INFOPRESS

Practical DATA SCIENCE 1 เล่ม

with RapidMiner Studio

RAPIDMINER

No-Code/Low-Code Data Science Platforms

อธิบายการสร้าง Basic Machine Learning Models
ตั้งแต่ต้นจนจบ เหมาะสำหรับผู้เชี่ยวชาญสาขา
ทางธุรกิจที่อยากเป็น Citizen Data Scientist

WORKSHOP DATASET :
SERAZU.COM

ผู้แต่ง ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
บรรณาธิการ ภีรพล คชาเจริญ

IDC
PREMIER

มีเพียง “**ความรู้**” เท่านั้นที่มนุษย์ใช้พลิก “**โลก**”
และเปลี่ยน “**ชีวิต**” เราจึงสร้างสรรค์ และส่งมอบ “**ความรู้**”
ในรูปแบบที่ดีกว่า เพื่อให้คนไทย “**เรียนรู้**” ได้ตลอดชีวิต

Only “**Knowledge**” can help human
change “**The World**” and “**Their Lives**”.
With this truth, it drives us to deliver
“**Knowledge**” for Thai being able to
“**Learn**” better everyday.

Think Beyond
INFOPRESS
Read comics
Little HEART

Practical Data Science with RapidMiner Studio เล่ม 1

Writer	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
Editor	ภีรพล คชาเจริญ
Production Manager	วรพล ณธิกุล
Graphic Designer	ชวนันท์ รัตนะ
Page Layout	สุรัสวดี วงศ์จันทร์สุข
Proofreader	สุนทรี บรรลือศักดิ์
Publishing Coordinators	สุพัตรา อาจปรุ, สุรีย์รัตน์ จิอู๋
Product Specialist	ศรันย์ ชาติสุทธิผล

สร้างสรรค์โดย

INFOPRESS

ข้อมูลทางบรรณานุกรม
ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
Practical Data Science with RapidMiner Studio เล่ม 1
นนทบุรี : ไอดีซีฯ, 2565
264 หน้า

1. โปรแกรมสำหรับระบบการประมวลผลข้อมูลเฉพาะชนิด I ชื่อเรื่อง
005.37

ISBN 978-616-487-359-9

พิมพ์ครั้งที่ 1 ตุลาคม 2565

ราคา 385 บาท

จัดพิมพ์และจัดจำหน่ายโดย

IDC PREMIER

บริษัท ไอดีซี พรีเมียร์ จำกัด
200 หมู่ 4 ชั้น 19 ห้อง 1901
อาคารจัสมินอินเตอร์เนชั่นแนลทาวเวอร์
ถ.แจ้งวัฒนะ อ.ปากเกร็ด จ.นนทบุรี 11120
โทรศัพท์ 0-2962-1081 (อัตโนมัติ 10 คู่สาย)
โทรสาร 0-2962-1084

สมาชิกสัมพันธ์
โทรศัพท์ 0-2962-1081-3 ต่อ 121
โทรสาร 0-2962-1084

ร้านค้าและตัวแทนจำหน่าย
โทรศัพท์ 0-2962-1081-3 ต่อ 112-114
โทรสาร 0-2962-1084

พิมพ์ที่ บริษัท พงษ์วรินการพิมพ์ จำกัด
299-299/1 หมู่ 10 ถ.สุขุมวิท 107 ต.สำโรงเหนือ
อ.เมืองสมุทรปราการ จ.สมุทรปราการ 10270
โทรศัพท์ 0-2399-4525-31 โทรสาร 0-2399-4524

PREFACE

ผมเองอยู่ในแวดวง Data มาเกือบ 20 ปีแล้วครับ เห็นหลายอย่างผ่านไปอย่างรวดเร็ว ตั้งแต่ช่วงที่คนทำงานด้าน Data ไม่มีตำแหน่งงานที่ชัดเจน จนมาถึงยุคปัจจุบันที่หลายองค์กรตระหนักถึงความสำคัญของ Data และอยากหาประโยชน์จากสิ่งเหล่านี้ จึงเกิดตำแหน่งงานใหม่ขึ้นมา เช่น Data Scientist ซึ่งเป็นอาชีพที่ใครหลายคนใฝ่ฝัน แต่การจะเป็น Data Scientist จริงๆ แล้วนั้น จะต้องมีความรู้ความสามารถหลายอย่าง ไม่ว่าจะเป็นความรู้ในเรื่อง Data Science และ Machine Learning ความรู้ด้านการเขียนโปรแกรม และความรู้ทางธุรกิจที่เกี่ยวข้องกับงานที่รับผิดชอบ

ดังนั้น ตำแหน่งนี้จึงดูเหมือนจะขาดแคลนอยู่เป็นจำนวนมาก จนมีแนวทางใหม่เกิดขึ้นที่เรียกว่า Citizen Data Scientist ซึ่งจะเน้นไปที่การนำบุคลากรที่มีความรู้ หรือประสบการณ์ทำงานในด้านนั้นๆ เช่น นักบัญชี นักการตลาด มาเรียนรู้เทคนิคการวิเคราะห์ข้อมูลเพิ่มเติม และใช้ซอฟต์แวร์ที่เรียกว่า No Code/Low Code Data Science Platform แทนการเขียนโปรแกรม ซึ่ง RapidMiner ก็เป็นซอฟต์แวร์ในกลุ่ม No Code/Low Code ที่ได้รับการยอมรับและใช้งานกันอย่างแพร่หลายตัวหนึ่งเลยครับ

ผมเองเริ่มต้นศึกษา RapidMiner มาตั้งแต่เวอร์ชัน 6 และจัดอบรมการใช้งานและให้คำปรึกษาเกี่ยวกับ RapidMiner เรื่อยมาจนพบว่า เวอร์ชัน 9 มีฟีเจอร์ใหม่ที่ทำให้เราใช้งานได้มากกว่าแต่ก่อน นั่นคือ Turbo Prep ที่ช่วยให้เราเตรียมข้อมูลได้ง่ายขึ้น และ Auto Model ที่ทำให้เราสร้างและเปรียบเทียบโมเดลทาง Machine Learning ได้เพียงแค่ไม่กี่ขั้นตอนเท่านั้น ด้วยเหตุผลนี้ ผมจึงอยากแชร์ประสบการณ์ใช้งาน RapidMiner Turbo Prep และ Auto Model ด้วยการยกตัวอย่าง Use Case ต่างๆ ไว้ในหนังสือเล่มนี้ครับ

สุดท้ายนี้ ถ้าใครเคยใช้ RapidMiner มาก่อนแล้ว ผมอยากให้ลองอ่านหนังสือเล่มนี้ แล้วคุณจะพบกับความง่ายในการใช้งานมากกว่าเดิม และใครที่เริ่มสนใจทางด้าน Data Science ผมก็ขอแนะนำ RapidMiner ให้เป็นอีกหนึ่งทางเลือกนอกจากการเขียนโปรแกรมต่างๆ ครับ เพราะในตลาดมีบริษัทที่ต้องการความรวดเร็วในการวิเคราะห์ข้อมูลมากกว่าค่าใช้จ่ายอยู่อีกมากครับ

ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

Data Science & RapidMiner Expert

Co-Founder and Data Science Team Lead

Cube Analytics Consulting Co., Ltd.

rapidminer

This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Machine Learning Master Certification

The Master level is the second level after Professional and stands for high proficiency in the topic. Some of the topics tested at this level are:

Complex Predictive Models, **Cross Validation and Correct Validation**, **Parameter Optimization**, **Model Selection**, **Feature Engineering**, **Time Series and Forecasting** and **Integrating R and Python models.**

This certificate can be verified using the open badge which was issued to the certificate holder.
The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck

Knut Makowski
(Head of Education)

Peter Lee
(CEO RapidMiner)

rapidminer

This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Data Engineering Master Certification

The Master level is the second level after Professional and stands for high proficiency in the topic. Some of the topics tested at this level are:

Loops and Branches, **Advanced Text Processing**, **Exception Handling**, **Logging**, **Data Cleansing**, **Regular Expressions**, **Macros**, **Web APIs**, and **Scripting**.

