การจัดการความรู้

ความรู้เบื้องต้นเกี่ยวกับ Big Data และ Machine Learning

ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร
กรมการจัดหางาน

คำนำ

การจัดการความรู้ (Knowledge Management : KM) เรื่อง ความรู้เบื้องต้นเกี่ยวกับ Big Data และ Machine Learning ซึ่งสนับสนุนประเด็นยุทธศาสตร์การพัฒนาระบบเทคโนโลยีสารสนเทศ นวัตกรรม ในการบริหารจัดการ และการบริการ ของกรมการจัดหางานประจำปี พ.ศ. ๒๕๖๒ และสอดคล้องกับยุทธศาสตร์ที่ ๕ ของแผนพัฒนาดิจิทัลเพื่อเศรษฐกิจและสังคม ในเรื่องการพัฒนากำลังคนให้พร้อมเข้าสู่เศรษฐกิจและสังคมดิจิทัล เพื่อให้สามารถใช้ประโยชน์จากเทคโนโลยีดิจิทัลได้อย่างรอบรู้ เท่าทันการเปลี่ยนแปลงของเทคโนโลยียุคปัจจุบันและอนาคต

ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร จึงได้จัดทำเอกสารการจัดการความรู้เกี่ยวกับความรู้เบื้องต้นเกี่ยวกับ Big Data และ Machine Learning เพื่อเป็นประโยชน์อย่างยิ่งกับบุคลากรภายในกรมการจัดหางานให้มีความรู้ ความเข้าใจเกี่ยวกับข้อมูลขนาดใหญ่และการวิเคราะห์ข้อมูล เพื่อนำความรู้ที่ได้นำไปประยุกต์ใช้กับงานที่รับผิดชอบได้

6 มิถุนายน 2562

คณะผู้จัดทำ

ศูนย์เทคโนโลยีสารสนเทศ กรมการจัดหางาน

สารบัญ

หน้า

คำนำ.....2
บทที่ 1 ความหมายของ Big Data.....4
บทที่ 2 ข้อมูล Big Data.....8
บทที่ 3 บุคคลากรและทักษะที่ใช้ในการทำ Big Data 12
บทที่ 4 เริ่มทำ Big Data Project.....15
บทที่ 5 Machine Learning เบื้องต้น.....21
เอกสารอ้างอิง.....24

บทที่ 1 ความหมายของ Big Data

คำว่า Big Data เข้ามาในประเทศไทยเมื่อไหร่ไม่มีการสำรวจแน่ชัด แต่แนวโน้มของการใช้ Big Data ในไทยเริ่มเห็นเด่นชัดเมื่อปี 2016 ในช่วงที่รัฐบาลมีการผลักดันให้เกิด Thailand Digital 4.0 แต่อย่างไรก็ตาม คนส่วนใหญ่ยังคงติดภาพว่า Big Data คือการเก็บข้อมูลเอาไว้เยอะๆ โดยที่ยังไม่มีเป้าหมายชัดเจนของการทำงาน

Big Data คือ ปริมาณข้อมูลจำนวนมหาศาลที่มีอยู่ในองค์กรทุกรูปแบบ ไม่ว่าแหล่งที่มาจะมาจากภายในองค์กรหรือภายนอกก็ตาม ทั้งนี้แบ่งออกเป็นข้อมูลที่มีโครงสร้างชัดเจน (Structured Data) และข้อมูลที่มีโครงสร้างไม่ชัดเจน (Unstructured Data) โดยมีองค์ประกอบ 5V ดังปรากฏตามรูปภาพที่ 1

รูปที่ 1 THE 5 Vs OF BIG DATA [7]

1. Volume หมายถึง ขนาดของข้อมูล แน่นอนว่าคำว่า “Big Data” ก็ทำให้เราเห็นภาพแล้วว่าต้องมีขนาดใหญ่ แต่ไม่มีการระบุความใหญ่ที่ชัดเจนได้ เนื่องจากข้อมูลที่เป็น Big Data ย่อมสามารถขยายตัวต่อไปได้ไม่หยุดอยู่กับที่ จากปริมาณข้อมูลที่มากทำให้การจัดการเก็บข้อมูลไม่สามารถใช้วิธีการจัดการแบบปกติได้ เช่น การจัดเก็บข้อมูลใน Excel หรือ การจัดเก็บข้อมูลลงฐานข้อมูล

2. Velocity หมายถึง ความเร็วทั้งจากการสร้างข้อมูลและการประมวลผลข้อมูล ที่สามารถจัดการข้อมูลมีขนาดใหญ่ ได้อย่างต่อเนื่องแบบ Real-Time เพื่อให้การจัดการดำเนินการสำเร็จ

อย่างรวดเร็ว เราต้องมีการวางกระบวนการทำงานที่ชัดเจนอีกด้วย เพราะหากเกิดความผิดพลาดต้องมีการแก้ไขได้โดยเร็วเช่นกัน

3. Variety หมายถึง ความหลากหลายของข้อมูลและชนิดของข้อมูล ทั้งข้อมูลประเภทที่มีโครงสร้าง (Structured) ได้แก่ ข้อมูลแบบตารางที่เก็บไว้ในฐานข้อมูล ประกอบไปด้วยข้อมูลที่เป็นตัวเลข ตัวหนังสือ และวันเดือนปี เป็นต้น และข้อมูลที่ไม่มีโครงสร้าง (Unstructured) ได้แก่ ข้อมูลที่เป็นรูปภาพ เสียง วีดีโอ และข้อมูลการแสดงความคิดเห็นต่างๆผ่านทางโซเชียลมีเดีย (Social Media) เป็นต้น

4. Veracity หมายถึง ความถูกต้องแม่นยำ เพราะข้อมูลประเภท Big Data มีขนาดใหญ่ที่ต้องการความเร็วในการใช้งาน และมีความหลากหลายสูง ดังนั้นในตัวของข้อมูลเองจะมีความไม่แน่นอนรวมอยู่ด้วย ซึ่งเกิดจาก Error ต่างๆ ระหว่างการสร้างข้อมูล หรือเป็นข้อมูลที่อยู่เกินขอบเขตของข้อมูลที่สามารถเป็นไปได้ ตัวอย่างเช่น ข้อมูลจังหวัด กรุงเทพมหานคร สามารถเขียนได้ว่า กทม. หรือ กรุงเทพ หรือ กรุงเทพฯ เป็นต้น แน่นอนว่าการทำให้ข้อมูลสะอาด ไม่มีการซ้ำซ้อนของชุดข้อมูล เป็นเรื่องที่ยากลำบากที่สุด และเป็นขั้นตอนที่ใช้เวลานานที่สุด แต่ถือว่าเป็นสิ่งสำคัญที่สุดของการทำ Big Data Project

5. Value หมายถึง ข้อมูลที่มีคุณค่า สามารถนำไปใช้ประโยชน์ได้ หรือมีมูลค่าและความสำคัญต่อธุรกิจในการนำมาใช้ประโยชน์ เช่นการนำข้อมูลไปวิเคราะห์ การสรุปผลเพื่อที่จะนำข้อมูลที่วิเคราะห์ได้ไปวางแผนการขับเคลื่อนธุรกิจ เพื่อสร้างมูลค่าของสินค้าหรือจูงใจให้อยากใช้บริการ

<u>การนำ Big Data เข้ามาใช้ในภาครัฐ</u>

การใช้ประโยชน์จาก Big Data ในภาคธุรกิจนั้นเป็นประโยชน์โดยตรงในการนำข้อมูลมาหาค่าเชิงสถิติ และพัฒนาผลิตภัณฑ์ให้ตรงกับความต้องการและพฤติกรรมผู้บริโภค และเกิดผลิตภัณฑ์ใหม่ (Enabling New Products) ส่วนในภาครัฐนั้นยังมีการใช้ประโยชน์จาก Big Data น้อยมากเมื่อเทียบกับภาคธุรกิจ โดยการใช้ประโยชน์จาก Big Data ของภาครัฐ คือ การนำมาพัฒนาการบริการภาครัฐให้ตรงต่อความต้องการของประชาชนให้มีประสิทธิภาพสูงขึ้น และใช้งบประมาณน้อยลง เช่น นำมาวิเคราะห์ข้อมูลสภาพอากาศ ที่มีปริมาณมหาศาล ทั้งข้อมูลจากเครื่องวัดมากมายบนโลก ทั้งดาวเทียม เรดาร์ และยานตรวจอากาศ ข้อมูลที่มากมายมหาศาลเหล่านี้ นำมาสู่การพยากรณ์อากาศที่แม่นยำ หรือการถอดรหัสพันธุกรรม เพื่อทำแผนที่ของสิ่งมีชีวิตต่างๆ บนโลก ซึ่งนำไปสู่การค้นพบตัวยารักษาโรคใหม่ๆ เป็นต้น

การนำไปใช้ประโยชน์ด้านต่างๆในประเทศไทย เช่น

1. การนำข้อมูลภัยพิบัติจากธรรมชาติ โดยการวิเคราะห์ที่ก้าวหน้าการจัดการโมเดล เช่น การวิเคราะห์สถานการณ์จากโมเดลที่คาดว่าจะส่งผลอันตรายมาพยากรณ์เหตุการณ์หรืออาจจะพัฒนาโปรแกรมแจ้งเตือนภัย เป็นต้น

2. การนำข้อมูลมาปรับปรุง และวิธีการฟื้นฟูแก้ไขสภาพแวดล้อมให้กลับไปสู่สภาพเดิม โดยใช้เทคโนโลยีมาวิเคราะห์ข้อมูล การทดลอง และการทำนายเพื่อช่วยในการตัดสินใจในการแก้ไขปัญหาเพื่อวางแผนยุทธศาสตร์ ทั้งในระยะยาวและระยะสั้น

3. การนำข้อมูลมาสร้างนวัตกรรม และความรู้พื้นฐานขั้นสูง เพื่อการสร้างโครงสร้างพื้นฐานของการบริการ และการนำข้อมูลเหล่านั้นมารวบรวม และสร้างเป็นฐานข้อมูลขนาดใหญ่ (Big Data) และปรับปรุงฐานความรู้ที่กว้างขวาง เพื่อเตรียมรับและตอบสนองต่อสังคม

การประยุกต์ใช้งาน Big Data นั้นสามารถนำมาใช้งานได้หลายๆ หน่วยงาน เช่น ด้านสาธารณสุข ด้านวิทยาศาสตร์ ด้านความมั่นคง ด้านการเงิน ด้านการบริการประชาชน ด้านเกษตรกรรม ด้านคมนาคม และด้านแรงงาน นอกจากนี้ยังสามารถนำมาใช้ในด้านการบริหารเงินงบประมาณ และรายได้ต่างๆ ของภาครัฐให้มีประสิทธิภาพมากขึ้นอีกด้วย

