

บทที่ 4

ผลการดำเนินงาน

ผลการดำเนินงานของโครงการเรื่อง การประยุกต์ใช้เทคนิคการวิเคราะห์การจัดกลุ่ม เพื่อทำความเข้าใจความหลากหลายของ Data Scientist วิเคราะห์ข้อมูลเงินเดือน ในช่วงปี 2020-2024 เพื่อใช้สำหรับเผยแพร่ข้อมูลบนเว็บไซต์

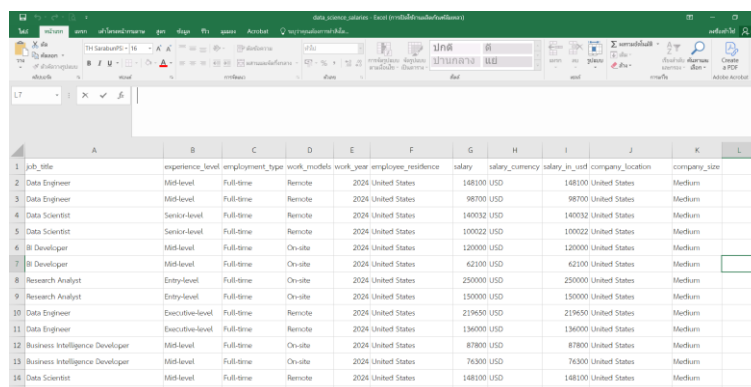
4.1 ผลการวิเคราะห์ข้อมูลด้วย CRISP-DM

4.1.1 ศึกษาข้อมูลเงินเดือนของ Data Scientists (Business Understanding) ขั้นตอนแรกคือการศึกษและทำความเข้าใจปัจจัยที่มีผลต่อเงินเดือนของ Data Scientist เพื่อกำหนดวัตถุประสงค์และขอบเขตของการวิเคราะห์ ข้อมูลจากการศึกษาพบว่าปัจจัยสำคัญ ได้แก่ 1) ประสบการณ์ (Experience) Data Scientist ที่มีประสบการณ์สูงมักได้รับเงินเดือนมากกว่าผู้ที่เพิ่งเริ่มงานหรือมีประสบการณ์น้อย 2) การศึกษาและคุณวุฒิ (Education & Qualification) การศึกษาที่สูงขึ้นหรือปริญญาในสาขาที่เกี่ยวข้อง เช่น ปริญญาตรีหรือปริญญาโท มีผลต่อระดับเงินเดือน 3) สถานที่ทำงาน (Work Location) เงินเดือนแตกต่างกันตามภูมิภาค เช่น เมืองใหญ่หรือพื้นที่ที่มีค่าครองชีพสูงมักมีเงินเดือนสูงกว่า 4) ประเภทของบริษัท (Company Type) บริษัทขนาดใหญ่หรือบริษัทที่มีเทคโนโลยีทันสมัยมักจ่ายเงินเดือนสูงกว่าบริษัทขนาดเล็กหรือเพิ่งเริ่มต้น 5) ทักษะเฉพาะ (Specialized Skills) ความเชี่ยวชาญในเทคโนโลยีหรือเครื่องมือเฉพาะ เช่น การวิเคราะห์ข้อมูลทางการเงินหรือการใช้เครื่องมือวิเคราะห์ขั้นสูง สามารถเพิ่มมูลค่าเงินเดือน

4.1.2 รวบรวมและทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนการจัดเก็บรวบรวมข้อมูล ตรวจสอบความถูกต้องของข้อมูลที่มาจาก เว็บไซต์ Kaggle.com โดยเลือกที่จะใช้ข้อมูลบางส่วนในการวิเคราะห์ให้สอดคล้องกับวัตถุประสงค์ที่กำหนดไว้ ชุดข้อมูล “data_science_salaries” การรวบรวมข้อมูล เงินเดือน, ประสบการณ์, ทักษะ และตำแหน่งงานการตรวจสอบความสมบูรณ์ของข้อมูล

ในไฟล์ข้อมูลจะมีจำนวนข้อมูล 6,599 รายการ ประกอบไปด้วย 11 แอตทริบิวต์ ประกอบด้วย ตำแหน่งงาน,ระดับประสบการณ์,ประเภทการจ้างงาน,รูปแบบของการทำงาน,ปีที่

ทำงาน,ประเทศที่อาศัย,เงินเดือนสกุลเงินท้องถิ่น,หน่วยสกุลเงิน,เงินแปลงเป็น USD,ประเทศที่ตั้งบริษัท และขนาดบริษัท



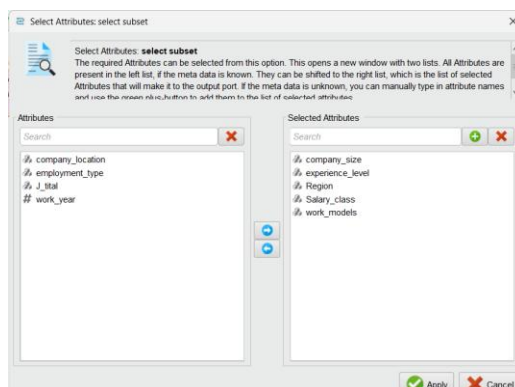
	A	B	C	D	E	F	G	H	I	J	K	L
	job_title	experience_level	employment_type	work_models	work_year	employee_residence	salary	salary_currency	salary_in_usd	company_location	company_size	
1	Data Engineer	Mid-level	Full-time	Remote	2024	United States	148100	USD	148100	United States	Medium	
2	Data Engineer	Mid-level	Full-time	Remote	2024	United States	98700	USD	98700	United States	Medium	
3	Data Scientist	Senior-level	Full-time	Remote	2024	United States	140032	USD	140032	United States	Medium	
4	Data Scientist	Senior-level	Full-time	Remote	2024	United States	100022	USD	100022	United States	Medium	
5	BI Developer	Mid-level	Full-time	On-site	2024	United States	120000	USD	120000	United States	Medium	
6	BI Developer	Mid-level	Full-time	On-site	2024	United States	42100	USD	42100	United States	Medium	
7	Research Analyst	Entry-level	Full-time	On-site	2024	United States	250000	USD	250000	United States	Medium	
8	Research Analyst	Entry-level	Full-time	On-site	2024	United States	150000	USD	150000	United States	Medium	
9	Data Engineer	Executive-level	Full-time	Remote	2024	United States	219450	USD	219450	United States	Medium	
10	Data Engineer	Executive-level	Full-time	Remote	2024	United States	136000	USD	136000	United States	Medium	
11	Business Intelligence Developer	Mid-level	Full-time	On-site	2024	United States	87800	USD	87800	United States	Medium	
12	Business Intelligence Developer	Mid-level	Full-time	On-site	2024	United States	76300	USD	76300	United States	Medium	
13	Data Scientist	Mid-level	Full-time	Remote	2024	United States	148100	USD	148100	United States	Medium	