This certificate can be verified using the open badge which was issued to the certificate holder.
The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck

Knut Makowski
(Head of Education)

Peter Lee
(CEO RapidMiner)

rapidminer

This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Applications & Use Cases Master Certification

The Master level is the second level after Professional and stands for high proficiency in the topic. Some of the topics tested at this level are:

Running Processes, **Deploying Models**, **Model Management** and **Web Apps**.

This certificate can be verified using the open badge which was issued to the certificate holder.
The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck

Knut Makowski
(Head of Education)

Peter Lee
(CEO RapidMiner)

rapidminer

This is to certify that

Eakasit Pacharawongsakda

successfully passed the examination for

Platform Administration Master Certification

The Master level stands for high proficiency in the topic. Some of the topics tested at this level are:

Platform Overview, **RapidMiner Studio Installation**, **RapidMiner Server Administration**, **RapidMiner Radoop**, **RapidMiner Real-Time Scoring**, and **RapidMiner Marketplace**.

This certificate can be verified using the open badge which was issued to the certificate holder.
The open badge can be uploaded and tested via this redirect: www.rapidminer.com/badgecheck

Knut Makowski
(Head of Education)

Peter Lee
(CEO RapidMiner)

EDITOR'S NOTE

ถ้าย้อนกลับไปเมื่อหลายปีก่อนที่ 'Data Science' จะกลายเป็น Buzzword ที่น่าตื่นเต้น ปลุกกระแส 'ปัญญาประดิษฐ์' หรือ 'AI' ให้กลับมาโลดแล่นอีกครั้ง หลังจากล้มลุกคลุกคลานมาแล้วหลายครั้ง เนื่องจากองค์ความรู้และเทคโนโลยียังเอื้อมไม่ถึง AI จึงเป็นแนวคิดที่มาก่อนกาล แต่เมื่อเกิด 'Machine Learning' และ 'Deep Learning' ซึ่งเป็นอัลกอริทึมเรียนรู้เอง สิ่งต่างๆ ที่ไม่สามารถทำได้ในอดีต แต่มันเป็นไปได้แล้วในปัจจุบัน เนื่องจากองค์ความรู้และองค์ประกอบทางเทคโนโลยีในตอนนี้เติบโตอย่างเต็มที่ จนทำให้ AI กลับสู่เส้นทางแห่งความสำเร็จอีกครั้ง

ศาสตร์ Data Science กลายเป็น MegaTrend ที่สร้างโลกเปลี่ยนอนาคต และอาชีพ Data Scientist ก็ถูกยกให้เป็น "The Sexiest Job of the 21st Century" ที่ Harvard Business ให้การยกย่อง กลายเป็น Content สร้างกระแสจากอดีตจนถึงปัจจุบัน

ในยุคแรก ศาสตร์ Data Science ตามหลักการพื้นฐานโดยผู้เชี่ยวชาญในตอนนั้น สรุปแนวคิดไว้ว่า มันประกอบด้วยองค์ความรู้แบบสหสาขา ทั้งคณิตศาสตร์และสถิติ (ศาสตร์), การเขียนโปรแกรม (ทักษะการเขียนโค้ด) และความรู้ทางธุรกิจ/ฟังก์ชันงาน (รู้ดีถึงปัญหาและความท้าทายที่ต้องเผชิญ) ถึงแม้ว่างาน Data Science นั้นจะประกอบด้วยทีม Data ที่มีสมาชิกเชี่ยวชาญกันคนละด้านก็ตาม แต่ Data Scientist และ Data Engineer ก็ยังต้องเป็น Expert ทางเทคนิคอล (Technical) ซึ่งกลายเป็นข้อจำกัด

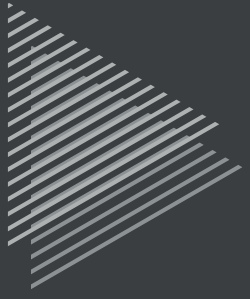
และด้วย Data Science Process มีหลายขั้นตอน มีรายละเอียดการทำงานที่ซับซ้อน จึงทำให้ต้นทุนการพัฒนาโปรเจ็กต์สูง ใช้เวลาปลุกปั้นนาน การปรับใช้ให้เกิดประโยชน์อย่างคุ้มค่าในเชิงธุรกิจยาก และที่สำคัญการฟอร์มทีมไม่ง่ายเลยในระยะแรก สิ่งเหล่านี้เป็นเหมือนยาขมที่ขวางทางการเข้าถึง Data Science ของคนในระดับองค์กร และในระดับธุรกิจอุตสาหกรรม

ในที่สุดก็เกิดเทรนด์ใหม่ที่เรียกว่า “Data Science and Machine Learning (DSML) Platforms” ที่เข้ามาเติมเต็มช่องว่าง เปลี่ยนโลก Data Science ดั้งเดิมที่เข้าถึงยาก และต้องใช้ผู้เชี่ยวชาญ มาเป็นแพลตฟอร์มที่เข้าถึงง่ายและมีความยืดหยุ่นมากขึ้น โดย DSML Platforms รองรับทั้ง Expert Data Scientists (Technical) และ Citizen Data Scientists (Nontechnical) และยังเติมขีดความสามารถให้กับ Data Engineers, Developers และ ML Engineers อีกทั้งคนสาย Tech และสาย Non-Tech สามารถทำงานร่วมกันได้บนแพลตฟอร์มเดียวกัน แค่มาเรียนรู้แอปพลิเคชันซอฟต์แวร์บน DSML Platforms ซึ่งในหนังสือเล่มนี้คือ RapidMiner ที่ Gartner ระบุไว้ในรายงานว่า “RapidMiner is a trustworthy choice.”

สุดท้ายนี้ ผมต้องขอขอบพระคุณ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา หนึ่งในผู้เชี่ยวชาญสาขา Data Science & RapidMiner Expert ที่ได้สร้างสรรค์หนังสือเชิงปฏิบัติการเล่มนี้ที่สำเร็จได้ยาก เพราะมีอุปสรรคในการเตรียม Datasets ที่เหมาะกับการสร้าง Machine Learning Models ที่สอดคล้องกับ Model Type การอธิบายกระบวนการตั้งแต่ต้นจนจบแบบ Step by Step รวมไปถึงการปรับใช้ในธุรกิจ ตัวอย่างในเล่มเกิดจากประสบการณ์ทั้งในฐานะอาจารย์และที่ปรึกษา หวังว่าผู้อ่านจะได้รับความรู้ตามที่คาดหวัง และอย่าลืมติดตามหนังสือ Practical Data Science with RapidMiner Studio เล่ม 2 (Advanced Model) และ Practical Data Engineer with RapidMiner Studio ในโอกาสต่อไป

