครงสร้างของ LSTM

การทำงานของ LSTM

ฟังก์ชัน Forgetting gate เป็นส่วนในการตัดสินใจจะลบข้อมูลไหนออกจาก memory นี้บ้าง

ฟังก์ชัน Input gate เป็นส่วนในการตัดสินใจว่าจะเอาข้อมูลใหม่ข้อมูลไหนเข้า memory

ฟังก์ชัน Output gate เป็นส่วนที่ตัดสินใจว่าจะเอาข้อมูลไหนมาส่งผลกระทบกับข้อมูลที่เราจะทำนาย

ข้อดี/ข้อเสีย

ข้อดี

1) ความสามารถในการจดจำ การจดจำ Patterns จาก Long Sequences ทำให้เหมาะกับ Applications ส่วนใหญ่ในทางปฏิบัติ

2) มีการนำไปใช้งานอย่างแพร่หลาย

3) มี Libraries ต่างๆรองรับ เช่น TensorFlow และ Keras ทำให้ง่ายต่อการนำไปใช้งาน

ข้อเสีย

1) ความซับซ้อนในการคำนวณ การ Train LSTMs ใช้เวลาและทรัพยากรในการคำนวณสูง

2) ต้องใช้ Dataset จำนวนมาก สำหรับการ Train Model ที่มีประสิทธิภาพ และหลีกเลี่ยง Overfitting จำเป็นต้องใช้ Data จำนวนมาก

3) Hyperparameter Tuning เหมือนกับ Neural Network อื่นๆ หากต้องการได้ประสิทธิภาพที่ดี จำเป็นต้องใช้ความพยายามกับการทำ Hyperparameter Tuning

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.13 วิธีวัดผลความแม่นยำโมเดล

1) ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error : MSE) เป็นการวัดค่าความคลาดเคลื่อน โดยการนำค่าความคลาดเคลื่อนมายกกำลัง แล้วนำไปหาค่าเฉลี่ย ในการวัดค่าความแม่นยำจากวิธีการนี้ยิ่งค่าที่ได้มีค่าน้อยแสดงว่าโมเดลที่ได้จะมีความแม่นยำมากโดยมีสมการดังนี้

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

โดยที่ MSE คือค่าความคาดเคลื่อนกำลังสองเฉลี่ย

n = คือจำนวนข้อมูลที่ใช้

Y_t = คือค่าจริงที่เวลา t ใด ๆ

$\widehat{Y}_t$ = คือค่าที่ได้จากการพยากรณ์ที่เวลา t ใด ๆ

2) ค่าสัมบูรณ์ของเปอร์เซ็นต์ของความคลาดเคลื่อน

ค่าสัมบูรณ์ของเปอร์เซ็นต์ของความคลาดเคลื่อน (Mean Absolute Percentage Error : MAPE) เป็นค่าความคลาดเคลื่อนที่แปลงให้อยู่ในรูปแบบของเปอร์เซ็นต์ในการวัดค่า ความแม่นยำจากวิธีการนี้ยิ่งค่าที่ได้มีค่าน้อยแสดงว่าโมเดลที่ได้จะมีความแม่นยำมาก ซึ่งยิ่งโดยมี สมการดังสมการ

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{Y_t - \widehat{Y}_t}{Y_t}\right| \times 100}{n_2}$$

โดยที่ MAPE คือค่าสัมบูรณ์ของเปอร์เซ็นต์ของความคลาดเคลื่อน

n = คือจำนวนข้อมูลที่ใช้

Y_t = คือค่าจริงที่เวลา t ใด ๆ

$\widehat{Y}_t$ = คือค่าที่ได้จากการพยากรณ์ที่เวลา t ใด ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย

ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation : MAD) หรือ (Mean Absolute Error : MAE) เป็นวิธีที่การวัดค่าความคลาดเคลื่อนที่นิยมอีกวิธีหนึ่ง ซึ่งวิธีนี้จะช่วยบอกถึงขนาดของความคลาดเคลื่อนรวมได้ โดยมีสมการในการวัดค่าความ แม่นยำจากวิธีการนี้ซึ่งค่าที่ได้มีค่าน้อยแสดงว่าโมเดลที่ได้จะมีความแม่นยำมาก

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_t - \widehat{Y_t}\right|$$

โดยที่ MAE คือค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย

n = คือจำนวนข้อมูลที่ใช้

Y_t = คือค่าจริงที่เวลา t ใด ๆ

$\widehat{Y_l}$ = คือค่าที่ได้จากการพยากรณ์ที่เวลา t ใด ๆ

4) ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error : RMSE) เป็นวิธีที่การวัดค่าความคลาดเคลื่อนแบบมาตรฐาน ซึ่งนิยมใช้กันอย่างแพร่หลาย โดยมีสมการดังสมการในการวัดค่าความแม่นยำจากวิธีการนี้ยิ่งค่า RMSE ที่ได้มีค่าน้อย แสดงว่าโมเดลที่ได้จะมีความแม่นยำมาก [7]

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{\left(\widehat{Y_i} - Y_i\right)^2}{n}}$$

โดยที่ RMSE คือค่ารากที่สองของค่าความคาดเคลื่อนกำลังสองเฉลี่ย

n = คือจำนวนข้อมูลที่ใช้

Y_t = คือค่าจริงที่เวลา t ใด ๆ

$\widehat{Y_l}$ = คือค่าที่ได้จากการพยากรณ์ที่เวลา t ใด ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2 งานวิจัยที่เกี่ยวข้อง

1) Forecasting of Vegetable Prices using STL-LSTM Method งานวิจัยนี้นำเสนอโดย Dong Jin, Helin Yin, Yeonghyeon Gu และ Seong Joan Yoo (2562) มีวัตถุประสงค์เพื่อสร้างตัวแบบทำนายราคากะหล่ำปลีและหัวไชเท้าในซีไฮนีสรายเดือน โดยข้อมูลที่ใช้ คือ ข้อมูลราคารายเดือน และข้อมูลอากาศ การพยากรณ์ที่ใช้ คือ seasonal-trend-loess (STL) และ long short-term memory (LSTM) ผลลัพธ์ที่ได้ คือ ตัวแบบการพยากรณ์กะหล่ำปลีและหัวไชเท้ามีความแม่นยำ 92.06% และ 88.74% ตามลำดับ

2) Machine Learning based Models for Fresh Produce Yield and Price Forecasting for Strawberry Fruit งานวิจัยนี้นำเสนอโดย Ifeanyi Okwuchi (2563) มีวัตถุประสงค์เพื่อสร้างตัวแบบการคาดการณ์ราคาสตรอเบอร์รี่ในแคลิฟอร์เนีย โดยตัวแบบที่ใช้ คือ การเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ข้อมูลที่ใช้ในการสร้างแบบจำลองการทำนาย ประกอบด้วย ข้อมูลสภาพอากาศ ผลผลิตสตรอเบอร์รี่ ราคาฟาร์มสตรอเบอร์รี่ และการซื้อของผู้ค้าปลีก การวัดประสิทธิภาพของตัวแบบใช้ AGM, MAE, MSE และ R^2 ผลการทดลองพบว่าตัวแบบที่ดีที่สุด คือ Attention CNN-LSTM (AC-LSTM) และ Attention ConvLSTM (ACV-LSTM)

3) Vegetable Price Forecasting Using STL and Attention Mechanism-Based LSTM งานวิจัยนี้นำเสนอโดย Helin Yin, Dong Jin, Yeong Hyeon Gu, Chang Jin Park, Sang Keun Han และ Seong Joon Yoo (2563) มีวัตถุประสงค์เพื่อทำนายราคาผัก 5 ชนิด คือ กะหล่ำปลี หัวไชเท้า หัวหอม พริกไทยร้อน และกระเทียม โดยข้อมูลที่ใช้ คือ สภาพอากาศของพื้นที่การผลิตหลัก, ปริมาณการซื้อขายในตลาด ผลลัพธ์ที่ได้ คือ ตัวแบบ STL-ATTLSTM ประสิทธิภาพดีที่สุด โดยมีค่า RMSE เท่ากับ 380 และค่า MAPE เท่ากับ 7%

จากงานวิจัยข้างต้นสามารถสรุปเป็นตาราง ดังตารางที่ 2.1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

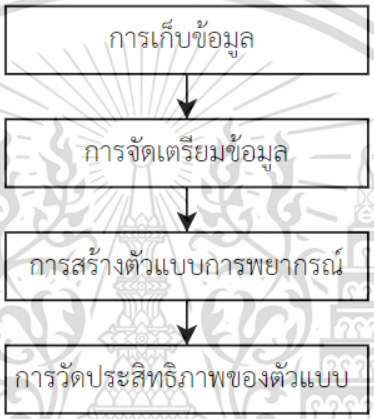
ตารางที่ 2.3 เปรียบเทียบงานวิจัยที่ได้ศึกษามา

