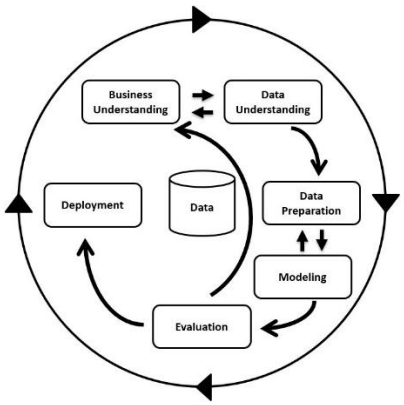
บทที่ 3
วิธีดำเนินการงานวิจัย

3.1 กระบวนการ CRISP–DM (Cross Industry Standard Process for Data Mining)

CRISP–DM เป็นกระบวนการหลักในการจัดทำเหมืองข้อมูลเพื่อการวิเคราะห์และใช้ประโยชน์ในทางธุรกิจ กระบวนการวิเคราะห์ข้อมูล ด้วย CRISP–DM หรือ (Cross Industry Standard Process for Data Mining) ประกอบด้วย 6 ขั้นตอน แต่ละขั้นตอนในรูปจะเป็นขั้นตอนที่ต่อเนื่องกันนั่นคือขั้นตอนถัดไปจะรอผลลัพธ์จากขั้นตอนก่อนหน้าซึ่งแสดงด้วยลูกศรที่เชื่อมระหว่างกล่องสี่เหลี่ยมแต่ละกล่อง ตัวอย่างเช่นเมื่อ ได้ผลลัพธ์จากขั้นตอนการเตรียมข้อมูล (Data Preparation) แล้วจะนำไปสร้างโมเดลจำแนกประเภทข้อมูลในขั้น Modeling และหลังจากนั้นอาจจะย้อนกลับมาเปลี่ยนแปลงข้อมูลให้ถูกต้องมากขึ้นเพื่อหวังว่าจะโมเดลที่ให้ความถูกต้องมากขึ้นก็ได้ เป็นต้น ในกระบวนการนี้ประกอบด้วย 6 ขั้นตอน ดังนี้



ภาพที่ 3.1 กระบวนการ CRISP–DM

3.1.1 ความเข้าใจในธุรกิจ (Business Understanding)

การพัฒนาบุคลากรสายวิชาการถือเป็นหัวใจสำคัญในการขับเคลื่อนสถาบันอุดมศึกษาของรัฐให้ก้าวสู่ความเป็นเลิศ เนื่องจากตำแหน่งทางวิชาการ ไม่ว่าจะเป็นผู้ช่วยศาสตราจารย์ (ผศ.) รองศาสตราจารย์ (รศ.) และอาจารย์ (อ.) ล้วนเป็นดัชนีชี้วัดศักยภาพ คุณภาพ และความก้าวหน้าของทั้งตัวบุคลากรและมหาวิทยาลัยในภาพรวม การที่บุคลากรมีตำแหน่งทางวิชาการที่สูงขึ้นไม่เพียงสะท้อนถึงความเชี่ยวชาญที่ได้รับการยอมรับ

สถาบัน อย่างไรก็ตาม มหาวิทยาลัยกรณีศึกษากำลังเผชิญกับความท้าทายจากจำนวนการขอตำแหน่งทางวิชาการของบุคลากรที่อยู่ในระดับน้อย ซึ่งเป็นปัญหาที่จำเป็นต้องได้รับการวิเคราะห์อย่างเป็นระบบ เพื่อทำความเข้าใจถึงปัจจัยที่นำไปสู่ความสำเร็จและอุปสรรคที่เกิดขึ้น โดยปัจจุบันยังขาดการวิเคราะห์เชิงข้อมูลที่ชัดเจนในประเด็นดังกล่าว โครงการวิจัยนี้จึงมีเป้าหมายในมุมมองขององค์กร เพื่อช่วยให้มหาวิทยาลัยสามารถเข้าใจถึงคุณลักษณะและปัจจัยร่วมของบุคลากรที่ประสบความสำเร็จในการขอตำแหน่งทางวิชาการได้อย่างลึกซึ้ง ซึ่งองค์ความรู้ที่ได้จะนำไปสู่การกำหนดนโยบายและกลยุทธ์ในการส่งเสริม สนับสนุน และวางแผนพัฒนาเส้นทางความก้าวหน้าทางวิชาการของบุคลากรได้อย่างตรงจุดและมีประสิทธิภาพยิ่งขึ้น

สำหรับเป้าหมายในเชิงการวิเคราะห์ข้อมูล (Data Mining Goal) โครงการนี้มุ่งเน้นการค้นหาความสัมพันธ์และปัจจัยสำคัญที่ซ่อนอยู่ในชุดข้อมูลบุคลากร โดยจะใช้เทคนิคการวิเคราะห์ความสัมพันธ์ (Correlation) เพื่อคัดเลือกตัวแปรที่สำคัญ เช่น ปีที่ปฏิบัติงาน ระดับการศึกษา และหน่วยงานที่สังกัด จากนั้นจะนำตัวแปรที่คัดเลือกมาสร้างแบบจำลองเพื่อจำแนกประเภท (Classification) บุคลากรที่มีศักยภาพและโอกาสสูงในการขอตำแหน่งทางวิชาการในอนาคต และอาจใช้เทคนิคการจัดกลุ่ม (Clustering) เพื่อแบ่งกลุ่มบุคลากรตามคุณสมบัติร่วม ซึ่งจะช่วยให้ผู้บริหารสามารถมองเห็นภาพรวมและระบุกลุ่มเป้าหมายในการส่งเสริมได้อย่างชัดเจน