ภาพที่ 4.1 แสดงข้อมูลเงินเดือน

4.1.3 คัดเลือกข้อมูล กลั่นกรองข้อมูลและแปลงรูปแบบข้อมูล (Data Preparation)

การคัดเลือกข้อมูล ที่เหมาะสมเพื่อนำไปใช้ในการวิเคราะห์ข้อมูลเลือกเฉพาะแอททริบิวท์ที่จำเป็นสำหรับการวิเคราะห์ การกลั่นกรองข้อมูลเป็นการทำความสะอาดข้อมูล ตรวจสอบ แก้ไขหรือลบรายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูลที่ดึงมา

4.1.3.1 ทำการคัดเลือกข้อมูล (Data Selection) คือ การคัดเลือกข้อมูลเป็นขั้นตอนสำคัญในการเตรียมข้อมูลสำหรับการวิเคราะห์ ผู้วิเคราะห์ข้อมูลได้ทำการเลือกและทำ Data Cleaning เพื่อตัดข้อมูลที่ไม่จำเป็นออก ให้เหลือเฉพาะข้อมูลที่เกี่ยวข้องกับการวิเคราะห์ความหลากหลายของ Data Scientist



ภาพที่ 4.2 คัดเลือกข้อมูล

4.1.3.2 ทำการกลั่นกรองข้อมูล (Data Cleaning) คือการทำความสะอาดข้อมูลเป็นกระบวนการการตรวจสอบและแก้ไข (หรือลบ) รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูลตารางหรือฐานข้อมูล ซึ่งเป็นหลักสำคัญของฐานข้อมูล จากการตรวจสอบข้อมูล ข้อมูลไม่มี ค่าผิดปกติ

4.1.3.3 การแปลงข้อมูล (Data Transformation) เป็นขั้นตอนที่นำข้อมูลที่รวบรวมมาและผ่านการเลือกแล้ว มาทำให้อยู่ในรูปแบบที่พร้อมใช้งานสำหรับการวิเคราะห์ในขั้นตอนถัดไป การแปลงนี้มีจุดประสงค์เพื่อให้ข้อมูลมีความถูกต้อง เหมาะสม และเข้าใจง่ายขึ้น โดยเฉพาะการแปลงค่าตัวเลขให้เป็นช่วง (Binning) และการจัดหมวดหมู่ (Categorical) เพื่อให้สอดคล้องกับการวิเคราะห์แบบ Classification ด้วยมี 3 คอลัมน์ Job_title, Salary_in_usd, company_location

job_title	engagement_level	employment_type	work_model	salary_in_usd	company_location	company_size
Data Engineer	Mid-level	Full-time	Remote	148100	United States	Medium
Data Engineer	Mid-level	Full-time	Remote	98700	United States	Medium
Data Scientist	DS/Analyst	Service-level	Full-time	140032	United States	Medium
Data Scientist	DS/Analyst	Service-level	Full-time	100022	United States	Medium
BI Developer	Other	Mid-level	Full-time	120000	United States	Medium
BI Developer	Other	Mid-level	Full-time	62100	United States	Medium
Research Analyst	Other	Entry-level	Full-time	250000	United States	Medium
Research Analyst	Other	Entry-level	Full-time	150000	United States	Medium
Data Engineer	Data Engineer	Executive-level	Full-time	219650	United States	Medium
Data Engineer	Data Engineer	Executive-level	Full-time	136000	United States	Medium
Business Intelligence Developer	Other	Mid-level	Full-time	87800	United States	Medium
Business Intelligence Developer	Other	Mid-level	Full-time	76300	United States	Medium
Data Scientist	DS/Analyst	Mid-level	Full-time	148100	United States	Medium

ภาพที่ 4.3 แสดงข้อมูลที่ถูกแปลง

salary_in_usd	Salary_class	company_location	Region
148100	High		
98700	Medium		
140032	High		
100022	High		
120000	High		
62100	Medium		
250000	High		
150000	High		
219650	High		
136000	High		
87800	Medium		

ภาพที่ 4.4 แสดงข้อมูลการกำหนดเงินเดือนให้อยู่ในรูปแบบ High Low และ Medium

การจัดกลุ่มเงินเดือน (Salary Grouping) เพื่อให้ง่ายต่อการวิเคราะห์ข้อมูลเงินเดือนของผู้ตอบแบบสอบถาม ทีมผู้วิจัยได้ทำการจัดกลุ่มเงินเดือน (Salary in USD) ต่อเดือน ออกเป็นสามกลุ่มหลัก ได้แก่ High, Medium, และ Low

หลักการจัดกลุ่ม

1) Low (ต่ำ) กลุ่มนี้รวมผู้ที่มีเงินเดือนรายเดือนต่ำกว่าเกณฑ์กลาง (เช่น ต่ำกว่า 2,000 USD ต่อเดือน) ใช้เพื่อวิเคราะห์ผู้ที่อยู่ในระดับเริ่มต้นหรือต่ำของตลาดแรงงาน

2) Medium (ปานกลาง) กลุ่มนี้รวมผู้ที่มีเงินเดือนอยู่ระหว่างเกณฑ์ต่ำและสูง (เช่น 2,000 – 5,000 USD ต่อเดือน) ใช้เพื่อเปรียบเทียบผู้ที่มีรายได้ในระดับทั่วไป