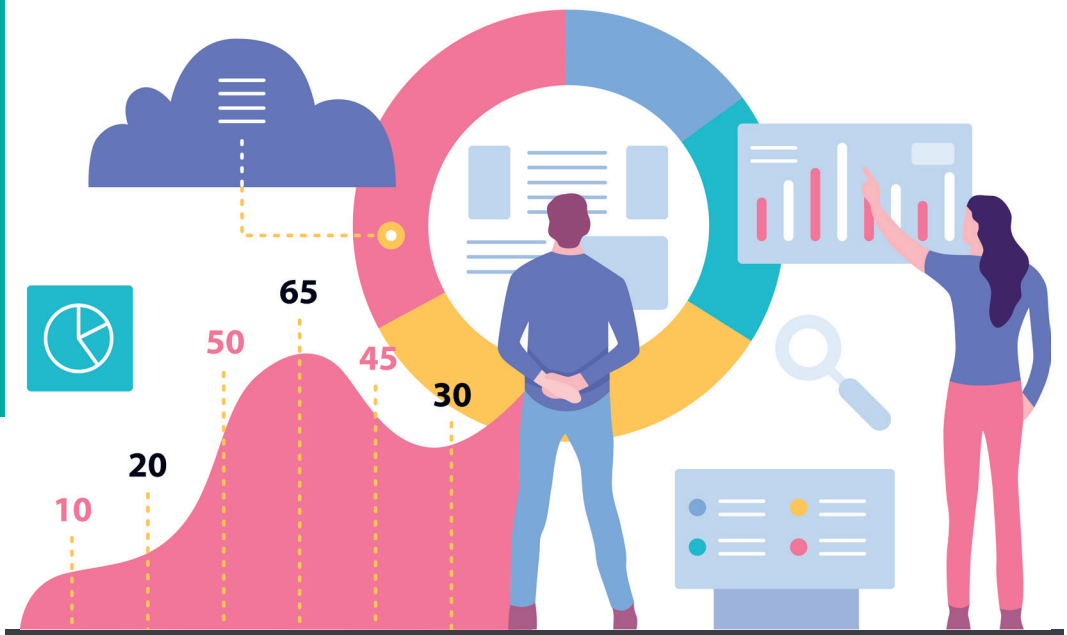
ภีรพล คชาเจริญ
บรรณาธิการ

CONTENT



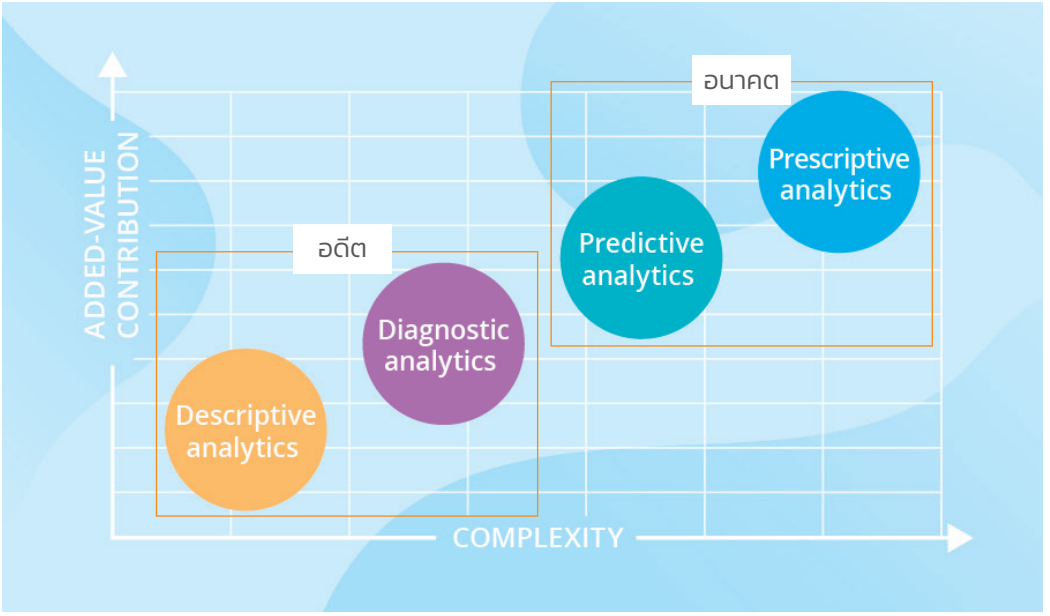
Chapter 1
DATA SCIENCE PLATFORMS ทางเลือกใหม่สำหรับ DATA SCIENTIST

Chapter 2
เริ่มต้นการใช้งาน RAPIDMINER STUDIO 9.10



DATA SCIENCE PLATFORMS ทางเลือกใหม่สำหรับ DATA SCIENTIST

องค์กรทุกขนาดในทุกอุตสาหกรรม ต่างมุ่งมั่นที่จะขับเคลื่อนด้วยข้อมูลมากขึ้น แต่ด้วยเกิดภาวะการขาดแคลน Data Scientists ที่จะมาตอบสนองได้ทันต่อความต้องการธุรกิจจึงสนใจแนวคิดใหม่ ที่จะมาเพิ่มขีดความสามารถให้กับคนที่มีอาชีพเป็นนักวิเคราะห์ข้อมูล ผู้เชี่ยวชาญด้านข้อมูล ผู้ที่มีประสบการณ์ในธุรกิจสาขาต่างๆ มาเรียนรู้แอปพลิเคชันซอฟต์แวร์บน Data Science and Machine Learning (DSML) Platforms โดยเรียกบุคลากรกลุ่มใหม่นี้ว่า “Citizen Data Scientists” และกลุ่ม Data Scientists ซึ่งเป็นผู้เชี่ยวชาญทางเทคนิคก็สามารถใช้ซอฟต์แวร์ DSML Platforms เพื่อ Develop & Deploy โมเดลไปปรับใช้ในองค์กรได้เช่นเดียวกัน



รูปที่ 1 – 4 การทำความเข้าใจการวิเคราะห์ข้อมูลในระดับที่ซับซ้อนและคุณค่าที่ได้รับ

เครดิตภาพ : www.scnsoft.com/blog/4-types-of-data-analytics

อยากให้เข้าใจว่า Citizen Data Scientist ไม่ได้จะเข้ามาแทน Data Scientist ตั้งเดิมแต่อย่างใด แต่คือส่วนที่เข้ามาเติมเต็มให้ทีม Data Science มีความยืดหยุ่นมากขึ้น เพราะการใช้แพลตฟอร์มหรือซอฟต์แวร์สำเร็จรูปนั้นทำได้ง่าย รวดเร็ว และประหยัด ที่สำคัญคือ หาบุคลากรที่เหมาะสมได้ง่ายกว่า ช่วยให้ธุรกิจสามารถปรับเปลี่ยนเพื่อเปลี่ยนแปลงด้วย Data Science ได้จริงมากขึ้นครับ

ด้วยแนวคิดดังกล่าว ผมจึงตั้งใจเขียนหนังสือเล่มนี้ขึ้นมา เพื่อให้โอกาสกับคนที่อยากทำงานด้าน Data Science ได้เข้าใจแนวคิดการวิเคราะห์ข้อมูลต่างๆ ทาง Machine Learning โดยเรียนรู้จากการลงมือปฏิบัติจริงโดยใช้ซอฟต์แวร์ที่ชื่อว่า **RapidMiner** ซึ่งถูกกล่าวถึงในรายงาน 2 ฉบับของ Gartner ว่า “RapidMiner is a trustworthy choice.”