เพื่อให้การดำเนินงานเป็นไปอย่างมีแบบแผนและบรรลุเป้าหมายที่วางไว้ โครงการวิจัยนี้จะดำเนินงานภายใต้กรอบกระบวนการ CRISP–DM (Cross–Industry Standard Process for Data Mining) โดยเริ่มต้นจากขั้นตอนการทำความเข้าใจในธุรกิจ (Business Understanding) ซึ่งคือเนื้อหาในส่วนนี้ จากนั้นจะเข้าสู่ขั้นตอนการทำความเข้าใจข้อมูล (Data Understanding) และการเตรียมข้อมูล (Data Preparation) ที่ได้รับจากมหาวิทยาลัยกรณีศึกษา ก่อนจะนำไปสู่การสร้างแบบจำลอง (Modeling) และการประเมินผล (Evaluation) เพื่อให้ได้โมเดลที่มีความแม่นยำและน่าเชื่อถือสูงสุด ท้ายที่สุด ผลลัพธ์และข้อมูลเชิงลึกที่ได้จากการวิเคราะห์จะถูกนำเสนอในรูปแบบของข้อเสนอแนะเชิงนโยบาย เพื่อสนับสนุนการตัดสินใจของผู้บริหารในการพัฒนาบุคลากรต่อไป

3.1.2 ขั้นตอนการเตรียมข้อมูล(Data Preparation)

1) การรวบรวมข้อมูลเบื้องต้น (Initial Data Collection)

ข้อมูลที่ใช้ในการวิเคราะห์นี้ได้รับความอนุเคราะห์จาก **หน่วยงานวิทยบริการของมหาวิทยาลัยกรณีศึกษา** หลังจากได้ดำเนินการยื่นเอกสารขอความอนุเคราะห์ข้อมูลอย่างเป็นทางการ โดยข้อมูลที่ได้รับมาอยู่ในรูปแบบไฟล์ ข้อมูลบุคลากรจากระบบ HR ณ 25 ต.ค. 67.xlsx จากการตรวจสอบอย่างละเอียด ข้อมูลชุดนี้ประกอบด้วยรายการบุคลากรจำนวนหนึ่ง และมีคุณลักษณะ (Attributes) ทั้งหมด 17 คอลัมน์ 1028 เรคคอร์ด ซึ่งครอบคลุมข้อมูลที่จำเป็นต่อการวิเคราะห์ปัจจัยในการขอตำแหน่งทางวิชาการ

วันที่ 28 เดือน ตุลาคม พ.ศ 2567

เรื่อง ขอความอนุเคราะห์เก็บข้อมูลเพื่อประกอบการทำโครงงาน

เรียน ผู้อำนวยการวิทยบริการ

ด้วยมหาวิทยาลัยได้มอบหมายให้ข้าพเจ้า ผู้ช่วยศาสตราจารย์ ดร.วรรณพร ทีเก่ง กำกับดูแลงานด้านพัฒนาบุคลากรฯ และส่งเสริมการขอตำแหน่งทางวิชาการของบุคลากรสายวิชาการ และในภาคเรียนที่ 2/2567 ได้เป็นอาจารย์ที่ปรึกษาโครงงานของนักศึกษาระดับปริญญาตรี สังกัดหลักสูตรระบบสารสนเทศทางธุรกิจ กลุ่มวิเคราะห์ข้อมูล รหัส 66 เรื่อง “การวิเคราะห์ปัจจัยที่มีผลต่อการขอตำแหน่งทางวิชาการของบุคลากร” ของนักศึกษา 2 ราย ได้แก่

1. นายกฤษณ์ แซ่ฟ่าน
2. นายพงศสรณ์ สุเดชา

เพื่อส่งเสริมการเรียนการสอนด้วยการวิเคราะห์ข้อมูลจริง จึงขอความอนุเคราะห์ข้อมูลพื้นฐานของบุคลากร มทร.ล้านนา เช่น ปีที่ทำงาน ปีที่เกิด ระดับการศึกษา ตำแหน่งทางวิชาการ สถาบันการศึกษาที่สำเร็จ หน่วยงานที่สังกัด โดยปกปิดข้อมูลส่วนบุคคลตาม พ.ร.บ.คุ้มครองข้อมูลส่วนบุคคล (PDPA) พร้อมแนบแบบข้อเสนอโครงงาน (FM-BIS-01) มาพร้อมนี้

จึงเรียนมาเพื่อโปรดพิจารณาอนุเคราะห์ และขอขอบพระคุณเป็นอย่างสูง ณ โอกาสนี้

(ลงชื่อ).....
(ผู้ช่วยศาสตราจารย์ ดร.วรรณพร ทีเก่ง)
อาจารย์ประจำหลักสูตรระบบสารสนเทศทางธุรกิจ

(ลงชื่อ).....
(นายกฤษณ์ แซ่ฟ่าน)
นักศึกษาหลักสูตรระบบสารสนเทศทางธุรกิจ

(ลงชื่อ).....
(นายพงศ์สรณ์ สุเดชา)