ลำดับ	ชื่องานวิจัย	ผู้เขียน	ปี	วัตถุประสงค์	เทคนิคหรือหลักการที่ใช้	ความเหมือน/ความต่างกับงานวิจัย
1	Forecasting of Vegetable Prices using STL-LSTM Method	Dong Jin, Helin Yin, Yeonghyeon Gu และ Seong Joan Yoo	2562	เพื่อสร้างตัวแบบทำนายราคากะหล่ำปลีและหัวไชเท้าในซีไฮนีสตลาดสินค้าเกษตรเกาหลี	1. long short-term memory (LSTM) 2. seasonal-trend-loess (STL)	ความเหมือน : ตัวแบบในการพยากรณ์ และการพยากรณ์ราคาสินค้าเกษตรกรรม ความแตกต่าง : งานวิจัยที่อ้างอิงนำราคาสินค้ามาวิเคราะห์เพียงอย่างเดียว แต่ในงานวิจัยนี้มีการนำตัวแปรอื่นๆมาร่วมวิเคราะห์ด้วย
2	Machine Learning based Models for Fresh Produce Yield and Price Forecasting for Strawberry Fruit	Ifeanyi Okwuchi	2563	เพื่อสร้างตัวแบบการคาดการณ์ราคาสตรอเบอร์รี่ในแคลิฟอร์เนีย	1. Xgboost 2. LSTM 3. CNN 4. CNN-LSTM 5. CLG 6. ConvLSTM 7. ACLG 8. ACV-LSTM 9. AC-LSTM 10. Stacking 11. Ensemble	ความเหมือน : ตัวแบบในการพยากรณ์ ความแตกต่าง : งานวิจัยที่อ้างอิงมีการเปรียบเทียบตัวแบบที่เป็นวิธีสถิติ แต่ในงานวิจัยนี้มีแต่ตัวแบบการเรียนรู้เชิงลึก
3	STL-ATTLSTM: Vegetable Price Forecasting Using STL and Attention Mechanism-Based LSTM	Helin Yin, Dong Jin, Yeong Hyeon Gu, Chang Jin Park, Sang Keun Han และ Seong Joon Yoo	2563	การพยากรณ์ราคาผักรายเดือนโดยใช้ข้อมูลประเภทต่างๆ เช่น ราคาผัก ข้อมูลสภาพอากาศของพื้นที่การผลิตหลัก และปริมาณการซื้อขายในตลาด	1. LSTM 2. Attention LSTM 3. STL-LSTM 4. STL-ATTLSTM	ความเหมือน : ตัวแบบในการพยากรณ์ และมีตัวแปรอื่นๆร่วมในการวิเคราะห์ ความแตกต่าง : ในงานวิจัยที่อ้างอิงมีตัวแบบที่หลากหลายกว่า ในงานวิจัยนี้จะเน้นไปที่ตัวแบบ LSTM แล้วมีการปรับไฮเปอร์พารามา Hyperparameters

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
วิธีดำเนินงานวิจัย

ในบทนี้จะนำเสนอวิธีดำเนินงานวิจัยของการพยากรณ์ราคาผักบุ้ง ด้วยวิธีการเรียนรู้เชิงลึก เพื่อหาตัวแบบที่เหมาะสมในการพยากรณ์ราคาผักบุ้ง โดยกระบวนการวิจัยแบ่งออกเป็น 4 ขั้นตอนหลัก ได้แก่ การเก็บข้อมูล การจัดเตรียมข้อมูล การสร้างตัวแบบการพยากรณ์ และการวัดประสิทธิภาพของตัวแบบ ดังรูปที่ 3.1



รูปที่ 3.1 แสดงขั้นตอนหลักในการดำเนินงานวิจัย

3.1 ชุดข้อมูล (Dataset)

ข้อมูลที่ใช้ในการศึกษางานวิจัยแบ่งออกเป็น 4 กลุ่ม จากแหล่งข้อมูลที่แตกต่างกัน ดังนี้

1) ข้อมูลราคาผักบุ้งไทย จากกรมการค้าภายใน ประกอบไปด้วย 4 คุณลักษณะ คือ วันที่, ราคา
ต่ำสุด, ราคาสูงสุด และราคาเฉลี่ย

2) ข้อมูลคุณภาพอากาศ จากเว็ปไซต์ www.wunderground.com ควบคุมโดย IBM ประกอบไปด้วย 16 คุณลักษณะ คือ วันที่, ค่าอุณหภูมิสูงสุด, ค่าอุณหภูมิเฉลี่ย, ค่าอุณหภูมิต่ำสุด, ค่าจุดน้ำค้างสูงสุด, ค่าจุดน้ำค้างเฉลี่ย, ค่าจุดน้ำค้างต่ำสุด, ค่าความชื้นในอากาศสูงสุด, ค่าความชื้นในอากาศเฉลี่ย, ค่าความชื้นในอากาศต่ำสุด, ค่าความเร็วลมสูงสุด, ค่าความเร็วลมเฉลี่ย, ค่าความเร็วลมต่ำสุด, ค่าความกดอากาศสูงสุด, ค่าความกดอากาศเฉลี่ย และค่าความกดอากาศต่ำสุด

3) ข้อมูลปริมาณน้ำฝน จากหน่วยงานสถาบันสารสนเทศทรัพยากรน้ำ ประกอบไปด้วย 2 คุณลักษณะ คือ วันที่ และปริมาณน้ำฝน

4) ข้อมูลปฏิทินวันหยุด จากเว็ปไซต์ www.myhora.com ประกอบไปด้วย 3 คุณลักษณะ คือ วันที่, ชื่อวันหยุดนักขัตฤกษ์ และวันเสาร์-อาทิตย์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นนำข้อมูลแต่ละกลุ่มมาประกอบกันด้วยคุณลักษณะวันที่ จะมีคุณลักษณะ(คอลัมภ์) ทั้งหมด 21 ตัว ดังแสดงในตารางที่ 3.1 และจำนวน 2,800 เรคอร์ด เป็นข้อมูลระหว่างวันที่ 1 มกราคม 2559 ถึง 31 สิงหาคม 2566

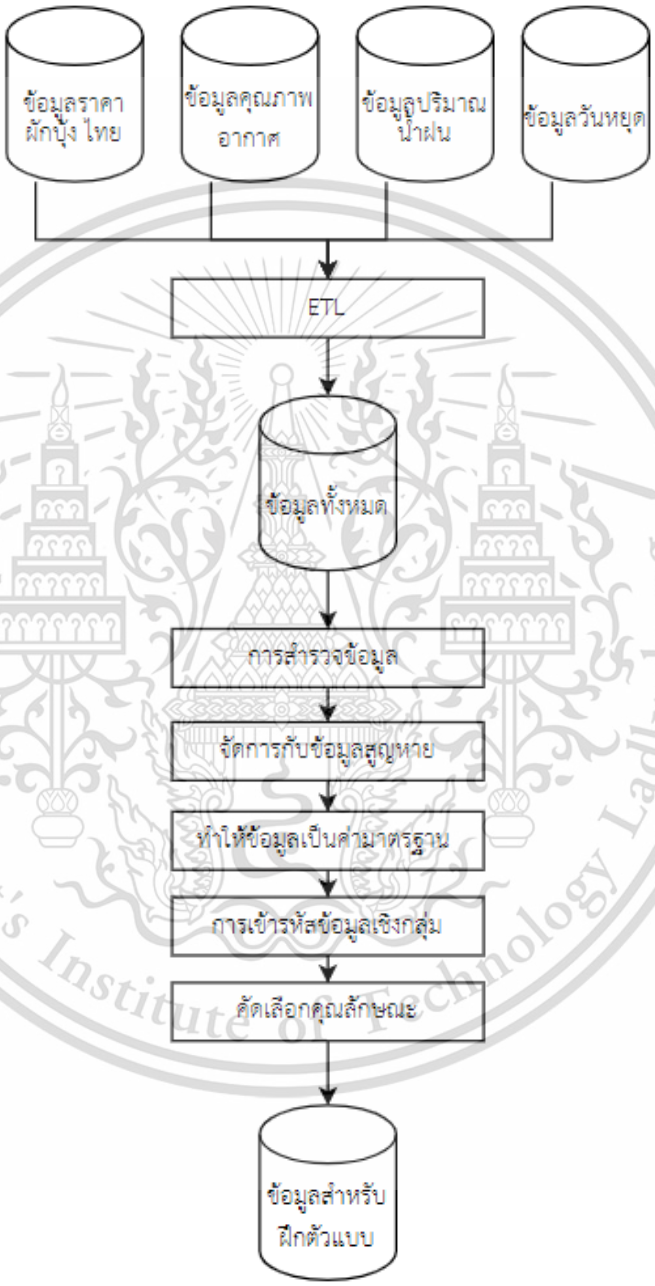
ตารางที่ 3.1 แสดงคุณลักษณะของชุดข้อมูลทั้งหมด

