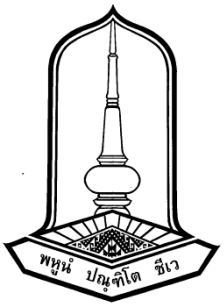


การพยากรณ์ค่าเฉลี่ยของอุณหภูมิจากการปล่อยคาร์บอน

ภูริภัทร สุ่นสุข
พชรพล แดงมณี

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาบัณฑิต สาขาวิชาวิทยาการข้อมูลประยุกต์
ตุลาคม 2567

การพยากรณ์ค่าเฉลี่ยของอุณหภูมิจากการปล่อยคาร์บอน

ภูริภัทร สุ่นสุข
พชรพล แดงมณี

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาบัณฑิต สาขาวิชาวิทยาการข้อมูลประยุกต์
ตุลาคม 2567

คณะกรรมการสอบปริญญานิพนธ์ ได้พิจารณาปริญญานิพนธ์ฉบับนี้ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาบัณฑิต สาขาวิชาวิทยาการข้อมูลประยุกต์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบปริญญานิพนธ์

..... ประธานกรรมการ
([ชื่อประธานกรรมการสอบ]) (ประธานกรรมการควบคุมปริญญานิพนธ์)

..... กรรมการ
([ชื่อกรรมการควบคุมฯ]) (ประธานกรรมการควบคุมปริญญานิพนธ์)

..... กรรมการ
([ชื่อกรรมการควบคุมปริญญานิพนธ์]) (ที่ปรึกษาผู้ควบคุมปริญญานิพนธ์)

มหาวิทยาลัยอนุมัติให้รับปริญญานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาบัณฑิต สาขาวิชาวิทยาการข้อมูลประยุกต์ ของมหาวิทยาลัยมหาสารคาม

....
(อาจารย์อุมาภรณ์ สายแสงจันทร์) (อาจารย์กวีพจน์ บรรลือวงศ์)
ผู้ดูแลโครงงาน ผู้ดูแลโครงงาน

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและการสนับสนุนจากบุคคลหลายท่านที่ข้าพเจ้าขอขอบคุณเป็นอย่างยิ่ง ขอขอบพระคุณอาจารย์กวีพจน์ บรรลือวงศ์ อาจารย์ที่ปรึกษา ซึ่งได้ให้คำแนะนำและข้อเสนอแนะที่เป็นประโยชน์อย่างยิ่งต่อการทำวิจัยในครั้งนี้ ขอขอบคุณคณะกรรมการสอบปริญญานิพนธ์ทุกท่านที่ได้ให้คำแนะนำและกำลังใจในการดำเนินการจนเสร็จสมบูรณ์ นอกจากนี้ ขอขอบคุณเพื่อนร่วมชั้นและครอบครัวที่ได้สนับสนุนและให้กำลังใจตลอดช่วงเวลาที่ผ่านมา ขอบคุณทุกท่านที่มีส่วนร่วมในการทำให้โครงงานนี้สำเร็จลุล่วงไปด้วยดี

ภูริภัทร สุ่นสุข
พชรพล แดงมณี

ชื่อเรื่อง การพยากรณ์ค่าเฉลี่ยของอุณหภูมิจากการปล่อยคาร์บอน
ผู้จัดทำ ภูริภัทร สุ่นสุข
พชรพล แดงมณี

ปริญญา วิทยาศาสตรบัณฑิต **สาขาวิชา** วิทยาการข้อมูลประยุกต์
อาจารย์ที่ปรึกษา อ. กวีพจน์ บรรลือวงศ์
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2567

บทคัดย่อ

โครงงานนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลการพยากรณ์ค่าเฉลี่ยของอุณหภูมิที่ได้รับผลกระทบจากการปล่อยก๊าซคาร์บอนไดออกไซด์ ด้วยการนำข้อมูลจากองค์การอาหารและการเกษตรแห่งสหประชาชาติ (FAO) และคณะกรรมการระหว่างรัฐบาลว่าด้วยการเปลี่ยนแปลงสภาพภูมิอากาศ (IPCC) มาสร้างชุดข้อมูลเพื่อใช้ทำนายผล อัลกอริทึมที่ใช้ในการเปรียบเทียบประสิทธิภาพการทำนายประกอบด้วย Linear Regression, Random Forest, และ Decision Tree ซึ่งได้แบ่งชุดข้อมูลออกเป็นชุดการฝึกสอนและชุดทดสอบ โดยใช้วิธีการ Cross-validation 10-Fold

ผลการทดลองแสดงให้เห็นว่า Random Forest มีประสิทธิภาพสูงสุดในการทำนายค่าเฉลี่ยของอุณหภูมิที่เปลี่ยนแปลงตามการปล่อยคาร์บอน ด้วยค่าคลาดเคลื่อนที่น้อยที่สุด การศึกษาแสดงให้เห็นถึงความสำคัญของการใช้โมเดลที่มีความซับซ้อนมากขึ้นในการทำนายผลกระทบต่อสิ่งแวดล้อมจากการปล่อยก๊าซเรือนกระจก ซึ่งสามารถนำไปประยุกต์ใช้ในการวางแผนนโยบายเพื่อควบคุมการปล่อยก๊าซและลดผลกระทบจากภาวะโลกร้อนได้อย่างมีประสิทธิภาพ

คำสำคัญ : การพยากรณ์อุณหภูมิ, การปล่อยก๊าซคาร์บอนไดออกไซด์, Decision Tree, Random Forest, Naïve Bayes

สารบัญ

หน้า

สารบัญภาพประกอบ

หน้า

สารบัญตาราง

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

(Cortez and Morais, 2007) สภาพอากาศในปัจจุบันมีการเปลี่ยนแปลงอย่างรวดเร็วและรุนแรงมากขึ้น อันเป็นผลมาจากการเปลี่ยนแปลงสภาพภูมิอากาศโลก ในช่วงทศวรรษที่ผ่านมา หลายประเทศทั่วโลกต้องเผชิญกับปรากฏการณ์สภาพอากาศสุดขั้ว เช่น คลื่นความร้อน พายุรุนแรง น้ำท่วมฉับพลัน และภัยแล้งที่ยาวนาน ส่งผลกระทบต่อระบบนิเวศ เศรษฐกิจ และความเป็นอยู่ของประชาชน การเปลี่ยนแปลงนี้เกิดจากการเพิ่มขึ้นของก๊าซเรือนกระจกในชั้นบรรยากาศ โดยเฉพาะจากการเผาไหม้เชื้อเพลิงฟอสซิล การตัดไม้ทำลายป่า และกิจกรรมอุตสาหกรรมต่างๆ

(Chalathip, 2567) งานวิจัยล่าสุดที่เพิ่งเผยแพร่ออกมาระบุว่าการปล่อยก๊าซคาร์บอนไดออกไซด์ที่ทำให้โลกร้อนส่วนใหญ่ นับตั้งแต่ปี 2559 จนถึงการบันทึกถึงปี 2565 มีสาเหตุมาจากกลุ่มผู้ผลิตเชื้อเพลิงฟอสซิลและซีเมนต์ 57 ราย ตามรายงานของ Carbon Majors โดย Influence Map ซึ่งเป็นองค์กรไม่แสวงผลกำไร ระบุว่าบริษัทเหล่านี้ทั้งของรัฐ และเอกชน เกี่ยวข้องกับการปล่อยก๊าซคาร์บอนฯในโลกถึง 80% ในรายงานได้เปิดเผยถึงบริษัทที่ปล่อยก๊าซคาร์บอนฯ 3 อันดับแรกของโลกในช่วงปีดังกล่าว ได้แก่

1. Saudi Aramco บริษัทน้ำมันของซาอุดิอาระเบียง
2. Gazprom บริษัทพลังงานยักษ์ใหญ่ของรัสเซีย
3. Coal India บริษัทผู้ผลิตถ่านหินอินเดียที่รัฐเป็นเจ้าของ

Daan Van Acker ผู้จัดการโครงการ Influence Map กล่าวว่า ข้อมูลจากรายงานสามารถใช้ได้ในหลายกรณี ตั้งแต่กระบวนการทางกฎหมายที่ต้องการควบคุมผู้ผลิตเหล่านี้ให้รับผิดชอบต่อความเสียหายต่อสภาพภูมิอากาศ หรือนักวิชาการสามารถนำมาใช้ในการวัดปริมาณการมีส่วนร่วมแก้ปัญหาสิ่งแวดล้อมของบริษัทเหล่านี้ รวมไปถึงนักลงทุนก็นำมาพิจารณาประกอบการลงทุนได้ ดังนั้น การนำเทคโนโลยีเข้ามาช่วยจำแนกและวิเคราะห์การปล่อยก๊าซคาร์บอนไดออกไซด์ (CO_2) จึงเป็นแนวทางหนึ่งที่จะช่วยให้การวิเคราะห์เกิดความแม่นยำมากขึ้น ด้วยความรวดเร็วในการประมวลผลจะทำให้ได้ข้อมูลที่ใกล้เคียงความเป็นจริงและน่าเชื่อถือมากขึ้น ดังนั้นจึงได้เกิดเทคโนโลยีในการวิเคราะห์ข้อมูลที่มีความสำคัญออกมาจากแหล่งเก็บข้อมูลขนาดใหญ่ เรียกเทคโนโลยีนี้ว่า การทำเหมืองข้อมูล (Data Mining)

โครงงานนี้ทำการศึกษาเกี่ยวกับการพยากรณ์การปล่อยก๊าซ CO_2 โดยศึกษาวิธีการจำแนกกลุ่มด้วยวิธีการต่างๆ เพื่อช่วยลดความเสี่ยงและป้องกันในรูปแบบที่มีความเหมาะสมและมีประสิทธิภาพให้ค่าผลลัพธ์ใกล้เคียงความเป็นจริงมากที่สุด เช่น วิธีการสุ่มป่าไม้ (Random Forest), วิธีการถดถอยเชิงเส้น (Linear Regressions),

วิธีต้นไม้ตัดสินใจ (Decision Tree) เพื่อให้ทราบถึงลักษณะของปัจจัยที่ส่งผลต่อการปล่อยก๊าซ CO_2 โดยเฉพาะในภาคส่วนต่างๆ เพื่อให้สามารถวางแผนและดำเนินการลดการปล่อยก๊าซเรือนกระจกได้อย่างถูกต้องและมีประสิทธิภาพ

1.2 วัตถุประสงค์ของโครงงาน

1. เพื่อสร้างโมเดลการทำนายค่าเฉลี่ยอุณหภูมิจากโมเดล Linear Regressions, Random Forest, Decision Tree

2. เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบ ระหว่าง Linear Regressions, Random Forest, Decision Tree

1.3 ขอบเขตของโครงงาน

1. ข้อมูลที่ใช้ในโครงงานเป็นข้อมูลกลุ่มตัวอย่างจำนวน 6965 ระเบียน ประกอบไปด้วย ตัวแปรทั้งหมด 31 แอตทริบิวต์ แบ่งเป็นตัวแปรต้น 30 แอตทริบิวต์ และตัวแปรตาม 1 แอตทริบิวต์

- Savanna fires: การปล่อยก๊าซจากไฟในระบบนิเวศทุ่งหญ้าสะวันนา
- Forest fires: การปล่อยก๊าซจากไฟในพื้นที่ป่า
- Crop Residues: การปล่อยก๊าซจากการเผาหรือการย่อยสลายเศษพืชหลังการเก็บเกี่ยว
- Rice Cultivation: การปล่อยก๊าซมีเทนจากการปลูกข้าว
- Drained organic soils (CO_2) การปล่อยก๊าซคาร์บอนไดออกไซด์เมื่อระบายน้ำในดินอินทรีย์
- Pesticides Manufacturing: การปล่อยก๊าซจากการผลิตสารกำจัดศัตรูพืช
- Food Transport: การปล่อยก๊าซจากการขนส่งสินค้าอาหาร
- Forestland: ที่ดินที่ปกคลุมด้วยป่าไม้
- Net Forest conversion: การเปลี่ยนแปลงพื้นที่ป่าเนื่องจากการตัดไม้ทำลายป่าและการปลูกป่า
- Food Household Consumption: การปล่อยก๊าซจากการบริโภคอาหารที่ระดับครัวเรือน
- Food Retail: การปล่อยก๊าซจากการดำเนินงานของร้านค้าปลีกที่ขายอาหาร
- On-farm Electricity Use: การใช้ไฟฟ้าในฟาร์ม
- Food Packaging: การปล่อยก๊าซจากการผลิตและการกำจัดบรรจุภัณฑ์อาหาร
- Agrifood Systems Waste Disposal: การปล่อยก๊าซจากการกำจัดขยะในระบบอาหารและเกษตร
- Food Processing: การปล่อยก๊าซจากการแปรรูปผลิตภัณฑ์อาหาร
- Fertilizers Manufacturing: การปล่อยก๊าซจากการผลิตปุ๋ย
- IPPU: การปล่อยก๊าซจากกระบวนการอุตสาหกรรมและการใช้ผลิตภัณฑ์
- Manure applied to Soils: การปล่อยก๊าซจากการใช้ปุ๋ยคอกในดินเกษตร

- Manure left on Pasture: การปล่อยก๊าซจากปุ๋ยคอกในทุ่งหญ้าหรือพื้นที่เลี้ยงสัตว์
- Manure Management: การปล่อยก๊าซจากการจัดการและการบำบัดปุ๋ยคอก
- Fires in organic soils: การปล่อยก๊าซจากไฟในดินอินทรีย์
- Fires in humid tropical forests: การปล่อยก๊าซจากไฟในป่าฝนเขตร้อนชื้น
- On-farm energy use: การใช้พลังงานในฟาร์ม
- Rural population: จำนวนคนที่อาศัยอยู่ในพื้นที่ชนบท
- Urban population: จำนวนคนที่อาศัยอยู่ในพื้นที่เมือง
- Total Population - Male: จำนวนประชากรชายทั้งหมด
- Total Population - Female: จำนวนประชากรหญิงทั้งหมด
- total_emission: การปล่อยก๊าซเรือนกระจกทั้งหมดจากแหล่งต่าง ๆ
- Average Temperature °C: อุณหภูมิเฉลี่ยที่เพิ่มขึ้น (ตามปี) ในองศาเซลเซียส

2. ข้อมูลถูกเก็บรวบรวมในรูปแบบไฟล์ CSV และสามารถดาวน์โหลดได้ที่

Agri-food CO2 emission dataset - Forecasting ML | Kaggle

3. แหล่งที่มาของข้อมูลได้รับข้อมูลมาจากองค์การอาหารและการเกษตร (FAO) และข้อมูลจาก IPCC

4. ทำการคัดเลือกคุณลักษณะและการลด Feature และการลด Dimension ของ Dataset ที่มีขนาดใหญ่ ด้วยการแปลง Variables ที่มีจำนวนมาก ให้มีจำนวนน้อยลงแต่ยัง Contains ข้อมูลส่วนใหญ่ของชุดข้อมูลไว้ได้

5. ทำการทำนายค่าเฉลี่ยอุณหภูมิเพื่อหาแบบจำลองที่ดีที่สุดโดยใช้อัลกอริทึม ซึ่งประกอบไป ด้วย Decision Tree, Random Forest, Linear Regressions

6. ทำการวัดประสิทธิภาพค่าเฉลี่ยอุณหภูมิโดยใช้วิธี Mean Absolute Error (ค่าคลาดเคลื่อนกำลังสองเฉลี่ย), Root Mean Squared Error (ค่าคลาดเคลื่อนกำลังสองเฉลี่ย) Mean Squared Error (ค่าคลาดเคลื่อนกำลังสองเฉลี่ย), R-square (ค่าสัมประสิทธิ์สหสัมพันธ์)

8. ทำการแบ่งข้อมูลแบบ 10-Fold Cross validation

9. เครื่องมือที่ใช้ในการทำโครงงาน Visual Studio Code และสร้างกราฟด้วย Power Bi

1.4 ขั้นตอนการดำเนินงาน

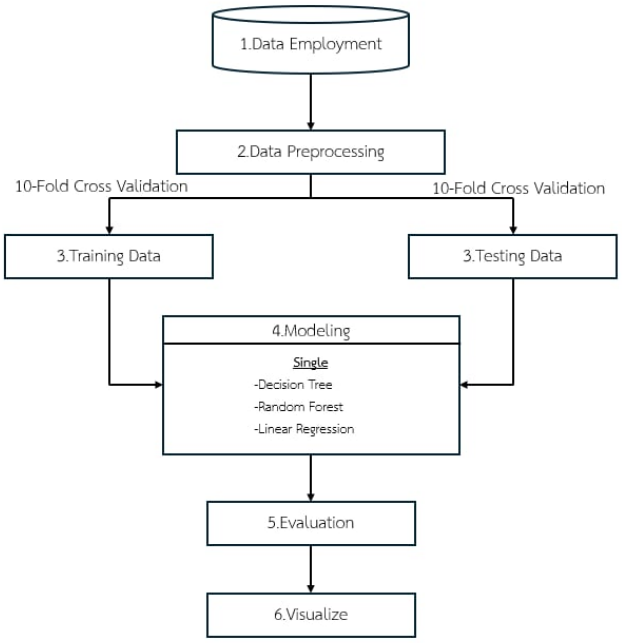
ขั้นตอนการดำเนินงานมีทั้งหมด 6 ขั้นตอน

1. เริ่มต้นจากการรวบรวมข้อมูลจาก Kaggle
2. เตรียมข้อมูลโดยการทำความสะอาดข้อมูล
3. แบ่งข้อมูล แบบ Training data และ Testing data ด้วย 10-Fold Cross validation