ภาพที่ 3.2 หนังสือขอความอนุเคราะห์

ตารางที่ 3.1 คุณลักษณะทั้ง 17 คอลัมน์และรายละเอียดชุดข้อมูล

ชื่อคุณลักษณะ (Attribute Name)	**คำอธิบาย (Description)**	**ประเภทข้อมูล (Data Type)**
ลำดับ	ลำดับที่ของรายการข้อมูล	String
เขตพื้นที่	ที่ตั้งของหน่วยงานที่บุคลากรสังกัด	Categorical
คณะ	คณะที่บุคลากรสังกัด	Categorical
สาขา	สาขาย่อยภายใต้คณะที่บุคลากรสังกัด	Categorical
สาขาวิชา	สาขาวิชาที่บุคลากรสำเร็จการศึกษา	String
อายุ	อายุของบุคลากร (ปี)	Numeric
ปีเกิด	ปี พ.ศ. ที่เกิดของบุคลากร	Numeric
อายุงานปัจจุบัน	ระยะเวลาการทำงาน (ปี)	Numeric
ประเภทบุคลากร	ประเภทการจ้างงาน เช่น ข้าราชการ, พนักงานฯ	Categorical
ประเภทสายงาน	สายงานหลักของบุคลากร เช่น สายวิชาการ, สายสนับสนุน	Categorical
ตำแหน่งงานปัจจุบัน	ตำแหน่งปัจจุบัน เช่น อาจารย์, ผู้ช่วยศาสตราจารย์	Categorical
ประเภทตำแหน่ง/กลุ่ม	กลุ่มของตำแหน่งงาน	Categorical
จำนวนผลงานอดีต	จำนวนผลงานตีพิมพ์ทั้งหมดที่ผ่านมา	Numeric
TCI	จำนวนผลงานที่ตีพิมพ์ในฐานข้อมูล TCI	Numeric
ISI	จำนวนผลงานที่ตีพิมพ์ในฐานข้อมูล ISI/Scopus	Numeric

ตารางที่ 3.1 คุณลักษณะทั้ง 17 คอลัมน์และรายละเอียดชุดข้อมูล (ต่อ)

ชื่อคุณลักษณะ (Attribute Name)	**คำอธิบาย (Description)**	**ประเภทข้อมูล (Data Type)**
ความพร้อม	ตัวแปรเป้าหมายระบุความพร้อมในการขอตำแหน่ง (Y/N)	Categorical
ระดับคุณวุฒิ	ระดับการศึกษาสูงสุดของบุคลากร	Categorical

2) จากขั้นตอนการรวบรวมข้อมูลเบื้องต้น พบว่าข้อมูลมีความไม่สมบูรณ์ด้าน Label ซึ่งไม่สามารถนำไปสร้างโมเดลได้โดยตรง ผู้จัดทำจึงได้กำหนด เกณฑ์ในการสร้าง Label ขึ้นมา โดยใช้สูตรการคำนวณดังนี้

โดยที่

$$\text{คะแนนรวม} = \text{จำนวนผลงานอดีต} + (TCI + \ (ISI\ X\ 2)) \qquad (5)$$

เกณฑ์การกำหนด Label

- ถ้าคะแนนรวม ≥ 3 คะแนน ⟶ กำหนดเป็น Label = Y (พร้อม)
- ถ้าคะแนนรวม < 3 คะแนน ⟶ กำหนดเป็น Label = N (ไม่พร้อม)

จากนั้นผู้จัดทำได้ทำการดำเนินการคัดเลือกคุณลักษณะ (Attribute Selection) โดยตัดคุณลักษณะที่ไม่เกี่ยวข้องออก เหลือไว้เพียงคุณลักษณะที่สัมพันธ์กับเกณฑ์การพิจารณาของ ก.พ.อ. รวมทั้งหมด 11 Attribute ได้แก่ ISI TCI ตำแหน่งงานปัจจุบัน จำนวนผลงานอดีต อายุงานปัจจุบัน ระดับคุณวุฒิ เขตพื้นที่ อายุ คณะ ประเภทบุคลากร และ ความพร้อม

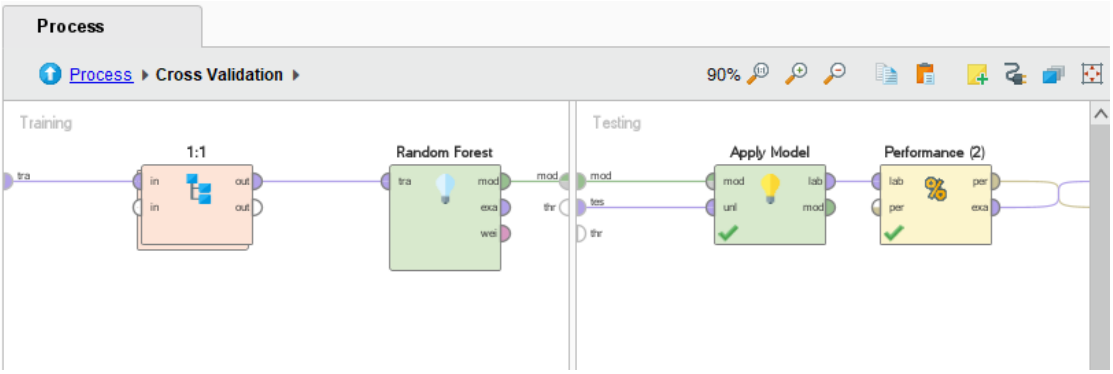
ตารางที่ 3.2 ข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูลสมบูรณ์

ชื่อแอตทริบิวต์	**รายละเอียด**
ISI	จำนวนผลงานที่ตีพิมพ์ในฐาน ISI/Scopus จำนวนผลงานที่ผลิตในปี 2565–2566
TCI	จำนวนผลงานที่อยู่ในฐาน Thai Citation Index จำนวนผลงานที่ผลิตในปี 2565–2566
ตำแหน่งงานปัจจุบัน	ระดับตำแหน่งสายวิชาการของบุคคล เช่น ผู้ช่วยศาสตราจารย์, รองศาสตราจารย์, อาจารย์ ฯลฯ

ตารางที่ 3.2 ข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูลสมบูรณ์ (ต่อ)

ชื่อแอตทริบิวต์	รายละเอียด
อายุงานปัจจุบัน	จำนวนปีที่อยู่ในตำแหน่ง/หน่วยงานปัจจุบัน
ระดับคุณวุฒิ	วุฒิการศึกษาสูงสุด
เขตพื้นที่	เขตพื้นที่ทางภูมิศาสตร์หรือเขตการปกครองที่สังกัดของบุคลากร เช่น ภาคกลาง ภาคเหนือ ฯลฯ
อายุ	อายุของบุคคล (หน่วยปี)
คณะ	คณะ/ส่วนงานที่สังกัด
ประเภทบุคลากร	ประเภทการจ้างงาน/สายงาน
จำนวนผลงานอดีต	จำนวนผลงานวิจัยหรือวิชาการที่เคยทำก่อนปี 2565