ลำดับ	ชื่อคุณลักษณะ	ประเภทข้อมูล	คำอธิบาย
1	date	datetime	วันที่
2	min_price	float	ราคาต่ำสุด
3	max_price	float	ราคาสูงสุด
4	price_avg	float	ราคาเฉลี่ย
5	rain	float	ปริมาณน้ำฝน
6	weekend	string	วันเสาร์-อาทิตย์
7	dayoff	string	วันหยุดนักขัตฤกษ์
8	temp _max	float	ค่าอุณหภูมิสูงสุด
9	temp _avg	float	ค่าอุณหภูมิเฉลี่ย
10	temp_min	float	ค่าอุณหภูมิต่ำสุด
11	dewpoint _max	float	ค่าจุดน้ำค้างสูงสุด
12	dewpoint_avg	float	ค่าจุดน้ำค้างเฉลี่ย
13	dewpoint _min	float	ค่าจุดน้ำค้างต่ำสุด
14	humidity_max	float	ค่าความชื้นในอากาศสูงสุด
15	humidity_avg	float	ค่าความชื้นในอากาศเฉลี่ย
16	humidity_min	float	ค่าความชื้นในอากาศต่ำสุด
17	windspeed_max	float	ค่าความเร็วลมสูงสุด
18	windspeed_avg	float	ค่าความเร็วลมเฉลี่ย
19	windspeed_min	float	ค่าความเร็วลมต่ำสุด
20	pressure_max	float	ค่าความกดอากาศสูงสุด
21	pressure_avg	float	ค่าความกดอากาศเฉลี่ย
22	Pressure_min	float	ค่าความกดอากาศต่ำสุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2 การจัดเตรียมข้อมูลเพื่อการประมวลผล

ในงานวิจัยนี้มีการจัดเตรียมข้อมูลเพื่อการประมวลผล โดยในขั้นแรกจะสกัดข้อมูลจาก 4 แหล่งข้อมูล และนำข้อมูลเหล่านี้มาแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์ จากนั้นทำการสำรวจข้อมูล จัดการกับข้อมูลสูญหาย ทำให้ข้อมูลเป็นค่ามาตรฐาน เข้ารหัสข้อมูลเชิงกลุ่ม และคัดเลือกคุณลักษณะ จากนั้นนำข้อมูลที่ได้ไปฝึกตัวแบบพยากรณ์ แสดงดังรูปที่ 3.2



รูปที่ 3.2 แสดงขั้นตอนการจัดเตรียมข้อมูลเพื่อการประมวลผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2.1 ETL

ในงานวิจัยนี้นำข้อมูลจากหลายแหล่งข้อมูลมาวิเคราะห์ร่วมกัน จึงต้องมีการแปลงรูปแบบให้เหมาะสม (Transform) แล้วเก็บข้อมูลไว้ที่เดียวกันเพื่อเตรียมสำหรับฝึกตัวแบบ ซึ่งการทำ ETL ประกอบด้วย 3 กลไก ดังนี้

1) การสกัดข้อมูลจากแหล่งข้อมูล (Extract) ในงานวิจัยนี้มีทั้งหมด 4 แหล่งข้อมูล

1.1) หน่วยงานกรมการค้า

1.2) สถาบันสารสนเทศทรัพยากรน้ำ

1.3) เว็ปไซต์ www.wunderground.com

1.4) เว็ปไซต์ www.myhora.com

2) การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม (Transform) ในแต่ละแหล่งข้อมูลจะมีการเก็บรูปแบบวันที่ที่แตกต่างกัน ดังนั้นจึงต้องมีการแปลงรูปแบบวันที่ให้เหมือนกัน เพื่อให้สามารถวิเคราะห์ร่วมกันได้

3) การโหลดข้อมูลไปยังปลายทาง (Load) เมื่อทำการแปลงรูปแบบข้อมูลให้เหมาะสมแล้วก็นำข้อมูลเหล่านี้ไปเก็บไว้ยังสถานที่ที่พร้อมต่อการนำไปฝึกสอนตัวแบบ

3.2.2 การสำรวจข้อมูล (data exploration)

เป็นการทำความเข้าใจเพื่อพิจารณาภาพรวมของข้อมูล โดยอาจใช้แผนภาพ หรือกราฟของข้อมูลในรูปแบบต่าง ๆ ระหว่างการสำรวจ อาจจะพบข้อผิดพลาดหรือปัญหาอื่น ๆ ในงานวิจัยนี้จะสำรวจข้อมูลดังนี้

1) ข้อมูลสถิติพื้นฐาน แสดงจำนวนแถว, ค่าเฉลี่ย, ค่าสูงสุด, ค่าต่ำสุด, เปอร์เซนไทล์, ค่าเบี่ยงเบนมาตรฐาน

2) กราฟเส้นแสดงจุดของข้อมูลในเวลานั้นๆ

3) ค่าสถิติสหสัมพันธ์ เป็นค่าที่บ่งชี้ถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว

3.2.3 การจัดการกับข้อมูลสูญหาย

เมื่อนำข้อมูลจากหลายแหล่งมารวมกัน จะพบว่ามีคุณลักษณะบางคุณลักษณะที่มีข้อมูลสูญหาย จึงต้องจัดการกับข้อมูลเหล่านี้ตามลักษณะของข้อมูล ดังตารางที่ 3.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.2 แสดงวิธีการจัดการกับข้อมูลสูญหาย

ชื่อคุณลักษณะ	วิธีการ
min_price	แทนค่าที่ว่างด้วยวิธี Interpolate
max_price	
price_avg	
rain	แทนค่าที่ว่างด้วยค่าเฉลี่ยของวันที่และเดือนเดียวกันในแต่ละปี

3.2.3 การทำให้ข้อมูลเป็นค่ามาตรฐาน (Data Normalization)

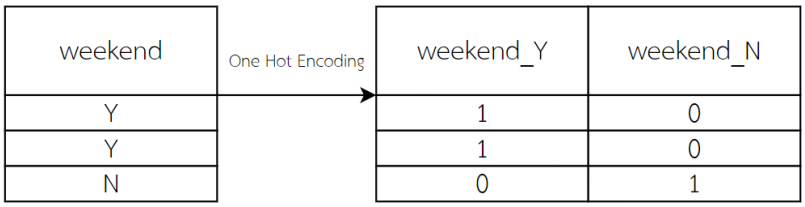
เป็นเทคนิคส่วนหนึ่งในการจัดเตรียมข้อมูลก่อนการสร้างแบบจำลอง ซึ่งเป้าหมายของการทำ Data Normalization เป็นการเปลี่ยนข้อมูลตัวแปรที่เป็นตัวเลขให้อยู่ในช่วงความถี่ที่เป็นมาตรฐานเดียวกันทั้งหมดเพื่อเพิ่มความแม่นยำในการพยากรณ์ ในงานวิจัยนี้ใช้วิธี MinMaxScaler ในการให้ข้อมูลเป็นค่ามาตรฐานแสดงตัวอย่างตามตารางที่ 3.3

ตารางที่ 3.3 แสดงตัวอย่างการทำข้อมูลให้เป็นมาตรฐาน

ชื่อคุณลักษณะ	ค่าต้นฉบับ	ค่าที่ได้จากการทำให้เป็นค่ามาตรฐาน
temp_avg	83.5	0.687919
dewpoint_avg	64.8	0.457421

3.2.4 การเข้ารหัสข้อมูล (One-Hot encoding)

ข้อมูลที่ใช้ในงานวิจัยมีข้อมูลที่เป็นข้อมูลเชิงคุณภาพอยู่ 2 คุณลักษณะ คือ วันเสาร์-อาทิตย์ (weekend) และ วันหยุดนักขัตฤกษ์ (dayoff) ซึ่งเป็นข้อมูลที่ถูกเก็บในลักษณะ Categorical ที่ไม่มีลำดับ (Nominal number) จึงต้องทำการเปลี่ยนให้อยู่ในรูปแบบของ Binary values ที่มีค่า 0 หรือ 1 เท่านั้น ซึ่งในงานวิจัยนี้ใช้วิธี One-Hot Encoding เนื่องจากเป็นรูปแบบที่เข้าใจง่าย เหมาะสำหรับข้อมูลที่เป็นการจัดหมวดหมู่ และไม่ต้องมีการปรับเปลี่ยนโครงสร้างโมเดลมาก ตัวอย่างการเข้ารหัสข้อมูลแสดงดังรูปที่ 3.4



รูปที่ 3.3 แสดงตัวอย่างการเข้ารหัสข้อมูลด้วยวิธี One-Hot Encoding ของคุณลักษณะ weekend คุณลักษณะเดิม (ซ้าย) คุณลักษณะที่สร้างขึ้นใหม่ (ขวา)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2.5 การคัดเลือกคุณลักษณะ (Feature selection)

ในงานวิจัยมีคุณลักษณะถึง 21 ตัว ซึ่งในบางครั้งการที่ใส่คุณลักษณะทั้งหมดไม่ได้ทำให้ตัวแบบมีประสิทธิภาพดีที่สุด ดังนั้นจึงจำเป็นต้องทำการคัดเลือกคุณลักษณะที่สำคัญมาใช้ในการฝึกตัวแบบ โดยในงานวิจัยนี้จะใช้วิธีการคัดเลือกแบบลำดับขั้น (Stepwise Selection)

3.3 การสร้างตัวแบบการพยากรณ์ (Modelling)

3.3.1 อัลกอริทึมที่ใช้

ในงานวิจัยนี้ต้องการสร้างตัวแบบพยากรณ์ราคาเฉลี่ยของผักบุ้งไทย ใช้อัลกอริทึม Long Short-Term Memory (LSTM) โดยแบ่งตัวแบบออกเป็น 2 กรณี คือ

1) ตัวแบบการเรียนรู้เชิงลึก (LSTM)

2) ตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะแบบลำดับขั้น (Stepwise Selection)

