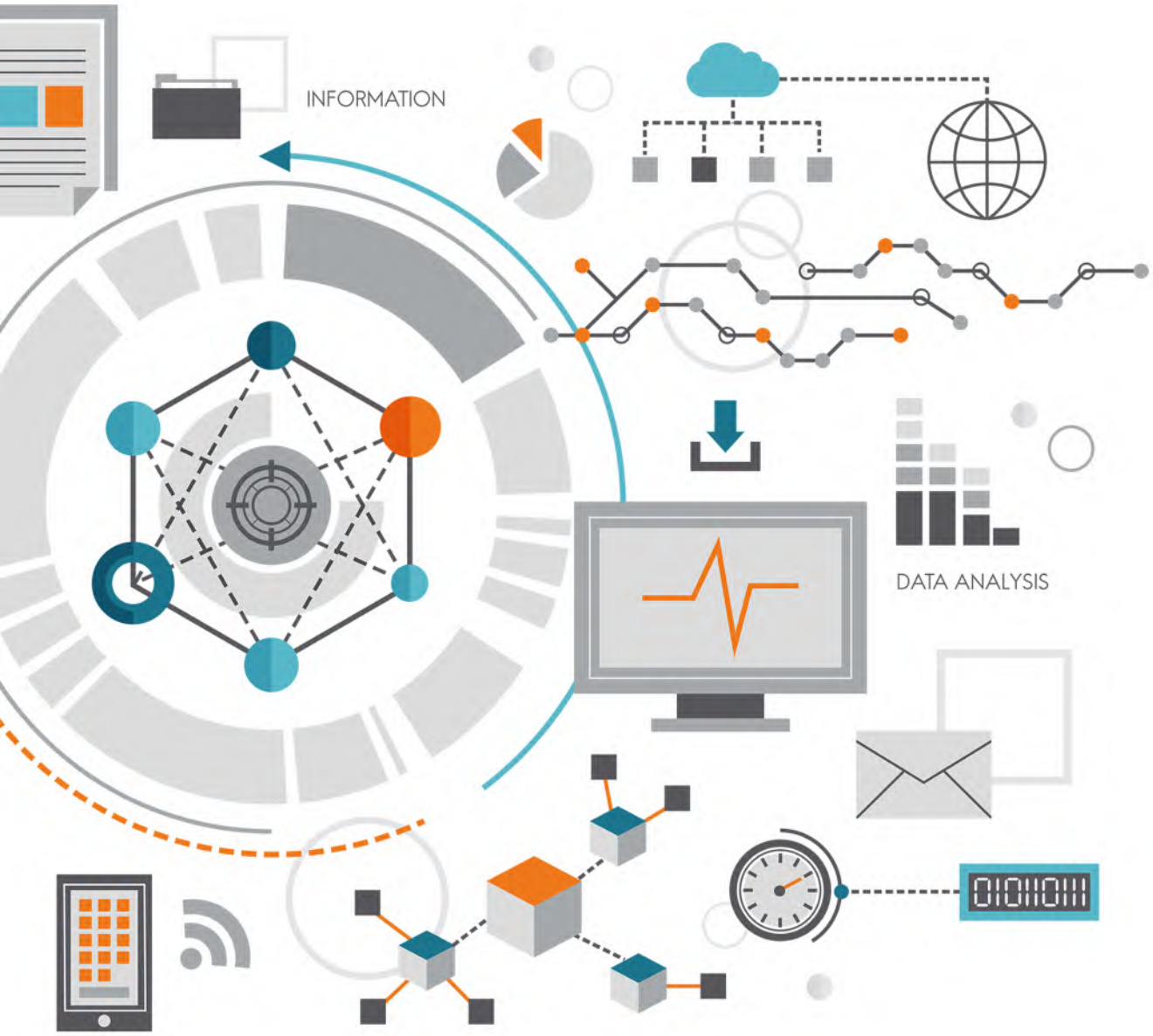
คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

การวิเคราะห์ข้อมูล

DATA ANALYTICS



ผศ.ดร.รัฐชัย ชาวอุทัย

การวิเคราะห์ข้อมูล

(Data Analytics)

ผู้ช่วยศาสตราจารย์ ดร. รัฐชัย ชาวอุทัย

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง



KMITL SCHOOL OF ENGINEERING

eBook STORE

ชื่อหนังสือ

ภาษาไทย : การวิเคราะห์ข้อมูล

ภาษาอังกฤษ : Data Analytics

รหัสวิชา : 01076598

ชื่อผู้แต่ง : ผู้ช่วยศาสตราจารย์ ดร.รัฐชัย ชาวอุทัย

ประเมินคุณภาพตำราเรียน : ผ่าน

จัดพิมพ์ประเภท : หนังสือและตำราเรียน

พิมพ์ครั้งที่ : 3 **เดือน** มกราคม 2566 **จำนวน** 100 เล่ม

รวมจำนวนการจัดพิมพ์ทั้งหมด : 300 เล่ม

ISBN : 978-616-338-169-9

วศ.สจล. : 294

จำนวนหน้า : 289 หน้า

จัดพิมพ์ประเภท : หนังสืออิเล็กทรอนิกส์ (e-book)

พิมพ์ครั้งที่ : 1 **เดือน** กุมภาพันธ์ 2567

ISBN : 5527300002510

วศ.สจล. e-Book : 001

จำนวนหน้า : 277 หน้า

ที่ปรึกษาฝ่ายผลิต

ผู้ช่วยศาสตราจารย์ ดร.อนรรฆพล แสนทน

ผู้ช่วยศาสตราจารย์ ดร. พิณิช ธนชัยโชคศิริกุล

นายชาคริต เทียนทอง

ประสานการจัดพิมพ์

น.ส.สุชาดา แดงอินทวัฒน์

ออกแบบรูปเล่ม

นายทนงศักดิ์ ใจซื่นแสน

จัดทำโดย : งานเทคโนโลยีการศึกษา คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เลขที่ 1 ซอย ฉลองกรุง 1 แขวงลาดกระบัง เขตลาดกระบัง กรุงเทพมหานคร 10520

โทร. 02-329-8301 (ต่อ 234)

จัดจำหน่ายโดย : ซีเอ็ดบุ๊ค (SE-ED)

พิมพ์ที่ : ห้างหุ้นส่วนจำกัด มีน เซอร์วิส ซัพพลาย

88/8 ถนนฉลองกรุง แขวงลำปลาทิว เขตลาดกระบัง กรุงเทพฯ 10520

โครงการส่งเสริมการผลิตหนังสือและตำราเพื่อความเป็นเลิศทางวิชาการ

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

สงวนลิขสิทธิ์ตามพระราชบัญญัติการพิมพ์

“

หากคุณค่าจากหนังสือเล่มนี้ สามารถเป็นบันไดอีกขั้น
ให้ผู้อ่านได้ใช้ประโยชน์จากการวิเคราะห์ข้อมูล
มาพัฒนาตนเอง องค์กร สังคม บ้านเมือง หรือสิ่งแวดล้อม ให้ดีขึ้น
ผู้เขียนขออุทิศความดีดังกล่าว ให้ครูบาอาจารย์ ครอบครัว มิตรสหาย
และผู้มีพระคุณต่อผู้เขียนทุกท่าน

”

คำนำ

จากคำพูดยอดนิยม “ข้อมูลเป็นแหล่งน้ำมันใหม่” (Data is the new oil.) ของ ไคลฟ์ ฮัมบี้ นักคณิตศาสตร์ชาวอังกฤษและผู้ประกอบการด้านการวิเคราะห์ข้อมูล เป็นคำพูดที่สะท้อนความจริงของการแข่งขันทางธุรกิจในยุคปัจจุบัน ซึ่งการแข่งขันเดิมคือยุทธศาสตร์ด้านแหล่งน้ำมัน แต่จากนี้ไปคือการถือครองสินทรัพย์ที่เป็นข้อมูลที่สามารถนำไปต่อยอดสร้างมูลค่าให้กับองค์กร ซึ่งหลายองค์กรระดับโลกที่ประสบความสำเร็จด้านเทคโนโลยี เช่น กูเกิล (Google), เฟสบุ้ค (Facebook), เน็ตฟลิกซ์ (Netflix) เป็นต้น ล้วนแต่สร้างรายได้เป็นกอบเป็นกำจากการถือครองข้อมูลมหาศาล ดังนั้นไม่ว่าใครก็อยากได้ข้อมูล แต่การมีข้อมูลเพียงอย่างเดียวนั้นไม่พอ ต้องนำมาวิเคราะห์ประมวลจนได้ผลลัพธ์ที่เป็นผลิตภัณฑ์ที่สร้างมูลค่าหรือสามารถสนับสนุนให้องค์กรตัดสินใจกำหนดทิศทางและกลยุทธ์ได้ถูกต้อง ก็จะเกิดมูลค่าอย่างสูงกับองค์กร

ด้วยตระหนักในความสำคัญดังที่กล่าวมานี้ หนังสือเล่มนี้จึงได้จัดทำขึ้นเพื่อถ่ายทอดความรู้ด้านการวิเคราะห์ข้อมูล โดยผู้เขียนได้ใช้ความรู้จากหนังสือ ตำรา และบทความวิชาการ ทั้งในและต่างประเทศ รวมถึงประสบการณ์การสอนวิชาด้านการวิเคราะห์ข้อมูลและปัญญาประดิษฐ์ การทำงานวิจัย และการให้บริการทางวิชาการ มาตกผลึกและเรียบเรียงเนื้อหาโดยการดำเนินเรื่องราวตามกระบวนการวิทยาศาสตร์ข้อมูล ตั้งแต่การรวบรวมข้อมูล, การจัดการประมวลผลข้อมูลให้พร้อมใช้, การสำรวจพฤติกรรมของข้อมูล, การวิเคราะห์ข้อมูลสร้างโมเดลเชิงพยากรณ์ วิเคราะห์จัดกลุ่ม และระบบแนะนำ ที่มีการใช้เทคนิคด้านปัญญาประดิษฐ์มาเป็นส่วนสำคัญ พร้อมทั้งการวัดประสิทธิภาพของโมเดล, และการนำผลการวิเคราะห์ไปใช้งาน รวมถึงกรณีศึกษาที่เป็นงานวิจัยของผู้เขียน โดยเนื้อหามีทั้งทฤษฎีและแนวทางการเขียนโปรแกรมภาษาไพธอน เพื่อให้ผู้อ่านสามารถเรียนรู้และฝึกปฏิบัติจนสามารถปฏิบัติงานจริงได้

