โครงงานคอมพิวเตอร์ 1

การพัฒนาตัวแบบจำแนกประเภทข้อมูลสภาพทางเศรษฐกิจครัวเรือนด้วยเทคนิคเหมืองข้อมูล

นำเสนอเค้าโครงงานโดย

นางสาวปิยวรรณ เรือนธรรม
รหัสนักศึกษา 62102105106
อาจารย์ที่ปรึกษา อาจารย์ ดร.นิภาพร ชนะมาร

สาขาคอมพิวเตอร์ หลักสูตรวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลัยราชภัฏสกลนคร

เนื้อหาที่นำเสนอ

1. หลักการและเหตุผล
2. วัตถุประสงค์ของโครงงาน
3. ขอบเขตของโครงงาน
4. เอกสารและงานวิจัยที่เกี่ยวข้อง
5. แผนการดำเนินงาน
6. ประโยชน์ที่คาดว่าจะได้รับ
7. กรอบการดำเนินงาน
8. บรรณานุกรม

หลักการและเหตุผล

ปัญหาความยากจน

การประยุกต์ใช้เทคนิคเหมืองข้อมูล

การแก้ปัญหา

วัตถุประสงค์ของโครงงาน

เพื่อพัฒนาแบบจำแนกประเภทข้อมูลเศรษฐกิจครัวเรือนของคนในชุมชนด้วยการประยุกต์ใช้

- เทคนิคต้นไม้ตัดสินใจ (Decision Tree)
- เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks – ANN)
- เทคนิคนาอีฟเบย์ (Naive Bayes)

ขอบเขตของโครงงาน

ด้านข้อมูล

ข้อมูลประชากรจากภาคครัวเรือนเฉพาะครัวเรือนในเขตพื้นที่ชนบท ของจังหวัดสกลนคร ซึ่งมี 20 หมู่บ้าน 12 ตำบล 12 อำเภอ โดยช่วงเวลาที่ทำการเก็บรวบรวมข้อมูล คือ ปี พ.ศ.2563 –2564 และจากฐานข้อมูลสภาพทางเศรษฐกิจครัวเรือนเป้าหมายตามโครงการจ้างงาน ประชาชนที่ได้รับผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID -19) มีการเก็บข้อมูลจากฐานเศรษฐกิจชุมชนซึ่งมีการเก็บข้อมูลออกเป็น 10 ส่วน

- ประชากรที่ใช้ในการศึกษา ได้แก่ ครัวเรือนตำบลที่อยู่ในช่วงปี พ.ศ. 2561- 2563 ได้มาจากข้อมูล 12 ตำบล ซึ่งมีจำนวน 17,933 ครัวเรือน
- กลุ่มตัวอย่างที่ใช้ในการศึกษาและวิเคราะห์ข้อมูล ได้แก่ กลุ่มครัวเรือนบ้านในช่วงปี พ.ศ. 2561-2563 ได้มาจากการเลือกแบบเจาะจง (Purposive Sampling) จำนวน 3,233 ครัวเรือน

ขอบเขตของโครงงาน

ด้านเทคนิค

การศึกษาครั้งนี้ได้ประยุกต์ใช้เทคนิคทางด้านการทำเหมืองข้อมูล ดังนี้

1) เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

2) เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks – ANN)

3) เทคนิคนาอีฟเบย์ (Naive Bayes)

สำหรับพัฒนาหาตัวแบบที่เหมาะสมในการจำแนกประเภทข้อมูลสภาพทางเศรษฐกิจครัวเรือน

4) ใช้กระบวนการ CRIPS-DM (Cross Reference Industry Standard for Data Mining)