3) High (สูง) กลุ่มนี้รวมผู้ที่มีเงินเดือนสูงกว่าค่าเฉลี่ยหรือเกณฑ์สูง (เช่น มากกว่า 5,000 USD ต่อเดือน) ใช้เพื่อวิเคราะห์ผู้ที่อยู่ในตำแหน่งงานระดับสูงหรือมีความเชี่ยวชาญเฉพาะด้าน

job_title	J_tital	experience_level	employment_type
Data Engineer	Data Engine	เรียงลำดับจาก ก ถึง ข	
Data Engineer	Data Engine	เรียงลำดับจาก ข ถึง ก	
Data Scientist	DS/Analyst	เรียงลำดับตามสี	
Data Scientist	DS/Analyst	ล้างตัวกรองออกจาก "J_tital"	
BI Developer	Other	กรองตามสี	
BI Developer	Other	ตัวกรองข้อมูลตาม	
Research Analyst	Other	ค้นหา	
Research Analyst	Other	☒ (เลือกทั้งหมด)	
Data Engineer	Data Engine	☒ Data Engineer	
Data Engineer	Data Engine	☒ DS/Analyst	
Business Intelligence Developer	Other	☒ Management	
		☒ ML Engineer	
		☒ Other	

ภาพที่ 4.5 แสดงข้อมูลการจัดกลุ่มตำแหน่งงานให้เหลือ 5 กลุ่ม

การจัดกลุ่มตำแหน่งงาน (Job Title Grouping) ในการวิเคราะห์ข้อมูลตำแหน่งงาน (Job Title) ของผู้ตอบแบบสอบถาม ผู้วิจัยได้ทำการจัดกลุ่มตำแหน่งงานเพื่อให้ง่ายต่อการวิเคราะห์และการสร้างโมเดล โดยใช้ สูตร Excel แบบ Nested IF ซึ่งสามารถจำแนกตำแหน่งงานออกเป็นกลุ่มหลัก ๆ ดังนี้

1) DS/Analyst รวมตำแหน่งงานที่เกี่ยวข้องกับ Data Science และ Data Analysis เช่น Data Scientist และ Data Analyst การจัดกลุ่มนี้ช่วยให้นักวิจัยสามารถวิเคราะห์กลุ่มงานด้านข้อมูล (Data-Driven Roles) ได้ง่ายขึ้น

2) ML Engineer รวมตำแหน่งที่เกี่ยวข้องกับการพัฒนาและประยุกต์ใช้ Machine Learning เช่น Machine Learning Engineer หรือคำย่อ ML การจัดกลุ่มนี้มีประโยชน์ในการศึกษาแนวโน้มด้านเทคโนโลยีและความเชี่ยวชาญด้าน Machine Learning

3) Data Engineer รวมตำแหน่งงานที่เน้นด้านโครงสร้างข้อมูล การจัดการและปรับปรุงข้อมูล เช่น Data Engineer กลุ่มนี้ช่วยให้สามารถวิเคราะห์บุคลากรด้านข้อมูลเชิงโครงสร้างได้อย่างชัดเจน

4) Management รวมตำแหน่งงานด้านการบริหาร เช่น Manager และ Director การจัดกลุ่มนี้ใช้เพื่อแยกข้อมูลบุคลากรระดับผู้บริหารจากกลุ่มงานปฏิบัติการ

5) Other สำหรับตำแหน่งงานอื่น ๆ ที่ไม่ตรงกับกลุ่มที่กล่าวมาใช้เพื่อรวบรวมข้อมูลตำแหน่งที่ไม่เกี่ยวข้องโดยตรงกับกลุ่มงานด้านข้อมูลหรือการบริหาร

กระบวนการจัดกลุ่ม

- 1) ใช้ฟังก์ชัน SEARCH เพื่อตรวจสอบว่าคำหลัก (Keyword) ปรากฏอยู่ในชื่อ Job Title หรือไม่
- 2) ใช้ฟังก์ชัน ISNUMBER เพื่อตรวจสอบผลลัพธ์ของ SEARCH ว่าเป็นตัวเลข (พบคำค้น) หรือไม่
- 3) ใช้ฟังก์ชัน IF แบบซ้อนกัน (Nested IF) เพื่อกำหนดค่า Output ตามกลุ่มที่พบเป็นลำดับขั้น
- 4) หากไม่พบคำค้นใด ๆ ในชื่อ Job Title จะถูกจัดให้อยู่ในกลุ่ม "Other"

company_location	Region	company_size
United States	NA	
United States	NA	
United States	NA	
United States	NA	
United States	NA	
United States	NA	
United States	NA	
United States	NA	
United States	NA	
United States	NA	
United States	NA	

ภาพที่ 4.6 แสดงข้อมูลสถานที่บริษัทให้เป็น ทวี

การจัดกลุ่มตามภูมิภาค (Region Grouping) เพื่อให้ง่ายต่อการวิเคราะห์ข้อมูลผู้ตอบแบบสอบถาม ผู้วิจัยได้จัดกลุ่มประเทศตาม ภูมิภาค (Region) โดยใช้สูตร Excel แบบ Nested IF ร่วมกับฟังก์ชัน OR ซึ่งสามารถจำแนกประเทศในคอลัมน์ H (เช่น H16) ออกเป็นกลุ่มภูมิภาคหลักๆ ดังนี้

กลุ่มภูมิภาคและประเทศที่มีในข้อมูล

1) NA (North America) รวมประเทศในทวีปอเมริกาเหนือ คือ United States, Canada, Bahamas

2) SA (South America) รวมประเทศในทวีปอเมริกาใต้ คือ Brazil, Chile, Colombia

3) EU (Europe) รวมประเทศในทวีปยุโรป คือ Slovenia, Czech Republic, Belgium, Portugal, Netherlands, Luxembourg, Italy, Hungary, Croatia, Greece, Spain, Denmark, Sweden, Germany, Ireland, Gibraltar

