บทที่ 3
ขั้นตอน และวิธีการดำเนินงาน

โครงงานเรื่อง การพัฒนาแบบจำลองการพยากรณ์แรงงานไทยในต่างประเทศด้วยเทคนิคเหมืองข้อมูล โดยเป็นการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล (Data Mining) และใช้กระบวนการวิเคราะห์ข้อมูลด้วยกระบวนการ CRISP-DM ที่สำคัญหลากหลายขั้นตอน เมื่อเสร็จสิ้นจากกระบวนการวิเคราะห์ข้อมูลแล้วจะเป็นขั้นตอนการสร้างโมเดลจำแนกประเภทข้อมูลในขั้นตอน Modeling ใช้เทคนิคดังต่อไปนี้ 1) Multiple Linear Regression 2) Gradient Boosted Trees (GBT) 3) Random Forest Regression และ 4) k-Nearest Neighbors (kNN) ซึ่งจะนำมาเปรียบเทียบความแม่นยำของแบบจำลองทั้งหมด โดยจะใช้ตัวชี้วัดมาตรฐาน ดังนี้ RMSE (Root Mean Squared Error) และ MAE (Mean Absolute Error) เป็นตัวเปรียบเทียบ จากนั้นทำการนำโมเดลที่ได้ไปทำการพัฒนาเว็บไซต์สำหรับการทำนายแสดงผล

3.1 ขั้นตอน และวิธีการดำเนินงานการวิเคราะห์ข้อมูลด้วย CRISP-DM

3.1.1 การทำความเข้าใจปัญหา (Problem Understanding)

3.1.2 การทำความเข้าใจข้อมูล (Data Understanding)

3.1.3 การเตรียมข้อมูล (Data Preparation)

3.1.4 การสร้างแบบจำลอง (Modeling)

3.1.5 การประเมินผล (Evaluation)

3.1.6 การนำไปใช้งาน (Deployment)

3.2 การวิเคราะห์และออกแบบระบบ

3.2.1 การวิเคราะห์และออกแบบฐานข้อมูล

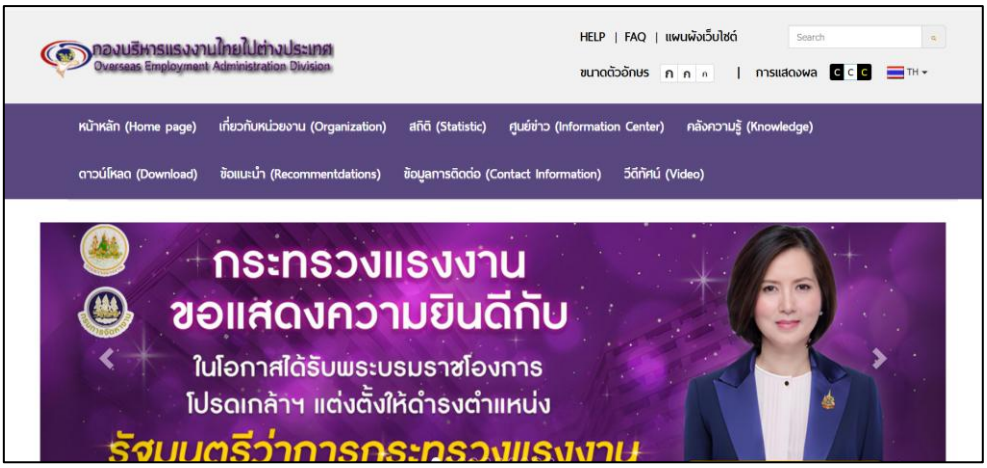
3.2.2 การออกแบบเว็บไซต์

3.1 ขั้นตอน และวิธีการดำเนินงานการวิเคราะห์ข้อมูลด้วย CRISP-DM

ขั้นตอนและรายละเอียดของ CRISP-DM กระบวนการ CRISP-DM ประกอบไปด้วย 6 ขั้นตอนหลัก ซึ่งแต่ละขั้นตอนมีบทบาทสำคัญต่อความสำเร็จของโครงการวิเคราะห์ข้อมูล ดังรายละเอียดต่อไปนี้

3.1.1 การทำความเข้าใจปัญหา (Problem Understanding) เป็นขั้นตอนแรก เริ่มด้วยการทำความเข้าใจปัญหาทางธุรกิจที่ต้องการแก้ไข ซึ่งจากการที่ได้ศึกษาในช่วงปี 2564–2567 จำนวนแรงงานไทยที่เดินทางไปทำงานต่างประเทศมีการเปลี่ยนแปลงขึ้นลงตลอด ขึ้นอยู่กับปัจจัยหลายอย่าง เช่น เพศ อายุ จังหวัดต้นทาง อาชีพ และประเทศที่ไป ซึ่งการศึกษาครั้งนี้มีวัตถุประสงค์เพื่อพยากรณ์ยอดรวม (Total) ของแรงงานไทยที่จะเดินทางไปทำงานต่างประเทศในอนาคต โดยมุ่งเน้นการเปรียบเทียบประสิทธิภาพของอัลกอริทึมต่าง ๆ เพื่อค้นหาโมเดลที่มีค่าความคลาดเคลื่อนน้อยที่สุด (Lowest Error Rate) ให้หน่วยงานที่เกี่ยวข้องสามารถนำผลพยากรณ์ไปวางแผนรองรับหรือบริหารจัดการทรัพยากรมนุษย์ได้อย่างเหมาะสม

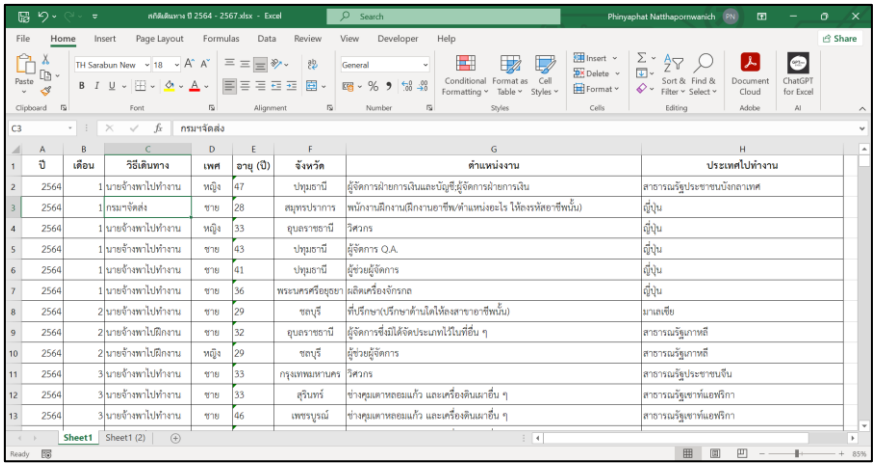
3.1.2 การทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนการจัดเก็บและรวบรวมข้อมูล ตลอดจนการพิจารณาตรวจสอบความถูกต้องของข้อมูลที่ได้รับ เริ่มต้นจากการรวบรวมข้อมูลที่เกี่ยวข้อทั้งหมด จากนั้น ทำการตรวจสอบคุณภาพของข้อมูล และวิเคราะห์เบื้องต้น โดยที่ว่าจะเลือกใช้ข้อมูลทั้งหมดหรือบางส่วนในการนำมาสิเคราะห์ให้สอดคล้องกับวัตถุประสงค์ที่กำหนดไว้และพิจารณาส่วนต่าง ๆ ผู้วิเคราะห์ได้ทำการรวบรวมข้อมูลที่จะนำมาใช้ในการพัฒนาแบบจำลองการพยากรณ์แรงงานไทยในต่างประเทศเป็นข้อมูลรายเดือนย้อนหลัง 48 เดือน ตั้งแต่ปี พ.ศ. 2564–2567 จำนวน 324,497 ระเบียน ที่ได้จากการขออนุเคราะห์ข้อมูลฝ่ายทะเบียน ฯ กองบริหารแรงงานไทยไปต่างประเทศ ซึ่งเป็นของกระทรวงแรงงานของไทยที่ได้ทำการส่งข้อมูลสถิติเกี่ยวกับการเดินทางไปทำงานต่างประเทศและพบปัญหาว่าข้อมูลเป็น Text เยอะ เช่น ชื่อจังหวัด, วิธีเดินทาง ข้อมูลยังไม่ถูกสรุปยอดรายเดือน (Aggregation) ทำให้ยังนำไปพยากรณ์ยอดรวมไม่ได้ทันทีมีตัวแปรที่อาจไม่จำเป็น (Noise) เช่น ชื่อจังหวัดรายบุคคล ซึ่งไม่มีผลต่อยอดรวมระดับประเทศ ดังนั้นเพื่อที่ผู้วิเคราะห์จะสามารถทำความเข้าใจกับข้อมูลเหล่านั้น และได้นำข้อมูลดังกล่าวมาทำการวิเคราะห์โดยผ่านเทคนิคการทำ Data Mining ต่อไป