3.3.2 การแบ่งข้อมูลสำหรับการฝึกและทดสอบ

แบ่งข้อมูลออกเป็น 2 ส่วน คือ 80:20 โดยข้อมูลส่วนที่หนึ่ง 80% ใช้ในการสร้างตัวแบบ และข้อมูลส่วนที่สอง 20% ใช้ในการทดสอบประสิทธิภาพของตัวแบบ

3.3.3 การสร้างตัวแบบการเรียนรู้เชิงลึก (Deep Learning Model)

การสร้างตัวแบบการเรียนรู้เชิงลึกในงานวิจัยนี้ ประกอบไปด้วย 3 ชั้น คือ ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออก ฟังก์ชันการตุ้นที่ใช้คือ linear ฟังก์ชันการสูญเสีย (loss function) คือ MSE, MAE, RMSE และ MAPE โดยในชั้นข้อมูลเข้ามีการเพิ่ม Dropout ที่มีอัตรา dropout เท่ากับ 0.2(20%) เพื่อลดการเรียนรู้ที่เกิดขึ้นในการฝึกโมเดลและลดความเสี่ยงของ overfitting

ตารางที่ 3.4 แสดงโครงสร้างตัวแบบของการเรียนรู้เชิงลึก

ลำดับที่	จำนวนนิวรอน	ฟังก์ชั่นกระตุ้น	ประเภทชั้น
1	100	linear	ชั้นข้อมูลเข้า
2	16	linear	ชั้นซ่อน
3	1	linear	ชั้นข้อมูลออก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3.4 การหาค่า Hyperparameter ที่เหมาะสม

ในการสร้างตัวแบบการเรียนรู้เชิงลึกจะต้องมีการกำหนด Hyperparameters ที่เกี่ยวข้องกับตัวแบบให้มีความเหมาะสม ในงานวิจัยนี้จะกำหนด Hyperparameters ดังนี้

ตารางที่ 3.5 แสดงการกำหนดค่า Hyperparameters ที่ใช้ในตัวแบบ

Hyperparameters	ค่าที่กำหนด
epochs	100
batch_size	32
optimizer	adam
dropout	0.2
loss	MSE, MAE,MAPE, RMSE
window_len	5
neurons	100

3.4 การวัดประสิทธิภาพของตัวแบบ (Model Evaluation)

การวัดประสิทธิภาพของตัวแบบ แบ่งออกเป็น 4 วิธี คือ

1) ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error : MSE)

2) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error : MAE)

3 ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error : RMSE)

4) ค่าสัมบูรณ์ของเปอร์เซ็นต์ของความคลาดเคลื่อน (Mean Absolute Percentage Error : MAPE)

โดยในงานวิจัยนี้ต้องการให้ค่าความเคลื่อน MSE, MAE, RMSE และ MAPE มีค่าน้อยที่สุด เพื่อให้ได้ตัวแบบที่ดีที่สุดที่เหมาะสมในการพยากรณ์ข้อมูลราคาเฉลี่ยผักบุ้ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการวิจัยการอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเพื่อหาตัวแบบการพยากรณ์ที่เหมาะสมของราคาผักบุ้งไทย โดยทางผู้วิจัยได้นำเทคนิคการเรียนรู้เชิงลึกมาใช้ในการวิจัย ซึ่งในบทนี้จะกล่าวถึงโครงสร้างของตัวแบบการเรียนรู้เชิงลึกและผลของการทดสอบประสิทธิภาพของตัวแบบ ในงานวิจัยนี้จะแบ่งการฝึกตัวแบบออกเป็น 2 แบบ คือ ตัวแบบการเรียนรู้เชิงลึก (LSTM) เพียงอย่างเดียว และ ตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก

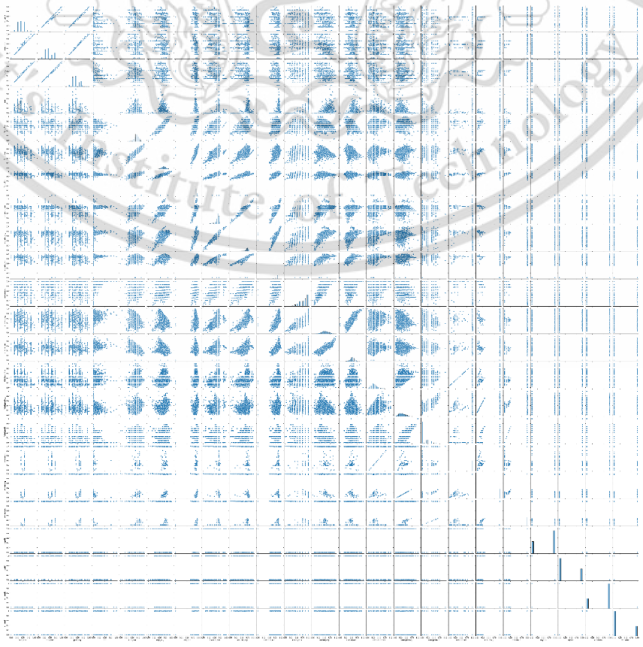
4.1 การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะ คือ การพิจารณาตัวแปรอิสระหรือปัจจัยที่มีผลต่อตัวแปรตาม แล้วทำการคัดเลือกตัวแปรที่มีความสำคัญเข้าไปอยู่ในตัวแบบ ซึ่งก่อนจะทำการคัดเลือกคุณลักษณะนั้นจะทำการตรวจสอบปัญหา Multicollinearity ที่เกิดจากการที่ตัวแปรอิสระมีความสัมพันธ์กันเอง ซึ่งอาจส่งผลให้ตัวแบบที่ได้มีประสิทธิภาพไม่ดี ผู้วิจัยจึงทำการทดสอบความสัมพันธ์ของตัวแปรอิสระโดยใช้ Scatter matrix, Correlation matrix และ ค่า Variance Inflation Factor (VIF)

4.1.1 การทดสอบ Multicollinearity

1) Scatter matrix

เป็นการ Plot scatter ของตัวแปรอิสระทุกๆตัว ในงานวิจัยนี้มีตัวแปรอิสระทั้งหมด 22 ตัว ดังนั้น ผลลัพธ์จึงออกมาเป็น Matrix ขนาด 22x22 แสดงดังรูปที่ 4.1



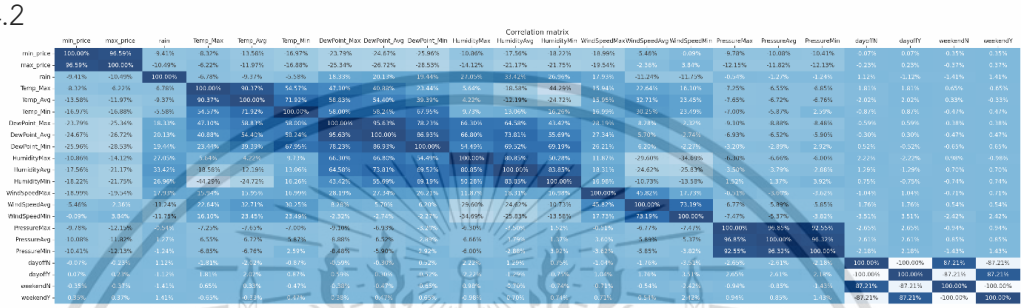
รูปที่ 4.1 ผลลัพธ์จากการทำ Scatter matrix

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ผลลัพธ์ที่ได้จากการหาความสัมพันธ์ของตัวแปรอิสระทุกตัวจากการทำ Scatter matrix ที่สนใจคือคู่ตัวแปรที่ให้ผลไปในทาง “เส้นตรง” แต่ผลลัพธ์ที่ได้จากการทำ Scatter matrix นั้นสรุปผลยากจากการมองด้วยตา

2) Correlation matrix

การหาความสัมพันธ์ของตัวแปรอิสระทุกตัวแปร ผลลัพธ์ที่ออกมาอยู่ในช่วง -1 ถึง 1 ซึ่งเครื่องหมายคือทิศทางของความสัมพันธ์ว่าเป็น Negative หรือ Positive relationship โดยในงานวิจัยนี้มีตัวแปรอิสระทั้งหมด 22 ตัว จะได้ผลลัพธ์ Correlation matrix ขนาด 22x22 แสดงดังรูปที่ 4.2

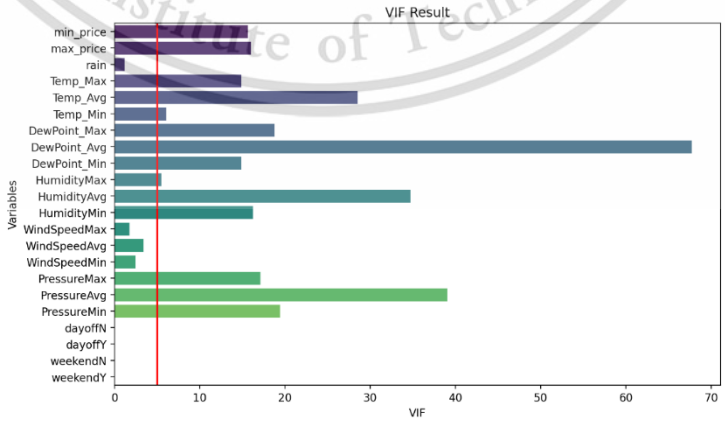


รูปที่ 4.2 ผลลัพธ์จากการทำ Correlation matrix

ผลลัพธ์ที่ได้จากการทำ Correlation matrix เมื่อนำไปร่วมวิเคราะห์กับผลลัพธ์จากการทำ Scatter matrix ถ้าคู่ตัวแปรใดมี Scatter matrix ที่ให้ผลไปในทางเส้นตรง และ Correlation matrix มีค่าสูงร่วมด้วย อาจสรุปได้แล้วว่าคู่ตัวแปรนี้เกิด Multicollinearity ขึ้น แต่ตัวแปรอื่น ๆ ที่ให้ค่าสูงเช่นกันแต่ Scatter plot ไม่ชัดเจน จะสรุปผลค่อนข้างยาก