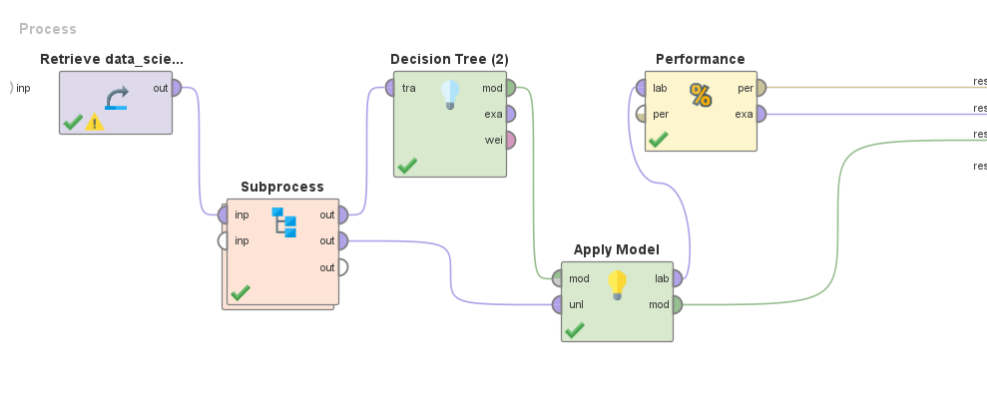
4) AS (Asia) รวมประเทศในทวีปเอเชีย คือ Thailand, Turkey, Russia, Mauritius, Lebanon, Vietnam, Saudi Arabia, India, Israel, Singapore, Hong Kong

5) OC (Oceania) รวมประเทศในทวีปโอเชียเนีย เช่น Australia, New Zealand

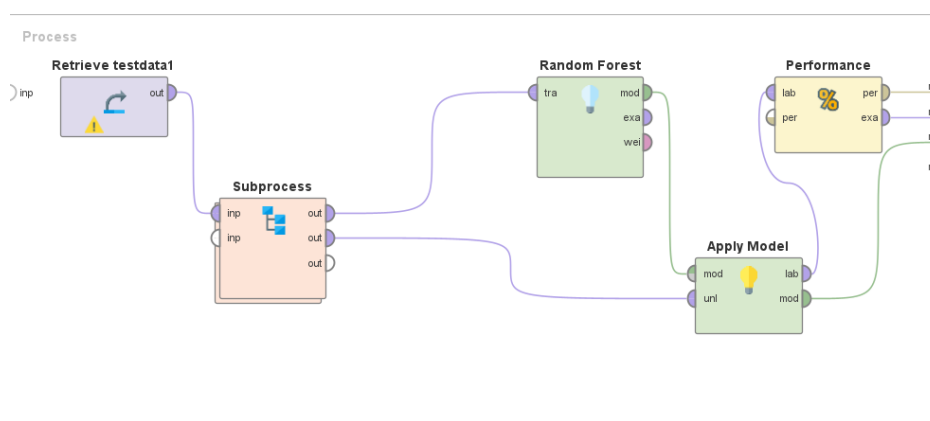
6) AF (Africa)สำหรับประเทศในทวีปแอฟริกา เช่น Nigeria

7) Other สำหรับประเทศอื่น ๆ ที่ไม่อยู่ในกลุ่มข้างต้น

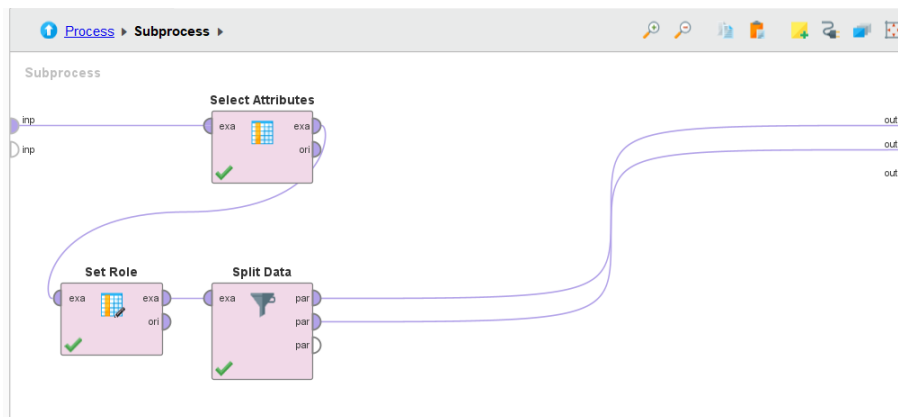
4.1.4 สร้างแบบจำลอง (Modeling) การสร้างแบบจำลองเป็นการใช้เทคนิคการเรียนรู้แบบกำกับดูแล (Supervised Learning) เพื่อจำแนกประเภท (Classification) โดยเลือกใช้สองโมเดลหลักคือ Decision Tree และ Random Forest ผ่านโปรแกรม RapidMiner



ภาพที่ 4.7 แสดงการทำโมเดล Decision Tree

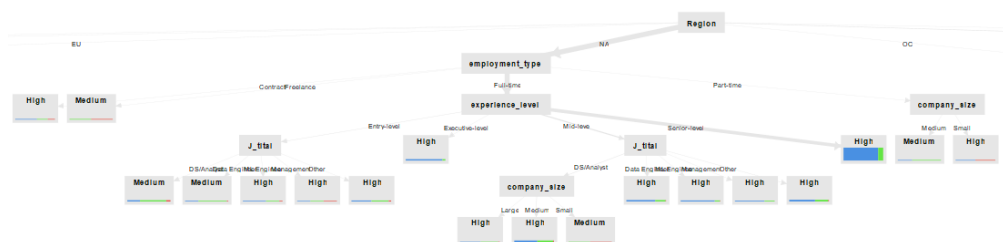


ภาพที่ 4.8 แสดงการทำโมเดล Random Forest



ภาพที่ 4.9 แสดง Operator ใน Subprocess

4.1.4.1 การสร้างโมเดล Decision Tree ใช้ Operator Decision Tree เพื่อสร้างโมเดลการจำแนกประเภท โดยตั้งค่า Parameter เช่น Criterion = Information Gain หรือ Gini Index Maximal Depth = กำหนดความลึกของต้นไม้เพื่อควบคุมความซับซ้อน ผลลัพธ์คือกฎการจำแนก (Classification Rules) ที่สามารถอธิบายได้ง่ายและมองเห็นโครงสร้างการตัดสินใจได้ชัดเจน



