

วิทยาการข้อมูล และการวิเคราะห์ข้อมูลด้วยโปรแกรม

Data Science and
Data Analysis with Program

R



พิมผกา ธานินพงศ์
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่



ประวัติและผลงาน

ผศ.ดร.พิมผกา ธานีพวงศ์

การศึกษา

- ▶ ปริญญาเอก สาขาวิชาวิทยาการคอมพิวเตอร์ คณะเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยมหิดล ปี พ.ศ. 2554
- ▶ ปริญญาโท สาขาวิชาสถิติประยุกต์ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์ ปี พ.ศ. 2544
- ▶ ปริญญาตรี สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ปี พ.ศ. 2538

การทำงาน

- ▶ อาจารย์ประจำภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ตั้งแต่วันที่ 1 กรกฎาคม พ.ศ. 2545 ปัจจุบันดำรงตำแหน่ง ผู้ช่วยศาสตราจารย์ สาขาวิชาสถิติประยุกต์
- ▶ หัวหน้าโครงการจัดตั้งศูนย์วิจัยวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ตั้งแต่วันที่ 16 เมษายน พ.ศ. 2564 – 31 พฤษภาคม พ.ศ. 2566
- ▶ รองหัวหน้าโครงการจัดตั้งศูนย์วิจัยวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ตั้งแต่วันที่ 23 พฤศจิกายน พ.ศ. 2563 – 15 เมษายน พ.ศ. 2564
- ▶ นักวิเคราะห์ระบบ (system analyst) บริษัท อาร์ ไอ เอส จำกัด ปี พ.ศ. 2542 – 2545
- ▶ โปรแกรมเมอร์ บริษัท ซอฟต์แวร์ 1999 จำกัด ปี พ.ศ. 2540 – 2542
- ▶ Management Trainee (IT) บริษัท ชันแวลเลย์ (ประเทศไทย) จำกัด ปี พ.ศ. 2538 – 2540

วิทยาการข้อมูล และการวิเคราะห์ข้อมูลด้วยโปรแกรม

Data Science and Data Analysis with Program

R



พิมผกา ธานินพงศ์
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

วิทยาการข้อมูลและการวิเคราะห์ข้อมูล ด้วยโปรแกรม R

Data Science and Data Analysis with Program R

บรรณาธิการ: วีระศักดิ์ ปัญญาพรวิทยา

ISBN (e-book): 978-616-620-118-5

ผู้เขียน: พิมพ์กา ธาณินพงศ์

เจ้าของและผู้จัดพิมพ์: สำนักพิมพ์มหาวิทยาลัยเชียงใหม่

สำนักงานบริหารงานวิจัย มหาวิทยาลัยเชียงใหม่

โทรศัพท์: 0 5394 3603-5

<https://cmupress.cmu.ac.th>,

E-mail: cmupress.th@gmail.com

พิมพ์ครั้งที่ 1: สิงหาคม 2569

ราคา: 360 บาท

หนังสือเล่มนี้ ผ่านการประเมินจากผู้ทรงคุณวุฒิในสาขาที่เกี่ยวข้องและต่างสถาบัน เพื่อตรวจสอบและรับรองการเผยแพร่ผลงานทางวิชาการ ประเภทหนังสือหรือตำรา ซึ่งจัดพิมพ์โดยสำนักพิมพ์มหาวิทยาลัยเชียงใหม่เรียบร้อยแล้ว

© สงวนลิขสิทธิ์ พ.ศ. 2537 ห้ามลอกเลียนแบบ ทำซ้ำ ดัดแปลง ส่วนหนึ่งส่วนใดของหนังสือเล่มนี้ รวมทั้งการจัดเก็บ ถ่ายทอดไม่ว่ารูปแบบหรือวิธีการใด ๆ ด้วยกระบวนการทางอิเล็กทรอนิกส์ การถ่ายภาพ การบันทึกหรือวิธีการอื่นใด โดยไม่ได้รับอนุญาต

คำนิยม



หนังสือ “วิทยาการข้อมูลและการวิเคราะห์ข้อมูลด้วยโปรแกรม R” เล่มนี้ นับเป็นผลงานทางวิชาการที่มีคุณค่าและสอดคล้องกับบริบทของโลกยุคดิจิทัลที่ข้อมูลมีบทบาทสำคัญต่อการตัดสินใจ ผู้เขียนได้ถ่ายทอดเนื้อหาทั้งด้านทฤษฎีและการประยุกต์ใช้งานไว้อย่างกระชับ ชัดเจน และเป็นระบบ พร้อมยกตัวอย่างที่น่าสนใจและ R codes ที่สามารถนำไปประยุกต์ใช้ ทั้งในด้านการจัดการข้อมูล การสำรวจข้อมูล การสร้างแบบจำลอง และการประยุกต์ใช้เทคนิคทางสถิติและการเรียนรู้ของเครื่อง ซึ่งช่วยให้ผู้อ่านสามารถเชื่อมโยงทฤษฎีกับการปฏิบัติได้อย่างชัดเจน ทำให้ผู้อ่านสามารถเข้าใจแนวคิดพื้นฐานของวิทยาการข้อมูล ตลอดจนการวิเคราะห์ข้อมูลด้วยโปรแกรม R ได้อย่างมีประสิทธิภาพ หนังสือเล่มนี้เหมาะสำหรับนักศึกษาระดับปริญญาตรี ปริญญาโท ผู้เริ่มต้นศึกษา ตลอดจนผู้สนใจทั่วไปที่ต้องการพัฒนาทักษะด้านการวิเคราะห์ข้อมูลและการใช้โปรแกรม R เพื่อประยุกต์ใช้ในการศึกษา การวิจัย หรือการทำงานในอนาคต

รองศาสตราจารย์ ดร.นายสัตวแพทย์ วีระศักดิ์ ปัญญาพรวิทยา
บรรณาธิการ

คำนิยม



หนังสือวิทยาการข้อมูลและการวิเคราะห์ข้อมูลด้วยโปรแกรม R เล่มนี้ แต่งโดย ผู้ช่วยศาสตราจารย์ ดร.พิมผกา ธานินพงศ์ ถือเป็นคลังความรู้ที่เหมาะสมกับยุคสมัยที่กำลังเปลี่ยนแปลงไปอย่างรวดเร็ว ด้วยข้อมูล และยุคที่ข้อมูลได้เข้ามามีบทบาทสำคัญในการดำเนินชีวิตของคนเรา หนังสือเล่มนี้จึงเหมาะสำหรับผู้อ่านที่มีความสนใจในวิทยาการข้อมูล เพียงผู้อ่านมีพื้นฐานทางสถิติ คณิตศาสตร์ และการใช้โปรแกรมสำเร็จรูปในระดับเบื้องต้น ก็สามารถทำความเข้าใจเนื้อหาของหนังสือเล่มนี้ได้ในทุกบท เนื่องจากผู้เขียนได้ร้อยเรียงเนื้อหาอย่างต่อเนื่อง และมีตัวอย่างของการวิเคราะห์ข้อมูลอย่างชัดเจน ซึ่งผู้อ่านสามารถทำความเข้าใจด้วยตนเองได้ และหนังสือเล่มนี้ยังมีความเหมาะสมที่จะใช้ประกอบการเรียนในระดับอุดมศึกษาที่มีการเรียนการสอนในรายวิชาที่เกี่ยวข้องกับ วิทยาการข้อมูลและการวิเคราะห์ข้อมูลด้วยโปรแกรม R อีกด้วย จุดเด่นของหนังสือเล่มนี้คือ การอธิบายเนื้อหาสำคัญของแต่ละบท ได้อย่างชัดเจน กระชับ และมีการแสดงผลการวิเคราะห์ข้อมูลพร้อมการใช้คำสั่งในโปรแกรม R ในตอนท้ายของแต่ละบท ผู้เขียนได้จัดทำแบบฝึกหัดครอบคลุมเนื้อหาอย่างครบถ้วน หากผู้อ่านได้ศึกษาและทำแบบฝึกหัดท้ายบทแล้ว จะมั่นใจได้ว่าผู้อ่านมีความรู้ความเข้าใจในวิทยาการข้อมูล และสามารถวิเคราะห์ข้อมูลเพื่อสร้างสารสนเทศ เพื่อนำไปสู่การตัดสินใจได้อย่างเหมาะสม

ผมขอแสดงความชื่นชม ผู้ช่วยศาสตราจารย์ ดร.พิมผกา ธานินพงศ์ ที่ได้สละเวลาอันมีค่าเขียนหนังสือเล่มนี้ด้วยความวิริยะ อุตสาหะ เพื่อเป็นแนวทางให้ผู้สนใจได้ศึกษา และนำความรู้มาใช้ในการพัฒนาตนเอง สังคม และประเทศชาติ ให้รู้เท่าทันข้อมูล และใช้ข้อมูลมาสร้างสารสนเทศ จนเกิดเป็นองค์ความรู้ที่มีค่าและนำมาใช้ในการตัดสินใจได้อย่างชาญฉลาด หนังสือเล่มนี้จึงเป็นเสมือนขุมทรัพย์ทางปัญญาอีกเล่มหนึ่งที่มีคุณค่าในสาขาวิทยาการข้อมูล

รองศาสตราจารย์พิษณุ เจียวกุณ

คำนำ



ปัจจุบันเป็นยุคที่ข้อมูลเป็นทรัพยากรที่มีมูลค่าสูง ทุกองค์กรทั้งภาคธุรกิจและภาครัฐต่างนำข้อมูลมาใช้ให้เป็นประโยชน์ วิทยาการข้อมูล (Data Science) จึงเป็นหนึ่งทักษะที่มีความสำคัญและเป็นที่ต้องการอย่างแพร่หลายในทุกองค์กร หนังสือ “วิทยาการข้อมูลและการวิเคราะห์ข้อมูลด้วยโปรแกรม R” เล่มนี้จัดทำขึ้นเพื่อเป็นคู่มือสำหรับผู้ที่มีความสนใจศึกษาเกี่ยวกับวิทยาการข้อมูล ได้เข้าใจในศาสตร์ที่เกี่ยวข้องกับการเก็บรวบรวมข้อมูล จัดเตรียมข้อมูล สืบค้นข้อมูล แสดงภาพข้อมูล วิเคราะห์ข้อมูล พัฒนาแบบจำลอง โดยอาศัยการเรียนรู้เชิงสถิติ การเรียนรู้ของเครื่อง เพื่อนำไปสู่การตัดสินใจที่มีประสิทธิภาพ โดยสรุปประกอบในหนังสือเล่มนี้จัดทำโดยผู้เขียน

