รายงานสรุปการอบรม/สัมมนา/พัฒนาความรู้/ประชุมเชิงปฏิบัติการ/และเป็นวิทยากร
กองนโยบายและแผนการใช้ที่ดิน กรมพัฒนาที่ดิน

****

ส่วนที่ 1 ข้อมูลทั่วไป

ชื่อ นายดิเรก นามสกุล คงแพ
ตำแหน่ง นักวิเคราะห์นโยบายและแผนชำนาญการพิเศษ กลุ่ม/ฝ่าย กลุ่มวางแผนบริหารจัดการพื้นที่ชุ่มน้ำ
หลักสูตร/หัวข้อเรื่องอบรม/สัมมนา/พัฒนาความรู้ฯ
หลักสูตร การวิเคราะห์และการประยุกต์ใช้งานข้อมูลสำหรับภาครัฐ
สถานที่อบรม/สัมมนา/พัฒนาความรู้ฯ
ระบบการฝึกอบรมผ่านสื่ออิเล็กทรอนิกส์ (https://bit.ly/ai-dga-slid)
หน่วยงานที่จัดฝึกอบรม/สัมมนา/พัฒนาความรู้ฯ
สำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน) สพร. หรือ DGA
ตั้งแต่วันที่ 15 เดือน กรกฎาคม พ.ศ. 2564 ถึงวันที่ 16 เดือน กรกฎาคม พ.ศ. 2564
เพื่อ ☑ อบรม ☐ สัมมนา ☐ อื่นๆ ระบุ.....

ส่วนที่ 2 สิ่งที่ได้รับจากการอบรม/สัมมนา/พัฒนาความรู้

2.1 รายงานสรุปเนื้อหาสาระสำคัญในการอบรม/ สัมมนา/พัฒนาความรู้ฯ

การอบรมโครงการฯ นี้ การวิเคราะห์และการประยุกต์ใช้งานข้อมูลสำหรับภาครัฐเพื่อเตรียมความพร้อมการใช้งานปัญญาประดิษฐ์มีเนื้อหาการบรรยายและ Workshop เช่นการตัวอย่างการประยุกต์ใช้การวิเคราะห์ข้อมูลและเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence : AI) กระบวนการในการวิเคราะห์ข้อมูล เรียนรู้เครื่องมือในการวิเคราะห์ข้อมูล ด้วยโปรแกรมสำเร็จรูป RapidMiner Studio โดยมีวัตถุประสงค์เพื่อพัฒนาความรู้และความเข้าใจกับบุคลากรของภาครัฐในทุกระดับสำหรับการพัฒนาหน่วยงานสู่รัฐบาลดิจิทัล และเพื่อต่อยอดในการประยุกใช้ปัญญาปรดิษฐ์ ในการปรับปรุงเพิ่มประสิทธิภาพการทำงานของภาครัฐที่สามารถตอบโจทย์ความต้องการของประชาชน ท่ามกลางสถานการณ์ของประเทศที่เปลี่ยนแปลงไปอย่างรวดเร็ว

อธิบายตามบทเรียน แบ่งได้เป็น

1) อย่างการประยุกต์ใช้การวิเคราะห์ข้อมูลและเทคโนโลยีปัญญาประดิษฐ์ (AI) ตัวอย่างการประยุกต์ใช้ AI ใช้ในชีวิตประจำวัน เช่นระบบตรวจจับวัตถุ (Image Recognition) ระบบตรวจวัดอุณหภูมิ ระบบรู้จำเสียง (Speech Recognition) ลำโพงอัจฉริยะ และ ระบบ Chat Bot เป็นต้น เช่น หุ่นอัจฉริยะของค่ายแอลจี อิเล็กทรอนิกส์ หรือ LG ในสนามบินนานาชาติอินชอน เกาหลีใต้ ที่ให้บริการข้อมูลแก่ผู้โดยสารจำนวนราว 57 ล้านคนจากทั่วโลก ผ่านเสียงพูด รวมทั้งสามารถให้ช่วยเหลือนำพาผู้โดยสารที่หลงทางไปสู่จุดหมายที่ถูกต้องภายในสนามบิน หุ่นยนต์น้องแสนดี (SAN:DEE Delivery Robot) ในเครือบริษัท “แสนสิริ” หุ่นยนต์ตัวแรกที่นำมาใช้ในวงการอสังหาริมทรัพย์ โดยหน้าที่ของแสนดี คือ อำนวยความสะดวกบริการส่งพัสดุ จดหมาย ถึงหน้าห้องลูกบ้าน ยกระดับเรื่องความปลอดภัย โดยป้องกันไม่ให้คนแปลกหน้าเดินขึ้นไปส่งของโดยตรงให้ลูกบ้าน Chatbot ผู้ช่วยคอยตอบคำถามเบื้องต้นของลูกค้าได้ตลอดเวลา ทำให้การทำธุรกิจดำเนินไปอย่างต่อเนื่องไม่ขาดตอน Apple Siri, Google Now, Microsoft Cortana ผู้ช่วยที่สั่งการด้วยเสียง AI Smart Home ระบบบ้านอัจฉริยะจากเทคโนโลยี AI FinTech ด้านการออม เพื่อวิเคราะห์พฤติกรรมการใช้เงิน ในบัญชีกระแสรายวันที่ไม่มีดอกเบี้ย หรือดอกเบี้ยต่ำโดยใช้เทคโนโลยี AI Robo-advisor ผู้ช่วยด้าน