สุดท้ายนี้ผู้เขียนจึงขอส่งมอบองค์ความรู้นี้ผ่านหนังสือเล่มนี้ สำหรับผู้ที่ต้องการนำศาสตร์ด้านการวิเคราะห์ข้อมูลไปใช้จริง ไม่ว่าจะเป็นด้านธุรกิจ หรือเป็นแหล่งอ้างอิงสำหรับงานวิจัย หรือนำไปพัฒนาสิ่งแวดล้อมให้ดีขึ้น อย่างไรก็ดีศาสตร์ด้านการวิเคราะห์ข้อมูลนี้มีความก้าวหน้าอยู่ตลอดเวลา ผู้เขียนยินดีน้อมรับความคิดเห็นจากผู้อ่านเพื่อปรับปรุงเนื้อหาให้ทันสมัยมากขึ้น

รัฐชัย ชาวอุทัย

สารบัญ

คำนำ **5**

สารบัญ **7**

สารบัญรูปภาพ **11**

สารบัญตาราง **15**

1. บทนำ **17**

1.1. การวิเคราะห์ข้อมูล 19

1.2. ระดับการวิเคราะห์ข้อมูล 30

1.3. กระบวนการของวิทยาศาสตร์ข้อมูล 33

1.4. ขอบเขตและโครงสร้างของหนังสือเล่มนี้ 36

1.5. สรุปท้ายบท 38

1.6. คำถามท้ายบท 39

1.7. เอกสารอ้างอิง 39

2. การจัดเก็บข้อมูล **41**

2.1. อินเทอร์เน็ตของสรรพสิ่ง 42

2.2. ประเภทข้อมูล 43

2.3. ระดับข้อมูล 47

2.4. การรวบรวมข้อมูล 49

2.5. ตัวอย่างข้อมูลที่ใช้ในหนังสือเล่มนี้ 52

2.6. สรุปท้ายบท 57

2.7. คำถามท้ายบท 58

2.8. เอกสารอ้างอิง 58

3. การจัดการประมวลผลข้อมูล **61**

3.1. เหตุผลที่ต้องจัดการข้อมูล 62

3.2. แนวทางการทำความสะอาดข้อมูล 63

3.3. แนวทางการแปลงรูปข้อมูล 68

3.4. การใช้ไพธอนสำหรับการวิเคราะห์ข้อมูล 72

3.5. โครงสร้างข้อมูลของไพธอน 75
3.6. การใช้ไลบรารีนัมไพ 78
3.7. การใช้ไลบรารีแพนดาส 80
3.8. สรุปท้ายบท 93
3.9. คำถามท้ายบท 94
3.10. เอกสารอ้างอิง 94

4. การสำรวจข้อมูลเบื้องต้น 97
4.1. ทบทวนฟังก์ชันสถิติพื้นฐาน 98
4.2. การสำรวจการกระจายของข้อมูล 101
4.3. ความสัมพันธ์ของข้อมูล 104
4.4. ตัวอย่างการสำรวจข้อมูล 107
4.5. สรุปท้ายบท 113
4.6. คำถามท้ายบท 113
4.7. เอกสารอ้างอิง 114

5. แนวทางการวิเคราะห์ข้อมูล 115
5.1. โมเดล 116
5.2. การเลือกเทคนิคสำหรับสร้างโมเดล 118
5.3. แนวทางการวัดผลโมเดล 122
5.4. กระบวนการสร้างโมเดล 126
5.5. สรุปท้ายบท 130
5.6. คำถามท้ายบท 131
5.7. เอกสารอ้างอิง 131

6. การวิเคราะห์การถดถอย 133
6.1. สมการถดถอยเชิงเส้น 134
6.2. การวัดผลการวิเคราะห์การถดถอย 138
6.3. การเลือกตัวแปรต้นหรือฟีเจอร์ 147
6.4. การปรับขนาดช่วงข้อมูล 149
6.5. ขั้นตอนการพัฒนาโมเดลสมการถดถอยเชิงเส้น 151
6.6. สรุปท้ายบท 156
6.7. คำถามท้ายบท 156

6.8. เอกสารอ้างอิง 157

7. การวิเคราะห์จำแนกประเภท 159

7.1. รู้จักการวิเคราะห์จำแนกประเภท 160

7.2. การวัดผลการวิเคราะห์จำแนกประเภท 161

7.3. ขั้นตอนการวิเคราะห์จำแนกประเภท 174

7.4. เทคนิคการวิเคราะห์จำแนกประเภท 177

7.5. การเปรียบเทียบประสิทธิภาพของโมเดล 198

7.6. สรุปท้ายบท 203

7.7. คำถามท้ายบท 204

7.8. เอกสารอ้างอิง 205

8. การวิเคราะห์จัดกลุ่ม 207

8.1. รู้จักการวิเคราะห์จัดกลุ่ม 208

8.2. การประเมินผลการวิเคราะห์จัดกลุ่ม 209

8.3. เทคนิคที่ใช้สำหรับการวิเคราะห์จัดกลุ่ม 210

8.4. สรุปท้ายบท 222

8.5. คำถามท้ายบท 222

8.6. เอกสารอ้างอิง 223

9. ระบบแนะนำ 225

9.1. รู้จักกับระบบแนะนำ 227

9.2. การวัดผลระบบแนะนำ 231

9.3. การพัฒนาระบบแนะนำด้วยวิธีพิจารณาการกรองร่วมกัน 236

9.4. การพัฒนาระบบแนะนำแบบผสม 241

9.5. สรุปท้ายบท 244

9.6. คำถามท้ายบท 245

9.7. เอกสารอ้างอิง 245

10. การนำเสนอข้อมูลเชิงภาพ 247

10.1. การใช้ภาพสื่อความหมายข้อมูล 248

10.2. การรับรู้ทางการมองเห็นของมนุษย์ 250

10.3. การแปลงข้อมูลให้เป็นภาพ 255

10.4. แนวทางการนำเสนอข้อมูลด้วยภาพ 258

10.5. สรุปท้ายบท 261
10.6. คำถามท้ายบท 261
10.7. เอกสารอ้างอิง 262

11. กรณีศึกษางานวิจัยด้านการวิเคราะห์ข้อมูล 263
11.1. การตรวจจับถนนชำรุดด้วยเซ็นเซอร์ไจโร 264
11.2. การตรวจใช้เซ็นเซอร์ไมโครเวฟสำหรับตรวจจับการเตะ 266
11.3. ระบบแนะนำนักชีววิทยาเพื่อค้นหาคู่การเกิดเห็ดราบนพืชและสัตว์ 269
11.4. ระบบแนะนำจุดจอดพักสำหรับรถ 271
11.5. สรุปท้ายบท 274
11.6. คำถามท้ายบท 275
11.7. เอกสารอ้างอิง 275

12. บทสรุป 277

อภิธานศัพท์ 281

บรรณานุกรม 283

ดัชนีคำศัพท์ 285

ประวัติผู้เขียน 287

สารบัญรูปภาพ

รูปที่ 1.1-1 : ทักษะของงานวิทยาศาสตร์ข้อมูล 29
รูปที่ 1.2-1 : แผนภาพแสดงระดับการวิเคราะห์ข้อมูล 31
รูปที่ 1.3-1 : แผนภาพกระบวนการวิทยาศาสตร์ข้อมูล 33
รูปที่ 2.0-1 : ขั้นตอนการจัดเก็บข้อมูลในกระบวนการวิทยาศาสตร์ข้อมูล 42
รูปที่ 2.1-1 : แผนภาพทั่วไปของสถาปัตยกรรมอินเทอร์เน็ตของสรรพสิ่ง 43
รูปที่ 2.2-1 : รูปภาพแสดงตัวอย่างข้อมูลอายุและส่วนสูงของนักเรียน 45
รูปที่ 2.2-2 : คลื่นเสียงสมมติตัวอย่างข้อมูลอายุและส่วนสูงของนักเรียน 46
รูปที่ 3.0-1 : ขั้นตอนการจัดการประมวลผลข้อมูลในกระบวนการวิทยาศาสตร์ข้อมูล 62
รูปที่ 3.2-1 : คะแนนมาตรฐานในส่วนโค้งปกติของข้อมูล 66
รูปที่ 3.3-1 : ตัวอย่างการแปลงข้อมูลลอการิทึม 68
รูปที่ 3.3-2 : ตัวอย่างการเข้ารหัสข้อมูล 69
รูปที่ 3.3-3 : ตัวอย่างการเข้ารหัสข้อมูลแยกคอลัมน์ 69
รูปที่ 3.3-4 : ตัวอย่างการระบุช่วงข้อมูล 70
รูปที่ 3.3-5 : ตัวอย่างการปรับขนาดช่วงข้อมูลให้ปกติตามค่าต่ำสุดสูงสุด 71
รูปที่ 3.3-6 : ตัวอย่างการตัดแบ่งข้อมูล 71
รูปที่ 3.4-1 : หน้าจอของกูเกิลโคแล็ปบนเว็บเบราเซอร์ 73
รูปที่ 3.4-2 : หน้าจอของอะนาคอนดาเนวิเกเตอร์ บนเครื่องคอมพิวเตอร์ 74
รูปที่ 3.4-3 : หน้าจอของแอปพลิเคชันสปายเดอร์บนเครื่องคอมพิวเตอร์ 75
รูปที่ 4.0-1 : ขั้นตอนการสำรวจข้อมูลเบื้องต้นในกระบวนการวิทยาศาสตร์ข้อมูล 98
รูปที่ 4.2-1 : ตัวอย่างแผนภูมิฮิสโทแกรมของส่วนสูงนักเรียนโรงเรียนแห่งหนึ่ง 101
รูปที่ 4.2-2 : รูปแบบแผนภูมิรูปกล่อง 103
รูปที่ 4.3-1 : รูปแบบค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน 104
รูปที่ 4.3-2 : ตัวอย่างแผนภูมิสแกตเตอร์แสดงความสัมพันธ์ระหว่างน้ำหนักกับส่วนสูง 106
รูปที่ 5.0-1 : ขั้นตอนการสร้างโมเดลในกระบวนการวิทยาศาสตร์ข้อมูล 116
รูปที่ 5.2-1 : โครงสร้างโจทย์การทำงานด้านวิเคราะห์ข้อมูล 119
รูปที่ 5.3-1 : แนวทางการแบ่งข้อมูลเพื่อการทดสอบ 123
รูปที่ 5.3-2 : แนวทางการวัดผลด้วยวิธีการแบ่งส่วนทดสอบหลายรอบ (ที่ K=3) 125

