บทที่ 3

วิธีการดำเนินงานโครงงาน

โครงงานเรื่องการประยุกต์ใช้เทคนิคการวิเคราะห์ข้อมูลเพื่อทำความเข้าใจความหลากหลายของ Data Scientist มีวัตถุประสงค์เพื่อวิเคราะห์ความแตกต่างของระดับเงินเดือนและลักษณะเฉพาะของ Data Scientist พร้อมเผยแพร่ผลลัพธ์ผ่านเว็บไซต์เริ่มจากการรวบรวมข้อมูลจากชุดข้อมูล “data_science_salaries” บน Kaggle จากนั้นทำความสะอาดข้อมูลและคัดเลือกตัวแปรที่สำคัญ เช่น ประสบการณ์ ระดับการศึกษา สถานที่ทำงาน และทักษะ เพื่อเตรียมพร้อมสำหรับการสร้างโมเดล ต่อมาสร้างแบบจำลองด้วย Decision Tree และ Random Forest ผ่าน RapidMiner เพื่อจำแนกระดับเงินเดือนของ Data Scientist ตามปัจจัยสำคัญต่าง ๆ และประเมินความถูกต้องของโมเดลด้วยตัวชี้วัดมาตรฐาน เช่น Accuracy, Precision, Recall และ Confusion Matrix สุดท้าย ผลลัพธ์ถูกนำไปพัฒนาเป็น เว็บแอปพลิเคชัน พร้อมกราฟและตาราง ทำให้ผู้ใช้งานทั่วไปสามารถเข้าถึงข้อมูล วิเคราะห์แนวโน้มตลาดแรงงาน และวางแผนพัฒนาทักษะได้สะดวกและมีประสิทธิภาพ

3.1 การวิเคราะห์ข้อมูลด้วย CRISP – DM

3.2 การออกแบบเว็บไซต์

3.3 บทสรุป

3.1 การวิเคราะห์ข้อมูลด้วย CRISP – DM

กระบวนการวิเคราะห์ข้อมูลด้วย CRISP–DM หรือ Cross–industry standard process for data mining พัฒนาขึ้นในปี ค.ศ 1996 โดยความร่วมมือของ 3 บริษัทคือ Daimler, SPSS และ NCR ที่มีการพัฒนาเป็น Work flow มาตารฐานสำหรับการทำเหมืองข้อมูล ประกอบด้วย 6 ขั้นตอนหลักดังนี้

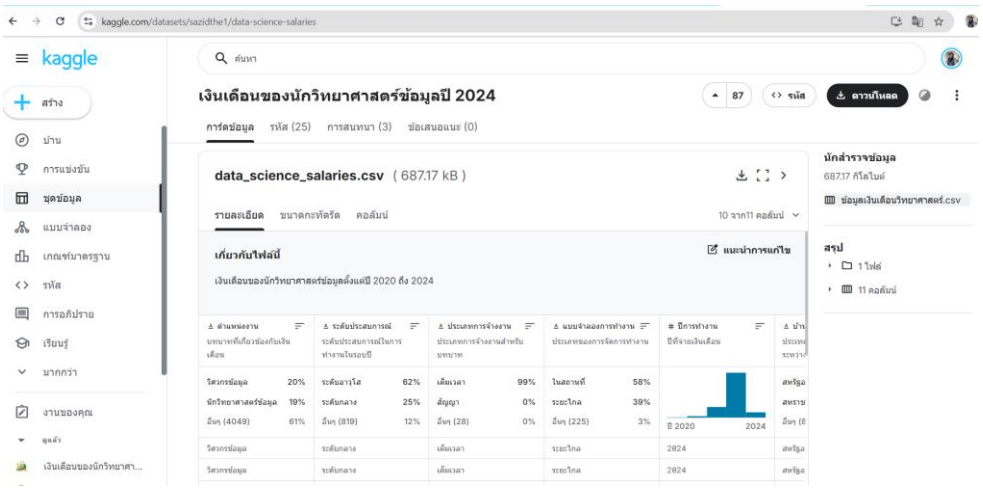
ภาพที่ 3.1 แสดง CRISP–DM

(ที่มา: https://kamboonchob.medium.com/)