บริการจัดการด้านการลงทุนโดยอัตโนมัติเป็นต้น ปัญญาประดิษฐ์ (AI) เป็นเทคโนโลยีการสร้างความสามารถได้แก่ เครื่องจักรและคอมพิวเตอร์ ด้วยอัลกอริทึมและกลุ่มเครื่องมือทางสถิติ เพื่อสร้างซอฟต์แวร์ทรงปัญญาที่สามารถเลียนแบบความสามารถของมนุษย์ที่ซับซ้อนได้ เช่นจดจำ แยกแยะ ให้เหตุผล ตัดสินใจ คาดการณ์ สื่อสาร กับมนุษย์ เป็นต้น ในบางกรณีอาจไปถึงขั้นเรียนรู้ได้ด้วยตนเอง AI บ่งบอกไปถึงความสามารถทางปัญญาของเครื่องจักร (machine) โดยมาตรฐานของ AI ถูกวัดด้วยสติปัญญาของมนุษย์ คำนึงถึง ความมีเหตุผล, การพูดและการมองเห็น แต่มาตรฐานนี้ยังห่างไกลอยู่มากจากปัจจุบัน AI ทำงานโดยรวบรวมข้อมูลปริมาณมหาศาลด้วยความเร็ว ประมวลผลซ้ำๆ ผ่านขั้นตอนการประมวลผลที่ชาญฉลาด ด้วยซอฟต์แวร์ที่สามารถเรียนรู้จากรูปแบบและลักษณะของข้อมูลได้อย่างอัตโนมัติบนพื้นฐานทางทฤษฎี สามารถแบ่งประเภทของ AI ได้ 3 ประเภท (1) Artificial Narrow Intelligence (ANI) หรือ “ปัญญาประดิษฐ์แบบเบา (Weak AI)” (2) Artificial General Intelligence (AGI) หรือ “ปัญญาประดิษฐ์แบบเข้ม (Strong AI)” และ (3) Artificial Super Intelligence (ASI) หรือ “ปัญญาประดิษฐ์แบบทรงปัญญา ความสามารถของปัญญาประดิษฐ์ (AI) ที่นำมาใช้ (1) Machine Learning (การเรียนรู้ของเครื่องจักร) เป็นการทำให้เครื่องสามารถเรียนรู้ได้ด้วยตนเอง ในการสร้างแบบจำลองการวิเคราะห์แบบอัตโนมัติ โดยใช้วิธีการจากโครงข่ายประสาทเทียม สถิติ การวิจัยดำเนินการ (operations research) และหลักฟิสิกส์ในการค้นหาข้อมูลเชิงลึกที่ซ่อนอยู่ในข้อมูลโดยไม่จำเป็นต้องเขียนโปรแกรมในการค้นหา (2) Natural Language Processing (NLP) (การประมวลผลภาษาธรรมชาติ) เป็นเทคนิคทำให้เครื่องทำความเข้าใจภาษามนุษย์ คือความสามารถของคอมพิวเตอร์ในการวิเคราะห์ ทำความเข้าใจและสร้างภาษามนุษย์ ซึ่งรวมถึงคำพูดด้วย ขั้นถัดไปของ NLP คือการโต้ตอบด้วยภาษาธรรมชาติ ซึ่งช่วยให้มนุษย์สามารถสื่อสารกับคอมพิวเตอร์ได้โดยใช้ภาษาเพื่อดำเนินการงานต่างๆ (3) Expert System (การวิเคราะห์แบบผู้เชี่ยวชาญ) ให้เลียนแบบความสามารถในการตัดสินใจที่เชี่ยวชาญอย่างมนุษย์ (4) Computer Vision ให้เครื่องสามารถเข้าใจคุณลักษณะของภาพคล้ายคลึงกับความสามารถในการมองเห็นของมนุษย์ (5) Speech Recognition การรู้จำเสียงและคำพูดเป็นความสามารถในการระบุคำและวลีในการพูด (6) Planning (การวางแผน) ให้เครื่องสามารถตัดสินใจเลือกการดำเนินงานในการบรรลุเป้าหมายที่กำหนด (7) Robotics (หุ่นยนต์) พัฒนาเครื่องจักรให้มีรูปร่างและเคลื่อนไหวแตกต่างกันไปตามวัตถุประสงค์การใช้งาน ปัญญาประดิษฐ์ (AI) “ในอนาคตเทคโนโลยี AI จะเป็นอีกหนึ่ง ‘**เครื่องมือ**’ ที่ช่วยเพิ่มประสิทธิภาพการทำงานของคน เวลาที่เราได้ยินคนพูดถึง AI ในปัจจุบัน โดยส่วนมาก จะหมายถึง Machine Learning นั่นเอง มีแหล่งข้อมูลที่ใช้ในการวิเคราะห์คือ (1) ศูนย์กลางข้อมูลเปิดภาครัฐ (Open Government Data) URL: https://opendata.data.go.th และ (2) Kaggle เป็นที่รวบรวมการแข่งขัน (competition) ทางด้าน Data Science/Machine Learning URL: https://www.kaggle.com/datasets