ภาพที่ 3.1 แสดงเว็บไซต์กองบริหารแรงงานไทยไปต่างประเทศ

ที่มา https://www.doe.go.th/

รายละเอียดของข้อมูลที่นำมาใช้ในการพัฒนาแบบจำลองการพยากรณ์แรงงานไทยในต่างประเทศเป็นข้อมูลรายเดือนย้อนหลัง 48 เดือน ตั้งแต่ปี พ.ศ. 2564–2567 จำนวน 324,497 ระเบียน ประกอบด้วย 8 แอทริบิวต์ คือ ปี เดือน วิธีการเดินทาง เพศ อายุ จังหวัด ตำแหน่งงาน ประเทศไปทำงาน

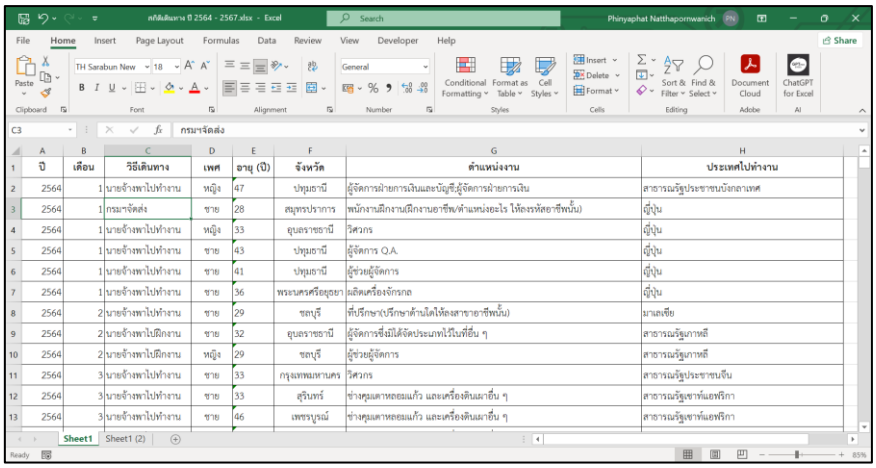


ปี	เดือน	วิธีเดินทาง	เพศ	อายุ (ปี)	จังหวัด	ตำแหน่งงาน	ประเทศไปทำงาน
2564	1	นายจ้างพาไปทำงาน	หญิง	47	ปทุมธานี	ผู้จัดการฝ่ายการเงินและบัญชี/ผู้จัดการฝ่ายการเงิน	สาธารณรัฐประชาชนบังกลาเทศ
2564	1	กรมาจัดส่ง	ชาย	28	สมุทรปราการ	พนักงานฝึกงาน(ฝึกงานอาชีพ/ตำแหน่งอะไร ให้ลงทัศอาชีพนั้น)	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	หญิง	33	อุบลราชธานี	วิศวกร	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	ชาย	43	ปทุมธานี	ผู้จัดการ Q.A.	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	ชาย	41	ปทุมธานี	ผู้ช่วยผู้จัดการ	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	ชาย	36	พระนครศรีอยุธยา	ผลิตเครื่องจักรกล	ญี่ปุ่น
2564	2	นายจ้างพาไปทำงาน	ชาย	29	ชลบุรี	ที่ปรึกษา(ปรึกษาด้านไดให้องสาขาอาชีพนั้น)	มาเลเซีย
2564	2	นายจ้างพาไปฝึกงาน	ชาย	32	อุบลราชธานี	ผู้จัดการซึ่งมิได้จัดประเภทไว้ในที่อื่น ๆ	สาธารณรัฐเกาหลี
2564	2	นายจ้างพาไปฝึกงาน	หญิง	29	ชลบุรี	ผู้ช่วยผู้จัดการ	สาธารณรัฐเกาหลี
2564	3	นายจ้างพาไปทำงาน	ชาย	33	กรุงเทพมหานคร	วิศวกร	สาธารณรัฐประชาชนจีน
2564	3	นายจ้างพาไปทำงาน	ชาย	33	สุรินทร์	ช่างศูนย์หล่อแก้ว และเครื่องดินเผาอื่น ๆ	สาธารณรัฐเชาท์แอฟริกา
2564	3	นายจ้างพาไปทำงาน	ชาย	46	เพชรบูรณ์	ช่างศูนย์หล่อแก้ว และเครื่องดินเผาอื่น ๆ	สาธารณรัฐเชาท์แอฟริกา

ภาพที่ 3.2 แสดงข้อมูลสถิติเดินทาง ปี 2564–2567

3.1.3 การเตรียมข้อมูล (Data Preparation) ขั้นตอนการแปลงข้อมูลที่ได้รวบรวมแล้วเลือกไว้ ให้อยู่ในรูปแบบที่พร้อมสำหรับการนำไปวิเคราะห์ในขั้นตอนต่อไป โดยการทำให้เป็นข้อมูลที่ถูกต้อง (Data Cleaning) โดยมีขั้นตอนดังนี้

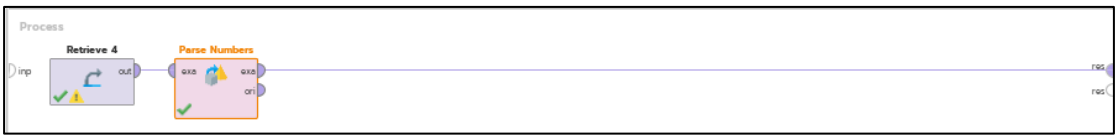
3.1.3.1 นำเข้าข้อมูลสำหรับวิเคราะห์ข้อมูลเป็นข้อมูลรายเดือนย้อนหลัง 48 เดือน ตั้งแต่ปี พ.ศ. 2564–2567 จำนวน 324,497 ระเบียน



ปี	เดือน	วิธีเดินทาง	เพศ	อายุ (ปี)	จังหวัด	ตำแหน่งงาน	ประเทศไปทำงาน
2564	1	นายจ้างพาไปทำงาน	หญิง	47	ปทุมธานี	ผู้จัดการฝ่ายการเงินและบัญชี,ผู้จัดการฝ่ายการเงิน	สาธารณรัฐประชาชนบังกลาเทศ
2564	1	กรมการจัดส่ง	ชาย	28	สมุทรปราการ	พนักงานฝึกงาน(ฝึกงานอาชีพ/ตำแหน่งอะไร ให้ลงรหัสอาชีพนั้น)	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	หญิง	33	อุบลราชธานี	วิศวกร	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	หญิง	43	ปทุมธานี	ผู้จัดการ Q.A.	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	ชาย	41	ปทุมธานี	ผู้ช่วยผู้จัดการ	ญี่ปุ่น
2564	1	นายจ้างพาไปทำงาน	ชาย	36	พระนครศรีอยุธยา	ผลิตเครื่องจักรกล	ญี่ปุ่น
2564	2	นายจ้างพาไปทำงาน	ชาย	29	ชลบุรี	ที่ปรึกษา(ปรึกษาด้านใดให้ลงสาขาอาชีพนั้น)	มาเลเซีย
2564	2	นายจ้างพาไปฝึกงาน	ชาย	32	อุบลราชธานี	ผู้จัดการซึ่งมิได้จัดประเภทไว้ในที่อื่น ๆ	สาธารณรัฐเกาหลี
2564	2	นายจ้างพาไปฝึกงาน	หญิง	29	ชลบุรี	ผู้ช่วยผู้จัดการ	สาธารณรัฐเกาหลี
2564	3	นายจ้างพาไปทำงาน	ชาย	33	กรุงเทพมหานคร	วิศวกร	สาธารณรัฐประชาชนจีน
2564	3	นายจ้างพาไปทำงาน	ชาย	33	สุรินทร์	ช่างศูนย์เหล่อแก้ว และเครื่องดินเผาอื่น ๆ	สาธารณรัฐเชาท์แอฟริกา
2564	3	นายจ้างพาไปทำงาน	ชาย	46	เพชรบูรณ์	ช่างศูนย์เหล่อแก้ว และเครื่องดินเผาอื่น ๆ	สาธารณรัฐเชาท์แอฟริกา