ในการวิเคราะห์ข้อมูลและสร้างตัวแบบโมเดล

ขอบเขตของโครงงาน

ด้านเครื่องมือในการพัฒนา

ซอฟต์แวร์

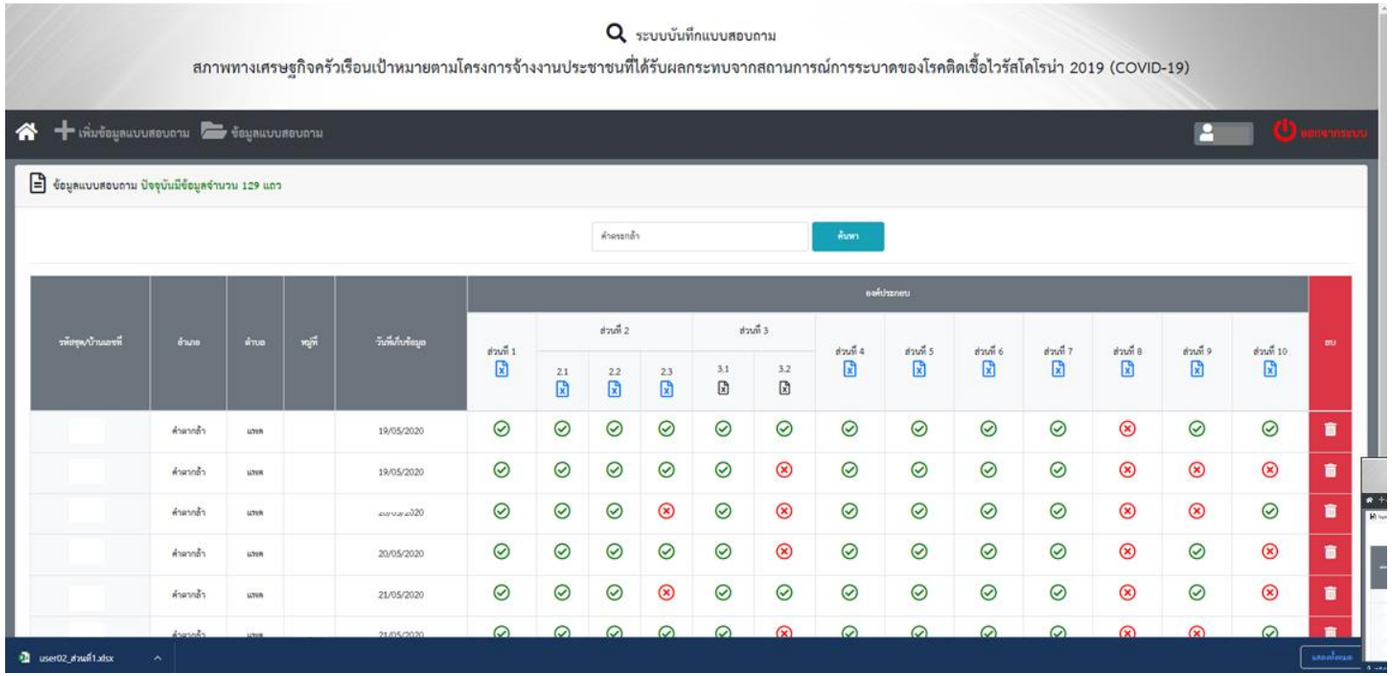
การศึกษาครั้งนี้ได้ทำการทดลองดำเนินการผ่านโปรแกรม RapidMiner Studio เวอร์ชัน 9.10 เป็นโปรแกรมที่ออกแบบมาสำหรับการวิเคราะห์ข้อมูล ของบริษัท RapidMiner คือซอฟต์แวร์ Data Science ใช้สำหรับการเตรียมข้อมูล การเรียนรู้เครื่อง การเรียนรู้ลึก การทำเหมืองข้อความ และการวิเคราะห์การทำนาย (Predictive Analysis) เป็นซอฟต์แวร์ที่ช่วยในการจัดส่งข้อมูล และลดข้อผิดพลาดจนแทบจะไม่จำเป็นต้องเขียนโค้ดเพิ่ม