2) กระบวนการในการวิเคราะห์ข้อมูล Machine Learning (ML) เรียนรู้จากสิ่งที่เราส่งเข้าไปกระตุ้น แล้วจดจำเอาไว้เป็นมันสมอง ส่งผลลัพธ์ออกมาเป็นตัวเลข หรือ code ที่ส่งต่อไปแสดงผล หรือให้เจ้าตัว AI นำไปแสดงการกระทำ Machine Learning เองสามารถเอาไปใช้งานได้หลายรูปแบบ ต้องอาศัยกลไกที่เป็นโปรแกรม หรือเรียกว่า Algorithm ที่มีหลากหลายแบบ ดังนั้น ML เป็นการสอนให้ระบบคอมพิวเตอร์ทำการเรียนรู้ได้ด้วยตนเองโดยการใช้ 'ข้อมูล' อาจจะทำความเข้าใจง่ายๆ ตามชื่อเลยก็คือ การสอน Algorithm ให้เรียนรู้ทำความเข้าใจและตัดสินใจได้ด้วยตัวเองจาก 'ข้อมูล' ที่ป้อนให้ การเรียนรู้ของ Machine นั้นเป็นไปใน 3 รูปแบบคือ (1) การเรียนรู้โดยมีผู้สอน (Supervised) คือการเรียนรู้ โดยมี data มาสอน ชัดๆ เลยก็คือ เด็กน้อยต้องไปสอบแยกแยะประเภทหมาแมว โดย Data Scientist จึงสร้าง Model ที่จะทำให้คอมพิวเตอร์รู้จักแยกแยะประเภทหมาแมว เช่น Model Regression, Support vector machine, Naive Bayes, Gradient boosting, Classification trees/random forest เป็นต้น (2) การเรียนรู้โดยไม่มีผู้สอน (Unsupervised) นั้นเครื่องจะเรียนรู้และทำนายผลได้จากการจำแนกและสร้างแพทเทิร์นของมันจากข้อมูลที่ได้รับเมื่อเครื่องสามารถทำนายผลลัพธ์จากชุดข้อมูลจำนวนมากได้มากเท่าไรเช่น Model K Nearest Neighbors, K Mean เป็นต้น (3) Reinforcement Learning ในบรรดา Machine Learning

ทั้งหมด Reinforcement Learning คือสิ่งที่ดูเป็น AI ที่แท้จริงที่สุดเพราะจะเรียนรู้และเปลี่ยนไปตามสิ่งแวดล้อม วิธีนี้เหมาะมากกับโจทย์บางประเภท คือการหากลยุทธ์ที่ทำให้ชนะเกม เช่นเดินออกจากเขาวงกต Model ที่ใช้กันบ่อยๆ คือ Markov Decision Processes (MDP), Q-learning เป็นต้น และแสดงความสามารถในการเรียนรู้เชิงลึก (Deep Learning) มากขึ้นไปด้วย