ภาพที่ 4.10 แสดงผลลัพธ์ของ Node Decision Tree


```

Tree
Region = AF
| experience_level = Entry-level: Low (High=0, Medium=0, Low=2)
| experience_level = Mid-level: Medium (High=0, Medium=2, Low=0)
| experience_level = Senior-level: High (High=1, Medium=1, Low=0)
Region = AS
| company_size = Large
| | experience_level = Entry-level: Low (High=0, Medium=2, Low=6)
| | experience_level = Mid-level: Low (High=0, Medium=3, Low=14)
| | experience_level = Senior-level: Medium (High=0, Medium=5, Low=4)
| company_size = Medium: Low (High=1, Medium=7, Low=15)
| company_size = Small
| | experience_level = Entry-level: Low (High=0, Medium=1, Low=3)
| | experience_level = Mid-level: Low (High=1, Medium=0, Low=2)
| | experience_level = Senior-level: High (High=1, Medium=0, Low=1)
Region = EU
| experience_level = Entry-level
| | company_size = Large
| | | J_title = DS/Analyst: Low (High=0, Medium=2, Low=4)
| | | J_title = Data Engineer: Medium (High=0, Medium=2, Low=2)
| | | J_title = ML Engineer: Low (High=0, Medium=1, Low=2)
| | | J_title = Other: Low (High=1, Medium=1, Low=4)
| | company_size = Medium: Low (High=0, Medium=2, Low=14)
| | company_size = Small
| | | employment_type = Full-time: Medium (High=0, Medium=4, Low=1)
| | | employment_type = Part-time: Low (High=0, Medium=0, Low=3)
| experience_level = Executive-level: High (High=3, Medium=2, Low=0)
| experience_level = Mid-level
| | J_title = DS/Analyst: Medium (High=1, Medium=16, Low=11)
| | J_title = Data Engineer: Medium (High=0, Medium=12, Low=5)
| | J_title = ML Engineer: Medium (High=2, Medium=10, Low=3)
| | J_title = Other

```

ภาพที่ 4.11 แสดงผลลัพธ์ของกฎ Decision Tree

จากผลการสร้างโมเดล Decision Tree เพื่อจำแนกระดับเงินเดือน (High, Medium, Low) ตามตัวแปรต่าง ๆ ได้แก่ ภูมิภาค (Region), ประสบการณ์ (Experience Level), และ ขนาดองค์กร (Company Size) พบว่าได้กฎการตัดสินใจดังต่อไปนี้

ภูมิภาคแอฟริกา (AF)

กฎที่ 1 หากผู้ปฏิบัติงานอยู่ในภูมิภาค Africa และมีระดับประสบการณ์ Entry-level ค่าจ้างอยู่ในระดับ Low (High = 0, Medium = 0, Low = 2) ผู้ที่เริ่มทำงานใหม่ในภูมิภาคแอฟริกาได้รับค่าตอบแทนค่อนข้างต่ำ

กฎที่ 2 หากผู้ปฏิบัติงานอยู่ในภูมิภาค Africa และมีระดับประสบการณ์ Mid-level ค่าจ้างอยู่ในระดับ Medium (High = 0, Medium = 2, Low = 0) ผู้ที่มีประสบการณ์ปานกลางในแอฟริกาได้รับเงินเดือนอยู่ในช่วงระดับกลาง

กฎที่ 3 หากผู้ปฏิบัติงานอยู่ในภูมิภาค Africa และมีระดับประสบการณ์ Senior-level ค่าจ้างอยู่ในระดับ High (High = 1, Medium = 1, Low = 0) ผู้ที่มีประสบการณ์สูงในภูมิภาคแอฟริกา มีแนวโน้มได้รับเงินเดือนสูง

ภูมิภาคเอเชีย (AS)

กฎที่ 4 หากอยู่ในภูมิภาค Asia บริษัทมีขนาด Large และระดับประสบการณ์ Entry-level ค่าจ้างอยู่ในระดับ Low (High = 0, Medium = 2, Low = 6) บริษัทขนาดใหญ่ในเอเชียจ่ายเงินเดือนต่ำให้กับพนักงานระดับเริ่มต้น

กฎที่ 5 หากอยู่ในภูมิภาค Asia บริษัทมีขนาด Large และระดับประสบการณ์ Mid-level ค่าจ้างอยู่ในระดับ Low (High = 0, Medium = 3, Low = 14) พนักงานระดับกลางในบริษัทใหญ่ของเอเชียส่วนใหญ่ยังได้รับค่าจ้างต่ำ

กฎที่ 6 หากอยู่ในภูมิภาค Asia บริษัทมีขนาด Large และระดับประสบการณ์ Senior-level ค่าจ้างอยู่ในระดับ Medium (High = 0, Medium = 5, Low = 4) ผู้ที่มีประสบการณ์สูงในบริษัทใหญ่ของเอเชียเริ่มมีรายได้อยู่ในระดับปานกลาง

กฎที่ 7 หากอยู่ในภูมิภาค Asia บริษัทมีขนาด Medium ค่าจ้างอยู่ในระดับ Low (High = 1, Medium = 7, Low = 15) บริษัทขนาดกลางในเอเชียส่วนใหญ่ให้ค่าตอบแทนระดับต่ำ

กฎที่ 8 หากอยู่ในภูมิภาค Asia บริษัทมีขนาด Small และระดับประสบการณ์ Entry-level ค่าจ้างอยู่ในระดับ Low (High = 0, Medium = 1, Low = 3) พนักงานใหม่ในบริษัทขนาดเล็กของเอเชียได้รับเงินเดือนต่ำที่สุดในภูมิภาค

ภูมิภาคยุโรป (EU)

กฎที่ 9 หากอยู่ในภูมิภาค Europe และมีระดับประสบการณ์ Entry-level ค่าจ้างอยู่ในระดับ Medium (High = 0, Medium = 2, Low = 0) ผู้เริ่มต้นทำงานในยุโรปได้รับเงินเดือนระดับปานกลาง สะท้อนถึงมาตรฐานค่าจ้างที่สูงกว่าภูมิภาคอื่น