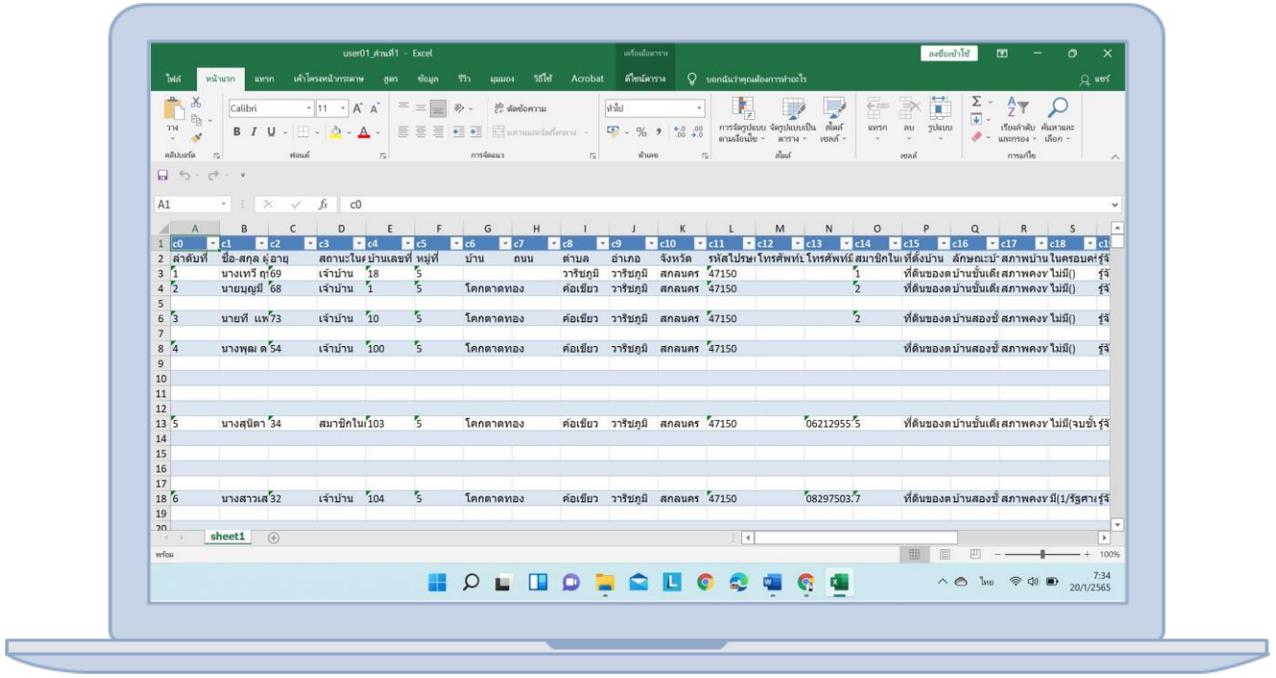
ฮาร์ดแวร์

โน๊ตบุ๊ค Lenovo

แผนการดำเนินงาน

1. กำหนดหัวข้อและนำเสนอหัวข้อ
2. ค้นหาปัญหา โอกาสและเป้าหมาย
3. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง
4. เสนอเค้าโครงโครงงาน
5. ศึกษาและวิเคราะห์ข้อมูล
6. ทำความเข้าใจข้อมูลและเตรียมข้อมูล
7. ดำเนินการพัฒนาโมเดล
8. ประเมินประสิทธิภาพการพัฒนาโมเดล
9. จัดทำเอกสารประกอบโครงงาน
10. นำเสนอโครงงานจบ





ประโยชน์ที่คาดว่าจะได้รับ

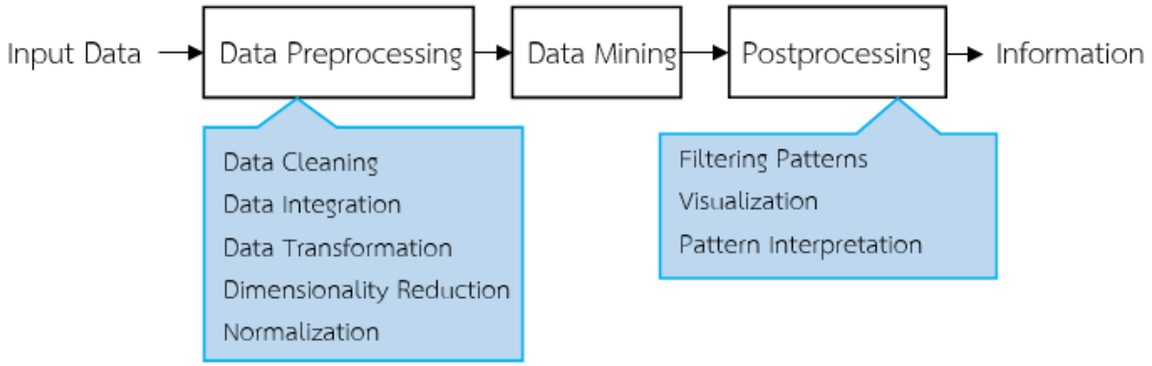
ได้พัฒนาตัวแบบโมเดล และทราบถึงประสิทธิภาพความถูกต้องแม่นยำในการจำแนกประเภทข้อมูลสภาพทางเศรษฐกิจครัวเรือนด้วยเทคนิคเหมืองข้อมูล เพื่อสนับสนุนหรือเป็นข้อมูลประกอบการตัดสินใจในการพัฒนาชุมชนท้องถิ่นสำหรับนักวิจัย