เนื้อหาหนังสือประกอบด้วยแนวคิดพื้นฐานทางด้านวิทยาการข้อมูล การจัดการข้อมูลและการแปลงข้อมูล การสำรวจข้อมูลและการแสดงภาพข้อมูล การวิเคราะห์ความสัมพันธ์ การจำแนกประเภทข้อมูลด้วยการเรียนรู้เชิงสถิติและการเรียนรู้ของเครื่อง การจัดกลุ่มข้อมูล และการพยากรณ์ โดยมุ่งเน้นการประยุกต์ใช้โปรแกรม R ซึ่งเป็นเครื่องมือนิยมใช้ในการวิเคราะห์ข้อมูลทางสถิติที่รู้จักอย่างกว้างขวางตลอดจนการประยุกต์ใช้งานผ่านกรณีศึกษา เนื้อหาในหนังสือถูกออกแบบให้สามารถนำไปประยุกต์ใช้ได้จริงพร้อมตัวอย่างการเขียนโปรแกรม ซึ่งสามารถทดลองปฏิบัติตามได้ทันที

หนังสือเล่มนี้ถูกออกแบบมาเพื่อยกระดับทักษะสำหรับผู้ที่มีความรู้พื้นฐานทางด้านสถิติและมีทักษะการเขียนโปรแกรมมาบ้างแล้ว จึงเหมาะสำหรับนักศึกษาในระดับปริญญาตรีและระดับปริญญาโท สาขาวิทยาการข้อมูลตลอดจนผู้ที่มีความสนใจในด้านวิทยาการข้อมูลและต้องการเพิ่มทักษะการวิเคราะห์เชิงลึกด้วยโปรแกรม R เพื่อนำไปประยุกต์ใช้งานจริง

ผู้เขียนขอกราบขอบพระคุณพ่อ-แม่ที่วางรากฐานการศึกษาของผู้เขียนเป็นอย่างดี กราบขอบพระคุณครู-อาจารย์ทุกท่านที่ได้ประสิทธิ์ประสาทความรู้แก่ผู้เขียน กราบขอบพระคุณรองศาสตราจารย์สุรินทร์ ขนาศักดิ์ที่กรุณาให้คำแนะนำในการเขียนหนังสือ ตลอดจนเพื่อนร่วมงานที่ให้คำแนะนำแก้ไขเนื้อหาขอขอบคุณภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่ให้การสนับสนุนแก่ผู้เขียนในการผลิตหนังสือเล่มนี้ สุดท้ายนี้ขอขอบคุณครอบครัวของผู้เขียนที่คอยให้กำลังใจ ขอขอบคุณลูกชายที่ช่วยออกแบบปกหนังสือเล่มนี้

ผู้เขียนขอขอบคุณผู้ประเมินหนังสือเล่มนี้ที่ได้ให้ข้อเสนอแนะอันเป็นประโยชน์ต่อการปรับปรุงหนังสือเล่มนี้ให้มีความสมบูรณ์มากยิ่งขึ้น และขอขอบคุณสำนักพิมพ์มหาวิทยาลัยเชียงใหม่ ที่ได้ให้ความกรุณาในการจัดพิมพ์หนังสือเล่มนี้

หากหนังสือเล่มนี้มีข้อบกพร่องหรือความผิดพลาดประการใด หรือหากผู้อ่านหรือผู้ที่นำไปใช้มีข้อเสนอแนะใด กรุณาแจ้งให้ผู้เขียนทราบที่ email: phimpaka.t@cmu.ac.th ผู้เขียนขออภัยเป็นอย่างสูงในข้อบกพร่อง/ความผิดพลาดนั้นและจะนำไปปรับ/แก้ไขในโอกาสต่อไป ผู้เขียนยินดีรับฟังข้อคิดเห็นต่างๆ และขอขอบคุณมา ณ โอกาสนี้

พิมพกา ธาณินพงศ์

สิงหาคม 2569

สารบัญ



	หน้า
คำนิยาม	III
คำนำ	V
สารบัญ	VII
สารบัญรูป	XIII
สารบัญตาราง	XIX

»» บทที่ 1 แนวคิดพื้นฐานทางด้านวิทยาการข้อมูล	1
1.1 วิทยาการข้อมูลคืออะไร	1
1.2 ข้อมูลขนาดใหญ่	3
1.3 กระบวนการของวิทยาการข้อมูล	4
1.4 วิทยาการข้อมูลในการทำงาน	6
1.5 การติดตั้งและการเขียนโปรแกรม R เบื้องต้นสำหรับวิทยาการข้อมูล	8
1.5.1 การติดตั้งโปรแกรม R	8
1.5.2 การติดตั้งแพ็คเกจ	11
1.5.3 การคำนวณค่าในโปรแกรม R	12
1.5.4 ตัวดำเนินการทางคณิตศาสตร์ในโปรแกรม R	13
1.5.5 วัตถุในโปรแกรม R	14
1.5.6 การเรียกใช้งานฟังก์ชันในโปรแกรม R	19
1.5.7 การเขียนฟังก์ชันในโปรแกรม R	20
1.5.8 การเขียนเงื่อนไข if..else ในโปรแกรม R	21
1.5.9 การเขียนคำสั่งวนซ้ำในโปรแกรม R	22
1.5.10 การเขียนข้อมูลลงไฟล์ในโปรแกรม R	23
1.6 สรุป	23
แบบฝึกหัดบทที่ 1	24
เอกสารอ้างอิง	26

»» บทที่ 2 การสำรวจข้อมูลและแสดงภาพข้อมูล	29
2.1 การสำรวจข้อมูล	29
2.1.1 มาตรวัดของข้อมูล	30
2.1.2 ประเภทของตัวแปร	30
2.1.3 การสำรวจข้อมูลและสรุปผลข้อมูลสำหรับตัวแปรเดียว	31

	หน้า
2.1.4 การสำรวจข้อมูลและสรุปผลข้อมูลสำหรับสองตัวแปร	39
2.2 การแสดงภาพข้อมูล	44
2.2.1 การออกแบบหน้าจอแสดงผลข้อมูล	45
2.2.2 การแสดงภาพข้อมูลสำหรับตัวแปรเดียว	49
2.2.3 การแสดงภาพข้อมูลเพื่อแสดงความสัมพันธ์ของตัวแปร	55
2.2.4 การแสดงภาพข้อมูลสำหรับการเปรียบเทียบ	60
2.2.5 ก้อนเมฆกลุ่มคำ	64
2.3 สรุป	66
2.4 กรณีศึกษาการใช้โปรแกรม R ในการแสดงภาพข้อมูล	66
แบบฝึกหัดบทที่ 2	75
เอกสารอ้างอิง	78
» บทที่ 3 การจัดการข้อมูลและการแปลงข้อมูล	81
3.1 การจัดการข้อมูล	81
3.1.1 รูปแบบไฟล์ข้อมูล	82
3.1.2 ประเภทข้อมูลจำแนกตามโครงสร้าง	88
3.2 การทำความสะอาดข้อมูล	89
3.2.1 การตรวจสอบข้อมูลเชิงกลุ่ม	90
3.2.2 การแทนค่าสูญหาย	90
3.2.3 การแก้ไขข้อมูลที่มีค่าผิดปกติ	99
3.2.4 การแก้ไขข้อมูลวันที่และเวลา	103
3.2.5 การสร้างตัวแปรใหม่จากวันที่และเวลา	105
3.2.6 การแก้ไขข้อมูลไม่สอดคล้องกัน	105
3.2.7 การเชื่อมโยงข้อมูลจากหลายแหล่ง	109
3.3 การแปลงข้อมูล	113
3.3.1 การสกัดคุณลักษณะ	114
3.3.2 การสร้างตัวแปรใหม่	117
3.3.3 การคัดเลือกชุดตัวแปร	123
3.4 สรุป	124
3.5 กรณีศึกษาการใช้โปรแกรม R เพื่อจัดการข้อมูล	124
แบบฝึกหัดบทที่ 3	128
เอกสารอ้างอิง	133

» บทที่ 4 การวิเคราะห์ความสัมพันธ์

หน้า

137

4.1 การวิเคราะห์ความสัมพันธ์	137
4.2 การวิเคราะห์สหสัมพันธ์	138
4.3 การทดสอบไคกำลังสอง	142
4.4 การวิเคราะห์กฎความสัมพันธ์	146
4.4.1 ขั้นตอนวิธีอะพริโอไร	150
4.4.2 การขุดกฎความสัมพันธ์	153
4.4.3 การใช้โปรแกรม R เพื่อการขุดกฎความสัมพันธ์	153
4.5 สรุป	160
4.6 กรณีศึกษาการแนะนำสินค้าโดยใช้กฎความสัมพันธ์	160
แบบฝึกหัดบทที่ 4	168
เอกสารอ้างอิง	173

» บทที่ 5 การจำแนกประเภทข้อมูลโดยวิธีการเรียนรู้เชิงสถิติ

177

5.1 การจำแนกประเภทข้อมูล	177
5.1.1 การแบ่งข้อมูลสำหรับการสร้างตัวแบบจำแนกประเภท	178
5.1.2 การประเมินผลตัวแบบจำแนกประเภท	183
5.1.3 ปัญหา overfitting และ underfitting	186
5.2 การจำแนกประเภทข้อมูลโดยวิธีเบส	187
5.2.1 การสร้างตัวแบบเนอโฟบีส	188
5.2.2 การใช้โปรแกรม R ในการสร้างตัวแบบเนอโฟบีส	197
5.3 การถดถอยลอจิสติก	205
5.3.1 การสร้างตัวแบบการถดถอยลอจิสติก	205
5.3.2 การทำนายโดยใช้ตัวแบบการถดถอยลอจิสติก	208
5.3.3 การคัดเลือกตัวแปรกรณีตัวแปรทำนายมีหลายตัวแปร	210
5.3.4 การประเมินประสิทธิภาพของตัวแบบการถดถอยลอจิสติก	211
5.3.5 การเปรียบเทียบตัวแบบการถดถอยลอจิสติก	212
5.3.6 การใช้โปรแกรม R ในการสร้างตัวแบบการถดถอยลอจิสติก	212
5.4 ต้นไม้ตัดสินใจ	221
5.4.1 โครงสร้างของต้นไม้ตัดสินใจ	222
5.4.2 เกณฑ์วัดความบริสุทธิ์ของข้อมูล	223
5.4.3 การสร้างแผนภาพต้นไม้โดยขั้นตอนวิธี ID3	227
5.4.4 การสร้างกฎการตัดสินใจในการจำแนกกลุ่ม	234