3.1.1 ศึกษาข้อมูลเงินเดือนของ Data Scientists (Business Understanding) ขั้นตอนแรกคือการศึกษาและทำความเข้าใจปัจจัยที่มีผลต่อเงินเดือนของ Data Scientist เพื่อกำหนดวัตถุประสงค์และขอบเขตของการวิเคราะห์ ข้อมูลจากการศึกษาพบว่าปัจจัยสำคัญได้แก่ 1) ประสบการณ์ (Experience) – Data Scientist ที่มีประสบการณ์สูงมักได้รับเงินเดือนมากกว่าผู้ที่เพิ่งเริ่มงานหรือมีประสบการณ์น้อย 2) การศึกษาและคุณวุฒิ (Education & Qualification) – การศึกษาที่สูงขึ้นหรือปริญญาในสาขาที่เกี่ยวข้อง เช่น ปริญญาตรีหรือปริญญาโท มีผลต่อระดับเงินเดือน 3) สถานที่ทำงาน (Work Location) – เงินเดือนแตกต่างตามภูมิภาค เช่น เมืองใหญ่หรือพื้นที่ที่มีค่าครองชีพสูงมักมีเงินเดือนสูงกว่า 4) ประเภทของบริษัท (Company Type) – บริษัทขนาดใหญ่หรือบริษัทที่มีเทคโนโลยีทันสมัยมักจ่ายเงินเดือนสูงกว่าบริษัทขนาดเล็กหรือเพิ่งเริ่มต้น5) ทักษะเฉพาะ (Specialized Skills) ความเชี่ยวชาญในเทคโนโลยีหรือเครื่องมือเฉพาะ เช่น การวิเคราะห์ข้อมูลทางการเงินหรือการใช้เครื่องมือวิเคราะห์ขั้นสูง สามารถเพิ่มมูลค่าเงินเดือน

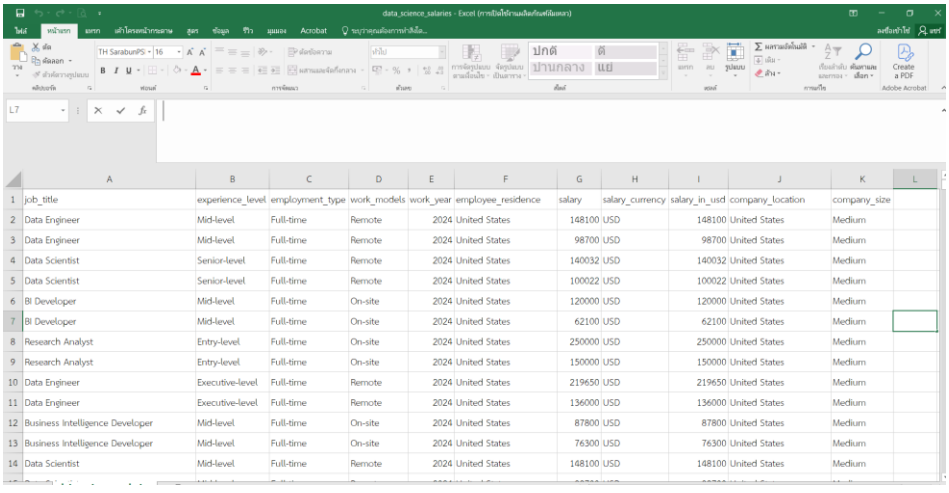
3.1.2 รวบรวมและทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนการจัดเก็บรวบรวมข้อมูล ตรวจสอบความถูกต้องของข้อมูลที่มาจาก เว็บไซต์ Kaggle.com โดยเลือกว่าจะใช้ข้อมูลบางส่วนในการวิเคราะห์ให้สอดกับวัตถุประสงค์ที่กำหนดไว้ ชุดข้อมู “data_science_salaries”