กฎที่ 10 หากอยู่ในภูมิภาค Europe และมีระดับประสบการณ์ Mid-level ค่าจ้างอยู่ในระดับ High (High = 3, Medium = 1, Low = 0) ผู้มีประสบการณ์ระดับกลางในยุโรปมีแนวโน้มได้รับเงินเดือนสูง แสดงถึงความคุ้มค่าของทักษะในตลาดแรงงานยุโรป

กฎที่ 11 หากอยู่ในภูมิภาค Europe และมีระดับประสบการณ์ Senior-level ค่าจ้างอยู่ในระดับ High (High = 5, Medium = 0, Low = 0) ผู้ที่มีประสบการณ์สูงในยุโรปได้รับค่าตอบแทนสูงสุดเมื่อเทียบกับระดับอื่น

ภูมิภาคอเมริกาเหนือ (NA)

กฎที่ 12 หากอยู่ในภูมิภาค North America และมีระดับประสบการณ์ Entry-level ค่าจ้างอยู่ในระดับ Medium (High = 1, Medium = 4, Low = 0) พนักงานเริ่มต้นในอเมริกาเหนือได้รับค่าจ้างปานกลาง ซึ่งสะท้อนถึงโอกาสการเติบโตในสายงานที่ดี

กฎที่ 13 หากอยู่ในภูมิภาค North America และมีระดับประสบการณ์ Mid-level ค่าจ้างอยู่ในระดับ High (High = 5, Medium = 1, Low = 0) ผู้ที่มีประสบการณ์ระดับกลางในอเมริกาเหนือได้รับเงินเดือนสูง แสดงถึงความต้องการแรงงานทักษะในตลาด

กฎที่ 14 หากอยู่ในภูมิภาค North America และมีระดับประสบการณ์ Senior-level ค่าจ้างอยู่ในระดับ High (High = 8, Medium = 0, Low = 0) พนักงานระดับอาวุโสในอเมริกาเหนือมีรายได้สูงที่สุดในกลุ่มทั้งหมด

ภูมิภาคอเมริกาใต้ (SA)

กฎที่ 15 หากอยู่ในภูมิภาค South America และมีระดับประสบการณ์ Entry-level ค่าจ้างอยู่ในระดับ Low (High = 0, Medium = 0, Low = 3) พนักงานเริ่มต้นในอเมริกาใต้ได้รับเงินเดือนต่ำเมื่อเทียบกับภูมิภาคอื่น

กฎที่ 16 หากอยู่ในภูมิภาค South America และมีระดับประสบการณ์ Mid-level ค่าจ้างอยู่ในระดับ Medium (High = 0, Medium = 2, Low = 0) ผู้ที่มีประสบการณ์ระดับกลางในอเมริกาใต้ได้รับค่าจ้างปานกลาง

กฎที่ 17 หากอยู่ในภูมิภาค South America และมีระดับประสบการณ์ Senior-level ค่าจ้างอยู่ในระดับ High (High = 2, Medium = 0, Low = 0) พนักงานอาวุโสในอเมริกาใต้มีแนวโน้มได้รับค่าตอบแทนสูงขึ้นตามประสบการณ์

ภูมิภาคโอเชียเนีย (OC)

กฎที่ 18 หากอยู่ในภูมิภาค Oceania และมีระดับประสบการณ์ Entry-level ค่าจ้างอยู่ในระดับ Medium (High = 0, Medium = 2, Low = 0) ผู้เริ่มต้นทำงานในโอเชียเนียได้รับเงินเดือนระดับปานกลาง

กฎที่ 19 หากอยู่ในภูมิภาค Oceania และมีระดับประสบการณ์ Mid-level ค่าจ้างอยู่ในระดับ High (High = 2, Medium = 0, Low = 0) ผู้ที่มีประสบการณ์ระดับกลางในภูมิภาคนี้ได้รับค่าจ้างสูง

กฎที่ 20 หากอยู่ในภูมิภาค Oceania และมีระดับประสบการณ์ Senior-level ค่าจ้างอยู่ในระดับ High (High = 3, Medium = 0, Low = 0) พนักงานระดับสูงในโอเชียเนียได้รับค่าตอบแทนสูงสุดในภูมิภาค

บทสรุปภาพรวมของกฎทั้งหมด

ภูมิภาค (Region) เป็นตัวแปรสำคัญที่สุดที่มีอิทธิพลต่อระดับเงินเดือน พื้นที่ยุโรป (EU), อเมริกาเหนือ (NA) และโอเชียเนีย (OC) มีแนวโน้มจ่ายเงินเดือนสูง ขณะที่เอเชีย (AS) และแอฟริกา (AF) มีระดับเงินเดือนโดยเฉลี่ยต่ำกว่า ประสบการณ์ (Experience Level) มีผลโดยตรงต่อรายได้

Entry-level ☐ เงินเดือนต่ำ

Mid-level ☐ ปานกลางถึงสูง

Senior-level ☐ สูงสุดในทุกภูมิภาค

ขนาดองค์กร (Company Size) มีผลเฉพาะในเอเชีย บริษัทใหญ่ (Large) มักจ่ายต่ำกว่าที่คาด เนื่องจากจำนวนพนักงานมากส่วนบริษัทขนาดเล็กและกลางก็มีค่าจ้างใกล้เคียงกันโดยอยู่ในระดับต่ำ

PerformanceVector

```

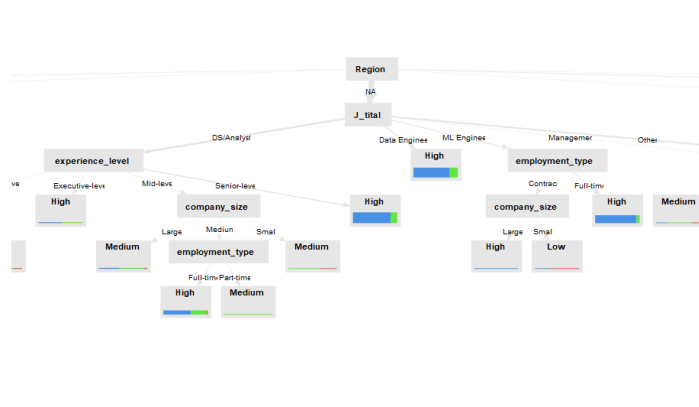
PerformanceVector:
accuracy: 89.17%
ConfusionMatrix:
True:   Medium  Large  Small
Medium: 1168    106    29
Large:   4       7      3
Small:   0       1      2
kappa: 0.116
ConfusionMatrix:
True:   Medium  Large  Small
Medium: 1168    106    29
Large:   4       7      3
Small:   0       1      2