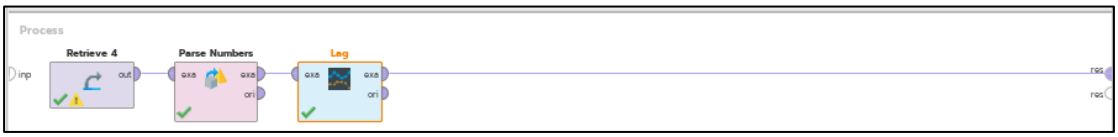
ภาพที่ 3.3 แสดงข้อมูลสถิติเดินทาง ปี 2564–2567

3.1.3.2 การแปลงชนิดข้อมูล (Data Type Conversion) ใช้ Operator Parse Numbers เพื่อแปลงข้อมูลตัวเลขที่อาจถูกอ่านเป็นข้อความ ให้กลับมาเป็นตัวเลข (Numerical) เพื่อให้สามารถคำนวณทางคณิตศาสตร์ได้



ภาพที่ 3.4 แสดง Operator Parse Numbers

3.1.3.3 สร้างตัวแปรใหม่ (Lag Features) สร้างข้อมูลย้อนหลัง (Lag) ใช้เทคนิค Windowing ผ่าน Operator Lag เพื่อสร้างตัวแปรทำนายจากอดีต ได้แก่ ยอดรวมย้อนหลัง 1 เดือน (Total–1), และยอดรายประเภทการเดินทางย้อนหลัง (เช่น Re–Entry–1)



ภาพที่ 3.5 Operator Lag

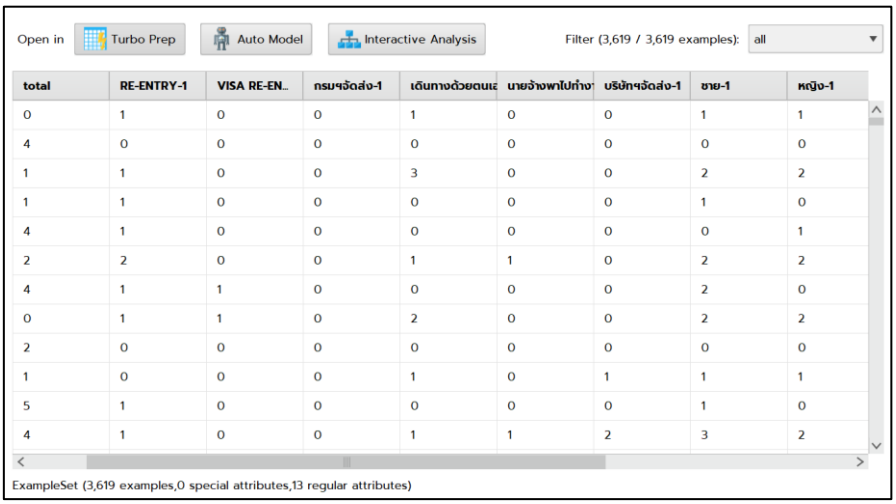
3.1.3.4 การคัดเลือกคุณลักษณะ (Feature Selection) เพื่อลดความซ้ำซ้อนของข้อมูล (Data Redundancy) และคัดกรองตัวแปรที่ส่งผลกระทบต่อประสิทธิภาพของแบบจำลอง โดยดำเนินการประเมินค่าน้ำหนักความสำคัญของตัวแปร (Feature Importance) ด้วยอัลกอริทึม Random Forest เพื่อพิจารณาคัดกรองตัวแปรกลุ่มย่อย ได้แก่ ข้อมูลจำแนกตามเพศ (ชาย/หญิง) ข้อมูลช่วงอายุ และวิธีการเดินทาง ที่มีลักษณะความซ้ำซ้อนกัน

(Multicollinearity) เนื่องจากข้อมูลเหล่านี้เป็นส่วนประกอบย่อยของตัวแปรเป้าหมาย (Total) ซึ่งหากนำเข้าสู่แบบจำลองทั้งหมด อาจทำให้แบบจำลองเกิดความสับสนและเรียนรู้ข้อมูลรบกวน (Noise) จึงเลือกเฉพาะตัวแปรที่มีค่าน้ำหนักความสำคัญสูงจึงช่วยเพิ่มประสิทธิภาพให้แบบจำลองได้ดี

ภาพที่ 3.6 แสดงการ (Feature Importance) ด้วยอัลกอริทึม Random Forest

3.1.3.5 การแบ่งชุดข้อมูลเพื่อใช้ในการสร้างและทดสอบประสิทธิภาพของแบบจำลอง ดำเนินการโดยใช้ Operator Split Data ด้วยวิธีการสุ่มแบบลำดับเวลา (Linear Sampling) เพื่อให้สอดคล้องกับลักษณะของข้อมูลที่เป็นอนุกรมเวลา (Time Series) โดยแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลสำหรับสอนโมเดล (Training Set) ร้อยละ 70 และข้อมูลสำหรับทดสอบ (Test Set) ร้อยละ 30 รวมถึงมีการนำแนวคิด Walk-forward Validation มาประยุกต์ใช้เพื่อการประเมินผลที่มีความแม่นยำตามลำดับเหตุการณ์จริง เหตุผลและแนวคิดทางทฤษฎี การกำหนดสัดส่วนในการแบ่งชุดข้อมูลมีความสำคัญต่อความสมดุลระหว่างการเรียนรู้และการประเมินผล หากกำหนดชุดฝึกสอนมากเกินไป (เช่น 95%) อาจทำให้ชุดทดสอบมีจำนวนไม่เพียงพอต่อการวัดผลที่น่าเชื่อถือ หรือมีความแปรปรวนสูง (High Variance) ในทางตรงกันข้าม หากกำหนดชุดทดสอบมากเกินไป (เช่น 50%) จะทำให้ข้อมูลสำหรับฝึกสอนมีน้อยเกินไป จนโมเดลไม่สามารถเรียนรู้รูปแบบที่ซับซ้อนของข้อมูลได้ ซึ่งอาจนำไปสู่ความเบี่ยงเบนสูง (High Bias) จากการศึกษาและงานวิจัยเชิงประจักษ์ในอดีต พบว่าสัดส่วนที่เหมาะสมและเป็นที่ยอมรับโดยทั่วไป (Rule of Thumb) คือการกำหนดชุดทดสอบให้อยู่ในช่วงร้อยละ 20–30 ของข้อมูลทั้งหมด โดยงานวิจัยที่อ้างอิงเรื่อง “Why 70/30 or 80/20 Relation Between Training and Testing Sets” และ “Optimal Ratio for Data Splitting” ระบุว่าสัดส่วน 70:30 หรือ 80:20 เป็นอัตราส่วนที่มีความสมดุลเหมาะสมสำหรับงานวิจัยส่วนใหญ่ ทั้งนี้ การเลือกใช้อัตราส่วน 70:30 ในการศึกษานี้ พิจารณาจากขนาดของชุดข้อมูลและวัตถุประสงค์เพื่อต้องการให้โมเดลมีข้อมูลเพียงพอสำหรับการเรียนรู้แพทเทิร์นของฤดูกาล (Seasonality) ในขณะเดียวกันก็ยังมีข้อมูลช่วงเวลาล่าสุดที่มากพอสำหรับการทดสอบความแม่นยำ ตัวอย่างการคำนวณการแบ่งข้อมูล สมมติฐานชุดข้อมูลแรงงานทั้งหมดมีจำนวน N ระเบียน (Records) การคำนวณสัดส่วนจะเป็นดังนี้ : ชุดข้อมูลทั้งหมด (Total Records): 100,000 ระเบียนข้อมูลสำหรับสอนโมเดล (Training Set 70%): คำนวณได้จาก $0.70 \times 100,000