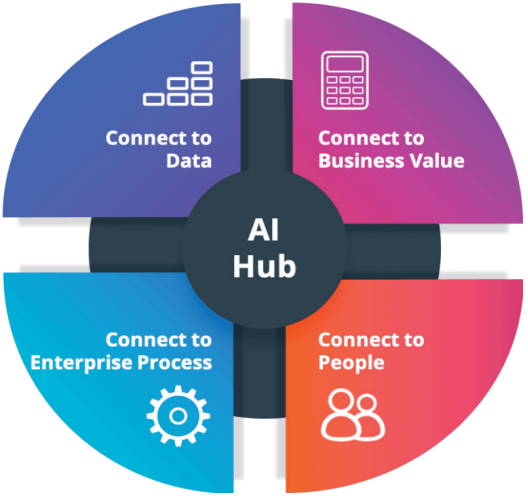
อ้างอิง : https://rapidminer.com/downloads/2022-gartner-market-guides/

RapidMiner AI Hub

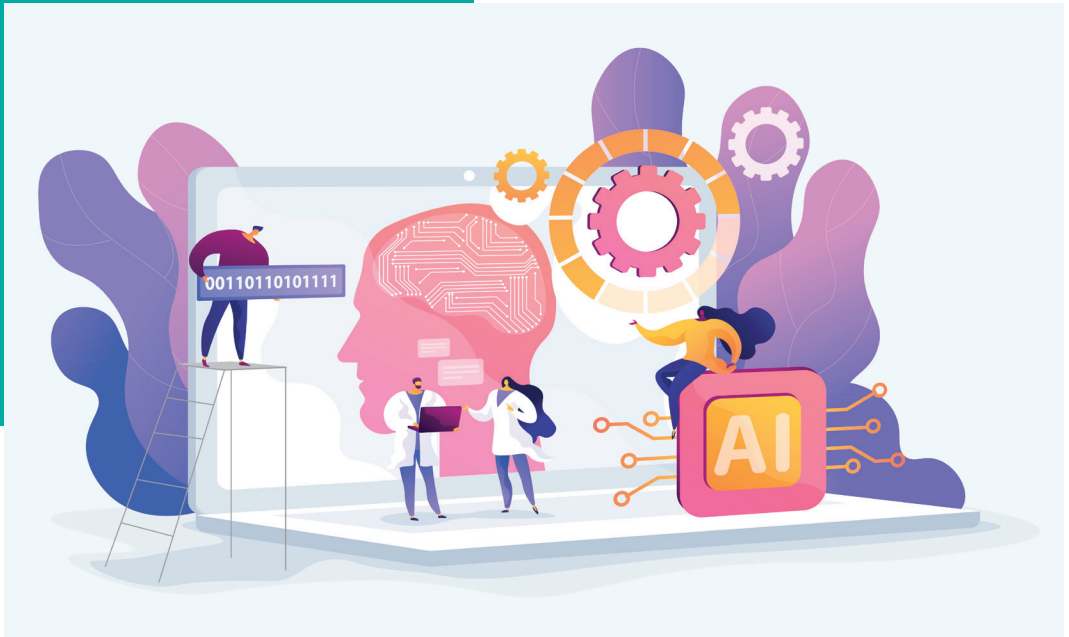
RapidMiner AI Hub

Data Science Platform for Enterprise

ถูกพัฒนาขึ้นเพื่อให้บริการสำหรับองค์กรแทนที่ RapidMiner Server เนื่องจากเทคโนโลยีคลาวด์ได้รับความนิยมในปัจจุบัน AI Hub มีเครื่องมือที่ช่วยเชื่อมโยงทุกคนในองค์กรเข้ากับการจัดการข้อมูล และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ และนำความสามารถของ AI ไปสร้างผลลัพธ์ที่มีคุณค่าต่อธุรกิจได้ รองรับการทำงานแบบ Multi-User คือ การทำงานพร้อมกันได้หลายยูสเซอร์ และ Multi-Tasking คือ สามารถรันโพรเซส (Process) ต่างๆ ได้พร้อมกัน และยังสามารถสร้าง Web Service เพื่อใช้ในการติดต่อกับโปรแกรมอื่นๆ ได้อีกด้วย เป็นแพลตฟอร์มที่เอื้อต่อการทำงานเป็นทีม โดยรวมเครื่องมือทุกอย่างไว้บนแพลตฟอร์มเดียวเพื่อใช้งานร่วมกัน เช่น



- **RapidMiner Go :** เครื่องมือที่ใช้งานง่ายสำหรับผู้ใช้ที่ไม่ใช่โปรแกรมเมอร์
- **JupyterLab :** เครื่องมือเขียนโปรแกรมสำหรับผู้ใช้ที่เป็นโปรแกรมเมอร์
- **RapidMiner Studio :** ในกรณีผู้ใช้มีการติดตั้งตัวนี้อยู่ในเครื่อง ก็สามารถเข้าถึงสิ่งที่ถูกแชร์บนแพลตฟอร์มได้ หรือเปลี่ยนไปประมวลผลบนคลาวด์ซึ่งมีฮาร์ดแวร์ที่ดีกว่าได้



เรียนรู้กระบวนการ DATA SCIENCE ในเชิงธุรกิจ

Data Science สามารถนำไปประยุกต์ใช้ในเชิงธุรกิจได้หลากหลายรูปแบบ ในบทนี้ผมจะขออธิบายเพิ่มถึงขั้นตอนในการทำโครงการที่เกี่ยวข้องกับทาง Data Science ซึ่งมีอยู่หลายแนวทาง แต่แนวทางที่ผมมักจะใช้งานก็จะเป็นไปตามแนวคิดของ CRISP-DM ซึ่งย่อมาจาก CRoss Industry Standard Process for Data Mining แม้ว่าจะมีคำว่า Data Mining อยู่ แต่ก็สามารถนำมาใช้กับทาง Data Science ได้ครับ

ขั้นตอนการทำ Data Science Project

หากจะให้คำจำกัดความของคำว่า **Data Mining** หรือ **การทำเหมืองข้อมูล** ในแบบที่ทุกคนสามารถเข้าใจได้ คงนิยามได้ว่า Data Mining คือ กระบวนการในการทำให้ข้อมูลของบริษัทมีประโยชน์ต่อความต้องการทางธุรกิจ

แต่ถ้าจะนิยามด้วยภาษาทางเทคโนโลยีแล้ว อาจจะกล่าวได้ว่า Data Mining คือ กระบวนการค้นหาข้อมูลเชิงลึก (Insight) จากรูปแบบและความสัมพันธ์ภายในชุดข้อมูลขนาดใหญ่ เพื่อคาดการณ์ผลลัพธ์ด้วยการใช้เทคนิคที่หลากหลาย ซึ่งเราสามารถใช้ประโยชน์จากข้อมูลเชิงลึกนี้ได้ในหลายทาง เช่น เพื่อเพิ่มรายได้ ลดต้นทุน ปรับปรุงความสัมพันธ์กับลูกค้า ลดความเสี่ยง และอื่นๆ

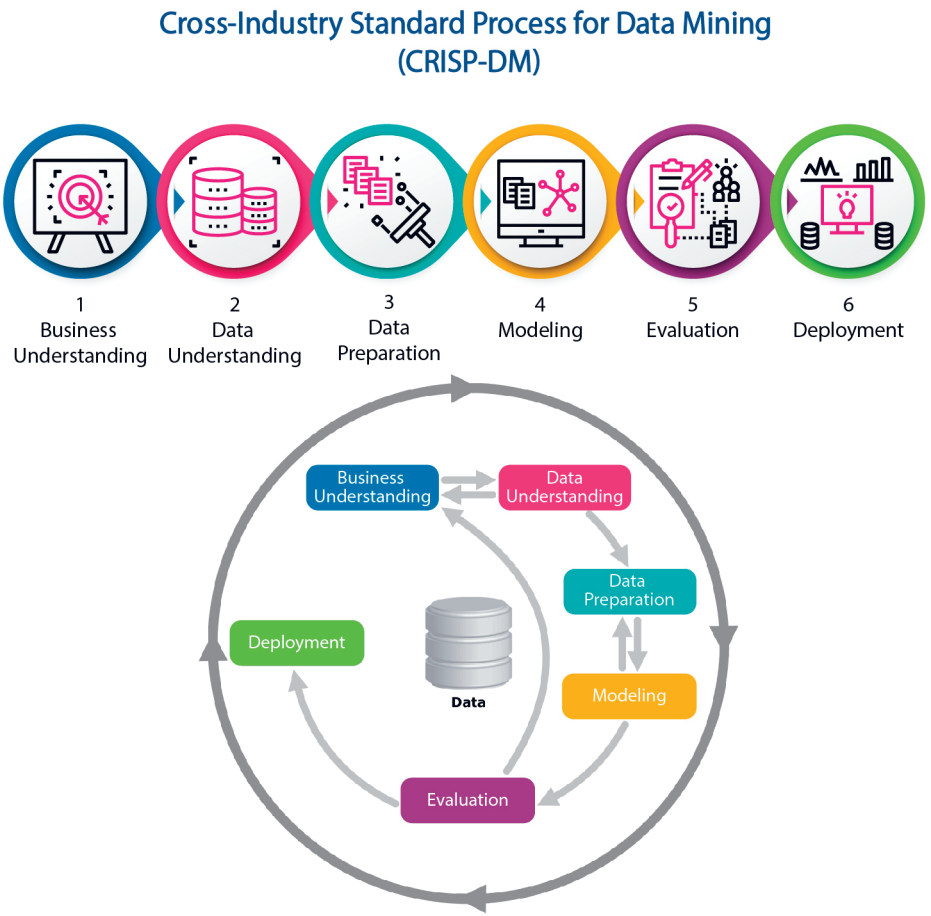
นอกจากนี้ กระบวนการทาง Data Mining ยังถูกใช้ในการสร้างโมเดล Machine Learning (ML) ซึ่งเป็นระบบการเรียนรู้และตอบสนองโดยอัตโนมัติ ที่พบได้ในแอปพลิเคชันซอฟต์แวร์เชิงปัญญาประดิษฐ์ (Artificial Intelligence : AI) เช่น อัลกอริทึมของเครื่องมือค้นหา (Search Engine) และระบบแนะนำ (Recommendation Systems) เป็นต้น