<u>นโยบายรัฐบาลเกี่ยวกับ Big Data</u>

รัฐบาล พลเอกประยุทธ์ จันทร์โอชา นายกรัฐมนตรีได้ผลักดันภาครัฐ ให้สู่ความเป็นเลิศตามวิสัยทัศน์ประเทศไทยปี พ.ศ. 2558-2563 โดยมีวัตถุประสงค์ “มั่นคง มั่งคั่ง และยั่งยืน” ของประเทศ มีการปรับเปลี่ยนภาครัฐสู่การเป็นรัฐบาลดิจิทัล การพัฒนาให้เกิดการใช้ข้อมูลมหาศาล หรือ Big Data ของภาครัฐ จะนำไปสู่การบูรณาการด้านโครงสร้างข้อมูล การจัดการข้อมูล การวิเคราะห์หาคุณค่าจากข้อมูล การเลือกใช้ข้อมูลที่มีความเหมาะสม และส่งเสริมให้เกิดการบริหารจัดการข้อมูลภาครัฐให้ได้ประโยชน์สูงสุดอย่างมีประสิทธิภาพ มิติการใช้ข้อมูลที่เกี่ยวข้องกับคน ได้แก่ เศรษฐกิจ สังคม ภัยพิบัติ สาธารณสุข การศึกษา และแรงงาน

<u>การเริ่มทำ Big Data Project ของหน่วยงาน</u>

เมื่อหลายองค์กร พยายามจะผลักดัน Big Data Project โดยที่ไม่มีเป้าหมายในการทำงานที่ชัดเจน โดยหัวข้อนี้จะพูดถึงหลักการง่ายๆ ในการทำ Big Data Project ดังนี้

1. สำรวจกระบวนงานแต่ละส่วนขององค์กร ว่าส่วนใดที่ยังเป็นแบบ Manual หรือเกิดการตัดสินใจโดย “มนุษย์” อยู่บ้าง

ในความเป็นจริงแล้วประสบการณ์ รวมถึงความคิดของมนุษย์ต่างเป็นกลไกการประมวลผลรูปแบบหนึ่ง การทำ Big Data Project จะสามารถเข้ามาเป็นเครื่องมือในการสร้าง Model เพื่อช่วยเรื่องการตัดสินใจทำให้กระบวนการทำงานเป็นไปด้วยความรวดเร็ว ลดความผิดพลาดในการตัดสินใจโดยมนุษย์มากขึ้น เพราะมนุษย์แต่ละคน มีการตัดสินใจที่ไม่เหมือนกัน ทำให้การควบคุมคุณภาพของผลงานไม่ชัดเจน

2. สำรวจดูว่าองค์กรมีข้อมูลอยู่ตรงไหน อย่างไรบ้าง หรือมีแนวทางการเก็บข้อมูลเพิ่มเติมจากส่วนใดได้บ้าง

ศึกษาข้อมูลที่สามารถนำมาวิเคราะห์ต่อยอดทำให้เกิดประโยชน์ต่อไป ในกรณีองค์กรยังไม่มีการเก็บข้อมูลมาก่อน หรือไม่รู้ว่าจะนำข้อมูลส่วนไหนมาใช้ในการทำ Big Data อาจจะต้องใช้แนวทางของ Design Thinking หรือการออกแบบการเก็บข้อมูล เพื่อให้ได้ข้อมูลที่ต้องการ ในแบบที่ผู้ให้ข้อมูลไม่รู้สึกลำบากใจในการให้ข้อมูล ตัวอย่างในการให้ข้อมูลด้านการหางาน เช่น อัตราเงินเดือนที่ต้องการ งานที่ต้องการ เพื่อให้ทราบข้อมูลความต้องการ (Demand) ที่แท้จริงเพื่อนำไปออกแบบระบบการจัดหางานภายในประเทศที่ถูกต้องต่อไป

3. เข้าร่วมสัมมนาหาความรู้พัฒนาทักษะในด้านข้อมูล หรือหาที่ปรึกษาเข้ามานำเสนอเทคโนโลยีใหม่ๆ

ในการทำให้ Big Data เกิดผลลัพธ์ได้นั้น จะต้องมีส่วนที่เป็นการวิเคราะห์ไม่ว่าจะเป็นการ Analytics , Machine Learning และ Data Science ซึ่งเป็นการวิเคราะห์ที่มีการใช้เครื่องมือและแนวทางการคิดที่แตกต่างกัน ดังนั้นหากภายในองค์กรต้องการเปลี่ยนข้อมูล (Data) ธรรมดาเป็นสารสนเทศ (Information) และเปลี่ยนเป็นข้อมูลที่เกี่ยวข้องกับกลุ่มเป้าหมาย (Insight) และเปลี่ยนเป็นข้อมูลที่ช่วยในการตัดสินใจ (Right Decision) ได้นั้น องค์กรต้องมีบุคลากรที่เข้าใจเรื่องการวิเคราะห์ หรือสร้างโมเดลทางคณิตศาสตร์อย่างถ่องแท้ และตามทันการเปลี่ยนแปลงของเทคโนโลยีอย่างรวดเร็วสามารถใช้เครื่องมือได้อย่างถูกต้อง

4. เปลี่ยน Mindset หรือแนวคิดในการทำงาน

การปรับเปลี่ยนทัศนคติองค์กร ต้องมีการปรับความคิดของผู้นำ ไม่ยึดติดอยู่กับสิ่งเดิม ต้องเปิดใจยอมรับถึงการเปลี่ยนแปลง และอย่ากลัวการผิดพลาด เพราะความผิดพลาดอาจจะนำพามาซึ่งการพัฒนาแบบก้าวกระโดด ที่สำคัญคือต้องพยายามให้เกิดผลงานโดยเร็ว เพื่อนำพาองค์กรไปสู่ความสำเร็จในการให้บริการประชาชน เพราะเทคโนโลยีต่างๆ เปลี่ยนแปลงไปอย่างรวดเร็ว

บทที่ 2 ข้อมูล Big Data

ข้อมูล (Data) คือ ข้อเท็จจริงที่เกิดขึ้น ข้อมูลอาจจะอยู่ในรูปแบบข้อความหรือตัวเลข ซึ่งข้อความเหล่านี้อาจเป็นเรื่องที่เกี่ยวข้องกับ คน สัตว์ สิ่งของ เช่น ความคิดเห็นของคนเกี่ยวกับการเลือกตั้ง เป็นต้น

โดย Big Data จะสามารถแบ่งข้อมูลออกได้เป็น 3 ประเภท

1. Structure Data หรือข้อมูลแบบมีโครงสร้าง เป็นข้อมูลที่มีลักษณะบ่งบอกชัดเจน สามารถเก็บในรูปแบบของ Relational Database หรือระบบฐานข้อมูลเชิงสัมพันธ์ ซึ่งเป็นการเก็บข้อมูลให้อยู่ในรูปแบบของตาราง (Table) ที่มีการจัดเรียงอย่างเป็นรูปแบบที่ชัดเจนและเป็นระเบียบ สามารถนำวิเคราะห์ได้เลย โดยส่วนประกอบของตารางจะแบ่งออกเป็นแถว (Row) และคอลัมน์ (Column) สามารถใช้ภาษา SQL ในการบริหารจัดการข้อมูลได้ เช่น ข้อมูลที่เก็บไว้ในฐานข้อมูล หรือข้อมูลที่เก็บไว้ในโปรแกรม spreadsheet อย่าง Microsoft Excel เป็นต้น

	ภายในอาเซียน (1)			การย้ายถิ่นรวม (2)			สัดส่วน (1) / (2) (%)	
	ออก	เข้า	ออก/เข้า	ออก	เข้า	ออก/เข้า	ออก	เข้า
บรูไน	9,313	120,578	0.08	24,343	148,123	0.16	38.26	81.40
กัมพูชา	53,722	320,573	0.17	350,485	335,829	1.04	15.33	95.46
อินโดนีเซีย	1,518,687	158,485	9.58	2,504,297	397,124	6.31	60.64	39.91
ลาว	82,788	10,134	8.17	366,663	18,916	19.38	22.58	53.58
มาเลเซีย	1.195.566	1,882,987	0.63	1,481,202	2,357,603	0.63	80.72	79.87
เมียนมาร์*	321,100	814	394.47	514,667	98,008	5.25	62.39	0.83
ฟิลิปปินส์	335,407	9,096	36.87	4,275,612	435,423	9.82	7.84	2.09
สิงคโปร์	122,254	1,162,960	0.11	297,234	1,966,865	0.15	41.13	59.13
ไทย	262,721	448,218	0.59	811,123	1,157,263	0.70	32.39	38.73
เวียดนาม	221,956	21,511	10.32	2,226,401	69,307	32.12	9.97	31.04
รวม	**4,123,515**	**4,135,357**	**1.00**	**12,852,027**	**6,984,461**	**1.84**	**32.08**	**59.21**

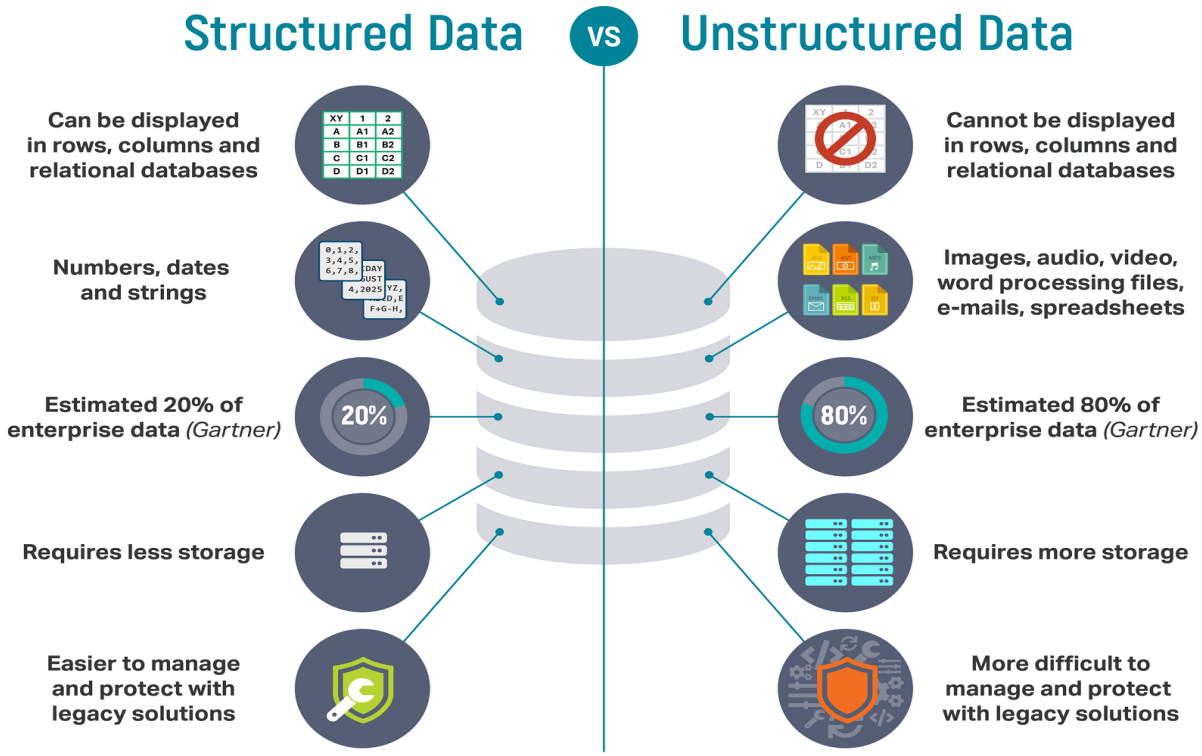
รูปที่ 2 ตัวอย่างข้อมูลแบบ Structure

2. Semi Structure Data หรือข้อมูลแบบกึ่งโครงสร้าง เป็นข้อมูลที่ถูกจัดเก็บอย่างมีรูปแบบ ในระดับหนึ่ง และข้อมูลที่สามารถค้นหา (Search) หรือแท็ก (Tag) ได้ เช่น เว็บเพจที่มีการระบุชื่อเพจ คำสำคัญในเพจ และวันที่อัพเดทข้อมูล