กระบวนการวิเคราะห์ข้อมูลด้วยเทคนิค CRISP-DM เป็นเทคนิคที่นิยมในการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining เป็นการวิเคราะห์ข้อมูลมาตรฐานซึ่งเป็นเหมือน blueprint ที่ใช้กันอย่างกว้างขวาง เช่นเดียวกันกับกระบวนการ ISOในโรงงานอุตสาหกรรม หรือกระบวนการ CMMI ซึ่งเป็นมาตรฐานในการพัฒนาซอฟต์แวร์ กระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านดาต้า ไมน์นิงนี้ พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือกันของ 3 บริษัท คือ DaimlerChrysler SPSS และNCR กระบวนการทำงานนี้เรียกว่า “ Cross-Industry Standard Process for Data Mining” หรือเรียกย่อว่า “ CRISP-DM” โดยในกระบวนการ CRISP-DM นี้จะประกอบด้วย 6 ขั้นตอน แต่ละขั้นตอนมีขั้นตอนที่ต่อเนื่องกันนั่นคือขั้นตอนถัดไปจะรอผลลัพธ์จากขั้นตอนก่อนหน้าที่เชื่อม ตัวอย่างเช่นเมื่อได้ผลลัพธ์จากขั้นตอนการเตรียมข้อมูล (Data Preparation) แล้วจะนำไปสร้างโมเดลจำแนกประเภทข้อมูลในขั้นModeling และหลังจากนั้นอาจจะย้อนกลับมาเปลี่ยนแปลงข้อมูลให้ถูกต้องมากขึ้นเพื่อหวังว่าจะโมเดลที่ให้ความถูกต้องมากขึ้นก็ได้ เป็นต้น ขั้นตอนในกระบวนการ CRISP-DM มีดังนี้

(1) Business Understanding เป็นขั้นตอนแรกในกระบวนการ CRISP-DM ซึ่งเน้นไปที่การเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิงพร้อมทั้งวางแผนในการดำเนินการคร่าวๆ ตัวอย่างการนำเทคนิคดาต้า ไมน์นิงไปใช้ในการวิเคราะห์ด้านต่างๆ มีดังนี้ เช่น การใช้เทคนิคดาต้า ไมน์นิงเพื่อพัฒนาคุณภาพชีวิต การประยุกต์ใช้โครงข่ายประสาทเทียมโดยใช้ตัวชี้วัดทางเทคนิค เพื่อการลงทุนในตลาดหลักทรัพย์แห่งประเทศไทย ระบบจำแนกและค้นคืนข้อมูลเว็บกระทู้ข่าว ด้วยโครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น การเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล โดยวิธีการเลือก ลักษณะสำคัญแบบพลวัตเพื่อเพิ่มประสิทธิภาพของอัลกอริทึม การจัดกลุ่มบนปริภูมิย่อย การพยากรณ์ความต้องการพลังงานไฟฟ้าสำหรับศูนย์จัดการ ความต้องการพลังงานไฟฟ้าของประเทศไทย Bagging Model with Cost Sensitive Analysis on Diabetes Data เป็นต้น

(2) Data Understanding ขั้นตอนนี้เริ่มจากการเก็บรวบรวมข้อมูล หลังจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์

(3) Data Preparation ขั้นตอนนี้เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (raw data) ให้กลายเป็นข้อมูลที่สามารถนำไป วิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (data cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป เป็นต้น โดยขั้นตอนนี้จะเป็นขั้นตอนที่ใช้เวลามากที่สุดของกระบวนการ CRISP-DM

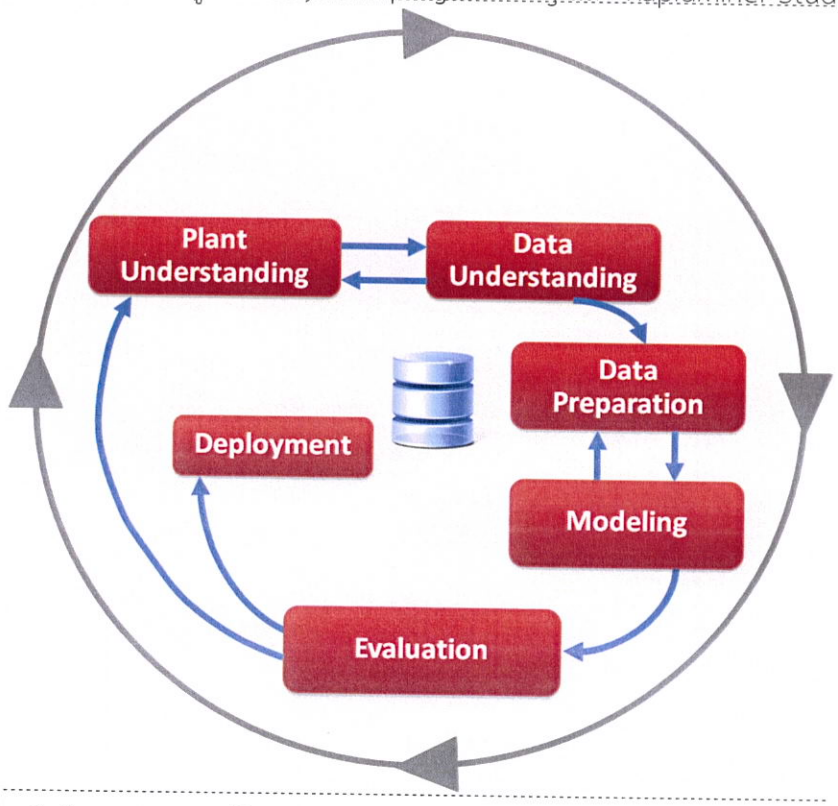
(4) Modeling ขั้นตอนนี้จะเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้า ไมน์นิง ที่ได้แนะนำไปแล้ว เช่น การจำแนกประเภทข้อมูล หรือ การแบ่งกลุ่มข้อมูล ซึ่งในขั้นตอนนี้หลายเทคนิคจะถูกนำมาใช้เพื่อให้ได้คำตอบที่ดีที่สุด ดังนั้นในบางครั้งอาจจะต้องมีการย้อนกลับไปที่ขั้นตอนที่ (3) Data Preparation เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย ตัวอย่างเทคนิคในการวิเคราะห์ข้อมูลต่างๆ เช่น (1) การแบ่งกลุ่มข้อมูล (Clustering) (2) การหากฎความสัมพันธ์ (Association Rules) และ (3) การจำแนกประเภทข้อมูล (Classification) ตัวอย่างเช่นเทคนิค Decision Tree เทคนิค Naive Bayes เทคนิค Neural Network และ เทคนิค Support Vector Machines (SVM)