3) Variance Inflation Factor (VIF)

เพื่อให้เกิดการสรุปผลที่แน่นอนมากกว่า Correlation matrix จึงควรมี Statistic 1 ค่าเพื่อเทียบกับ Threshold ว่าเกิด Multicollinearity ขึ้นหรือไม่ วิธีการที่นิยมใช้เป็นสากลคือ Variance Inflation Factor (VIF) โดยในงานวิจัยนี้กำหนดค่า Threshold มีค่าเท่ากับ 5 ได้ผลลัพธ์แสดงดังรูปที่ 4.3



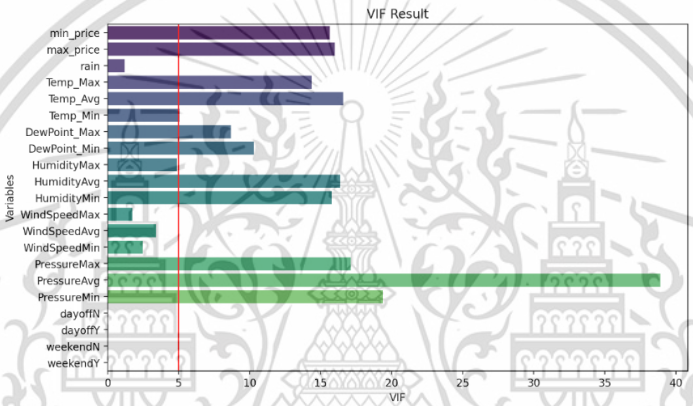
รูปที่ 4.3 ผลลัพธ์จากการทำ Variance Inflation Factor (VIF)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากกราฟเห็นได้ว่าตัวแปร DewPoint_Avg เกิด Multicollinearity ด้วยค่า VIF ที่เยอะที่สุด แต่ตัวแปรmin_price, max_price, Temp_Max, Temp_Avg, Temp_Min, DewPoint_Max, DewPoint_Avg, DewPoint_Min, HumidityMax, HumidityAvg, HumidityMin, PressureMax, PressureAvg, PressureMin, dayoffN, dayoffY, weekendN, weekendY ก็มี Multicollinearity ด้วยเช่นกัน ในกรณีที่เกิดเหตุการณ์เช่นนี้ ต้องมีลำดับขั้นตอนในการจัดการ

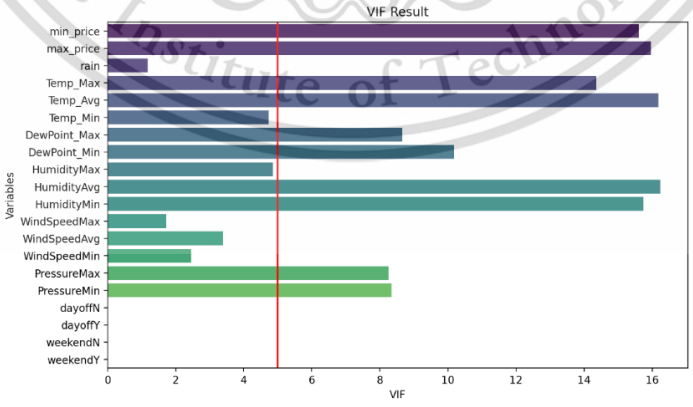
ขั้นตอนการจัดการกับปัญหา Multicollinearity คือ Drop ตัวแปรที่ VIF สูงที่สุดแล้ว รัน VIF เพื่อทดสอบ Multicollinearity อีกรอบ ถ้ายังมีตัวแปรที่ VIF สูงกว่า Threshold ก็ให้ Repeat ซ้ำไปเรื่อย ๆ จนกว่าทุกตัวแปรจะมีค่า VIF อยู่ภายใต้ Threshold

3.1) จากการผลลัพธ์ในรูปที่ 4.3 ตัวแปร DewPoint_Avg มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร DewPoint_Avg แสดงดังรูปที่ 4.4



รูปที่ 4.4 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 1

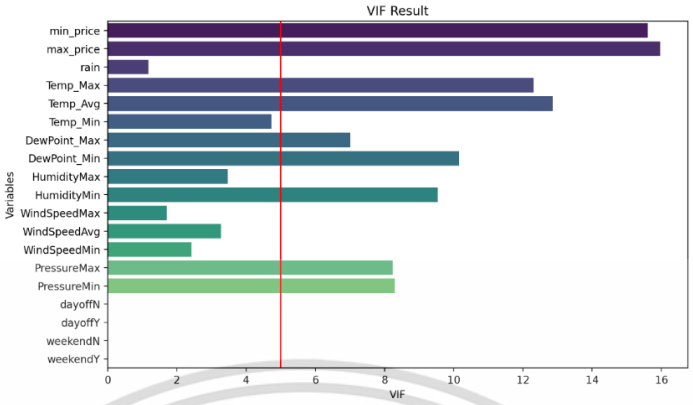
3.2) จากการผลลัพธ์ในรูปที่ 4.4 ตัวแปร PressureAvg มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร PressureAvg แสดงดังรูปที่ 4.5



รูปที่ 4.5 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 2

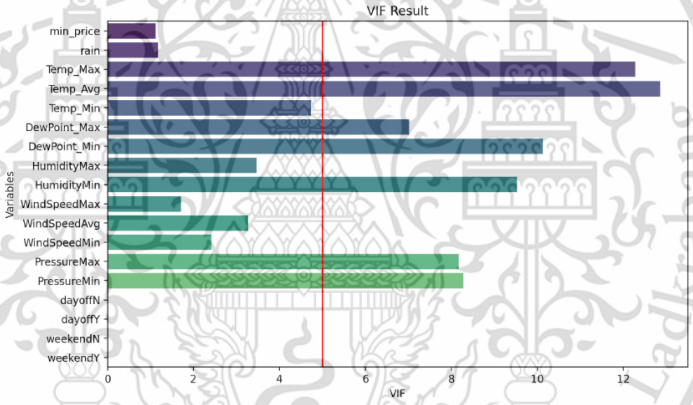
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3) จากการผลลัพธ์ในรูปที่ 4.5 ตัวแปร HumidityAvg มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร HumidityAvg แสดงดังรูปที่ 4.6



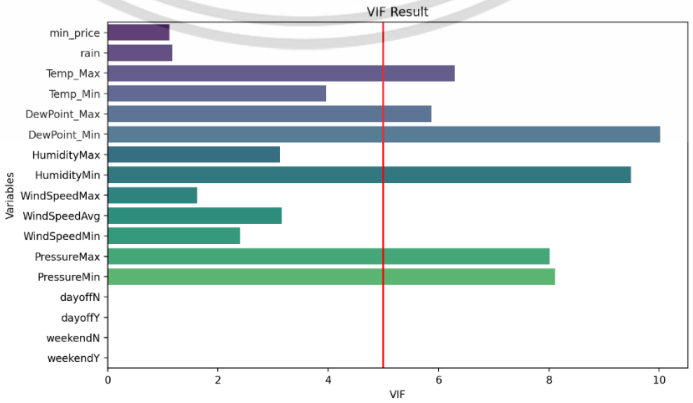
รูปที่ 4.6 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 3

3.4) จากการผลลัพธ์ในรูปที่ 4.6 ตัวแปร max_price มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร max_price แสดงดังรูปที่ 4.7



รูปที่ 4.7 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 4

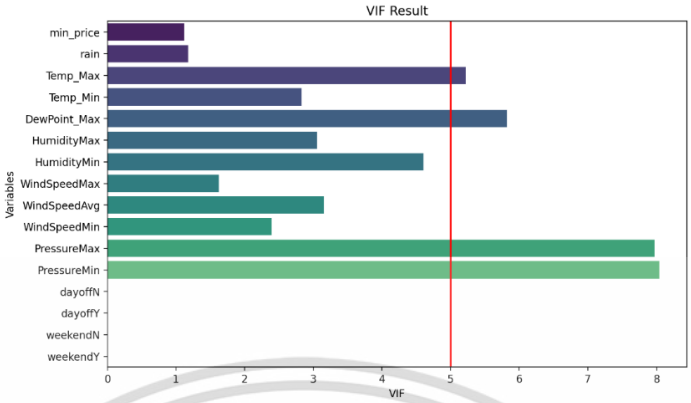
3.5) จากการผลลัพธ์ในรูปที่ 4.7 ตัวแปร Temp_Avg มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร Temp_Avg แสดงดังรูปที่ 4.8