4. สร้างแบบจำลองโดยใช้อัลกอริทึม Decision Tree, Random Forest และ Linear Regressions โดยใช้ IDE ทำการวัดประสิทธิภาพ

5. วิเคราะห์ผลการวัดประสิทธิภาพของการจำแนก

6. จากนั้นจะได้แบบจำลองที่ดีที่สุดแล้วนำมาแสดงผลการทดลองด้วย Visualization ดังภาพประกอบที่ 1



ภาพประกอบที่ 1 แผนผังขั้นตอนการดำเนินงาน

1.5 ประโยชน์ที่คาดว่าจะได้รับ

สามารถศึกษาการเปรียบเทียบการการเปลี่ยนแปลงของอุณหภูมิเพื่อช่วยเฝ้าระวังและเตรียมความพร้อมกันการรับมือกับสภาพอากาศที่ร้อนจัดหรือมลผิดทางอากาศ เช่น การเดินทางหรือทำกิจกรรมกลางแจ้ง

1.6 อุปกรณ์และเครื่องมือที่ใช้ในการดำเนินงาน

1. ฮาร์ดแวร์ อุปกรณ์ที่เป็นสิ่งสำคัญและต้องนำมาใช้ในการทำโปรเจค ประกอบด้วย

คอมพิวเตอร์เครื่องที่ 1 หน่วยประมวลผลกลาง (CPU) รุ่น Intel(R) Core(TM) i5-1135G7 ความเร็ว 2.40 GHz, 2.42 GHz

หน่วยความจำหลัก RAM 8.00 GB

ระบบปฏิบัติการ Windows 11 Home Single Language

คอมพิวเตอร์เครื่องที่ 2 หน่วยประมวลผลกลาง (CPU) AMD Ryzen 5 4600H

ความเร็ว 3.00 GHz

หน่วยความจำหลัก RAM 16.0 GB

ระบบปฏิบัติการ Windows 11 Home Single Language

2. ซอฟต์แวร์ที่ใช้ Visual Studio Code ในการสร้างอัลกอริทึมและ Power Bi สร้างกราฟแสดงผลข้อมูล

1.7 แผนการดำเนินงาน

โครงงานปริญญานิพนธ์ฉบับนี้ ดำเนินงาน ณ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างเดือน มิถุนายน 2567 ถึง พฤษจิกายน 2567 ดังตารางที่ 1

ตารางที่ 1 แผนการดำเนินงาน

กิจกรรม	เดือน						
	มิ.ย.	ก.ค.	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.
1. ศึกษาและรวบรวมข้อมูล							
2. วิเคราะห์และกำหนดขอบเขต							
3. ออกแบบขั้นตอนการดำเนินการ							
4. พัฒนาโครงงาน							
5. ทดสอบและวัดประสิทธิภาพ							
6. ทำรายงานสรุป							

7. นำเสนอโครงงาน							

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การพยากรณ์

การพยากรณ์เป็นการคาดการณ์เกี่ยวกับลักษณะหรือ แนวโน้มของสิ่งใดสิ่งหนึ่งที่สนใจที่จะเกิดขึ้นในอนาคต เพื่อใช้เป็นสารสนเทศ (Information) ประกอบการตัดสินใจ ซึ่งการพยากรณ์จะต้องดำเนินการเป็นส่วนแรกสุดที่จะต้องทำก่อนการวางแผน หรือการเตรียมการที่จะริเริ่มการใดๆ เพื่อความถูกต้อง และแม่นยำในการตัดสินใจ ดังนั้นในการดำเนินธุรกิจภายใต้ความไม่แน่นอน จำเป็นที่จะต้องทราบถึงความเป็นไปในอนาคตโดยอาศัยเทคนิค หรือวิธีการพยากรณ์ต่างๆ ซึ่งอาจจะนำหลายๆ วิธีมาใช้โดยขึ้นอยู่กับสถานการณ์ด้วย เช่น นำข้อมูล ในอดีตมาพยากรณ์หาเหตุการณ์ในอนาคตด้วยการอาศัยหลักการทางคณิตศาสตร์เข้ามาช่วย ทั้งนี้อาจจะใช้ดุลพินิจของผู้พยากรณ์เพียงอย่างเดียว หรืออาจใช้หลายๆ วิธีเข้าด้วยกันเพื่อให้ผลการพยากรณ์มีความแม่นย่ามากที่สุด

ประเภทของการพยากรณ์

การพยากรณ์เชิงคุณภาพ (Qualitative Forecasting**)** ใช้วิจารณญาณหรือความคิดเห็นของผู้เชี่ยวชาญในการคาดการณ์ เหมาะสำหรับสถานการณ์ที่มีข้อมูลเชิงปริมาณน้อย เช่น การคาดการณ์แนวโน้มทางสังคม

การพยากรณ์เชิงปริมาณ (Quantitative Forecasting) ใช้ข้อมูลเชิงตัวเลขและแบบจำลองทางคณิตศาสตร์ในการพยากรณ์ เช่น การพยากรณ์ยอดขายโดยอิงจากข้อมูลในอดีตหรือการวิเคราะห์อนุกรมเวลา (Time Series)

2.1.2 คาร์บอนไดออกไซด์

คาร์บอนไดออกไซด์ (Carbon Dioxide) เป็นหนึ่งในแก๊สเรือนกระจก (Greenhouse gases) ซึ่งเป็นแก๊สที่มีอยู่ตามธรรมชาติและไม่เป็นอันตรายหากมีในปริมาณน้อยแต่ทว่าในสถานการณ์ปัจจุบันได้มีการเพิ่มขึ้นของคาร์บอนไดออกไซด์เป็นอย่างมาก จนทำให้ถึงระดับที่ส่งผลต่อการเปลี่ยนแปลงทางธรรมชาติต่าง ๆ ซึ่งสาเหตุของการเพิ่มขึ้นส่วนใหญ่ก็มาจากฝีมือของมนุษย์ เช่น การผลิตไฟฟ้า การใช้น้ำมันในยานพาหนะต่าง ๆ ในทางองค์ประกอบทางเคมีคาร์บอนไดออกไซด์ประกอบด้วย คาร์บอน 1 ส่วน และ ออกซิเจน 2 ส่วน คาร์บอนไดออกไซด์เรียกได้ว่าเป็นหนึ่งในแก๊สที่มีความสำคัญมากถึงมากที่สุดของโลกเพราะว่า คาร์บอนไดออกไซด์นั้นเป็นส่วนหนึ่งในกระบวนการสร้างอาหารของพืช

(photosynthesis) หรือก็คือพืชพันธุ์ต่าง ๆ มีชีวิตอยู่ได้ก็ด้วยคาร์บอนไดออกไซด์ เช่นเดียวกับที่มนุษย์มีชีวิตอยู่ได้จากการหายใจด้วยแก๊สออกซิเจน นั้นหมายความว่าหากไม่มีคาร์บอนไดออกไซด์ ก็จะไม่มีพืช และถ้าไม่มีพืช สิ่งมีชีวิตต่าง ๆ เองก็คงจะอยู่ไม่ได้

แต่ทั้งนี้ทุกอย่างเมื่อมีข้อดีก็ย่อมจะมีข้อเสียตามมา เพราะว่าคาร์บอนไดออกไซด์นั้นเป็นสาเหตุส่วนหนึ่งของการทำให้เกิดภาวะโลกร้อน (warming effect) ซึ่งเป็นปรากฏการณ์ที่ทำให้เกิดการเปลี่ยนแปลงทางสภาวะอากาศ หรือภัยพิบัติทางธรรมชาติต่าง ๆคาร์บอนไดออกไซด์ (Wikipedia 2024) หรือ CO_2 เป็นก๊าซไม่มีสี ซึ่งหากได้รับก๊าซนี้เข้าไปในปริมาณมากจะรู้สึกเปรี้ยวที่ปาก เกิดการระคายเคืองที่จมูกและคอ เนื่องจากอาจเกิดการละลายของแก๊สนี้ในเมือกในอวัยวะ ก่อให้เกิดกรดคาร์บอนิกอย่างอ่อนคาร์บอนไดออกไซด์มีความหนาแน่น 1.98 kg/m3 ซึ่งเป็นประมาณ 1.5 เท่าของอากาศ โมเลกุลประกอบด้วยพันธะคู่ 2 พันธะ (O=C=O) หรือ CO_2 น้ำหนักโมเลกุล 44.01 ไม่ติดไฟและไม่ทำปฏิกิริยา

คาร์บอนไดออกไซด์ในสถานะของแข็ง เรียกอีกชื่อหนึ่งว่า คาร์บอนไดออกไซด์แข็ง หรือ solid carbon dioxide เตรียมได้จากการนำแก๊สคาร์บอนไดออกไซด์มาผ่านกระบวนการอัดและทำให้เย็นลงภายใต้ความดันสูงกลายเป็นคาร์บอนไดออกไซด์เหลว แล้วลดความดันลงอย่างรวดเร็วโดยการพ่นคาร์บอนไดออกไซด์เหลวสู่ความดันบรรยากาศ ผลที่ได้คือเกล็ดน้ำแข็งคล้ายเกล็ดหิมะแล้วจึงนำมาอัดเป็นรูป คาร์บอนไดออกไซด์จะกลายเป็นของแข็งที่มีสีขาวอุณหภูมิ -78 องศาเซลเซียส โดยไม่ผ่านการเป็นของเหลวก่อน หากต้องการทำให้คาร์บอนไดออกไซด์เป็นของเหลว ต้องใช้ความดันไม่น้อยกว่า 5.1 บรรยากาศ คาร์บอนไดออกไซด์สามารถละลายน้ำได้ 1 เปอร์เซ็นต์ของสารละลายนั้นจะกลายเป็นกรดคาร์บอนิกซึ่งจะเปลี่ยนรูปเป็นไบคาร์บอเนตและคาร์บอเนตในภายหลัง (ทัศนะนาคะจิตต์, 2562)

1. พิษจากคาร์บอนไดออกไซด์

พิษคาร์บอนไดออกไซด์เกิดขึ้นเมื่อหายใจเอาอากาศที่มีคาร์บอนไดออกไซด์ 5% ขึ้นไปตามปริมาตร อาการที่พบได้บ่อยที่สุดของพิษคาร์บอนไดออกไซด์ ได้แก่ ปวดศีรษะ เวียนศีรษะอ่อนแรง เจ็บหน้าอก และสับสน หากไม่สามารถรับอากาศบริสุทธิ์ได้ทันที อาจเกิดภาวะหายใจไม่ออกได้ยังไม่มีผลกระทบใดๆ จากการที่ก๊าซ CO_2 สัมผัสกับดวงตาหรือผิวหนัง แม้ว่าก๊าซ CO_2 จะเป็นก๊าซที่ไม่มีกลิ่น แต่หลายคนก็บอกว่ากลิ่นของก๊าซ CO_2 ในระดับที่สูงกว่าเป็นกลิ่นฉุนหรือมีกลิ่นเปรี้ยว นั่นเป็นเพราะว่าก๊าซ CO_2 จะสร้างกรดคาร์บอนิกในร่างกายของคุณต่างจากก๊าซ CO_2 ก๊าซ CO_2 ที่เป็นของเหลวหรือแข็งตัว (เรียกว่าน้ำแข็งแห้ง) เป็นอันตรายเมื่อสัมผัส ควรสวมถุงมือหุ้มฉนวนและหน้ากากทุกครั้งที่สัมผัสน้ำแข็งแห้ง (co2meter, 2024)

2. ปริมาณ CO_2 เท่าไหร่ที่เป็นอันตราย

ในธรรมชาติ คาร์บอนไดออกไซด์ (CO_2) มีเพียงประมาณ 0.04% ของปริมาตรรวมของก๊าซในอากาศบริสุทธิ์ อย่างไรก็ตาม เมื่อเปลี่ยนจากของเหลวหรือของแข็งเป็นก๊าซ มันจะขยายตัวเป็น 535 เท่าของปริมาตรเดิม นั่นหมายความว่าในพื้นที่ปิด แม้เพียงการรั่วไหลเล็กน้อยจากถังหรือกระบอก CO_2 ก็สามารถเพิ่มระดับ CO_2 ให้สูงถึง 5% หรือมากกว่าได้อย่างรวดเร็ว ซึ่งอาจนำไปสู่อาการหายใจลำบากหรือการขาดอากาศหายใจ

แม้จะไม่อันตรายแต่ระดับ CO_2 ที่สูงขึ้นในอาคารเนื่องจากการหายใจตามปกติของมนุษย์ก็สามารถส่งผลกระทบต่อเราได้ ในแต่ละครั้งที่เราหายใจออก ลมหายใจจะมี CO_2 ประมาณ 3% การศึกษาแสดงให้เห็นว่าในห้องปิด แม้ที่ระดับเกิน 950 ส่วนในล้านส่วน (ppm) ของ CO_2 ก็สามารถนำไปสู่อาการปวดศีรษะ และสมาธิลดลง
สำนักงานบริหารความปลอดภัยและอาชีวอนามัยแห่งสหรัฐอเมริกา (OSHA) ได้กำหนดช่วงการหายใจที่เหมาะสมที่สุดอยู่ระหว่าง 19.5 ถึง 23.5 เปอร์เซ็นต์ของออกซิเจน ผลข้างเคียงที่ร้ายแรงอาจเกิดขึ้นได้หากระดับออกซิเจนอยู่นอกเขตปลอดภัย ที่ระดับ 17 เปอร์เซ็นต์หรือต่ำกว่าความสามารถทางสมองของคุณจะเริ่มบกพร่อง

เมื่อเราพูดถึงความเป็นพิษและอันตรายของ CO_2 การศึกษายังแสดงให้เห็นถึงปัญหาเมื่อบุคคลสัมผัสกับระดับที่สูงกว่า 5,000 ppm เป็นเวลาหลายชั่วโมง เนื่องจากการใช้ระบบ CO_2 อัดแรงดันสำหรับเครื่องดื่ม เหตุการณ์ที่เกี่ยวข้องกับ CO_2 ได้เพิ่มขึ้นในร้านอาหาร โรงเบียร์ สถานที่เพาะปลูกในร่ม และสนามกีฬาที่ให้บริการน้ำอัดลมหรือเบียร์ โปรดจำไว้เสมอว่าในพื้นที่จำกัดหรือเมื่อหายใจในสภาพแวดล้อมปิด CO_2 สามารถสะสมได้อย่างรวดเร็ว และสุขภาพโดยรวมอาจตกอยู่ในความเสี่ยงได้ (co2meter, 2024)

การเตรียมข้อมูลสำหรับการจำแนก

1. การทำความสะอาดข้อมูล เป็นการทำงานที่เกี่ยวกับการตรวจสอบและแก้ไขข้อมูลเพื่อให้ ข้อมูลอยู่ใน รูปแบบที่ถูกต้องและสมบูรณ์ โดยขั้นตอนของการทำความสะอาดข้อมูลนับเป็นขั้นตอนแรก ที่สำคัญของการเตรียมพร้อมข้อมูล ซึ่งถ้าข้อมูลไม่มีความถูกต้อง หรือไม่สมบูรณ์ อาจจะส่งผลให้คำตอบ หรือข้อสรุปที่ได้ไม่สามารถนำไปใช้ประโยชน์ได้

2. การแปลงข้อมูล เป็นอีกขั้นตอนหนึ่งที่สำคัญในกระบวนการเตรียมข้อมูลเพื่อนำไปวิเคราะห์ ด้วยว่า ข้อมูลที่ได้มานั้นบางครั้งอาจได้มาจากหลายแหล่ง ซึ่งแต่ละแหล่งอาจมีการจัดเก็บที่แตกต่างกัน และทำให้อยู่ในรูปตัวเลขเพื่อให้ง่ายต่อการนำไปใช้งานอัลกอริทึม

2.1.3 การทำเหมืองข้อมูล (data mining)

การทำเหมืองข้อมูล (amazon, 2023) คือเทคนิคที่ใช้คอมพิวเตอร์ช่วยในการวิเคราะห์เพื่อประมวลผลและสำรวจชุดข้อมูลขนาดใหญ่ เมื่อใช้เครื่องมือและวิธีการทำเหมืองข้อมูล องค์กรสามารถค้นพบรูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในข้อมูลของตน การทำเหมืองข้อมูลแปลงข้อมูลดิบเป็นความรู้เชิงปฏิบัติ บริษัทใช้ความรู้นี้ในการแก้ไขปัญหา วิเคราะห์ผลกระทบในอนาคตของการตัดสินใจทางธุรกิจ และเพิ่มขอบเขตกำไรของบริษัท