5.4.5 การเขียนโปรแกรม R เพื่อสร้างต้นไม้ตัดสินใจ	235
5.5 สรุป	238
5.6 กรณีศึกษาการเขียนโปรแกรม R เพื่อการจำแนกอีเมลขยะ	239
แบบฝึกหัดบทที่ 5	246
เอกสารอ้างอิง	256
»» บทที่ 6 การจำแนกประเภทข้อมูลโดยวิธีการเรียนรู้ของเครื่อง	261
6.1 การจำแนกประเภทข้อมูลโดยวิธีซัพพอร์ตเวกเตอร์แมชชีน	261
6.1.1 กรณีข้อมูลสามารถแยกกลุ่มได้โดยใช้เส้นตรง	263
6.1.2 กรณีข้อมูลไม่สามารถแยกกลุ่มได้โดยใช้เส้นตรง	268
6.1.3 การจำแนกประเภทข้อมูลโดยใช้ตัวแบบ SVM	270
6.1.4 ซัพพอร์ตเวกเตอร์แมชชีนแบบหนึ่งกลุ่ม	271
6.1.5 การจำแนกประเภทข้อมูลหลายกลุ่ม	273
6.1.6 การใช้โปรแกรม R ในการสร้างตัวแบบ SVM	274
6.2 การจำแนกประเภทข้อมูลโดยวิธีโครงข่ายประสาทเทียม	281
6.2.1 การสร้างตัวแบบโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า	284
6.2.2 วิธีการแพร่ย้อนกลับ	285
6.2.3 การจำแนกประเภทโดยใช้ตัวแบบ ANN	289
6.2.4 การใช้โปรแกรม R ในการสร้างตัวแบบ ANN	292
6.3 สรุป	296
6.4 กรณีศึกษาการเขียนโปรแกรม R เพื่อการจำแนกอีเมลขยะ	296
6.4.1 การจำแนกประเภทโดยใช้ตัวแบบ SVM	297
6.4.2 การจำแนกประเภทโดยใช้ตัวแบบ ANN	299
แบบฝึกหัดบทที่ 6	302
เอกสารอ้างอิง	309
»» บทที่ 7 การจัดกลุ่มข้อมูล	313
7.1 แนวคิดของการจัดกลุ่มข้อมูล	313
7.2 การวัดความคล้ายและความต่างของจุดข้อมูล	314
7.3 การจัดกลุ่มข้อมูลโดยใช้เทคนิคการแบ่งข้อมูล	317
7.3.1 วิธีการแบ่งข้อมูลด้วยขั้นตอนวิธี k-means	317
7.3.2 วิธีการแบ่งข้อมูลด้วยขั้นตอนวิธี k-mediods	322
7.3.3 การกำหนดค่า k ที่เหมาะสม	323

	หน้า
7.3.4 การใช้โปรแกรม R สำหรับขั้นตอนวิธี k-means	324
7.4 การจัดกลุ่มข้อมูลแบบลำดับขั้น	327
7.4.1 การสร้างผังลำดับขั้นของข้อมูลโดยแนวทางแบบรวมกลุ่ม	327
7.4.2 การใช้โปรแกรม R ในการจัดกลุ่มข้อมูลแบบลำดับขั้น	333
7.5 การประเมินคุณภาพของการจัดกลุ่มข้อมูล	338
7.5.1 ดัชนี silhouette	339
7.5.2 การใช้โปรแกรม R ในการประเมินคุณภาพการจัดกลุ่มข้อมูล	343
7.6 สรุป	345
7.7 กรณีศึกษาการเขียนโปรแกรม R เพื่อการจัดกลุ่มลูกค้า	345
แบบฝึกหัดบทที่ 7	353
เอกสารอ้างอิง	358

» บทที่ 8 การพยากรณ์ **361**

8.1 แนวคิดเกี่ยวกับการพยากรณ์	361
8.2 การวิเคราะห์การถดถอยเชิงเส้น	362
8.2.1 การถดถอยเชิงเส้นอย่างง่าย	362
8.2.2 การถดถอยเชิงเส้นพหุ	368
8.2.3 การทดสอบความเหมาะสมของตัวแบบการถดถอยเชิงเส้น	370
8.2.4 การใช้โปรแกรม R สำหรับการวิเคราะห์การถดถอยเชิงเส้น	375
8.3 การวิเคราะห์อนุกรมเวลา	379
8.3.1 ส่วนประกอบของอนุกรมเวลา	379
8.3.2 การสร้างกราฟข้อมูลอนุกรมเวลา	382
8.3.3 การแยกส่วนประกอบของอนุกรมเวลา	386
8.3.4 การปรับให้เรียบโดยวิธีเฉลี่ยเคลื่อนที่	386
8.3.5 การปรับให้เรียบเอกซ์โปเนนเชียลชั้นเดียว	392
8.3.6 ตัวแบบแนวโน้มเชิงเส้น	396
8.3.7 ตัวแบบของ Holt และ Winter	404
8.4 การพยากรณ์โดยใช้โครงข่ายประสาทเทียม	413
8.4.1 โครงข่ายประสาทเทียมสำหรับตัวแปรพหุ	413
8.4.2 การใช้โปรแกรม R ในการสร้างตัวแบบ ANN สำหรับตัวแปรพหุ	418
8.4.3 โครงข่ายประสาทเทียมสำหรับข้อมูลอนุกรมเวลา	420
8.4.4 โครงข่ายประสาทเทียมสำหรับอนุกรมเวลาที่มีตัวแปรนอกระบบ	424

	หน้า
8.5 การวัดความแม่นยำของตัวแบบพยากรณ์	426
8.6 สรุป	427
8.7 กรณีศึกษาการเขียนโปรแกรม R เพื่อการพยากรณ์	428
แบบฝึกหัดบทที่ 8	442
เอกสารอ้างอิง	448
ภาคผนวก	453
รายการไฟล์ข้อมูล	453
แนวทางอธิบายแบบฝึกหัด	454
บรรณานุกรม	499
ดัชนี	501

สารบัญรูป



	หน้า
รูปที่ 1.1 Data Science Venn Diagram	2
รูปที่ 1.2 สิ่งที่มีบทบาทสำคัญต่อ Big Data	4
รูปที่ 1.3 กระบวนการของวิทยาการข้อมูล	5
รูปที่ 1.4 เว็บไซต์สำหรับดาวน์โหลดโปรแกรม R	8
รูปที่ 1.5 เว็บไซต์สำหรับดาวน์โหลดโปรแกรม RStudio	9
รูปที่ 1.6 หน้าจอ RStudio	10
รูปที่ 1.7 หน้าต่างสำหรับติดตั้งแพ็คเกจ	11
รูปที่ 1.8 การแสดงผลในส่วนของ environment pane	12
รูปที่ 1.9 ข้อมูลกลุ่มบุคคล	17
รูปที่ 2.1 แผนภาพ density ของตัวแปร Petal.Length	34
รูปที่ 2.2 แผนภาพกล่องของตัวแปร Petal.Length	34
รูปที่ 2.3 ฮิสโทแกรมของตัวแปรอิสระ	38
รูปที่ 2.4 แผนภาพกล่องเปรียบเทียบของตัวแปรอิสระจำแนกตามตัวแปร species	38
รูปที่ 2.5 แผนภาพ mosaic แสดงจำนวนกล้วยไม้จำแนกตามการรดน้ำ และการอยู่รอด	42
รูปที่ 2.6 แผนภาพการกระจายแสดงความสัมพันธ์ของตัวแปรเชิงปริมาณ	43
รูปที่ 2.7 แผนภาพการกระจายระหว่างตัวแปรในชุดข้อมูล iris	44
รูปที่ 2.8 ตำแหน่ง รูปร่าง ขนาด สี และรูปแบบเส้น ในการแสดงภาพข้อมูล	46
รูปที่ 2.9 ตัวอย่างระดับสีสำหรับใช้แบ่งกลุ่มข้อมูล	47
รูปที่ 2.10 ตัวอย่างระดับสีแบบไล่ระดับ	47
รูปที่ 2.11 ตัวอย่างระดับสีที่แสดงความมากน้อยในสองทิศทาง (ColorBrewer “RdBu”)	48
รูปที่ 2.12 ตัวอย่างระดับสีแบบอ่อนเข้ม (ColorBrewer “Accent”)	48
รูปที่ 2.13 แผนภูมิแท่งแสดงจำนวนกล้วยไม้จำแนกตามการอยู่รอดของกล้วยไม้ (Died : ไม่สามารถอยู่รอด, Survived : อยู่รอด)	50
รูปที่ 2.14 แผนภูมิวงกลมแสดงสัดส่วนของกล้วยไม้จำแนกตามการอยู่รอด (Died : ไม่สามารถอยู่รอด, Survived : อยู่รอด)	51
รูปที่ 2.15 แผนภูมิวงกลมแสดงร้อยละกล้วยไม้จำแนกตามการอยู่รอด (Died : ไม่ สามารถอยู่รอด, Survived : อยู่รอด) เมื่อปรับแต่งแล้ว	52
รูปที่ 2.16 แผนภูมิเส้นแสดงปริมาณ PM2.5 ณ สถานี 35T	53