รูปที่ 4.8 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 5

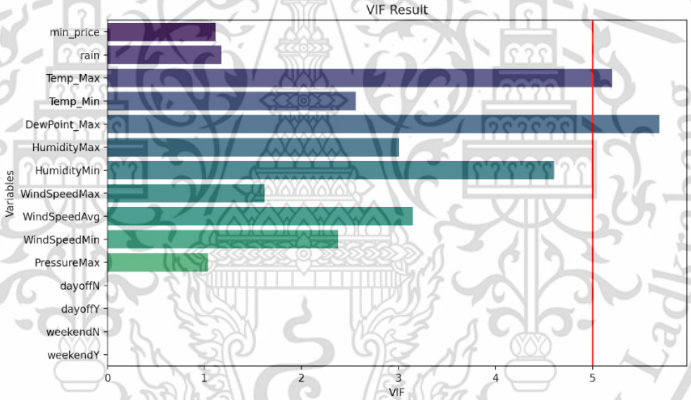
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6) จากการผลลัพธ์ในรูปที่ 4.8 ตัวแปร DewPoint_Min มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร DewPoint_Min แสดงดังรูปที่ 4.9



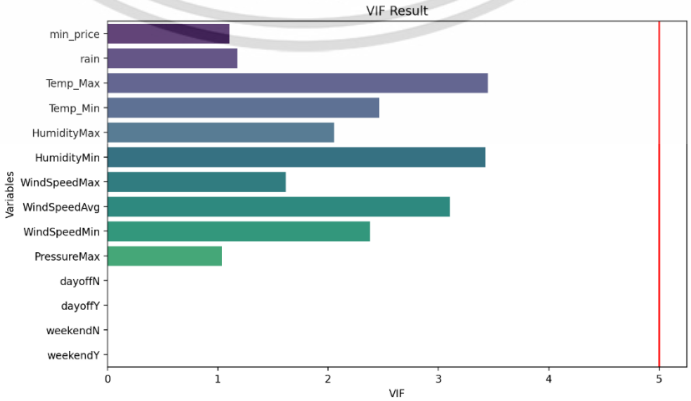
รูปที่ 4.9 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 6

3.7) จากการผลลัพธ์ในรูปที่ 4.9 ตัวแปร PressureMin มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร PressureMin แสดงดังรูปที่ 4.10



รูปที่ 4.10 ผลลัพธทดสอบ Multicollinearity ดวย Variance Inflation Factor (VIF) ครั้งที่ 7

3.8) จากการผลลัพธ์ในรูปที่ 4.10 ตัวแปร DewPoint_Max มีค่า VIF สูงที่สุดจึงทำการตัดตัวแปรนี้ทิ้งไป ผลลัพธ์จากการตัดทิ้งตัวแปร DewPoint_Max แสดงดังรูปที่ 4.11



รูปที่ 4.11 ผลลัพธ์ทดสอบ Multicollinearity ด้วย Variance Inflation Factor (VIF) ครั้งที่ 8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากผลลัพธ์ที่ได้ในรูปที่ 4.11 ตัวแปรทุกตัวอยู่ภายใต้ Threshold จึงทำการหยุดรัน พบว่าเหลือตัวแปรทั้งหมด 14 ตัว จาก 22 ตัวในตอนแรก คือ min_price, rain, Temp_Max, Temp_Min, HumidityMax, HumidityMin, WindSpeedMax, WindSpeedAvg, WindSpeedMin, PressureMax, dayoffN, dayoffY, weekend และ weekend

4.1.2 การเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน(Stepwise Selection)

เป็นวิธีการคัดเลือกผสมผสานระหว่างวิธีการคัดเลือกตัวแปรอิสระทั้งแบบการเพิ่มตัวแปรและการลดตัวแปรเข้าด้วยกัน โดยดูที่ค่า p-value ของตัวแปรทุกตัวต้องผ่าน Significant threshold ที่ 0.05 ผลลัพธ์จากการเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน แสดงดังรูปที่ 4.12

OLS Regression Results

Dep. Variable:	price_avg	R-squared:	0.103
Model:	OLS	Adj. R-squared:	0.101
Method:	Least Squares	F-statistic:	80.01
Date:	Sun, 12 Nov 2023	Prob (F-statistic):	2.31e-64
Time:	07:48:47	Log-Likelihood:	1194.2
No. Observations:	2800	AIC:	-2378.
Df Residuals:	2795	BIC:	-2349.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0589	0.043	24.691	0.000	0.975	1.143
HumidityMin	-0.4178	0.034	-12.440	0.000	-0.484	-0.352
Temp_Max	-0.2548	0.030	-8.462	0.000	-0.314	-0.196
PressureMax	-0.1936	0.029	-6.743	0.000	-0.250	-0.137
WindSpeedMax	-0.1348	0.021	-6.535	0.000	-0.175	-0.094

Omnibus:	28.519	Durbin-Watson:	0.111
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28.223
Skew:	0.226	Prob(JB):	7.44e-07
Kurtosis:	2.805	Cond. No.	31.2

รูปที่ 4.12 ผลลัพธ์จากการคัดเลือกตัวแปรด้วยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน

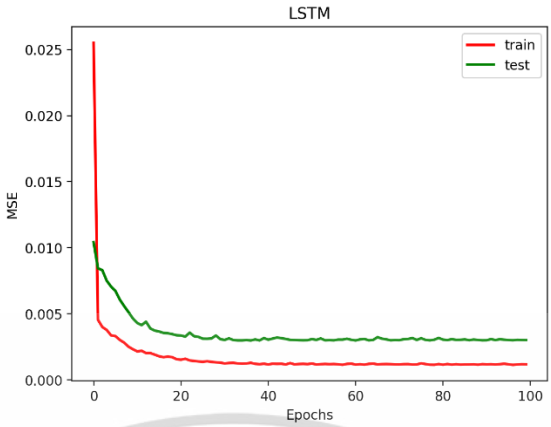
จากรูปที่ 4.12 ตัวแปรที่ผ่านการคัดเลือกคุณลักษณะโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอนมีทั้งหมด 4 ตัวแปร คือ HumidityMin, Temp_Max, PressureMax และ WindSpeedMax ที่ค่า p-value เท่ากับ 0.000

4.2 ตัวแบบการพยากรณ์

ในงานวิจัยนี้จะแบ่งการฝึกตัวแบบออกเป็น 2 แบบ คือ ตัวแบบการเรียนรู้เชิงลึก (LSTM) เพียงอย่างเดียว และตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก โดยกำหนดค่า Hyperparameters ที่ epochs เท่ากับ 100, window_len เท่ากับ 5, test_size เท่ากับ 0.2, neurons เท่ากับ 100, batch_size เท่ากับ 32, dropout เท่ากับ 0.2 และ optimizer เท่ากับ adam

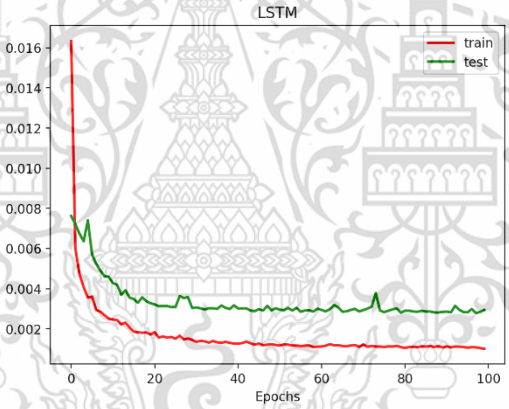
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1 ตัวแบบการเรียนรู้เชิงลึก (LSTM) เพียงอย่างเดียว พบว่าค่าสูญเสียลดลงตามรูปที่4.13



รูปที่ 4.13 แสดงค่าสูญเสียของการฝึก 0 - 100 รอบ (LSTM)

4.2.2 ตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก 4 ตัว คือ HumidityMin, Temp_Max, PressureMax และ WindSpeedMax พบว่าค่าสูญเสียลดลงตามรูปที่ 4.14



รูปที่ 4.14 แสดงค่าสูญเสียของการฝึก 0 - 100 รอบ (LSTM+ปัจจัยภายนอก)

ประสิทธิภาพของตัวแบบการเรียนรู้เชิงลึก (LSTM) เพียงอย่างเดียว และตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก พบว่าตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก ให้ค่าความคลาดเคลื่อน MSE, RMSE, MAPE และ MPE น้อยกว่าตัวแบบการเรียนรู้เชิงลึก (LSTM) เพียงอย่างเดียว แสดงดังตารางที่ 4.1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.1 แสดงประสิทธิภาพของตัวแบบการพยากรณ์