1. เทคนิคการทำเหมืองข้อมูล

   เทคนิคการทำเหมืองข้อมูลอิงจากสาขาวิชาต่างๆ ที่ทับซ้อนกัน รวมถึงการวิเคราะห์ทางสถิติ แมชชีนเลิร์นนิง (ML) และคณิตศาสตร์

2. การทำเหมืองตามกฎความเกี่ยวข้อง

   การทำเหมืองกฎการเชื่อมโยงเป็นกระบวนการในการค้นหาความสัมพันธ์ระหว่างชุดข้อมูลสองชุดที่ดูเหมือนไม่เกี่ยวข้องกัน คำสั่ง if-then แสดงให้เห็นถึงความน่าจะเป็นของความสัมพันธ์ระหว่างจุดข้อมูลสองจุด นักวิทยาศาสตร์ข้อมูลจะวัดความถูกต้องของผลลัพธ์โดยใช้เกณฑ์การสนับสนุนและความมั่นใจ การสนับสนุนวัดความถี่ที่องค์ประกอบที่ เกี่ยวข้องปรากฏในชุดข้อมูล ในขณะที่ความมั่นใจจะแสดงจำนวนครั้งที่คำสั่ง if-then นั้นถูกต้อง ตัวอย่างเช่น เมื่อลูกค้าซื้อสินค้า พวกเขาก็มักจะซื้อสินค้าที่เกี่ยวข้องกันเป็นลำดับที่สอง ผู้ค้าปลีกสามารถใช้การเชื่อมโยงข้อมูลการซื้อที่ผ่านมาเพื่อระบุความสนใจของลูกค้าใหม่ พวกเขาใช้ผลการทำเหมืองข้อมูลเพื่อเติมส่วนที่แนะนำของร้านค้าออนไลน์

2.1.4 Exploratory Data Analysis

Exploratory Data Analysis หรือ EDA คือ กระบวนการตรวจสอบข้อมูล หรือ การสำรวจข้อมูลเบื้อต้น เพื่อทำความเข้าใจของชุดข้อมูลที่จะใช้ทำงานได้ดียิ่งขึ้น

1. ประโยชน์ของการทำ EDA
   - เพื่อช่วยให้มีความเข้าใจและความรู้พื้นฐานเกี่ยวกับข้อมูลนั้นๆ
   - เพื่อช่วยในการตรวจสอบสมมุติฐานเบื้องต้นและตรวจสอบความผิดพลาดของข้อมูล
   - เพื่อให้เราเห็นค่าที่โดดออกมาจากค่าปกติ (Outlier) เพื่อป้องกันความผิดเพี้ยนตอนนำข้อมูลไปวิเคราะห์ หรือคำนวณในภายหลัง

- เพื่อให้เราเข้าใจข้อมูล มองเห็น Trends, Patterns หรือ Insights ต่างๆ เพื่อนำไป Take Action หรือต่อยอดธุรกิจได้อย่างรวดเร็ว

2. กระบวนการ EDA

- Data Transformation

คือการจัดเตรียมข้อมูลเพื่อให้พร้อมและดูเข้าใจสำหรับการวิเคราะห์ เพราะถ้ายังไม่ได้ถูกนำมาแปลงและวิเคราะห์อย่างเหมาะสม ก็แทบจะไม่มีค่าเลย การที่เราจะนำข้อมูลไปวิเคราะห์หรือสร้างแบบจำลองทางสถิติต่อได้ง่าย ต้องผ่านการเตรียมข้อมูล หรือ Data Cleansing ให้พร้อมก่อน โดยขั้นตอนการทำความสะอาดข้อมูลนี้ถือเป็นขั้นตอนที่ต้องใช้เวลามาก แต่เป็นขั้นตอนที่สำคัญที่สุด เช่น การทำให้ค่าทั้งหมดอยู่ในมาตราฐานเดียวกัน หรือที่เรียกว่าการ Normalize values, การเปลี่ยน Common Data types ให้เหมาะสม หรือ การจัดการกับแถวที่มีข้อมูลหายไป ด้วยการลบหรือเติมค่าลงไปให้เหมาะสม (Handle missing values)

- Data analysis

กระบวนการ เป็นการวิเคราะห์เพื่อหา Insight มาต่อยอด ไม่ว่าจะเป็น ความสัมพันธ์ต่างๆ ของตัวแปร, Patterns ที่ปรากฏขึ้นมา,การคำนวณค่าสถิติต่างๆ

- Data Visualization

เป็นขั้นตอนของการทำ Data Visualization หรือการนำข้อมูลออกมาแสดงให้ทั้งเราและคนที่เกี่ยวข้องเข้าใจกันได้ง่าย ๆ เมื่อการวิเคราะห์ข้อมูลออกมาได้แล้วนั้น สิ่งที่สำคัญมาก ๆ ต่อมาก็คือ การสื่อสารออกไปให้คนอื่นในทีมเห็นเป็นภาพเดียวกัน การทำ Data Visualization คือการสร้างกราฟ หรือ Chart ต่าง ๆ เพื่อให้เรานำเสนอข้อมูลในรูปแบบที่เข้าใจ insights ได้ง่ายขึ้น เป็นการแปลงข้อมูลให้เป็นภาพที่แค่มองครั้งแรกก็เข้าใจถึงสิ่งที่ต้องการจะสื่ออย่างชัดเจน

2.1.5 Time Series

เป็นเทคนิคทางสถิติที่ใช้ในการวิเคราะห์และทำความเข้าใจข้อมูลที่ขึ้นอยู่กับเวลา รวมถึงการแยกองค์ประกอบของอนุกรมเวลาออกเป็นแนวโน้ม (Trend) ฤดูกาล (Seasonality) และค่าคลาดเคลื่อน (Residuals) โดยการวิเคราะห์องค์ประกอบเหล่านี้ ทำให้สามารถระบุรูปแบบ ตรวจจับสิ่งผิดปกติ และตัดสินใจบนพื้นฐานของข้อมูลได้อย่างมีประสิทธิภาพ

1. องค์ประกอบสำคัญของ Time Series Analysis

แนวโน้ม (Trend): การเคลื่อนไหวในระยะยาวของอนุกรมเวลา มีแนวโน้มเพิ่มขึ้นหรือลดลง แนวโน้มอาจเป็นแบบเชิงเส้นหรือไม่เชิงเส้น และสามารถแสดงการเติบโตการเสื่อมลง หรือรูปแบบวัฏจักร

ฤดูกาล (Seasonality): รูปแบบที่เกิดขึ้นซ้ำในช่วงเวลาที่แน่นอน เช่น รายสัปดาห์ รายเดือน รายไตรมาส หรือรายปี ฤดูกาลอาจได้รับอิทธิพลจากปัจจัยต่างๆ เช่น สภาพอากาศ วันหยุด หรือพฤติกรรมของผู้บริโภค

วัฏจักร (Cyclicity): การเปลี่ยนแปลงระยะยาวที่ยืดเยื้อนานกว่าช่วงฤดูกาลเพียงอย่างเดียว วัฏจักรอาจเกิดจากปัจจัยทางเศรษฐกิจแนวโน้มของอุตสาหกรรม หรือปัจจัยภายนอกอื่นๆ

ค่าคลาดเคลื่อน (Residuals): การเปลี่ยนแปลงที่เหลืออยู่ในอนุกรมเวลาหลังจากคำนึงถึงแนวโน้ม ฤดูกาล และวัฏจักรแล้ว โดยทั่วไปค่าคลาดเคลื่อนจะถูกสมมติให้เป็นสัญญาณรบกวนหรือการเปลี่ยนแปลงที่ไม่สามารถอธิบายได้

2. การประยุกต์ใช้ในอุตสาหกรรม

การเงิน: การวิเคราะห์ราคาหุ้น อัตราแลกเปลี่ยน และดัชนีทางการเงิน เพื่อตัดสินใจในการลงทุนและบริหารความเสี่ยง

พลังงาน: การพยากรณ์ความต้องการใช้พลังงาน การจัดสรรทรัพยากรอย่างมีประสิทธิภาพ และตรวจสอบรูปแบบการใช้พลังงาน

การผลิต: การตรวจสอบกระบวนการผลิต ระบุปัญหาคุณภาพ และการบริหารจัดการสินค้าคงคลัง

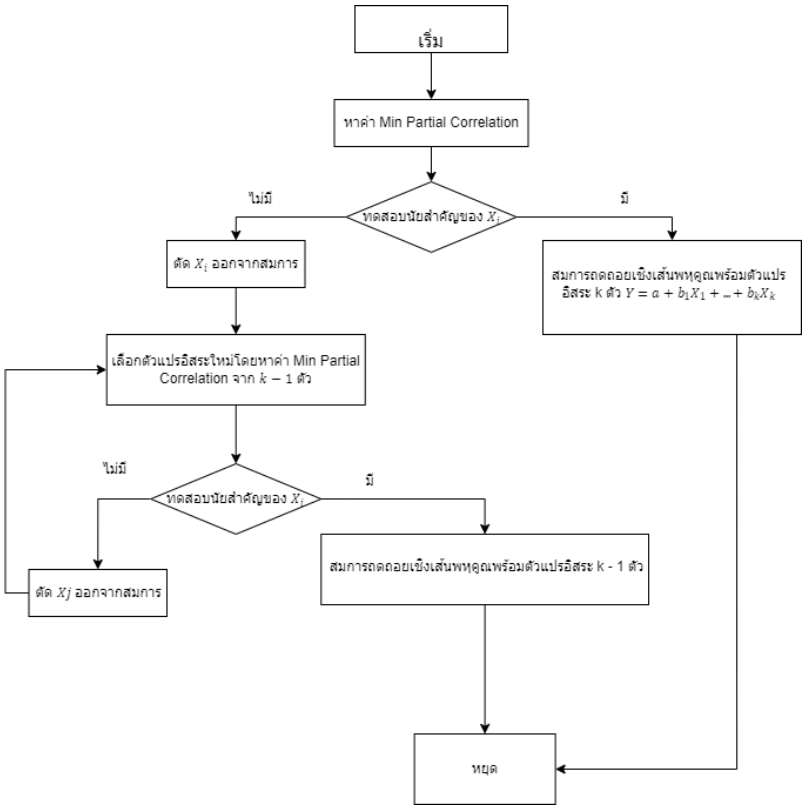
ค้าปลีก: การพยากรณ์ยอดขาย ทำความเข้าใจพฤติกรรมของลูกค้า และวางแผนโปรโมชั่นและแคมเปญการตลาด

สิ่งแวดล้อม: การวิเคราะห์ข้อมูลสภาพภูมิอากาศ พยากรณ์รูปแบบสภาพอากาศ และติดตามคุณภาพอากาศและน้ำ

2.1.6 วิธีการคัดเลือกตัวแปรอิสระแบบการกำจัดแบบถอยหลัง (Backward eliminate)

การคัดเลือกตัวแปรอิสระแบบการกำจัดแบบถอยหลัง (Backward Elimination) เป็นกระบวนการทางสถิติที่ใช้ในการสร้างแบบจำลองทางสถิติที่เหมาะสมโดยการลบตัวแปรที่ไม่มีนัยสำคัญออกจากแบบจำลองทีละตัว จากสมการถดถอยที่ประกอบด้วยตัวแปรอิสระ k ตัว โดยจะเริ่ม

พิจารณาจากตัวแปรอิสระที่มีความสัมพันธ์กับ Y น้อยที่สุด แล้วนำตัวแปรอิสระดังกล่าวมาทดสอบนัยสำคัญ ถ้าการทดสอบพบว่าไม่มีนัยสำคัญ แสดงว่าตัวแปรอิสระตัวนั้นจะถูกคัดออก และทำการคัดเลือกตัวแปรอิสระตัวที่ 2 ต่อไป แต่ในกรณีที่ทดสอบแล้วพบว่ามีนัยสำคัญจะหยุดทำการทดสอบและสรุปผลว่า สมการถดถอยประกอบด้วยตัวแปรอิสระทั้ง k ตัว ซึ่งมีขั้นตอนการวิเคราะห์ดังแผนผังการทำงานในภาพประกอบที่ 2



ภาพประกอบที่ 2 แผนผังการวิเคราะห์ข้อมูลโดยวิธีการคัดเลือกตัวแปรอิสระแบบการกำจัดแบบถอยหลัง

2.1.7 อัลกอริทึม Decision tree – Regression (การถดถอยของต้นไม้ตัดสินใจ)

Decision Tree Regression เป็นวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ไม่ใช้พารามิเตอร์ (Non-parametric) สำหรับการพยากรณ์ค่าตัวเลขต่อเนื่อง ซึ่งสร้างโมเดลในรูปแบบของโครงสร้างต้นไม้ ซึ่งโหนดภายใน (Internal Nodes) แทนการทดสอบบนคุณลักษณะ (Features) กิ่ง (Branches) แทนผลลัพธ์ของการทดสอบ และใบไม้ (Leaf Nodes) แทนค่าที่พยากรณ์ได้สุดท้าย เป็นแบบจำลองที่ใช้ในการทำนายผลลัพธ์ที่เป็นค่าต่อเนื่อง โดยมันทำงานผ่านการแบ่งชุดข้อมูลออกเป็นส่วนย่อยๆ ตามค่าคุณลักษณะต่างๆ ซึ่งจะสร้างโครงสร้างที่คล้ายกับต้นไม้ ที่แต่ละโหนด (Node) ข้อมูล

จะถูกแบ่งเพื่อทำให้ความแปรปรวนของตัวแปรเป้าหมายในแต่ละส่วนลดลง จุดประสงค์คือการทำให้ข้อมูลในโหนดย่อยๆ มีความเป็นเอกภาพมากที่สุด โหนดปลาย (Leaf Node) ซึ่งเป็นจุดสิ้นสุดของแต่ละสาขาจะแสดงค่าที่ใช้ในการทำนาย สำหรับข้อมูลหนึ่งชุด การทำนายจะเกิดขึ้นโดยการเดินตามโครงสร้างต้นไม้ไปตามค่าคุณลักษณะของข้อมูลนั้นจนถึงโหนดปลายที่ให้ค่าพยากรณ์

1. การทำงานของต้นไม้ตัดสินใจสำหรับการถดถอย (Decision Tree for Regression)

การเลือกตัวแปรและค่าตัดแบ่ง: เริ่มจากโหนดราก (Root Node) โดยการเลือกคุณลักษณะ (Feature) ที่ดีที่สุดในการแบ่งข้อมูลออกเป็นสองกลุ่ม วิธีการเลือกตัวแปรและค่าตัดแบ่งนั้นมักพิจารณาจากการลดความแปรปรวน (Variance) ของค่าตัวแปรเป้าหมายในแต่ละกลุ่มหลังจากการแบ่ง ข้อมูลในกลุ่มย่อยจะมีความเป็นเอกภาพมากขึ้น และค่าที่ทำนายจะใกล้เคียงกับค่าจริงมากขึ้น

การแบ่งข้อมูลซ้ำๆ: เมื่อข้อมูลถูกแบ่งแล้ว โหนดแต่ละโหนดจะทำหน้าที่เหมือนโหนดรากของชุดข้อมูลย่อย และกระบวนการแบ่งจะทำซ้ำกับข้อมูลที่เหลือ โดยต้นไม้จะสร้างโหนดใหม่ เรื่อยๆ จนกว่าจะถึงเงื่อนไขที่กำหนด เช่น การมีจำนวนข้อมูลในโหนด ต่ำกว่าขั้นต่ำที่กำหนด หรือความลึกของต้นไม้ถึงค่าที่กำหนด

โหนดปลาย (Leaf Node): เมื่อไม่สามารถแบ่งข้อมูลได้อีกต่อไป โหนดนั้นจะกลายเป็นโหนดปลาย ซึ่งโหนดปลายจะมีค่าเป็นค่าเฉลี่ยของตัวแปรเป้าหมายในข้อมูลย่อยของโหนดนั้น ค่าทำนายของต้นไม้ตัดสินใจจะเป็นค่าของโหนดปลายที่ข้อมูลนั้นเดินทางมาถึง

การทำนาย: เมื่อต้องการทำนายค่าจากต้นไม้ที่สร้างขึ้น ข้อมูลใหม่จะถูกป้อนเข้าต้นไม้และเดินทางผ่านโหนดต่างๆ ตามเงื่อนไขของคุณลักษณะ จนกระทั่งไปถึง โหนดปลายที่ให้ค่าทำนายเป็นค่าเฉลี่ยของข้อมูลในโหนดนั้น

2. สมการที่ใช้ในต้นไม้ตัดสินใจสำหรับการถดถอย (Decision Tree for Regression)

2.1 การคำนวณค่าเฉลี่ยในโหนดปลาย ดังสมการที่ (2-1)

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (2\text{-}1)$$

โดยที่

$\hat{y}$ คือค่าทำนายของโหนดปลาย

N คือจำนวนตัวอย่างในโหนดนั้น

y_i คือค่าของตัวแปรเป้าหมาย (target variable) ของตัวอย่างที่ i

2.2 การลดความแปรปรวน ดังสมการที่ (2-2)