เอกสารและงานวิจัยที่เกี่ยวข้อง

การทำเหมืองข้อมูล (Data Mining)

คือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหา รูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคมต่างๆ

เป็นขั้นตอนหนึ่งของกระบวนการค้นพบองค์ความรู้ในฐานข้อมูลขนาดใหญ่ (Knowledge Discovery In Database: KDD) ที่นำข้อมูลดิบ (Raw Data) มาผ่านกระบวนการต่างๆ จนได้สารสนเทศ (Information) หรือองค์ความรู้ที่สามารถใช้ประโยชน์ได้



กระบวนการค้นหาความรู้ในฐานข้อมูลขนาดใหญ่

เอกสารและงานวิจัยที่เกี่ยวข้อง

การทำเหมืองข้อมูล (Data Mining) มี 2 ประเภท คือ

1. Supervised Learning การเรียนรู้แบบมีผู้สอน

การจำแนกประเภทข้อมูล (Data Classification)

- ต้นไม้ตัดสินใจ (Decision Tree)
- โครงข่ายประสาทเทียม (Artificial Neural Networks – ANN)
- นาอีฟเบย์ (Naive Bayes)

2. Unsupervised Learning การเรียนรู้แบบไม่มีผู้สอน

- กฎความสัมพันธ์ (Association Rule)
- การแบ่งกลุ่มข้อมูล (Clustering Algorithm)
- Time Series Algorithm

เอกสารและงานวิจัยที่เกี่ยวข้อง

การจำแนกประเภทข้อมูล (Data Classification)

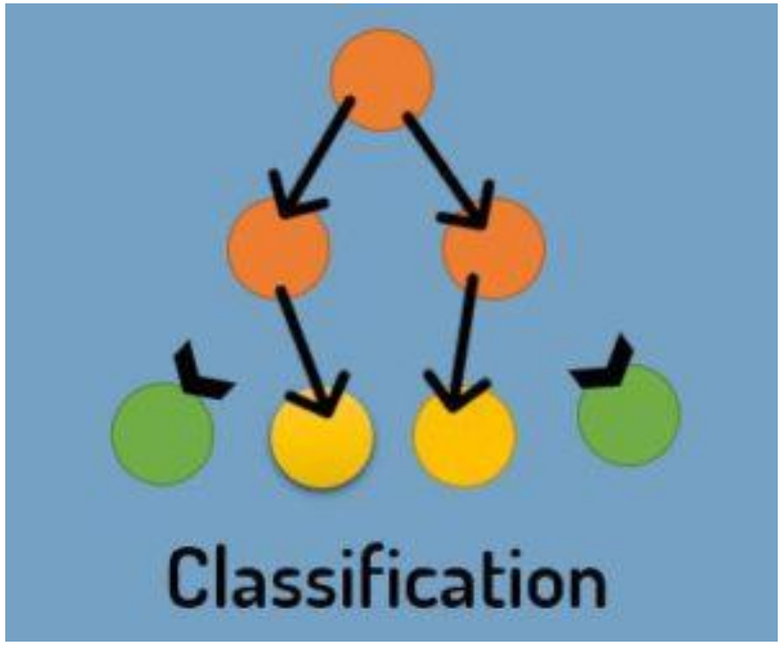
เป็นกระบวนการทำเหมืองข้อมูลชนิดหนึ่งที่มีการเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยมีขั้นตอนหลัก ๆ อยู่ 2 ขั้นตอน คือ