3. Unstructured Data หรือ ข้อมูลแบบไม่มีโครงสร้าง เป็นข้อมูลที่ไม่สามารถระบุรูปแบบได้แน่นอน ข้อมูลประเภทนี้เป็นได้ทั้ง ข้อความ รูปภาพ วิดีโอ และเสียงและไม่สามารถจัดเก็บในรูปแบบ Relational Database ได้ ดังนั้นจึงจำเป็นต้องเลือกที่จะเก็บข้อมูลในรูปแบบ Non-Relational Database และสามารถใช้เครื่องมือที่เป็น No-SQL (No-only SQL) จัดการ ซึ่งมีข้อดี

ในส่วนที่สามารถรองรับเรื่องการขยายตัวของข้อมูลอีกด้วย เช่น ข้อมูลการแสดงความคิดเห็นบนเฟสบุค (Facebook) ที่มีทั้งข้อความในรูปแบบของข้อความ รูปภาพ วีดีโอ เสียง สติ๊กเกอร์ เป็นต้น

ปัจจุบันข้อมูลแบบมีโครงสร้าง (Structure) มีอัตราส่วนอยู่ที่ประมาณ 20% ของข้อมูลทั้งหมดที่มีอยู่ในโลกใบนี้ ซึ่งเป็นรูปแบบข้อมูลที่สามารถนำไปจัดการได้ง่าย เนื่องจากมีความชัดเจนในตัวข้อมูลอยู่แล้ว และข้อมูลแบบไม่มีโครงสร้าง มีอัตราส่วนอยู่ที่ประมาณ 80% ของข้อมูลทั้งหมดที่มีอยู่ในโลกใบนี้ ซึ่งเป็นข้อมูลที่จัดการได้ยาก โดยต้องมีการจัดโครงสร้างของข้อมูลเสียก่อน



รูปที่ 3 Structured Data VS Unstructured Data [11]

<u>แหล่งข้อมูล (Data Source)</u>

การได้มาซึ่งข้อมูลที่จะนำมาวิเคราะห์ได้นั้น มีวิวัฒนาการมาจากเดิมที่ต้องจดบันทึกในกระดาษในการจดบันทึกข้อมูล การบันทึกข้อมูลลงในแบบฟอร์มที่สร้างขึ้น หรือการจัดเก็บข้อมูลจากเว็บไซต์ลงฐานข้อมูล แต่ในปัจจุบันแหล่งข้อมูล (Data Source) ได้เปลี่ยนมาเป็นการใช้เครื่องมืออุปกรณ์ต่างๆ ที่สามารถรับข้อมูลได้ เช่นเครื่องวัดอัตราการเต้นของหัวใจ เครื่องสแกนลายนิ้วมือ ข้อมูลที่รับส่งทางอีเมล์ โดยเครื่องมือต่างๆเหล่านี้มักมีหน่วยความจำของตัวเอง และสามารถเชื่อมต่อกับเครือข่ายอินเตอร์เน็ตได้ เครื่องมือชนิดนี้ถูกเรียกว่า Internet Of Think (IoT) ได้แก่ โทรศัพท์มือถือ ซึ่งสามารถแบ่งปันข้อมูลต่างๆ ได้ผ่านทาง Social เช่น ข้อมูลพฤติกรรมความต้องการต่างๆ หรือนาฬิกา Smart Watch ที่สามารถเก็บข้อมูลอัตราการเต้นของหัวใจขณะออกกำลังกาย หรือข้อมูลการนอนหลับ เป็นต้น

การจัดเก็บข้อมูล (Data Lake) [13]

เมื่อได้ข้อมูลมาแล้วสิ่งที่ควรคำนึงถึงคือจะนำข้อมูลเหล่านั้นมาจัดเก็บที่ไหนใน Big Data ใช้แนวทางในการเก็บข้อมูลแบบ Data Lake ซึ่งเป็นพื้นที่ในการจัดเก็บข้อมูลที่มีโครงสร้าง และไม่มีโครงสร้างทุกขนาด สามารถจัดเก็บข้อมูลตามที่ต้องการโดยไม่ต้องวางโครงสร้างที่แน่นอน สามารถรองรับการวิเคราะห์ข้อมูลแบบ Real-time ได้ เมื่อเปรียบเทียบ Data Lake กับการคลังเก็บข้อมูลแบบธรรมดา รายละเอียดปรากฏตามตารางที่ 1

ตารางที่ 1 ตารางการเปรียบเทียบ Data Lake กับการคลังเก็บข้อมูล [13]

คุณลักษณะ	คลังเก็บข้อมูล	Data Lake
ข้อมูล	ข้อมูลเชิงสัมพันธ์จากระบบธุรกรรม ฐานข้อมูลการปฏิบัติงาน และแอปพลิเคชันหน่วยธุรกิจ	ข้อมูลที่ไม่ใช่เชิงสัมพันธ์และเชิงสัมพันธ์จากอุปกรณ์ IoT เว็บไซต์ แอปมือถือ โซเชียลมีเดีย และแอปพลิเคชันองค์กร
สคีมา	ได้รับการออกแบบก่อนการนำ DW ไปใช้ (สคีมาที่กำหนดไว้ล่วงหน้า)	เขียนเมื่อมีการวิเคราะห์ (สคีมาที่กำหนดเมื่อใช้)
ราคา/ประสิทธิภาพ	ผลการสืบค้นที่รวดเร็วที่สุดโดยใช้พื้นที่จัดเก็บที่มีต้นทุนสูงกว่า	ผลการสืบค้นที่รวดเร็วยิ่งขึ้นโดยใช้พื้นที่จัดเก็บที่มีต้นทุนต่ำ
คุณภาพข้อมูล	ข้อมูลที่จัดเตรียมอย่างดีซึ่งใช้เป็นความจริงพื้นฐาน	ข้อมูลใดๆ ที่อาจได้รับหรือไม่ได้รับการจัดเตรียม (เช่น ข้อมูลดิบ)
ผู้ใช้	นักวิเคราะห์ทางธุรกิจ	นักวิทยาศาสตร์ข้อมูล, นักพัฒนาข้อมูล และนักวิเคราะห์ทางธุรกิจ (โดยใช้ข้อมูลที่จัดเตรียม)
การวิเคราะห์	การรายงานแบบกลุ่ม, BI และการแสดงภาพ	Machine Learning การวิเคราะห์เชิงคาดการณ์ การค้นพบข้อมูล และจัดทำโปรไฟล์

แหล่งจัดเก็บข้อมูล (Data Storage)

เนื่องจาก Big Data เป็นข้อมูลที่มีขนาดใหญ่ และขยายเพิ่มมากขึ้นเรื่อยๆตลอดเวลา จึงไม่สามารถเก็บในอุปกรณ์ที่มีความจุที่จำกัดได้ ปัจจุบัน Cloud Storage เข้ามามีบทบาทเกี่ยวกับระบบสารสนเทศ (Information Technology) ด้วยการออกแบบระบบ Server less หรือไม่จำเป็นต้องมี Physical Data Center อีกต่อไป โดย cloud สามารถแบ่งออกเป็น 3 ประเภท ดังนี้ [14]

1. Public Cloud คือ ระบบบริการผู้ให้บริการออกแบบมาเพื่อให้คนทั่วไปสามารถเข้าถึงและใช้งานร่วมกันได้ เช่น Cloud ของ Google หรือ Cloud ของ AWS เป็นต้น

2. Private Cloud คือ ระบบที่องค์กรแต่ละองค์กรจัดทำขึ้นเพื่อรองรับการทำงานขององค์กรใด องค์กรหนึ่งหรือเฉพาะกลุ่มที่ได้รับอนุญาตเท่านั้น

3. Hybrid Cloud คือ ระบบที่มีผสมผสานการใช้งานแบบ Public Cloud และ Private Cloud เข้าด้วยกัน เพื่อความยืดหยุ่นในการใช้งาน

Cloud แบ่งตามการให้บริการออกเป็น 3 ประเภท [15] คือ

1. Software as a Service (Saas) คือ การให้บริการทางด้าน Software หรือ Application ผ่านทางระบบอินเตอร์เน็ต อาจจะไม่คิดค่าใช้บริการ หรือคิดค่าบริการตามลักษณะการใช้บริการ เช่น Microsoft Office 365 Google Drive และ Google Calendar เป็นต้น

2. Platform as a Service (Paas) คือ การให้บริการด้าน Platform สำหรับผู้ใช้บริการด้านการพัฒนาโปรแกรม เพื่อใช้ Software หรือ Application ที่ผู้ให้บริการจัดเตรียมไว้ให้ในการพัฒนา Application เช่น Database Server Web Application หรือ platform ที่รองรับการพัฒนา Application บน Mobile เป็นต้น

3. Infrastructure as a Service (IaaS) คือ การให้บริการทางด้านโครงสร้างพื้นฐานทางด้านไอที (Infrastructure) และระบบการจัดเก็บข้อมูลขององค์กร (Storage) เพื่อรองรับการใช้งาน Software ขององค์กร

นอกจากการเก็บข้อมูลบน Cloud แล้ว Big Data ยังสามารถใช้เทคโนโลยี Hadoop ซึ่งเป็น Software แบบ Open-Source ที่สร้างขึ้นมาเพื่อเป็น Frame work ในการจัดการ Big Data โดยเฉพาะ จุดเด่นของ Hadoop คือ สามารถใช้กับเครื่องคอมพิวเตอร์ (Server) ที่ไม่ต้องแรงมากได้อีกด้วย การขยาย Scale ในอนาคตก็สามารถเพิ่มเครื่องคอมพิวเตอร์เข้าไปได้ง่าย และยังมีระบบสำรองข้อมูลอัตโนมัติ โดยระบบการทำงานของ Hadoop จะทำงาน โดยเมื่อรับข้อมูลจากภายนอกเข้ามาจะแบ่งส่วนของข้อมูลออกเป็น 3 ข้อมูลที่เหมือนกัน (3 Node Replicate) บน Node หรือ Hardware หลายตัวพร้อมกัน ทำให้การประมวลผลของ Hadoop สามารถทำได้อย่างรวดเร็ว โดยเฉพาะข้อมูลที่มีขนาดใหญ่ นอกจากนี้ Hadoop ยังมีการประมวลผลแบบ MapReduce ซึ่งเป็น Framework ในการเขียนโปรแกรมแบบหนึ่งที่ช่วยในการประมวลผลที่มี Data Set จำนวนมาก หลักการทำงานจะเป็นการทำงานแบบคู่ขนาน (Parallel) ซึ่งจะอาศัยเครื่องคอมพิวเตอร์หลายๆ เครื่องช่วยกันประมวลผล โดย Map จะเป็นการจับคู่ Key/ Value ที่ต้องการ และส่งข้อมูลให้ Reduce ประมวลผลเพื่อให้ได้ผลลัพธ์ตามที่ต้องการ ยกตัวอย่างการนับจำนวนสัตว์ที่อยู่ในข้อมูลขนาดใหญ่ โดยกำหนดให้ Key คือ ชนิดของสัตว์เป็น และกำหนด Value คือ 1 โดยส่งค่า Key และ Value ไปยัง File ต่างๆ เช่น หมู สามารถนับเป็น 1 ตัวได้ 2 ครั้งหมายถึง 2 File แล้วจึงนำมาประมวลผลรวมกันคือ หมู นับได้ 2 ตัวเป็นต้น รายละเอียดปรากฏในรูปภาพที่ 4