รูปที่ 3 – 1 Data Mining กระบวนการทำให้ข้อมูลของบริษัทมีประโยชน์ต่อความต้องการทางธุรกิจ

การทำ Data Science ด้วยกระบวนการเหมืองข้อมูล (CRISP-DM)

การทำ CRISP-DM Process ประกอบด้วย 6 ขั้นตอน ดังรูปที่ 3 - 2 ซึ่งได้อธิบายรายละเอียดไว้ในหนังสือ A Little Book of Big Data and Machine Learning แล้วส่วนหนึ่ง และในบทนี้เลยขอสรุปสั้นๆ เป็นตารางที่ 3 - 1 ซึ่งยกมาจากหนังสือ Text Book ที่ชื่อ

"Applied Predictive Analytics: Principle and Techniques for the Professional Data Analyst"

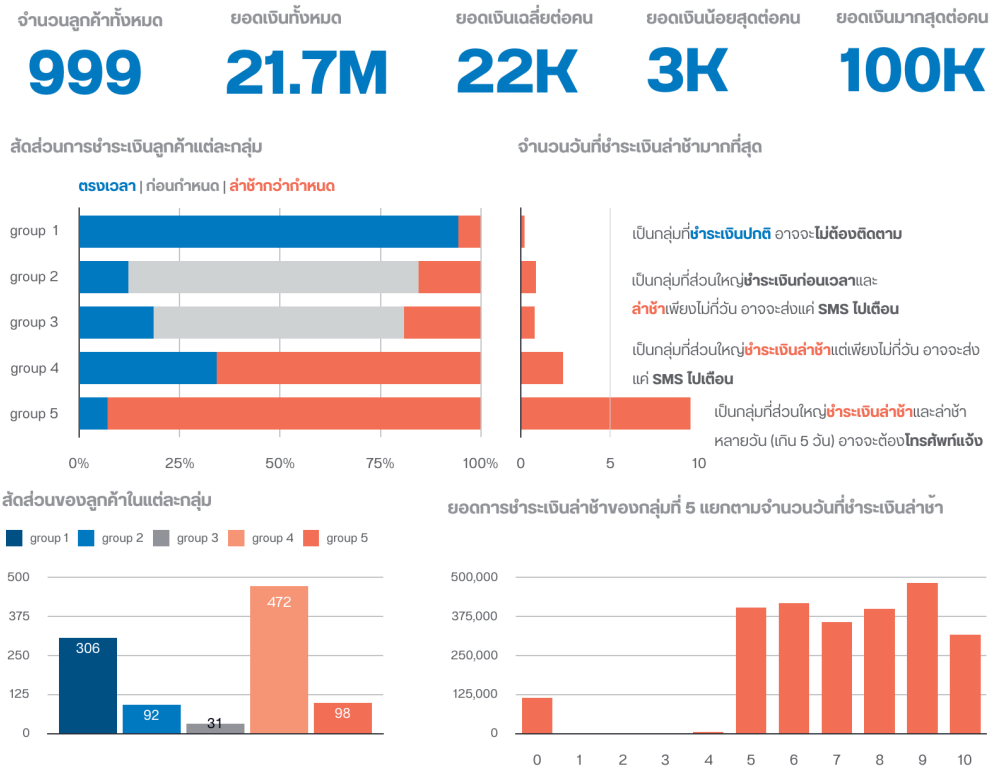


รูปที่ 3 - 2 แสดงขั้นตอนมาตรฐานของ CRISP-DM Process

จะเห็นว่า การแบ่งกลุ่มลูกค้าแบบนี้ จะทำให้บริษัทสามารถเลือกวิธีการติดต่อได้ตรงกับลูกค้าแต่ละรายมากขึ้น และช่วยลดค่าใช้จ่ายลงไปได้อีกด้วย

STEP 6 Deployment

หลังจากทำการวิเคราะห์ข้อมูลได้เรียบร้อยแล้ว ควรจะมีการนำผลลัพธ์ที่ได้มาสื่อสารด้วยข้อมูลในแบบที่เข้าใจง่าย (Data Storytelling) เพื่อให้ผู้เกี่ยวข้องเห็นภาพเดียวกันและนำไปใช้ประโยชน์ต่อไป ดังรูปที่ 3 – 5



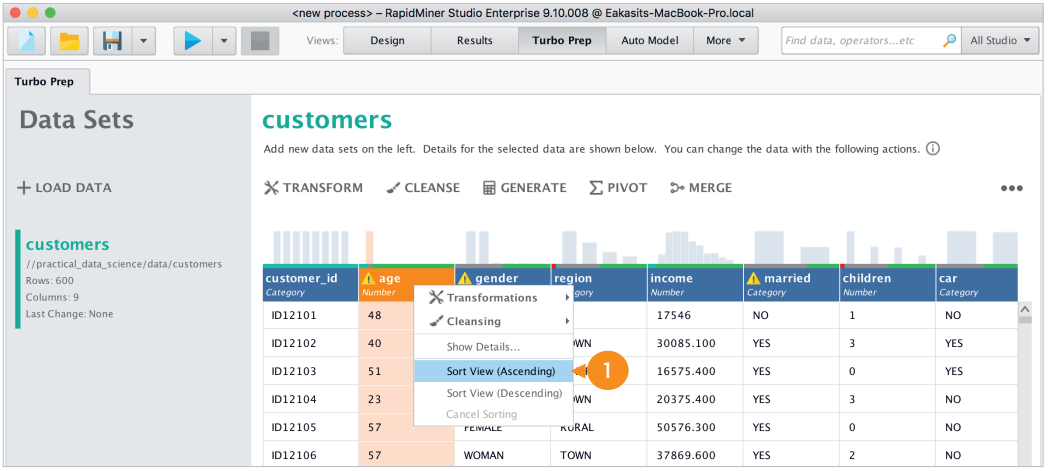
รูปที่ 3 – 5 Dashboard แสดงภาพรวมการชำระเงินของลูกค้า