1. การสร้างแบบจำลอง (Model Construction)

   โดยเซตของตัวอย่างที่ใช้สร้างแบบจำลองจะเรียกว่าชุดข้อมูลสอน (Training Set) ซึ่งแต่ละตัวอย่างจะมีคุณลักษณะบอกค่าประเภทไว้ล่วงหน้า

2. การนำแบบจำลองที่ได้ไปใช้ (Model Usage)

   สำหรับการจำแนก ประเภท ตัวอย่างในอนาคตโดยจะต้องมีการประมาณค่าความแม่นยำ (Accuracy)

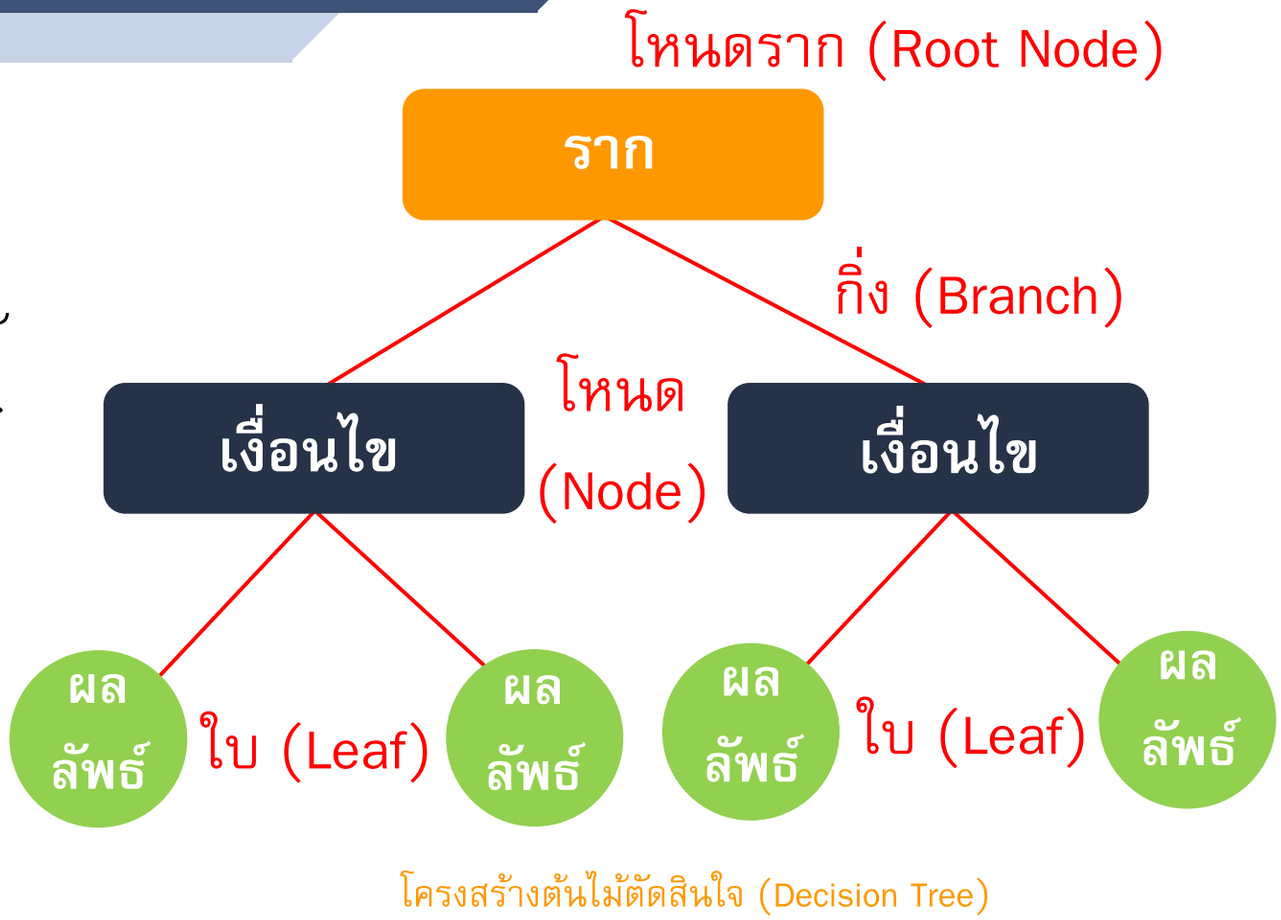


โครงสร้างการจำแนกประเภทข้อมูล (Classification)