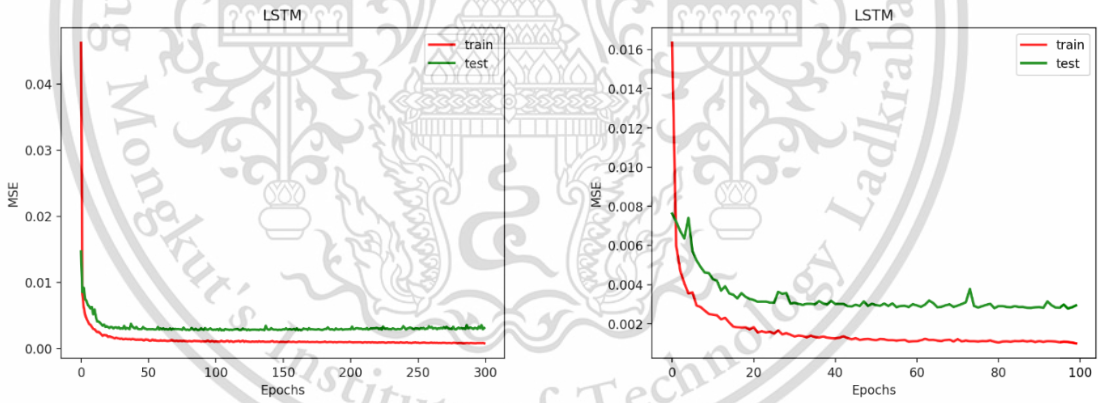
ตัวแบบการพยากรณ์	epochs	window	neurons	batch size	ค่าคลาดเคลื่อน			
					MSE	RMSE	MAPE	MAE
LSTM	100	5	100	32	0.0012	0.0299	3.8514	0.0162
LSTM + ปัจจัยภายนอก	100	5	100	32	0.0010	0.0274	3.7803	0.0158

4.3 การปรับค่า Hyperparameters

จากผลลัพธ์ในตารางที่ 4.1 ผู้วิจัยเลือกตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก มาทำการปรับค่า Hyperparameters ใหม่เพื่อให้ได้ตัวแบบที่มีประสิทธิภาพดีที่สุด

1. Epoch

จากการทดลองด้วยการฝึก 300 รอบ พบว่าค่าสูญเสียลดลงตามรูปที่ 4.15 เมื่อดูจากรูปพบว่า Epoch ที่ตั้งไว้ 50 รอบเมื่อเพิ่มจำนวน Epoch เป็น 300 รอบ ค่าสูญเสียปรับตัวลดลงอย่างมีนัยสำคัญ ผู้วิจัยจึงทำการเลือกค่าที่ Epoch เท่ากับ 300



รูปที่ 4.15 แสดงค่าสูญเสียของการฝึก 0 - 300 รอบ (ด้านซ้าย) เปรียบเทียบกับ 0 – 100 รอบ (ด้านขวา)

ประสิทธิภาพของตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอกที่ epoch เท่ากับ 100 เปรียบเทียบกับ epoch เท่ากับ 300 โดยวัดค่าความคลาดเคลื่อน MSE, RMSE ผลลัพธ์ที่ได้ตัวแบบที่มีการเพิ่มค่า epoch เท่ากับ 300 มีค่าความคลาดเคลื่อนน้อยกว่าตัวแบบที่มีค่า epoch เท่ากับ 100 แสดงดังตารางที่ 4.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.2 แสดงประสิทธิภาพของตัวแบบจากการทดลองกำหนดค่า epoch

ตัวแบบการพยากรณ์	epochs	window	neurons	batch size	ค่าคลาดเคลื่อน			
					MSE	RMSE	MAPE	MAE
LSTM + ปัจจัยภายนอก	100	5	100	32	0.0010	0.0274	3.7803	0.0158
LSTM + ปัจจัยภายนอก	300	5	100	32	0.0008	0.0258	3.9509	0.0162

2. Batch size

จากการทดลองค่า Batch size ที่ค่า 32 และ 64 กับ epoch เท่ากับ 300 ประสิทธิภาพของตัวแบบที่ Batch size เท่ากับ 32 ให้ค่า MSE และ RMSE น้อยกว่าตัวแบบที่ Batch size เท่ากับ 64 แสดงดังตารางที่ 4.3

ตารางที่ 4.3 แสดงประสิทธิภาพของตัวแบบจากการทดลองกำหนดค่า Batch size

ตัวแบบการพยากรณ์	epochs	window	neurons	batch size	ค่าคลาดเคลื่อน			
					MSE	RMSE	MAPE	MAE
LSTM + ปัจจัยภายนอก	300	5	100	32	0.0008	0.0258	3.9509	0.0162
LSTM + ปัจจัยภายนอก	300	5	100	64	0.0008	0.0267	4.0250	0.0164

3. window_len

จากการทดลองกำหนดค่า window_len เท่ากับ 5, 7, 14, 21 และ 30 ประสิทธิภาพของตัวแบบที่ window เท่ากับ 5 ให้ค่า MSE และ RMSE ต่ำที่สุด แสดงดังตารางที่ 4.4

ตารางที่ 4.4 แสดงประสิทธิภาพของตัวแบบจากการทดลองกำหนดค่า window

ตัวแบบการพยากรณ์	epochs	window	neurons	batch size	ค่าคลาดเคลื่อน			
					MSE	RMSE	MAPE	MAE
LSTM + ปัจจัยภายนอก	300	5	100	32	0.0008	0.0258	3.9509	0.0162
LSTM + ปัจจัยภายนอก	300	7	100	32	0.0010	0.0292	4.0572	0.0168
LSTM + ปัจจัยภายนอก	300	14	100	32	0.0009	0.0276	3.8916	0.0162
LSTM + ปัจจัยภายนอก	300	21	100	32	0.0008	0.0269	3.8377	0.0157
LSTM + ปัจจัยภายนอก	300	30	100	32	0.0008	0.0267	4.0250	0.0164

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4 ผลการศึกษาและการวัดประสิทธิภาพ

จากการทดลองนำตัวแบบการเรียนรู้เชิงลึก (LSTM) ร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก มาทำการปรับค่า Hyperparameters ของตัวแบบ ตัวแบบที่ให้ค่าความคลาดเคลื่อน MSE ต่ำที่สุด คือตัวแบบที่ 2, 6 ตัวแบบที่ให้ค่าความคลาดเคลื่อน RMSE ต่ำที่สุด คือตัวแบบที่ 2 ตัวแบบที่ให้ค่าความคลาดเคลื่อน MAPE ต่ำที่สุด คือตัวแบบที่ 1 และตัวแบบที่ให้ค่าความคลาดเคลื่อน MAE ต่ำที่สุด คือตัวแบบที่ 6 แสดงดังตารางที่ 4.5

ตารางที่ 4.5 แสดงการเปรียบเทียบประสิทธิภาพของตัวแบบ

ตัวแบบที่	ตัวแบบการพยากรณ์	epochs	window	neurons	batch size	ค่าคลาดเคลื่อน			
						MSE	RMSE	MAPE	MAE
1	LSTM + ปัจจัยภายนอก	100	5	100	32	0.0010	0.0274	3.7803*	0.0158
2	LSTM + ปัจจัยภายนอก	300	5	100	32	0.0008*	0.0258*	3.9509	0.0162
3	LSTM + ปัจจัยภายนอก	300	5	100	64	0.0009	0.0286	3.8942	0.0158
4	LSTM + ปัจจัยภายนอก	300	7	100	32	0.0010	0.0292	4.0572	0.0168
5	LSTM + ปัจจัยภายนอก	300	14	100	32	0.0009	0.0276	3.8916	0.0162
6	LSTM + ปัจจัยภายนอก	300	21	100	32	0.0008*	0.0269	3.8377	0.0157*

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยภายนอกที่มีผลต่อราคาผักบุ้งไทย และเพื่อหาตัวแบบการพยากรณ์ที่เหมาะสม โดยใช้วิธีการเรียนรู้เชิงลึกร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก สามารถสรุปผลการวิจัยและข้อเสนอแนะได้ดังนี้

5.1.1 จากตัวแปรปัจจัยภายนอกทั้งหมด 22 ตัว ตัวแปรที่ส่งผลกระทบต่อราคาผักบุ้งมี 4 ตัว ได้แก่ ที่ค่าความชื้นในอากาศสูงสุด ค่าอุณหภูมิสูงสุด ค่าความกดอากาศสูงสุด และค่าความเร็วลมต่ำสุด

5.1.2 ตัวแบบการพยากรณ์ที่ใช้การเรียนรู้เชิงลึกเพียงอย่างเดียว พบว่าประสิทธิภาพของตัวแบบดีที่สุด โดยมีค่าความคลาดเคลื่อน MSE เท่ากับ 0.0012 RMSE เท่ากับ 0.0299 MAPE เท่ากับ 3.8514 และ MAE เท่ากับ 0.0162

5.1.3 ตัวแบบการพยากรณ์ที่ใช้การเรียนรู้เชิงลึกร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก ได้ผลลัพธ์ปัจจัยภายนอก 4 ตัว คือ ค่าความชื้นในอากาศสูงสุด ค่าอุณหภูมิสูงสุด ค่าความกดอากาศสูงสุด และค่าความเร็วลมต่ำสุด พบว่าประสิทธิภาพของตัวแบบเพิ่มขึ้นจากการพยากรณ์ที่ใช้การเรียนรู้เชิงลึกเพียงอย่างเดียว โดยมีค่าความคลาดเคลื่อนน้อยลง MSE เท่ากับ 0.0010 RMSE เท่ากับ 0.0274 MAPE เท่ากับ 3.7803 และ MAE เท่ากับ 0.0158