รูปที่ 4 MapReduce Process [21]

บทที่ 3 บุคคลากรและทักษะที่ใช้ในการทำ Big Data

การเริ่มทำ Big Data Project แน่นอนว่าไม่สามารถทำให้สำเร็จได้ด้วยบุคลากรเพียงคนเดียว ดังนั้นในแต่ละหน้าที่ แต่ละส่วนงานต้องใช้บุคลากรที่มีความเชี่ยวชาญเฉพาะด้าน ซึ่งในบางหน่วยงาน ยังขาดบุคคลากรที่มีทักษะในด้านข้อมูลจำนวนมาก องค์กรจึงต้องวางแผนพัฒนาบุคลากรและส่งเสริมให้มีการเรียนรู้ อบรมให้บุคลากรเกิดความรู้ความเข้าใจเฉพาะด้าน โดยสามารถแบ่งบุคลากรผู้เชี่ยวชาญออกเป็น 3 หน้าที่ ที่ชัดเจน ได้ดังนี้

1. Data Engineer (วิศวกรข้อมูล) คือ บุคคลที่ทำหน้าที่ติดตั้งวางระบบ Server ระบบความปลอดภัย Security และดูแลจัดการข้อมูลทั้งหมดของระบบ ตั้งแต่ระบุชนิดของข้อมูล วางโครงสร้างของข้อมูล รวมถึงการทำข้อมูลให้มีความถูกต้องและพร้อมใช้งาน เพื่อส่งข้อมูลต่อไปให้ Data Scientist นำไปต่อยอดในการทำงานได้ หน้าที่นี้ค่อนข้างมีความท้าทาย เพราะต้องอาศัยทักษะหลายด้าน แต่ตั้งการเขียน Data Flow Diagram (DFD) เพื่อดูการเข้า-ออกของข้อมูล การจัดการระบบ Cloud หรือการจัดการระบบ Data Storage ในแบบอื่นๆด้วย เช่น Hadoop เป็นต้น การออกแบบ Database การทำ ETL (กระบวนการตรวจสอบคุณภาพของข้อมูลในระบบ Data Warehouse) การทำความสะอาดข้อมูล (Cleansing Data) การดึงข้อมูลออกจากระบบเพื่อสำรองข้อมูล (Backup Data) ทั้งแบบ Batch และ Real-Time รวมถึงต้องเขียน API ในการดึงข้อมูลกับโปรแกรมอื่น รวมถึงการ Transform Model ขึ้นไป Run บนระบบ Production ทักษะของ Data Engineer คือต้องมีความรู้ความเข้าใจในหลากหลาย Platform สามารถเปรียบเทียบข้อดี ข้อเสียของการใช้งานแต่ละ Platform ได้อย่างละเอียด เพื่อประโยชน์ในการตัดสินใจในด้านงบประมาณต่อไปในอนาคต

ทักษะของ Data Engineer

SQL เป็นภาษาที่ใช้เพื่อจัดการข้อมูลในระบบ Relational Database Management System (RDBMS) ซึ่งมีลักษณะการเก็บข้อมูลแบบมีโครงสร้างหรือฐานข้อมูลเชิงสัมพันธ์ (Relational Database)

NoSQL (Not Only SQL) สำหรับใช้จัดการข้อมูลแบบ Non-Relational Database ที่เป็นข้อมูลส่วนใหญ่ในระบบข้อมูลขนาดใหญ่ (Big Data) โดยหลักการของ NoSQL สามารถจัดการข้อมูลได้หลายรูปแบบ อาทิ Key-Value Graph หรือ Document เป็นต้น

Python เป็นภาษาถูกพัฒนามาให้ไม่ยึดติดกับ Platform และเป็นภาษาที่ถูกใช้มากใน Data Science เนื่องจากมี Packet ที่สนับสนุนหลากหลาย อาทิ Pandas (สำหรับ Data Wrangling) Scikit-learn (สำหรับทำ Machine Learning Model) Tensorflow (สำหรับทำ Deep Learning) สามารถใช้ Pyspark ที่ทำให้สามารถเชื่อมต่อกับ Spark เพื่อบริหารจัดการข้อมูลบน Hadoop cluster ได้

โดยตรง นอกจากนี้ Python สามารถทำ ETL ได้โดยการเขียน Script ขึ้นมาเพื่อเรียกข้อมูลจากแหล่งต่างๆ ไปใส่ในโมเดลได้อีกด้วย

Hadoop หรือ (HDFS) Hadoop File System ซึ่งเป็น Platform หลักที่ออกแบบมาเพื่อใช้งานกับระบบข้อมูลขนาดใหญ่ (Big Data) สามารถจัดการสร้าง Hadoop Cluster และเข้าใจหลักการทำงานแบบ Node ที่สามารถแบ่งไฟล์ออกเป็นส่วน ให้สามารถทำงานพร้อมกันได้ และไม่เหมาะที่จะนำมาใช้งานกับข้อมูลแบบ Real-Time เหมาะกับการเก็บข้อมูลขนาดใหญ่ (Big Data) เพื่อนำมาวิเคราะห์มากกว่า

Cloud computing สามารถบริหารจัดการและ การเลือกใช้บริการที่หลากหลายของ Service รวมถึงการประเมินงบประมาณที่ใช้งานในแต่ละ Service เพราะ Cloud Computing มีลักษณะการใช้งานแบบจ่ายเท่าที่ใช้งาน Pay as you go รวมทั้งหากต้องติดตั้ง Hadoop บน Cloud จะต้องสามารถวางระบบเครือข่าย หรือความปลอดภัยของแหล่งจัดเก็บข้อมูลอีกด้วย

ระบบหลักที่ใช้ในการบริการขององค์กร ตั้งแต่ระบบที่ใช้บริการประชาชนแต่ละด้านขององค์กร การเข้าใจความสำคัญของข้อมูลที่นำมาเก็บใน Data Storage รวมถึงสามารถทราบถึงแหล่งข้อมูล (Data Source) หรือการวางแผน เพื่อให้ได้มาซึ่งข้อมูลที่ชัดเจน นอกเหนือจากข้อมูลที่มีอยู่แล้วในระบบบริการ เช่น ข้อมูลการตอบแบบสอบถามความพึงพอใจ ข้อมูลความต้องการทางด้านอาชีพโดยอาศัยข้อมูลผ่านทางโซเชียลมีเดีย (Social Media)

2. Data Scientist หรือนักวิทยาศาสตร์ข้อมูล คือ บุคคลที่สามารถนำข้อมูลมาหาความสัมพันธ์ จากการวิเคราะห์เชิงลึก มีความรู้ด้านการวิเคราะห์ข้อมูลและสร้างมูลค่าให้กับข้อมูล โดยการทำงานของ Data Scientist นั้นเริ่มจากการตั้งโจทย์การเลือกข้อมูลที่ตอบโจทย์ รวมถึงการนำข้อมูลที่ได้มาสร้างโมเดล และพัฒนาโมเดลให้แม่นยำมากที่สุด เพื่อนำโมเดลนั้นไปประยุกต์ใช้งานได้จริงและเกิดประโยชน์สูงสุด

ทักษะของ Data Scientist

พื้นฐานทางคณิตศาสตร์ แคลคูลัส (Calculus) สถิติ ความน่าจะเป็น พีชคณิตเชิงเส้น (Linear algebra) ตรรกศาสตร์ (Logic) การเพิ่มประสิทธิภาพ (Optimization) การออกแบบการทดลอง (Design of Experiment)

พื้นฐานทางการเขียนโปรแกรม คือ ความรู้ ความเข้าใจทางด้านภาษาที่ใช้ในการวิเคราะห์ข้อมูล ได้แก่ ภาษา R และ Python

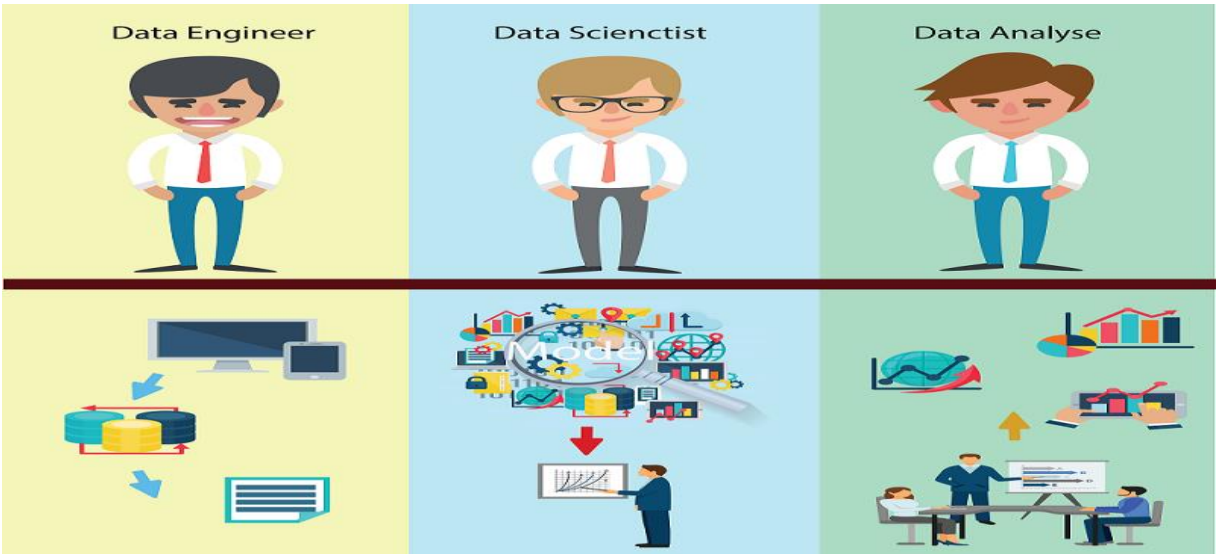
พื้นฐานความรู้เฉพาะทาง (Domain Knowledge) คือ ความรู้ในสิ่งที่ต้องคิดหรือการตั้งหลักการเฉพาะเรื่อง ในการแก้ปัญหา หรือการตัดสินใจ หรือความรู้ทางด้านธุรกิจ เป็นต้น

ปัจจุบันสายงาน Data Scientist มีน้อยมากโดยเฉพาะในประเทศไทย เนื่องจากสายงานนี้ต้องใช้ความรู้เชิงลึกในด้านการทำ Big Data Project เช่น การทำ Image Processing การระบุตัวตนโดยภาพ การทำ Text Mining การใช้วิเคราะห์ภาษาเพื่อแปล หรือสะกดคำที่สามารถไปต่อยอดเป็น ChatBot การประดิษฐ์ AI เป็นต้น