รูปที่ 5.4-1 : แผนภาพแสดงกระบวนการสร้างโมเดล 127
รูปที่ 6.0-1 : การวิเคราะห์ถดถอยในกระบวนการวิทยาศาสตร์ข้อมูล 134
รูปที่ 6.1-1 : แผนภูมิสแกตเตอร์ของตัวอย่างข้อมูลน้ำหนักและส่วนสูงของนักเรียน 136
รูปที่ 6.1-2 : ตัวอย่างโมเดลสมการถดถอยเชิงเส้นสำหรับพยากรณ์ส่วนสูงของนักเรียน 137
รูปที่ 7.0-1 : การวิเคราะห์จำแนกประเภทในกระบวนการวิทยาศาสตร์ข้อมูล 160
รูปที่ 7.2-1 : แผนภาพลำดับการหาค่าวัดผลวิเคราะห์จำแนกประเภท 162
รูปที่ 7.2-2 : ตัวอย่างการอธิบายเมทริกซ์คอนฟิวชัน 165
รูปที่ 7.2-3 : ตัวอย่างการอธิบายค่าพรีซิชันและรีคอล 169
รูปที่ 7.4-1 : กราฟจากสมการถดถอยโลจิสติก 178
รูปที่ 7.4-2 : แผนภาพตัวอย่างโมเดลต้นไม้การตัดสินใจ 182
รูปที่ 7.4-3 : แผนภาพตัวอย่างการทำงานของเทคนิคเคเนียร์เรสเนเบอร์ 192
รูปที่ 7.4-4 : แผนภาพโครงสร้างของเพอร์เซปตรอน 194
รูปที่ 7.4-5 : ตัวอย่างฟังก์ชันกระตุ้น 195
รูปที่ 7.4-6 : แผนภาพโครงสร้างของโครงข่ายประสาทเทียม 196
รูปที่ 8.0-1 : การวิเคราะห์จัดกลุ่มในกระบวนการวิทยาศาสตร์ข้อมูล 208
รูปที่ 8.1-1 : ตัวอย่างการจัดกลุ่ม 209
รูปที่ 8.3-1 : ตัวอย่างการจัดกลุ่มด้วยเทคนิคเคมีนส์ 211
รูปที่ 8.3-2 : ตัวอย่างการจัดกลุ่มที่ซับซ้อนขึ้น 219
รูปที่ 8.3-3 : หลักการทำงานของเทคนิคดีบีสแกน 220
รูปที่ 8.3-4 : ตัวอย่างผลวิเคราะห์การจัดกลุ่มด้วยเทคนิคดีบีสแกน 220
รูปที่ 9.0-1 : ขอบเขตระบบแนะนำในกระบวนการวิทยาศาสตร์ข้อมูล 226
รูปที่ 9.1-1 : ตัวอย่างการแนะนำสินค้าในระบบพานิชย์อิเล็กทรอนิกส์ 228
รูปที่ 9.1-2 : ภาพรวมของระบบแนะนำ 228
รูปที่ 9.1-3 : ตัวอย่างการแนะนำไอเท็มตามความเหมือนของข้อมูลผู้ใช้ 229
รูปที่ 9.1-4 : ตัวอย่างการแนะนำไอเท็มตามพฤติกรรมร่วมกันของการซื้อไอเท็ม 229
รูปที่ 9.1-5 : ตัวอย่างการแนะนำไอเท็มตามความคล้ายของข้อมูลไอเท็ม 230
รูปที่ 9.1-6 : ตัวอย่างการแนะนำไอเท็มตามองค์ความรู้ 231
รูปที่ 9.3-1 : ตัวอย่างพฤติกรรมการให้คะแนนไอเท็มของผู้ใช้ 239
รูปที่ 9.4-1 : ตัวอย่างระบบแนะนำของแอปพลิเคชันวิดีโอแห่งหนึ่ง 244
รูปที่ 10.0-1 : ขั้นตอนการนำเสนอข้อมูลด้วยภาพในกระบวนการวิทยาศาสตร์ข้อมูล 248
รูปที่ 10.1-1 : แผนภูมิเส้นแสดงยอดขายรายเดือนของแต่ละผลิตภัณฑ์ 249

รูปที่ 10.1-2 : แผนภูมิแท่งซ้อนสามมิติแสดงยอดขายรายเดือนของแต่ละผลิตภัณฑ์ 250
รูปที่ 10.2-1 : การมองเห็นของขนาด 251
รูปที่ 10.2-2 : ตัวอย่างแผนภูมิที่ใช้ขนาด 251
รูปที่ 10.2-3 : ตัวอย่างแผนภูมิแท่งแสดงคะแนนภาพยนตร์ 252
รูปที่ 10.2-4 : การมองเห็นตำแหน่ง และตัวอย่างแผนภูมิ 252
รูปที่ 10.2-5 : การใช้ตำแหน่งบนแผนที่ 253
รูปที่ 10.2-6 : หลักการของการใช้สี 254
รูปที่ 10.4-1 : ตัวอย่างการนำเสนอข้อมูลเชิงภาพแสดงสถิติการใช้เว็บไซต์หนึ่ง 258
รูปที่ 11.1-1 : ตำแหน่งที่ตรวจพบถนนชำรุดจากโมเดลต้นไม้การตัดสินใจที่พัฒนาขึ้น 266
รูปที่ 11.2-1 : สัญญาณที่ได้รับจากเซ็นเซอร์ไมโครเวฟตามกิจกรรมต่าง ๆ 267
รูปที่ 11.4-1 : แผนภาพแสดงลักษณะการเลือกจุดจอดถัดไป 272
รูปที่ 12.0-1 : ขั้นตอนการวิเคราะห์ข้อมูลในกระบวนการวิทยาศาสตร์ข้อมูล 278

สารบัญตาราง

ตาราง 2.2-1 : ตารางแสดงตัวอย่างข้อมูลอายุและส่วนสูงของนักเรียน....44
ตาราง 2.5-1 : ตารางแสดงคำอธิบายชุดข้อมูลพนักงาน....53
ตาราง 2.5-2 : ตารางแสดงข้อมูลพนักงาน....53
ตาราง 2.5-3 : ตารางแสดงคำอธิบายชุดข้อมูลประเภทรถอย่างง่าย....54
ตาราง 2.5-4 : ตารางแสดงข้อมูลประเภทรถอย่างง่าย....54
ตาราง 2.5-5 : ตารางแสดงคำอธิบายชุดข้อมูลการวินิจฉัยไข้หวัดใหญ่....55
ตาราง 2.5-6 : ตารางแสดงข้อมูลการวินิจฉัยไข้หวัดใหญ่....56
ตาราง 2.5-7 : ตารางแสดงคำอธิบายชุดข้อมูลราคาบ้านในเมืองบอสตัน....56
ตาราง 2.5-8 : ตารางแสดงตัวอย่างข้อมูลราคาบ้านในเมืองบอสตัน....57
ตาราง 3.2-1 : ตารางแสดงตัวอย่างข้อมูลที่ไม่พร้อมใช้....63
ตาราง 3.9-1 : ตารางตัวอย่างข้อมูลพนักงาน....94
ตาราง 5.2-1 : ตารางสรุปสำหรับระบุโจทย์การวิเคราะห์ข้อมูล....121
ตาราง 6.1-1 : ตารางแสดงตัวอย่างน้ำหนักและความสูงของนักเรียน 4 คน....136
ตาราง 6.2-1 : ตารางแสดงตัวอย่างการพยากรณ์ด้วยสมการถดถอยเชิงเส้นสำหรับใช้วัดผล....140
ตาราง 6.7-1 : ข้อมูลสำหรับคำถามท้ายบท โจทย์การวิเคราะห์ถดถอยเชิงเส้น....157
ตาราง 7.2-1 : ตารางเมทริกซ์คอนฟิวชัน....164
ตาราง 7.2-2 : ตัวอย่างผลการพยากรณ์ของระบบนับรถยนต์....166
ตาราง 7.2-3 : ตารางเมทริกซ์คอนฟิวชัน....167
ตาราง 7.4-1 : ตัวอย่างการคำนวณความน่าจะเป็นแต่ละพจน์ของนาอีฟเบย์....187
ตาราง 8.3-1 : การนำเสนอขั้นตอนการจัดกลุ่มด้วยเทคนิคเคมีนส์....212
ตาราง 9.2-1 : ตัวอย่างผลลัพธ์รายการแนะนำไอเท็มสำหรับทดสอบ....231
ตาราง 9.3-1 : ตัวอย่างข้อมูลดิบการให้คะแนนไอเท็มของแต่ละคน....237
ตาราง 9.3-2 : เมทริกซ์ ผู้ใช้-ไอเท็ม แสดงข้อมูลคะแนน....237
ตาราง 9.4-1 : ตารางเปรียบเทียบเทคนิคการวิเคราะห์แนะนำแต่ละแบบ....242
ตาราง 10.1-1 : ยอดขายของแต่ละผลิตภัณฑ์ของบริษัทแห่งหนึ่งในปี 2020....248
ตาราง 10.3-1 : ตารางแสดงการรับรู้ของส่วนประกอบรูป....256
ตาราง 11.1-1 : ตัวอย่างข้อมูลเซ็นเซอร์สำหรับการวิเคราะห์ตรวจจับถนนชำรุด....265

ตาราง 11.1-2 : ผลการวิเคราะห์ตรวจจับถนนชำรุดด้วยค่า F1 266
ตาราง 11.2-1 : ตัวอย่างข้อมูลสัญญาณคลื่นไมโครเวฟตรวจจับการเตะเท้า 267
ตาราง 11.2-2 : ผลการวัดผลการวิเคราะห์ตรวจจับการเตะเท้า 268
ตาราง 11.3-1 : ตัวอย่างข้อมูลการพบเห็ดราบนเจ้าบ้าน 269
ตาราง 11.3-2 : ผลการทดสอบการวิเคราะห์ข้อมูลการพบเห็ดราบนเจ้าบ้าน 271
ตาราง 11.4-1 : ตัวอย่างข้อมูลจีพีเอสการวิ่งของรถ 272
ตาราง 11.4-2 : ตัวอย่างข้อมูลลำดับการจอดรถเพื่อนำไปวิเคราะห์ 273

บทที่

1. บทนำ

Introduction

“Data is the new oil.”