การรวบรวมข้อมูล เงินเดือน, ประสบการณ์, ทักษะ และตำแหน่งงานการตรวจสอบความสมบูรณ์ของข้อมูล



ภาพที่ 3.2 เว็บไซต์ Kaggle.com ชุดข้อมูล “data_science_salaries”

ในไฟล์ข้อมูลจะมีจำนวนข้อมูล 6599 รายการ ประกอบไปด้วย 11 แอตทริบิวท์ ประกอบด้วย ตำแหน่งงาน,ระดับประสบการณ์,ประเภทการรจ้างงาน,รูปแบบของการทำงาน,ปีที่ทำงาน,ประเทศที่อาศัย,เงินเดือนสกุลเงินท้องถิ่น,หน่วยสกุลเงิน,เงินแปลงเป็น USD,ประเทศที่ตั้งบริษัท และขนาดบริษัท ดังภาพที่ 3.3



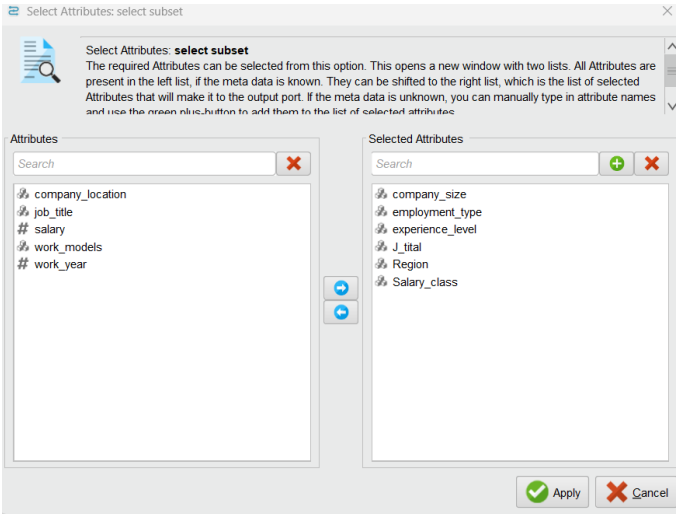
	job_title	experience_level	employment_type	work_models	work_year	employee_residence	salary	salary_currency	salary_in_usd	company_location	company_size
2	Data Engineer	Mid-level	Full-time	Remote	2024	United States	148100	USD	148100	United States	Medium
3	Data Engineer	Mid-level	Full-time	Remote	2024	United States	98700	USD	98700	United States	Medium
4	Data Scientist	Senior-level	Full-time	Remote	2024	United States	140032	USD	140032	United States	Medium
5	Data Scientist	Senior-level	Full-time	Remote	2024	United States	100022	USD	100022	United States	Medium
6	BI Developer	Mid-level	Full-time	On-site	2024	United States	120000	USD	120000	United States	Medium
7	BI Developer	Mid-level	Full-time	On-site	2024	United States	62100	USD	62100	United States	Medium
8	Research Analyst	Entry-level	Full-time	On-site	2024	United States	250000	USD	250000	United States	Medium
9	Research Analyst	Entry-level	Full-time	On-site	2024	United States	150000	USD	150000	United States	Medium
10	Data Engineer	Executive-level	Full-time	Remote	2024	United States	219650	USD	219650	United States	Medium
11	Data Engineer	Executive-level	Full-time	Remote	2024	United States	136000	USD	136000	United States	Medium
12	Business Intelligence Developer	Mid-level	Full-time	On-site	2024	United States	87800	USD	87800	United States	Medium
13	Business Intelligence Developer	Mid-level	Full-time	On-site	2024	United States	76300	USD	76300	United States	Medium
14	Data Scientist	Mid-level	Full-time	Remote	2024	United States	148100	USD	148100	United States	Medium

