องค์ความรู้

Big Data และการประยุกต์ใช้ในการวิเคราะห์จัดทำนโยบายและแผน

สำนักงานเศรษฐกิจอุตสาหกรรม | OFFICE OF INDUSTRIAL ECONOMICS

การคาดการณ์สภาพเศรษฐกิจอุตสาหกรรม ด้วยข้อมูลที่มีขนาดใหญ่ (Big Data)

การพัฒนาเศรษฐกิจอุตสาหกรรมมีความสำคัญต่อการพัฒนาประเทศไทย จึงจำเป็นต้องเก็บรวบรวมข้อมูลที่สำคัญที่มีขนาดใหญ่ เช่น ทางเศรษฐกิจ และการผลผลิต เพื่อคาดการณ์แนวโน้มสภาพเศรษฐกิจอุตสาหกรรมในอนาคต

บิ๊กเดต้า (Big data) คือ ข้อมูลที่มีขนาดใหญ่ ซึ่งเป็นการรวบรวมข้อมูลทั้งข้อมูลที่มีโครงสร้าง (Structured) และข้อมูลที่ไม่มีโครงสร้าง (Unstructured) ที่มีอยู่ในองค์กรหรือหน่วยงานต่าง ๆ มาทำการประมวลวิเคราะห์ข้อมูลและนำไปใช้ประโยชน์ในการคาดการณ์สถานการณ์ จัดกลุ่มความสัมพันธ์ของข้อมูล หรือจำแนกประเภท ตามวัตถุประสงค์ขององค์กร หรือธุรกิจ สำหรับข้อมูลที่จะถูกเรียกว่าบิ๊กเดต้าหรือข้อมูลขนาดใหญ่มีองค์ประกอบสำคัญ 4 ข้อ ได้แก่

รูปที่ 1 คุณลักษณะของบิ๊กเดต้า

องค์ความรู้

Big Data และการประยุกต์ใช้ในการวิเคราะห์จัดทำนโยบายและแผน

สำนักงานเศรษฐกิจอุตสาหกรรม | OFFICE OF INDUSTRIAL ECONOMICS

กระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านข้อมูลขนาดใหญ่

พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือกันของ 3 บริษัท คือ DaimlerChrysler SPSS และ NCR กระบวนการทำงานนี้เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือเรียกย่อ ว่า “CRISP-DM” ซึ่งเป็นที่นิยมใช้อย่างแพร่หลายสำหรับการวิเคราะห์ข้อมูลขนาดใหญ่ มีขั้นตอนหรือกระบวนการโดยแบ่งเป็น 2 กลุ่มการทำงานได้แก่กลุ่มผู้เชี่ยวชาญทางด้านเศรษฐศาสตร์ และผู้เชี่ยวชาญทางด้านสถิติคอมพิวเตอร์ และคณิตศาสตร์ โดยจะใช้ฐานข้อมูลเดียวกันเป็นตัวเชื่อมทั้ง 2 กลุ่มเข้าด้วยกัน ซึ่งจะแบ่งขั้นตอนได้ทั้งหมด 7 ขั้นตอน ซึ่งจะอธิบายขั้นตอนการทำงานดังต่อไปนี้