= 70,000$ ระเบียน ข้อมูลสำหรับทดสอบ (Test Set 30%) : คำนวณได้จาก $0.30 \times 100,000 = 30,000$ ระเบียน

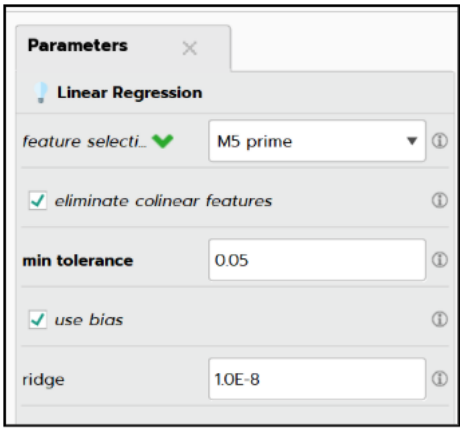


total	RE-ENTRY-1	VISA RE-EN..	กรมฯจัดส่ง-1	เดินทางด้วยตนเอ	นายจ้างพาไปทำง	บริษัทฯจัดส่ง-1	ชาย-1	หญิง-1
0	1	0	0	1	0	0	1	1
4	0	0	0	0	0	0	0	0
1	1	0	0	3	0	0	2	2
1	1	0	0	0	0	0	1	0
4	1	0	0	0	0	0	0	1
2	2	0	0	1	1	0	2	2
4	1	1	0	0	0	0	2	0
0	1	1	0	2	0	0	2	2
2	0	0	0	0	0	0	0	0
1	0	0	0	1	0	1	1	1
5	1	0	0	0	0	0	1	0
4	1	0	0	1	1	2	3	2

ภาพที่ 3.7 แสดงตัวอย่างชุดข้อมูลที่ผ่านการเตรียมข้อมูล (Data Preparation)

3.1.4 การสร้างแบบจำลอง (Modeling) เมื่อได้ข้อมูลที่พร้อมแล้ว จากขั้นตอนที่ 3.1.3 จึงนำเข้าสู่กระบวนการสร้างโมเดล โดยแบ่งข้อมูลเป็นชุดสอน (Training 70%) และชุดทดสอบ (Testing 30%) และเลือกใช้ 4 อัลกอริทึมเพื่อเปรียบเทียบ

3.1.4.1 Multiple Linear Regression (MLR) ผู้ศึกษาได้พัฒนาโดยแบ่งข้อมูลเป็น Training Set ร้อยละ 70 และ Testing Set ร้อยละ 30 จากนั้นได้เลือกใช้วิธีคัดเลือกตัวแปรแบบ M5 Prime เพื่อช่วยลดตัวแปรที่ไม่จำเป็นและลดความซ้ำซ้อนของข้อมูล ตามแนวคิดการลดมิติข้อมูลในงานเหมืองข้อมูล (Quinlan, 1992) พร้อมทั้งกำจัดตัวแปรที่มีความสัมพันธ์กันสูงและกำหนดค่า minimum tolerance เท่ากับ 0.05 เพื่อป้องกันปัญหา Multicollinearity ซึ่งเป็นข้อสมมติฐานสำคัญของการวิเคราะห์การถดถอยเชิงเส้น (Gujarati, 2003) เลือกใช้ use bias เพื่อให้สมการมีค่าคงที่ตามหลักของสมการถดถอย และกำหนดค่า ridge เท่ากับ 1.0E-8 เพื่อช่วยให้การคำนวณมีความเสถียร

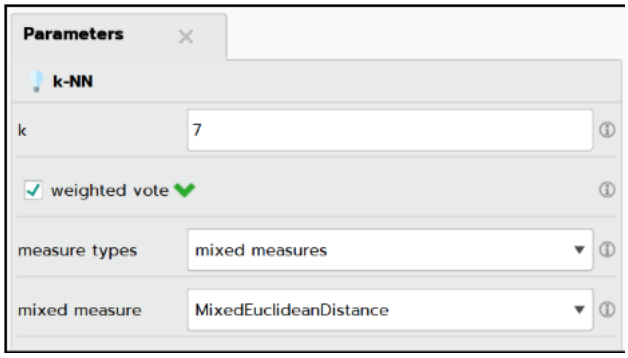


ภาพที่ 3.8 แสดงการปรับค่า Parameter โมเดล Multiple Linear Regression (MLR)

3.1.4.2 Gradient Boosted Trees (GBT) ใช้วิธีสร้างต้นไม้ตัดสินใจ (Decision Tree) ผู้ศึกษาได้พัฒนาโมเดลด้วยโปรแกรม AI Studio โดยกำหนดจำนวนต้นไม้ (number of trees) เท่ากับ 400 เพื่อเพิ่มความสามารถในการเรียนรู้รูปแบบข้อมูลที่ซับซ้อน พร้อมกำหนดค่า learning rate เท่ากับ 0.01 เพื่อให้การปรับแก้ค่าความผิดพลาดของแต่ละต้นไม้เป็นไปอย่างค่อยเป็นค่อยไป ซึ่งช่วยลดความเสี่ยงในการเกิด Overfitting ตามหลักการของ Boosting (Friedman, 2001) นอกจากนี้ได้กำหนดค่า maximal depth เท่ากับ 20 เพื่อควบคุมความลึกของต้นไม้ไม่ให้ซับซ้อนจนเกินไป และกำหนดค่า min rows เท่ากับ 20 เพื่อป้องกันการแตกกิ่งจากข้อมูลจำนวนน้อยเกินไป รวมถึงกำหนดค่า sample rate เท่ากับ 0.8 เพื่อสุ่มข้อมูลบางส่วนในแต่ละรอบการเรียนรู้ ซึ่งช่วยเพิ่มความเสถียรของแบบจำลอง ทั้งนี้ได้ตั้งค่า random seed เพื่อให้ผลลัพธ์สามารถทำซ้ำได้ (reproducible) และช่วยเพิ่มความน่าเชื่อถือของการทดลอง

3.1.4.3 Random Forest Regression ในการสร้างแบบจำลอง Random Forest ผู้ศึกษาได้พัฒนาโมเดลด้วยโปรแกรม AI Studio โดยกำหนดจำนวนต้นไม้ (number of trees) เท่ากับ 500 เพื่อเพิ่มความเสถียรของผลลัพธ์ เนื่องจาก Random Forest อาศัยหลักการรวมผลจากต้นไม้หลายต้น (Ensemble Learning) ซึ่งยิ่งมีจำนวนต้นไม้มาก ผลลัพธ์จะมีความนิ่งมากขึ้น (Breiman, 2001) โดยเลือกเกณฑ์การแบ่งข้อมูลเป็นแบบ least square ให้เหมาะสมกับงานพยากรณ์เชิงปริมาณ และกำหนดค่า maximal depth เท่ากับ 20 เพื่อควบคุมความซับซ้อนของต้นไม้ไม่ให้ลึกเกินไป เปิดใช้การตัดแต่งกิ่งล่วงหน้า (apply prepruning) พร้อมกำหนดค่า minimal gain เท่ากับ 0.01 และ minimal size for split เท่ากับ 4 เพื่อป้องกันการแตกกิ่งจากข้อมูลจำนวนน้อยเกินไป ซึ่งช่วยลดความเสี่ยงในการเกิด Overfitting