ภาพที่ 3.3 แสดงข้อมูลเงินเดือน

3.1.3 คัดเลือกข้อมูล กลั่นกรองข้อมูลและแปลงรูปแบบข้อมูล (Data Preparation)

การคัดเลือกข้อมูล ที่เหมาะสมเพื่อนำไปใช้ในการวิเคราะห์ข้อมูลเลือกเฉพาะแอททริบิวท์ที่จำเป็นสำหรับการวิเคราะห์ การกลั่นกรองข้อมูลเป็นการทำความสะอาดข้อมูล ตรวจสอบ แก้ไข หรือลบรายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูลที่ดึงมา โดยการทำให้เป็นข้อมูลที่ถูกต้อง (Data cleaning) มักใช้เวลาค่อนข้างมาก โดยมีขั้นตอนดังนี้

3.1.3.1 ทำการคัดเลือกข้อมูล (Data Selection) คือ การคัดเลือกข้อมูลเป็นขั้นตอนสำคัญในการเตรียมข้อมูลสำหรับการวิเคราะห์ ผู้วิเคราะห์ข้อมูลได้ทำการ เลือกและทำ Data Cleaning เพื่อตัดข้อมูลที่ไม่จำเป็นออก ให้เหลือเฉพาะข้อมูลที่เกี่ยวข้องกับการวิเคราะห์ความหลากหลายของ Data Scientist

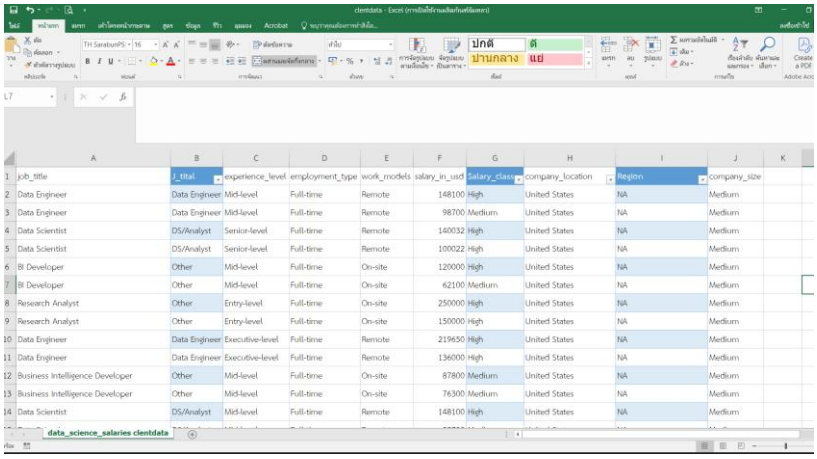


ภาพที่ 3.4 แสดงข้อมูลที่ถูกเลือก

ผู้วิเคราะห์ข้อมูลทำการคัดเลือกข้อมูล และทำการ Data Cleaning เพื่อศึกษาปัจจัยที่มีผลต่อเงินเดือนของสายอาชีพ Data Scientist โดยเลือกข้อมูลที่จำเป็นจำนวน 6,599 รายการ ข้อมูลประกอบด้วย 11 แอตทริบิวท์ ได้แก่ ตำแหน่งงาน ระดับประสบการณ์ ประเภทการจ้างงาน รูปแบบการทำงาน ปีที่ทำงาน ประเทศที่อาศัย เงินเดือนสกุลเงินท้องถิ่น หน่วยสกุลเงิน เงินเดือนแปลงเป็น USD ประเทศที่ตั้งบริษัท และขนาดบริษัท ข้อมูลเหล่านี้ถือเป็นตัวแปรสำคัญที่ใช้ในการวิเคราะห์และสร้างโมเดล Clustering