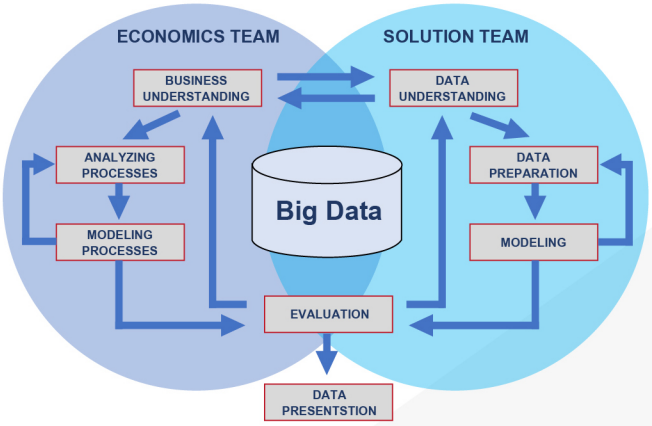
1 Business Understanding

เป็นกระบวนการในการกำหนดขอบเขตที่เกี่ยวข้องกับธุรกิจหรือสิ่งที่สนใจ โดยกำหนดวัตถุประสงค์และเป้าหมายของการทำเหมืองข้อมูล ซึ่งในขั้นตอนนี้ต้องอาศัยกลุ่มนักเศรษฐศาสตร์และผู้ที่ปฏิบัติงานเกี่ยวกับธุรกิจนั้น ๆ

2 Data Understanding

เป็นขั้นตอนของการเก็บรวบรวมข้อมูลต่าง ๆ ที่ถูกคัดเลือกจากแหล่งข้อมูลเพื่อนำมาใช้ในการวิเคราะห์โดยจะต้องคำนึงถึงลักษณะของข้อมูล ความเหมาะสมของข้อมูล คุณภาพของข้อมูลที่จะนำมาใช้ในการวิเคราะห์สำหรับขั้นตอนนี้ต้องใช้ความสามารถของกลุ่มนักสถิติและคอมพิวเตอร์ เพื่อใช้สำหรับการเก็บรวบรวมข้อมูลจากหลายแหล่งข้อมูล การตรวจสอบคุณภาพของข้อมูลและการคัดเลือกตัวแปรที่เหมาะสม

การคัดเลือกตัวแปรที่เหมาะสมนั้นสามารถใช้หลักการทางสถิติเพื่อใช้ในการคัดเลือกตัวแปรที่มีความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรผลลัพธ์ด้วยการทดสอบค่าสหสัมพันธ์ของเพียรสัน (Pearson Correlation Coefficient) โดยแสดงสมการทางคณิตศาสตร์ที่ใช้ในการคำนวน และเกณฑ์การตัดสินใจ ดังต่อไปนี้



รูปที่ 2 กระบวนการมาตรฐานอุตสาหกรรมสำหรับการทำเหมืองข้อมูล

สมการทางคณิตศาสตร์ที่ใช้ในการคำนวน

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

r_{xy} = ค่าสัมประสิทธิ์สหสัมพันธ์เพียรสัน

$\sum X$ = ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่หนึ่ง (X)

$\sum Y$ = ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่สอง (Y)

N = ขนาดของกลุ่มตัวอย่าง

เกณฑ์การพิจารณาค่าสัมประสิทธ์สหสัมพันธ์

ค่า r = 0.00 – 0.50 แสดงว่าตัวแปรมีความสัมพันธ์ต่ำ

ค่า r = 0.50 – 0.70 แสดงว่าตัวแปรมีความสัมพันธ์ปานกลาง

ค่า r = 0.70 – 1.00 แสดงว่าตัวแปรมีความสัมพันธ์สูง

องค์ความรู้

Big Data และการประยุกต์ใช้ในการวิเคราะห์จัดทำนโยบายและแผน

สำนักงานเศรษฐกิจอุตสาหกรรม | OFFICE OF INDUSTRIAL ECONOMICS

3 Analyzing Processes

เป็นขั้นตอนในการวิเคราะห์ของกลุ่มนักเศรษฐศาตร์เพื่อหาปัจจัยบางประเภทที่จะส่งผลกระทบต่อตัวแปรผลลัพธ์โดยใช้หลักการทางเศรษฐศาสตร์ โดยจะส่งตัวแปรที่มีความสำคัญทางเศรษฐศาสตร์ไปยังขั้นตอน Data Preparation