- Clive Humby

ในยุคนี้ ข้อมูลถือว่าเป็นผู้เล่นสำคัญของการพัฒนาแทบทุกภาคส่วน ไม่ว่าจะเป็น ด้านธุรกิจ, ด้านการศึกษา, ด้านการแพทย์, ด้านการทหาร, ด้านการคมนาคม, และอื่น ๆ ตัวอย่างเช่น

- ร้านสะดวกซื้อเซเว่นอีเลฟเว่น (7-Eleven) ในประเทศญี่ปุ่นทำการวิเคราะห์ยอดขายตามปัจจัย เช่น หนึ่งวันก่อนหน้า วันนี้เมื่อปีก่อนหน้า วันที่มีสภาพอากาศแบบเดียวกัน และสินค้าที่ขายในร้านอื่น ๆ เพื่อวางแผนว่าในวันต่อไปจะมีปริมาณการซื้อสินค้าในแต่ละสาขาเท่าใด เพื่อนำมาวางแผนการจัดการคลังสินค้า ส่งผลให้ร้านสะดวกซื้อนี้เป็นร้านค้าปลีกที่ทำกำไรดีที่สุดในญี่ปุ่นกว่า 30 ปี (Ross et al., 2013)

- บริษัทอะเมซอนดอทคอม (Amazon.com) บริษัทพาณิชน์อิเล็กทรอนิกส์รายใหญ่ของโลก เคยพบปัญหามีสินค้าให้เลือกจำนวนมากจนทำให้ลูกค้าใช้เวลาเลือกนานจนตัดสินใจซื้อลำบาก จึงพัฒนาระบบแนะนำสินค้าโดยวิเคราะห์จากพฤติกรรมการเรียกดูรายการสินค้า คาดเดารายได้จากทำเลที่อยู่อาศัย ทำให้สามารถแนะนำสินค้าได้แม่นยำจนมียอดขายเพิ่มขึ้น (Marr, 2021)

- ธนาคารกรุงศรีได้วิเคราะห์ข้อมูลลูกหนี้เพื่อพยากรณ์หาแนวโน้มการจ่ายเงินของลูกหนี้ ส่งผลให้เพิ่มประสิทธิภาพการทวงหนี้ และลดการทำงานของฝ่ายทวงหนี้ได้ถึงร้อยละ 20 (Muangtum, 2020)

- ทีมงานสปลังค์ (Splunk) ชนะเลิศการแข่งขัน Deutsche Bahn IoT Hackathon ในการแก้ปัญหาความผิดเพี้ยนของตำแหน่งรางรถไฟโดยคาดการณ์ตำแหน่งที่อาจจะเกิดขึ้นในอนาคต ทำให้ทางบริษัทสามารถระบุปัจจัยหลักที่ทำให้ตำแหน่งของรางคาดเคลื่อนได้จากวัสดุที่ใช้ในการก่อสร้าง โดยบริเวณที่มีการใช้ไม้จะมีค่าความเบี่ยงเบนที่สูงกว่าปกติ และเมื่อข้ามไปวิเคราะห์รางบริเวณที่ใช้คอนกรีตเป็นหลักพบว่ามีปัญหาน้อยกว่าไม้ มากกว่าร้อยละ 18 ซึ่งทำให้บริษัทมั่นใจว่าการลงทุนที่ใช้คอนกรีตแทนจะทำให้อายุการใช้งานของรางยืนยาวขึ้น และทำการวิเคราะห์ข้อมูลเพิ่มเติมถึงการพยากรณ์อนาคตว่าควรจะมีการแจ้งเตือนบริษัทล่วงหน้าว่าตำแหน่งไหนของรางที่มีโอกาสจะเกิดปัญหาขึ้นบ้าง ก็ทำให้บริษัท มีแนวโน้มในการบูรณะซ่อมแซมรางรถไฟหรือตรวจสอบสภาพได้เป็นอย่างดี (Drieger, 2015)

- แอปพลิเคชันแกร็บ (Grab) ในตลาดเอเชียตะวันออกเฉียงใต้ ได้วิเคราะห์ข้อมูลเพื่อเลิกการแสดงผลตามรายบุคคลตามปัจจัยดังนี้ ความยอดนิยมของผลิตภัณฑ์, ความชอบส่วนบุคคล, กลุ่มคนที่มีพฤติกรรมลักษณะเดียวกัน, พฤติกรรมจากการใช้แอปพลิเคชัน, และการแนะนำ ส่งผลให้แอปพลิเคชันนี้เป็นที่นิยมในการสั่งซื้ออาหาร สินค้าต่าง ๆ และบริการการเดินทาง ด้วยการแนะนำสินค้าอุปโภคและบริโภคได้ตรงตามความต้องการ (Justin, 2019)

- บริการเพลงของสป็อตติไฟ (Spotify) พบปัญหาการเลือกเพลงไม่ตรงใจลูกค้า จึงใช้ข้อมูลพฤติกรรมการเลือกฟังเพลงและประวัติการค้นหา จนได้ระบบแนะนำเพลงให้กับผู้ใช้งานแต่ละคน ซึ่งได้แนวเพลงตรงตามความต้องการฟังของผู้ใช้งาน และได้ผลตอบรับที่ดีขึ้นจากผู้ใช้งาน (Tyagi, 2020)

จากกรณีศึกษาที่ยกมา การวิเคราะห์ข้อมูลเป็นศาสตร์ในการหาคำตอบที่ต้องการจากข้อมูล ในทางปฏิบัติจะเริ่มโดยการกำหนดโจทย์ก่อน แล้วจึงทำการเลือกใช้ข้อมูล หลักจากนั้นจึงมีการใช้เทคนิคและเครื่องมือเพื่อหาคำตอบ และสุดท้ายเป็นการนำเสนอผลการวิเคราะห์เพื่อนำไปทำประโยชน์ต่อไป

โดยรายละเอียดในศาสตร์การวิเคราะห์ข้อมูลได้ดำเนินการอธิบายในหนังสือเล่มนี้ พร้อมด้วยการแสดงการเขียนโค้ดภาษาไพธอน (Python) ประกอบ และในบทนำนี้กจะกล่าวถึงการวิเคราะห์ข้อมูล ประเภทการวิเคราะห์ข้อมูล กระบวนการทางวิทยาศาสตร์ข้อมูล และโครงสร้างของหนังสือเล่มนี้

1.1. การวิเคราะห์ข้อมูล

เริ่มต้นในหัวข้อนี้เป็นการอธิบายถึงสิ่งที่ต้องทราบในศาสตร์ด้านการวิเคราะห์ข้อมูล เริ่มจากการทำความเข้าใจประโยชน์ของการวิเคราะห์ข้อมูล, การตั้งโจทย์ในการวิเคราะห์ข้อมูลซึ่งเป็นสิ่งสำคัญในการกำหนดทิศทางการวิเคราะห์ข้อมูลทั้งในเรื่องการเลือกใช้เทคนิคการวิเคราะห์และวิธีการวัดผล, และการตั้งทีมงานในการวิเคราะห์ข้อมูล

1.1.1. ประโยชน์ของการวิเคราะห์ข้อมูล

จากคำพูดที่ว่า “ข้อมูลคือแหล่งน้ำมันใหม่” (Data is the new oil.) คงเป็นคำพูดที่ไม่เกินความจริงในปัจจุบัน เนื่องจากการขับเคลื่อนทางอุตสาหกรรมล้วนแต่ต้องใช้พลังงานโดยเฉพาะอย่างยิ่งแหล่งน้ำมัน ดังนั้นใครที่มีแหล่งน้ำมันคือผู้ที่ได้เปรียบในการแข่งขันทางธุรกิจ แต่ในยุคนี้ความก้าวหน้าทางการขับเคลื่อนธุรกิจนั้นได้มีข้อมูลมาเป็นผู้เล่นสำคัญ จึงกล่าวได้ว่าผู้ใดที่มีข้อมูลคือผู้ที่ได้เปรียบในการแข่งขันทางธุรกิจเช่นกัน เนื่องจากการตัดสินใจหลาย ๆ อย่างในทิศทางทางธุรกิจต้องอาศัยข้อมูลเป็นสำคัญ ยิ่งมีข้อมูลมากและเป็นข้อมูลที่มีคุณภาพ จะสามารถนำมาวิเคราะห์เพื่อให้ได้ข้อสรุปที่สามารถสร้างทิศทางในการตัดสินใจได้ (Provost & Fawcett, 2013) โดยการคิดถึงประโยชน์หรือเป้าหมายของการวิเคราะห์ข้อมูลจะมีหลัก ๆ 3 ประการคือ การลดต้นทุน, การเพิ่มคุณค่า, และการสร้างนวัตกรรม

การลดต้นทุน (Cost Reduction)

คำว่าต้นทุน (หรือ cost) ในที่นี้คือทรัพยากร (resource) และเวลา (time) เช่น ถ้าทำแล้วสามารถรลดค่าใช้จ่ายหรือสามารถลดจำนวนการใช้แรงงานคนได้ถือว่าเป็นการลดการใช้ทรัพยากร หรือถ้าทำแล้วสามารถทำให้ทำงานได้เร็วขึ้นก็ถือว่าเป็นการลดเวลา ตัวอย่างเช่นกรณีของร้านสะดวกซื้อเซเว่นอีเลฟเว่น (7-Eleven) ในประเทศญี่ปุ่นที่กล่าวมาได้นำข้อมูลไปวิเคราะห์พยากรณ์ยอดขายสินค้าในแต่ละวัน ส่งผลให้สามารถลดต้นทุนในการกักเก็บสินค้าคงเหลือ เป็นต้น

นอกจากนี้การวิเคราะห์ข้อมูลเพื่อลดเวลาก็สามารถทำได้ เช่น กูเกิลแมพ (Google Map) ซึ่งเป็นแอปพลิเคชันแผนที่ของค่ายกูเกิล ที่มีระบบนำทาง จากเดิมถ้าไม่มีระบบนี้ผู้ขับขี่รถส่วนใหญ่ต้องเสียเวลาดูแผนที่ซึ่งเป็นหนังสือหรือแผ่นพับเพื่อวางแผนเส้นทาง แต่ด้วยศักยภาพของกูเกิลแมพที่นำข้อมูลแผนที่และข้อมูลการเคลื่อนที่ของอุปกรณ์มือถือทั่วโลกสามารถนำมาวิเคราะห์หาเส้นทางที่ดีที่สุดให้ผู้ใช้ด้วยเวลาที่รวดเร็ว ดังนั้นจึงเป็นการลดระยะเวลาในการเลือกหาเส้นทางของผู้ใช้ได้อย่างดี

การเพิ่มคุณค่า (Value Adding)

คำว่าคุณค่าในที่นี้ในทางธุรกิจหมายถึงผลประโยชน์ ซึ่งก็คือรายได้และกำไร และรวมไปถึงคุณค่าที่สร้างให้มนุษย์ สังคม หรือสิ่งแวดล้อม เป็นต้น เช่น การวิเคราะห์ข้อมูลผู้ใช้ของอะเมซอนดอทคอม (Amazon.com) ทำให้สามารถแนะนำสินค้าได้ตรงใจผู้ใช้ ส่งผลให้มียอดขายเพิ่มขึ้น, การวิเคราะห์ข้อมูลในการวินิจฉัยโรคจากค่าต่าง ๆ ที่ตรวจพบในร่างกาย จนสามารถตรวจพบแนวโน้มที่จะเป็นมะเร็งแต่เนิ่น ๆ ทำให้สามารถเตรียมการรักษาได้ทันท่วงที ซึ่งเป็นคุณค่าต่อมนุษย์และทางการแพทย์, และการวิเคราะห์สรุปข้อมูลมลพิษทางอากาศทำให้ช่วยในการตัดสินใจวางนโยบายด้านสิ่งแวดล้อมในหลายเมืองใหญ่ ถือเป็นการเพิ่มคุณค่าให้สิ่งแวดล้อม เป็นต้น