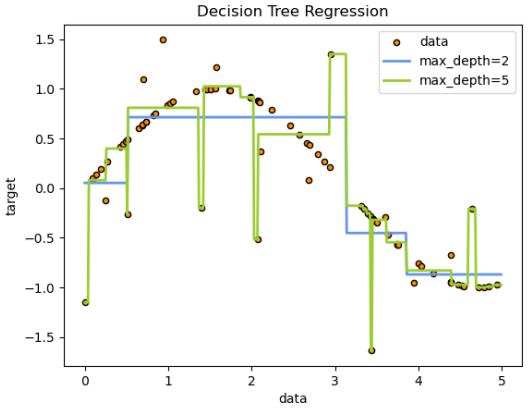
$$Variance\ Reduction = Variance\ before\ split - (\frac{N_L}{N}\ x\ Variance\ + \frac{N_R}{N}\ x\ Variance\ _r) \quad (2\text{-}2)$$

โดยที่

NL และ NR คือจำนวนตัวอย่างในโหนดซ้าย (Left) และโหนดขวา (Right)

VarianceL และ VarianceR คือค่าความแปรปรวนในแต่ละโหนดหลังการแบ่ง

N คือจำนวนตัวอย่างทั้งหมดก่อนการแบ่ง



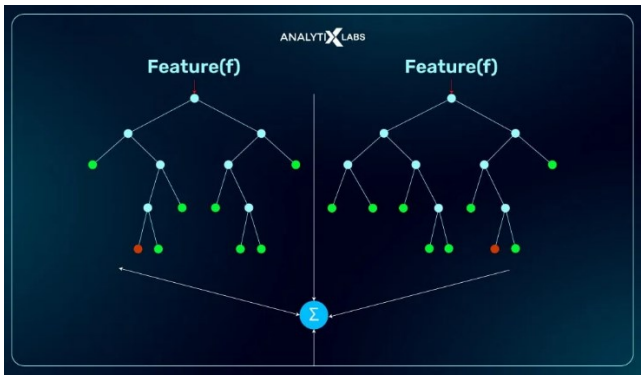
ภาพประกอบที่ 3 decision tree regression (developers, 2024)

2.1.8 อัลกอริทึม Random Forest – Regression (การถดถอยของป่าไม้สุ่ม)

Random Forest Regression การถดถอยแบบป่าไม้สุ่ม (Random Forest Regression) เป็นเครื่องมือในวิทยาศาสตร์ข้อมูลที่ช่วยให้สามารถพยากรณ์ข้อมูลได้อย่างแม่นยำและวิเคราะห์ชุดข้อมูลที่ซับซ้อนด้วยอัลกอริทึมการเรียนรู้ของเครื่องที่มีประสิทธิภาพ โมเดลการถดถอยแบบป่าไม้สุ่มสร้างขึ้นจากการรวมต้นไม้ตัดสินใจ (Decision Trees) หลายต้นเข้าด้วยกัน โดยต้นไม้แต่ละต้นในป่าจะสร้างขึ้นจากข้อมูลย่อยที่ถูกสุ่มเลือกมา และให้การพยากรณ์เป็นรายต้น จากนั้นการพยากรณ์สุดท้ายจะพิจารณาจากค่าเฉลี่ยหรือค่าเฉลี่ยถ่วงน้ำหนักของการพยากรณ์จากต้นไม้ตัดสินใจทั้งหมด (AnalytixLabs, 2023)

การทำงานการถดถอยแบบป่าไม้สุ่ม

การถดถอยแบบป่าไม้สุ่มเป็นอัลกอริทึมการเรียนรู้แบบมีผู้สอน โดยใช้ชุดของต้นไม้ตัดสินใจเพื่อพยากรณ์ตัวแปรเป้าหมายแบบต่อเนื่องโมเดลต้นไม้ตัดสินใจแต่ละตัวจะถูกสร้างขึ้นโดยใช้กระบวนการ bagging ซึ่งเป็นการสุ่มเลือกชุดย่อยของข้อมูลฝึกฝนเพื่อสร้างต้นไม้ตัดสินใจขนาดเล็ก หลังจากขั้นตอน bagging โดยจะรวมโมเดลขนาดเล็กเหล่านี้เข้าด้วยกันเพื่อสร้างโมเดลป่าไม้สุ่ม ซึ่งให้ค่าพยากรณ์เดียว วิธีนี้ช่วยลดความแปรปรวนและเพิ่มความแม่นยำโดยการรวมผลลัพธ์จากต้นไม้ตัดสินใจหลายต้น



ภาพประกอบที่ 4 การถดถอยแบบป่าไม้สุ่มทำงานอย่างไร? (AnalytixLabs, 2023)

ในกรณีของการถดถอยแบบป่าไม้สุ่ม (Random Forest Regression) สมการที่ใช้ในการคำนวณการพยากรณ์โดยทั่วไปสามารถอธิบายได้ในรูปของการรวมผลลัพธ์จากต้นไม้ตัดสินใจหลายต้น ดังสมาการที่ (2-3)

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \qquad (2\text{-}3)$$

โดยที่

y^ คือ ค่าการพยากรณ์สุดท้าย

T คือ จำนวนต้นไม้ตัดสินใจในป่า (Random Forest)

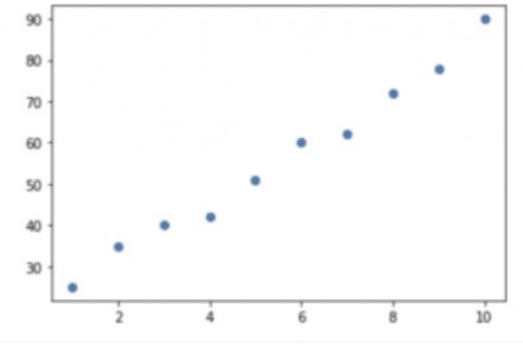
ft(x) คือ ผลลัพธ์จากต้นไม้ตัดสินใจที่ t เมื่อป้อนข้อมูล x แต่ละต้นไม้จะให้ผลลัพธ์การพยากรณ์ของตัวเอง ft(x) แล้วค่าเฉลี่ยของผลลัพธ์เหล่านี้จะถูกใช้เป็นค่าพยากรณ์สุดท้าย y^ ซึ่งช่วยให้โมเดลมีความแม่นยำและลดความแปรปรวนของผลลัพธ์

2.1.9 อัลกอริทึม Linear Regression

Linear regression คือเทคนิคพื้นฐานทางสถิติที่ถูกใช้เพื่อสร้างโมเดลความสำพันธ์ระหว่างเซตของตัวแปรอิสระและหนึ่งตัวแปรไม่อิสระ มันเป็นโมเดลที่ง่ายต่อการตีความความสัมพันธ์และถูกใช้เป็นเครื่องมือกันอย่างกว้างขวางในการวิเคราะห์ข้อมูล ทำนายและการทดสอบทางสถิติ(Chomchit, 2023)

Linear Regression ถือว่าเป็น Machine Learning ประเภท Supervised Learning หรือ การเรียนรู้แบบมีผู้สอน ชนิดแบบ Statistical Regression ที่เราจะต้องใส่ชุดข้อมูลเข้าไปให้โปรแกรมเรียนรู้ก่อน โดยโปรแกรมจะนำตัวแปรต้นและตัวแปรตามไปคำนวณด้วยสถิติทางคณิตศาสตร์ แล้วก็จะได้ข้อมูลกลับมาเป็นตัวเลข ความสัมพันธ์ของ Linear Regression หลัก ๆ จะเป็นดังต่อไปนี้(Zheng, 2021)

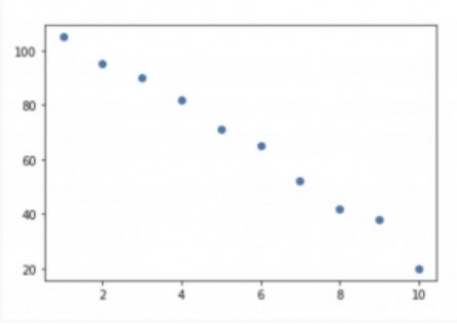
1. Positive Linear Relationship



ภาพประกอบที่ 5 Positive Linear Relationship (Zheng, 2021)

ถ้าค่า r เข้าใกล้ค่า +1.0 จะถือว่าเป็น Positive linear relationship ซึ่งเป็นการแปรผันตรง เมื่อค่าของตัวแปรต้น (x) เพิ่ม ค่าของตัวแปรตาม (y) ก็จะเพิ่มเช่นกัน

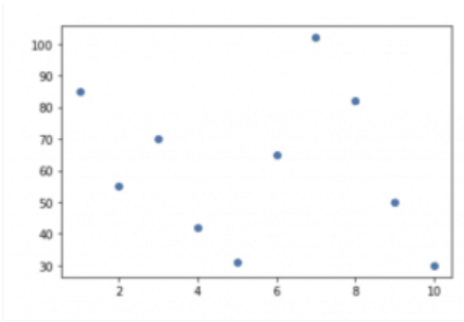
2. Negative Linear Relationship



ภาพประกอบที่ 6 Negative Linear Relationship (Zheng 2021)

ถ้าค่า r เข้าใกล้ค่า -1.0 จะถือว่าเป็น Negative linear relationship ซึ่งเป็นการแปรผกผัน เมื่อค่าของตัวแปรต้น (x) เพิ่ม ค่าของตัวแปรตาม (y) จะลดลง

3. No Apparent Linear Relationship



ภาพประกอบที่ 7 No Apparent Linear Relationship (Zheng, 2021)

ถ้าค่า r เข้าใกล้ค่า 0 จะถือว่าเป็น No apparent linear relationship รูปแบบที่ตัวแปรต้น (x) และตัวแปรตาม (y) ไม่มีความสัมพันธ์กันชัดเจน จึงไม่สามารถระบุได้ว่าเป็นความสัมพันธ์แบบไหนนั่นเอง

2.1.10 การวัดประสิทธิภาพ

(hitexts, 2022) การประเมินผลนั้นจำเป็นกับการทำ Machine Learning อย่างมาก เพราะเป็นสิ่งที่เอาไว้วัดว่าโมเดลทำงานอย่างไร ประเมินผลการทำงานน่าพึงพอใจหรือไม่ รวมถึงไว้ใช้เปรียบเทียบระหว่างโมเดลเพื่อเลือกใช้โมเดลได้อย่างเหมาะสม ซึ่งการวัดผลสำหรับโมเดลนั้นบทความนี้จะมานำเสนอการประเมินผลสำหรับงานด้าน Regression (ทำนายค่า)

Model Evaluation สำหรับงานด้าน Regression

Root Mean Squared Error (RMSE) เป็นการวัดความแตกต่างเฉลี่ยระหว่างค่าที่ทำนายและค่าจริงโดยการยกกำลังสองของความผิดพลาด แล้วนำมาหารเฉลี่ยก่อนจะถอดรากที่สองออกมา โดยที่ RMSE ต่ำ หมายถึงโมเดลมีความผิดพลาดน้อย โมเดลที่มี RMSE ต่ำกว่าจะมีความแม่นยำมากกว่า

Mean Squared Error (MSE) คือ ค่าเฉลี่ยของกำลังสองของความแตกต่างระหว่างค่าจริง (Actual Value หรือ Observed Value) กับค่าที่โมเดลทำนาย (Predicted Value) กล่าวอีกนัยหนึ่งคือ เป็นการวัดโมเดลทำนายค่าออกมาแตกต่างจากค่าจริงมากน้อยเพียงใด โดยมีหน่วยเป็น “กำลังสองของหน่วยต้นทาง” ค่าที่ต่ำแสดงว่าโมเดลมีความผิดพลาดน้อย

Mean Absolute Error (MAE) เป็นอีกหนึ่งมาตรวัดประสิทธิภาพของโมเดลการทำนาย โดยเฉพาะในปัญหา Regression เช่นเดียวกับ Mean Squared Error (MSE) แต่ MAE วัดค่าความผิดพลาดด้วยการใช้ค่า "ค่าสัมบูรณ์" ของความแตกต่างระหว่างค่าทำนาย (Predicted Value) และค่าจริง (Actual Value) ซึ่งทำให้มีความแตกต่างกับ MSE ที่ใช้กำลังสองของค่าความผิดพลาด ยิ่งค่า MAE ต่ำแสดงว่าโมเดลมีความผิดพลาดน้อย

R^2 หรือ R-Squared คือตัวสถิติที่ใช้วัดว่าตัวแบบคณิตศาสตร์ที่ได้นี้มีความสมรูปกับข้อมูลมากน้อยอย่างไร หรือรู้จักกัน ในอีกความหมายหนึ่งว่าเป็นค่าสัมประสิทธิ์แสดงการตัดสินใจ (Coefficient of Determination) หรือค่าสัมประสิทธิ์แสดงการตัดสินใจเชิงซ้อน (Coefficient of Multiple Determination) สำหรับการวิเคราะห์การถดถอยแบบพหุคูณ (Multiple Regression)

- ค่าที่ใกล้ 1 หมายถึง โมเดลสามารถอธิบายความแปรผันของข้อมูลได้ดี (เหมาะสมกับข้อมูลมาก)

-ค่าที่ใกล้ 0 หมายถึง โมเดลอธิบายความแปรผันของข้อมูลได้ไม่ดี (ไม่เหมาะสมกับข้อมูล)

2.1.11 การแสดงข้อมูลด้วยภาพ

การแสดงข้อมูลด้วยภาพ เป็นกระบวนการใช้ส่วนประกอบของศิลปะที่มองเห็นได้ เช่น แผนภูมิ กราฟ หรือแผนที่ในการแสดงข้อมูล ซึ่งเป็นการแปลข้อมูลที่ซับซ้อน ข้อมูลจำนวนมาก หรือข้อมูลตัวเลขให้แสดงเป็นภาพเพื่อง่ายในการประเมินผล เครื่องมือการแสดงข้อมูลด้วยภาพช่วยปรับปรุงกระบวนการสื่อสารด้วยภาพและทำให้เป็นระบบอัตโนมัติ เพื่อความถูกต้องและรายละเอียดที่ชัดเจน คุณสามารถใช้การแสดงด้วยภาพเพื่อดึงข้อมูลเชิงลึกที่นำไปใช้ได้จริงจากข้อมูลดิบ(Service, 2023)

เทคนิคประเภทต่างๆ ในการแสดงข้อมูลด้วยภาพ

แม้ว่าแผนภูมิและกราฟจะเป็นรูปแบบทั่วไป แต่ก็สามารถใช้วิธีการแสดงข้อมูลด้วยภาพได้หลากหลายวิธี วิธีการแสดงข้อมูลด้วยภาพมีหลักๆ อยู่ห้าประเภทดังต่อไปนี้

1. การแสดงข้อมูลด้วยภาพแบบชั่วคราว

การแสดงข้อมูลด้วยภาพแบบชั่วคราวใช้เพื่อแสดงถึงวัตถุเชิงเส้นหนึ่งมิติ เช่น กราฟเส้น แผนภูมิเส้น หรือไทม์ไลน์ ตัวอย่างเช่น คุณสามารถใช้แผนภูมิเส้น

เพื่อแสดงการเปลี่ยนแปลงที่เกิดขึ้นอย่างต่อเนื่องในช่วงเวลาที่กำหนด เส้นหลายๆ เส้นในแผนภูมิเส้นแสดงให้เห็นถึงความผันแปรของปัจจัยต่างๆ ในช่วงเวลาเดียวกัน

2. การแสดงข้อมูลด้วยภาพแบบเป็นลำดับชั้น

การแสดงข้อมูลด้วยภาพแบบเป็นลำดับชั้น หมายถึงกลุ่มหรือชุดของรายการ ที่มีความเชื่อมโยงทั่วไปกับรายการหลัก คุณสามารถใช้โครงสร้างข้อมูลแบบต้นไม้เหล่านี้ในการแสดงคลัสเตอร์ของข้อมูล ตัวอย่างเช่น คุณสามารถแสดงปริมาณข้อมูลสินค้า คงคลังเป็นแผนผังที่มีโหนดพ่อแม่ (เสื้อผ้า) และโหนดลูก (เสื้อเชิ้ต กางเกงขายาว และถุงเท้า)

3. การแสดงข้อมูลด้วยภาพแบบเครือข่าย

การแสดงข้อมูลด้วยภาพแบบเครือข่ายมีประโยชน์ในการแสดงความสัมพันธ์ที่ซับซ้อนระหว่างข้อมูลประเภทต่างๆ ที่เกี่ยวข้องกัน ตัวอย่างเช่น แผนภูมิจุดแบบกระจัดกระจายที่แสดงข้อมูลเป็นจุดบนกราฟ กราฟบับเบิ้ลที่เพิ่มปัจจัยของข้อมูลที่สามให้กับแผนภูมิจุดแบบกระจัดกระจาย Word Cloud ที่แสดงความถี่ของคำโดยใช้คำที่มีขนาดต่างๆกัน