รูปที่ 2.17 แผนภูมิแท่งแนวนอนแสดงจำนวนนักท่องเที่ยวที่เดินทางเข้าประเทศ ตั้งแต่เดือน มกราคม-มิถุนายนปี พ.ศ. 2568	54
รูปที่ 2.18 แผนภาพการกระจายแบบเมทริกซ์ 2 มิติ ของตัวแปรทั้งหมดของชุดข้อมูล iris	55
รูปที่ 2.19 correlogram จากคำสั่ง corplot() โดยกำหนด method = 'number'	58
รูปที่ 2.20 correlogram จากคำสั่ง corplot() โดยกำหนด method = 'color'	58
รูปที่ 2.21 correlogram จากคำสั่ง corplot() โดยกำหนด method = 'ellipse'	59
รูปที่ 2.22 แผนภาพฟองอากาศ	60
รูปที่ 2.23 แผนภูมิแท่งแบบซ้อนกันจำนวนต้นกล้วยไม้จำแนกตามการรอดชีพ (Died : ไม่สามารถอยู่รอด, Survived : อยู่รอด) และการรดน้ำ (Once : เฉพาะช่วงเช้า, Twice : รดน้ำเช้าบ่าย)	61
รูปที่ 2.24 แผนภูมิเส้นแสดงปริมาณ PM2.5 ณ สถานี 35T และสถานี 36T	62
รูปที่ 2.25 แผนภูมิพื้นที่แสดงจำนวนนักศึกษาตามน้ำหนักจำแนกตามเพศ	64
รูปที่ 2.26 ก้อนเมฆของกลุ่มคำที่เกี่ยวข้องกับวิทยาการข้อมูล	65
รูปที่ 2.27 รายละเอียดของข้อมูล house_prices	67
รูปที่ 2.28 การสร้าง Flex Dashboard	69
รูปที่ 2.29 โครงร่าง R markdown สำหรับสร้าง Dashboard	69
รูปที่ 2.30 กระดานแสดงภาพข้อมูล (dashboard) ของ house_prices	74
รูปที่ 3.1 ตัวอย่างไฟล์ข้อมูล person.csv	83
รูปที่ 3.2 ตัวอย่างไฟล์ข้อมูล person.txt	85
รูปที่ 3.3 ตัวอย่างไฟล์ข้อมูล person.json	87
รูปที่ 3.4 ข้อมูลกลุ่มบุคคลใน data frame "person"	93
รูปที่ 3.5 แผนภาพกล่องแสดงค่าผิดปกติในข้อมูลส่วนสูง	95
รูปที่ 3.6 แผนภาพการกระจายแสดงความสัมพันธ์ระหว่างน้ำหนักและส่วนสูงของเพศหญิง	97
รูปที่ 3.7 แผนภาพกล่องแสดงค่าผิดปกติของตัวแปรอายุ	102
รูปที่ 3.8 ตัวอย่างไฟล์กฎที่ใช้ในการตรวจสอบชื่อว่า "editrules.txt"	106
รูปที่ 3.9 การเชื่อมโยงตารางข้อมูลสองตาราง	109
รูปที่ 3.10 ข้อมูล new_person	110
รูปที่ 3.11 ผลลัพธ์ที่ได้จากการใช้ left_join()	111
รูปที่ 3.12 ผลลัพธ์ที่ได้จากการใช้ full_join()	112
รูปที่ 3.13 ผลลัพธ์ที่ได้จากการใช้ right_join()	112
รูปที่ 3.14 ผลลัพธ์ที่ได้จากการใช้ inner_join()	113

	หน้า
รูปที่ 3.15 ตัวอย่างรูปตัวเลขเขียนเลข 5	116
รูปที่ 3.16 เวกเตอร์คุณลักษณะของรูปเลข 5	116
รูปที่ 3.17 แผนภาพกล่องของจำนวนห้องนอน	125
รูปที่ 4.1 แผนภาพการกระจายแสดงทิศทางความสัมพันธ์เชิงเส้นและไม่เป็นเชิงเส้น	139
รูปที่ 4.2 แผนภาพการกระจายระหว่างตัวแปร Petal Length และ Sepal Length	141
รูปที่ 4.3 อาณาเขตวิกฤตของการทดสอบไคกำลังสอง	144
รูปที่ 4.4 ตัวอย่างข้อมูล fastfoods.csv	154
รูปที่ 4.5 กราฟเชื่อมโยงกฎความสัมพันธ์	158
รูปที่ 4.6 หน้าจอ Viewer Zoom เพื่อปรับเส้นเชื่อม	159
รูปที่ 4.7 หน้าจอ Viewer Zoom เพื่อเลือกแสดงเฉพาะบางกฎความสัมพันธ์	159
รูปที่ 4.8 กราฟแท่งแสดงความถี่ของสินค้าที่มีการซื้อซ้ำบ่อยที่สุด	162
รูปที่ 4.9 กราฟแท่งแสดงความถี่ของสินค้าที่มีการซื้อซ้ำบ่อยที่สุดเมื่อกำหนด ค่าสนับสนุนต่ำสุดที่ 0.15	163
รูปที่ 5.1 การแบ่งชุดข้อมูลแบบ (ก) train-test-split และ (ข) train-validation-test-split	178
รูปที่ 5.2 การแบ่งข้อมูลแบบ 5-fold cross validation	180
รูปที่ 5.3 bias และ variance สำหรับ underfitting และ overfitting	187
รูปที่ 5.4 (ก) แสดงการประมาณความน่าจะเป็นของการผิנדชำระหนี้ โดยวิธีการถดถอยเชิงเส้นอย่างง่าย (ข) แสดงการประมาณความน่าจะเป็น ของการผิנדชำระหนี้โดยวิธีการถดถอยลอจิสติก	206
รูปที่ 5.5 ฟังก์ชันลอจิสติก	206
รูปที่ 5.6 แผนภาพกล่องเปรียบเทียบอายุ	214
รูปที่ 5.7 แผนภาพกล่องเปรียบเทียบระดับความดันโลหิตค่าบน	214
รูปที่ 5.8 ต้นไม้ตัดสินใจเพื่อจำแนกประเภทความเสี่ยงการเป็นโรคความดันโลหิตสูง	221
รูปที่ 5.9 องค์ประกอบของต้นไม้ตัดสินใจ	222
รูปที่ 5.10 ความบริสุทธิ์ของกลุ่มหลังจากเลือกคุณลักษณะที่ใช้แบ่งกลุ่ม	223
รูปที่ 5.11 แผนภาพต้นไม้ตัดสินใจเมื่อเลือกตัวแปร temperature เป็นโน้ตดราก	231
รูปที่ 5.12 แผนภาพต้นไม้เพื่อจำแนกการอยู่รอดของต้นกล้วยไม้	234
รูปที่ 5.13 แผนภาพต้นไม้ตัดสินใจจากคำสั่ง J48()	236
รูปที่ 5.14 แผนภาพต้นไม้ตัดสินใจจากคำสั่ง rpart()	238

	หน้า
รูปที่ 5.15 กราฟ ROC และค่า AUC ของตัวแบบเนอรัลเน็ต	242
รูปที่ 5.16 กราฟ ROC และค่า AUC ของตัวแบบการถดถอยลอจิสติก	245
รูปที่ 6.1 แสดงจุดข้อมูลในพื้นที่สองมิติ	262
รูปที่ 6.2 แสดงเส้นแบ่งข้อมูลหลากหลายเส้น	263
รูปที่ 6.3 (ก) ระบายแบ่งที่มีความกว้างแคบ (ข) ระบายแบ่งที่มีความกว้างมาก	263
รูปที่ 6.4 ระบายแบ่งที่เหมาะสมที่สุด	264
รูปที่ 6.5 ระยะขอบ	265
รูปที่ 6.6 รูปแบบขอบอ่อนซัพพอร์ตเวกเตอร์แมชชีน	267
รูปที่ 6.7 ตัวอย่างข้อมูลไม่สามารถแบ่งด้วยเส้นตรง	268
รูปที่ 6.8 การแปลงข้อมูลจากพื้นที่ข้อมูลนำเข้าเป็นพื้นที่ที่มีมิติสูงขึ้น	269
รูปที่ 6.9 ระบายการตัดสินใจเมื่อใช้ฟังก์ชันเคอร์เนลแตกต่างกัน	270
รูปที่ 6.10 แผนภาพแสดงจุดข้อมูลที่ต้องการจำแนกกลุ่ม	271
รูปที่ 6.11 เส้นตัดสินใจสำหรับซัพพอร์ตเวกเตอร์แมชชีนแบบหนึ่งกลุ่มในพื้นที่คุณลักษณะ	272
รูปที่ 6.12 โครงสร้างเซลล์ประสาทของสมองมนุษย์	282
รูปที่ 6.13 โครงสร้างของโครงข่ายประสาทเทียม	283
รูปที่ 6.14 ตัวอย่างโครงข่ายประสาทเทียมแบบ simple perceptron	283
รูปที่ 6.15 ตัวอย่างโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า	284
รูปที่ 6.16 โหนดในชั้นนำเข้า	286
รูปที่ 6.17 โหนดในชั้นซ่อน	286
รูปที่ 6.18 โครงข่ายประสาทเทียมเพื่อทำนายการอยู่รอดของต้นกล้วยไม้	290
รูปที่ 6.19 โครงข่ายประสาทเทียมเพื่อทำนายการอยู่รอดของต้นกล้วยไม้โดยใช้คำสั่ง plot()	295
รูปที่ 6.20 กราฟ ROC และค่า AUC ของตัวแบบ SVM	298
รูปที่ 6.21 กราฟ ROC และค่า AUC ของตัวแบบ ANN	300
รูปที่ 7.1 กราฟแสดงจุดข้อมูลของชุดข้อมูล	319
รูปที่ 7.2 ผลการแบ่งกลุ่ม	322
รูปที่ 7.3 ค่า WSS เมื่อกำหนดจำนวนกลุ่ม (k) ต่าง ๆ	324
รูปที่ 7.4 กราฟจุดข้อมูลและผลการแบ่งกลุ่ม	326
รูปที่ 7.5 ค่า WSS เมื่อกำหนดจำนวนกลุ่ม (k) ต่าง ๆ โดยวิธี k-means	327
รูปที่ 7.6 ขั้นตอนการสร้างผังลำดับชั้นแบบรวมกลุ่ม	328
รูปที่ 7.7 ความแตกต่างระหว่างแนวทางแบบรวมกลุ่มกับแตกกลุ่ม	328