เอกสารและงานวิจัยที่เกี่ยวข้อง

เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

เป็นหนึ่งในเทคนิคการทำเหมืองข้อมูลในรูปแบบวิธีการจัดหมวดหมู่ ที่รู้จักกันดีที่สุดโดยมักใช้ตรวจสอบข้อมูลและสร้างต้นไม้เพื่อการพยากรณ์ สำหรับโครงสร้างของต้นไม้ตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ทั่วไป โดยการแตกแขนงไปตามเงื่อนไข หรือเส้นทางของกิ่งไม้และข้อมูลที่คาดคะเนไว้ว่าจะเกิดขึ้น ซึ่งจะใช้กฎในรูปแบบ “ถ้า (เงื่อนไข) แล้ว (ผลลัพธ์)”(If-then Rule) มาประกอบการสร้างโครงสร้างต้นไม้ตัดสินใจ



โครงสร้างต้นไม้ตัดสินใจ (Decision Tree)

เอกสารและงานวิจัยที่เกี่ยวข้อง

เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks – ANN)

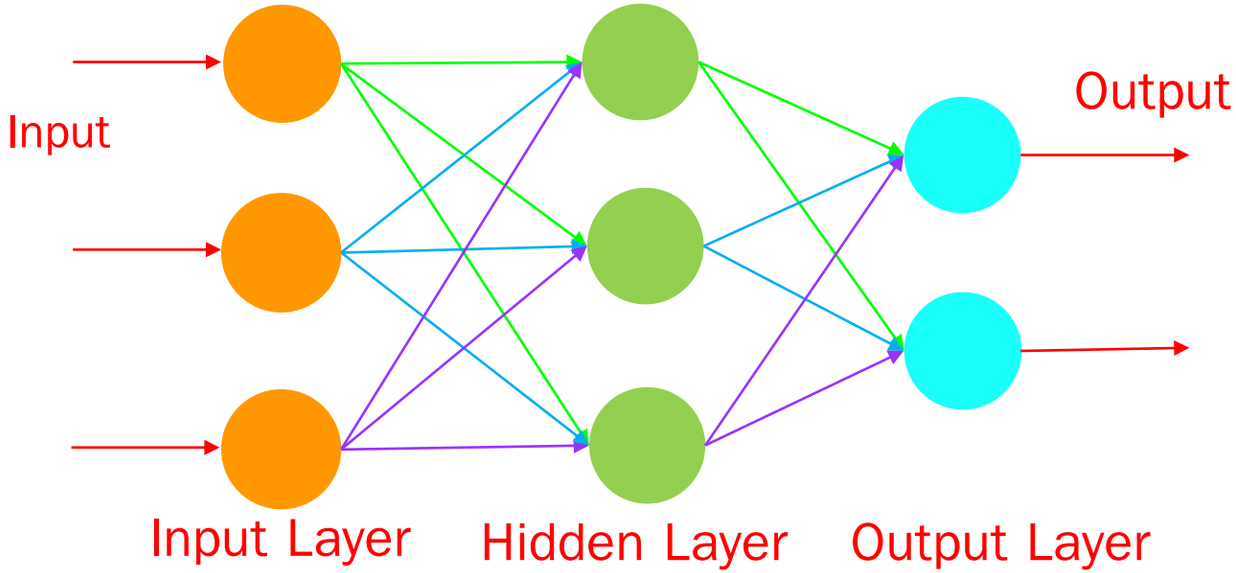
แบบจำลองทางคณิตศาสตร์ที่เลียนแบบกระบวนการทำงานของระบบประสาทในมนุษย์

- ชั้นรับข้อมูลเข้า

  หรืออินพุตเลเยอร์ (Input Layer)
- ชั้นซ่อน

  หรือฮิดเดนเลเยอร์ (Hidden Layer)
- ชั้นส่งข้อมูลออก

  หรือเอาต์พุตเลเยอร์ (Output Layer)