4. การแสดงข้อมูลด้วยภาพแบบหลายมิติ

การแสดงข้อมูลด้วยภาพแบบหลายมิติจะแสดงตัวแปรข้อมูลตั้งแต่สองตัวขึ้นไปให้เป็นภาพแบบสองหรือสามมิติ แผนภูมิแท่ง แผนภูมิวงกลม และแผนภูมิแท่งแบบต่อกันเป็นตัวอย่างที่นิยมในการแสดงภาพเหล่านี้ ตัวอย่างเช่น แผนภูมิแท่งจะเปรียบเทียบปัจจัยข้อมูลตั้งแต่สองปัจจัยขึ้นไป และแสดงให้เห็นการเปลี่ยนแปลง ของตัวแปรหนึ่งตัวในช่วงระยะเวลาหนึ่ง แผนภูมิวงกลมแสดงภาพส่วนต่างๆของทั้งหมดในแต่ละหมวดหมู่

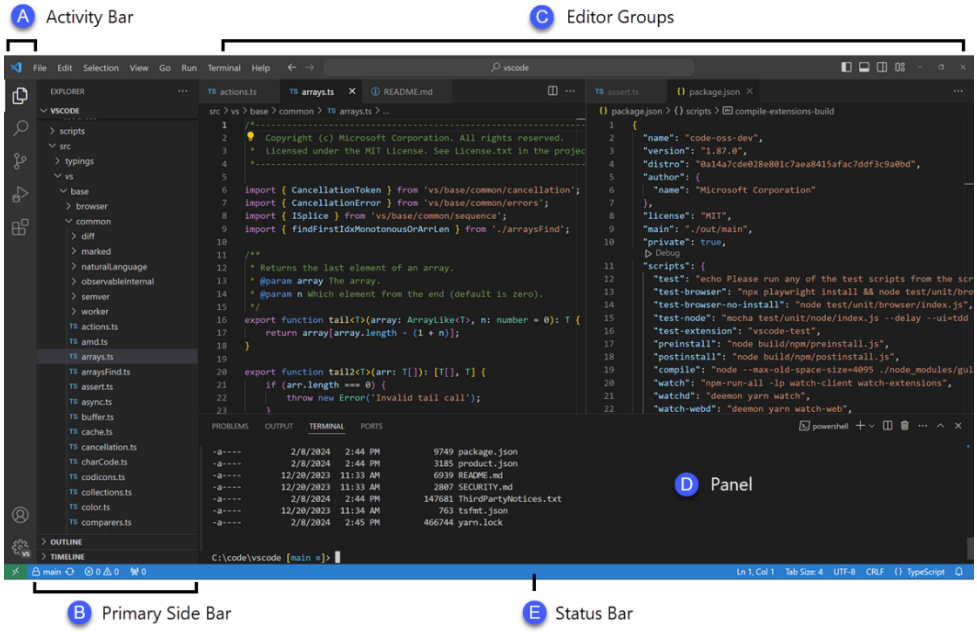
5. การแสดงข้อมูลด้วยภาพที่ระบุตำแหน่งที่ตั้ง

การแสดงข้อมูลด้วยภาพที่ระบุตำแหน่งที่ตั้ง ตัวอย่างเช่น การไฮไลท์สีที่ต่างกันคล้ายอุณหภูมิลงในการแสดงข้อมูล แผนที่แสดงความหนาแน่น หรือแผนที่คาร์โตแกรม ซึ่งเป็นการนำเสนอข้อมูลที่สัมพันธ์กับตำแหน่งในโลกความเป็นจริง

ตัวอย่างเช่น การแสดงข้อมูลด้วยภาพที่บอกถึงจำนวนลูกค้าที่เข้าชมสาขาของร้านค้าปลีกต่างๆ

2.1.12 Visual Studio Code

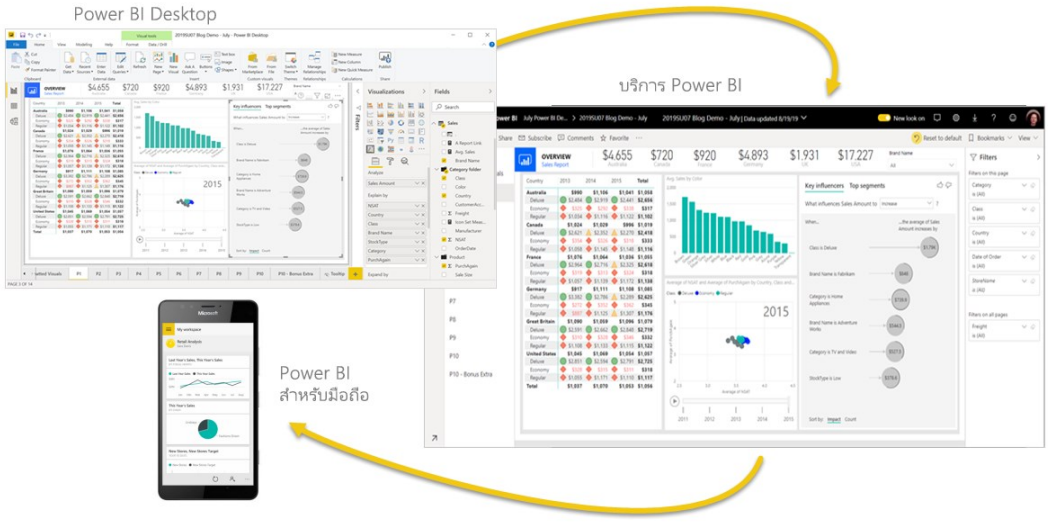
Vscode (Visual studio code) คือ โปรแกรมประเภท Editor หรือจะให้เรียกง่ายๆก็เครื่องมือในการเขียนโปรแกรมนั่นเอง โดยจะใช้ในการแก้ไข Code ที่มีขนาดเล็ก แต่มีประสิทธิภาพสูง เหมาะสำหรับนักพัฒนาโปรแกรมในทุกระดับ ไม่ว่าจะอยู่ในระดับเริ่มต้นยันมืออาชีพเลย รองรับการใช้งานทั้ง Windows, MacOS และ Linux รวมทั้งรองรับได้หลายภาษาไม่ว่าจะเป็น JavaScript, TypeScript, Python, C++ และ อื่นๆ สามารถนำมาใช้งานได้ง่ายไม่ซับซ้อน ซึ่งมีส่วนขยายหรือเครื่องมืออำนวยความสะดวกที่ให้เลือกใช้อยู่เยอะมาก



ภาพประกอบที่ 8 Visual studio code (visualstudio)

2.1.13 Power Bi

Power BI เป็นคอลเลกชันของบริการซอฟต์แวร์ แอป และตัวเชื่อมต่อที่ทำงานร่วมกันเพื่อเปลี่ยนแหล่งข้อมูลของคุณที่ไม่เกี่ยวข้องกันให้เป็นข้อมูลเชิงลึกที่สอดคล้องกัน เกี่ยวข้องกับผู้ใช้และโต้ตอบได้ ข้อมูลของคุณอาจเป็นสเปรดชีต Excel หรือคอลเลกชั่นของคลังข้อมูลระบบคลาวด์และคลังข้อมูลแบบไฮบริดภายในองค์กร Power BI ช่วยให้เชื่อมต่อกับแหล่งข้อมูลของคุณ แสดงภาพ และค้นพบเรื่องสำคัญ รวมถึงแชร์สิ่งนั้นกับบุคคลหรือทุกคนที่คุณต้องการได้อย่างง่ายดาย(microsoft, 2024)



ภาพประกอบที่ 9 Power Bi (microsoft, 2024)

2.1.14 งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้อง มีงานวิจัยเกี่ยวข้องกับการทำนายการปล่อยคารบอนมีตัวอย่างดังต่อไปนี้

(Cesar de Lima Nogueira, 2023) ได้นำเสนอการตรวจสอบประสิทธิภาพของแบบจำลอง Random Forest (RF) ใหม่ในการทำนายตัวแปรจากชุดข้อมูลการทดลองของเครื่องยนต์ดีเซลที่ดัดแปลงให้ใช้ได้ทั้งก๊าซธรรมชาติอัดและน้ำมันดีเซล โดย ดัดแปลงเครื่องยนต์ดีเซล 6 สูบให้ใช้ได้ทั้งก๊าซธรรมชาติอัดและน้ำมันดีเซล มีการศึกษาตัวแปร 5 ตัว โดยพิจารณาปัจจัยต่างๆ เช่น มุมฉีดน้ำมันเชื้อเพลิง อัตราส่วนอากาศต่อเชื้อเพลิง ใช้ Tree Structured Parzen Estimator และวิธี Feature Engineering 6 วิธีในการปรับแต่งพารามิเตอร์ของแบบจำลอง RF และ ใช้วิธี Shapley Additive Explanation (SHAP) เพื่อแปลผลลัพธ์ของแบบจำลอง RF ผลลัพธ์คือแบบจำลอง RF สามารถทำนายสัญญาณเอาต์พุตของเครื่องยนต์ดีเซลได้อย่างแม่นยำ โดยมีค่าสัมประสิทธิ์การตัดสินใจ (R^2) อยู่ระหว่าง 0.8842 ถึง 0.9811 สำหรับตัวแปรทั้ง 5 ตัวที่ศึกษา ซึ่งแสดงถึงประสิทธิภาพและความเป็นไปได้ในการใช้งานจริง (Cesar de Lima Nogueira, Och et al. 2023)

(Fuji, 2024) ได้ศึกษาวิธีการทำนายปริมาณการดูดซับ CO2 ของสารละลายเอมีน เพื่อเร่งการพัฒนาสารละลายเอมีนประสิทธิภาพสูงสำหรับการดูดซับ CO2 ทางเคมีในกระบวนการดักจับใช้ประโยชน์ และกักเก็บคาร์บอน โดย รวบรวมและจัดระเบียบข้อมูลจากงานวิจัยที่มีอยู่ (45 เอมีน, 3,151 จุดข้อมูล) ใช้การถดถอยแบบ Random Forest กับข้อมูลและตัวบ่งชี้ที่คำนวณด้วยซอฟต์แวร์ HSPiP หรือ RDkit วิเคราะห์ความสำคัญของคุณลักษณะ (Feature importance analysis) พัฒนารายการตัวบ่งชี้ที่ง่ายขึ้น โดยใช้อุณหภูมิ, ความดันย่อยของ CO2 และประจุย่อยของอะตอม N ที่คำนวณด้วยทฤษฎีฟังก์ชันความหนาแน่น ประเมินประสิทธิภาพการทำนายของแบบจำลองด้วยวิธี leave-one-group-out ผลลัพธ์ที่ได้ การถดถอยปริมาณการดูดซับ CO2 ของสารละลายเอมีนเดี่ยวมีความแม่นยำสูง (R^2 = 0.943, RMSE = 0.072–0.073 สำหรับข้อมูลตรวจสอบ) แบบจำลองที่ใช้ตัวบ่งชี้เพียง 5 ตัวสามารถทำนายปริมาณการดูดซับ CO2 ของสารละลายเอมีนเดี่ยวได้แม่นยำใกล้เคียงกับแบบจำลองแรก (R^2 = 0.931, RMSE = 0.079) แบบจำลองที่ใช้ตัวบ่งชี้ 8 ตัวสามารถทำนายปริมาณการดูดซับ CO2 ของทั้งสารละลายเอมีนเดี่ยวและผสมได้อย่างแม่นยำ (R^2 = 0.944, RMSE = 0.073) (Fujii, Sako et al. 2024)

(Lin, Jinyao, 2021) งานวิจัยนี้วิเคราะห์ตัวชี้วัดโครงสร้างอาคารที่ส่งผลต่อการปล่อย CO2 ในพื้นที่หนาแน่น โดยเน้นการออกแบบเมืองแนวตั้งเพื่อลดการปล่อย CO2 ในขั้นแรก ได้วิเคราะห์

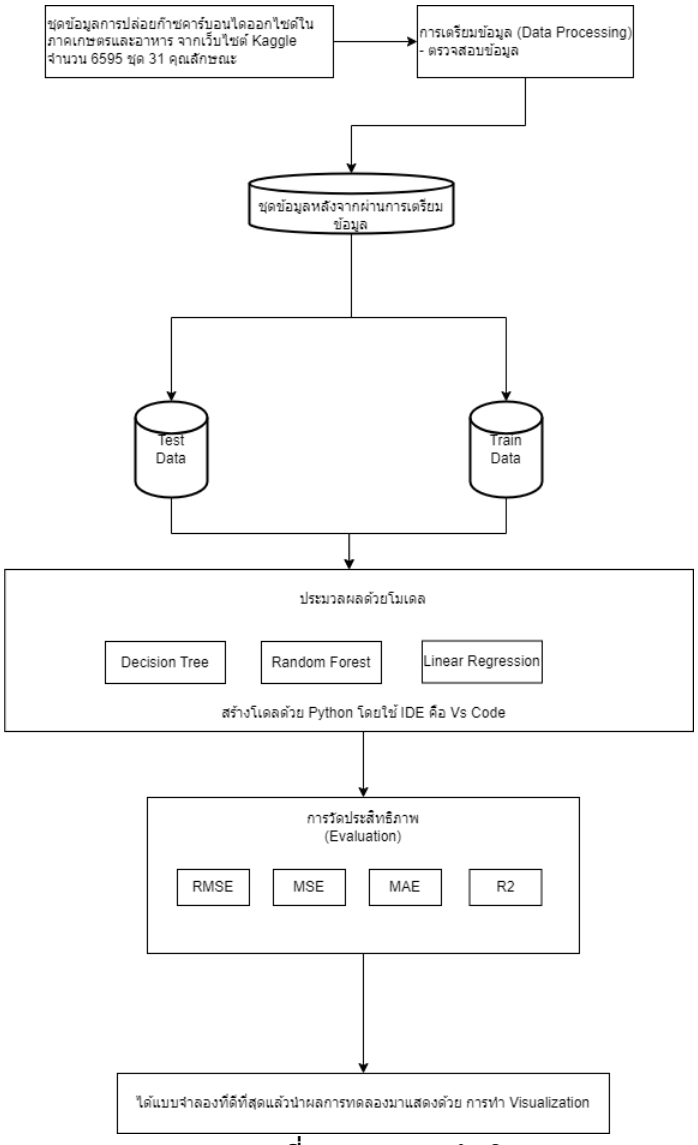
ความสัมพันธ์ระหว่างการปล่อย CO2 กับปัจจัยเชิงพื้นที่ด้วยการทดสอบเพียร์สัน (Pearson correlation) จากนั้นใช้การถดถอยแบบสุ่มฟอเรสต์ (Random Forest Regression) เพื่อประเมินว่าตัวชี้วัดอาคารสามารถอธิบายการเปลี่ยนแปลงของ CO2 ได้ดีเพียงใด ผลพบว่า อัตราการครอบคลุมอาคาร จำนวนอาคารเฉลี่ย ระดับความแออัด และอัตราพื้นที่ชั้น มีผลอย่างมากต่อการปล่อย CO2 โดยโมเดลปรับปรุงสามารถลดความคลาดเคลื่อนสัมพัทธ์เชิงรากที่สองจาก 34.68% เหลือ 32.53% ทำให้เป็นข้อมูลสำคัญในการวางแผนนโยบายเมืองที่ยั่งยืน (Lin, Lu et al. 2021)

บทที่ 3

วิธีดำเนินโครงการ

บทนี้เป็นการอธิบายขั้นตอนการดำเนินโครงการ ตั้งแต่การเตรียมข้อมูล การสร้างและประเมินโมเดลพยากรณ์อุณหภูมิ โดยใช้เทคนิคต่าง ๆ เช่น การทำความสะอาดข้อมูล การเลือกตัวแปร และ Cross-validation เพื่อตรวจสอบประสิทธิภาพของโมเดลที่สร้างขึ้น