รูปที่ 7.8 แผนภาพ dendrogram ของวิธี agglomerative	333
รูปที่ 7.9 แผนภาพ dendrogram จากคำสั่ง agnes()	335
รูปที่ 7.10 แผนภาพ dendrogram จากคำสั่ง diana()	337
รูปที่ 7.11 กราฟจุดข้อมูลและผลการแบ่งกลุ่มโดยคำสั่ง diana()	338
รูปที่ 7.12 (ก) แสดงการวัดค่าความคล้ายคลึงกันภายในกลุ่ม (cohesion) (ข) การวัดค่าความแตกต่างกันระหว่างกลุ่ม (separation)	338
รูปที่ 7.13 ตัวอย่างผลการแบ่งกลุ่ม	341
รูปที่ 7.14 แผนภาพกล่องของ Recency Frequency และ Monetary	348
รูปที่ 7.15 ฮิสโทแกรมของ Recency Frequency และ Monetary	349
รูปที่ 7.16 ตารางแสดงผลศูนย์กลางของคลัสเตอร์	350
รูปที่ 7.17 ผลการแบ่งกลุ่มลูกค้า	351
รูปที่ 7.18 ผลลัพธ์จากคำสั่ง plot_ly()	352
รูปที่ 8.1 แผนภาพการกระจายและสมการถดถอยเชิงเส้น	363
รูปที่ 8.2 แผนภาพการกระจายแสดงความสัมพันธ์ระหว่างน้ำหนักและส่วนสูงข้าง	366
รูปที่ 8.3 แสดง correlogram ของตัวแปรในชุดข้อมูล Combined Cycle Power Plant	375
รูปที่ 8.4 การวิเคราะห์ค่าส่วนเหลือ	378
รูปที่ 8.5 แสดงตัวอย่าง (ก) ข้อมูลรายวัน (ข) ข้อมูลรายเดือน (ค) ข้อมูลรายปี	380
รูปที่ 8.6 อนุกรมเวลาที่มีวัฏจักร	381
รูปที่ 8.7 แสดงข้อมูลราคาทองคำรายวันที่มีค่าสูญหาย	382
รูปที่ 8.8 แสดงข้อมูลราคาทองคำรายวันที่เติมค่าสูญหายแล้ว	383
รูปที่ 8.9 แสดงข้อมูลจำนวนผู้ป่วยไข้เลือดออกเดงกี	384
รูปที่ 8.10 แสดงข้อมูลอัตราแลกเปลี่ยนเงินดอลลาร์สหรัฐ	385
รูปที่ 8.11 ส่วนประกอบของอนุกรมเวลา	386
รูปที่ 8.12 ตัวแบบเฉลี่ยเคลื่อนที่ของราคาทองคำ	389
รูปที่ 8.13 ตัวแบบการปรับให้เรียบเอกซ์โปเนนเชียลขั้นเดียว	395
รูปที่ 8.14 ตัวแบบแนวโน้มเชิงเส้นระยะยาว	399
รูปที่ 8.15 ค่าพยากรณ์แนวโน้มเชิงเส้นระยะสั้น	403
รูปที่ 8.16 ค่า fitted ของตัวแบบพยากรณ์แนวโน้มระยะสั้น	404
รูปที่ 8.17 (ก) กราฟความแปรปรวนตามฤดูกาลคงที่ (ข) ความแปรปรวนตามฤดูกาล เพิ่มขึ้นหรือลดลงเป็นสัดส่วนกับแนวโน้มเมื่อเวลาเปลี่ยนไป	404
รูปที่ 8.18 ตัวแบบ Holt และ Winter และค่าพยากรณ์	410

	หน้า
รูปที่ 8.19 ตัวแบบ Holt และ Winter และค่าพยากรณ์ปริมาณฝุ่น PM2.5 ของสถานี 35T	412
รูปที่ 8.20 ตัวแบบ Holt และ Winter และค่าพยากรณ์ปริมาณฝุ่น PM2.5 ของสถานี 36T	412
รูปที่ 8.21 โครงข่ายประสาทเทียมสำหรับการพยากรณ์	413
รูปที่ 8.22 โครงข่ายประสาทเทียมหลังจากการปรับปรุงค่าน้ำหนัก	416
รูปที่ 8.23 โครงข่ายประสาทเทียมเพื่อทำนายน้ำหนักข้าง	417
รูปที่ 8.24 โครงสร้างของตัวแบบโครงข่ายประสาทเทียมเพื่อทำนายน้ำหนักข้าง	419
รูปที่ 8.25 โครงข่ายประสาทเทียมสำหรับข้อมูลอนุกรมเวลา	420
รูปที่ 8.26 ตัวแบบโครงข่ายประสาทเทียมและค่าพยากรณ์ปริมาณฝุ่น PM2.5 ของสถานี 35T	423
รูปที่ 8.27 ตัวแบบโครงข่ายประสาทเทียมและค่าพยากรณ์ปริมาณฝุ่น PM2.5 ของสถานี 36T	423
รูปที่ 8.28 ตัวแบบโครงข่ายประสาทเทียมและค่าพยากรณ์ปริมาณฝุ่น PM2.5 ของสถานี 36T เมื่อพิจารณาตัวแปรจุดความร้อนร่วมด้วย	426
รูปที่ 8.29 เมทริกซ์สัมประสิทธิ์สหสัมพันธ์	429
รูปที่ 8.30 แผนภาพกล่องแสดงการแจกแจงของราคาบ้านแยกตามตัวแปร waterfront	430
รูปที่ 8.31 แผนภาพกล่องแสดงการแจกแจงของราคาบ้านแยกตามตัวแปร view	430
รูปที่ 8.32 แผนภาพกล่องแสดงการแจกแจงของราคาบ้านแยกตามตัวแปร condition	431
รูปที่ 8.33 แผนภาพกล่องแสดงการแจกแจงของราคาบ้านแยกตามตัวแปร grade	432
รูปที่ 8.34 การวิเคราะห์ค่าส่วนเหลือ	437

XIX

	หน้า
ตารางที่ 4.10 อาร์กิวเมนต์ของคำสั่ง apriori()	155
ตารางที่ 4.11 ตัวอย่างรายการสินค้าในชุดข้อมูล basket	160
ตารางที่ 4.12 ตัวอย่างข้อมูลในไฟล์ข้อมูล basket	161
ตารางที่ 5.1 ตารางการตัดสินใจหรือเมทริกซ์ความสับสน	183
ตารางที่ 5.2 ข้อมูลตัวอย่างการเลี้ยงกล้วยไม้เขตร้อน	189
ตารางที่ 5.3 ความน่าจะเป็นก่อนหน้า	189
ตารางที่ 5.4 ความน่าจะเป็นแบบมีเงื่อนไข	190
ตารางที่ 5.5 ตัวแบบเนอปีเบส์	193
ตารางที่ 5.6 ข้อมูลตัวอย่างการเลี้ยงกล้วยไม้เขตร้อน	195
ตารางที่ 5.7 ความน่าจะเป็นก่อนหน้า	196
ตารางที่ 5.8 การแจกแจงความน่าจะเป็นแบบมีเงื่อนไขของตัวแปร temp	196
ตารางที่ 5.9 ความน่าจะเป็นแบบมีเงื่อนไขของตัวแปร water	196
ตารางที่ 5.10 การแจกแจงความน่าจะเป็นแบบมีเงื่อนไขของตัวแปร humidity	196
ตารางที่ 5.11 อาร์กิวเมนต์ของคำสั่ง naiveBayes()	198
ตารางที่ 5.12 เมทริกซ์ความสับสนของตัวแบบเนอปีเบส์	203
ตารางที่ 5.13 แสดงวิธีการคำนวณค่า odds และค่าอัตราส่วน odds	209
ตารางที่ 5.14 อาร์กิวเมนต์ของคำสั่ง glm()	213
ตารางที่ 5.15 ตัวอย่างข้อมูลผู้ป่วยจำนวน 10 ราย	226
ตารางที่ 5.16 ข้อมูลตัวอย่างกล้วยไม้เขตร้อน	227
ตารางที่ 5.17 ข้อมูลตัวอย่างกล้วยไม้เรียงตามตัวแปร temperature	228
ตารางที่ 5.18 ข้อมูลตัวอย่างกล้วยไม้เรียงตามตัวแปร water	229
ตารางที่ 5.19 ข้อมูลตัวอย่างกล้วยไม้เรียงตามตัวแปร humidity	230
ตารางที่ 5.20 ข้อมูลย่อยของ temperature = ">40" เรียงตามตัวแปร water	232
ตารางที่ 5.21 ข้อมูลย่อยของ temperature = ">40" เรียงตามตัวแปร humidity	233
ตารางที่ 5.22 ข้อมูลต้นกล้วยไม้ที่ต้องการจำแนกประเภท	234
ตารางที่ 5.23 อาร์กิวเมนต์ของคำสั่ง J48()	235
ตารางที่ 5.24 อาร์กิวเมนต์ของคำสั่ง rpart()	237
ตารางที่ 6.1 ตัวอย่างข้อมูลสองมิติ จำนวน 12 จุดข้อมูล	262
ตารางที่ 6.2 อาร์กิวเมนต์ของคำสั่ง SVM()	275
ตารางที่ 6.3 ตารางการตัดสินใจหรือเมทริกซ์ความสับสนสำหรับสามกลุ่ม	279