การสร้างนวัตกรรม (Innovation Creation)

นอกจากการวิเคราะห์ข้อมูลเพื่อเพิ่มคุณค่าหรือลดต้นทุนแล้ว บ่อยครั้งที่ผลการวิเคราะห์ทำให้พบปัญหาที่คาดไม่ถึง หรือเข้าใจความต้องการของผู้บริโภคมากขึ้น ทำให้เกิดผลิตภัณฑ์หรือสิ่งประดิษฐ์ใหม่ที่ตอบโจทย์ผู้ใช้งาน หรือจะเรียกได้ว่าการวิเคราะห์ข้อมูลสามารถทำให้เกิดนวัตกรรมขึ้นได้ ตัวอย่างเช่นการพัฒนารถยนต์ขับเคลื่อนด้วยตนเอง (Self-Driving Car) ซึ่งเป็นนวัตกรรมที่อาศัยการประมวลผลข้อมูลมหาศาล หลายปัจจัย หลายมิติข้อมูล และใช้เทคโนโลยีขั้นสูง เพื่อให้ได้นวัตกรรมที่ใช้งานได้จริง สร้างคุณค่าตอบโจทย์ผู้ใช้ และปลอดภัย

เนื่องจากการวิเคราะห์ข้อมูลในแต่ละครั้งต้องมีการลงทุนและลงแรงเป็นอย่างมาก เช่น การหาข้อมูล การนำเข้าข้อมูล การจัดเก็บข้อมูล การจัดการประมวลผลข้อมูล การวิเคราะห์ข้อมูล และการตรวจสอบความถูกต้อง ล้วนแต่ต้องอาศัยเงินทุนและกำลังคน ซึ่งถือว่าเป็นรายจ่ายของการดำเนินการ ยิ่งถ้าข้อมูลมีขนาดใหญ่หรือมีจำนวนมากยิ่งต้องใช้ค่าใช้จ่ายที่มากขึ้น ดังนั้นในการวิเคราะห์ข้อมูลแต่ละครั้งจำเป็นต้องกำหนดเป้าหมายให้ชัดเจนว่าเป็นประโยชน์ในด้านการลดต้นทุน การเพิ่มคุณค่า หรือการสร้างนวัตกรรม ซึ่งทั้งนี้ทั้งนั้นผลลัพธ์ที่ได้รับต้องคุ้มค่าในการปฏิบัติงาน

ด้วยเหตุนี้ในการเริ่มวิเคราะห์ข้อมูลจึงจำเป็นต้องตั้งโจทย์ให้ดี ให้ชัดเจน และเกิดประโยชน์ ซึ่งต้องอาศัยความรู้ ความเชี่ยวชาญ และประสบการณ์ของผู้ที่เชี่ยวชาญในด้านนั้น ๆ เพราะโจทย์ที่ดีจะนำไปสู่ทิศทางและกระบวนการวิเคราะห์ข้อมูลที่มีประสิทธิภาพ ดังที่อธิบายในหัวข้อต่อไปนี้

1.1.2. การตั้งโจทย์ในการวิเคราะห์ข้อมูล

ในการตัดสินใจแต่ละครั้งตั้งแต่การตัดสินใจเรื่องส่วนตัวจนไปถึงระดับประเทศ ย่อมมีโจทย์เกิดขึ้นในใจผู้จะตัดสินใจเสมอ (Provost & Fawcett, 2013) ตัวอย่างเช่น ถ้าจะไปถึงที่ทำงานตรงเวลา ควรจะออกจากบ้านเวลาใด? ตรงนี้แม้เป็นคำถามเล็ก ๆ แต่ก็ตอบไม่ง่าย จะต้องแยกย่อยคำถามไปต่อว่าระยะทางจากบ้านถึงที่ทำงานเป็นเท่าใด? รถติดเพียงใด? ในชั่วโมงเร่งด่วนจะสามารถขับรถด้วยความเร็วเท่าใด? คำถามเหล่านี้ไม่สามารถใช้กฎ หรือความรู้สึกตอบได้แน่นอน ถ้าจะตอบให้ได้ต้องอาศัยการวิเคราะห์และสรุปผลข้อมูลที่ถูกต้องจึงจะสามารถได้คำตอบที่แม่นยำออกมา โดยคำถามที่ต้องใช้การวิเคราะห์ข้อมูลจะสามารถแบ่งเป็น 3 วิธี คือการตั้งโจทย์เพื่อหาข้อสรุป, การตั้งโจทย์เชิงพยากรณ์, และการตั้งโจทย์เชิงแนะนำเพื่อประกอบการตัดสินใจ โดยที่มีการประยุกต์ใช้กับด้านต่าง ๆ คือ ด้านการค้า, ด้านสุขภาพ, ด้านการศึกษา, ด้านการเกษตร, และด้านการคมนาคมขนส่ง มาเป็นตัวอย่างของโจทย์

การตั้งโจทย์เพื่อหาข้อสรุป (Descriptive Analytics)

การตั้งโจทย์เพื่อหาข้อสรุปเป็นลักษณะคำถามที่การหาคำตอบเป็นเพียงการสรุปเหตุการณ์ต่าง ๆ ที่เกิดขึ้นมาแล้วโดยใช้ข้อมูลในอดีต จะมักเป็นลักษณะคำถามมีคำว่า เท่าใด ประเภทใด สิ่งไหน เป็นต้น ตัวอย่างคำถามสำหรับการทำธุรกิจขายสินค้าออนไลน์มีดังต่อไปนี้

ด้านการค้า

- ยอดขายเฉลี่ยต่อเดือนในปีที่แล้วเท่าใด?
- ลูกค้า 100 คนแรกที่จ่ายเงินซื้อสินค้ามากที่สุดมีใครบ้าง?
- เดือนใดทำยอดขายเครื่องปรับอากาศได้สูงที่สุด?
- มีลูกค้าร้อยละเท่าใดที่คลิกเข้าไปดูสินค้าแล้วกดซื้อ?
- มีสินค้าประเภทใดบ้างที่ลูกค้าสนใจน้อยที่สุด?

ด้านสุขภาพ

- เดือนใดมีผู้ป่วยไข้หวัดใหญ่สูงที่สุด?
- ช่วงอายุเท่าใดของผู้ป่วยที่นิยมเป็นไข้หวัดใหญ่?
- คนช่วงอายุ 40-50 ปีที่ทำงานในเมือง เป็นโรคใดมากที่สุด?

- ช่วงเวลาใดของแต่ละวันที่ผู้ป่วยนอกใช้เวลารับบริการในโรงพยาบาลนานที่สุด?
- อาชีพใดที่ผิดนัดพบแพทย์บ่อยที่สุด?

ด้านการศึกษา

- เกรดเฉลี่ยของนักศึกษาแต่ละคนในปีการศึกษาที่แล้วเป็นเท่าใด?
- มีนักศึกษาร้อยละเท่าใดที่มีเกรดเฉลี่ยแต่ละเทอมสูงขึ้นในทุก ๆ เทอม?
- วิชา 10 วิชาที่ได้ผลการประเมินสูงที่สุดในปีการศึกษาที่แล้วคือวิชาอะไรบ้าง?
- วิชา 10 วิชาที่ให้เกรดเฉลี่ยสูงที่สุดในปีการศึกษาที่แล้วมีวิชาอะไรบ้าง?
- นักศึกษาจากโรงเรียนใดที่ทำเกรดได้สูงสุดใน 10 ปีทีผ่านมา?

ด้านการเกษตร

- ผลผลิตข้าวในปีที่ผ่านมาของจังหวัดสุพรรณบุรีเป็นเท่าใด?
- พืชผลใดที่ผลิตได้มากที่สุดในแต่ละเดือนของปีที่ผ่านมา?
- ใน 5 ปีที่ผ่านมา 10 จังหวัดใดที่นำเข้าสารเคมีเพื่อการเกษตรสูงที่สุด?
- ใน 5 ปีที่ผ่านมา เดือนใดที่มีจำนวนเกษตรกรมากู้เงินมากที่สุด?
- ใน 5 ปีที่ผ่านมา การเกษตรประเภทใดที่ทำรายได้ในแต่ละเดือนสูงที่สุด?

ด้านการคมนาคม

- ช่วงเวลาใดในกรุงเทพฯ ที่มีรถติดสูงที่สุด?
- สถานที่ท่องเที่ยว 10 ที่ที่รถทัวร์นิยมมาเที่ยว ในเดือนที่ผ่านมา มีที่ใดบ้าง?
- ช่วงถนนใดที่มีการเคลื่อนตัวสะดวกที่สุดในชั่วโมงเร่งด่วน?
- ในปีที่ผ่านมา รถประเภทใดที่มีสัดส่วนการขับรถเร็วเกิน 90 กิโลเมตรต่อชั่วโมง บ่อยที่สุด?
- จุดใดบ้างที่มีอุบัติเหตุทางจราจรสูงที่สุด?

ซึ่งโจทย์เหล่านี้เป็นเพียงคำถามที่สามารถหาได้จากข้อสรุปด้วยวิธีการทางสถิติอย่างง่าย ๆ เช่น การหาค่าเฉลี่ย, ค่าสูงสุด, ค่าต่ำสุด เป็นต้น แต่คำตอบจากคำถามเหล่านี้ยังไม่สามารถสร้างมูลค่าทางธุรกิจได้ในทันที ยังต้องมีการประชุมระดมสมองอีกเพื่อจะลงมือกระทำอะไรสักอย่างให้ได้ผล เช่น หากทราบว่าเดือนมกราคมขายเครื่องปรับอากาศได้ดีที่สุด ก็จะมีคำถามต่อว่า “แล้วอย่างไรต่อ?” เป็นต้น

การตั้งโจทย์เชิงพยากรณ์หรือทำนาย (Predictive Analytics)

การตั้งโจทย์เชิงพยากรณ์เป็นคำถามที่เป็นการทำนายคำตอบในอนาคต ซึ่งส่วนใหญ่จะเป็นการพยากรณ์ค่าหรือพยากรณ์ประเภทของแต่ละกรณี โดยในคำถามมักจะมีคำว่า “จะ” อยู่ในคำถาม ตัวอย่างเช่น

ด้านการค้า

- ยอดขายเดือนหน้าจะเป็นเท่าใด?
- ไตรมาสหน้าจะขายเครื่องปรับอากาศได้กี่เครื่อง?
- ลูกค้ารายนี้จะมีโอกาสกลับมาเข้าร้านอีกหรือไม่?
- ความคิดเห็นสินค้าแบบนี้ ผู้แสดงความคิดเห็นน่าจะชอบสินค้าหรือไม่?