3.1 แผนการดำเนินงาน



ภาพประกอบที่ 10 แผนการดำเนินงาน

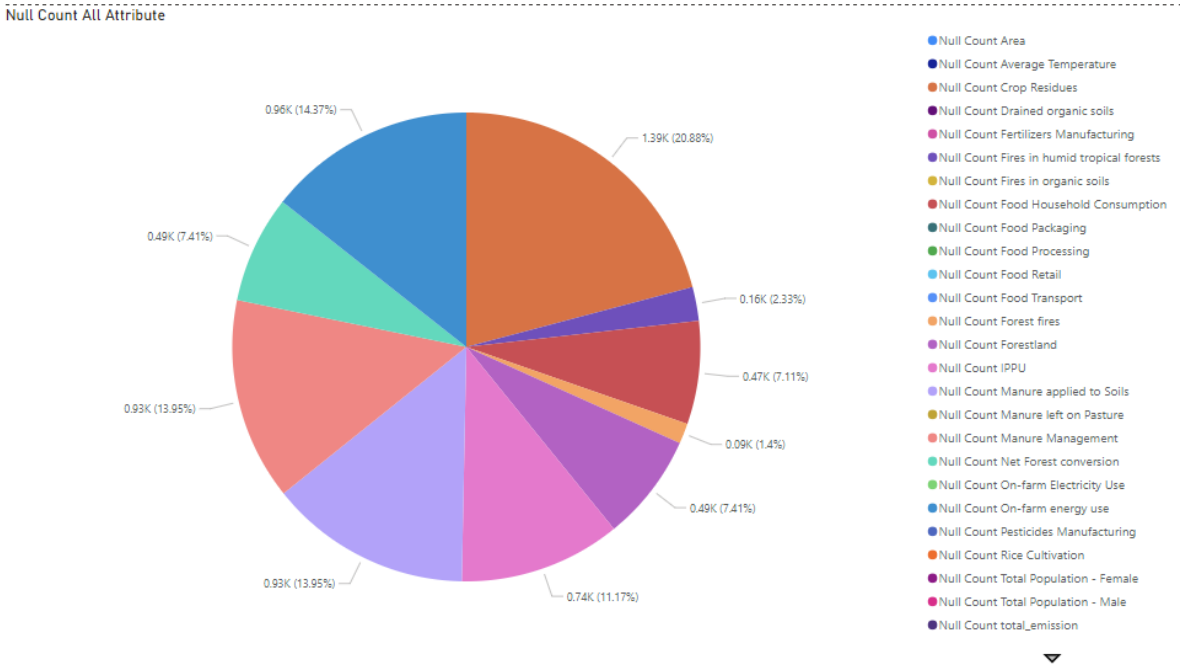
3.2 การเตรียมข้อมูล

ศึกษาปัจจัยต่าง ๆ ที่สามารถปล่อยคาร์บอนสู่ชั้นบรรยากาศหรือก่อให้เกิดก๊าซเรือนกระจก ได้แก่ การปล่อยก๊าซจากไฟในระบบนิเวศทุ่งหญ้าสะวันนา, การปล่อยก๊าซจากไฟในพื้นที่ป่า, การปล่อยก๊าซจากการเผาหรือการย่อยสลายเศษพืชหลังการเก็บเกี่ยว, การปล่อยก๊าซมีเทนจากการปลูกข้าว, การปล่อยก๊าซ, คาร์บอนไดออกไซด์เมื่อระบายน้ำในดินอินทรีย์, การปล่อยก๊าซจากการผลิตสารกำจัดศัตรูพืช,การปล่อยก๊าซจากการขนส่งสินค้าอาหาร เป็นต้น เพื่อนำข้อมูลดังกล่าวมาทำเหมืองข้อมูล โดยใช้ Python ในการสร้าง Model เพื่อนำมาวิเคราะห์ ซึ่งข้อมูลนี้จัดอยู่ในรูปแบบ CSV เป็นข้อมูลกลุ่มตัวอย่างจำนวน 6965 ระเบียน ประกอบไปด้วยตัวแปรทั้งหมด 31 แอตทริบิวต์ แบ่งเป็นตัวแปรต้น 30 แอตทริบิวต์ และตัวแปรตาม 1 แอตทริบิวต์ โดยมีปัจจัยที่สามารถนำมาวิเคราะห์ได้ ดังตารางที่ 2

ตารางที่ 2 ตัวอย่างข้อมูล

Area	Year	Savanna fires	Forest fires	Crop Residues
Afghanistan	1990	14.7237	0.0557	205.6077
Afghanistan	1991	14.7237	0.0557	209.4971
Afghanistan	1992	14.7237	0.0557	196.5341
Afghanistan	1993	14.7237	0.0557	230.8175

ตรวจสอบ Attribute จะเห็นว่ามี Attribute ชื่อ Savanna fires, Forest fires ,Crop Residues และตัวแปรอื่นๆมีค่า Missing จึงจำเป็นต้องกำจัดออกด้วยการทำความสะอาดข้อมูล



โดยตัวแปรที่มาค่า Missing มีดังตารางที่ 3

ตารางที่ 3 ค่า Missing

Attribute	Missing
Savanna fires	31
Forest fires	93
Crop Residues	1389
Forestland	493
Net Forest conversion	493
Food Household Consumption	473
IPPU	743
Manure applied to Soils	928
Manure Management	928
Fires in humid tropical forests	155
On-farm energy use	956

3.3 การทำความสะอาดข้อมูล

ข้อมูลที่นำมาตรวจสอบนั้นมีบางแอตทริบิวต์มีค่าที่ผิดปกติที่ไม่มีข้อมูลที่สามารถนำไปประมวลผล

ซึ่งค่าดังกล่าวอาจเกิดจากความผิดพลาดจากการเก็บรวบรวมข้อมูล ดังนั้น โครงงานนี้จึงทำการตรวจสอบข้อมูลที่สูญหายไปด้วยคำสั่ง fillna จะทำการเพิ่มค่าข้อมูลที่สูญหายไปด้วยค่าเฉลี่ย

	Savanna fires	Forest fires	Crop Residues	Forestland	Net Forest conversion	Food Household Consumption	IPPU	Manure applied to Soils	Manure Management	Fires in humid tropical forests
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
6960	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6961	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6962	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6963	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6964	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

ภาพประกอบที่ 11 การตรวจสอบข้อมูลที่สูญหาย

```
df["Savanna fires"].fillna(df["Savanna fires"].mean(), inplace = True)
df["Forest fires"].fillna(df["Forest fires"].mean(), inplace = True)
df["Crop Residues"].fillna(df["Crop Residues"].mean(), inplace = True)
df["Forestland"].fillna(df["Forestland"].mean(), inplace = True)
df["Net Forest conversion"].fillna(df["Net Forest conversion"].mean(), inplace = True)
df["Food Household Consumption"].fillna(df["Food Household Consumption"].mean(), inplace = True)
df["IPPU"].fillna(df["IPPU"].mean(), inplace = True)
df["Manure applied to Soils"].fillna(df["Manure applied to Soils"].mean(), inplace = True)
df["Manure Management"].fillna(df["Manure Management"].mean(), inplace = True)
df["Fires in humid tropical forests"].fillna(df["Fires in humid tropical forests"].mean(), inplace = True)
df["On-farm energy use"].fillna(df["On-farm energy use"].mean(), inplace = True)
✓ 0.0s
```

ภาพประกอบที่ 12 การแทนค่าว่างด้วย fillna ด้วยค่า mean

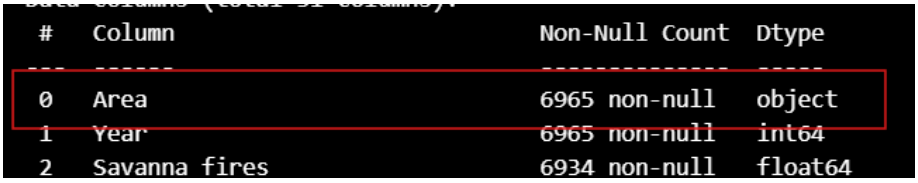
Savanna fires	Forest fires	Crop Residues	Forestland	Net Forest conversion	Food Household Consumption	IPPU	Manure applied to Soils	Manure Management	Fires in humid tropical forests	On-farm energy use
14.7237	0.0557	205.6077	-2388.8030	0.0000	79.0851	209.9778	260.1431	319.1763	0.0	3008.982252
14.7237	0.0557	209.4971	-2388.8030	0.0000	80.4885	217.0388	268.6292	342.3079	0.0	3008.982252
14.7237	0.0557	196.5341	-2388.8030	0.0000	80.7692	222.1156	264.7898	349.1224	0.0	3008.982252
14.7237	0.0557	230.8175	-2388.8030	0.0000	85.0678	201.2057	261.7221	352.2947	0.0	3008.982252
14.7237	0.0557	242.0494	-2388.8030	0.0000	88.8058	182.2905	267.6219	367.6784	0.0	3008.982252
...	...	...	...	...	...	...	...	...	...	...
1190.0089	232.5068	70.9451	76500.2982	10662.4408	251.2681	858.9820	96.1332	282.5994	0.0	417.315000
1431.1407	131.1324	108.6262	76500.2982	10662.4408	203.1236	889.4250	81.2314	255.5900	0.0	398.164400
1557.5830	221.6222	109.9835	76500.2982	10662.4408	211.1539	966.2650	81.0712	257.2735	0.0	465.773500
1591.6049	171.0262	45.4574	76500.2982	10662.4408	228.6381	945.9420	85.7211	267.5224	0.0	444.233500
481.9027	48.4197	108.3022	76500.2982	10662.4408	213.9211	940.4200	85.3143	266.7316	0.0	444.233500

ภาพประกอบที่ 13 หลังการเติมค่าว่าง

จะเห็นได้ว่าชุดตัวเลขได้ถูกเติมโดยเป็นชุดค่าเฉลี่ยจากแต่ละคอลัมล์และพร้อมใช้งานในขั้นตอนต่อไป

3.4 การแปลงข้อมูล

ทำการตรวจสอบข้อมูลเพื่อหา Attribute เพื่อนำไปเป็นตัวแปรทำนายผล และทำการแปลงข้อมูล Attribute นั้น ซึ่งเป็นตัวหนังสือที่เป็นข้อความ (categorical) ให้กลายเป็นตัวเลขจากนั้นทำการเลือก Label ซึ่งเป็นตัวแปรตามมาเป็นค่าที่ใช้ทำนายได้ ดังภาพประกอบที่ 15 และ 16



ภาพประกอบที่ 14 Attribute ที่จะทำการแปลง

ภาพประกอบที่ 15 Attribute ที่จะใช้เป็น Label

	Area	Year	Savanna fires	Forest fires	Crop Residues	Rice Cultivation	D o
0	Afghanistan	1990	14.7237	0.0557	205.6077	686.00	
1	Afghanistan	1991	14.7237	0.0557	209.4971	678.16	
2	Afghanistan	1992	14.7237	0.0557	196.5341	686.00	
3	Afghanistan	1993	14.7237	0.0557	230.8175	686.00	
4	Afghanistan	1994	14.7237	0.0557	242.0494	705.60	
5	Afghanistan	1995	14.7237	0.0557	243.8152	666.40	
6	Afghanistan	1996	38.9302	0.2014	249.0364	686.00	
7	Afghanistan	1997	30.9378	0.1193	276.2940	705.60	
8	Afghanistan	1998	64.1411	0.3263	287.4346	705.60	
9	Afghanistan	1999	46.1683	0.0895	247.4980	548.80	

0 rows × 31 columns

ภาพประกอบที่ 16 ตัวอย่างตารางก่อนแปลงข้อมูล

	Area	Year	Savanna fires	Forest fires	Crop Residues	Rice Cultivation
0	0	1990	14.7237	0.0557	205.6077	686.00
1	0	1991	14.7237	0.0557	209.4971	678.16
2	0	1992	14.7237	0.0557	196.5341	686.00
3	0	1993	14.7237	0.0557	230.8175	686.00
4	0	1994	14.7237	0.0557	242.0494	705.60
5	0	1995	14.7237	0.0557	243.8152	666.40
6	0	1996	38.9302	0.2014	249.0364	686.00
7	0	1997	30.9378	0.1193	276.2940	705.60
8	0	1998	64.1411	0.3263	287.4346	705.60
9	0	1999	46.1683	0.0895	247.4980	548.80

ภาพประกอบที่ 17 ตัวอย่างตารางหลังแปลงข้อมูล

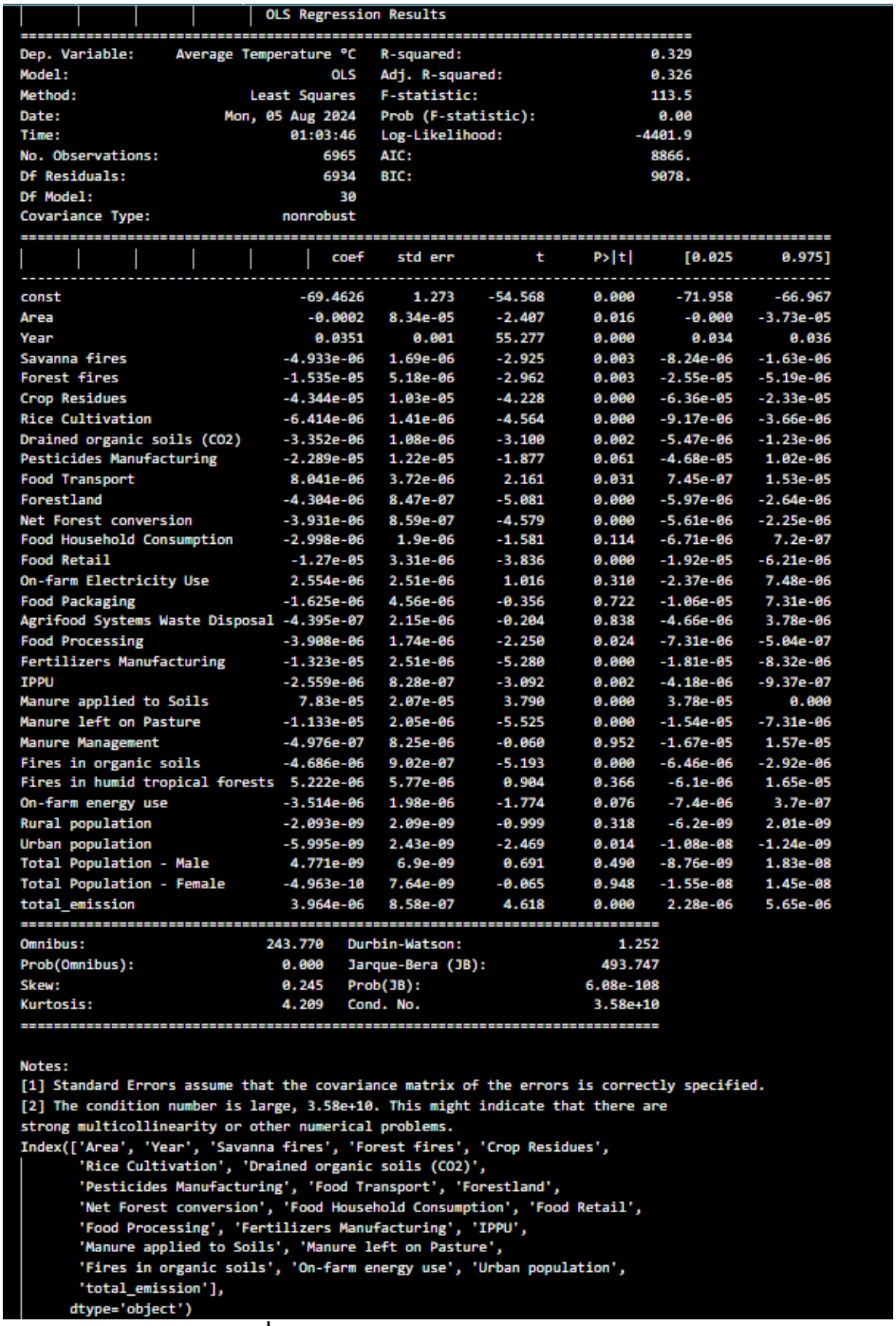
3.5 Backward-Elimination

Backward Elimination เป็นเทคนิคหนึ่งในการเลือกตัวแปร (feature selection) ที่ใช้ในการสร้างโมเดลเชิงสถิติ โดยเฉพาะอย่างยิ่งในการสร้างโมเดลการถดถอย (regression model) กระบวนการนี้ช่วยลด

จำนวนตัวแปรที่ไม่จำเป็นหรือไม่มีนัยสำคัญทางสถิติ ซึ่งจะช่วยให้โมเดลของมีความเรียบง่ายและทำงานได้มีประสิทธิภาพมากขึ้นโดยทำการคัดเลือกตัวแปรอิสระออกทีละ 1 ตัว จากสมการถอยหลังที่ประกอบด้วยตัวแปรอิสระ k ตัว โดยจะเริ่มพิจารณาตัวแปรอิสระที่มีความสัมพันธ์กับ Y น้อยที่สุดแล้วนำตัวแปรอิสระดังกล่าวมาทดสอบนัยสำคัญ ถ้าการทดสอบพบว่าไม่มีนัยสำคัญ แสดงว่าตัวแปรอิสระตัวนั้นจะถูกถอดออก และทำการคัดเลือกตัวแปรอิสระตัวที่ 2 ต่อไป แต่ในกรณีที่ทดสอบแล้วพบว่ามีนัยสำคัญจะหยุดทดสอบและสรุปผลว่า สมการถอยหลังประกอบด้วยตัวแปรอิสระทั้ง k ตัว ซึ่งขั้นตอนการวิเคราะห์ดังแสดงการทำงานดังนี้

```python
y = df['Average Temperature °C']
X = df.drop(['Average Temperature °C'], axis=1)

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
print(model.summary())

pmax = 1
while (len(X.columns) > 0):
    pmax = max(model.pvalues)
    if pmax > 0.05:
        feature_with_p_max = model.pvalues.idxmax()
        X = X.drop([feature_with_p_max], axis=1)
        model = sm.OLS(y, X).fit()
    else:
        break

selected_features = X.columns
selected_features = selected_features.drop('const')
print(selected_features)
```