CHAPTER 4

การเตรียมข้อมูลให้มีคุณภาพ ด้วย TURBO PREP

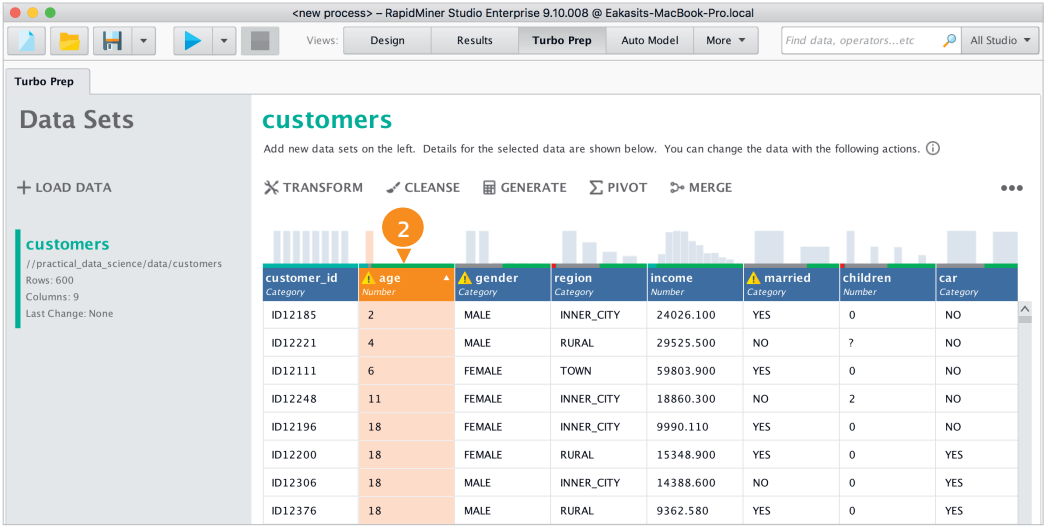
ในการวิเคราะห์ข้อมูลหลายๆ ครั้ง เรามักจะดึงข้อมูลมาจากหลาย Data Sources ที่มีหลาย Formats ที่แตกต่างกันมารวมกัน ซึ่งเรามักพบว่าข้อมูลที่มีที่มาต่างกัน มักจะยังไม่ถูกต้องเสียทั้งหมด เป็นข้อมูลคุณภาพต่ำที่ยังไม่เหมาะจะนำไปใช้งาน จำเป็นต้องมีการทำ Data Blending & Cleansing เพื่อรวบรวมและแปลงรูปแบบข้อมูลให้สอดคล้องกัน และนำไปใช้ประโยชน์ได้จริง ช่วยเปลี่ยนข้อมูลที่ไร้ค่าให้เป็นข้อมูลที่มีประโยชน์ เรียกขั้นตอนนี้ว่า การเตรียมข้อมูล (Data Preparation) ในบทที่ 4 นี้ เราจะมาดูการจัดการและแก้ไขข้อมูลแบบง่ายๆ ด้วย Turbo Prep ใน RapidMiner Studio กันครับ

1. คลิกขวาที่คอลัมน์ **age** (เปลี่ยนเป็นสีส้ม) แล้วเลือก **Sort View (Ascending)** เพื่อเรียงลำดับข้อมูลจากอายุน้อยสุดไปยังมากสุด ดังแสดงในรูปที่ 4 - 9



รูปที่ 4 - 9 แสดงเมนูสำหรับการเรียงลำดับข้อมูล (Sort) จากน้อยไปมาก

2. จากรูปที่ 4 - 10 จะเห็นว่ามีข้อมูลที่ผิดปกติคือ คนที่มีอายุเท่ากับ 2, 4, 6 และ 11 ปี ซึ่งอาจจะเป็นข้อมูลที่ผิดปกติ เพราะมีอายุน้อยผิดปกติ และบางคนสมรสแล้วทั้งที่ยังเป็นเด็ก



รูปที่ 4 - 10 แสดงข้อมูลที่เรียงลำดับตามอายุจากน้อยไปมากแล้ว

CHAPTER 5

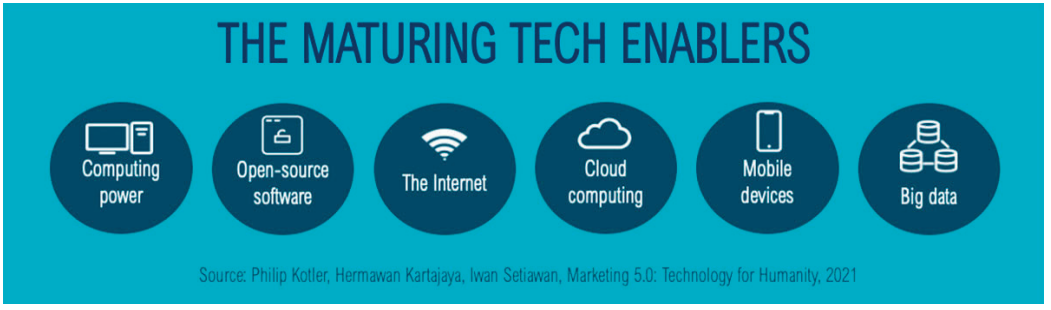
การแบ่งกลุ่มลูกค้า CUSTOMER (RFM) SEGMENTATION ด้วย TURBO PREP

ในบทที่ 4 เราได้แสดงขั้นตอนการเตรียมข้อมูล (Data Preparation) โดยใช้เครื่องมือ Turbo Prep ของ RapidMiner ในการจัดการและแก้ไขข้อมูลต่างๆ กันไปแล้ว ในบทนี้จะเริ่มเข้าสู่การวิเคราะห์ข้อมูลเบื้องต้น โดยเริ่มจากเรื่องของการทำ Segmentation โดยใช้ Turbo Prep มาช่วยในการแบ่งกลุ่มตามแนวคิดของ RFM Segmentation พร้อมกันนี้จะได้แสดงวิธีการใช้ฟังก์ชัน Generate เพื่อสร้างชุดข้อมูลใหม่จากสูตรคำนวณ ซึ่งเป็นหนึ่งในฟังก์ชันพื้นฐานที่จำเป็นที่ยังไม่ได้กล่าวถึงในบทที่ 4 ด้วยครับ

แนวคิดการแบ่งกลุ่มลูกค้าในยุค Marketing 5.0

การแบ่งกลุ่มลูกค้า (Customer Segmentation) เป็นแนวคิดทางการตลาดกระแสหลักที่ถูกนำมาใช้ในทางธุรกิจมานานแล้ว โดยเป็นการแบ่งส่วนตลาด หรือแบ่งกลุ่มลูกค้าที่มีบางสิ่งร่วมกันไว้เป็นกลุ่มเดียวกัน แต่ในยุคสมัยปัจจุบันที่เราโฟกัสเรื่อง “Digital & Technology” ทางด้าน Data จึงมีเครื่องมือที่ทันสมัยมากขึ้นบนแนวคิดเดิม ซึ่งในบทนี้จะแสดงการแบ่งข้อมูลดิจิทัลของ Customer Data ด้วย RapidMiner Turbo Prep เพื่อแบ่งข้อมูลออกเป็นกลุ่มย่อยๆ โดยมีการระบุเงื่อนไขไว้แต่แรกแล้ว ทั้งนี้เพื่อหาทางตอบสนองด้วยแคมเปญที่จูงใจได้อย่างแม่นยำ