ด้านสุขภาพ

- ไตรมาสหน้าจะมีผู้ป่วยไข้หวัดใหญ่ประมาณกี่คน?
- ไตรมาสหน้าควรจะมีเลือดสำรองเป็นปริมาณเท่าใด?
- ถ้ามีอาการปวดหัวและอาเจียน น่าจะเป็นไข้หวัดใหญ่หรือไม่?
- จากผลการตรวจเลือด ผู้ป่วยมีโอกาสเป็นโรคไขมันพอกตับหรือไม่?

ด้านการศึกษา

- เกรดเฉลี่ยเทอมหน้าของนักเรียนคนนี้จะเป็นเท่าใด?
- นักเรียนคนนี้ ควรเรียนวิศวกรรมคอมพิวเตอร์หรือไม่?
- จากผลการเรียนตลอด 4 ปี นักศึกษาคนนี้มีโอกาสเรียนต่อปริญญาเอกหรือไม่?
- จากผลการเรียนที่ผ่านมา ถ้านักศึกษาคนนี้เลือกวิชาเลือกนี้ จะได้เกรดเท่าไร?

ด้านการเกษตร

- ไตรมาสหน้า จะมีการผลิตทุเรียนเท่าใด?
- พื้นที่นี้ควรขุดบ่อน้ำเพียงใดถึงจะมีน้ำเลี้ยงพื้นที่การเกษตรได้ทั้งปี?
- จากรูปถ่ายทางอากาศที่กำหนด พื้นที่นี้เป็นไร่ข้าวโพดหรือไม่?
- จากสีของใบข้าว ในนาข้าวมีธาตุไนโตเจนเพียงพอหรือไม่?

ด้านการคมนาคม

- ความเร็วเฉลี่ยของจุดนี้ในอีก 5 นาที จะมีค่าเป็นเท่าใด?
- ในวันหยุดยาวที่จะถึงนี้ คาดว่ารถในช่วงถนนนี้ จะมีความเร็วเฉลี่ยเท่าใด?

- การจราจรจุด ณ นี้ กำลังมีอุบัติเหตุเกิดอยู่หรือไม่?
- รถที่วิ่งผ่านจุดนี้ เวลานี้ เป็นรถที่ขนส่งผลผลิตทางการเกษตรหรือไม่?

โจทย์เหล่านี้จะต้องระบุถึงสิ่งที่จะทำนายหรือพยากรณ์ให้ชัดเจน และใช้เทคนิคทางด้านสถิติขั้นสูงร่วมการพัฒนาโปรแกรมเพื่อหารูปแบบ (pattern) ของข้อมูลแล้วนำมาพยากรณ์หาคำตอบ โดยคำตอบนั้นจะเป็นการระบุว่า มีค่าเท่าใดหรืออยู่ในกลุ่มไหนเ เช่น ปีหน้าราคายางพาราจะมีค่าเท่าใด เป็นต้น แต่จะไม่ถามว่า ปีหน้าควรจะปลูกอะไรดี เพราะคำถามอย่างหลังนี้จะเป็นคำถามเพื่อการตัดสินใจในหัวข้อต่อไป

การตั้งโจทย์เชิงแนะนำเพื่อประกอบการตัดสินใจ (Prescriptive Analytics)

การตั้งโจทย์เชิงแนะนำเพื่อประกอบการตัดสินใจเป็นคำถามที่ใช้เพื่อแนะนำแนวทางหรือทางเลือกที่ดีในอนาคตโดยมีเป้าหมายที่ชัดเจน โดยคำถามส่วนใหญ่มักจะมีคำว่า อะไรดี อันใดดี หรือสิ่งไหนดี เป็นต้น ดังหัวข้อต่อไปนี้

ด้านการค้า

- เดือนหน้าจะนำสินค้าใดมาขายดีเพื่อให้ได้ลูกค้าวัยรุ่นเข้ามาใช้บริการเพิ่มมากขึ้น?
- จะแนะนำสินค้าชิ้นใดบ้างให้ลูกค้า เพื่อให้ลูกค้าอยากกดซื้อมากที่สุด?
- จะเลือกรูปสินค้ามุมไหนดี ลูกค้าถึงอยากคลิกมากที่สุด?

ด้านสุขภาพ

- สัปดาห์นี้ควรลงพื้นที่ตรวจโรคที่ไหนดี?
- อาการแบบนี้ควรจะเลือกกายภาพบำบัดวิธีใดดี?
- หลังเกษียณอายุ ควรจะทำอะไรดี ถึงจะมีสุขภาพดี?

ด้านการศึกษา

- คะแนนแบบนี้ควรเข้าคณะไหนดีจึงจะเหมาะกับผลการเรียนมัธยม?
- จากคะแนนที่ผ่านมา หลักสูตรควรแนะนำวิชาเลือกใดให้นักศึกษาคนนี้ดี?
- ภาควิชาควรไปโปรโมทหลักสูตรใหม่นี้ที่โรงเรียนไหนดี?

ด้านการเกษตร

- ปีหน้าควรปลูกอะไรบ้างดีเพื่อให้สร้างรายได้จากผลผลิตและไม่ทำลายธาตุในดิน?
- ควรส่งผลผลิตให้ตลาดไหนดีเพื่อให้ได้กำไรสูง?
- ระบบควรแนะนำให้เกษตรกรปลูกพืชอะไรดี ร่วมกับพืชที่ปลูกอยู่ เพื่อลดศัตรูพืช?

ด้านการคมนาคม

- ควรตัดถนนเพิ่มบริเวณไหนดีเพื่อลดปริมาณจราจรคับคั่งในชั่วโมงเร่งด่วน?
- ควรปรับช่องทางนี้ให้เดินรถทางเดียวช่วงเวลาไหนดี เพื่อแก้ปัญหารถติด?
- ควรเลือกเส้นทางไหนดีในการเดินทางเข้าเมืองพรุ่งนี้ดีโดยใช้เวลาน้อยและคุ้มค่าใช้จ่าย?

สำหรับคำถามเพื่อการตัดสินใจจะมีลักษณะการพยากรณ์ตัวเลือกต่าง ๆ แล้วเลือกทางเลือกที่ดีที่สุดให้ผู้ใช้ หรือชุดทางเลือกที่ดีให้ผู้ใช้ได้เลือก เช่น ระบบแนะนำสินค้าของระบบบัญชีอิเล็กทรอนิกส์ หรือระบบแนะนำคลิปวิดีโอของระบบดูคลิปวิดีโอ เป็นต้น ซึ่งในการเลือกนี้จะใช้เทคนิคทางสถิติขั้นสูงรวมถึงการเขียนโปรแกรมที่ซับซ้อนยิ่งขึ้น เพื่อให้การแนะนำประกอบการตัดสินใจมีความแม่นยำสูงสุด

สรุปแนวทางการตั้งโจทย์

เมื่อถึงจุดนี้ผู้อ่านสามารถคิดตั้งโจทย์และเชื่อมโยงแนวทางของแต่ละคำถามได้ว่ามีลักษณะใด ซึ่งในการตอบโจทย์แต่ละประเภทจะใช้เทคนิคและวิธีการที่ยากง่ายแตกต่างกันออกไป ดังนั้นในทางธุรกิจจะต้องใช้เวลาในการพิจารณาเรื่องนี้ให้รอบคอบ เพราะจะช่วยสร้างประโยชน์ต่อธุรกิจ และนำไปสู่ความได้เปรียบในการแข่งขันอีกด้วย

เพื่อเพิ่มความเข้าใจของผู้อ่าน จึงได้ยกตัวอย่างการเชื่อมโยงคำถามแต่ละประเภท ในรูปของคำถามเกี่ยวกับการเล่นหุ้นของธุรกิจหนึ่ง ดังต่อไปนี้

โจทย์เพื่อหาข้อสรุป

- หุ้นของบริษัทนี้ในไตรมาสที่ผ่านมาขึ้นหรือลงเท่าไร?

โจทย์เชิงพยากรณ์หรือทำนาย

- หุ้นของบริษัทนี้ในไตรมาสหน้าจะมีค่าเท่าใด?
- หุ้นของบริษัทนี้ในไตรมาสหน้าจะขึ้นหรือลง ถ้าอัตราแลกเปลี่ยนเงินบาทสูงขึ้น?

โจทย์เชิงแนะนำเพื่อประกอบการตัดสินใจ

- ควรซื้อหุ้นตัวไหนดีเพื่อทำกำไรในไตรมาสหน้า
- เวลานี้ควรขายหุ้นในพอร์ทตัวไหนดีเพื่อให้ได้ผลกำไรสูง

สำหรับคำถามเพื่อหาข้อสรุปนั้นจะง่ายที่สุด เพราะว่าสามารถใช้ข้อมูลที่มีอยู่มาหาข้อสรุปเป็นคำตอบหรือแสดงในรูปแผนภูมิได้ทันที นอกจากนั้น สำหรับคำถามเชิงพยากรณ์และการตัดสินใจนั้นถือว่ามีความยากตามลำดับ ซึ่งหลายคนมักสับสนในลักษณะการตั้งคำถาม เช่น หัวข้อโครงงานบางเรื่องเขียนไว้

ว่าระบบแนะนำหุ้น แต่ตัวระบบจริงให้ผู้ใช้กรอกโค้ดของหุ้นแล้วระบบจะบอกว่าหุ้นนั้น ๆ จะมีค่าในอนาคตเท่าใด เพื่อให้ผู้ใช้ตัดสินใจซื้อหรือขาย ซึ่งมาถึงจุดนี้ผู้อ่านพอจะทราบว่าโครงงานนี้ควรจะเป็นระบบสำหรับการพยากรณ์หรือการทำนายราคาหุ้นที่เลือก แต่ถ้าจะให้เป็นระบบแนะนำหุ้นจริงนั้น ระบบจะต้องเลือกหุ้นที่คาดว่าจะมีอนาคตที่ดีมาให้ผู้ใช้ได้ตัดสินใจซื้อเลย เป็นต้น ดังนั้นการตั้งคำถามที่ถูกต้องตามประเภทถือเป็นจุดเริ่มสำคัญของการวิเคราะห์ข้อมูล เพราะว่าจะเป็นตัวชี้นำไปถึงการเลือกใช้ข้อมูล, การเลือกใช้เทคนิคการวิเคราะห์, การเลือกใช้วิธีการวัดผล, และการนำเสนอผลการวิเคราะห์ข้อมูล ซึ่งมีผลต่อการสร้างประโยชน์ และการลงทุนลงแรงในการวิเคราะห์ข้อมูลอีกด้วย