3.1.3.2 ทำการกลั่นกรองข้อมูล (Data Cleaning) คือการทำความสะอาดข้อมูลเป็นกระบวนการการตรวจสอบและแก้ไข (หรือลบ) รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล ตารางหรือฐานข้อมูล ซึ่งเป็นหลักสำคัญของฐานข้อมูล จากการตรวจสอบข้อมูล ข้อมูลไม่มี ค่าผิดปกติ

3.1.3.3 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	experience_level	employment_type	work_models	salary_in_usd	Salary_class	company_location	Region	company_size
Data Engineer	Mid-level	Full-time	Remote	148100	High	United States	NA	Medium
Data Engineer	Mid-level	Full-time	Remote	98700	Medium	United States	NA	Medium
Data Scientist	Senior-level	Full-time	Remote	140032	High	United States	NA	Medium
Data Scientist	Senior-level	Full-time	Remote	100022	High	United States	NA	Medium
BI Developer	Mid-level	Full-time	On-site	120000	High	United States	NA	Medium
BI Developer	Mid-level	Full-time	On-site	62100	Medium	United States	NA	Medium
Research Analyst	Entry-level	Full-time	On-site	250000	High	United States	NA	Medium
Research Analyst	Entry-level	Full-time	On-site	150000	High	United States	NA	Medium
Data Engineer	Executive-level	Full-time	Remote	219650	High	United States	NA	Medium
Data Engineer	Executive-level	Full-time	Remote	136000	High	United States	NA	Medium
Business Intelligence Developer	Mid-level	Full-time	On-site	87800	Medium	United States	NA	Medium
Business Intelligence Developer	Mid-level	Full-time	On-site	76300	Medium	United States	NA	Medium
Data Scientist	Mid-level	Full-time	Remote	148100	High	United States	NA	Medium

ภาพที่ 3.5 แสดงข้อมูลที่ถูกแปลง

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

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

1) เลือกชุดข้อมูลที่นำเข้าสู่โปรแกรม Rapid Miner

2) เลือกข้อมูลที่จะนำไปวิเคราะห์ โดยใช้ operators ที่ชื่อว่า select Attiributes

3) ทำการกำหนดตัวแปรโดยใช้ operators ที่ชื่อว่า Set Role

4) แบ่งข้อมูลออกเป็น 2 ส่วนเพื่อให้ได้ 80% ในการสร้างโมเดลและเพื่อให้ได้โมเดลที่มีประสิทธิภาพ โดยใช้ operators ที่ชื่อว่า Split data

5) สร้างตัวแบบ Decision Tree โดยใช้ operators ที่ชื่อว่า Decision Tree

6) การวัดประสิทธิภาพ operators ที่ชื่อว่า Appy Model

7) แสดงผลการวัดประสิทธิภาพ operators ที่ชื่อว่า Performance (Regression)

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

1) เลือกชุดข้อมูลที่นำเข้าสู่โปรแกรม Rapid Miner

2) เลือกข้อมูลที่จะนำไปวิเคราะห์ โดยใช้ operators ที่ชื่อว่า select Attiributes

3) ทำการกำหนดตัวแปรโดยใช้ operators ที่ชื่อว่า Set Role

4) แบ่งข้อมูลออกเป็น 2 ส่วนเพื่อให้ได้ 80% ในการสร้างโมเดลและเพื่อให้ได้โมเดลที่มีประสิทธิภาพ โดยใช้ operators ที่ชื่อว่า Split data

5) สร้างตัวแบบ Random Forest โดยใช้ operators ที่ชื่อว่า Random Forest

6) การวัดประสิทธิภาพ operators ที่ชื่อว่า Appy Model

7) แสดงผลการวัดประสิทธิภาพ operators ที่ชื่อว่า Performance (Regression)