(5) Evaluation ในขั้นตอนนี้เราจะได้ผลการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้า ไมน์นิงแล้วแต่ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานต่อไปก็จะต้องมีการวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรก หรือ มีความน่าเชื่อถือมากน้อยเพียงใด ซึ่งอาจจะย้อนกลับไปยังขั้นตอน

ก่อนหน้าเพื่อเปลี่ยนแปลงแก้ไขเพื่อให้ได้ผลลัพธ์ตามที่ต้องการได้ สำหรับการสร้างโมเดลด้วยเทคนิค Classification มีการทดสอบประสิทธิภาพของโมเดลอยู่ 3 แบบใหญ่ คือ Self-consitency test Split test Cross-validation test

(6) Deployment ในกระบวนการทำงานของ CRISP-DM นั้นไม่ได้หยุดเพียงแค่ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้าไมน์นิงเท่านั้น แม้ว่าผลลัพธ์ที่ได้จะแสดงถึงองค์ความรู้ที่มีประโยชน์ แต่จะต้องนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริงในองค์กรหรือบริษัท ตัวอย่างเช่น การสร้างรายงานเพื่อให้ผู้บริหารหรือนักการตลาดเข้าใจได้ง่ายและสามารถนำไปออกโปรโมชั่นได้ เป็นต้น

กระบวนการทั้ง 6 ขั้นตอนนี้ไม่ได้ดำเนินการทำอยู่เป็นประจำ แต่จะนำโมเดลที่ได้ไปใช้ในการวิเคราะห์ข้อมูลทางธุรกิจ จนกว่าสภาพแวดล้อมทางธุรกิจเปลี่ยนไปอย่างเห็นได้ชัด หรือโมเดลที่ได้จัดทำไว้เริ่มมีความแม่นยำน้อยลง ซึ่งนั่นหมายความว่าอาจจะมีตัวแปรใหม่ๆ หรือปัจจัยบางอย่างที่มีความสำคัญน้อยลงไป สำหรับผู้ที่สนใจศึกษาเพิ่มเติมได้ในหนังสือเกี่ยวกับ Data Mining และ Data Analytics ส่วนซอฟท์แวร์ที่ช่วยในการจัดการข้อมูล อาทิ Python programming และ RapidMiner Studio เป็นต้น



3) เครื่องมือในการวิเคราะห์ข้อมูล

ซอฟต์แวร์ Rapidminer ใช้สำหรับการเตรียมข้อมูล การเรียนรู้เครื่อง การเรียนรู้ลึก การทำเหมืองข้อมูล และการวิเคราะห์การทำนาย (Predictive analysis) เป็นซอฟต์แวร์ที่ช่วยในการจัดส่งข้อมูลและลดข้อผิดพลาดจนแทบจะไม่จำเป็นต้องเขียนโค้ดเพิ่ม แต่ที่ทำให้เป็นเครื่องมือที่เหล่า Data Scientist นิยมเลือกใช้เป็นเพราะว่าตัว RapidMiner มีขั้นตอนพร้อมสำหรับการทำ Data mining (ขุดข้อมูล) และ Machine learning ซึ่งรวมไปถึงการโหลดและการแปลงข้อมูล (ETL) การประมวลผลล่วงหน้าและการวาดภาพจากข้อมูล การวิเคราะห์เชิงพยากรณ์และการสร้างแบบจำลองทางสถิติสามารถวิเคราะห์ข้อมูลด้าน Data Science ได้หลากหลาย เช่น (1) Customer Segmentation เป็นการแบ่งกลุ่มข้อมูลลูกค้าออกเป็นกลุ่มต่างๆ ซึ่งก็จะทำให้บริษัทเข้าใจพฤติกรรมของลูกค้าได้มากขึ้น เช่น ลูกค้า กลุ่มนี้เป็นกลุ่มที่มาใช้บริการบ่อย ใช้จ่ายเยอะ ก็ถือว่าเป็นกลุ่มลูกค้าชั้นดีของบริษัท (2) Demand Forecasting เป็นการคาดการณ์การผลิตสินค้า หรือการเตรียม stock สินค้าแต่ละประเภท ไว้ครับ ซึ่งส่วนใหญ่ก็เป็นลักษณะของ time series ที่ใช้ข้อมูลในอดีตมาคาดการณ์ว่าในอนาคตจะต้องผลิตสินค้าหรือ stock สินค้าแต่ละประเภทเท่าไร (3) Text