โครงสร้างโครงข่ายประสาทเทียม (Artificial Neural Networks – ANN)

เอกสารและงานวิจัยที่เกี่ยวข้อง

เทคนิคนาอีฟเบย์ (Naive Bayes)

เครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็นตามทฤษฎีของเบย์ (Bayes Theorem)ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อนเป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ หลักการของนาอีฟเบย์ใช้การคำนวณหาความน่าจะเป็นในการทำนายผลเป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

ตัวอย่างสมการทฤษฎีของเบย์

เอกสารและงานวิจัยที่เกี่ยวข้อง

ความแม่นยำ (Accuracy)

เกี่ยวข้องกับการวัดได้ใกล้เคียงกับค่าการตรวจมาตรฐานสูงสุด (Gold Standard) หรือค่าที่ตั้งใจจะวัดหรือค่าจริงหรือไม่ อีกนัยหนึ่งกล่าวได้ว่า ความแม่นยำ คือความถูกต้องของค่าที่วัดได้ เป็นความใกล้เคียงกับค่าจริงหรือใกล้เคียงกับค่าจากเครื่องมือมาตรฐาน

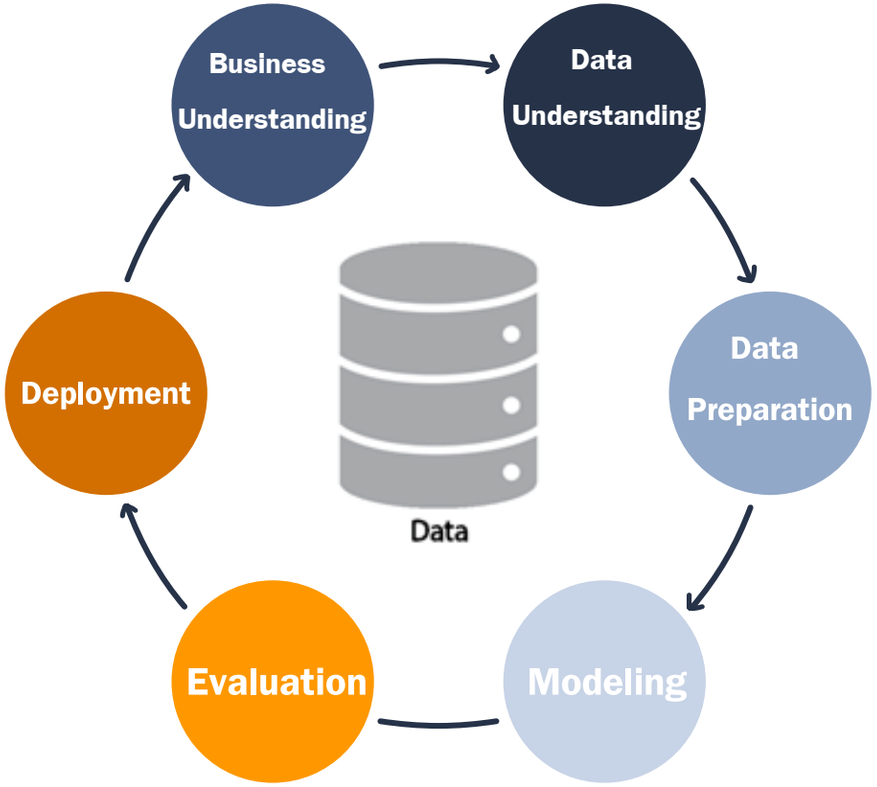
การวัดประสิทธิภาพโมเดลการจำแนกข้อมูลด้วย K-Fold Cross Validation

Cross-validation Test ใช้ในการทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัด ประสิทธิภาพด้วยวิธี Cross-validation นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน (มักจะแสดงด้วยค่า k) เช่น 5-fold cross-validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูล 1 ส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้ จนครบจำนวนที่แบ่งไว้

เอกสารและงานวิจัยที่เกี่ยวข้อง

CRIPS-DM (Cross Reference Industry Standard for Data Mining)