ภาพประกอบที่ 18 การทำ backward-elimination โดยใช้ python



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Average Temperature °C   R-squared:                       0.329
Model:                                OLS   Adj. R-squared:                  0.326
Method:                     Least Squares   F-statistic:                     113.5
Date:                    Mon, 05 Aug 2024   Prob (F-statistic):               0.00
Time:                            01:03:46   Log-Likelihood:                -4401.9
No. Observations:                    6965   AIC:                             8866.
Df Residuals:                        6934   BIC:                             9078.
Df Model:                              30
Covariance Type:                nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          -69.4626      1.273    -54.568      0.000     -71.958     -66.967
Area                            -0.0002   8.34e-05     -2.407      0.016      -0.000   -3.73e-05
Year                             0.0351      0.001     55.277      0.000       0.034       0.036
Savanna fires                -4.933e-06   1.69e-06     -2.925      0.003   -8.24e-06   -1.63e-06
Forest fires                 -1.535e-05   5.18e-06     -2.962      0.003   -2.55e-05   -5.19e-06
Crop Residues                -4.344e-05   1.03e-05     -4.228      0.000   -6.36e-05   -2.33e-05
Rice Cultivation             -6.414e-06   1.41e-06     -4.564      0.000   -9.17e-06   -3.66e-06
Drained organic soils (CO2)  -3.352e-06   1.08e-06     -3.100      0.002   -5.47e-06   -1.23e-06
Pesticides Manufacturing     -2.289e-05   1.22e-05     -1.877      0.061   -4.68e-05    1.02e-06
Food Transport                8.041e-06   3.72e-06      2.161      0.031    7.45e-07    1.53e-05
Forestland                   -4.304e-06   8.47e-07     -5.081      0.000   -5.97e-06   -2.64e-06
Net Forest conversion        -3.931e-06   8.59e-07     -4.579      0.000   -5.61e-06   -2.25e-06
Food Household Consumption   -2.998e-06    1.9e-06     -1.581      0.114   -6.71e-06    7.2e-07
Food Retail                   -1.27e-05   3.31e-06     -3.836      0.000   -1.92e-05   -6.21e-06
On-farm Electricity Use       2.554e-06   2.51e-06      1.016      0.310   -2.37e-06    7.48e-06
Food Packaging               -1.625e-06   4.56e-06     -0.356      0.722   -1.06e-05    7.31e-06
Agrifood Systems Waste Disposal -4.395e-07   2.15e-06     -0.204      0.838   -4.66e-06    3.78e-06
Food Processing              -3.908e-06   1.74e-06     -2.250      0.024   -7.31e-06   -5.04e-07
Fertilizers Manufacturing    -1.323e-05   2.51e-06     -5.280      0.000   -1.81e-05   -8.32e-06
IPPU                         -2.559e-06   8.28e-07     -3.092      0.002   -4.18e-06   -9.37e-07
Manure applied to Soils       7.83e-05   2.07e-05      3.790      0.000    3.78e-05       0.000
Manure left on Pasture       -1.133e-05   2.05e-06     -5.525      0.000   -1.54e-05   -7.31e-06
Manure Management            -4.976e-07   8.25e-06     -0.060      0.952   -1.67e-05    1.57e-05
Fires in organic soils       -4.686e-06   9.02e-07     -5.193      0.000   -6.46e-06   -2.92e-06
Fires in humid tropical forests  5.222e-06   5.77e-06      0.904      0.366    -6.1e-06    1.65e-05
On-farm energy use           -3.514e-06   1.98e-06     -1.774      0.076    -7.4e-06     3.7e-07
Rural population             -2.093e-09   2.09e-09     -0.999      0.318    -6.2e-09    2.01e-09
Urban population             -5.995e-09   2.43e-09     -2.469      0.014   -1.08e-08   -1.24e-09
Total Population - Male       4.771e-09    6.9e-09      0.691      0.490   -8.76e-09    1.83e-08
Total Population - Female    -4.963e-10   7.64e-09     -0.065      0.948   -1.55e-08    1.45e-08
total_emission                3.964e-06   8.58e-07      4.618      0.000    2.28e-06    5.65e-06
==============================================================================
Omnibus:                      243.770   Durbin-Watson:                   1.252
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              493.747
Skew:                           0.245   Prob(JB):                    6.08e-108
Kurtosis:                       4.209   Cond. No.                     3.58e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.58e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
Index(['Area', 'Year', 'Savanna fires', 'Forest fires', 'Crop Residues',
       'Rice Cultivation', 'Drained organic soils (CO2)',
       'Pesticides Manufacturing', 'Food Transport', 'Forestland',
       'Net Forest conversion', 'Food Household Consumption', 'Food Retail',
       'Food Processing', 'Fertilizers Manufacturing', 'IPPU',
       'Manure applied to Soils', 'Manure left on Pasture',
       'Fires in organic soils', 'On-farm energy use', 'Urban population',
       'total_emission'],
      dtype='object')
```

ภาพประกอบที่ 19 ผลลัพท์ของการทำ backward-elimination

```
print("Selected Features:", selected_features)
✓ 0.0s                                                    Python

Selected Features: Index(['Area', 'Year', 'Savanna fires', 'Forest fires', 'Crop Residues',
       'Rice Cultivation', 'Drained organic soils (CO2)',
       'Pesticides Manufacturing', 'Food Transport', 'Forestland',
       'Net Forest conversion', 'Food Household Consumption', 'Food Retail',
       'Food Processing', 'Fertilizers Manufacturing', 'IPPU',
       'Manure applied to Soils', 'Manure left on Pasture',
       'Fires in organic soils', 'On-farm energy use', 'Urban population',
       'total_emission'],
      dtype='object')


print("Selected Features:", total_selected_features)
✓ 0.0s                                                    Python

Selected Features: 22
```

ภาพประกอบที่ 20 features จากการทำ backward-elimination

3.6 Cross validation

Cross validation คือการ แบ่งข้อมูลไป Train ในงานวิจัยนี้ได้ Setting ค่าแบ่งข้อมูลเป็น 10 folds เพื่อแบ่งข้อมูลใช้สำหรับ Train และ Test โดยสลับให้ข้อมูลแต่ละส่วนนำมาเป็นชุด Test เพื่อทดสอบประสิทธิภาพ โดยงานวิจัยนี้จะทำการทดสอบทั้งสี่โมเดล คือ Decision Tree, Random Forest และ Linear Regression ทั้งหมด 10 ครั้งเพื่อวัดประสิทธิภาพของทั้งสี่โมเดล

```
dt = DecisionTreeRegressor()
rf = RandomForestRegressor()
lr = LinearRegression()


kf = KFold(n_splits=10, shuffle=True, random_state=42)
```

ภาพประกอบที่ 21 Cross validation

3.7 การสร้าง Model ในการจำแนก

จะใช้ IDE (VS Code) เพื่อใช้สร้างตัวแบบในการวิเคราะห์ข้อมูลเพื่อทำนายการปล่อยคาร์บอน โดยใช้เทคนิค Decision Tree, Random Forest, Linear Regression และ Vote Ensemble แล้วนำผลพยากรณ์มาเปรียบเทียบหาค่าความถูกต้องโดยใช้ผลที่มีความเชื่อถือมากที่สุด

1. ขั้นตอนการสร้าง Model Decision tree, Random Forest และ Linear regression

เป็นเทคนิคที่นำข้อมูลแต่ละโนด (Node) ของแอททริบิวท์ (Attribute) มาทำการตัดสินใจ จากนั้นจะแสดงข้อมูลออกมาเป็นกิ่ง (Branch) และแสดงค่าออกมาเป็นใบ (Leaf) โดยใช้ information Gain มาหาความสัมพันธ์ในแต่ละโนดและทำให้ต้นไม้การตัดสินใจมีความซับซ้อนไม่มาก การสร้างโมเดล Decision Tree, Random Forest และ Linear regression โดยใช้ Python นั้นทำได้โดยใช้ไลบรารี scikit-learn (sklearn) ซึ่งเป็นไลบรารีที่นิยมใช้ในงาน Machine Learning ในที่นี้จะอธิบายขั้นตอนตั้งแต่การเตรียมข้อมูล การสร้างโมเดล การฝึกสอนโมเดล ไปจนถึงการประเมินผล

1.1 เตรียมข้อมูล

ในขั้นตอนนี้เราจะโหลดข้อมูลที่ใช้ในการฝึกสอนโมเดล โดยที่ X คือการนำ feature ที่ผ่านการทำ Backward-elimination มาแล้ว และ y คือ Label ที่จะใช้ในนการทำนาย

```
X = df[selected_features]
y = df['Average Temperature °C']
```

ภาพประกอบที่ 22 การแบ่งข้อมูล

1.2 สร้างและฝึกสอนโมเดล Decision Tree, Random Forest และ Linear regression

ในขั้นตอนนี้เราจะสร้างโมเดล และฝึกสอนโมเดลด้วยข้อมูลที่แบ่งไว้โดยการใช้ K-fold โดยที่

n_splits=10 คือการแบ่งข้อมูลเป็น 10 ส่วน (หรือ 10 folds) ซึ่งหมายความว่าแต่ละครั้งของการ cross-validation จะมี 1 fold ที่ใช้สำหรับการทดสอบ และอีก 9 fold ที่เหลือใช้สำหรับการฝึกสอน (training) กระบวนการนี้จะถูกทำซ้ำทั้งหมด 10

ครั้ง โดยแต่ละครั้งจะใช้ fold ที่ต่างกันในการทดสอบ จนครบทุก fold shuffle=True คือการตั้งค่า shuffle=True หมายความว่าข้อมูลจะถูกสุ่มก่อนที่จะแบ่งเป็น folds การสุ่มข้อมูลช่วยให้แน่ใจว่าข้อมูลแต่ละ fold มีความหลากหลายและไม่เกิดการแบ่งข้อมูลที่ไม่สมดุล เช่น การที่ข้อมูลที่อยู่ติดกันมีลักษณะคล้ายกันมากเกินไป

random_state=42 ใช้ในการตั้ง seed สำหรับการสุ่มหมายเลข ซึ่งทำให้การสุ่มข้อมูลเป็นแบบ deterministic (สามารถทำซ้ำได้) หมายเลข seed สามารถเป็นค่าตัวเลขใดก็ได้ เช่น 42 ในที่นี้เป็นค่า seed ที่มักใช้เพื่อให้การสุ่มเป็นแบบ reproducible (สามารถได้ผลลัพธ์เดิมเมื่อทำซ้ำ) การใช้ random_state จะช่วยให้ทุกครั้งที่รันคำสั่งนี้ ผลลัพธ์ของการสุ่มจะเหมือนเดิม ทำให้การทดสอบโมเดลมีความเสถียร

```
kf = KFold(n_splits=10, shuffle=True, random_state=42)
```

ภาพประกอบที่ 23 K-fold

2 ทำนายผลและประเมินผลโมเดล

หลังจากฝึกสอนโมเดลแล้ว เราจะใช้ข้อมูลชุดทดสอบในการทำนายผลและประเมินประสิทธิภาพของโมเดล

2.1 การทำ Cross-Validation สำหรับ MSE

mse_scores = cross_val_score(model, X, y, cv=kf, scoring=mse_scorer) โดยที่ cross_val_score: ฟังก์ชันนี้ทำ cross-validation สำหรับโมเดลที่กำหนดmodel: โมเดลปัจจุบันในลูป, X: ฟีเจอร์ของข้อมูล, y: เป้าหมาย (target), cv=kf: ใช้การแบ่งข้อมูลแบบ KFold ที่เราตั้งไว้ก่อนหน้านี้, scoring=mse_scorer: ใช้ MSE (Mean Squared Error) เป็นตัวชี้วัดประสิทธิภาพ

2.2 คำนวณ RMSE

rmse_scores = np.sqrt(mse_scores) โดยใช้ np.sqrt เพื่อคำนวณรากที่สองของค่า MSE สำหรับแต่ละ fold ซึ่งจะได้ค่า RMSE (Root Mean Squared Error)

2.3 การทำ Cross-Validation สำหรับ MAE

mae_scores=cross_val_score(model,X,y, cv=kf,scoring=mae_scorer) โดยใช้การทำ cross-validation เช่นเดียวกับ MSE แต่เปลี่ยนตัวชี้วัดเป็น MAE (Mean Absolute Error)

2.4 การทำ Cross-Validation สำหรับ R^2:

r2_scores = cross_val_score(model, X, y, cv=kf, scoring=r2_scorer) โดยใช้การทำ cross-validation เช่นเดียวกับ MSE แต่เปลี่ยนตัวชี้วัดเป็น R^2 (R-squared score)

2.5 คำนวณค่าเฉลี่ยและพิมพ์ผลลัพธ์:

model.__class__.__name__: ชื่อของคลาสโมเดล เช่น DecisionTreeRegressor

np.mean(mse_scores): ค่าเฉลี่ยของ MSE จากการ cross-validation

np.mean(rmse_scores): ค่าเฉลี่ยของ RMSE จากการคำนวณค่า RMSE

np.mean(mae_scores): ค่าเฉลี่ยของ MAE จากการ cross-validation

np.mean(r2_scores): ค่าเฉลี่ยของ R^2 จากการ cross-validation

```
models = [dt, rf, lr]
for model in models:
    mse_scores = cross_val_score(model, X, y, cv=kf, scoring=mse_scorer)
    rmse_scores = np.sqrt(mse_scores)
    mae_scores = cross_val_score(model, X, y, cv=kf, scoring=mae_scorer)
    r2_scores = cross_val_score(model, X, y, cv=kf, scoring=r2_scorer)