```

ภาพที่ 4.12 แสดงผลลัพธ์ของ PerformanceVector Decision Tree

โมเดล แม่นยำกับ Class Medium มากที่สุด (เพราะข้อมูลส่วนใหญ่เป็น Medium class imbalance) แต่มีปัญหาการทำนาย Large และ Small เนื่องจากจำนวนข้อมูลน้อย □ ทำให้โมเดลเรียนรู้ได้ไม่ดี จึงได้ Accuracy สูง แต่ Kappa ต่ำ

4.1.4.2 การสร้างโมเดล Random Forest ใช้ Operator Random Forest โดยสุ่มตัวแปรและสร้างต้นไม้หลายต้นเพื่อลดความเอนเอียง (Bias) และเพิ่มความแม่นยำ (Accuracy) การตั้งค่าที่สำคัญ Number of Trees = จำนวนต้นไม้ Maximal Depth = ความลึกสูงสุดของแต่ละต้นไม้ Voting Type = Majority Vote เพื่อใช้ผลรวมของการทำนาย



ภาพที่ 4.13 แสดงผลลัพธ์ของ Node Random Forest

PerformanceVector

```

PerformanceVector:
accuracy: 88.82%
ConfusionMatrix:
True:   Medium  Large  Small
Medium: 1054    93     26
Large:  11      10     4
Small:  0        0     1
kappa: 0.135
ConfusionMatrix:
True:   Medium  Large  Small
Medium: 1054    93     26
Large:  11      10     4
Small:  0        0     1

```

ภาพที่ 4.14 แสดงผลลัพธ์ของ PerformanceVector ของ Radom Forest

หาค่า Precision และ Recall ของ Radom Forest

1) Precision (ความแม่นยำ)

PrecisionHigh = $\frac{TP}{TP + FP}$
 $\text{Precision}_{\text{High}} = \frac{TP}{TP + FP}$

TP (True Positive) = จำนวนตัวอย่างที่ทำนายถูกต้องว่าเป็น High = 830

FP (False Positive) = จำนวนตัวอย่างที่ไม่เคยทำนายเป็น High แต่จริง ๆ ไม่ใช่ High
 = 27 + 2 = 29

PrecisionHigh = $\frac{830}{830+29} = \frac{830}{859} \approx 0.966$
 $\text{Precision}_{\text{High}} = \frac{830}{830+29} = \frac{830}{859} \approx 0.966$

2) Recall (ความครอบคลุม / Sensitivity)

RecallHigh = $\frac{TP}{TP + FN}$
 $\text{Recall}_{\text{High}} = \frac{TP}{TP + FN}$

FN (False Negative) = จำนวนตัวอย่างที่จริงเป็น High แต่โมเดลทำนายผิด = 184 + 7 = 191

RecallHigh = $\frac{830}{830+191} = \frac{830}{1021} \approx 0.813$
 $\text{Recall}_{\text{High}} = \frac{830}{830+191} = \frac{830}{1021} \approx 0.813$

3.1.5 การประสิทธิภาพของตัวแบบ (Evaluation) วัดผลประสิทธิภาพทั้ง 2 เทคนิค คือ เทคนิคตัดสินใจ (Decision tree) และ เทคนิคแรนดอมฟอเรสต์ (Random Forest) เลือกเทคนิคที่มีค่าความแม่นยำที่สุด

ตารางที่ 4.1 ตารางเปรียบเทียบประสิทธิภาพของโมเดล

โมเดล	Accuracy	kappa	Precision			Recall		
			H	M	L	H	M	L
Decision tree	89.17%	1.110	0.954	0.310	0.261	0.815	0.556	0.419
Random Forest	88.82%	0.135	0.966	0.266	0.413	0.813	0.565	0.542

จากตารางแสดงให้เห็นว่า Decision Tree มีค่า Accuracy สูงกว่าเล็กน้อย (89.17%) เมื่อเทียบกับ Random Forest (88.82%) และยังมีค่า RMSE และ MAE ต่ำกว่า จึงสรุปว่า Decision Tree เป็นโมเดลที่เหมาะสมที่สุด

3.1.5 การนำโมเดลไปใช้งานจริง (Deployment) ได้ทำการวัดผลประสิทธิภาพของทั้งสองเทคนิค ได้แก่ Decision Tree และ Random Forest โดยใช้ตัวชี้วัดด้านการทำนาย Accuracy เพื่อพิจารณาเลือกโมเดลที่มีประสิทธิภาพดีที่สุด

1) การเลือกโมเดลที่เหมาะสม

จากการวิเคราะห์โดยใช้โปรแกรม RapidMiner พบว่า Decision Tree เป็นโมเดลที่มี Accuracy สูงสุด และมีประสิทธิภาพโดยรวมดีกว่า จึงถูกเลือกให้นำไปใช้งานจริง

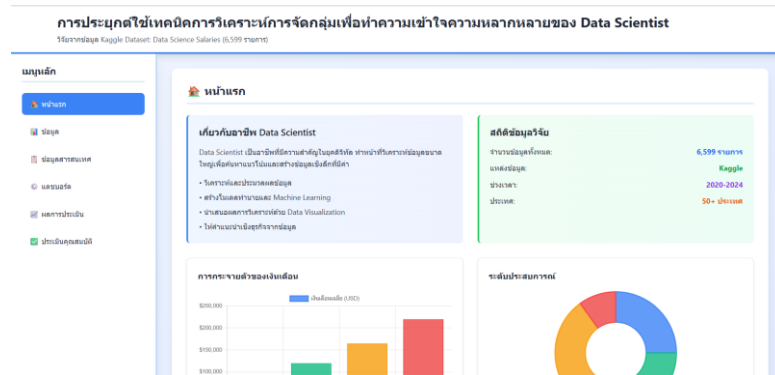
2) การประยุกต์ใช้งาน

การวิเคราะห์ตลาดแรงงาน (Labor Market Analysis) ใช้จำแนกและพยากรณ์ช่วงเงินเดือนของ Data Scientist ตามตำแหน่งงาน ระดับประสบการณ์ และสถานที่ทำงาน