	หน้า
ตารางที่ 6.4 ข้อมูลตัวอย่างการเลี้ยงกล้วยไม้เขตร้อน	289
ตารางที่ 6.5 ค่าต่ำสุดและสูงสุดของตัวแปร temp และ humidity	290
ตารางที่ 6.6 ข้อมูลนำเข้า	291
ตารางที่ 6.7 น้ำหนักบนเส้นเชื่อมจากโหนดนำเข้าและโหนดในชั้นซ่อนและ การคำนวณในแต่ละโหนด	291
ตารางที่ 6.8 ผลลัพธ์ในโหนดชั้นซ่อน	291
ตารางที่ 6.9 น้ำหนักบนเส้นเชื่อมจากโหนดนำเข้าและโหนดในชั้นซ่อนและ การคำนวณในแต่ละโหนด	291
ตารางที่ 6.10 อาร์กิวเมนต์ของคำสั่ง neuralnet()	292
ตารางที่ 6.11 เปรียบเทียบประสิทธิภาพของตัวแบบจำแนกประเภทอีเมลขยะ	301
ตารางที่ 7.1 ตารางการจรรยาบรรณสำหรับคำนวณสัมประสิทธิ์ของ Jaccard	315
ตารางที่ 7.2 ข้อมูลตัวอย่างยอดจำหน่ายสินค้า	319
ตารางที่ 7.3 จุดศูนย์กลางเริ่มต้น (Centroid)	319
ตารางที่ 7.4 การจัดจุดข้อมูลให้อยู่ในกลุ่มที่ใกล้ที่สุดรอบที่ 1	320
ตารางที่ 7.5 ปรับจุดศูนย์กลาง (Centroid) รอบที่ 1	320
ตารางที่ 7.6 การจัดจุดข้อมูลให้อยู่ในกลุ่มที่ใกล้ที่สุดรอบที่ 2	320
ตารางที่ 7.7 ปรับจุดศูนย์กลาง (Centroid) รอบที่ 2	321
ตารางที่ 7.8 การจัดจุดข้อมูลให้อยู่ในกลุ่มที่ใกล้ที่สุดรอบที่ 3	321
ตารางที่ 7.9 สมาชิกและจุดศูนย์กลางของแต่ละกลุ่ม	321
ตารางที่ 7.10 อาร์กิวเมนต์ของคำสั่ง kmeans()	324
ตารางที่ 7.11 เมตริกซ์ระยะทางโดยใช้ระยะทาง Euclidean	329
ตารางที่ 7.12 เลือกรวมจุดข้อมูล B1 กับ B3 เพราะมีระยะทางน้อยที่สุด	330
ตารางที่ 7.13 การปรับปรุงระยะทางในเมตริกซ์ระยะทางหลังจากรวม B1 และ B3	330
ตารางที่ 7.14 เมตริกซ์ระยะทางใหม่หลังจากรวม B1 และ B3	330
ตารางที่ 7.15 การปรับปรุงระยะทางในเมตริกซ์ระยะทางหลังจากรวม (B1, B3) และ B6	331
ตารางที่ 7.16 เมตริกซ์ระยะทางใหม่หลังจากรวม (B1, B3) และ B6	331
ตารางที่ 7.17 การปรับปรุงระยะทางในเมตริกซ์ระยะทางหลังจากรวม B4 และ B5	331
ตารางที่ 7.18 เมตริกซ์ระยะทางใหม่หลังจากรวม B4 และ B5	332
ตารางที่ 7.19 การปรับปรุงระยะทางในเมตริกซ์ระยะทางหลังจากรวม (B4, B5) และ B7	332
ตารางที่ 7.20 เมตริกซ์ระยะทางใหม่หลังจากรวม (B4, B5) และ B7	332
ตารางที่ 7.21 อาร์กิวเมนต์ของคำสั่ง agnes()	333

ตารางที่ 7.22 อาร์กิวเมนต์ของคำสั่ง <code>diana()</code>	334
ตารางที่ 7.23 ตัวอย่างข้อมูล	340
ตารางที่ 7.24 เมทริกซ์ระยะทาง	341
ตารางที่ 7.25 คำนวณค่า $a(i)$ ของคลัสเตอร์ A โดยคำนวณค่าเฉลี่ยในแต่ละคอลัมน์	341
ตารางที่ 7.26 คำนวณค่า $a(i)$ ของคลัสเตอร์ B โดยเฉลี่ยแต่ละคอลัมน์	342
ตารางที่ 7.27 คำนวณค่า $b(i)$ ของทุกจุดในคลัสเตอร์ A ไปยังจุดในคลัสเตอร์ B	342
ตารางที่ 7.28 คำนวณค่า $b(i)$ ของทุกจุดในคลัสเตอร์ B ไปยังจุดในคลัสเตอร์ A	342
ตารางที่ 7.29 การคำนวณค่า $s(i)$	342
ตารางที่ 7.30 อาร์กิวเมนต์ของคำสั่ง <code>silhouette()</code>	343
ตารางที่ 7.31 คำอธิบายตัวแปรในชุดข้อมูล	345
ตารางที่ 7.32 แสดงจุดศูนย์กลางของแต่ละกลุ่ม	352
ตารางที่ 8.1 ส่วนสูงวัดจากรูปถ่ายและน้ำหนักข้าง	365
ตารางที่ 8.2 อาร์กิวเมนต์ของคำสั่ง <code>lm()</code>	367
ตารางที่ 8.3 อาร์กิวเมนต์ของคำสั่ง <code>predict()</code>	367
ตารางที่ 8.4 ตารางวิเคราะห์ความแปรปรวน	374
ตารางที่ 8.5 เทคนิคการวิเคราะห์อนุกรมเวลาแยกตามการมีส่วนร่วมประกอบ	387
ตารางที่ 8.6 การพยากรณ์โดยการปรับให้เรียบวิธีเฉลี่ยเคลื่อนที่อย่างง่าย	388
ตารางที่ 8.7 การคำนวณวิธีเฉลี่ยเคลื่อนที่สองครั้ง	390
ตารางที่ 8.8 การพยากรณ์โดยวิธีเฉลี่ยเคลื่อนที่สองครั้ง	391
ตารางที่ 8.9 การพยากรณ์ด้วยวิธีการปรับให้เรียบแบบเอกซ์โปเนนเชียลชั้นเดียว	393
ตารางที่ 8.10 ข้อมูลยอดขายสินค้ารายปี	397
ตารางที่ 8.11 การพยากรณ์ด้วยวิธีแนวโน้มเชิงเส้นของ Holt	401
ตารางที่ 8.12 ข้อมูลยอดขายรายไตรมาส (หน่วย : ล้านบาท)	406
ตารางที่ 8.13 การพยากรณ์โดยวิธีปรับให้เรียบของ Holt และ Winter รูปแบบบวก	407
ตารางที่ 8.14 การพยากรณ์โดยวิธีปรับให้เรียบของ Holt และ Winter รูปแบบคูณ	409
ตารางที่ 8.15 ข้อมูลความสูง พื้นที่ผิวด้านข้างและน้ำหนักข้าง	417
ตารางที่ 8.16 อาร์กิวเมนต์ของคำสั่ง <code>nnetar()</code>	421
ตารางที่ 8.17 เปรียบเทียบความแม่นยำของตัวแบบพยากรณ์ราคาบ้าน	441

บทที่ 1

แนวคิดพื้นฐานทางด้านวิทยาการข้อมูล

บทนี้มุ่งเน้นการสร้างความเข้าใจพื้นฐานเกี่ยวกับวิทยาการข้อมูล (data science) และข้อมูลขนาดใหญ่ (big data) เพื่อให้ผู้อ่านสามารถอธิบายกระบวนการทางวิทยาการข้อมูลแนวทางการประยุกต์ใช้งานจริง รวมถึงพื้นฐานการเขียนโปรแกรมภาษา R เบื้องต้นสำหรับวิทยาการข้อมูล

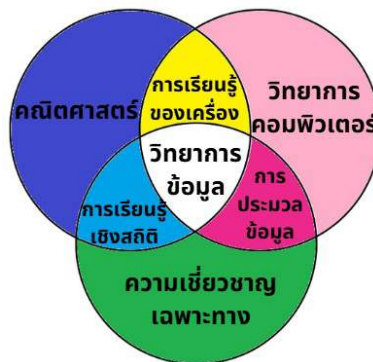
1.1 วิทยาการข้อมูลคืออะไร

ปัจจุบันวิทยาการข้อมูลเป็นที่รู้จักกันอย่างแพร่หลายถึงความสามารถในการนำข้อมูลที่มีอยู่มาวิเคราะห์ให้เกิดมูลค่าต่อองค์กร งานที่เกี่ยวข้องกับวิทยาการข้อมูลนั้นต้องอาศัยความรู้พื้นฐานหลายด้าน ไม่เพียงแต่ความรู้ด้านวิทยาการคอมพิวเตอร์เท่านั้น เนื่องจากวิทยาการข้อมูลเป็นการศึกษาเกี่ยวกับข้อมูลและสกัดข้อมูลเชิงลึกที่มีความหมายต่อการบริหารองค์กร จึงผสมผสานความรู้ทางด้านคณิตศาสตร์ สถิติ วิทยาการคอมพิวเตอร์ การวิเคราะห์ขั้นสูง ปัญญาประดิษฐ์ (artificial intelligence) การเรียนรู้ของเครื่อง (machine learning) และอาศัยความเชี่ยวชาญเฉพาะทางในด้านที่ต้องการนำวิทยาการข้อมูลไปประยุกต์ เช่น การวิเคราะห์ข้อมูลทางด้านสุขภาพ ต้องมีความเข้าใจเกี่ยวกับโรคและปัจจัยที่ทำให้เกิดโรค การวิเคราะห์ข้อมูลการขาย ต้องมีความเข้าใจเกี่ยวกับข้อมูลสินค้าและบริการ มีความเชี่ยวชาญด้านการส่งเสริมการตลาด เพื่อสกัดข้อมูลเชิงลึกภายใต้ข้อมูลขนาดใหญ่ที่องค์กรมีอยู่ โดยข้อมูลเชิงลึกเหล่านี้สามารถใช้เป็นแนวทางในการตัดสินใจและการวางแผนเชิงกลยุทธ์ได้