3.1.5 การประสิทธิภาพของตัวโมเดล (Evaluation) วัดผลประสิทธิภาพทั้ง 2 เทคนิค คือ เทคนิคดิซิชั่นทรี (Decision tree) และ เทคนิคแรนดอมฟอเรสต์ (Random Forest) เลือกเทคนิคที่มีค่าความแม่นยำที่สุด

ตารางที่ 3.1 ตารางเปรียบเทียบประสิทธิภาพของโมเดล

โมเดล	***Accuracy***	***kappa***	***Precision***	***Recall***
Decision tree				
Random Forest				

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

1) การเลือกโมเดลที่เหมาะสม

การเลือกโมเดลที่เหมาะสมเป็นขั้นตอนสำคัญในการนำไปใช้งานจริง โดยทำการสร้างและเปรียบเทียบโมเดลหลายเทคนิค ได้แก่ Decision Tree และ Random Forest ซึ่งทั้งสองโมเดลถูกประเมินด้วยตัวชี้วัดมาตรฐาน เช่น Accuracy, Precision, Recall เพื่อพิจารณาประสิทธิภาพของการทำนาย

2) การประยุกต์ใช้งาน

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

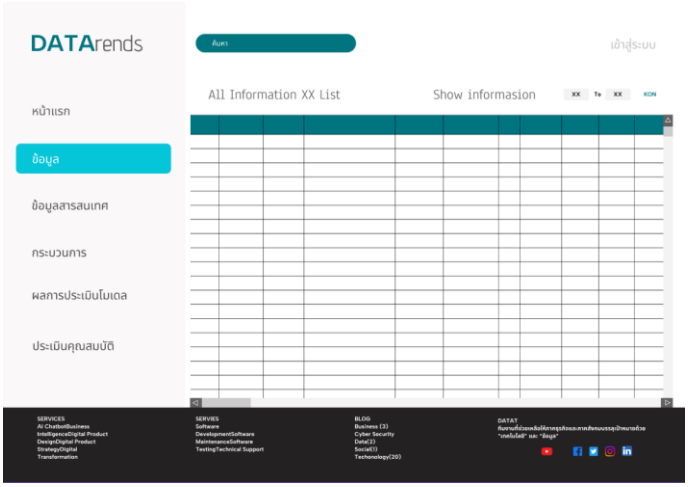
3.2 การออกแบบเว็บไซต์

3.2.1 การออกแบบ Wireframe

1) หน้าโฮมเพจของเว็บไซต์ แสดงเมนูต่างๆ ของหน้าเว็บไซต์ ข่าวสาร และ Dashboard

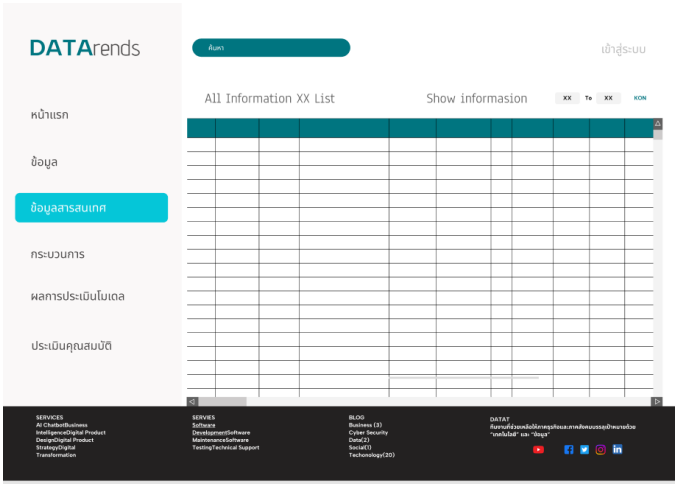
ภาพที่ 3.6 หน้าโฮมเพจของเว็บไซต์

2) หน้าเว็บเพจที่ 2 ชุดข้อมูล



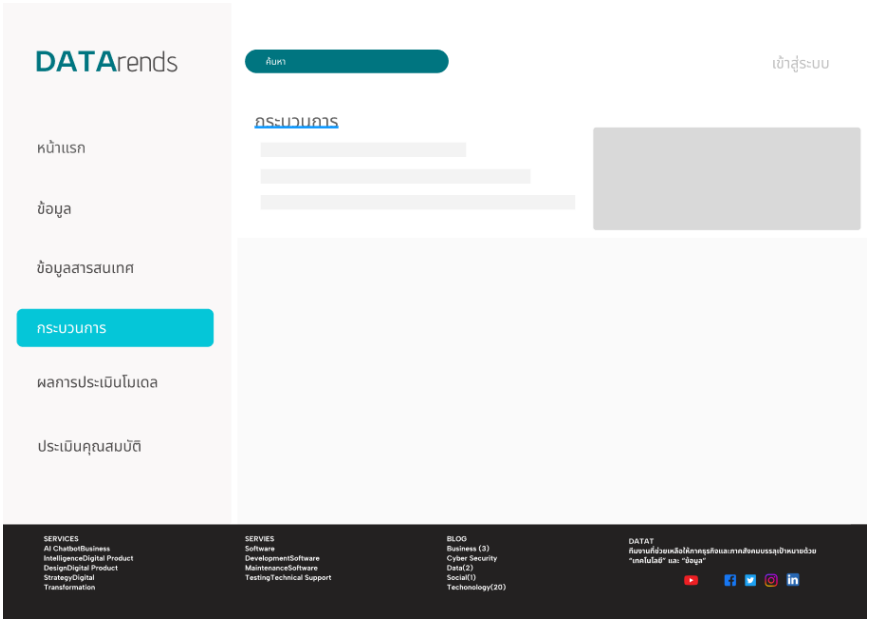
ภาพที่ 3.7 หน้าเว็บเพจที่ 2 ชุดข้อมูล

3) หน้าเว็บเพจที่ 3 ชุดข้อมูลสารสนเทศ



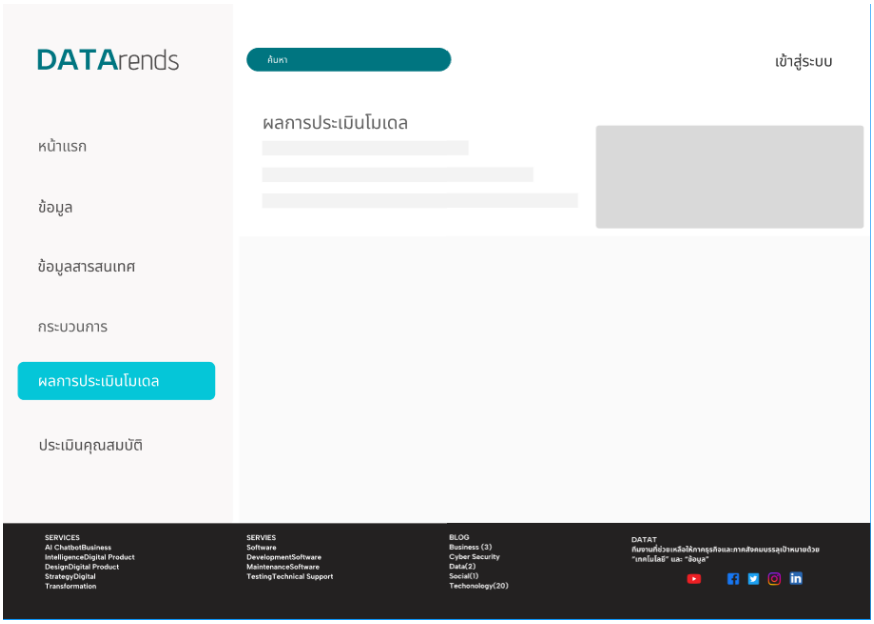
ภาพที่ 3.8 หน้าเว็บเพจที่ 3 ชุดข้อมูลสารสนเทศ