1.1.3. ทีมวิเคราะห์ข้อมูล

ในการวิเคราะห์ข้อมูลที่มีประสิทธิภาพ จำเป็นต้องมีทีมงานที่มีความเชี่ยวชาญแต่ละด้านคือ นักวิเคราะห์ข้อมูลธุรกิจ (Business Data Analyst), นักวิศวกรรมข้อมูล (Data Engineer), นักวิทยาศาสตร์ข้อมูล (Data Scientist), และนักพัฒนา (Developer) มาทำงานร่วมกันตั้งแต่การตั้งโจทย์, การจัดเตรียมข้อมูล, การวิเคราะห์ข้อมูล, และการนำผลการวิเคราะห์ข้อมูลไปพัฒนาใช้งานต่อไป

นักวิเคราะห์ข้อมูลธุรกิจ (Business Data Analyst)

นักวิเคราะห์ข้อมูลธุรกิจเป็นผู้ที่มีความรู้ด้านธุรกิจนั้น ๆ เป็นอย่างดี โดยคำว่าธุรกิจหมายถึงเรื่องเฉพาะทางที่มีความเชี่ยวชาญ เช่น ด้านการเงิน, ด้านการเกษตร, ด้านการศึกษา, การขนส่ง, ด้านการแพทย์, และอื่น ๆ เป็นต้น และยังต้องมีความรู้ความเข้าใจในการใช้ข้อมูลเพื่อเสริมศักยภาพของธุรกิจเป็นอย่างดี โดยมีหน้าที่หลักคือการตั้งโจทย์เพื่อเป็นแนวทางการวิเคราะห์ข้อมูลที่สามารถนำไปสร้างคุณค่าให้กับธุรกิจนั้น ๆ โดยมีแนวทางคือ การลดต้นทุน, การเพิ่มกำไร, หรือการสร้างผลิตภัณฑ์ใหม่ ๆ เป็นต้น

นักวิทยาศาสตร์ข้อมูล (Data Scientist)

นักวิทยาศาสตร์ข้อมูลคือผู้ที่ทำหน้าที่หลักในการพัฒนาโมเดลในการวิเคราะห์ข้อมูลที่ตอบโจทย์ทางธุรกิจ โดยสร้างผลการวิเคราะห์ที่มีความแม่นยำสูง และนำไปใช้งานได้จริง ซึ่งมีลักษณะการทำงานตามกระบวนการทางวิทยาศาสตร์ โดยมีขั้นตอนดังนี้คือ

- **การกำหนดปัญหา**

เมื่อได้รับโจทย์ในการวิเคราะห์ข้อมูลแล้ว นักวิทยาศาสตร์ข้อมูลจะต้องตีความโจทย์ให้ได้ ว่าเป็นปัญหาประเภทใด เช่น เป็นโจทย์เพื่อหาข้อสรุป, เป็นโจทย์เชิงพยากรณ์หรือทำนาย, หรือเป็นโจทย์เชิงแนะนำประกอบการตัดสินใจ เป็นต้น โดยเฉพาะถ้าเป็นโจทย์เชิงพยากรณ์ จะต้องระบุลงลึกต่อไปให้ได้ว่าพยากรณ์ค่าตัวเลขต่อเนื่อง (continuous number) เช่นทำนายราคาหุ้น ซึ่งค่าที่พยากรณ์ได้เป็นตัวเลข

จำนวนจริง, หรือจะเป็นการพยากรณ์ข้อมูลไม่ต่อเนื่อง (discrete data) หรือจัดกลุ่ม (classification) เช่น ทำนายว่าหุ้นตัวนี้จะขึ้นหรือลง ซึ่งค่าที่พยากรณ์มี 2 กลุ่มคือขึ้นหรือลง

- **การตั้งสมมติฐาน**

เป็นการคาดคะเนคำตอบของปัญหาอย่างมีเหตุผล ระหว่างตัวแปรต้นหรือตัวแปรอิสระ กับตัวแปรตาม ซึ่งในทางปฏิบัติ ตัวแปรต้นมักจะเป็นข้อมูลตั้งต้นและเทคนิคการวิเคราะห์ และตัวแปรตามมักจะเป็นสิ่งที่ไปตอบโจทย์ทางธุรกิจ, ความแม่นยำที่สูงขึ้น, หรือเวลาในการประมวลผลที่เร็วขึ้น เป็นต้น ตัวอย่างเช่น ในการพยากรณ์ยอดขายเครื่องปรับอากาศ สามารถตั้งสมมติฐานได้ว่า “อุณหภูมิของสิ่งแวดล้อมมีผลต่อยอดขายเครื่องปรับอากาศ” ในสมมติฐานนี้บอกตัวแปรต้นคืออุณหภูมิและตัวแปรตามคือยอดขายเครื่องปรับอากาศ แต่สามารถระบุให้มากขึ้นถึงเทคนิคได้คือ “การใช้ค่าของอุณหภูมิร่วมกับเทคนิคสมการถดถอยเชิงเส้น (Linear Regression) สามารถเพิ่มความแม่นยำในการทำนายยอดขายได้มากขึ้น” ซึ่งการตั้งสมมติฐานแบบนี้จะนำไปสู่การหาเก็บข้อมูลอุณหภูมิ และใช้วิธีสมการถดถอยเชิงเส้นในการเพิ่มค่าความแม่นยำในการทำนายยอดขายเครื่องปรับอากาศ ดังนั้นการตั้งสมมติฐานที่ดี จะส่งผลต่อการเลือกใช้ข้อมูล เลือกใช้เทคนิคการวิเคราะห์ และเป้าหมายของการวัดผล ที่ชัดเจน

- **การตรวจสอบสมมติฐาน**

หลังจากตั้งสมมติฐานแล้วลำดับต่อไปคือการเลือกใช้ข้อมูลกับเทคนิคเพื่อหาคำตอบ หรือเพิ่มความแม่นยำ หรือลดเวลาในการวิเคราะห์ แล้วก็เข้าสู่กระบวนการตรวจสอบสมมติฐาน ซึ่งโดยหลักคือการทดลองที่เป็นการใช้เครื่องมือวิเคราะห์ข้อมูลหรือเขียนโปรแกรมเพื่อวิเคราะห์ข้อมูลตามแนวทางที่ได้ระบุในสมมติฐานนั้นไว้

- **การวิเคราะห์ผล**

หลังจากทำการทดลองเพื่อตรวจสอบสมมติฐานเรียบร้อยแล้ว จึงนำผลการทดลองมาวิเคราะห์ต่อ เพื่อสำรวจว่าผลของการตรวจสอบสมมติฐานที่ได้มานั้นมีความถูกต้องหรือไม่ กระบวนการวัดผลน่าเชื่อถือเพียงใด และนำผลที่ได้มานั้นตอบโจทย์ที่ได้รับมาถูกต้องหรือไม่

- **การสรุปผล**

สุดท้ายเป็นการสรุปผลการวิเคราะห์ข้อมูลเพื่อระบุว่ายอมรับสมมติฐานนั้นหรือไม่ ถ้ายอมรับก็นำโมเดลที่พัฒนาขึ้นไปเป็นต้นแบบในการพัฒนาจริง หรือถ้าไม่ยอมรับก็ต้องระบุสาเหตุและแนวทางการแก้ปัญหาต่อไป

ในการทำงานของนักวิทยาศาสตร์ข้อมูลนั้นจะทำงานกับบุคคลอื่นด้วย โดยมีหน้าที่ร่วมกับทีมดังนี้

- **หน้าที่ร่วมกับนักวิเคราะห์ข้อมูลธุรกิจ**

ทำโดยการที่นักวิทยาศาสตร์ข้อมูลร่วมทำความเข้าใจโจทย์ที่ได้รับจากนักวิเคราะห์ข้อมูลธุรกิจ, วางแผนขั้นตอนการทำงาน, วางแนวทางการเลือกใช้ข้อมูล และนำผลการวิเคราะห์ไปอภิปรายร่วมกันเพื่อเกิดคุณประโยชน์ตามที่เป้าหมายที่ตั้งไว้

- **หน้าที่ร่วมกับนักวิศวกรรมข้อมูล**

ทำโดยที่นักวิทยาศาสตร์ข้อมูลชี้แจงและทำความเข้าใจร่วมกันว่าจะใช้ข้อมูลอะไรบ้าง เพื่อให้นักวิศวกรรมข้อมูลช่วยจัดเตรียมชุดข้อมูลที่พร้อมใช้ (clean dataset) ให้มาใช้วิเคราะห์

- **หน้าที่ร่วมกับนักพัฒนา**

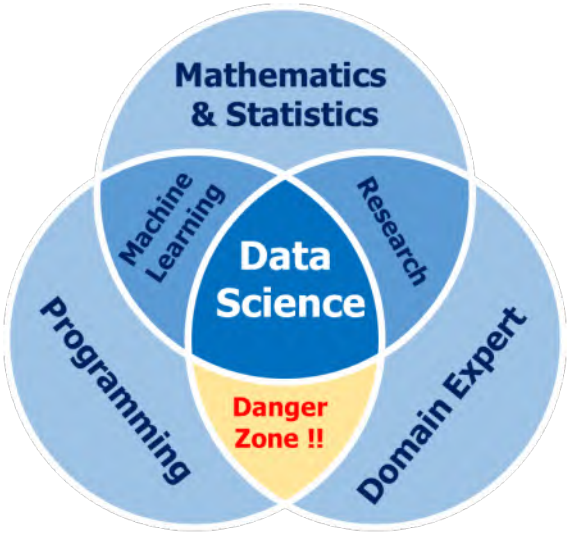
ทำโดยการอธิบายโมเดลที่วิเคราะห์ได้ซึ่งเป็นแบบจำลองทางคณิตศาสตร์ เพื่อให้นักพัฒนาสามารถนำไปเขียนโปรแกรมเพื่อพัฒนาซอฟต์แวร์ตามที่ธุรกิจต้องใช้

จะเห็นได้ว่านักวิทยาศาสตร์ข้อมูลมีบทบาทหลักในกระบวนการวิเคราะห์ข้อมูล ดังนั้นอาชีพนี้ต้องมีความรู้รอบด้าน ทั้งด้านคณิตศาสตร์และสถิติศาสตร์, การเขียนโปรแกรม, ความรู้ด้านธุรกิจนั้น ๆ, และการสื่อสาร เป็นอย่างดี และต้องประสานงานร่วมกันกับบุคคลากรในทีมวิเคราะห์ข้อมูลได้อย่างดีด้วย