3. Data Analyst หรือนักวิเคราะห์ข้อมูล คือ บุคคลที่ใช้ข้อมูลในการวิเคราะห์แนวโน้ม หรือแก้ปัญหาจากสิ่งที่ต่างไปจากแนวโน้มเดิม โดยใช้หลักสถิติเพื่อวิเคราะห์ทั่วไปและนำมาแสดงผล และแก้ไขปัญหาภายในองค์กร หรือกำหนดยุทธศาสตร์ภายในองค์กร โดยการทำ Data Analyst ไม่ใช่การวิเคราะห์ในเชิงลึกเหมือน Data Scientist แต่ก็ต้องอาศัยประสบการณ์จากความรู้เฉพาะด้านนั้นๆ เช่น ความรู้ทางด้านการตลาด (Marketing Analyst) หรือความรู้ทางด้านการจัดหางาน (Employment knowledge) เป็นต้น และมุมมองที่เฉียบขาดเพื่อให้การแก้ไขปัญหา รวมทั้งสามารถกำหนดขอบเขตของปัญหาได้อย่างชัดเจน เพื่อช่วยในการกำหนดแนวทางการตัดสินใจของผู้บริหารระดับสูงขององค์กร

<u>ทักษะของ Data Analyst</u>

ส่วนใหญ่ทักษะของ Data Analyst จะเป็นการใช้เครื่องมือต่างๆ เพื่อนำมาวิเคราะห์โดยเฉพาะ เช่น Microsoft Excel SAS หรือเครื่องมือในการทำ BI ต่างๆ เช่น Power BI Tableau Rapid Miner เป็นต้น และอาจจะต้องมีความสามารถทางด้านการเขียนโปรแกรมที่นำมาใช้ในการวิเคราะห์ข้อมูล การใช้ภาษา R เพื่อนำมาช่วยในการวิเคราะห์ข้อมูลเชิงสถิติ พื้นฐานการจัดการข้อมูลบน Database เข้าใจหลักการ Machine Leaning เบื้องต้น การทำ Data Visualization หรือการนำเสนอข้อมูลที่ผ่านการวิเคราะห์แล้ว ในรูปแบบต่างๆ กัน เช่น ตาราง กราฟ และการจัดทำสื่อต่างๆ เพื่อให้สามารถเข้าใจความสำคัญของข้อมูลนั้นได้โดยง่าย สวยงาม และสร้างสรรค์



รูปที่ 5 บุคลลากรในการทำ Big Data Project

บทที่ 4 เริ่มทำ Big Data Project

เมื่อพูดถึงการนำ Big Data เข้ามาใช้ในองค์กร การทำให้ Big Data Project อาจไม่ใช่เรื่องง่ายนัก โดยทั่วไปการทำให้ Project สำเร็จ คือ การปิด Project ให้ได้ตามที่ตกลง ต้องมีการส่งมอบระบบ และใช้งานได้จริงได้อย่างเสถียรภายในเวลาที่กำหนด ในทางปฏิบัติจริง การส่งงานตามเวลาที่กำหนดอาจจะไม่ใช่การปิด Project แต่ต้องดูด้วยว่าสุดท้ายแล้ว องค์กรได้รับระบบที่ตอบโจทย์การใช้งานของผู้ใช้งาน หรือ User นั้นจริงๆ โดยที่ User มีความเข้าใจในระบบ และสามารถปรับจูนโมเดล หรือแก้ไขงานที่เชื่อมโยง Workflow ได้อย่างยืดหยุ่น การปฏิบัติงานได้จริงมากกว่าแค่ภาคทฤษฎี และมีการพัฒนาโมเดลอย่างต่อเนื่อง

กรณีศึกษาตัวอย่างได้แก่ Netflix บริการสตรีมมิ่งทางอินเตอร์เน็ตเพื่อรับชมรายการทีวี ภาพยนตร์ และสารคดี สัญชาติอเมริกา ที่ก่อตั้งมาเมื่อปี 1997 ผลิตภัณฑ์ของ Netflix คือ วีดีโอ นั้นทำให้ Netflix มีข้อมูลเกี่ยวกับวีดีโอมากมายรวมเข้ากับข้อมูลของลูกค้า และข้อมูลการเข้ารับชมวีดีโอของลูกค้า จึงต้องมีการพัฒนา Platform ให้มีประสิทธิภาพที่ดีมากขึ้น ในแต่ละปี Netflix จะจัดงานให้ Data Scientist พัฒนาได้แข่งขันกันทำ Algorithm เพื่อแนะนำวีดีโอ โดยจะต้องมีความแม่นยำมากขึ้น แต่โมเดลที่ดีที่สุดเมื่อทดลองกับระบบแล้วพบว่า ไม่เหมาะกับสภาพแวดล้อมจริงของระบบ เหตุผลที่ไม่สามารถนำไปใช้งานได้จริง อาจเกิดจากความผิดพลาดในการสร้างโมเดล ไม่ว่าจะเป็นการเก็บข้อมูลไม่ครบ ข้อมูลที่นำมาทดสอบน้อยเกินไป ทดสอบโมเดลไม่ละเอียด ความต้องการของผู้เกี่ยวข้องที่ไม่ชัดเจน รวมถึงสภาพแวดล้อมที่ไม่เอื้ออำนวยต่อการใช้งานโมเดล

1. ทำความเข้าใจปัญหาและความต้องการขององค์กร

การวางเป้าหมายในการทำงานถือเป็นขั้นตอนแรก และเป็นขั้นตอนที่สำคัญที่สุดในการกำหนดแนวทางของการทำงาน โดยเริ่มจากปัญหาที่พบให้ชัดเจน หรือข้อสงสัยต่างๆ ที่ทำให้ต้องหาทางแก้ไข การทำความเข้าในส่วนงานต่างๆขององค์กรว่ามีส่วนใดที่สามารถเข้ามาทำให้เกิดการพัฒนาด้วยระบบดิจิทัล เพื่อสามารถต่อยอดการทำงานได้ อาจต้องมีการเพิ่มการสังเกตการณ์ (Observation) ไม่ว่าจะเป็นการเข้าร่วมทำงานกับสภาพแวดล้อมการทำงานจริง หรือแม้แต่การทำตัวเป็นผู้ใช้บริการ และการให้ผู้มีส่วนเกี่ยวข้องกับการใช้งานของข้อมูลทั้งหมด เข้ามาช่วยกำหนดความต้องการเพื่อแปลงความต้องการทั้งหมด ออกมาเป็นปัจจัยหรือตัวแปลหนึ่งในโมเดล เพื่อให้สามารถทำโมเดลที่สามารถใช้งานได้จริงใน Big Data Project

อย่างไรก็ตามในบางองค์กรก็ยังไม่สามารถตั้งเป้าหมายของ Big Data Project ได้ เพราะยังไม่มีความเข้าใจเรื่องการใช้งาน Big Data และไม่มีพื้นฐานทางด้าน Data Science หรือเทคโนโลยี ต่างๆ ในการบริหารข้อมูลอีกด้วย

2. ทำความเข้าใจข้อมูล

การทำความเข้าใจข้อมูลเป็นขั้นตอนที่ 2 ของการสร้างโมเดลเพื่อวิเคราะห์ข้อมูล ซึ่งหากทราบเป้าหมายในการใช้ข้อมูล ก็ไม่สามารถทราบได้ว่าต้องใช้ข้อมูลอะไรบ้าง

การทำความเข้าใจข้อมูล ไม่ใช่แค่การทำความสะอาดข้อมูล (Cleansing Data) แต่ต้องทำความเข้าใจว่าข้อมูลใดมีความสำคัญ ในบางกรณีการทำความเข้าใจอาจจะเริ่มจากการนำข้อมูลทั้งหมดที่มีมารวมกัน และหาความสัมพันธ์ระหว่างข้อมูล

ดังนั้น สิ่งที่สำคัญที่สุดในการทำความเข้าใจข้อมูล คือ คำนิยามของข้อมูล (Heading) แล้วตามมาด้วยรายละเอียดภายใน หากเป็นข้อมูลประเภทที่มีโครงสร้าง (Structured Data) สามารถเข้าใจง่าย แต่หากเป็นข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) จะต้องมีการแปลงโครงสร้างให้ชัดเจน สามารถเข้าได้เสียก่อน ขั้นตอนนี้อยู่ที่มุมมองและประสบการณ์ของ Data Scientist ในการคัดเลือกข้อมูลมาใช้งานที่ตอบโจทย์ในแต่ละโจทย์

3. การเตรียมข้อมูล

การเตรียมข้อมูลเพื่อทำการวิเคราะห์ถือว่าเป็นขั้นตอนที่ใช้เวลานานที่สุดของการทำ Big Data Project อาจจะใช้เวลาประมาณ 75%-80% ของการทำ Project ทั้งหมด โดยหลักการแล้ว Data Engineer และ Data Scientist เป็นคนที่มีบทบาทในส่วนนี้มากที่สุด เหตุผลที่ขั้นตอนนี้ใช้เวลานานที่สุด เพราะข้อมูลที่มีขนาดใหญ่มาจากการรวมกันจากหลายแหล่งข้อมูล และเจ้าของข้อมูลไม่เหมือนกัน ทำให้เกิดความยุ่งยากเกิดขึ้น การเตรียมข้อมูล แบ่งออกเป็นการทำความสะอาดข้อมูล (Cleansing Data) และการจัดรูปแบบข้อมูลให้พร้อมใช้งาน

Data Engineer มีหน้าที่ในการทำให้ข้อมูลอยู่ในรูปแบบที่สามารถนำไปต่อยอดต่อได้ หากเป็นข้อมูลที่ไม่มีโครงสร้าง เปลี่ยนแปลงให้เป็นข้อมูลที่มีโครงสร้าง เช่น การทำ Normalization หรือการทำ Meaning Extraction เพื่อจัดทำให้เป็นข้อมูลที่ได้มีโครงสร้างตามต้องการเพื่อนำไปสร้างเป็นโมเดลในขั้นตอนต่อไป นอกจากนี้ยังมีหน้าที่ทำความสะอาดข้อมูล การตรวจสอบข้อมูลที่มีความผิดปกติอีกด้วย

Data Scientist มีหน้าที่ในการกรองข้อมูลเพื่อนำข้อมูลไปเป็นต้นแบบในการทำโมเดล เช่น การหาข้อมูลที่อยู่นอกความเป็นไปได้ของข้อมูลที่เกิดขึ้น (Detect Outlier) การจัดการกับข้อมูลที่ขาดหายไป (Missing Value) การเลือกขนาดของข้อมูล การเลือกช่วงเวลาที่จะนำข้อมูลไปวิเคราะห์เป็นโมเดล เป็นต้น

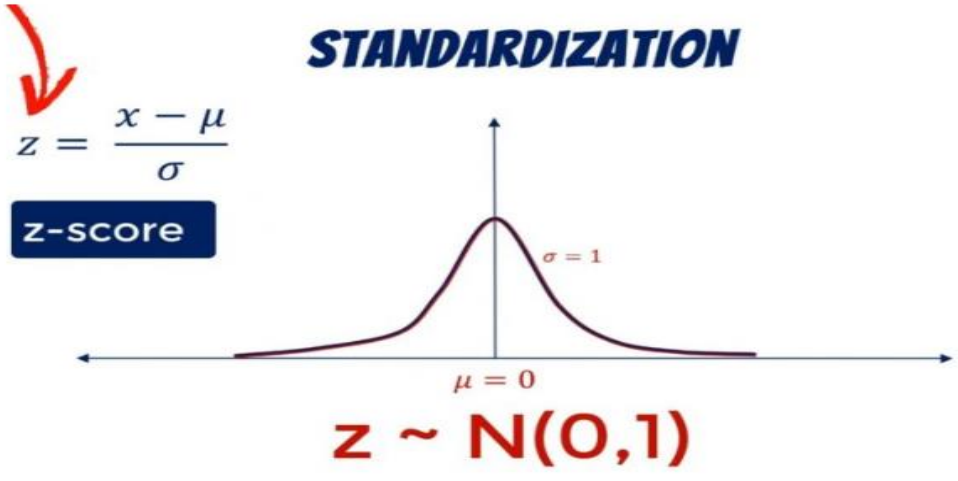
ขั้นตอนการทำความสะอาดข้อมูล (Cleansing Data)