3.1.4.4 k–Nearest Neighbors (k–NN) ใช้วิธีนำข้อมูลในปัจจุบันไปเปรียบเทียบกับข้อมูลในอดีต เพื่อค้นหาช่วงเวลาที่มีลักษณะเหตุการณ์คล้ายคลึงกันที่สุด (เช่น ยอดเดือนนี้คล้ายกับเดือนเดียวกันของปีก่อน) แล้วนำตัวเลขของช่วงเวลานั้นมาใช้เป็นคำตอบในการพยากรณ์ ผู้ศึกษาได้กำหนดค่า k เท่ากับ 7 เพื่อให้การพยากรณ์อ้างอิงจากข้อมูลใกล้เคียงจำนวนพอเหมาะ ไม่มากหรือน้อยเกินไป ช่วยลดความไวต่อค่าผิดปกติของข้อมูล ตามหลักการของ k–NN (Cover & Hart, 1967) พร้อมทั้งเลือกใช้วิธี weighted vote เพื่อให้น้ำหนักกับข้อมูลที่อยู่ใกล้มากกว่าข้อมูลที่อยู่ไกล และกำหนดการวัดระยะทางแบบ Mixed Euclidean Distance เนื่องจากข้อมูลมีทั้งตัวแปรเชิงปริมาณและเชิงหมวดหมู่ การกำหนดระยะทางเพื่อช่วยให้การคำนวณความใกล้เคียงมีความเหมาะสมกับลักษณะข้อมูล



ภาพที่ 3.9 แสดงการปรับค่า Parameter โมเดล k–NN

3.1.5 การประเมินผล (Evaluation) เป็นกระบวนการสำคัญในการประเมินคุณภาพของกฎความสัมพันธ์ที่สร้างขึ้นจากทั้ง 4 โมเดล เพื่อเลือกโมเดลที่ดีที่สุดสำหรับการนำไปวิเคราะห์และเปรียบเทียบผลลัพธ์ ผู้จัดทำได้ใช้โปรแกรมของ RapidMiner ในการวัดประสิทธิภาพของโมเดล โดยพิจารณาตัวชี้วัดหลัก ได้แก่ ค่า

3.1.5.1 RMSE (Root Mean Squared Error) วัดค่าความคลาดเคลื่อนโดยให้โทษหนักกับ error ที่สูงผิดปกติ

$$\boldsymbol{RMSE} = \frac{\sqrt{\sum_{i=1}^{n}(Y - \hat{Y})^2}}{n}$$

3.1.5.2 MAE (Mean Absolute Error) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย ไม่ยกกำลังสอง ไม่เน้นลงโทษความผิดพลาดใหญ่เหมือน MSE/RMSE เหมาะเมื่อข้อมูลมี outliers ที่ไม่อยากให้มีผลมากเกินไป

$$MAE = \frac{1}{n}\sum_{t-1}^{n}|y_t - \hat{y}_t|$$

ตารางที่ 3.1 ตารางเปรียบเทียบประสิทธิภาพระหว่างโมเดล

ตัวแบบ	RMSE	MAE
Multiple Linear Regression		
Gradient Boosted Trees (GBT)		
Random Forest Regression		
k–Nearest Neighbors (kNN)		

3.1.6 การนำไปใช้งาน (Deployment) ในการนำโมเดลที่ผ่านการพัฒนาและเปรียบเทียบมาใช้งานจริง มีการดำเนินการดังต่อไปนี้ดังนี้เลือกตัวแบบที่เหมาะสมจากการวิเคราะห์ผ่านโปรแกรม RapidMiner ได้ทำการเปรียบเทียบค่าความคลาดเคลื่อน (Error) ของตัวแบบที่มีค่าความคลาดเคลื่อนต่ำที่สุดโดยใช้เกณฑ์ RMSE (Root Mean Squared Error), MAE (Mean Absolute Error)

3.1.6.1 การพยากรณ์และประยุกต์ใช้งาน ตัวแบบที่ได้รับการเลือกจะถูกนำมาใช้ในการพยากรณ์ข้อมูลในอนาคต โดยอาศัยข้อมูลจากชุดข้อมูลใหม่ เพื่อสร้างผลลัพธ์ที่มีความแม่นยำและสอดคล้องกับเป้าหมายของงานที่กำหนดไว้

3.1.6.2 นำเสนอผลลัพธ์ผ่านเว็บไซต์ ผลลัพธ์จากการพยากรณ์ด้วยตัวแบบจะถูกนำมาแสดงในรูปแบบกราฟต่าง ๆ เพื่อแสดงผลผ่านบนเว็บไซต์ โดยใช้ ภาษา HTML CSS

3.2 การวิเคราะห์และออกแบบระบบ

การวิเคราะห์และออกแบบฐานข้อมูล สำหรับระบบวิเคราะห์ข้อมูลและการพัฒนาแบบจำลองการพยากรณ์แรงงานไทยในต่างประเทศด้วยเทคนิคเหมืองข้อมูลโดยอธิบายการวิเคราะห์และออกแบบเว็บไซต์ดังนี้ ฟังก์ชั่นจะประกอบไปด้วย

3.2.1 แผนภาพบริบท (Context Diagram)

3.2.1.1 ผู้ใช้งานระบบ

1) ผู้ใช้งานทั่วไป

2) ผู้ดูแลระบบ

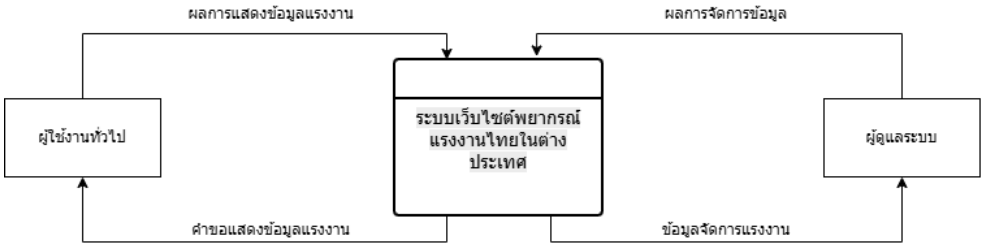
3.2.1.2 ความต้องการในระบบ

1) ผู้ดูแลระบบ

- สามารถเข้าสู่ระบบได้ด้วยชื่อระหัสผ่านผู้ใช้งาน
- สามารถ เพิ่ม ลบ แก้ไขข้อมูลได้
- สามารถเปลี่ยนรหัสผ่านผู้ใช้งานได้

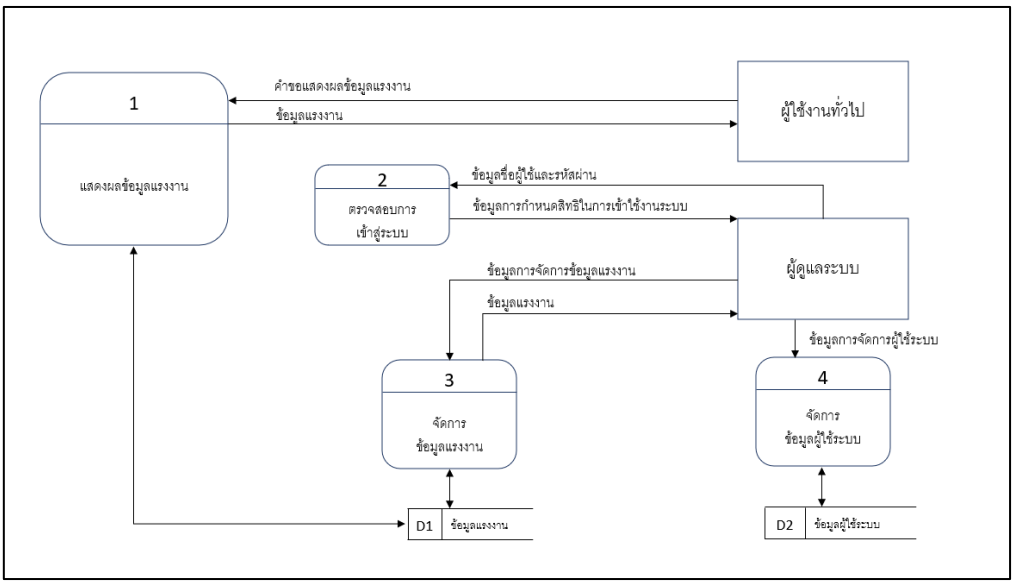
2) ผู้ใช้งานทั่วไป

- สามารถเข้าชมข้อมูลได้



ภาพที่ 3.10 แผนภาพบริบท (Context Diagram)