4 Data Preparation

เป็นขั้นตอนที่สำคัญในการเตรียมข้อมูลเพื่อที่จะใช้ในการวิเคราะห์ให้เหมาะสมและมีข้อผิดพลาดน้อยที่สุดเพื่อทำให้การวิเคราะห์ได้ผลลัพธ์ที่ใกล้เคียงกับความเป็นจริง โดยจะมีขั้นตอนคือ กำหนดตัวแปรที่จะใช้ในการวิเคราะห์ การตรวจสอบความผิดปกติของข้อมูลหรือการที่ข้อมูลในบางตัวแปรขาดหายไป สำหรับปัญหาที่พบมากที่สุดในขั้นตอนการเตรียมข้อมูลคือการพบว่าข้อมูลในบางตัวแปรนั้นขาดหายไปในบางช่วงเวลาซึ่งอาจจะส่งผลกระทบต่อการวิเคราะห์ โดยจะมีวิธีการในแก้ปัญหานี้หลากหลายวิธี เช่น การใช้ค่ามากที่สุดหรือน้อยที่สุด (Max-Min) ของตัวแปรนั้นเป็นค่าที่แทนค่าที่ขาดหาย การใช้ค่าเฉลี่ย ค่าเฉลี่ยถ่วงน้ำหนัก ค่ามัธยฐาน ค่าฐานนิยม หรือการใช้แบบจำลองการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple linear Regression Model) ในการแทนค่าที่ขาดหายไป

5 Modeling

เป็นขั้นตอนที่จะเลือกวิธีหรือกระบวนการในการวิเคราะห์ข้อมูลและทำการวิเคราะห์ เพื่อให้ได้ผลลัพธ์ที่พึงพอใจ ซึ่งเป็นกระบวนการที่ต้องใช้ความรู้ความสามารถในการคัดเลือกตัวแบบที่เหมาะสมกับปัจจัยที่ถูกคัดเลือกหรือผลลัพธ์ที่คาดหวังและตอบวัตถุประสงค์ สำหรับแบบจำลองที่นิยมใช้นั้นสามารถแบ่งออกเป็น 2 กลุ่มใหญ่ ๆ ได้แก่ 1. แบบจำลองการเรียนรู้ที่ต้องมีผู้ฝึกสอน (Supervised Learning Model) และ 2. แบบจำลองการเรียนรู้ที่ไม่ต้องมีผู้ฝึกสอน (Unsupervised Learning Model) โดยในครั้งนี้จะนำเสนอ แบบจำลองการเรียนรู้ที่ต้องมีผู้ฝึกสอนเนื่องจากนิยมใช้ในการคาดการณ์ระบบเศรษฐกิจต่าง ๆ ทั้งหมด 3 วิธี ได้แก่

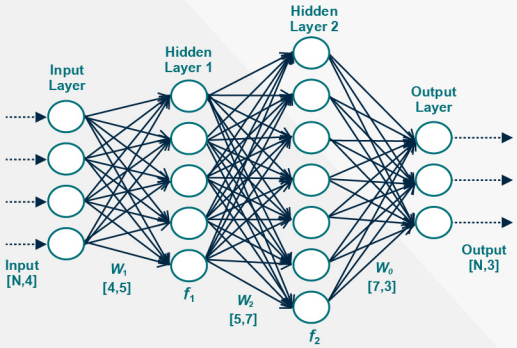
5.1 แบบจำลองการถดถอยเชิงเส้นแบบพหุ (Multiple Linear Regression) การวิเคราะห์การถดถอยเป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระ (x) กับตัวแปรตาม (y) จะเป็นการศึกษาความสัมพันธ์เชิงเส้นตรง ถ้าตัวแปรอิสระมีมากกว่าหนึ่งตัวซึ่งเป็นวิธีที่นิยมใช้ในการคาดการณ์ทางด้านเศรษฐกิจ แต่จะมีข้อเสียคือมีข้อบังคับทางสถิติที่ต้องคำนึงถึงหลายประเด็นจึงทำให้มีความซับซ้อนในการวิเคราะห์ และมีสมการคือ

$$\hat{y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon_i$$

5.2 แบบจำลอง Exponential smoothing (Holt-winters Multiplicative) เป็นวิธีที่เหมาะสมกับการคาดการณ์ระยะสั้นและปานกลาง โดนวิธีนี้เป็นวิธีที่คำนึงถึงอิทธิพลของข้อมูลในอดีตที่แตกต่างกันไปตามช่วงเวลามีสมการคือ