เมื่อพูดถึงทักษะงานของวิทยาศาสตร์ข้อมูล (Data Science) แล้ว ในวงการวิทยาศาสตร์ข้อมูลจะกล่าวถึง 3 ทักษะดังที่กล่าวมาแล้ว หลักตามแผนภาพที่เป็นที่นิยมพูดถึงใน**รูปที่ 1.1-1** คือมีทักษะด้านคณิตศาสตร์และสถิติศาสตร์ (Mathematics & Statistics), ทักษะด้านการเขียนโปรแกรม (Programming), และความรู้ด้านธุรกิจนั้น ๆ หรือความเชี่ยวชาญเฉพาะด้าน (Domain Expert) แต่ไม่ได้กล่าวถึงทักษะด้านการสื่อสาร (Communication Skills) ด้วยถือเป็นทักษะด้านการทำงาน (Soft Skill) ที่ทุกอาชีพควรจะมีอยู่แล้ว ตามรูปนี้ได้พยายามสื่อสารว่างานวิทยาศาสตร์ข้อมูลที่ดีจะต้องมีทักษะครบทั้ง 3 ด้านนี้ แต่ถ้าขาดความความรู้ด้านใดด้านหนึ่งเช่น ขาดความรู้ด้านการเขียนโปรแกรมก็จะสามารถผลิตงานวิจัยได้, ถ้าขาดความเชี่ยวชาญทางธุรกิจเฉพาะด้านก็ยังสามารถใช้ทักษะทางคณิตศาสตร์ สถิติศาสตร์และการเขียนโปรแกรม ผลิตผลงานด้านการเรียนรู้ของเครื่อง (Machine Learning) ที่ไปเป็นพื้นฐานของธุรกิจต่าง ๆ ได้, แต่ทว่าหากมีความความรู้ด้านธุรกิจและเขียนโปรแกรมเป็น แต่ขาดความเชี่ยวชาญด้านคณิตศาสตร์และสถิติศาสตร์ขั้นสูง ซึ่งส่วนใหญ่เป็นบุคคลากรไอทีที่ไม่ได้ทุ่มเทศึกษาคณิตศาสตร์และสถิติศาสตร์อย่างถ่องแท้ แต่สนใจในงานธุรกิจ ถือเป็นจุดที่มีความเสี่ยงสูงหรือเขตอันตราย (Danger Zone) เนื่องจากบุคคลนั้นมีแนวโน้มที่จะเลือกใช้เทคนิคหรือวิธีการตามผู้อื่นจากการค้นคว้าในอินเทอร์เน็ตแล้วนำมาเขียนโปรแกรมวิเคราะห์ข้อมูลทันที โดยไม่คำนึงถึงลักษณะพฤติกรรมของข้อมูล คุณลักษณะของเทคนิคที่ใช้ และลักษณะของโจทย์ที่ได้รับให้ถ่องแท้เสียก่อน ซึ่งกรณีนี้อาจจะสามารถจะสร้างโมเดลที่ได้ผลดี

ในช่วงการทดลอง แต่มีโอกาสที่จะทำให้เกิดความเข้าใจของความสัมพันธ์ของข้อมูลที่คลาดเคลื่อนส่งผลให้เกิดความผิดพลาดเมื่อใช้งานจริง จนเกิดความเสียหายทางธุรกิจได้

ดังนั้น การเป็นนักวิทยาศาสตร์ข้อมูลที่ดีจะต้องมีพื้นฐานความรู้ด้านคณิตศาสตร์, สถิติศาสตร์, การเขียนโปรแกรม, ความรู้ในธุรกิจหรือหัวเรื่องนั้น ๆ, และความสามารถในการสื่อสาร เพื่อให้สามารถทำงานกับทีมเพื่อสร้างโมเดลที่แม่นยำ และสามารถสร้างกระบวนการวิเคราะห์ที่ตอบโจทย์ได้ตามความต้องการที่แท้จริง



รูปที่ 1.1-1 : ทักษะของงานวิทยาศาสตร์ข้อมูล

นักวิศวกรรมข้อมูล (Data Engineer)

ถ้าเปรียบนักวิทศาสตร์ข้อมูลเป็นพ่อครัวที่ทำหน้าที่ปรุงอาหาร นักวิศวกรรมข้อมูลก็เป็นดังผู้จัดเตรียมวัตถุดิบที่ดีให้นักวิทยาศาสตร์ข้อมูล ซึ่งในกระบวนการวิเคราะห์ข้อมูลนั้น นักวิศวกรรมข้อมูลจะทำหน้าที่รวบรวมข้อมูลและประมวลผลข้อมูลเบื้องต้นเป็นหลัก เพื่อให้ได้ชุดข้อมูลที่พร้อมจะนำไปวิเคราะห์ โดยในการปฏิบัติงานจริงนั้น งานของนักวิศวกรรมข้อมูลมีบทบาทมากและใช้ทรัพยากรมาก เนื่องจากข้อมูลมีมหาศาลและไม่ได้อยู่ในรูปที่พร้อมใช้งานได้จริง ดังนั้นตำแหน่งนี้จึงต้องเป็นผู้เลือกใช้เทคโนโลยีและเครื่องมือที่มีประสิทธิภาพที่สุด ในการนำเข้าข้อมูล รวบรวมข้อมูล จัดเก็บข้อมูล และประมวลผลข้อมูลเบื้องต้น เพื่อให้พร้อมนำไปวิเคราะห์ข้อมูลต่อไป

นักพัฒนา (Developer)

เมื่อนักวิทยาศาสตร์ข้อมูลทำการทดลองจนได้โมเดลที่เหมาะสมแล้ว นักพัฒนาซึ่งก็คือนักพัฒนาระบบซอฟต์แวร์ ต้องนำโมเดลซึ่งเป็นฟังก์ชันบรรจุแบบจำลองทางคณิตศาสตร์ นำไปพัฒนาเพื่อใช้งานจริง

ในระบบที่ใช้งานจริง ที่เป็นเช่นนี้เนื่องจากในการทดลองของนักวิทยาศาสตร์ข้อมูลอาจใช้เครื่องมือสำหรับการปฏิบัติการทดลอง เช่น MATLAB, R, RapidMiner เป็นต้น แต่ในการใช้งานจริงอาจจะมีสภาพแวดล้อมต่างจากเครื่องมือที่นักวิทยาศาสตร์ข้อมูลทดลงใช้ เช่น พัฒนาบนระบบฝังตัว (Embed System) หรือแอปพลิเคชันบนโทรศัพท์มือถือ (Mobile Application) ทำให้นักพัฒนาต้องนำแนวคิดของโมเดลจากห้องปฏิบัติการ นำมาพัฒนาจริงบนเครื่องมือหรือภาษาคอมพิวเตอร์ที่ต่างออกไปเพื่อให้สามารถทำงานในระบบนั้น ๆ ได้

1.2. ระดับการวิเคราะห์ข้อมูล

สำหรับการวิเคราะห์ข้อมูลนั้นทำขึ้นเพื่อให้เกิดการกระทำ (Action) อย่างหนึ่งเพื่อให้เกิดประโยชน์ตามเป้าหมาย เช่น เพื่อจะโฆษณาสินค้านี้ที่ไหน เวลาไหนถึงจะมียอดขายเพิ่ม, เพื่อจะเดินทางไปยังสถานที่นั้นด้วยเส้นทางไหนเพื่อให้เดินทางทันเวลาและประหยัดค่าใช้จ่าย, เพื่อจะลดพนักงานกี่คนโดยให้บริษัทสามารถพ้นวิกฤติ เป็นต้น ดังนั้นเพื่อให้เกิดการกระทำ (Action) จึงต้องใช้ข้อมูล (Data) สนับสนุน แต่ลำพังข้อมูลดิบนั้นยังไม่สามารถช่วยอะไรได้ ด้วยเหตุนี้ข้อมูลเหล่านี้จึงต้องผ่านการวิเคราะห์ข้อมูล (Analytics) เสียก่อนเพื่อให้ได้ผลลัพธ์ไปใช้ตัดสินใจ (Decision) แล้วจึงเกิดการกระทำ (Action) ตามมา ดัง**รูปที่ 1.2-1** โดยแผนภาพนี้จะแสดงระดับการวิเคราะห์ข้อมูล ประกอบด้วยการวิเคราะห์เชิงพรรณนา (Descriptive Analytics), การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics), และการวิเคราะห์เชิงแนะนำ (Prescriptive Analytics) โดยแต่ละวิธีในแผนภาพจะแสดงให้เห็นถึงความซับซ้อนของเทคนิคในการวิเคราะห์ข้อมูล ซึ่งถ้ามีความซับซ้อนไม่มากดังเช่นการวิเคราะห์เชิงพรรณนาซึ่งเป็นเพียงหาข้อสรุปข้อมูล เช่น ยอดขายรายวัน รายเดือน รายปี นั้นไม่อาจช่วยให้ตัดสินใจได้ทันทีว่าจะต้องลงทุนเก็บสินค้าสำหรับแต่ละสัปดาห์เท่าใด จึงจำเป็นต้องมีการออกแรงระดมสมองเพิ่มเติมโดยมนุษย์ (More Human Input) จึงจะสามารถตัดสินใจ (Decision) ได้ว่าควรทำอะไร แล้วจึงลงมือทำ (Action) นั้น ดังนั้นยิ่งเป็นระดับการวิเคราะห์ข้อมูลขั้นสูงขึ้น เช่น การวิเคราะห์เชิงพยากรณ์ หรือเชิงแนะนำ จะใช้เทคนิคและความรู้ทางคณิตศาสตร์ที่มีความซับซ้อนขึ้น กล่าวคือต้องลงทุนลงแรงไปกับกระบวนการวิเคราะห์อย่างหนัก แต่ได้ผลลัพธ์ที่สามารถลดภาระมนุษย์ในการทำงานต่อน้อยลง (Human Input) เพื่อให้ได้ข้อตัดสินใจและการลงมือทำ ดังนั้นนักวิเคราะห์ข้อมูลต้องเลือกให้ได้ว่าโจทย์ที่กำลังวิเคราะห์อยู่ระดับใด เพราะแต่ละระดับมีการลงทุนลงแรงที่ต่างกัน ใช้เทคนิคและเครื่องมือที่ต่างกัน และความคาดหวังในผลลัพธ์และการออกแรงระดมสมองเพิ่มเติมเพื่อให้ได้แนวทางการตัดสินใจด้วย