เรียกว่า แนวคิดกระบวนการมาตรฐานอุตสาหกรรม ประกอบด้วย 6 ขั้นตอนหลัก ดังนี้



CRISP-DM process model

1) การทำความเข้าใจโจทย์ (Business Understanding)
2) การทำความเข้าใจข้อมูล (Data Understanding)
3) การเตรียมข้อมูล (Data Preparation)
4) การสร้างตัวแบบ (Modeling)
5) การประเมินผล (Evaluation)
6) การใช้งาน (Deployment)

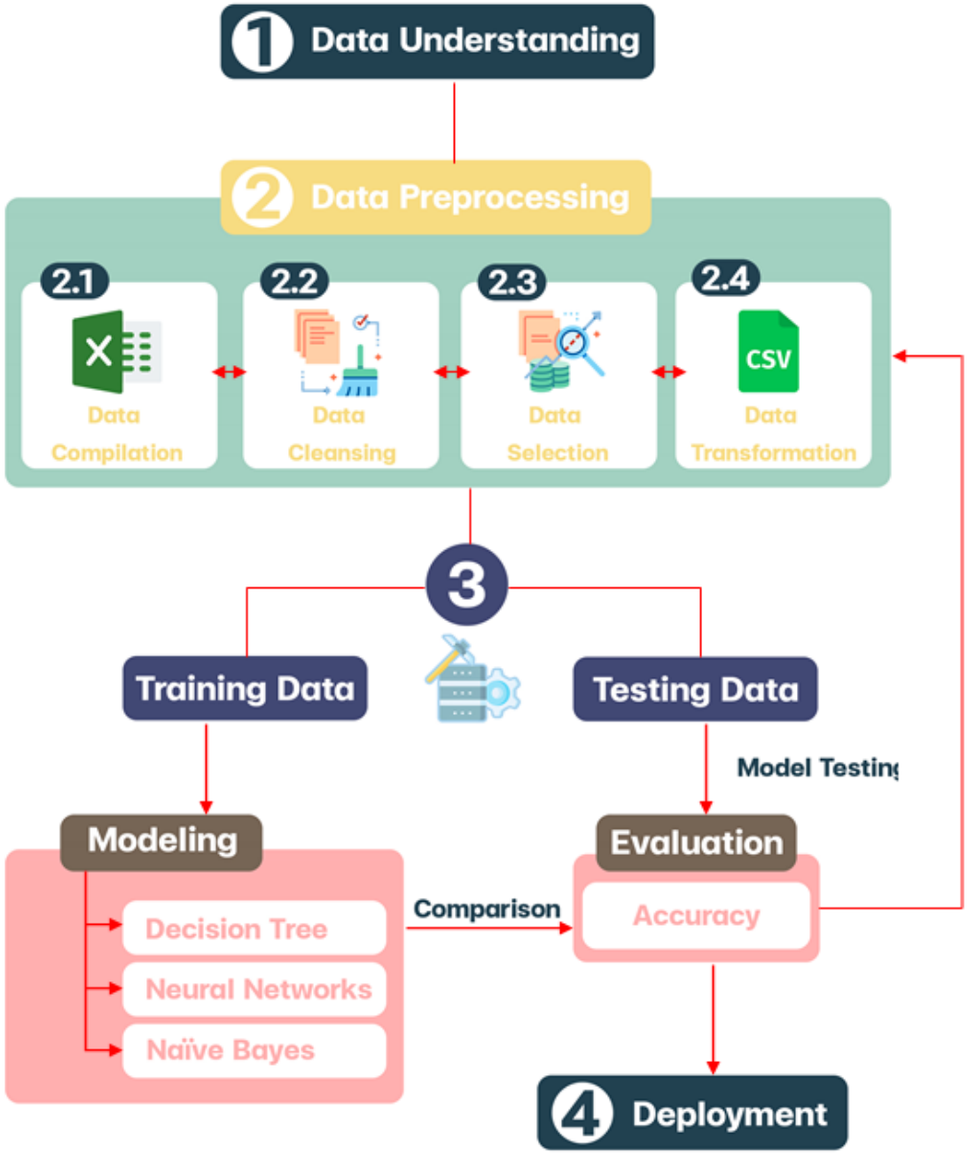
เอกสารและงานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้อง

เอกสารและงานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้อง (ต่อ)

กรอบการดำเนินงาน



บรรณานุกรม

THANKS!