Mining เป็นการวิเคราะห์ข้อความเพื่อหาทัศนคติ (sentiment) หรือการแบ่งกลุ่มข้อความออกเป็นประเภท (category) ต่าง ๆ ซึ่งตัว RapidMiner เองสามารถทำงานเหล่านี้ได้อยู่แล้วกับข้อความภาษาอังกฤษ แต่ถ้าจะใช้กับข้อมูลภาษาไทยก็อาจจะมีขั้นตอนเพิ่มเล็กน้อยเนื่องจากภาษาไทยเพราะมีความยากในการตัดคำ (tokenize) จึงต้องใช้โมดูลต่าง ๆ ของ Python มาช่วย เช่น PyThaiNLP หรือ DeepCut โดยสามารถเขียน code ภาษา Python เข้าไปในตัว RapidMiner ได้เลย เป็นต้น RapidMiner platform มี 3 โมดูลใหญ่ๆ คือ โมดูลแรก : RapidMiner Studio ใช้สำหรับการออกแบบการวิเคราะห์ข้อมูลผ่านทางหน้า GUI ซึ่งสามารถทำการจัดการข้อมูล และสร้างโมเดลแบบต่าง ๆ ได้ โมดูลสอง : RapidMiner Server เป็นโมดูลที่รองรับการทำงานของหลายๆ user ได้ครับ ช่วยในเรื่องการตั้งเวลาให้ทำงาน (scheduler) หรือสร้าง web service หรือ web application ได้ด้วย โมดูลสาม : RapidMiner Radoop เป็นโมดูลที่ใช้ในการจัดการข้อมูลที่มีขนาดใหญ่ๆ แบบ Big Data โดยการทำงานจะไปประมวลผลบน Hadoop/Spark แต่ไม่ต้องเขียน code เพิ่มในหน้าต่างทำงานของซอฟต์แวร์ RapidMiner มีรายละเอียด ดังนี้ (ดาวน์โหลดฟรีได้จาก http://www.rapidminer.com)

(1) Repository : ส่วนนี้จะใช้ในการจัดการไฟล์ต่างๆ ของ RapidMiner Studio โดยจะเก็บไฟล์ข้อมูล หรือ Process ต่าง ๆ ไว้ใน Folder เพื่อความสะดวกในการเรียกใช้งานครั้งถัดไป

(2) Operators : ส่วนนี้จะเก็บ Operators ในการใช้งานต่างๆ ไว้เป็นกลุ่มตามหน้าที่ที่คล้ายคลึงกัน สามารถค้นหา Operators ที่ต้องการได้ในช่อง Search เพื่อความสะดวกในการเรียกใช้งาน

(3) Process : ส่วนนี้เป็นส่วนที่สำคัญของ RapidMiner Studio เพราะเป็นการนำเอา Operators ต่างๆ มา ประกอบกันให้เป็น Process ขึ้นมาใช้งาน

(4) Parameters : ส่วนนี้จะเป็นส่วนที่แสดงพารามิเตอร์ (parameter) ที่เกี่ยวข้องกับแต่ละ Operator เพื่อให้ผู้ใช้งานปรับแต่งตามที่ต้องการ ถ้าเปรียบเทียบกับ tools อื่น RapidMiner มีทั้ง free license และ commercial license แต่ราคาก็ไม่แพงเท่ากับซอฟต์แวร์ของ SAS หรือ IBM

4) การเตรียมข้อมูล (Data Preparation)