3) หลังจากขั้นตอนการเตรียมข้อมูลเรียบร้อยแล้ว พบว่าข้อมูลมีความไม่สมดุล กล่าวคือ จำนวนบุคลากรที่ไม่มีความพร้อมมีมากกว่าบุคลากรที่มีความพร้อม ความไม่สมดุลนี้อาจส่งผลให้การวิเคราะห์ขาดประสิทธิภาพ ดังนั้นคณะผู้วิจัยจึงแก้ไขปัญหาด้วยวิธีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย (SMOTE: Synthetic Minority Over-sampling Technique) เพื่อปรับสมดุลของข้อมูล และทำให้จำนวนข้อมูลของคลาสรอง (Minority class) มีจำนวนใกล้เคียงกับคลาสหลัก (Majority class) ก่อนนำข้อมูลเข้าสู่กระบวนการวิเคราะห์ต่อไป



ภาพที่ 3.3 Processการปรับสมดุลข้อมูล

การเพิ่มจำนวนข้อมูลตัวอย่างด้วยวิธี SMOTE (Synthetic Minority Over-sampling Technique) เป็นเทคนิคที่ได้รับความนิยมในการแก้ไขปัญหาความไม่สมดุลของข้อมูล (Data Imbalance) โดยเฉพาะอย่างยิ่งในกรณีที่จำนวนตัวอย่างของกลุ่มข้อมูลชนกลุ่มน้อย (Minority Class) มีไม่เพียงพอ ซึ่งอาจส่งผลให้โมเดลการเรียนรู้ของเครื่องขาดความแม่นยำในการจำแนก

กลุ่มดังกล่าว ตัวอย่างเช่น การจำแนกโรครากเน่าโคนเน่าของต้นทุเรียนด้วยอัลกอริทึม Decision Tree พบว่าเมื่อมีการสร้างความสมดุลของข้อมูลด้วย SMOTE แล้ว ประสิทธิภาพการจำแนกเพิ่มขึ้นจากเดิม 97.33% เป็น 99.33% (Sanitphonklang, 2023)

นอกจากนี้ เพื่อเพิ่มความน่าเชื่อถือของการแก้ไขปัญหาข้อมูลไม่สมดุล คณะผู้วิจัยยังได้ใช้ SMOTE แบบ Hybrid method ในอัตตราส่วน 1:3 ของ1026 เรคคอร์ด ซึ่งเป็นการผสมผสานระหว่าง การสุ่มเพิ่มข้อมูลสังเคราะห์ในกลุ่มส่วนน้อย และ การลดจำนวนข้อมูลในกลุ่มที่มีมากเกินไป (Under–sampling) ทำให้สามารถปรับสมดุลของข้อมูลได้ทั้งสองด้าน กล่าวคือ ไม่เพียงแต่เพิ่มข้อมูลของกลุ่มที่มีน้อย แต่ยังช่วยลดความซ้ำซ้อนของข้อมูลในกลุ่มที่มีมาก วิธีการแบบ Hybrid จึงช่วยเพิ่มคุณภาพของชุดข้อมูลให้เหมาะสมกับการนำไปสร้างแบบจำลอง ตลอดจนช่วยลดความลำเอียงของโมเดล และส่งเสริมให้ผลการวิเคราะห์มีความแม่นยำมากยิ่งขึ้น

ตารางที่ 3.3 การเปรียบเทียบจำนวนข้อมูลก่อนและหลังการทำการปรับสมดุลข้อมูล 1:1

กลุ่ม	**ข้อมูลก่อนการทำ SMOTE**	**ข้อมูลหลังการทำ SMOTE**
พร้อม	949	513
ไม่พร้อม	77	513
จำนวนทั้งหมด	1026	1026

3.1.3 การทำความสะอาดข้อมูล (Data Cleaning)

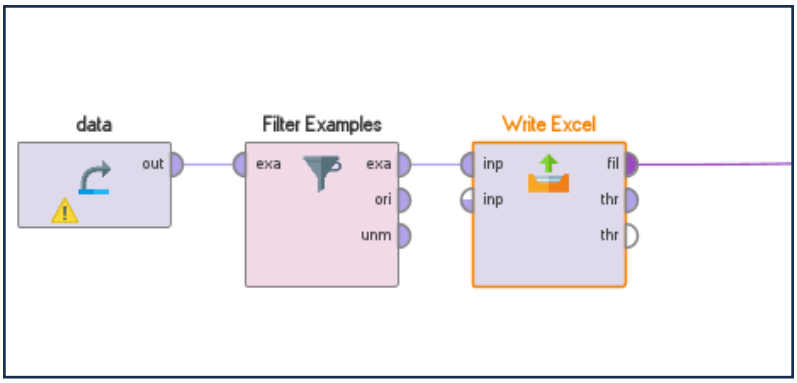
1) การจัดการข้อมูลที่ขาดหาย (Handling Missing Data) เพื่อจัดการกับค่าว่างในชุดข้อมูล ซึ่งอาจส่งผลกระทบต่อการวิเคราะห์ทางสถิติ ผู้วิจัยได้ทำการสำรวจข้อมูลในคอลัมน์ที่เกี่ยวข้องกับผลงานตีพิมพ์ (TCI, ISI_Scopus) และพบว่ามีค่าว่าง (Null values) จำนวนมาก จากการพิจารณาบริบทของข้อมูล สันนิษฐานได้ว่าค่าว่างดังกล่าวหมายถึง "การไม่มีผลงานตีพิมพ์" ดังนั้น ผู้วิจัยจึงได้ทำการแทนที่ค่าว่างทั้งหมดในคอลัมน์เหล่านี้ด้วยค่า 0 เพื่อให้ข้อมูลอยู่ในรูปแบบตัวเลขที่สมบูรณ์และสามารถนำไปใช้ในการคำนวณได้