วิทยาการข้อมูลนั้นเป็นศาสตร์ที่มีรากฐานมาจากเทคนิคการทำเหมืองข้อมูล (data mining) ที่อาศัยทฤษฎีทางคณิตศาสตร์ (mathematical theory) การเรียนรู้เชิงสถิติ (statistical learning) และการเรียนรู้ของเครื่อง และเกี่ยวข้องกับการรวบรวมข้อมูล การเตรียมข้อมูลเพื่อวิเคราะห์ การแสดงภาพข้อมูล การจัดการข้อมูล ในอดีตข้อมูลในองค์กรเก็บไว้ในระบบฐานข้อมูลทั้งในรูปแบบรวมศูนย์และกระจาย ทำให้ปัจจุบันฐานข้อมูลในองค์กรมีหลายแหล่งและมีขนาดใหญ่แต่ขาดการสกัดข้อมูลเชิงลึกมาใช้ในการบริหารงาน อีกทั้งข้อมูลในปัจจุบันไม่ได้จัดเก็บเฉพาะในรูปของระบบฐานข้อมูลเชิงสัมพันธ์ที่มีรูปแบบตาราง (table) แถว (tuple) และคอลัมน์ (column) (Date, 2003) ซึ่งเป็นข้อมูลที่มีโครงสร้าง (structured data) ยังมีข้อมูลที่ไม่มีโครงสร้าง (unstructured data) อีกหลายประเภท เช่น ข้อความ วิดีโอ ภาพ เสียง เป็นต้น ตลอดจนฐานข้อมูลแบบ NoSQL (NoSQL database) ที่ไม่ได้มีโครงสร้าง

แบบฐานข้อมูลเชิงสัมพันธ์ ดังนั้นวิทยาการข้อมูลจึงมีความเชื่อมโยงกับวิทยาการคอมพิวเตอร์เป็นอย่างมาก เนื่องจากต้องอาศัยคอมพิวเตอร์เป็นเครื่องมือและอาศัยวิศวกรข้อมูลในการจัดการข้อมูล ที่มาจากหลายแหล่งและหลายรูปแบบให้พร้อมต่อการนำไปวิเคราะห์ด้วยเทคนิคทางด้านวิทยาการข้อมูล นักวิทยาการข้อมูลจึงต้องมีความรู้ทางด้านวิทยาการคอมพิวเตอร์ มีทักษะการเขียนโปรแกรมคอมพิวเตอร์ และมีทักษะอีกหลายด้าน เช่น การสร้างตัวแบบเชิงคณิตศาสตร์ ตัวแบบเชิงสถิติ ตัวแบบการเรียนรู้ของเครื่อง ตลอดจนทักษะเฉพาะทาง เช่น ด้านธุรกิจ ด้านการแพทย์ เป็นต้น (Stanton, 2012; Tayefi et al., 2017)

สาเหตุที่วิทยาการข้อมูลเป็นศาสตร์ที่ได้รับความนิยมเพิ่มขึ้นอย่างกว้างขวางในหลายสาขาวิชา เนื่องจากปริมาณข้อมูลขนาดใหญ่และการนำปัญญาประดิษฐ์ มาใช้ในการประมวลผลข้อมูล ซึ่งทิศทางในอนาคตจะมีการนำปัญญาประดิษฐ์มาใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่ มากขึ้น นอกจากนี้วิทยาการข้อมูลสามารถนำมาประยุกต์ในการทำงานได้จริง โดยผู้ที่มีวิชาชีพเกี่ยวกับวิทยาการข้อมูล ได้แก่ นักวิทยาการข้อมูล นักสถิติ นักวิเคราะห์ข้อมูล นักคณิตศาสตร์ หรือผู้ที่ทำงานด้านปัญญาประดิษฐ์ กล่าวโดยสรุป วิทยาการข้อมูลคือ การรวบรวมข้อมูล การเข้าใจข้อมูล การสร้างตัวแบบ เพื่อวิเคราะห์ แก้ไขปัญหาและเพื่อช่วยในการตัดสินใจ



รูปที่ 1.1 Data Science Venn Diagram

จากที่ได้กล่าวมาแล้วว่าวิทยาการข้อมูลมีรากฐานมาจากการทำเหมืองข้อมูลซึ่งอาศัยทฤษฎีทางคณิตศาสตร์ การเรียนรู้เชิงสถิติและการเรียนรู้ของเครื่อง นอกจากนี้วิทยาการคอมพิวเตอร์ และความเชี่ยวชาญเฉพาะทางยังมีบทบาทสำคัญอย่างมากในด้านวิทยาการข้อมูล (Cleveland, 2001) ดังนั้นองค์ประกอบที่สำคัญของวิทยาการข้อมูลสามารถอธิบายได้โดยแผนภาพเวนน (Palmer, 2015) ดังรูปที่ 1.1 โดยการประมวลผลข้อมูลคือ ทักษะการเขียนโปรแกรมเพื่อดึงข้อมูลที่อยู่ในรูปแบบหลากหลายไม่ว่าจะเป็นข้อมูลที่ถูกบันทึกในระบบฐานข้อมูล หรือข้อมูลที่ไม่มีโครงสร้างเพื่อเตรียมข้อมูลให้ครบถ้วนและเหมาะสมสำหรับการวิเคราะห์ข้อมูล หลังจากรวบรวมข้อมูลและทำความสะอาดข้อมูลเรียบร้อยแล้ว

แล้วจะต้องทำความเข้าใจข้อมูล มองหาข้อมูลเชิงลึกที่จะตอบโจทย์ปัญหาโดยการประยุกต์ใช้วิธีการทางคณิตศาสตร์ สถิติศาสตร์ การเรียนรู้ของเครื่อง และความเชี่ยวชาญในงานเฉพาะทาง ดังนั้นผู้ที่ทำงานในสาขานี้ไม่เพียงสามารถใช้เครื่องมือหรือซอฟต์แวร์ในการสร้างตัวแบบทางคณิตศาสตร์ หรือตัวแบบทางสถิติได้ แต่ต้องสามารถอธิบายผลลัพธ์ของตัวแบบนั้นได้ด้วย เช่น การสร้างตัวแบบการถดถอยเชิงเส้นอย่างง่าย ต้องสามารถอธิบายสัมประสิทธิ์การถดถอยของตัวแบบการถดถอยเชิงเส้นอย่างง่ายได้ ยกตัวอย่างตัวแบบการถดถอยเชิงเส้นอย่างง่ายที่แสดงความสัมพันธ์ระหว่างยอดขายสินค้ากับราคาสินค้า สัมประสิทธิ์การถดถอยอธิบายอัตราการเปลี่ยนแปลงของยอดขายเมื่อราคาสินค้าเปลี่ยนไป 1 หน่วย ดังนั้นนักวิทยาการข้อมูลจึงจำเป็นต้องมีทักษะทั้งด้านวิทยาการคอมพิวเตอร์ คณิตศาสตร์ สถิติศาสตร์ การเรียนรู้ของเครื่องและมีความเข้าใจหรือความเชี่ยวชาญงานที่กำลังแก้ไขปัญหา

1.2 ข้อมูลขนาดใหญ่

ปัจจุบันรูปแบบการสื่อสารระหว่างผู้คนเปลี่ยนไปคนนิยมสื่อสารผ่านโทรศัพท์มือถือ แท็บเล็ต คอมพิวเตอร์พกพา โดยสนทนาผ่านข้อความสั้นในมือถือ อีเมล เครือข่ายสังคม เช่น Facebook Line, Twitter และ LinkedIn เป็นประจำ ทำให้ทุกองค์กรต้องเปลี่ยนแปลงรูปแบบการสื่อสารระหว่างบริษัท กับลูกค้าเช่นกัน ดังจะเห็นได้ว่าทุกองค์กรจะมี Facebook Page, Line Official, Instagram ในการโฆษณาสินค้าและบริการ นอกจากนี้ยังมีการพิมพ์ QR code บนผลิตภัณฑ์หรือแผ่นพับโฆษณาเพื่อให้ลูกค้าเข้าถึงสินค้าและบริการของบริษัทได้ง่ายขึ้น สามารถสั่งซื้อสินค้าผ่านระบบออนไลน์ เหล่านี้เป็นการทำการตลาดดิจิทัล (digital marketing) ของบริษัทในยุคปัจจุบัน นอกจากนี้การใช้ระบบออนไลน์ยังช่วยให้บริษัทสามารถรับฟังเสียงของลูกค้าเกี่ยวกับสินค้าหรือบริการได้อย่างรวดเร็ว ดังนั้นข้อมูลการซื้อสินค้าหรือบริการและเสียงของลูกค้านี้เองทำให้ทุกองค์กรต่าง ๆ มีข้อมูลขนาดใหญ่และต้องการเทคโนโลยีสำหรับจัดการข้อมูลขนาดใหญ่เหล่านี้ เนื่องจากการประมวลผลข้อมูลที่รวดเร็วเพื่อการตัดสินใจในองค์กรเป็นสิ่งสำคัญและสามารถสร้างมูลค่าทางเศรษฐกิจแก่บริษัทได้มากมาย

ข้อมูลขนาดใหญ่ถูกนิยามด้วย 3 V ได้แก่ ปริมาตรหรือปริมาณข้อมูลมหาศาล (Volume) ความเร็ว (Velocity) ข้อมูลถูกสร้างอย่างรวดเร็วและเปลี่ยนแปลงอย่างต่อเนื่อง และความหลากหลาย (Variety) ข้อมูลมีหลากหลายประเภท เช่น วิดีโอ เสียง ภาพ เครือข่ายสังคม เว็บไซต์ อีเมล ระบบเซนเซอร์ จากรายงาน Data never sleeps 12.0 ของเว็บไซต์ domo.com (Domo, 2024) พบว่าในพ.ศ. 2567 ประชากรที่เข้าถึงอินเทอร์เน็ตมีประมาณ 67.5% ของประชากรโลก หรือคิดเป็นประมาณ 5.5 พันล้านคน และคาดการณ์ว่าข้อมูลที่ถูกสร้างและจัดเก็บทั่วโลกจะเพิ่มขึ้นเป็น 394 เซตตะไบต์ ภายในปีพ.ศ. 2571 ข้อมูลขนาดใหญ่เหล่านี้ทำให้บริษัทผู้ให้บริการทางอินเทอร์เน็ตสามารถทราบข้อมูลส่วนตัว เช่น สิ่งที่เราสนใจ รสนิยมของเรา สถานที่ที่เราชอบเดินทางไป เว็บไซต์ที่ถูกใจ และเครือข่ายเพื่อนของเราเป็นใครบ้าง เป็นต้น