ข้อมูลที่มีอยู่ในตารางต่าง ๆ ที่ประกอบไปด้วยแถวและคอลัมม์ ซึ่งจะเรียกในแถวเป็น ตัวอย่าง (Example) ส่วนคอลัมพ์เรียก แอททริบิวต์ (Attribute) มีหน้าที่ (role) 3 แบบ คือ ID , Attribute เป็นแอททริบิวต์ปกติที่จะใช้ในการสร้างโมเดลหรือเรียกว่าฟีเจอร์ (feature) หรือตัวแปรต้น (independent) , label คือเป็น Attribute ที่เป็นคำตอบ Value type การเตรียมข้อมูล เพื่อให้การนำเข้าข้อมูลมีความถูกต้อง ก่อนการนำเข้าจึงต้องมีการจัดการข้อมูล (preprocessing) ดังนี้ (1) จัดการข้อมูลที่มีความผิดพลาดในชุดข้อมูล เช่น ข้อมูลมีค่าไม่ตรง ข้อมูลขาดหายไป (missing value) ข้อมูลแปลกแยก (outlier) เป็นต้น (2) แปลงข้อมูล เช่น Discretization แปลงข้อมูล numeric ให้เป็น nominal โดยการแบ่งข้อมูลออกเป็นช่วงๆ ได้แก่ แบ่งตามเงื่อนไขที่กำหนด แบ่งตามช่วงของข้อมูลที่เท่ากัน (equal width) และ แบ่งตามข้อมูลที่มีความถี่เท่ากัน (equal frequency) Normalization แปลงข้อมูล numeric ให้มี scale ที่เท่ากัน การแปลงข้อมูลจากฐานข้อมูล relation database ให้เป็นฐานข้อมูล transaction database การหากฎความสัมพันธ์ (association rules) คือ การสร้างจาก item ที่เกิดขึ้นบ่อยๆ โดยเทคนิคการหากฎความสัมพันธ์ ได้แก่ Apriori และ FP Growth

5) การแบ่งกลุ่มข้อมูล (Clustering)

การแบ่งกลุ่มข้อมูล (clustering) คือ การแบ่งกลุ่มข้อมูล โดยข้อมูลที่มีลักษณะคล้ายๆ กัน อยู่กลุ่มเดียวกัน และข้อมูลที่อยู่คนละกลุ่มจะมีลักษณะที่แตกต่างกันมากๆ เทคนิคการแบ่งกลุ่มข้อมูล ได้แก่ K-Means, Agglomerative Clusstering และ DBScan การจำแนกประเภทข้อมูล (classification)คือ การนำข้อมูลเดิมที่มีคำตอบที่สนใจ คือ คลาส (class) มาสร้างเป็น โมเดล (model) เพื่อหาคำตอบให้กับข้อมูลใหม่ (unseen data) โดยคลาสคำตอบเป็น ประเภท (nomimal) เช่น ฝนตกหรือไม่ตก, spam email หรือ normal email เป็นต้น เทคนิคการจำแนกประเภทข้อมูล ได้แก่ Linear Regression, Naive Bayes,

Decision Tree, K-Nearest Neighbours, Neural Networks และ Support Vector Machines การเปรียบเทียบประสิทธิภาพของโมเดลจากเทคนิค classification ต่าง ๆ และ t-test การทำ Text Mining การทำ Image Mining