ในหนังสือ **Marketing 5.0 : Technology for Humanity** โดย **Philip Kotler** กล่าวถึงเหตุผลที่สำคัญคือ สิ่งต่างๆ ที่ไม่สามารถทำได้ในอดีต แต่มันเป็นไปได้แล้วในปัจจุบัน เนื่องจากองค์ประกอบทางเทคโนโลยีในตอนนี้เติบโตอย่างเต็มที่ ไม่ว่าจะเป็นพลังการคำนวณ (Computational Power), ซอฟต์แวร์โอเพ่นซอร์ส (Open Source Software), การประมวลผลบนคลาวด์ (Cloud Computing), การสื่อสารทางอินเทอร์เน็ตด้วยความเร็วสูง (Internet & Broadband Speeds), อุปกรณ์มือถือ (Mobile Devices) และบิ๊กดาต้า (Big Data) ฯลฯ ล้วนเป็นปัจจัยผลักดันให้การตลาดเปลี่ยนจาก Marketing 4.0 (Moving to Digital) เป็น Marketing 5.0 (Marketing in Digital World) ซึ่งเป็นช่วงที่การตลาดเข้าสู่ระยะที่เทคโนโลยีขั้นสูงจะถูกนำไปใช้ในการทำ Marketing เช่น AI, IoT, หุ่นยนต์บริการ เป็นต้น ซึ่งจะเป็นลักษณะของ Deepen Marketing นั่นเองครับ



รูปที่ 5 – 1 เมื่อเทคโนโลยีต่างๆ เติบโตอย่างเต็มที่จึงส่งเสริมกิจกรรมทางการตลาดมากยิ่งขึ้น

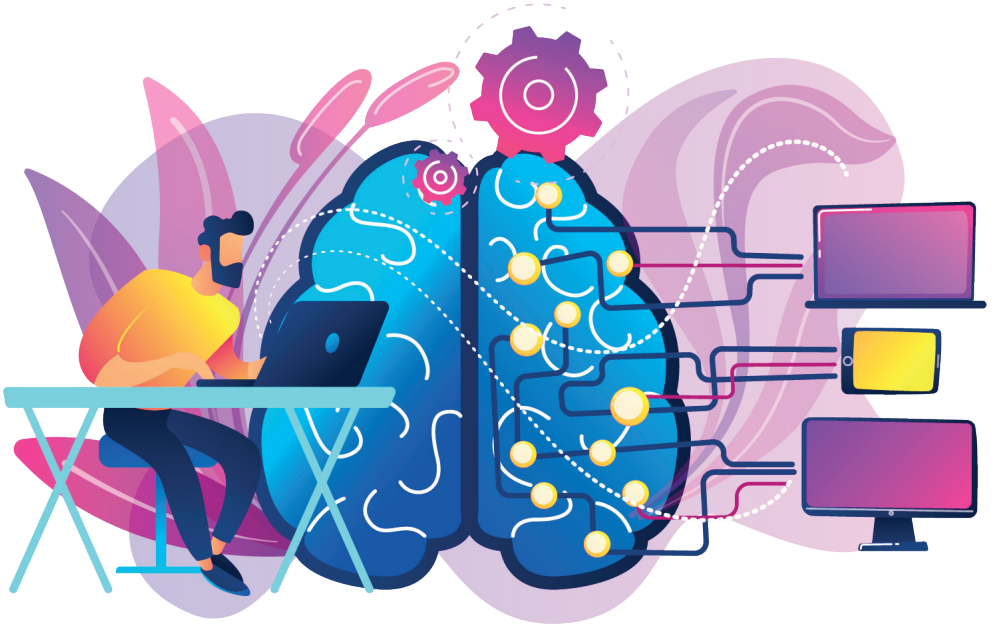
STEP 4 : การใช้ประโยชน์จาก RFM Segmentation (Benefits)

หลังจากการแบ่งกลุ่มลูกค้าตามแนวคิดของ RFM แล้ว จะได้กลุ่มตั้งแต่ 111 จนถึง 555 คิดเป็น 125 กลุ่ม ซึ่งในการใช้งานเราไม่จำเป็นต้องพิจารณาทุกกลุ่ม แต่เลือกกลุ่มที่น่าสนใจมาพิจารณา เช่น กลุ่ม 555 อาจจะเรียกว่า “กลุ่ม Champion” หรือ “กลุ่ม VIP” เป็นลูกค้ากลุ่มที่ซื้อสินค้าเมื่อไม่นานมานี้ ซื้อสินค้าบ่อย และซื้อสินค้ามูลค่าสูงอีกด้วย บริษัทจึงควรพิจารณาแคมเปญที่เหมาะสำหรับลูกค้ากลุ่มนี้เป็นพิเศษ และควรรักษาลูกค้ากลุ่ม VIP เหล่านี้เอาไว้ให้นานที่สุด

นอกจากนี้ ยังมีแนวทางในการตั้งชื่อกลุ่มต่างๆ เช่น ในรูปที่ 5 - 51 ที่มีการแบ่งกลุ่มลูกค้าออกเป็น 12 กลุ่ม สามารถอ่านรายละเอียดเพิ่มเติมได้จากเว็บไซต์ **“การตลาดวันละตอน”** ในหัวข้อ **“12 Customer Segments จาก RFM Model ที่นักการตลาดยุคดาต้า 5.0 ต้องรู้”**

รูปที่ 5 - 51 กลุ่มลูกค้าหลังจากแบ่งกลุ่มด้วยวิธี RFM Segmentation

เครดิต : https://www.everydaymarketing.co/target/mass/12-strategy-for-12-customer-segments-from-rfm-model/

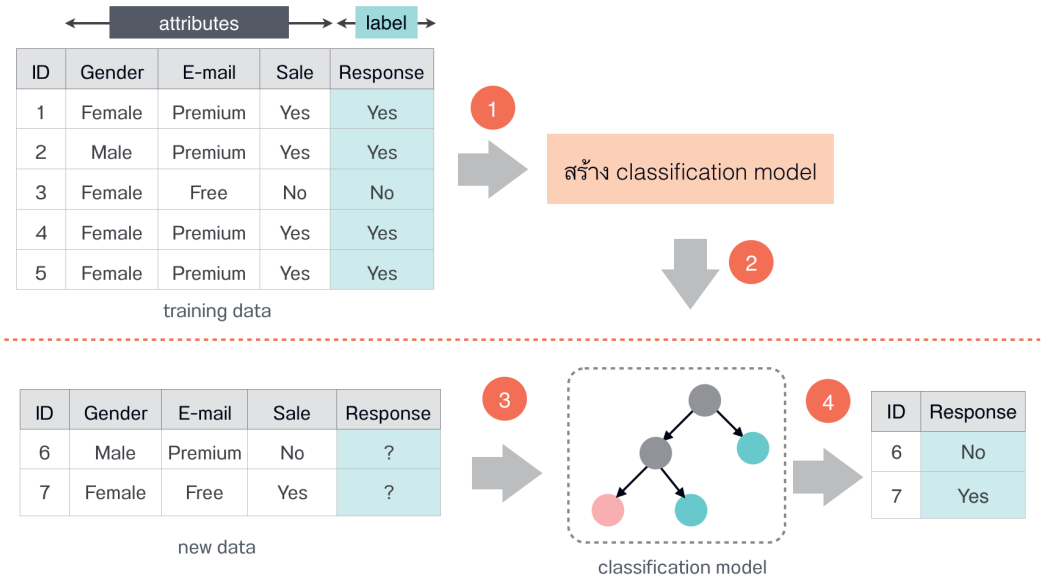


การสร้าง MACHINE LEARNING ด้วย RAPIDMINER AUTO MODEL

ในบทนี้จะอธิบายสรุปแต่ละเทคนิคแล้วทำ Workshop ด้วย RapidMiner ของเทคนิคนั้นๆ โดยจะเริ่มจาก Machine Learning ขั้นพื้นฐาน การสร้างโมเดลด้วยเทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยใช้ RapidMiner Auto Model และในหนังสือเล่มถัดไป จะอธิบายถึงการวิเคราะห์ข้อมูลด้วยเทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ที่มีความซับซ้อนมากขึ้น