    print(f'{model.__class__.__name__} Mean MSE: {np.mean(mse_scores)}, Mean RMSE: \
      {np.mean(rmse_scores)}, Mean MAE: {np.mean(mae_scores)}, Mean R^2: {np.mean(r2_scores)}')
    print("\n")
```

ภาพประกอบที่ 24 การทำนายและประเมินผลโมเดล

3.8 สมการที่ใช้ในการวัดประสิทธิภาพ

1. Mean Absolute Error (MAE)

$$MAE = {}^{1}/_{n} * \sum|prediction - actual| \quad (1)$$

MAE คือค่าเฉลี่ยของความคลาดเคลื่อนสัมบูรณ์ระหว่างค่าที่คาดการณ์ (prediction) และค่าจริง (actual) เพื่อบอกให้เราทราบว่าค่าคาดการณ์เฉลี่ยแตกต่างจากค่าจริงมากเพียงใดโดยไม่สนใจทิศทาง (บวกหรือลบ) ดังสมการที่ (1) โดยค่าของ MAE อยู่ในหน่วยเดียวกับข้อมูลจริง และค่า MAE ที่น้อยกว่าแสดงให้เห็นว่าการคาดการณ์นั้นแม่นยำมาก โดยที่

- 1/n คือการหารด้วยจำนวนตัวอย่างทั้งหมด เพื่อหาค่าเฉลี่ย
- **Σ** คือการรวมผลลัพธ์ของทุกตัวอย่าง
- |prediction - actual| คือค่าสัมบูรณ์ของความแตกต่างระหว่างค่าทำนายและค่าจริง

2. Mean Squared Error (MSE)

$$MSE = {}^{1}/_{n} * \sum(prediction - actual)^2 \qquad (2)$$

MSE เป็นค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสองระหว่างค่าคาดการณ์และค่าจริง เป็นการยกกำลังสองตัดเครื่องหมายลบ ทำให้เน้นความแตกต่างขนาดใหญ่ (เพราะค่าความคลาดเคลื่อนที่ใหญ่จะถูกยกกำลังสอง)ดังสมการที่ (2) โดยค่าที่น้อยกว่าของ MSE บ่งชี้ว่าการคาดการณ์มีความแม่นยำมาก โดยที่

- การยกกำลังสอง (prediction - actual)² ทำให้ค่าความผิดพลาดที่มากมีน้ำหนักมากขึ้น
- MSE มีหน่วยเป็นกำลังสองของหน่วยดั้งเดิม

3. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{{}^{1}/_{n} * \sum(prediction - actual)^2} \qquad (3)$$

RMSE คือค่ารากที่สองของ MSE RMSE อยู่ในหน่วยเดียวกับข้อมูลจริง ทำให้เราสามารถเข้าใจได้ง่ายว่าการคาดการณ์มีความคลาดเคลื่อนเฉลี่ยเท่าใดRMSE ที่น้อยกว่าหมายถึงความแม่นยำในการคาดการณ์ที่ดีกว่า และเน้นความคลาดเคลื่อนขนาดใหญ่เช่นเดียวกับ MSE ดังสมการที่ (3) โดยที่

- การใช้รากที่สอง (√) ทำให้หน่วยกลับมาเป็นหน่วยเดียวกับข้อมูลดั้งเดิม
- RMSE ยังคงคุณสมบัติของ MSE ในการให้น้ำหนักกับค่าผิดพลาดขนาดใหญ่มากกว่า

4. R2

$$R2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (4)$$

R2 คือค่าสัดส่วนของความผันแปรของข้อมูลจริงที่อธิบายได้โดยโมเดลการคาดการณ์ ดังสมการที่ (4) โดยที่

- SSres คือผลรวมของค่าความคลาดเคลื่อนกำลังสองระหว่างค่าคาดการณ์กับค่าจริง (Residual Sum of Squares)
- SStot คือผลรวมของค่าความคลาดเคลื่อนกำลังสองของค่าจริงกับค่าเฉลี่ยของข้อมูลจริง (Total Sum of Squares)

ค่า R^2 ที่ใกล้ 1 หมายถึงโมเดลสามารถอธิบายความผันแปรของข้อมูลได้ดี แต่ถ้า R^2 ใกล้ 0 แสดงว่าโมเดลไม่สามารถอธิบายความผันแปรของข้อมูลได้ดีเท่าไร

บทที่ 4
ผลทดลองและการอภิปราย

บทนี้นำเสนอผลการทดลองที่เกิดขึ้นจากการสร้างแบบจำลองทำนายค่าเฉลี่ยอุณหภูมิจากชุดข้อมูล โดยพิจารณาการใช้โมเดลสามประเภท ได้แก่ Decision Tree, Random Forest, และ Linear Regression ในการวิเคราะห์ผลการทำนาย จะมีการวัดประสิทธิภาพโดยใช้ดัชนีชี้วัดที่สำคัญ เช่น Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) และค่า R^2 รวมถึงการอภิปรายเปรียบเทียบประสิทธิภาพของแต่ละโมเดล

4.1 ผลการทดลอง

การทดลองนี้ใช้ข้อมูลจากตัวอย่างที่มีทั้งหมด 6965 ระเบียน ประกอบด้วย 31 แอตทริบิวต์ ข้อมูลนี้ผ่านการเตรียมพร้อมด้วยการทำความสะอาด การจัดกลุ่มข้อมูล และการแปลงข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปใช้กับโมเดลทำนายได้ จากนั้นได้ทำการทดสอบโดยใช้วิธี 10-Fold Cross Validation ซึ่งเป็นวิธีแบ่งข้อมูลออกเป็น 10 ส่วนและทำการทดสอบโมเดลในแต่ละรอบ โดยใช้ข้อมูลส่วนหนึ่งเป็นชุดทดสอบและส่วนที่เหลือเป็นชุดฝึก เทคนิคที่ใช้ทำนายในการทดสอบครั้งนี้ประกอบด้วย

1. **Decision Tree**: โมเดลต้นไม้ตัดสินใจ เป็นโมเดลที่สร้างขึ้นโดยการแบ่งข้อมูลออกเป็นโหนดต่างๆ เพื่อหาความสัมพันธ์ในข้อมูล
2. **Random Forest**: โมเดลป่าแห่งการทำนาย เป็นการสร้างหลายๆ Decision Tree และนำผลลัพธ์จากต้นไม้ต่างๆ มาโหวตให้ได้ผลลัพธ์ที่แม่นยำที่สุด
3. **Linear Regression**: การถดถอยเชิงเส้น เป็นวิธีการคำนวณโดยใช้เส้นตรงที่เชื่อมโยงระหว่างตัวแปรอิสระและตัวแปรตาม

4.2 ค่าความคลาดเคลื่อนที่ได้รับจากแต่ละโมเดล

จากการทดสอบโดยใช้โมเดลทั้งสาม สามารถเปรียบเทียบค่าความคลาดเคลื่อน (Error) และประสิทธิภาพของโมเดลแต่ละโมเดลได้ดังตารางต่อไปนี้

ตารางที่ 4 การวัดประสิทธิภาพ

Model	การวัดประสิทธิภาพ			
	MSE	RMSE	MAE	R^2

Decision Tree	0.2321	0.4813	0.3505	0.2503
Random Forest	**0.1263**	**0.3550**	**0.2586**	**0.5900**
Linear Regression	0.2091	0.4570	0.3448	0.3220

จากตารางแสดงผลการทดลองสามารถสรุปได้ว่า โมเดล Random Forest มีประสิทธิภาพดีที่สุดในการทำนายค่าเฉลี่ยอุณหภูมิ เนื่องจากมีค่าความคลาดเคลื่อนต่ำที่สุด ทั้งค่า MSE (0.1263), RMSE (0.3550) และ MAE (0.2586) ซึ่งบ่งบอกว่าโมเดลสามารถทำนายค่าได้ใกล้เคียงกับค่าจริงมากกว่าทุกโมเดลที่ทดสอบ อีกทั้งค่า R^2 ของ Random Forest ที่สูงถึง (0.5900) สะท้อนถึงความสามารถในการอธิบายความผันแปรของข้อมูลได้อย่างแม่นยำกว่าโมเดลอื่นๆ

ในทางกลับกัน โมเดล Decision Tree มีค่า MSE สูงสุด (0.2321) และ R^2 ที่ต่ำสุด (0.2503) ซึ่งแสดงให้เห็นว่ามีความคลาดเคลื่อนในการทำนายมากที่สุด และมีความแม่นยำน้อยกว่าโมเดลอื่นๆ โมเดลนี้อาจมีปัญหาเรื่อง overfitting ทำให้เกิดความไม่เสถียรเมื่อทำการทดสอบกับข้อมูลชุดใหม่

Linear Regression มีผลการทำนายอยู่ในระดับปานกลาง ค่า MSE (0.2091) และ R^2 (0.3220) แสดงถึงความสามารถในการทำนายที่ดีขึ้นกว่า Decision Tree แต่ยังไม่เทียบเท่า Random Forest แม้ว่าการถดถอยเชิงเส้นจะเป็นวิธีที่ง่ายและเหมาะสมในกรณีที่ความสัมพันธ์ระหว่างตัวแปรเป็นเชิงเส้น แต่ในข้อมูลนี้ Random Forest กลับให้ผลลัพธ์ที่แม่นยำกว่า เนื่องจากข้อมูลอาจมีความซับซ้อนมากกว่าเส้นตรงธรรมดา

จากผลการทดลอง สามารถสรุปได้ว่า Random Forest เหมาะสมที่สุดสำหรับการทำนายข้อมูลชุดนี้ เนื่องจากสามารถลดค่าความคลาดเคลื่อนและทำนายได้แม่นยำกว่าโมเดลอื่นๆ

บทที่ 5
สรุปผลอภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองในการทำนายค่าเฉลี่ยอุณหภูมิจากชุดข้อมูลโดยใช้โมเดล Decision Tree, Random Forest และ Linear Regression พร้อมกับการเปรียบเทียบประสิทธิภาพของโมเดลแต่ละประเภท โดยใช้ดัชนีชี้วัด Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) และค่า R^2 ในการวิเคราะห์ผลการทำนาย ซึ่งผลการศึกษาสามารถสรุปได้ดังนี้

5.1 ผลการศึกษา

จากการทดลองทำนายค่าเฉลี่ยอุณหภูมิด้วยโมเดลทั้งสาม พบว่า Random Forest ให้ผลลัพธ์ที่ดีที่สุด มีค่า MSE ต่ำและค่า R^2 สูงสุด แสดงถึงความแม่นยำในการทำนายและความสามารถในการอธิบายความผันแปรของข้อมูลได้ดีที่สุด โมเดล Linear Regression มีความแม่นยำในระดับปานกลาง เหมาะสำหรับข้อมูลที่มีความสัมพันธ์เชิงเส้นตรง ในขณะที่ Decision Tree มีความแม่นยำน้อยที่สุดและมีความเสี่ยงต่อการเกิด Overfitting แม้จะเข้าใจและตีความโครงสร้างข้อมูลได้ง่ายก็ตาม สรุปได้ว่า Random Forest เป็นโมเดลที่เหมาะสมที่สุดสำหรับการทำนายข้อมูลในชุดนี้

5.2 ข้อสรุปจากการวิจัย

จากผลการวิเคราะห์ พบว่าโมเดล Random Forest เหมาะสมที่สุดสำหรับการทำนายค่าเฉลี่ยอุณหภูมิจากชุดข้อมูลที่มีความซับซ้อน เนื่องจากสามารถลดความคลาดเคลื่อนในการทำนายและให้ผลลัพธ์ที่แม่นยำที่สุด ในขณะที่ Linear Regression เหมาะสำหรับการวิเคราะห์ข้อมูลที่มีความสัมพันธ์เชิงเส้นตรงและง่ายต่อการใช้งาน ส่วน Decision Tree เหมาะสำหรับการวิเคราะห์ข้อมูลเบื้องต้นที่ต้องการการตีความที่ง่ายและตรงไปตรงมา

5.3 ข้อเสนอแนะสำหรับการวิจัยในอนาคต

1. การศึกษาเพิ่มเติมอาจเกี่ยวกับการใช้ โมเดลการทำนายอื่นๆ เช่น Support Vector Machine (SVM) หรือ Neural Networks เพื่อเปรียบเทียบประสิทธิภาพกับโมเดลที่ใช้ในงานวิจัยนี้

2. การใช้ Feature Engineering เพื่อปรับปรุงประสิทธิภาพของโมเดล เช่น การเลือกแอตทริบิวต์ที่สำคัญในการทำนายและการปรับแต่งพารามิเตอร์ของโมเดล Random Forest

3. การนำข้อมูลเพิ่มเติมที่มีความหลากหลายมาทดสอบโมเดลเพื่อประเมินความสามารถในการปรับตัวของโมเดลในการทำนายข้อมูลประเภทต่างๆ

5.4 ข้อจำกัดของการวิจัย

1. ข้อจำกัดด้านข้อมูล เนื่องจากข้อมูลที่ใช้มีขนาดและประเภทจำกัด อาจทำให้ผลการทำนายไม่สามารถสะท้อนสภาพแวดล้อมหรือข้อมูลจริงในบางกรณีได้ทั้งหมด

2. การใช้พารามิเตอร์มาตรฐาน ในการตั้งค่าโมเดลทำให้อาจมีความแม่นยำที่ต่ำกว่าที่เป็นไปได้หากทำการปรับแต่งพารามิเตอร์ให้เหมาะสมมากขึ้น

การวิจัยครั้งนี้ได้แสดงให้เห็นถึงความสำคัญของการเลือกโมเดลทำนายที่เหมาะสมกับลักษณะของข้อมูลและประสิทธิภาพของ Random Forest ในการจัดการกับข้อมูลที่มีความซับซ้อน ทั้งนี้ ข้อเสนอแนะในการศึกษาต่อสามารถช่วยให้มีการปรับปรุงและพัฒนาโมเดลทำนายได้ดียิ่งขึ้นในอนาคต

5.5 ข้อเสนอแนะ

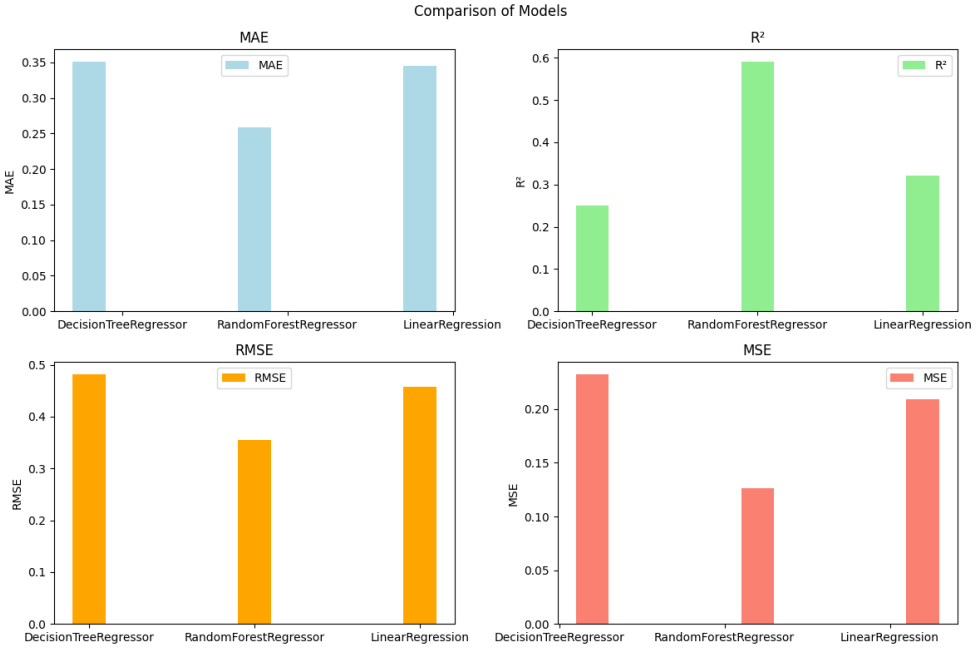
จากการวิเคราะห์และเปรียบเทียบโมเดลทั้งสาม สามารถเสนอแนวทางการใช้งานโมเดลต่างๆ ได้ดังนี้

1. สำหรับข้อมูลที่มีความซับซ้อนสูงและต้องการความแม่นยำสูง แนะนำให้ใช้ Random Forest เนื่องจากสามารถลดความคลาดเคลื่อนและให้ผลลัพธ์ที่แม่นยำได้ดีกว่าโมเดลอื่น

2. ในกรณีที่ข้อมูลมีความสัมพันธ์เชิงเส้นชัดเจน สามารถเลือกใช้ Linear Regression เพื่อความสะดวกและประหยัดทรัพยากรคอมพิวเตอร์ โมเดลนี้สามารถคำนวณได้รวดเร็วและให้ผลลัพธ์ที่ยอมรับได้

3. Decision Tree อาจเหมาะสมสำหรับการวิเคราะห์ข้อมูลที่ไม่ซับซ้อน หรือเมื่อต้องการการตีความผลลัพธ์ที่ง่าย เช่น การใช้ในการอธิบายกระบวนการตัดสินใจของข้อมูล เนื่องจากโครงสร้างของโมเดลสามารถอธิบายได้ชัดเจน

การปรับปรุงเพิ่มเติมอาจเกี่ยวข้องกับการทดสอบโมเดลอื่นๆ เช่น Support Vector Machine (SVM) หรือ Neural Networks เพื่อเปรียบเทียบความแม่นยำในการทำนาย นอกจากนี้ การปรับแต่งพารามิเตอร์ของโมเดล Random Forest หรือการใช้ Feature Engineering เพื่อปรับปรุงประสิทธิภาพของการทำนายก็เป็นแนวทางที่ควรพิจารณาในงานวิจัยต่อไป



ภาพประกอบที่ 25 กราฟแสดงการเปรียบเทียบการวัดประสิทธิภาพ

เอกสารอ้างอิง

amazon. (2023). "การทำเหมืองข้อมูลคืออะไร." Retrieved 30, 2024, from https://aws.amazon.com/th/what-is/data-mining/.

AnalytixLabs. (2023). "Random Forest Regression — How it Helps in Predictive Analytics?", from https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4.

Cesar de Lima Nogueira, S., S. H. Och, L. M. Moura, E. Domingues, L. d. S. Coelho and V. C. Mariani (2023). "Prediction of the NOx and CO2 emissions from an experimental dual fuel engine using optimized random forest combined with feature engineering." Energy **280**: 128066.

Chalathip. (2567). "รายงานฟาด 57 บริษัทตัวการโลกร้อน ปล่อยก๊าซคาร์บอนฯรวม 80%." Retrieved 28, 2567, from https://workpointtoday.com/sustainability-climate-change-co2-emissions/.

Chomchit, P. (2023). "คู่มือสำหรับ Linear Regression: แนวคิด และ การใช้งาน." Retrieved 28/7/2024, 2024.

co2meter. (2024). "Dangers of CO2: What You Need to Know." Retrieved 28, 2024, from https://www.co2meter.com/blogs/news/dangers-of-co2-what-you-need-to-know.

Cortez, P. and A. d. J. R. Morais (2007). A data mining approach to predict forest fires using meteorological data.

developers, s.-l. (2024). "Decision Tree Regression." from https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html.

Fujii, T., M. Sako, K. Ishihama, Y. Kohno, T. Makino, N. Yasuo and S. Kawauchi (2024). "Prediction of CO2 absorbing performance of amine aqueous solution using random forest models." Gas Science and Engineering **129**: 205417.

hitexts. (2022). "Model Evaluation มีตัววัดผลอะไรบ้างและทำไปเพื่ออะไร." from https://www.hitexts.com/data/model-evaluation/.

Lin, J., S. Lu, X. He and F. Wang (2021). "Analyzing the impact of three-dimensional building structure on CO2 emissions based on random forest regression." Energy **236**: 121502.

microsoft. (2024). "Power BI คืออะไร."
Power BI คืออะไร - Power BI | Microsoft Learn

Service, A. W. (2023). "การแสดงข้อมูลด้วยภาพคืออะไร." Retrieved 29, 2024, from https://aws.amazon.com/th/what-is/data-visualization/.

visualstudio. "Visual Studio Code." from https://code.visualstudio.com/docs/getstarted/userinterface.

Wikipedia. (2024). "คาร์บอนไดออกไซด์." Retrieved 28, 2024.
คาร์บอนไดออกไซด์ - วิกิพีเดีย

Zheng, P. (2021). "ทำความรู้จัก “Linear Regression” Algorithm ที่คนทำ Machine Learning ยังไงก็ต้องได้ใช้!" Retrieved 30, 2024.

ทัศนะนาคะจิตต์, ภ. (2562). "คาร์บอนไดออกไซด์ไม่มีประโยชน์เลยหรือ." Retrieved 28, 2567, from https://www.scimath.org/article-physics/item/9827-2019-02-21-08-51-20.