3.2.2 แผนภาพกระแสข้อมูลระดับที่ 0 (Data Flow Diagram Level 0 : DFD Level 0)



ภาพที่ 3.11 แผนภาพกระแสข้อมูล

ตารางที่ 3.2 คำอธิบายการแสดงข้อมูลแรงงาน

Process Description	
System	ระบบเว็บไซต์การพัฒนาแรงงานไทย
DFD Number	1
Process Name	แสดงผลข้อมูลแรงงาน
Input Data Flow	คำขอแสดงผลข้อมูลแรงงาน (จากผู้ใช้งานทั่วไป) , ข้อมูลแรงงานจาก (D1)
Output Data Flow	ข้อมูลแรงงาน (ไปยังผู้ใช้งานทั่วไป)
Data Store Used	D1 ข้อมูลแรงงาน
Description	เป็นกระบวนการที่ผู้ใช้ร้องขอข้อมูลแรงงาน และระบบดึงข้อมูลจากฐานข้อมูลมาแสดงผลผ่านแดชบอร์ด

ตารางที่ 3.3 อธิบายกระบวนการเข้าสู่ระบบ

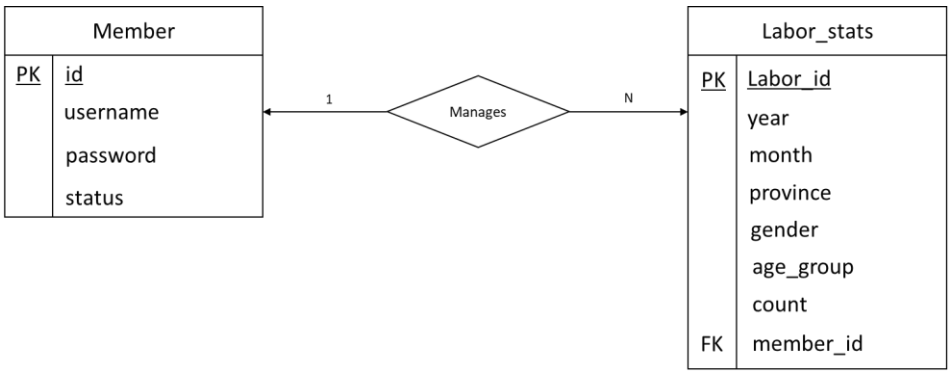
Process Description	
System	ระบบเว็บไซต์การพัฒนาแรงงานไทย
DFD Number	2
Process Name	ตรวจสอบการเข้าสู่ระบบ
Input Data Flow	ข้อมูลชื่อผู้ใช้และรหัสผ่าน (จากผู้ดูแลระบบ)
Output Data Flow	ข้อมูลการกำหนดสิทธิในการเข้าใช้งานระบบ (ไปยังผู้ดูแลระบบ)
Data Store Used	D2 ข้อมูลผู้ใช้ระบบ
Description	เป็นกระบวนการตรวจสอบความถูกต้องของสิทธิการใช้งาน เพื่อให้ผู้ดูแลระบบสามารถเข้าถึงเครื่องมือการจัดการได้ตามสิทธิ

ตารางที่ 3.4 คำอธิบายกระบวนการจัดการข้อมูลแรงงาน

Process Description	
System	ระบบเว็บไซต์การพัฒนาแรงงานไทย
DFD Number	3
Process Name	จัดการข้อมูลแรงงาน
Input Data Flow	ข้อมูลการจัดการข้อมูลแรงงาน (จากผู้ดูแลระบบ)
Output Data Flow	ข้อมูลแรงงาน (บันทึกใน D1 และแสดงผลกลับไปยังผู้ดูแลระบบ)
Data Store Used	D1 ข้อมูลแรงงาน
Description	เป็นกระบวนการที่ผู้ดูแลระบบทำการ เพิ่ม ลบ แก้ไข ข้อมูลสถิติแรงงานไทยในต่างประเทศและบันทึกลงในฐานข้อมูลแรงงาน

ตารางที่ 3.5 คำอธิบายกระบวนการจัดการข้อมูลผู้ใช้ระบบ

Process Description	
System	ระบบเว็บไซต์การพัฒนาแรงงานไทย
DFD Number	4
Process Name	จัดการข้อมูลผู้ใช้ระบบ
Input Data Flow	ข้อมูลการจัดการผู้ใช้ระบบ (จากผู้ใช้ระบบ)
Output Data Flow	ข้อมูลผู้ใช้ระบบ (บันทึกลง D2)
Data Store Used	D2 ข้อมูลผู้ใช้ระบบ
Description	เป็นกระบวนการที่ผู้ดูแลระบบจัดการข้อมูลเกี่ยวกับบัญชีผู้ใช้งานระบบ เช่น การเพิ่ม ลบ แก้ไข ข้อมูลของผู้ดูแลระบบคนอื่น ๆ



ภาพที่ 3.12 แผนภาพ ERD

ตารางที่ 3.6 ตารางอธิบาย Entity : Member

ชื่อฟิลด์	ชนิดข้อมูล	คีย์	รายละเอียด
id	INT	PK	รหัสผู้ดูแลระบบ ใช้ระบุข้อมูลผู้ใช้แต่ละคน
Username	VARCHAR	–	ชื่อผู้ใช้สำหรับเข้าสู่ระบบ
Password	VARCHAR	–	รหัสผ่านสำหรับเข้าสู่ระบบ
status	VARCHAR	–	สถานะหรือสิทธิ์ของผู้ใช้งาน เช่น admin

ตาราง Member ใช้สำหรับจัดเก็บข้อมูลผู้ดูแลระบบที่สามารถเข้าสู่ระบบและจัดการข้อมูลภายในเว็บไซต์ได้

ตารางที่ 3.7 ตารางอธิบาย Entity : Labor_stats

ชื่อฟิลด์	ชนิดข้อมูล	คีย์	รายละเอียด
labor_id	INT	PK	รหัสข้อมูลแรงงาน ใช้ระบุข้อมูลสถิติแรงงานแต่ละรายการ
Year	INT	–	ปีที่บันทึกข้อมูลแรงงาน
Mounth	INT	–	เดือนที่บันทึกข้อมูลแรงงาน
Province	VARCHAR	–	จังหวัดของแรงงาน
Gender	VARCHAR	–	เพศของแรงงาน
Age_group	VARCHAR	–	ช่วงอายุของแรงงาน เช่น 15–25 ปี
Count	INT	–	จำนวนแรงงานในกลุ่มข้อมูลนั้น

Member_id	INT	FK	รหัสผู้ดูแลระบบที่ทำการเพิ่มหรือแก้ไขข้อมูล

ตาราง Labor_stats ใช้สำหรับจัดเก็บข้อมูลสถิติแรงงานไทยในต่างประเทศ โดยข้อมูลจะถูกบันทึกตามปี เดือน จังหวัด เพศ และช่วงอายุ

ตารางที่ 3.8 ตารางความสัมพันธ์ (Relationship)