เหตุผลที่ข้อมูลไม่สะอาดเกิดจากหลายสาเหตุ ตั้งแต่การพิมพ์ตก การพิมพ์ผิด เครื่องมือเกิดความผิดพลาด (Error) ข้อมูลที่อยู่นอกกลุ่มเช่น คนมีอายุ 500 ปี หรือ คนน้ำหนัก 1000 กิโลกรัม เป็นต้น โดยทางเทคนิคจะเรียกข้อมูลที่มีลักษณะแบบนี้ว่า “Outlier” นอกจากการทำความสะอาดข้อมูลแล้ว ยังต้องหาวิธีจัดการข้อมูลที่ตกหล่นหายไป หรือที่เราเรียกว่า “Missing Value” อีกด้วย ขั้นตอนการ Clean ข้อมูล 4 ขั้นตอน ได้แก่

1. Passing คือ การแจกแจงข้อมูลตามประเภทของข้อมูล หรือใช้หัวข้อของข้อมูล ความสำคัญของขั้นตอนนี้ไม่ใช่แค่เพียงการเลือกหัวของข้อมูล แต่เป็นการทำความเข้าใจคำจำกัดความของชุดข้อมูลนั้นๆ รวมถึงเข้าใจค่า และความหมายของข้อมูล

2. Correcting คือ การแก้ไขข้อมูลที่ผิดพลาด เช่น ในช่องข้อมูลต้องเป็นตัวเลขจำพวกอายุส่วนสูง แต่ใส่ตัวอักษรลงไป หรือในช่องของเพศมีการใส่ตัวเลขลงไป หรือข้อมูลมีค่าเกินความเป็นไปได้ เช่น ใส่ข้อมูลในช่องอายุ 500 ปี เป็นต้น วิธีการแก้ไขข้อมูลจะต้องใช้วิธีการทางสถิติเพื่อทำการแก้ไข เช่น การหาค่าเฉลี่ย การหาค่าเบี่ยงเบนมาตรฐาน (standard deviation) และการแก้ไขข้อมูลต้องมีการพิจารณาเลือกที่จะแก้ไขข้อมูลโดยการแทนที่ด้วยข้อมูลที่น่าจะเป็นไปได้ หรือตัดข้อมูลชุดที่ผิดนั้นออกทั้งแถว (Row)

3. Standardizing คือ การทำข้อมูลให้เป็นรูปแบบเดียวกัน ตัวอย่างจังหวัดในหลากหลายรูปแบบ เช่น กรุงเทพมหานคร กทม. หรือกรุงเทพฯ ซึ่งการประมวลผลของคอมพิวเตอร์ไม่สามารถทราบได้ว่าข้อมูลจังหวัดเดียวกัน ต้องทำการกำหนดค่ามาตรฐานในการใช้ข้อมูลร่วมกัน และหากเป็นข้อมูลตัวเลขที่มีช่วงค่า หรือหน่วยของตัวเลขแตกต่างกัน หรือมีความกว้างของข้อมูลที่ไม่เหมือนกัน สามารถใช้วิธีของ Standardization เพื่อเรียงข้อมูลให้อยู่ในรูปแบบระฆังคว่ำ เป็นการทำให้ข้อมูลอยู่ในช่วงค่าที่กำหนด คือระหว่าง 0 ถึง 1 ดังปรากฏในรูปที่ 5



รูปที่ 6 การทำ Standardization [35]

4. Duplicate Elimination คือ การลบข้อมูลที่มีความซ้ำซ้อนกันทิ้ง ซึ่งอาจจะต้องเขียนอัลกอลิทึม (Algorithm) เพื่อใช้ระบุชุดข้อมูลที่มีความซ้ำซ้อนกัน เช่น ข้อมูลทุกอย่างเหมือนกันทั้งหมด ตั้งแต่ ชื่อ นามสกุล และอยู่ภายในไฟล์หรือฐานข้อมูลเดียวกัน เป็นต้น

4. การสร้างโมเดล

การสร้างโมเดลถือเป็นขั้นตอนที่สำคัญที่สุด โดย Data Scientist ที่มีประสบการณ์จะสามารถพิจารณาได้ว่าแต่ละโจทย์จะเลือกใช้อัลกอริทึม (Algorithm) และสามารถเลือกตัวแปรที่เหมาะสมทำให้ใช้เวลาน้อยในการสร้างโมเดลหนึ่งโมเดล แต่การสร้างโมเดลเพียงโมเดลเดียวอาจไม่ตอบโจทย์ เพราะฉะนั้นอาจจะต้องมีการสร้างโมเดลหลายโมเดล เพื่อเปรียบเทียบหาโมเดลที่ดีที่สุด ดังนั้น Data Scientist ส่วนใหญ่จะทำการสร้างโมเดลเพิ่มหรือปรับปรุงไปเรื่อยๆ จนกว่าจะหมดเวลาหรือต้องส่งมอบงาน และการสร้างโมเดลนั้นจะต้องจะต้องเข้าใจรายละเอียดของผลลัพธ์ด้วย

<u>ตัวอย่างการสร้างโมเดล</u>

1. Classification คือ การทำนาย (Prediction) ว่าข้อมูลแต่ละข้อมูลของประชากรควรจัดอยู่ในกลุ่มใด โดยแต่ละกลุ่มมีการกำหนดชื่อไว้แล้วอยู่แล้ว เช่น เสื้อที่มีขนาด ลาย และสี สามารถกำหนดได้ว่าเป็นเสื้อชนิดใด

2. Regression คือ การประมาณ (Value Estimation) ดูว่าข้อมูลแต่ละตัวควรมีค่าเชิงตัวเลขเป็นเท่าไร เช่น การพยากรณ์การเพิ่มจำนวนอัตราการมีงานทำของคนไทยว่ามีอัตราเพิ่มขึ้นกี่เปอร์เซ็นต์ เป็นต้น

3. Similarity Matching คือ การหาอัตลักษณ์ที่เหมือนกัน (Similar Identifying) ตามมิติต่างๆกันของข้อมูลในประชากรทั้งหมด ซึ่งวิธีการนี้ถูกนำมาทำโมเดลระบบการแนะนำสินค้าหรือบริการ (Recommendation System) เช่น พฤติกรรมการซื้อสินค้า A คนส่วนใหญ่จะซื้อสินค้า B ด้วย ดังนั้นเมื่อมีคนซื้อสินค้า A ระบบก็จะมีการแนะนำสินค้า B ด้วย เป็นต้น

รูปที่ 7 Recommendation Product [37]

นอกจากนี้ระบบ Recommendation System ยังสามารถนำมาประยุกต์ใช้กับงานจัดหางานภายในประเทศของกรมการจัดหางาน โดยการประเมินพฤติกรรมการสมัครงานของ นาย B ซึ่งมีประวัติการสมัครคล้ายกับนาย A ซึ่งทำให้เราสามารถเดาได้ว่างานที่นาย B ต้องการสมัครต่อไปจะเป็นงานลักษณะใด โดยสามารถสร้างระบบแนะนำงานที่เหมาะสมและตรงกับความต้องการของนาย B จะทำให้ระบบการจัดหางานภายในประเทศไทยมีประสิทธิภาพมากยิ่งขึ้น เป็นต้น

4. Clustering คือ การจัดกลุ่มหรือการกระจุกตัวของข้อมูล ซึ่งมีความแตกต่างจาก Classification ตรงที่ Clustering ไม่มีการกำหนดจำนวนกลุ่มไว้ล่วงหน้า จำนวนกลุ่มที่ได้มาจากการคำนวณผ่านอัลกอริทึม (Algorithm) ที่เลือกมาทั้งหมด เช่น การจำนวนกลุ่มผู้สมัครงานจากจำนวนทั้งหมด เป็นต้น

5. Link Prediction คือ การทำนายความเชื่อมโยง ระหว่างข้อมูลแต่ละข้อมูลว่ามีความสัมพันธ์กันหรือไม่ และมีความสัมพันธ์กันในระดับใด เช่น การเชื่อมโยงกันของเพื่อนในระบบ Social Media หากนาย A เป็นเพื่อนกับนาย B แล้วนาย B เป็นเพื่อนกับนางสาว C มีความเป็นไปได้ว่านาย A อาจจะรู้จักกับนางสาว C เป็นต้น

6. Profiling (Anatomy Detection) คือ การวิเคราะห์คุณลักษณะ (Characteristic) ที่เกี่ยวข้องกับพฤติกรรม (Behaviour) ในการทำกิจกรรมบางสิ่งบางอย่าง อาทิ ระบบตรวจจับบัตรเครดิตปลอม (Fraud Detection) โดย หากเกิดการใช้จ่ายบัตรเครดิตในสถานที่ห่างไกลจากพื้นที่ที่มีการใช้ประจำ หรือการซื้อสินค้ามีราคาสูงผิดปกติจากธรรมดา เจ้าหน้าที่ก็จะมีการโทรเข้าไปสอบถามเจ้าของบัตรเครดิตว่ามีการซื้อสินค้าจริงหรือไม่

รูปที่ 8 Profiling (Anatomy Detection)

5. การประเมินผลโมเดล

หลังจากการที่ได้โมเดลแล้ว ต้องทำการประเมินว่าโมเดลนั้นมีความแม่นยำมากหรือน้อยเพียงใด โดยการประเมินผลอาจจะเป็นทั้งในกรณีที่สามารถวัดออกมาเป็นค่าได้ หรือเรียกว่า การวัดเชิงปริมาณ เช่นการทำนายยอดขาย เป็นต้น และในกรณีที่ไม่สามารถวัดค่าได้ หรือกรณีที่ไม่มีข้อเปรียบเทียบระหว่างค่าที่โมเดลทำนายได้กับค่าจริง อาจจะต้องทำการทดสอบแบบจำลองสถานการณ์จริง เช่น การนำโมเดลไปประมวลผลข้อมูลจริงที่มีอยู่ เพื่อทำการวิเคราะห์ผลและเปรียบเทียบความถูกต้องออกมาเป็นอัตราร้อยละ หรือเรียกว่า การทดลองในระบบเสมือน (Simulation) เพื่อให้แน่ใจว่าโมเดลสามารถนำไปใช้งานได้จริง