2) ในการเตรียมข้อมูลเพื่อการวิเคราะห์ พบว่าคอลัมน์ ตำแหน่งงานปัจจุบัน มีค่าที่ไม่สอดคล้องกับขอบเขตของการศึกษา เช่น “รองอธิบดี” ซึ่งไม่ใช่ตำแหน่งทางวิชาการ คณะผู้วิจัยจึงใช้วิธี **Filter Example** เพื่อลบข้อมูลดังกล่าวออก ผลการทำความ

สะอาดทำให้ชุดข้อมูลมีความเหมาะสมต่อการวิเคราะห์ และเหลือจำนวนทั้งสิ้น 1,026 รายการ สำหรับการวิจัยต่อไป

จำนวนผลงานอดีต	จำนวนผลงาน65_66	TCI	ISI_Scopus
1	0	0	0
1	0	0	0
2	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
2	0	0	0
1	0	0	0
2	0	0	0
1	0	0	0
1	1	1	0
2	1	1	0
1	1	1	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
2	1	1	0
1	0	0	0

ภาพที่ 3.4 การแทนที่ค่าว่างทั้งหมด



ภาพที่ 3.5 Process การ Filter Example

3.1.4 ขั้นตอนการคัดเลือกคุณลักษณะข้อมูล (Feature Selection)

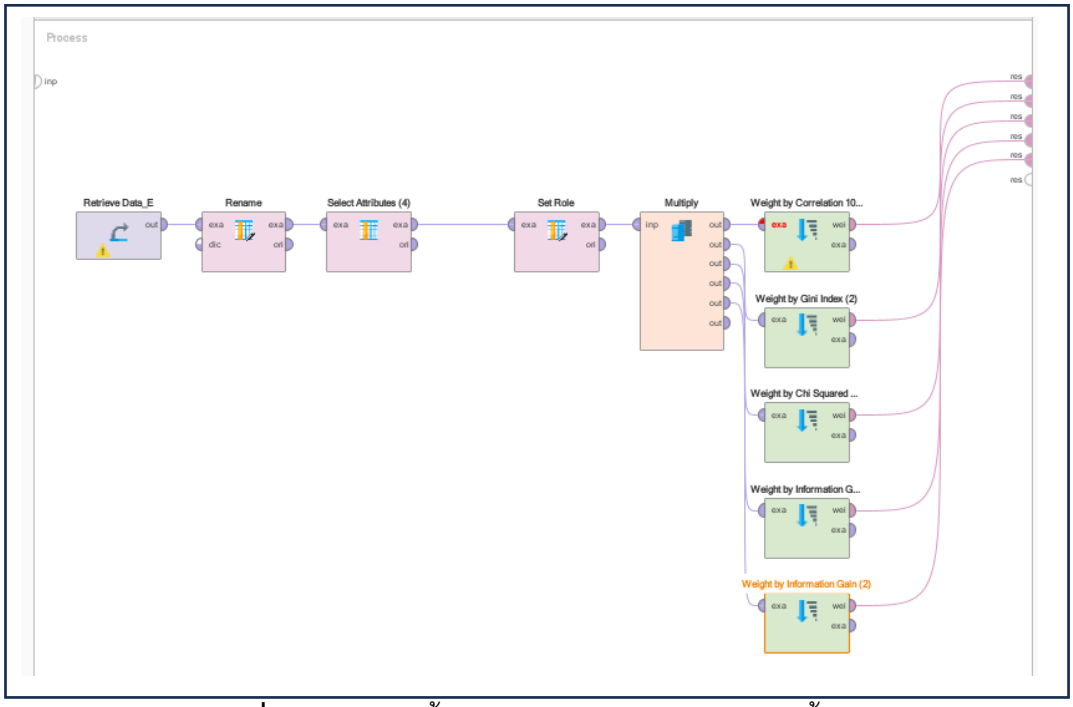
ขั้นตอนการคัดสรรลักษณะเฉพาะ โดยใช้แนวทางเชิงตัวกรองในการประเมินความสำคัญ โดยใช้เทคนิคการคำนวณหาค่าน้ำหนัก (Filter approach) ได้แก่ 1) Chi-Square Statistic 2) Gini Index 3) Gain Ratio 4) Information Gain และ 5) Correlation based เป็นการเลือกคุณลักษณะแบบการคัดกรอง (Filter-Base Feature Selection) จะใช้เงื่อนไขทางสถิติในการคัดกรอง โดยค่าความสัมพันธ์ระหว่างคุณลักษณะกับคลาสคำตอบเพื่อจัดลำดับคุณลักษณะตามค่านัยสำคัญทางสถิติ

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (6)$$

โดยที่

- Oi คือ ความถี่ที่สังเกตได้ (Observed frequency) สำหรับแต่ละหมวดหมู่
- Ei คือ ความถี่ที่คาดการณ์ไว้ (Expected frequency) สำหรับแต่ละหมวดหมู่
- n คือ จำนวนหมวดหมู่

ค่าสถิติ Chi-Square ใช้ในการทดสอบความสัมพันธ์ระหว่างตัวแปรเชิงหมวดหมู่สองตัว หรือเพื่อทดสอบว่าการกระจายของข้อมูลสังเกตการณ์แตกต่างจากการกระจายที่คาดหวังหรือไม่ โดยจะวัดความแตกต่างระหว่างความถี่ที่สังเกตได้ (Observed Frequencies) และความถี่ที่คาดการณ์ไว้ (Expected Frequencies)