$$y_{t+k} = (a + bk)c_{t+k}$$

5.3 แบบจำลอง Neural network เป็นแบบจำลองทางคณิตศาสตร์สำหรับการวิเคราะห์ข้อมูลต่าง ๆ โดยเลียนแบบกระบวนการการทำงานของสมองของมนุษย์ โดยจะมี 3 ชั้น ได้แก่ ชั้น Input คือชั้นของตัวแปรนำเข้า ชั้นที่ 2 จะถูกเรียกว่า Hidden layer ซึ่งเป็นชั้นที่มีความซับซ้อนโดยจะมีฟังก์ชั่นในการดำเนินงานเรียกว่า ซิกมอยส์ฟังก์ชั้น และชั้นสุดท้ายคือชั้นผลลัพธ์ และสามารถสอนให้แบบจำลองมีความแม่นยำมากยิ่งขึ้น เพราะฉะนั้นแบบจำลอง Neural network เป็นที่นิยมใช้ในการคาดการณ์ต่าง ๆ แต่มีข้อเสียคือ เนื่องจากกระบวนการภายในการวิเคราะห์นั้นมีความซับซ้อนอย่างมากจึงทำให้การหาปัจจัยที่ส่งผลต่อการคาดการณ์มีความซับซ้อนและเสียเวลาเป็นอย่างมาก และการสอนให้โมเดลแม่นยำมากก็จะทำให้เกิดปัญหาที่ถูกเรียกว่า Overfitting หรือ แบบจำลองไม่สามารถคาดการณ์ข้อมูลนำเข้าอื่นได้ในอนาคต โดยมีกระบวนการดังต่อไปนี้



รูปที่ 3 กระบวนการทำงานของ neural network

ที่มา: https://medium.com/coinmonks/the-artificial-neural-networks-handbook-part-1-f9ceb0e376b4

องค์ความรู้

Big Data และการประยุกต์ใช้ในการวิเคราะห์จัดทำนโยบายและแผน

สำนักงานเศรษฐกิจอุตสาหกรรม | OFFICE OF INDUSTRIAL ECONOMICS

6 Evaluation

เป็นขั้นตอนที่ใช้ในการประเมินผลลัพธ์จากกระบวนการในการทำเหมืองข้อมูล ที่ได้จากขั้นตอนที่ 5 ว่าผลลัพธ์ที่ได้มานั้นสามารถนำไปใช้งานได้จริงตามวัตถุประสงค์หรือเป้าหมายที่กำหนดไว้ได้หรือไม่ โดยจะใช้หลักทางสถิติ ในการวัดประสิทธิภาพของแบบจำลอง เช่น RMSE MAPE MAE เป็นต้น

7 Data Presentation

เป็นขั้นตอนสุดท้ายที่จะนำผลลัพธ์ทั้งหมดที่ได้จากการวิเคราะห์มานำเสนอในรูปแบบต่าง ๆ ทั้งเอกสารเผยแพร่ การนำเสนอผ่านเว็บไซต์ เพื่อนำความรู้ที่ได้จากกระบวนการต่าง ๆ มาใช้ประโยชน์ให้ได้มากที่สุด