Entity ที่ 1	**ความสัมพันธ์**	**Entity ที่ 2**	**ประเภทความสำคัญ**	**คำอธิบาย**
Member	Manages	Labor_stats	1:N	ผู้ดูแลระบบ 1 คน สามารถจัดการข้อมูลแรงงานได้หลายรายการ

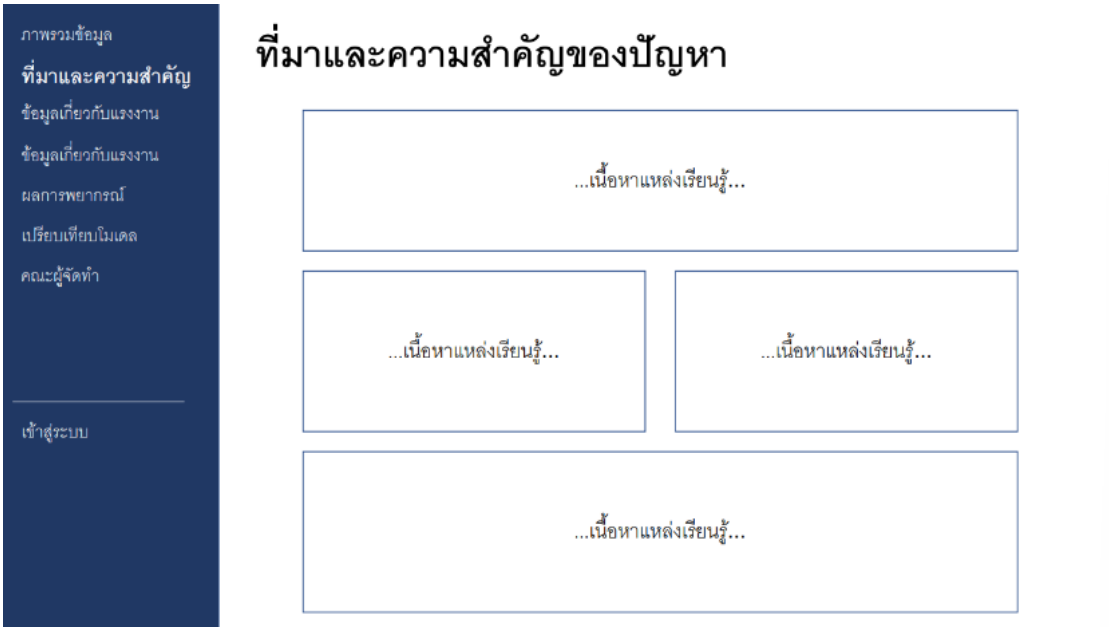
จากแผนภาพ ERD ระบบเว็บไซต์พยากรณ์แรงงานไทยในต่างประเทศประกอบด้วยเอนทิตีหลัก 2 เอนทิตี ได้แก่ Member และ Labor_stats โดยเอนทิตี Member ใช้จัดเก็บข้อมูลผู้ดูแลระบบ และเอนทิตี Labor_stats ใช้จัดเก็บข้อมูลสถิติแรงงาน ซึ่งทั้งสองเอนทิตีมีความสัมพันธ์กันแบบหนึ่งต่อหลาย (1:N) โดยผู้ดูแลระบบหนึ่งคนสามารถจัดการข้อมูลแรงงานได้หลายรายการ

3.2.3 การออกแบบเว็บไซต์สำหรับการเผยแพร่ข้อมูลจำนวนประชากรแรงงานไทยที่เดินทางไปทำงานในต่างประเทศออกแบบมาเพื่อแสดงข้อมูลของแรงงานไทยที่เดินทางออกไปทำงานในต่างประเทศ

1.หน้าแรกของเว็บไซต์แสดงข้อมูลต่าง ๆ ของหน้าเว็บ

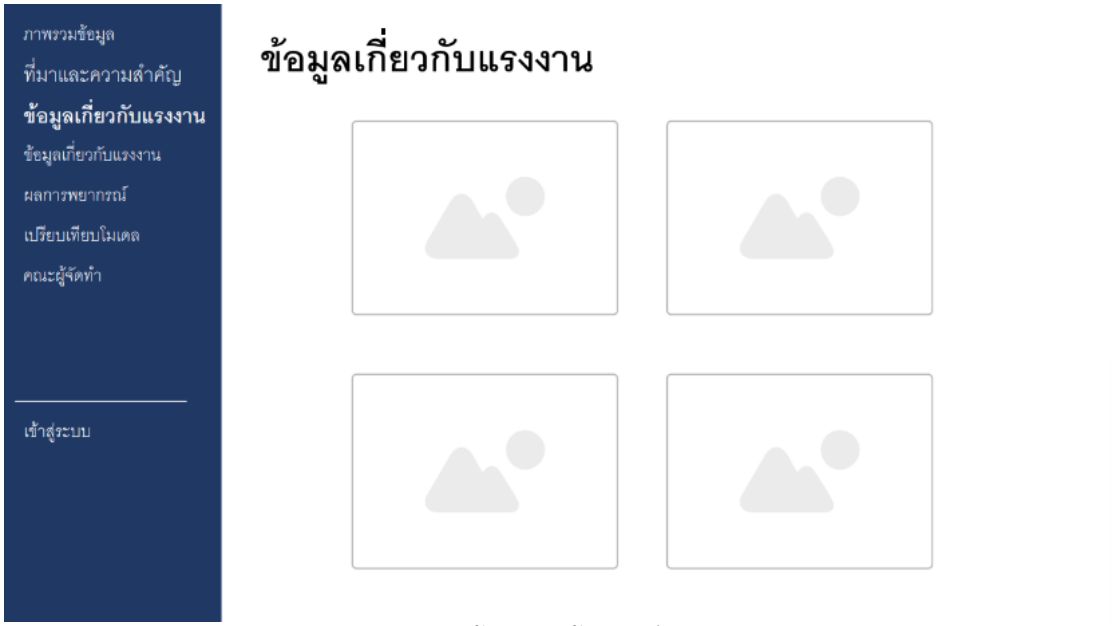
ภาพที่ 3.13 แสดงโครงร่างหน้าแรกของเว็บไซต์