6. การนำโมเดลไปใช้งานจริง

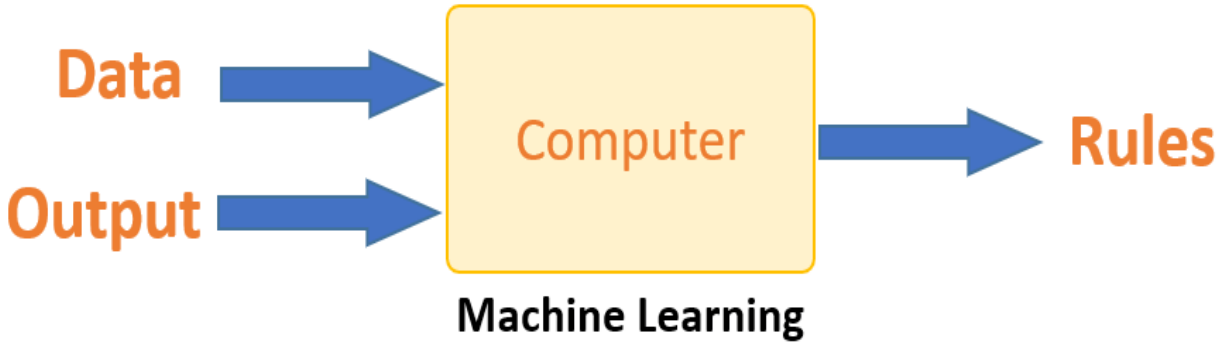
หลังจากที่ได้โมเดลที่มีคุณภาพและมีความแม่นยำตามที่ต้องการ ก็สามารถนำโมเดลนั้นมาประยุกต์ใช้กับงานจริง โดยอาจจะต้องมีการปรับปรุงเพื่อให้เหมาะสมกับสภาวะจริง เพื่อนำโมเดลมาสร้างเป็นผลิตภัณฑ์ (Product) ต่างๆ เช่น การสร้างระบบ (Application) ทำนายความมีโอกาสในการได้งานทำ หรืออาจร่วมกับระบบอื่น เช่น ระบบช่วยการตัดสินใจ (Decision Support System) เป็นต้น เพื่อให้การใช้โมเดลมีประโยชน์อย่างสูงสุดและมีความยั่งยืนอาจต้องมีปรับปรุงโมเดลใหม่เมื่อผลลัพธ์ไม่เป็นไปตามที่คาดหวัง หรือมีการปรับปรุงโมเดลเป็นระยะๆ อยู่เสมอเพราะว่าข้อมูลที่อยู่ภายในระบบข้อมูลขนาดใหญ่ (Big Data) และเทคโนโลยีต่างๆ มีการเปลี่ยนแปลงอยู่ตลอดเวลา

บทที่ 5 Machine Learning เบื้องต้น

จากที่ได้เรียนรู้เบื้องต้นเกี่ยวกับ Big Data Project เบื้องต้นแล้ว จะมีการกล่าวถึง Machine Learning บ้างแล้วในส่วนของการสร้างโมเดล และตัวอย่างจากการสร้างโมเดล ซึ่งในบทนี้จะมาขยายความเกี่ยวกับการเรียนรู้ของ Machine จากข้อมูลที่เข้ามาในลักษณะแบบต่างๆ และความเป็นมาของ Machine Learning

Machine learning เริ่มต้นมาตั้งแต่ปี ค.ศ.1950 เมื่อนักวิทยาศาสตร์คอมพิวเตอร์ คิดหาวิธีสอนคอมพิวเตอร์ให้เล่นหมากฮอส จากนั้นเมื่อวิวัฒนาการทางเทคโนโลยี ระบบการคำนวณค่าต่างๆ ของคอมพิวเตอร์เพิ่มขึ้น ทำให้คอมพิวเตอร์สามารถเข้าใจ และจดจำรูปแบบของค่าต่างๆ ที่ซับซ้อนได้แล้วจึงประยุกต์ไปสู่การคาดการณ์สถานการณ์รวมไปถึงการแก้ปัญหาด้วยตัวเอง

Machine learning คือ ระบบที่สามารถเรียนรู้จากตัวอย่างด้วยตนเอง โดยปราศจากการป้อนคำสั่งของโปรแกรมเมอร์ ความก้าวหน้าในครั้งนี้มาพร้อมกับความคิดที่ว่าเครื่องคอมพิวเตอร์สามารถเรียนรู้เพียงแค่จากข้อมูลอย่างเดียวเพื่อที่จะผลิตผลลัพธ์ที่แม่นยำออกมาได้



รูปที่ 9 กระบวนการทำงานของ Machine Learning [40]

จากรูปที่ 9 เป็นการอธิบายการทำงานของ Machine learning โดยการเรียนรู้ของ Machine คล้ายกับการเรียนรู้ของมนุษย์ โดยเรียนรู้จากประสบการณ์ ยิ่งรู้มากยิ่งง่ายต่อการพยากรณ์ว่าอะไรจะเกิดขึ้นต่อไป เมื่อประสบกับเหตุการณ์ที่ไม่เคยเจอมาก่อน มีความเป็นไปได้ที่พยากรณ์เหตุการณ์ได้น้อยลง เพื่อที่จะเพิ่มความแม่นยำในการพยากรณ์ โดยเมื่อเรามีข้อมูลตัวอย่างและผลลัพธ์ของข้อมูลให้ Machine ได้เรียนรู้ก็จะสามารถสร้างกฎหรือที่เรียกว่าโมเดลในการพยากรณ์ โดยการเรียนรู้ของ Machine learning แบ่งออกเป็น 2 ประเภทดังนี้

<u>การเรียนรู้แบบมีผู้สอน (Supervised learning)</u>

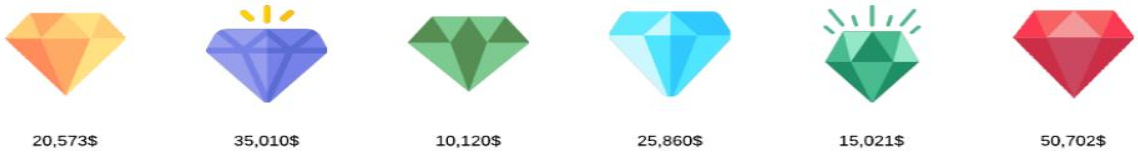
1. Supervised Learning (การเรียนรู้โดยมีผู้สอน) คือ การทำให้คอมพิวเตอร์หาคำตอบของปัญหาได้ด้วยตนเอง หลังจากเรียนรู้จากชุดข้อมูลตัวอย่างที่ประกอบไปด้วยข้อมูล และผลลัพธ์ของข้อมูล คล้ายกับการสอนเด็กด้วยรูปภาพโดยที่รูปภาพของสัตว์จะมีข้อมูลบ่งบอกด้วยว่าเป็นสัตว์

ชนิดใด และเมื่อนำภาพสัตว์ที่ไม่เคยเห็นก็สามารถแยกแยะได้ว่าเป็นสัตว์ชนิดใด เรียกกระบวนการสอนดังกล่าวว่า Classification โดยรายละเอียดปรากฏตามรูปที่ 10 – รูปที่ 11

รูปที่ 9 ผลลัพธ์จากการสอนแบบ Classification แบบไม่ซับซ้อน [41]

รูปที่ 10 ผลลัพธ์จากการสอบแบบ Cclassification แบบซับซ้อน [41]

การสอนเรื่องราคาเพชร โดยเมื่อหยิบเพชร ขนาด 2 กะรัต สีเหลือง ระดับความสะอาด VS2 แล้วบอกเด็กว่าราคา 20,000$ และหยิบอีกเม็ดขนาด 3 กะรัต สีแดง ระดับความสะอาด VS1 แล้วบอกเด็กว่าราคา 50,000$ ทำซ้ำไปเรื่อยๆ จนเด็กเกิดโมเดล (Model) หรือตรรกะ (Logic) ในการคาดเดาราคาของเพชรได้ แล้วจึงสุ่มหยิบเพชรเม็ดใหม่ขึ้นมา ก็อาจให้เด็กสามารถคาดเดาราคาได้เลย เรียกกระบวนการสอนดังกล่าวว่า Regression ดังปรากฏตามรูปที่ 11



รูปที่ 11 ผลลัพธ์จากการสอนแบบ Regression [41]

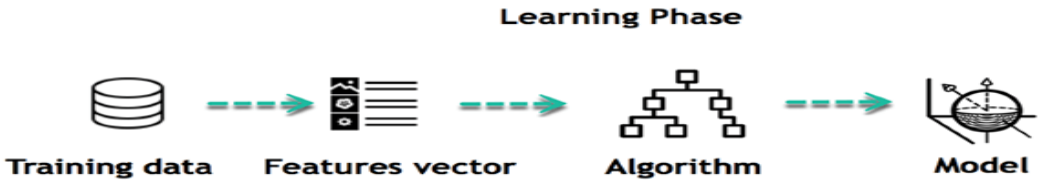
ปัจจุบันมีการนำโมเดลปัญญาประดิษฐ์ (AI Model) แบบมีผู้สอน (Supervise learning) ไปประยุกต์ใช้แก้ไขปัญหาหลากหลายรูปแบบมากๆ เช่น การนำไปพัฒนาเป็นระบบ Application ผู้ช่วยส่วนตัวในโทรศัพท์มือถือ เช่น Siri (Speech recognition) หรือพัฒนา Application ทางการแพทย์เพื่อตรวจสอบมะเร็งผิวหนังจากภาพถ่าย (Image Classification) หรือจะเป็นการโพสภาพลงสื่อ Social แล้วจะทำการ Tag อัตโนมัติว่าคนในภาพเป็นใคร (Face Detection) เป็นต้น

2. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) คือ การให้คอมพิวเตอร์เรียนรู้จากข้อมูลอย่างเดียวโดยปราศจากผลลัพธ์ของข้อมูล เช่นการสำรวจข้อมูลประชากรเพื่อหารูปแบบ (Pattern) ของข้อมูลนั้น สามารถใช้อัลกอริทึม (Algorithm) เพื่อหารูปแบบ (Pattern) ในการแบ่งกลุ่มข้อมูล (Clustering) เช่น การแบ่งกลุ่มข้อมูลแบบเคมีน (K-mean Clustering) เป็นการนำข้อมูลใส่ลงไปในกลุ่มข้อมูลทั้งหมด K กลุ่ม ซึ่งแต่ละข้อมูลมีลักษณะเหมือนกัน (ซึ่งถูกกำหนดขึ้นโดย

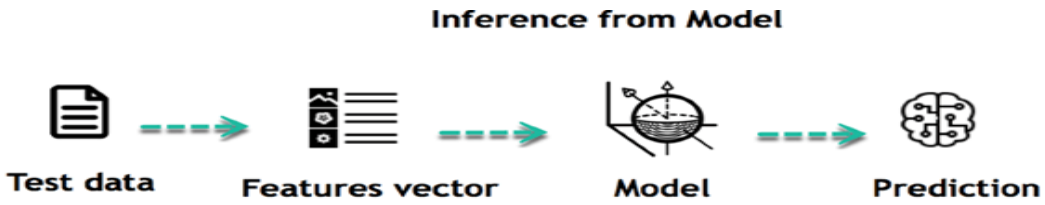
การประมวลผลของโมเดลไม่ได้เปลี่ยนแปลงตามที่มนุษย์สั่งการ) การแบ่งกลุ่มข้อมูลตามลำดับชั้น (Hierarchical Clustering) สามารถถูกใช้เพื่อแบ่งระดับชั้นของสมาชิกลูกค้าได้ ระบบให้คำแนะนำ (Recommendation System) เป็นต้น หรือการลดมิติของข้อมูลเพื่อการวิเคราะห์ข้อมูล (Dimension Reduction) โดยใช้ Algorithm PCA/T-SNE