ภาพที่ 3.6 Processขั้นตอนการคัดเลือกคุณลักษณะข้อมูล

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{7}$$

โดยที่

- D คือ ชุดข้อมูล
- m คือ จำนวนคลาส (categories) ที่เป็นไปได้ในชุดข้อมูล
- pi คือ สัดส่วนของเรคคอร์ดที่อยู่ในคลาส i ในชุดข้อมูล D

Gini Index หรือ Gini Impurity เป็นการวัดความไม่บริสุทธิ์ (impurity) หรือความไม่เป็นเนื้อเดียวกันของชุดข้อมูล มักใช้ใน Decision Tree Algorithms (เช่น CART) เพื่อช่วยในการเลือกคุณสมบัติที่ดีที่สุดในการแบ่งข้อมูล ยิ่งค่า Gini Index ต่ำเท่าไหร่ แสดงว่าข้อมูลยิ่งมีความบริสุทธิ์หรือเป็นเนื้อเดียวกันมากขึ้น

$$\text{Gain Ratio}(A) = \frac{\text{Information Gain}(A)}{\text{Split Information}(A)} \tag{8}$$

โดยที่

- A คือ คุณสมบัติ (attribute) ที่ใช้ในการแบ่ง
- Information Gain(A) คือ ค่า Information Gain ของคุณสมบัติ A (ดูหัวข้อถัดไป)
- Split Information(A) คือ ค่าที่วัดความกว้างของการกระจายของข้อมูลเมื่อถูกแบ่งด้วยคุณสมบัติ A คำนวณโดย

Gain Ratio เป็นการปรับปรุงจาก Information Gain เพื่อแก้ไขปัญหาที่ Information Gain มีแนวโน้มที่จะเลือกคุณสมบัติที่มีจำนวนค่าที่แตกต่างกันมากๆ (ซึ่งอาจทำให้เกิด Overfitting) Gain Ratio จะพิจารณา "Intrinsic Information" ของการแบ่งด้วยคุณสมบัตินั้น ๆ ด้วย ทำให้การเลือกคุณสมบัติมีความสมดุลมากขึ้น

$$\text{Information Gain}(D, A) = \text{Entropy}(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} \text{Entropy}(D_v) \tag{9}$$

โดยที่

- D คือ ชุดข้อมูลทั้งหมด
- A คือ คุณสมบัติ (attribute) ที่ต้องการคำนวณ Information Gain
- Values(A) คือ เซตของค่าที่เป็นไปได้ของคุณสมบัติ A

- Dv คือ ชุดข้อมูลย่อยของ D ที่มีคุณสมบัติ A เป็นค่า v
- |Dv| คือ จำนวนข้อมูลใน Dv
- |D| คือ จำนวนข้อมูลทั้งหมดใน D
- Entropy(S) คือ ค่าเอนโทรปีของชุดข้อมูล S คำนวณโดย

Information Gain เป็นการวัดว่าการทราบค่าของตัวแปรหนึ่งช่วยลดความไม่แน่นอน (entropy) ของตัวแปรเป้าหมายได้มากน้อยเพียงใด มักใช้ใน Decision Tree Algorithms (เช่น ID3, C4.5) เพื่อเลือกคุณสมบัติที่ดีที่สุดในการแบ่งข้อมูล โดยคุณสมบัติที่มี Information Gain สูงสุดจะถูกเลือกเป็นโหนดหลัก

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (10)$$

โดยที่

- rxy คือ สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร x และ y
- xi คือ ค่าของตัวแปร x สำหรับข้อมูลตัวที่ i
- yi คือ ค่าของตัวแปร y สำหรับข้อมูลตัวที่ i
- $\bar{x}$ คือ ค่าเฉลี่ยของตัวแปร x
- $\bar{y}$ คือ ค่าเฉลี่ยของตัวแปร y
- n คือ จำนวนข้อมูล

Correlation-based หมายถึงการใช้สัมประสิทธิ์สหสัมพันธ์เพื่อวัดความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรสองตัว ค่าสัมประสิทธิ์สหสัมพันธ์จะอยู่ระหว่าง −1 ถึง 1

- 1 สหสัมพันธ์เชิงบวกที่สมบูรณ์แบบ (เมื่อตัวแปรหนึ่งเพิ่มขึ้น อีกตัวก็เพิ่มขึ้นด้วยอัตราส่วนคงที่)
- −1 สหสัมพันธ์เชิงลบที่สมบูรณ์แบบ (เมื่อตัวแปรหนึ่งเพิ่มขึ้น อีกตัวกลับลดลงด้วยอัตราส่วนคงที่)
- 0 ไม่มีความสัมพันธ์เชิงเส้นตรง

3.1.5 ขั้นตอนการสร้างแบบจำลอง (Modeling)

การสร้างแบบจำลองเป็นกระบวนการสร้างรูปแบบทางคณิตศาสตร์หนึ่งอย่างหรือมากกว่า โดยอาศัยข้อมูลจากอดีตมาสร้างเป็นแบบจำลอง (Model) การพยากรณ์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ในงานวิจัยนี้นอกแบบและสร้างแบบจำลองการพยากรณ์จำนวน 3 แบบจำลอง ได้แก่ เทคนิคต้นไม้ตัดสินใจ แรนดอมฟอร์เรสต์ และ เทคนิค (K–NN)

1) **เทคนิค ต้นไม้ตัดสินใจ** (Decision Tree) เป็นอัลกอริทึมการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ใช้สำหรับการ จำแนกประเภท (Classification) และ การพยากรณ์ค่า (Regression) โดยสร้างแบบจำลองที่มีโครงสร้างคล้ายแผนผังต้นไม้ ซึ่งแต่ละโหนด (Node) ในต้นไม้จะแสดงถึงการทดสอบคุณสมบัติ (Attribute) ของข้อมูล กิ่ง (Branch) แสดงถึงผลลัพธ์ที่เป็นไปได้ของการทดสอบนั้น และใบ (Leaf Node) แสดงถึงผลลัพธ์สุดท้ายที่เป็นการจัดประเภทหรือค่าที่พยากรณ์ได้