4.2 การออกแบบเว็บไซต์

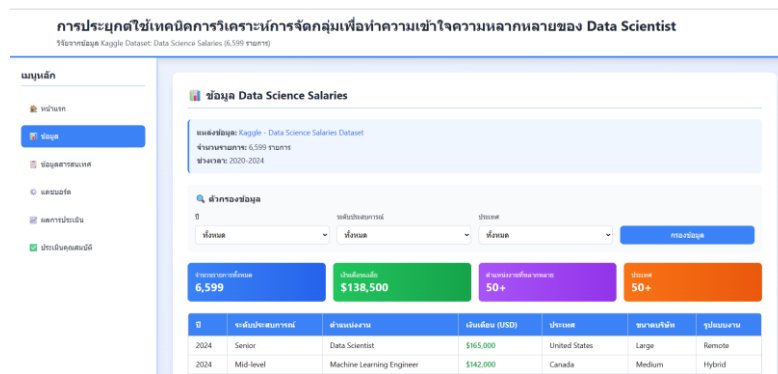
3.2.1 การออกแบบ Wireframe

1) หน้าโฮมเพจของเว็บไซต์ แสดงเมนูต่างๆ ของหน้าเว็บไซต์ ข่าวสาร และ Dashboard



ภาพที่ 4.15 หน้าโฮมเพจของเว็บไซต์

2) หน้าเว็บเพจที่ 2 ชุดข้อมูล



ภาพที่ 4.17 หน้าเว็บเพจที่ 2 ชุดข้อมูล

3) หน้าเว็บเพจที่ 3 ชุดข้อมูลสารสนเทศ

การประยุกต์ใช้เทคนิคการวิเคราะห์การจัดกลุ่มเพื่อทำความเข้าใจความหลากหลายของ Data Scientist

ข้อมูลจาก Kaggle Dataset: Data Science Salaries (6,599 รายการ)

เมนูหลัก

- หน้าแรก
- ข้อมูล
- ข้อมูลสารสนเทศ
- เอกสาร
- ผลการประเมิน
- หน้าเว็บเพจ

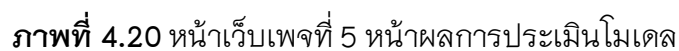
ข้อมูลสารสนเทศ

job_title	J_title	experience_level	employment_type	work_model	salary	Salary class	company_location	Region	company_size	work_year
Data Engineer	Data Engineer	Mid-level	Full-time	Remote	140100	High	United States	NA	Medium	2024
Data Engineer	Data Engineer	Mid-level	Full-time	Remote	98700	Medium	United States	NA	Medium	2024
Data Scientist	DS/Analyst	Senior-level	Full-time	Remote	140002	High	United States	NA	Medium	2024
Data Scientist	DS/Analyst	Senior-level	Full-time	Remote	100002	High	United States	NA	Medium	2024
BI Developer	Other	Mid-level	Full-time	On-site	120000	High	United States	NA	Medium	2024
BI Developer	Other	Mid-level	Full-time	On-site	62100	Medium	United States	NA	Medium	2024
Research Analyst	Other	Entry-level	Full-time	On-site	250000	High	United States	NA	Medium	2024
Research Analyst	Other	Entry-level	Full-time	On-site	130000	High	United States	NA	Medium	2024
Data Engineer	Data Engineer	Executive-level	Full-time	Remote	219600	High	United States	NA	Medium	2024
Data Engineer	Data Engineer	Executive-level	Full-time	Remote	136000	High	United States	NA	Medium	2024
Business Intelligence Developer	Other	Mid-level	Full-time	On-site	87800	Medium	United States	NA	Medium	2024
Business Intelligence Developer	Other	Mid-level	Full-time	On-site	87800	Medium	United States	NA	Medium	2024

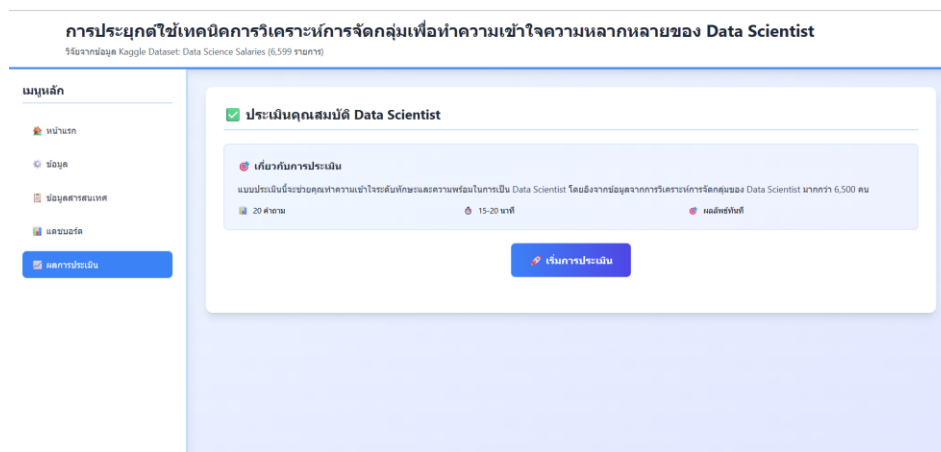
ภาพที่ 4.18 หน้าเว็บเพจที่ 3 ชุดข้อมูลสารสนเทศ



5) หน้าเว็บเพจที่ 5 แสดงหน้าผลการประเมินโมเดล



6) หน้าเว็บเพจที่ 6 แสดงหน้าการประเมินคุณสมบัติ



ภาพที่ 4.19 หน้าเว็บเพจที่ 6 แสดงหน้าการประเมินคุณสมบัติ