2. หน้าแสดงข้อมูลที่มาและความสำคัญของปัญหา



ภาพที่ 3.14 แสดงข้อมูลที่มาและความสำคัญของปัญหา

3. หน้าแสดงข้อมูลเกี่ยวกับแรงงาน

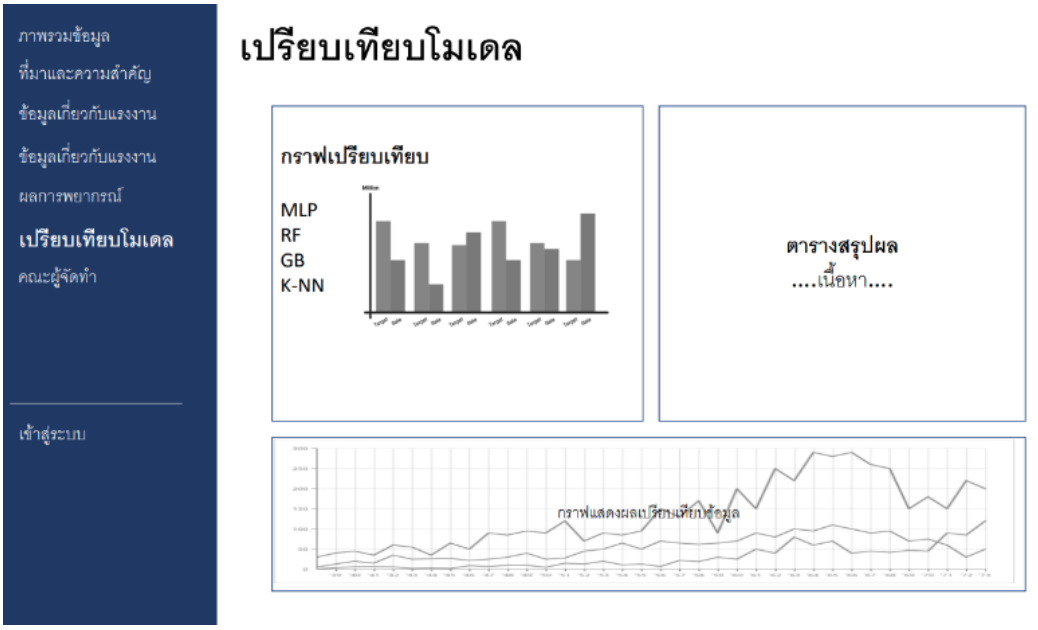


ภาพที่ 3.15 หน้าแสดงข้อมูลเกี่ยวกับแรงงาน

4. หน้าแสดงผลการพยากรณ์

ภาพที่ 3.16 หน้าแสดงผลการพยากรณ์

5. เปรียบเทียบโมเดล

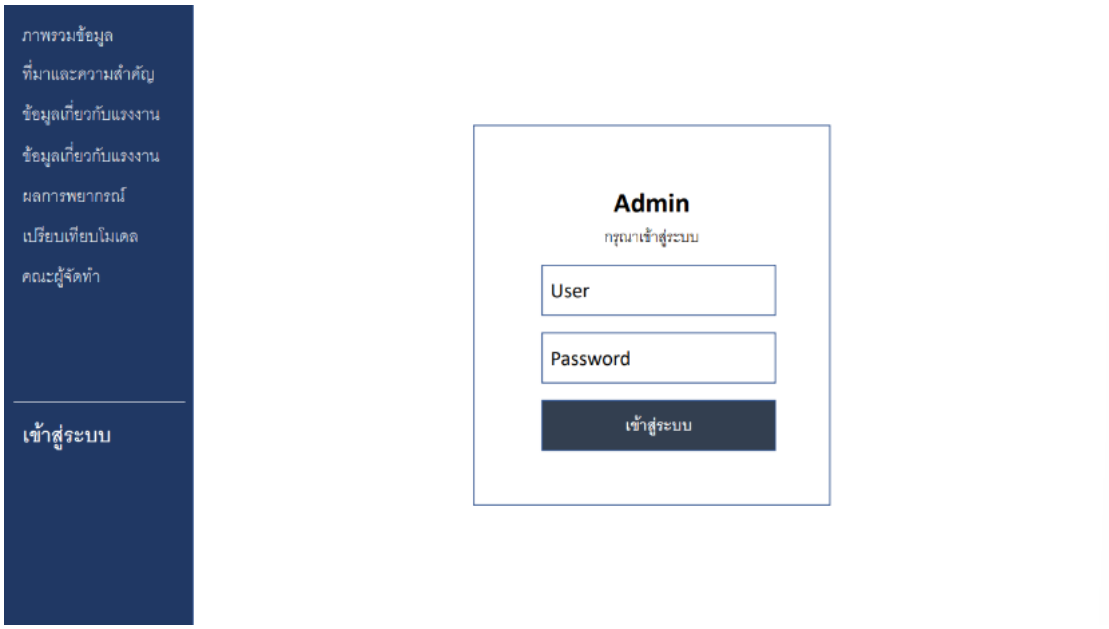


ภาพที่ 3.17 เปรียบเทียบโมเดล

6. คณะผู้จัดทำ

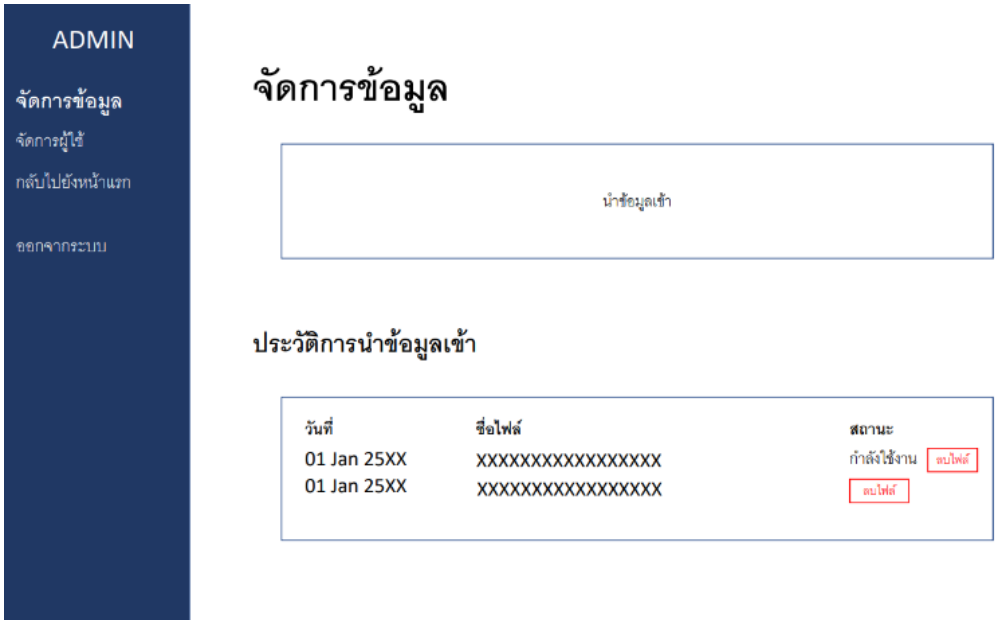
ภาพที่ 3.18 คณะผู้จัดทำ

8. ผู้ดูแลระบบ (Admin)



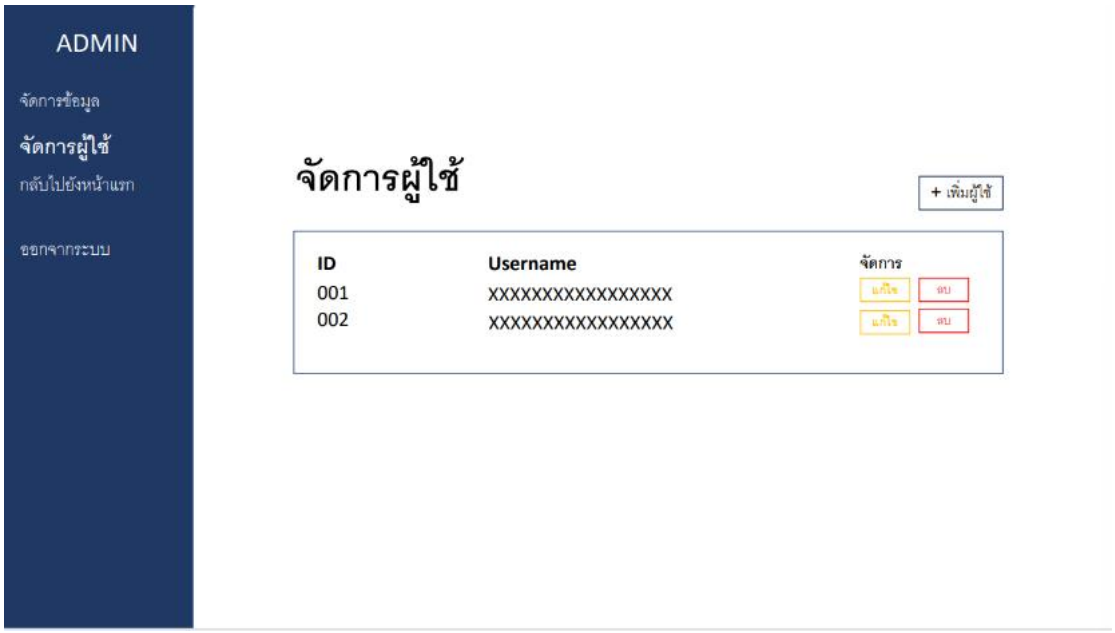
ภาพที่ 3.19 ผู้ดูแลระบบ (Admin)

9.จัดการข้อมูล



ภาพที่ 3.20 จัดการข้อมูล

10. จัดการผู้ใช้



ภาพที่ 3.21 จัดการผู้ใช้