2) เทคนิค แรนดอมฟอร์เรสต์ (Random Forest) เป็นการขยายมาจาก Decision Tree (Gehrke et al., 2000) โดยการรวมต้นไม้ตัดสินใจหลายต้นเข้าด้วยกันผ่านกระบวนการ Bagging (Bootstrap Aggregating) วิธีการนี้ช่วยลดความเอนเอียงของโมเดล (Bias) และเพิ่มความแม่นยำของผลลัพธ์ ทั้งยังใช้การสุ่มตัวแปรที่นำมาใช้สร้างต้นไม้แต่ละต้นทำให้โมเดลมีความหลากหลายและมีความทนทานต่อข้อมูลที่มี Noise จุดเด่นของ Random Forest คือความสามารถในการจัดการข้อมูลขนาดใหญ่และซับซ้อน อย่างไรก็ตาม ข้อจำกัดของวิธีการนี้คือการตีความผลลัพธ์ทำได้ยากกว่าการใช้ Decision Tree แบบเดี่ยว

3) เทคนิค K–Nearest Neighbors หรือ K–NN เป็นโมเดลที่ทำงานโดยการเปรียบเทียบข้อมูลใหม่กับข้อมูลตัวอย่างที่ใกล้เคียงในพื้นที่มิติของข้อมูล (Feature Space) โดยใช้ระยะทาง (Cheng et al., 2014) เช่น ระยะทางแบบ Euclidean หรือ Manhattan ในการวัดความใกล้เคียง หลักการของ K–NN คือการพิจารณาคลาสที่ปรากฏมากที่สุดจากข้อมูลตัวอย่างใกล้เคียงจำนวน k ตัว ข้อดีของ K–NN คือมีความง่ายในการใช้งานและสามารถรองรับข้อมูลที่มีรูปแบบหลากหลาย อย่างไรก็ตาม ข้อเสียคือความต้องการทรัพยากรค่อนข้างสูงเมื่อข้อมูลมีขนาดใหญ่ อีกทั้งประสิทธิภาพยังลดลงเมื่อข้อมูลมีจำนวนมิติมาก (High–dimensional data)

4)**เทคนิค การจัดกลุ่ม (Clustering)** เป็นวิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ที่ทำงานโดยการวัดความคล้ายคลึงของข้อมูลผ่านตัวชี้วัด เช่น

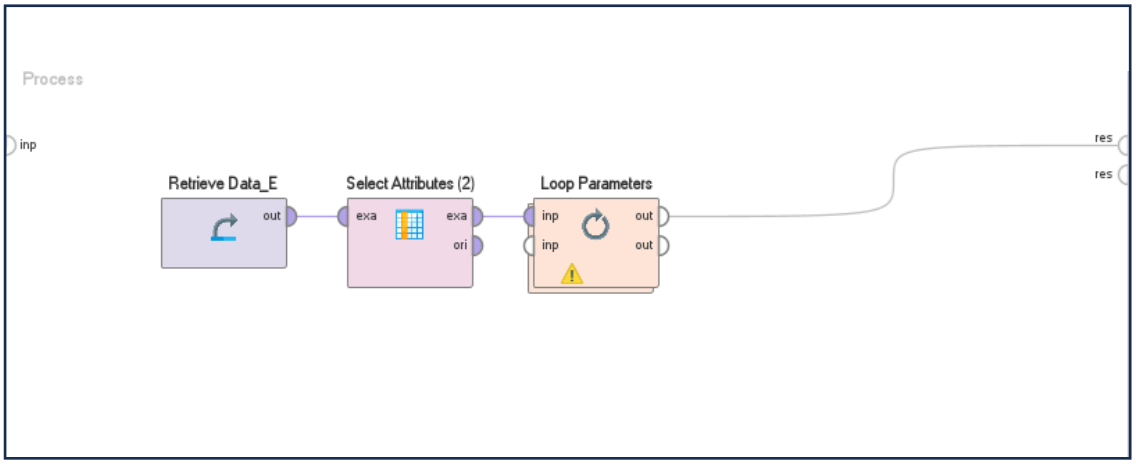
ระยะทางหรือความหนาแน่น จากนั้นจึงจัดข้อมูลที่มีลักษณะใกล้เคียงกันให้อยู่ในกลุ่มเดียวกัน และแยกข้อมูลที่แตกต่างออกไปให้อยู่ในกลุ่มอื่น ผลลัพธ์ที่ได้ช่วยสะท้อนโครงสร้างและรูปแบบที่ซ่อนอยู่ภายในข้อมูล ทำให้สามารถนำไปใช้ในการวิเคราะห์เชิงลึกและสนับสนุนการตัดสินใจได้อย่างมีประสิทธิภาพ

ในการวิจัยครั้งนี้ ผู้วิจัยได้คัดเลือกตัวแปรที่ใช้ในการวิเคราะห์จำนวน 4 ตัว ได้แก่ อายุวุฒิการศึกษา ผลงานทั้งหมด และอายุงาน โดยเหตุผลในการเลือกใช้ตัวแปรดังกล่าว เนื่องจากเป็นปัจจัยที่มีความสัมพันธ์โดยตรงกับการพิจารณาความก้าวหน้าทางวิชาการตามเกณฑ์ของ ก.พ.อ. ตัวแปร อายุ ใช้สะท้อนถึงวุฒิภาวะและประสบการณ์ที่สั่งสมในการทำงาน ตัวแปร วุฒิการศึกษา แสดงถึงคุณสมบัติด้านการศึกษาอันเป็นเงื่อนไขพื้นฐานของการขอตำแหน่ง ตัวแปร ผลงานทั้งหมด ใช้วัดศักยภาพด้านการสร้างองค์ความรู้และคุณภาพของงานวิจัย ขณะที่ตัวแปร อายุงาน ใช้แสดงถึงระยะเวลาที่ปฏิบัติงานทางวิชาการอย่างต่อเนื่อง ทั้งนี้ การเลือกตัวแปรทั้ง 4 ช่วยให้กระบวนการวิเคราะห์ข้อมูลมีความถูกต้อง สอดคล้องกับกรอบการประเมินจริง และสามารถสะท้อนศักยภาพของบุคลากรได้อย่างสมบูรณ์ยิ่งขึ้น