5.1.4 เมื่อนำตัวแบบการพยากรณ์ที่ใช้การเรียนรู้เชิงลึกร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก ได้ผลลัพธ์ปัจจัยภายนอก 4 ตัว คือ ค่าความชื้นในอากาศสูงสุด ค่าอุณหภูมิสูงสุด ค่าความกดอากาศสูงสุด และค่าความเร็วลมต่ำสุด มาปรับค่า Hyperparameters คือ epoch, batch size, window_len ผลลัพธ์ที่ได้คือตัวแบบที่ให้ค่าความเคลื่อน MSE ต่ำที่สุด โดยมีค่าเท่ากับ 0.0008 คือ ตัวแบบที่ 2 ที่ Epoch เท่ากับ 300, Batch size เท่ากับ 32 และ window_len เท่ากับ 5 และตัวแบบที่ 6 ที่ Epoch เท่ากับ 300, Batch size เท่ากับ 32 และ window_len เท่ากับ 21 ตัวแบบที่ให้ค่า RMSE ต่ำที่สุด โดยมีค่าเท่ากับ 0.0258 คือ ตัวแบบที่ 2 ที่ Epoch เท่ากับ 300, Batch size เท่ากับ 32 และ window_len เท่ากับ 5 ตัวแบบที่ให้ค่าความเคลื่อน MAPE ต่ำที่สุด โดยมีค่าเท่ากับ 3.7803 คือ ตัวแบบที่ 1 ที่ Epoch เท่ากับ 300, Batch size เท่ากับ 32 และ window_len เท่ากับ 5 และตัวแบบที่ให้ค่าความเคลื่อน MAE ต่ำที่สุด โดยมีค่าเท่ากับ 0.0157 คือ ตัวแบบที่ 6 ที่ Epoch เท่ากับ 300, Batch size เท่ากับ 32 และ window_len เท่ากับ 21

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากการศึกษาครั้งนี้พบว่าประสิทธิภาพของตัวแบบการพยากรณ์ที่ใช้การเรียนรู้เชิงลึกร่วมกับการคัดเลือกคุณลักษณะปัจจัยภายนอก ได้ผลลัพธ์ปัจจัยภายนอก 4 ตัว คือ ค่าความชื้นในอากาศสูงสุด ค่าอุณหภูมิสูงสุด ค่าความกดอากาศสูงสุด และค่าความเร็วลมต่ำสุด มีประสิทธิภาพดีกว่าตัวแบบการพยากรณ์ที่ใช้การเรียนรู้เชิงลึกเพียงอย่างเดียว และเมื่อมีการปรับค่า Hyperparameter ตัวแบบที่ได้ก็จะมีประสิทธิภาพดีขึ้น

5.2 ข้อเสนอแนะ

5.2.1 การปรับค่า Hyperparameter หากต้องการตัวแบบที่มีประสิทธิภาพที่ดี จำเป็นต้องใช้เวลาในการปรับค่า Hyperparameter

5.2.2 จำนวนข้อมูลในการฝึกสอนตัวแบบมีผลต่อประสิทธิภาพของตัวแบบการพยากรณ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

[1] Narut Soontranon. (2566). LSTM. สืบค้นเมื่อ 7 ตุลาคม 2566, จาก https://www.nerd-data.com/deep_learning_lstm

[2] Puechkaset. (2559). ผักบุ้งไทย สรรพคุณ และการปลูกผักบุ้งไทย. สืบค้นเมื่อ 7 ตุลาคม 2566, จาก https://puechkaset.com

[3] Surapong Kanoktipsatharporn. (2562). Normalization คืออะไร ปรับช่วงข้อมูล Feature Scaling ด้วยวิธี Normalization, Standardization ก่อนเทรน Machine Learning Preprocessing ep.2. สืบค้นเมื่อ 7 ตุลาคม 2566, จาก https://www.bualabs.com/archives/2100/what-is-normalization-feature-scaling-rescaling-normalization-standardization-feedforward-train-machine-learning-preprocessing-ep-2/

[4] Surapong Kanoktipsatharporn. (2563). One Hot Encoding คืออะไร ประโยชน์ ข้อดี ข้อเสีย ของ One Hot Encoding ทำไมต้องใช้ One Hot Encoding ใน Machine Learning. สืบค้นเมื่อ 12 ตุลาคม 2566, จาก https://www.bualabs.com/archives/1902/what-is-one-hot-encoding-benefit-one-hot-encoding-why-one-hot-encoding-in-machine-learning/

[5] กระทรวงพาณิชย์. (2563). เส้นทางราคาผักและผลไม้ที่มีผลต่อการใช้จ่ายของผู้บริโภค. สืบค้นเมื่อ 12 ตุลาคม 2566, จาก www.tpso.moc.go.th

[6] ปภังกร อิ่นแก้ว. (2566). วิทยาการข้อมูลเบื้องต้น. สืบค้นเมื่อ 20 พฤศจิกายน 2566, จาก https://www2.cs.science.cmu.ac.th/courses/204123/lib/exe/fetch.php?media=book_ch2.pdf

[7] ธีร์ธวัช แก้ววิจิตร. (2559). เกณฑ์ที่ใช้ในการวัดความแม่นยำของโมเดล. สืบค้นเมื่อ 12 ตุลาคม 2566, จาก http://sutir.sut.ac.th:8080/jspui/bitstream/ 123456789/6835/2/ Fulltext.pdf

[8] รัฐพงศ์ ชัยเอิก และคณะ. (2560). วิธีการคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ. สืบค้นเมื่อ 15 ตุลาคม 2566, จาก http://sc2.kku.ac.th/stat/statweb/images /Eventpic/ 60/Seminar/01_11_.pdf

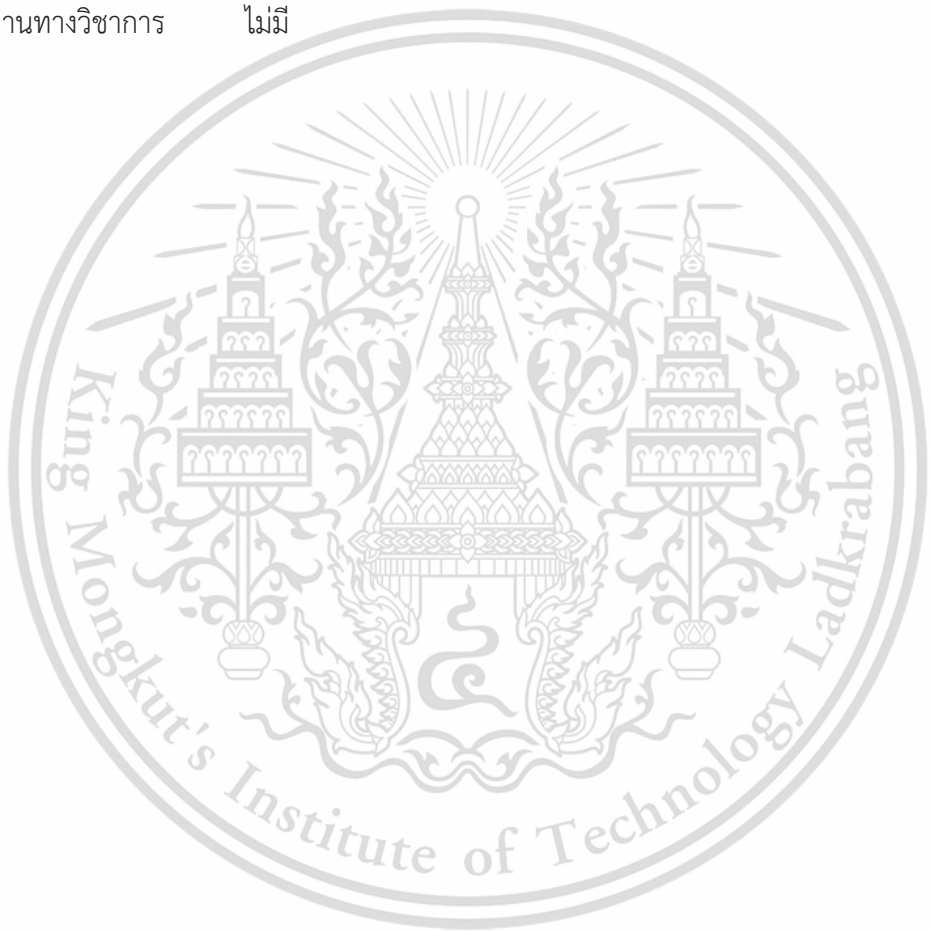
[9] สถาบันนวัตกรรมและธรรมาภิบาลข้อมูล. (2565). Correlation Coefficient. สืบค้นเมื่อ 15 ตุลาคม 2566, จาก https://digi.data.go.th/blog/what-is-correlation-coefficient

[10] สุวดี นำพาเจริญ. (2554). VIF คืออะไร. สืบค้นเมื่อ 15 ตุลาคม 2566, จาก https://www.solutioncenterminitab.com/blog/what-in-the-world-is-a-vif/

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ	นางสาวกนกวรรณ แววดี
วัน เดือน ปีเกิด	17 พฤศจิกายน 2540
ที่อยู่ปัจจุบัน	10/1 หมู่ 2 ต.หนองเพรางาย อ.ไทรน้อย จ.นนทบุรี 11150
ประวัติการศึกษา	(2559) วิทยาศาสตรบัณฑิต สาขาสถิติ เกรดเฉลี่ย 3.18
ทุนการศึกษาที่ได้รับ	ไม่มี
ผลงานทางวิชาการ	ไม่มี



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้