รูปที่ 1.2 สิ่งที่มีบทบาทสำคัญต่อ Big Data

ในการประมวลผลข้อมูลขนาดใหญ่เพื่อสกัดข้อมูลเชิงลึกที่จะเป็นประโยชน์ต่อการตัดสินใจขององค์กรต้องอาศัยระบบการจัดการข้อมูลสมัยใหม่ที่สามารถรองรับข้อมูลที่มีความแตกต่างทั้งด้านปริมาณและความหลากหลาย รูปที่ 1.2 แสดงสิ่งที่มีบทบาทสำคัญต่อการจัดการข้อมูลขนาดใหญ่ เช่น Hadoop ของ Yahoo (Apache Software Foundation, 2010) ซึ่งเป็นระบบการจัดเก็บข้อมูลขนาดใหญ่และมีโครงสร้าง MapReduce ที่มีความสามารถในการประมวลผลข้อมูลขนาดใหญ่ (Schönberger & Cukier, 2013) ทั้งที่เป็นข้อมูลที่มีโครงสร้าง เช่น ระบบฐานข้อมูลเชิงสัมพันธ์ และฐานข้อมูลที่ไม่มีโครงสร้างที่ไม่สามารถเข้าถึงโดยใช้ภาษาการจัดการฐานข้อมูลแบบ SQL เช่นในระบบฐานข้อมูลเชิงสัมพันธ์ จึงมีการพัฒนาฐานข้อมูลแบบ NoSQL ซึ่งเป็นภาษาที่ไม่จำเป็นต้องมีการกำหนดโครงสร้างของข้อมูลไว้ล่วงหน้า ในส่วนของการวิเคราะห์ข้อมูลขนาดใหญ่จำเป็นต้องใช้เทคโนโลยีการประมวลผลที่มีประสิทธิภาพสูง (high performance computing) หรือหากประมวลผลข้อมูลขนาดใหญ่ที่มาจากอุปกรณ์ IOT จะใช้ Edge computing และอาศัยวิธีการทางสถิติ คณิตศาสตร์และปัญญาประดิษฐ์ในการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อหาข้อสรุปหรือหาความเป็นไปได้ต่าง ๆ เช่น สามารถคาดการณ์ว่าลูกค้ามีแนวโน้มจะซื้อผลิตภัณฑ์ใหม่หรือไม่จากข้อมูลพฤติกรรมการซื้อสินค้าในอดีต สามารถคาดการณ์แนวโน้มการฆ่าตัวตาย จากข้อความที่สื่อสารผ่านทางเครือข่ายสังคมต่าง ๆ เป็นต้น

1.3 กระบวนการของวิทยาการข้อมูล

กระบวนการของวิทยาการข้อมูลเกี่ยวกับการประมวลผลข้อมูลขนาดใหญ่เพื่อสกัดสารสนเทศจากข้อมูลโดยกระบวนการประกอบด้วย 5 ขั้นตอนดังรูปที่ 1.3 โดยมีรายละเอียดดังนี้



รูปที่ 1.3 กระบวนการของวิทยาการข้อมูล

ขั้นตอนที่ 1 เป็นการทำความเข้าใจปัญหา ในขั้นตอนนี้นักวิทยาการข้อมูลต้องทำความเข้าใจในงานที่เกี่ยวข้อง เช่น ด้านธุรกิจ ต้องทำความเข้าใจธุรกิจ สิ่งที่ธุรกิจต้องการทราบ และข้อมูลใดบ้างที่จำเป็นต่อการแก้ไขปัญหา

ขั้นตอนที่ 2 เป็นการรวบรวมข้อมูลที่เกี่ยวข้อง ซึ่งได้จากการทำความเข้าใจปัญหาในขั้นตอนที่ 1

ขั้นตอนที่ 3 เป็นการเตรียมพร้อมก่อนการวิเคราะห์ข้อมูล ในขั้นตอนนี้ต้องทำการสำรวจข้อมูลและมองภาพข้อมูลเพื่อจะได้ทราบว่าข้อมูลมีปัญหาหรือไม่ ข้อมูลที่รวบรวมมานั้นอาจไม่ครบถ้วนและไม่สอดคล้องกัน ดังนั้นจึงต้องทำความสะอาดข้อมูล (cleaning) การรวบรวมข้อมูลและการเชื่อมโยงข้อมูลหากข้อมูลมาจากหลายแหล่ง (integration) ซึ่งในบางครั้งอาจต้องปรับโครงสร้างของข้อมูลใหม่เพื่อรองรับข้อมูลที่มาจากหลายแหล่ง การแปลงข้อมูล (data transformation) ให้อยู่ในรูปแบบที่เหมาะสมต่อการนำไปวิเคราะห์ข้อมูล นอกจากนี้อาจจำเป็นต้องทำการกรองข้อมูล (filtering) การลบข้อมูล (deleting) การตรวจสอบข้อมูล (verifying) เพื่อตรวจสอบค่าสูญหาย (missing value) และการแทนค่าข้อมูลสูญหาย (imputation) ตรวจสอบค่าผิดปกติ (outlier) และการแก้ไขค่าผิดปกติ

ขั้นตอนที่ 4 เป็นการวิเคราะห์ข้อมูลและสร้างตัวแบบ ในส่วนนี้อาจรวมถึงการวิเคราะห์ความสัมพันธ์ของตัวแปร เพื่อคัดเลือกตัวแปรที่มีความเหมาะสมในการสร้างตัวแบบ (modelling) ไปจนถึงการสร้างตัวแบบทางคณิตศาสตร์ สถิติหรือการเรียนรู้ของเครื่อง การวิเคราะห์ข้อมูลขั้นสูง

ขั้นตอนที่ 5 ทำการประเมินผลประสิทธิภาพของตัวแบบที่ได้จากขั้นตอนที่ 4 และนำเสนอผลการทดลองตัวแบบที่พัฒนา นำเสนอผลการวิเคราะห์ข้อมูล ตลอดจนการแสดงผลภาพของข้อมูล และนำผลการศึกษาไปใช้ในการทำงานต่อไป

1.4 วิทยาการข้อมูลในการทำงาน

ปัจจุบันวิทยาการข้อมูลได้ถูกนำมาประยุกต์ในการแก้ไขปัญหามากมายยกตัวอย่างเพียงบางด้าน ดังนี้

■ การสืบค้นข้อมูล

บริษัทที่ให้บริการสืบค้นข้อมูล เช่น Google ได้ใช้ประโยชน์จากวิทยาการข้อมูลในการที่จะส่งผลลัพธ์จากการค้นของลูกค้าให้ตรงที่สุดและเร็วที่สุดในเวลาไม่นาน นอกจากนี้ยังสามารถทำการสืบค้นผ่านรูปภาพและเสียงโดยใช้การเรียนรู้ของเครื่องและปัญญาประดิษฐ์ ได้แก่ การรู้จำรูปภาพ (image recognition) และการรู้จำเสียง (speech recognition)

■ การตลาดดิจิทัล

บริษัทต่าง ๆ ได้เก็บรวบรวมข้อมูลพฤติกรรมลูกค้าจากสื่อสังคมออนไลน์ และนำข้อมูลมาวิเคราะห์หาความเป็นไปได้ว่าถ้าบริษัทนำเสนอสินค้าผ่านโฆษณาของสื่อออนไลน์แล้วลูกค้าจะรับชมโฆษณาสินค้าหรือเลื่อนผ่าน และความเป็นไปได้ที่ลูกค้าจะตัดสินใจซื้อสินค้า เพื่อวางแผนการนำเสนอข่าวสารให้ตรงกับความสนใจของลูกค้าโดยอาศัยการโฆษณาผ่านสื่อออนไลน์ เช่น Facebook, YouTube หรือระบบโฆษณาของ google เป็นต้น

■ ระบบให้คำแนะนำแก่ลูกค้า

บริษัทต่างๆ มีเว็บไซต์ในการโฆษณา ขายสินค้าหรือบริการ เมื่อเราเข้าไปชมสินค้าเว็บไซต์ไหนก็ตามจะมีการแนะนำสินค้าที่เราอาจสนใจ ซึ่งเกิดจากการที่บริษัทนำวิทยาการข้อมูลมาช่วยในการวิเคราะห์ความสนใจของลูกค้าที่มีพฤติกรรมชมสินค้า การซื้อสินค้าที่คล้ายกับเรานั้นเอง ตัวอย่างเว็บไซต์ขายหนังสือชื่อดัง amazon.com เว็บไซต์จะทำการช่วยหาสินค้าที่ใกล้เคียงกับสินค้าลูกค้าเลือกจากจำนวนสินค้าที่มีมากมายในเวลาอันรวดเร็ว นอกจากนี้ยังอาศัยข้อมูลการคลิกของลูกค้าในอดีตมาวิเคราะห์เพื่อปรับปรุงการให้บริการและคำแนะนำแก่ลูกค้าท่านใหม่ เช่น เมื่อลูกค้าเลือกดูหนังสือเล่มหนึ่งระบบก็จะแนะนำหนังสือเล่มอื่นที่ลูกค้าคนอื่นสนใจเช่นกัน

■ ระบบสั่งการด้วยเสียง

ตัวอย่างของระบบสั่งการด้วยเสียงที่รู้จักกันเป็นอย่างดี คือ Siri Google voice ซึ่งนำมาช่วยในการตอบสนองความต้องการของผู้ใช้ผ่านการออกเสียงคำ เช่น การพูดออกเสียงผ่านแอปพลิเคชันเครือข่ายสังคมในโทรศัพท์มือถือแทนการพิมพ์ข้อความ และโปรแกรมจะเปลี่ยนเสียงให้เป็นข้อความที่โปรแกรมสามารถทำได้เพราะโปรแกรมได้เรียนรู้จากข้อความเสียงมากมายโดยการสร้างระบบรู้จำเสียงยังมีการเรียนรู้โดยใช้ข้อความเสียงจำนวนมากโปรแกรมจะยิ่งแปลงเสียงเป็นข้อความได้ถูกต้องมากขึ้นในปัจจุบัน Call center หลายแห่งได้นำระบบรู้จำเสียงมาช่วยในการตอบปัญหาของลูกค้า โดยลูกค้า