แนวคิดของการเรียนรู้แบบมีผู้สอน

การเรียนรู้แบบมีผู้สอน เป็นการเรียนรู้จากข้อมูลที่มีคำตอบ/ลาเบล (Label)/คลาส (Class) อยู่แล้ว ซึ่งเป็นข้อมูลในอดีต หรือให้ผู้เชี่ยวชาญเป็นผู้กำหนดลาเบลเหล่านี้ให้ ซึ่งข้อมูลนี้เรียกว่า ข้อมูลฝึกสอน (Training Data) แนวคิดการทำงานของเทคนิคในกลุ่มนี้ จะเริ่มจากการส่งข้อมูลฝึกสอนไปให้คอมพิวเตอร์เรียนรู้ และได้เป็นเงื่อนไข (Rules) หรือสมการทางคณิตศาสตร์ (Math Equation) ที่ช่วยในการพยากรณ์ค่าสำหรับข้อมูลใหม่ต่อไปได้ ซึ่งสามารถสรุปได้ดังรูปที่ 6 - 3



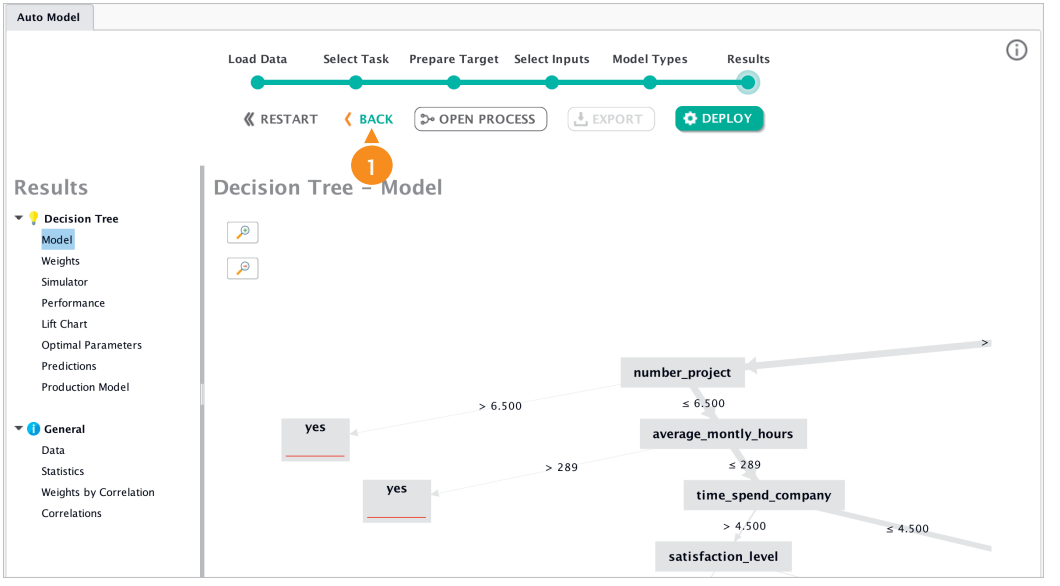
รูปที่ 6 - 3 แนวคิดของการเรียนรู้แบบมีผู้สอน

การแก้ปัญหา Imbalanced Data ทำได้หลายวิธี เช่น

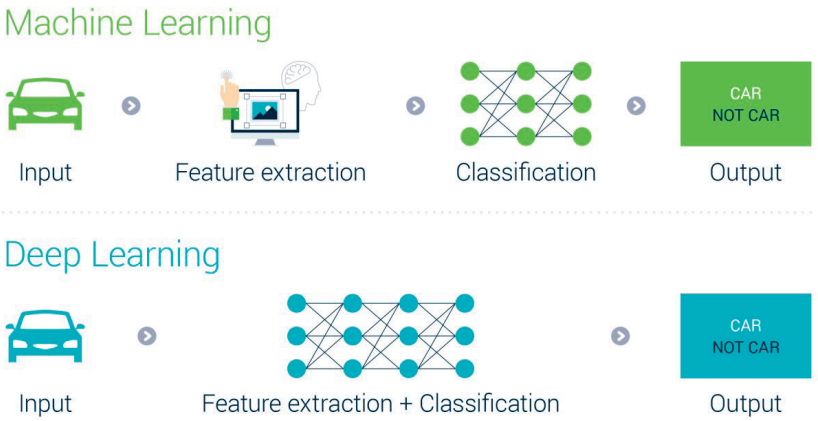
- **การทำ Under Sampling :** เป็นการสุ่มข้อมูลที่เป็น Majority Class ลดลงมาให้มีจำนวนใกล้เคียงกับกลุ่มที่เป็น Minority Class
- **การทำ Over Sampling :** เป็นการสร้างข้อมูลที่เป็น Minority Class ให้มีปริมาณมากขึ้น ซึ่งข้อมูลใหม่นี้อาจจะอยู่ใกล้เคียงกับข้อมูลเดิม (เรียกว่า Synthetic Minority Oversampling Technique หรือเรียกสั้นๆ ว่า SMOTE)
- **การสร้างโมเดลด้วยวิธี Cost Sensitive Learning :** โดยจะกำหนดค่า Cost หรือน้ำหนักให้กับข้อมูลต่างๆ โดยค่า Cost นี้จะเป็นตัวที่เป็น Penalty เวลาสร้างโมเดล นั่นหมายความว่าโมเดลจะพยายามให้ค่าพยากรณ์คำตอบนี้ผิดน้อยลง

สำหรับ Workshop นี้จะแสดงตัวอย่างการใช้ Auto Model ใน RapidMiner แก้ปัญหา Imbalanced Data ด้วยวิธี Cost Sensitive Learning เป็นหลัก โดยมีขั้นตอนการใช้งานดังนี้

1. คลิกปุ่ม **BACK** เพื่อย้อนขั้นตอนกลับไปที่ขั้นตอน **Prepare Target** ดังรูปที่ 6 – 38



รูปที่ 6 – 38 คลิกปุ่ม BACK เพื่อย้อนขั้นตอนกลับไปที่ขั้นตอน Prepare Target

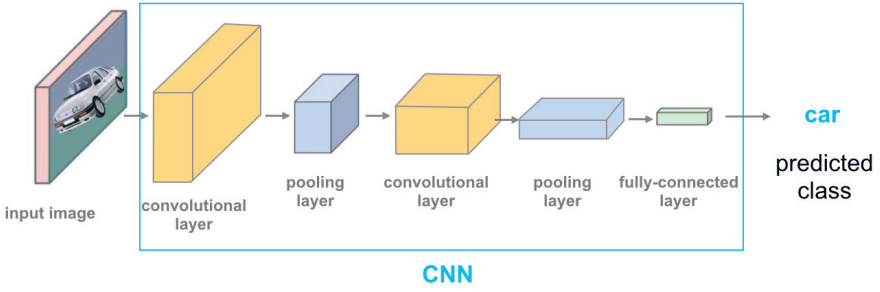


รูปที่ 6 - 128 แนวคิดของ Machine Learning และ Deep Learning

เครดิต : https://www.linkedin.com/pulse/lets-understand-difference-between-machine-learning-vs-gauri-bapat/

ใน RapidMiner Auto Model จะเป็นเทคนิค Deep Learning ที่มีสถาปัตยกรรมแบบ Convolutional Neural Network (CNN) โดยมีรายละเอียดแนวคิดดังต่อไปนี้

เป็นวิธีการพื้นฐานของการเรียนรู้เชิงลึกที่ใช้ในการจำแนกประเภทของรูปภาพ โดยจะรับข้อมูลรูปภาพเข้ามาและแปลงให้กลายเป็นตาราง หรือเมทริกซ์ (Matrix) ก่อนส่งไปประมวลผลในชั้นถัดๆ ไป จนได้คำตอบในชั้นท้ายสุด ดังแสดงในรูปที่ 6 - 129



รูปที่ 6 - 129 แนวคิดของโครงข่ายประสาทเทียมแบบ Convolutional Neural Network (CNN)

เครดิต : https://cezannec.github.io/Convolutional_Neural_Networks/