6) การจำแนกประเภทข้อมูล (classification) และการประมาณค่า (estimation) งานส่วนใหญ่ของ Data Mining จะเป็นในการทำ Classification ซึ่งสามารถพบในชีวิตประจำวัน เช่น การพยากรณ์อากาศ, เรื่อง Speech recognition, face recognition การอัพโหลดรูป แล้วบอกว่าเป็นหน้าใครเช่นในเฟสบุ๊ค, Spam e-mail การจำแนกประเภทข้อมูล (classification) มีวิธีการที่นิยมใช้คือ เทคนิค Decision Tree เทคนิค Naive Bayes เทคนิค K-Nearest Neighbors (k-NN) เทคนิค Linear Regression เทคนิค Neural Network และตัววัดประสิทธิภาพของโมเดล (Classification) Confusion Matrix เป็นการทำนายไว้ว่าถูกผิดเท่าไหร่ Precision ดูสิ่งที่เรา Predict ออกมา แล้วทายถูกได้กี่เปอร์เซ็นต์ Recall จำนวนที่ทำนายถูกกี่ตัว F-Measure ช่วยหาค่าเฉลี่ยของ Precision และ Recall Accuracy จำนวนข้อมูลที่ทำนายถูกของทุกคลาส ROC Graph & Area แสดงกราฟความสัมพันธ์ ทำนายถูกไปทางแกนตั้ง (Y) ถ้าทายผิดไปแนวแกนนอน (X) ROC Curve มีค่าเข้าใกล้ 1 จะแสดงว่ามีประสิทธิภาพดีกว่า จากกราฟ และ Area Under Curve พื้นที่ใต้กราฟ (AUC) ถ้าพื้นที่ใต้กราฟเข้าใกล้ 1 จะมีพื้นที่ใต้กราฟมาก ฉะนั้นมีค่ามาก (เข้าใกล้ 1) จะยิ่งดี สรุปแล้วเป็นตัววัดประสิทธิภาพเหล่านี้จะช่วยมองในมุมมองต่าง ๆ ได้ Validation การแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพของโมเดล วิธีการทดสอบโมเดลแบ่งออกเป็น 3 ตัว (1) Self-consistency test (use training set) เอา Training Data มาเป็นตัว Test เลย คือ ใช้ข้อมูลเดิมแล้วมันมีความถูกต้องมากน้อยแค่ไหน (2) Split test แบ่งออกเป็น 2 ส่วนเลย คือ เป็นโมเดล และส่วนทดสอบ เช่น 70% สร้างโมเดล 30% ใช้ทดสอบ หรือ 80:20 ใช้ข้อมูล 2 ชุด เป็น Training data สำหรับสร้าง Model และ Testing Data สำหรับทดสอบ ถ้า Sampling มาดีก็ได้ข้อมูลทดสอบดี Split Test จะดีเมื่อมีข้อมูลมาก ๆ และ (3) Cross-validation test ใช้ค่อนข้างเยอะเหมือนกัน การทำ Split หลาย ๆ รอบ การทำคล้ายๆ กับ Split แต่แบ่งจำนวน N ชุดเท่าๆ กัน เช่น N=5 , N= 10 แล้วทำงานโดยการสร้างโมเดลทั้งหมด N ตัว จนครบ Split Test ทำรอบเดียวแล้วจบไป คือเก็บ1, 2 ไว้ แต่ใน Cross จะเอา 3 ไปเป็นตัวทดสอบ แล้วใส่ 2 กลับเข้าไป แล้วเอา 1 เป็นตัวทดสอบ จะพบว่าทุกตัวจะถูกนำมาทดสอบหมด เราจะได้ความถูกต้องเฉลี่ยออกมาในแต่ละรอบ แต่ไม่เหมาะกับการทำแบบนี้กับกรณีที่ข้อมูลมากๆ ไม่เหมาะ

.....

2.2 ประสบการณ์/ประโยชน์ที่ได้รับ/การประยุกต์ใช้กับหน่วยงาน

☑ต่อตนเอง

ทำให้สามารถเรียนรู้การวิเคราะห์ข้อมูล Big data ผ่านโปรแกรม RapidMiner Studio ทำ Data Visualization การทำ Infographic จาก Data ให้สามารถ Storytelling ใช้โมดูลต่าง ๆ ของ Python มาช่วย เช่น PyThaiNLP หรือ DeepCut โดยสามารถเขียน code ภาษา Python เข้าไปในตัว RapidMiner ...

.....

.....

☑ต่อหน่วยงาน / การนำมาประยุกต์ใช้กับหน่วยงาน

.....แนวทางการวิเคราะห์ข้อมูล Big data ผ่านโปรแกรม RapidMiner Studio ทำ Data Visualization การทำ Infographic จาก Data ให้สามารถ Storytelling ได้การนำเสนอการวิเคราะห์ข้อมูลด้านเศรษฐกิจที่ดิน ที่เข้าใจยากกับนักวิชาการวิชาชีพอื่น การนำ RapidMiner Studio มาวิเคราะห์ข้อมูลและนำเสนอแบบ Infographic นับว่าเป็นสิ่งที่น่าสนใจในการพัฒนางานด้วยเทคนิคใหม่ ๆ

.....

2.3 ปัญหาและอุปสรรคในการอบรม/สัมมนา/พัฒนาความรู้ฯ

- การเรียนรู้เน้นด้านวิทยาศาสตร์ข้อมูล จำเป็นต้องฝึกทำแบบฝึกหัดมากกว่าปกติ และเป็นศาสตร์แขนงด้านวิทยาศาสตร์ข้อมูล จึงใช้เวลามาก

2.4 ข้อคิดเห็นและข้อเสนอแนะ

ควรให้นักวิชาการฝึกปฏิบัติโปรแกรม RapidMiner Studio หรือสนับสนุนการฝึกอบรม

ลงชื่อ.....

(นายดิเรก คงแพ)

ตำแหน่ง นักวิเคราะห์นโยบายและแผนชำนาญการพิเศษ

ผู้รายงาน

วันที่ ๒๙ เดือน ก.ค. พ.ศ. ๖๕

ส่วนที่ 3 ความเห็นของผู้บังคับบัญชา

(✓) ทราบ

ลงชื่อ.....

((นายสมศักดิ์ สุขจันทร์))

ตำแหน่ง ผู้อำนวยการกองนโยบายและแผนการใช้ที่ดิน

วันที่ ๒๒ เดือน ก.ค. พ.ศ. ๖๕