4) หน้าเว็บเพจที่ 4 Classification Model แสดงถึงขั้นตอนทำงาน



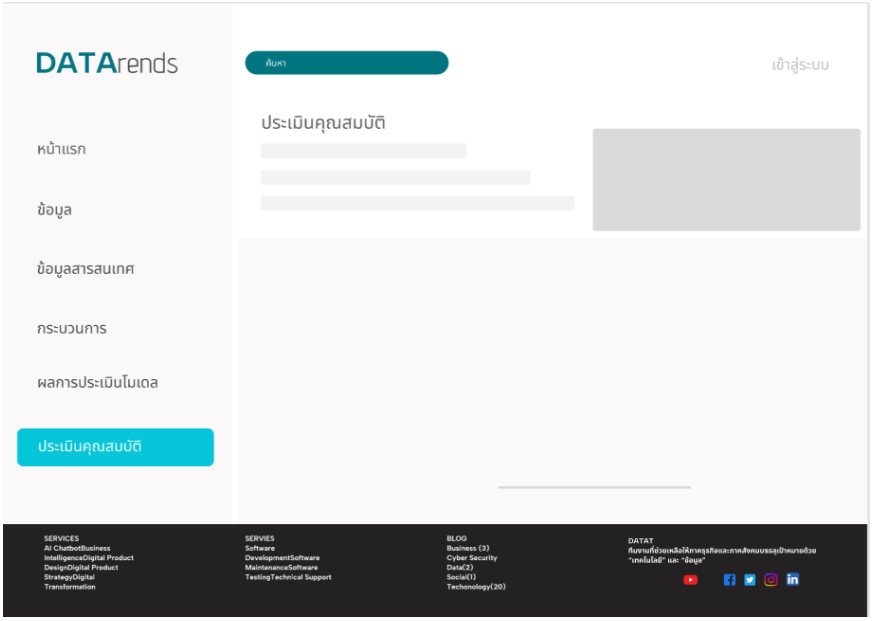
ภาพที่ 3.8 หน้าเว็บเพจที่ 4 Classification Model

5) หน้าเว็บเพจที่ 5 แสดงหน้าผลการประเมินโมเดล



ภาพที่ 3.9 หน้าเว็บเพจที่ 5 หน้าผลการประเมินโมเดล

6) หน้าเว็บเพจที่ 6 แสดงหน้าการประเมินคุณสมบัติ



ภาพที่ 3.10 หน้าเว็บเพจที่ 6 แสดงหน้าการประเมินคุณสมบัติ

3.3 บทสรุป

จากวิธีการดำเนินงานโครงงานทั้งหมด ผู้วิเคราะห์ข้อมูลได้พัฒนาแบบจำลองเพื่อวิเคราะห์ปัจจัยที่มีผลต่อเงินเดือนของ Data Scientist เช่น ระดับประสบการณ์ ระดับการศึกษา สถานที่ทำงาน ประเภทของบริษัท และทักษะเฉพาะด้าน โดยใช้เทคนิคการจำแนกประเภท (Classification) ได้แก่ Decision Tree และ Random Forest ตามขั้นตอนกระบวนการ CRISP-DM อย่างเป็นระบบ ผลลัพธ์จากการจำแนกช่วยให้สามารถมองเห็นปัจจัยที่มีอิทธิพลต่อการแบ่งกลุ่มช่วงเงินเดือน (ต่ำ / กลาง / สูง) ได้อย่างชัดเจน นอกจากนี้ ข้อมูลที่ได้ถูกนำมาจัดทำเป็น Visualization ในรูปแบบกราฟและตาราง รวมถึงออกแบบและเผยแพร่บนเว็บไซต์ เพื่อให้ผู้ใช้งานทั่วไปสามารถเข้าถึงและทำความเข้าใจข้อมูลได้อย่างสะดวก