ตัวอย่างการคาดการณ์ข้อมูลด้วยข้อมูลขนาดใหญ่

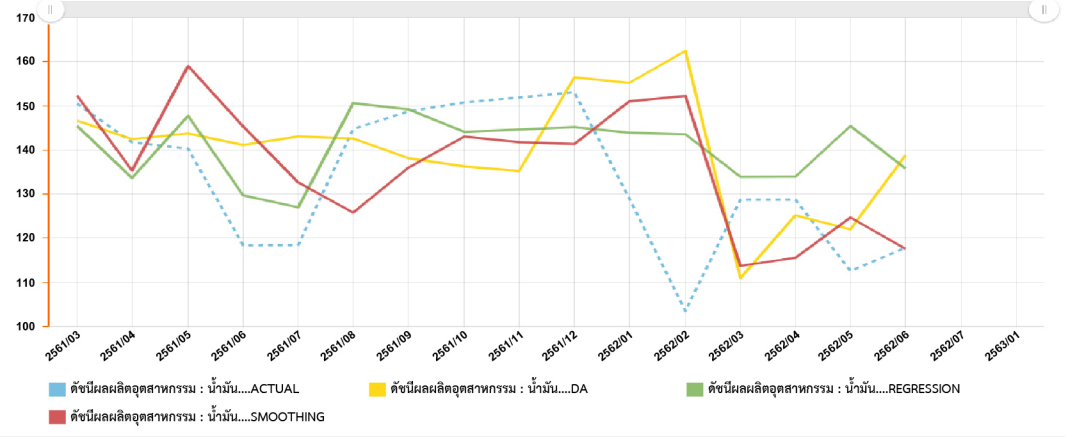
จากขั้นตอนของวิธีการ CRISP-DM ที่ได้นำเสนอทั้ง 7 ขั้นตอนสามารถนำมาประยุกต์ใช้กับการคาดการณ์ผลทางด้านเศรษฐกิจด้วยข้อมูลขนาดใหญ่ เช่น ผลดัชนีผลผลิตอุตสาหกรรม ซึ่งต้องมีการเก็บรวบรวมที่มีจำนวนมากและหลากหลาย เพื่อนำมาคัดเลือกตัวแปรและจัดเตรียมข้อมูลที่เหมาะสม ตลอดจนการสร้างแบบจำลองทางคณิตศาสตร์เพื่อคาดการณ์ จากรูปที่ 4 แสดงให้เห็นว่าดัชนีอุตสาหกรรมของน้ำมันปาล์มบริสุทธิ์มีแนวโน้มที่ลดลง

เอกสารอ้างอิง

[1] สุรพงศ์ เอื้อวัฒนมงคล, (2561), การทำเหมืองข้อมูล (ฉบับปรับปรุง), สำนักพิมพ์สถาบันบัณฑิตพัฒนบริหารศาสตร์.

[2] เรื่อง Bigdata คืออะไรhttps://www.salika.co/2019/03/12/big-data-introduction/

ลำดับ	ดัชนีผลผลิตอุตสาหกรรม : น้ำมันปาล์มบริสุทธิ์ ที่ใช้ในการคาดการณ์	หน่วย	ค่าที่ได้จากการคาดการณ์	MPE
1	ดัชนีผลผลิตอุตสาหกรรม : น้ำมันปาล์มบริสุทธิ์ DA	ดัชนี	138.78	1.69
2	ดัชนีผลผลิตอุตสาหกรรม : น้ำมันปาล์มบริสุทธิ์ REGRESSION	ดัชนี	135.79	10.44
3	ดัชนีผลผลิตอุตสาหกรรม : น้ำมันปาล์มบริสุทธิ์ SMOOTHING	ดัชนี	117.58	8.98



รูปที่ 4 ตัวอย่างการใช้เทคนิค CRISP-DM ในการคาดการณ์ผลทางด้านเศรษฐกิจด้วยข้อมูลขนาดใหญ่

ที่มา: www.piiu.oie.go.th

คณะทำงานจัดทำองค์ความรู้ Big Data และการประยุกต์ใช้ในการวิเคราะห์จัดทำนโยบายและแผน

๑. นางสาวเพียงใจ ไชยรังสินันท์	คณะทำงาน
๒. นางสาวภูริตา มณียม	คณะทำงาน
๓. นางสาววรรณพร บุณยรัตพันธุ์	คณะทำงาน
๔. นางสาวพัชราวดี คำรอด	คณะทำงาน
๕. นางสาวทิพจุฑา รวบยอด	คณะทำงาน
๖. นางสาวจันทิมา ยาเกิ้น	คณะทำงาน
๗. นางสาวปทิตตา เตชะศุภสิน	คณะทำงาน
๘. นางสาวพิมพ์ขวัญ ลิมปโสภา	คณะทำงาน
๙. นางสาววรนาถ ด่านผดุงทรัพย์	คณะทำงาน และเลขานุการ