วิธีการเรียนรู้ของ Machine Learning

ขั้นตอนของการเรียนรู้ของ Machine learning ขั้นแรกจะเป็นการแบ่งข้อมูลเพื่อนำมาวิเคราะห์ข้อมูลโดยแบ่งชุดข้อมูลออกเป็นชุดที่ให้ Machine ได้ฝึกเรียนรู้เรียกว่า Training Data Set จากนั้นทำการเรียนรู้ข้อมูลจนได้ออกมาเป็นโมเดล แล้วนำโมเดลที่ได้นำมาทำนายข้อมูลที่ไม่เคยเจอมาก่อนข้อมูลที่ถูกแบ่งจากชุดข้อมูลทั้งหมดเรียกว่า Testing Data Set แล้วจึงวัดค่าประสิทธิภาพจากการทำนายข้อมูลชุด Testing ดูว่ามีความแม่นยำมากน้อยเพียงใด



รูปที่ 12 ระยะการเรียนรู้ของ Machine learning [40]



รูปที่ 13 ระยะการนำโมเดลใช้ และวัดประสิทธิภาพโมเดล [40]

การประยุกต์ใช้ Machine learning

1. การทำงานอัตโนมัติ (Automation) การทำงานอัตโนมัติในด้านต่างๆ ยกตัวอย่าง เช่น หุ่นยนต์ (Robot) ที่ดำเนินการตามกระบวนการตามขั้นตอนในโรงงานการผลิต เป็นต้น

2. อุตสาหกรรมการเงิน (Finance Industry) ใช้ Machine learning เพื่อหารูปแบบ (pattern) ของข้อมูลเพื่อป้องกันการฉ้อโกงจะเกิดขึ้น

3. องค์การภาครัฐ (Government organization) ใช้ Machine learning ในการจัดการความปลอดภัยของระบบสาธารณูปโภคบริโภค ยกตัวอย่างเช่น มีการใช้ปัญญาประดิษฐ์ (Artificial Intelligence : AI) เพื่อช่วยให้การข้ามถนนอย่างปลอดภัย เป็นต้น

4. การตลาด (Marketing) มักใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data) เพื่อเพิ่มประสิทธิภาพการหาความสัมพันธ์ของลูกค้าและการโฆษณาด้านการตลาด

เอกสารอ้างอิง

[1] [Online] Big Data in Healthcare Available: http://bigdataexperience.org/big-data-healthcare/ [7 มิถุนายน 2562]

[2] [Online] what is Big Data Available: https://www.sas.com/th_th/insights/big-data/what-is-big-data.html [7 มิถุนายน 2562]

[3] [Online] “Big Data” สำคัญสำหรับธุรกิจยุค 4.0 อย่างไร Available: https://www.peerpower.co.th/blog/sme/big-data-for-sme/ [7 มิถุนายน 2562]

[4] [Online] big data Available: https://searchdatamanagement.techtarget.com/ definition/big-data [7 มิถุนายน 2562]

[5] [Online] Big Data ภาครัฐ สำนักวิชาการ สำนักงานเลขาธิการสภาผู้แทนราษฎร Available: https://library2.parliament.go.th/ejournal/content_af/2559/dec2559-4.pdf [7 มิถุนายน 2562]

[6] [Online] What is Big Data? Available: http://bigdata.black/featured/what-is-big-data/ [10 มิถุนายน 2562]

[7] [Online] Big Data and Hadoop Available: https://www.edureka.co/blog/what-is-big-data/ [10 มิถุนายน 2562]

[8] [Online] Big Data คืออะไร? Available: https://blog.goodfactory.co/big-data-คืออะไร-8ebf3a1a0050 [10 มิถุนายน 2562]

[9] [Online] What is Data? Available: https://www.mathsisfun.com/data/data.html [10 มิถุนายน 2562]

[10] [Online] คุณลักษณะ 6 ประการของฐานข้อมูลคุณภาพสูง Available: https://www.perceptra.tech/6-vs-of-big-data/ [10 มิถุนายน 2562]

[11] [Online] Structured Data vs. Unstructured Data: what are they and why care? Available: https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care [11 มิถุนายน 2562]

[12] [Online] ความหมายและชนิดของข้อมูล Available: https://sites.google.com/site/chokupgarage/khxmul-sarsnthes-laea-rabb-kherux-khay-khxmphiwtexr/khwam-hmay-laea-chnid-khxng-khxmul [11 มิถุนายน 2562]

[13] [Online] Data Lake คืออะไร Available: https://aws.amazon.com/th/big-data/datalakes-and-analytics/what-is-a-data-lake/ [11 มิถุนายน 2562]

[14] [Online] Cloud คืออะไร มีกี่ประเภท แบบไหนบ้าง? Available: http://blog.onestopware.com/cloud-คืออะไร-มีกี่ประเภท-แบบ/ [11 มิถุนายน 2562]

[15] [Online] Cloud computing คืออะไร? Cloud computing ดีอย่างไร? Available: https://www.it24hrs.com/2015/cloud-computing-and-cloud-definition/ [12 มิถุนายน 2562]

[16] [Online] Big data analytics สำคัญยังไงและช่วยอะไรเราได้บ้าง? Available: http://bigdataexperience.org/what-is-big-data-analytics/ [12 มิถุนายน 2562]
[17] [Online] Big Data คืออะไร ? + วิธีใช้ Hadoop/Spark บน Cloud Dataproc Available: http://www.siamhtml.com/getting-started-with-big-data-and-hadoop-spark-on-cloud-dataproc/ [12 มิถุนายน 2562]
[18] [Online] Hadoop Ecosystem สำหรับการพัฒนา Big Data Available: https://thanachart.org/2014/10/18/hadoop-ecosystem-สำหรับการพัฒนา-big-data/ [12 มิถุนายน 2562]
[19] [Online] หลักการทำงานของ MapReduce และตัวอย่างการใช้ Available: https://www.kanouivirach.com/2013/10/การทำงาน-mapreduce-และการใช้/ [13 มิถุนายน 2562]
[20] [Online] MapReduce Available: https://langisser.wordpress.com/2012/02/07/mapreduce/ [13 มิถุนายน 2562]
[21] [Online] MapReduce คืออะไร Available:https://www.howtoautomate.in.th/the-big-data-world-explanation-part-one/2018-05-29-21_58_49-mapreduce-คืออะไร_-bhuridech-sudsee-medium/ [13 มิถุนายน 2562]
[22] [Online] ทำไมธุรกิจจึงต้องใช้ Big Data Available: https://www.g-able.com/digital-review/why-using-big-data-in-business/ [13 มิถุนายน 2562]
[23] [Online] Data Scientist และเทคนิคการวิเคราะห์ Big Data Available: https://blog.goodfactory.co/data-scientist-และเทคนิคการวิเคราะห์-big-data-73dfbdcaa770 [13 มิถุนายน 2562]
[24] [Online] เริ่มเรียน Machine Learning 0–100 Available:https://medium.com/mmp-li/เริ่มเรียน-machine-learning-0-100-introduction-1c58e516bfcd [13 มิถุนายน 2562]
[25] [Online] แนะนำ Machine Learning เบื้องต้น Available: http://machinelearningth.actvee.com/2017/06/1-machine-learning.html [14 มิถุนายน 2562]
[26] [Online] Data Analytic: การเตรียมข้อมูลเพื่อการวิเคราะห์ Available: http://blog.dechathon.com/prepare-data-for-analytic/ [14 มิถุนายน 2562]
[27] [Online] การทำ Data Cleansing และ Database Marketing Available: https://www.affinity.co.th/data-cleansing/?lang=th [14 มิถุนายน 2562]
[28] [Online] ลักษณะของข้อมูลที่ใช้ในการวิเคราะห์ Available: http://www.dpu.ac.th/bigdata/variable_type.html [17 มิถุนายน 2562]
[29] [Online] การพัฒนาบุคลากรสำหรับงานทางด้าน Big Data Available: https://thanachart.org/2015/12/02/การพัฒนาบุคลากรสำหรับง/ [12 มิถุนายน 2562]

[30] [Online] วิทยาศาสตร์ข้อมูล Data Science คืออะไร? ต้องเริ่มต้นศึกษาอย่างไร? Available: https://www.appdisqus.com/2019/03/12/108-data-science-what-how-to-learning.html [17 มิถุนายน 2562]
[31] [Online] Data Science คืออะไร เรื่องที่ธุรกิจยุคดิจิทัลต้องรู้ Available: https://brandinside.asia/data-science-sexiest-job/ [17 มิถุนายน 2562]
[32] [Online] Data Engineer, Data Scientist และ Data Analyst ต่างกันอย่างไร Available: https://medium.com/@info_46914/data-engineer-data-scientist-และ-data-analyst-ต่างกันอย่างไร-4202b519183e [17 มิถุนายน 2562]
[33] [Online] 5 ปัจจัยส่งเสริม Big Data ให้ประสบความสำเร็จ Available: https://www.g-able.com/digital-review/5-factors-of-successful-big-data/ [17 มิถุนายน 2562]
[34] [Online] 4 ขั้นตอนการ Clean Data สำคัญไฉนWhy data quality is a KING? Available: https://www.coraline.co.th/single-post/why-data-quality-is-a-KING [17 มิถุนายน 2562]
[35] [Online] Explaining Standardization Step-By-Step Available: https://365datascience.com/standardization/ [19 มิถุนายน 2562]
[36] [Online] data science ก้าวที่ล้ำหน้ากว่า big data Available: https://www.g-able.com/digital-review/digital-transformation/big-data-analytics/data-science-ก้าวที่ล้ำหน้ากว่า-big-data-2/Reccomendation%20system [19 มิถุนายน 2562]
[37] [Online] Learning How Recommendation System Recommends Available: https://towardsdatascience.com/learning-how-recommendation-system-recommends-45ad8a941a5a [20 มิถุนายน 2562]
[38] [Online] Machine learning คืออะไร? Available: https://blog.finnomena.com/machine-learning-คืออะไร-fa8bf6663c07 [20 มิถุนายน 2562]
[39] [Online] 5 สิ่งที่ทุกคนควรรู้เกี่ยวกับ Machine Learning Available: https://www.quickserv.co.th/knowledge-base/technology/5%20สิ่งที่ทุกคนควรรู้เกี่ยวกับ%20Machine%20Learning/ [20 มิถุนายน 2562]
[40] [Online] Supervised learning คืออะไร? ทำงานยังไง? Available: https://medium.com/@every.phu/supervised-learning-คืออะไร-ทำงานยังไง-1c0e411a40a2 [20 มิถุนายน 2562]
[41] [Online] อะไรคือ การเรียนรู้ของเครื่อง (Machine Learning)? (ฉบับมือใหม่) Available: https://www.thaiprogrammer.org/2018/12/อะไรคือ-การเรียนรู้ของเ/ [20 มิถุนายน 2562]
[42] ดร. อสมา กุลวานิชไชยนันท์, 6/2018 Big Data Series 1 : Introduction to a Big Data Project ปฐมบทในการทำโปรเจคบิ๊กดาต้า [21 มิถุนายน 2562]

การจัดการความรู้

ความรู้เบื้องต้นเกี่ยวกับ Big Data และ Machine Learning

ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร
กรมการจัดหางาน
0 2354 0095