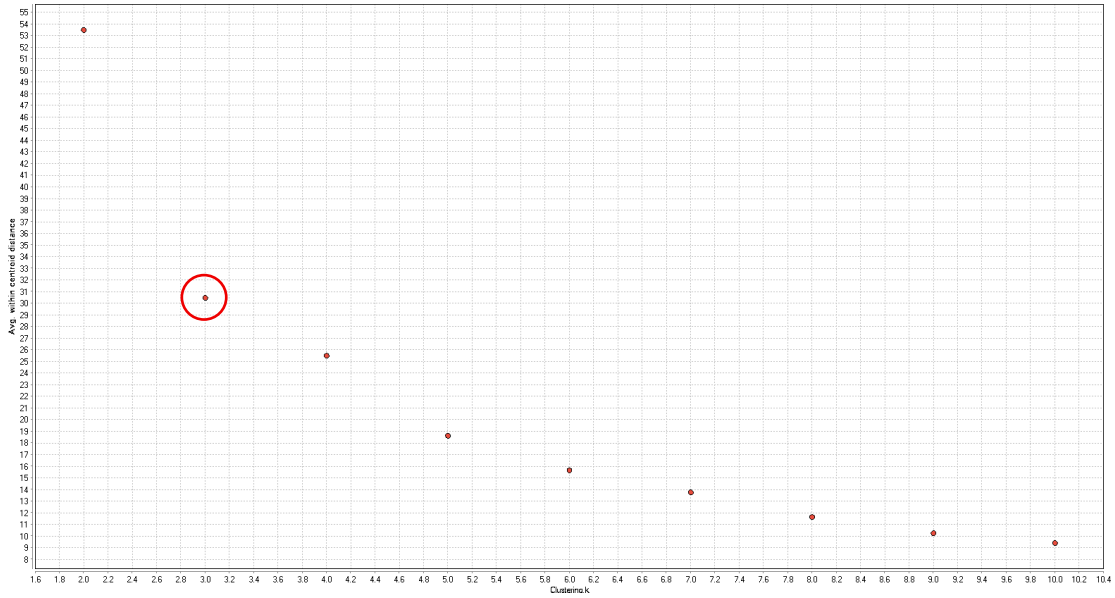
ตารางที่ 3.4 คุณลักษณะที่ใช้สำหรับทำการจัดกลุ่ม

ชื่อคุณลักษณะ (Attribute Name)	**คำอธิบาย** (Description)	**ประเภทข้อมูล** (Data Type)
วุฒิการศึกษา	จำนวนปีของอายุบุคคลตั้งแต่เกิดจนถึงปัจจุบัน	Numeric
ผลงานทั้งหมด	จำนวนผลงานทางวิชาการหรือวิจัยที่บุคคลผลิตได้รวมทั้งหมด	Numeric
อายุงาน	จำนวนปีที่บุคคลปฏิบัติงานในตำแหน่งหรือสายงานปัจจุบัน	Numeric
อายุ	จำนวนปีของอายุบุคคลตั้งแต่เกิดจนถึงปัจจุบัน	Numeric

3.2.2 การเลือกจำนวณกลุ่มโดยใช้วิธี Elbow Method



ภาพที่ 3.7 Process หาค่า K Elbow Method



ภาพที่ 3.8 ผล Elbow Method

Loop Parameters (9 rows, 3 columns)

iteration	Clustering.k	Avg. within centroid distance
1	2	53.491
2	3	30.472
3	4	25.505
4	5	18.632
5	6	15.661
6	7	13.768
7	8	11.650
8	9	10.259
9	10	9.416

ภาพที่ 3.9 ค่า Avg. within centroid distance

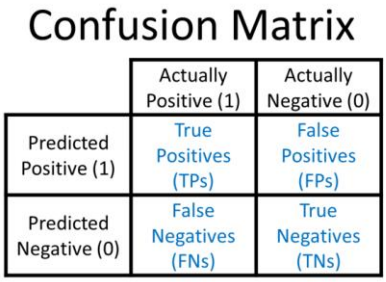
จากผลการประเมินค่า Avg. within centroid distance และกราฟ Elbow Method พบว่า จุดหักศอก (elbow point) ปรากฏชัดเจนที่ K = 3 โดยค่าเฉลี่ยระยะห่างภายในกลุ่มลดลงอย่างมากจาก 53.491 (K=2) เหลือ 30.472 (K=3) ก่อนที่จะเริ่มลดลงช้าลงในค่า K ที่สูงกว่า ซึ่งบ่งบอกว่าเมื่อใช้การแบ่งกลุ่มเป็น 3 กลุ่ม จะได้ความสมดุลระหว่าง ความกะทัดรัดของกลุ่ม (compactness) และ ความสามารถในการแยกแยะกลุ่ม (separation) ที่เหมาะสมที่สุด

จึงสรุปได้ว่าจากการแบ่งกลุ่ม K=3 พบว่า Cluster 1 มีศักยภาพสูงที่สุดในการขอตำแหน่ง เนื่องจากแม้อายุงานยังไม่มาก แต่มีผลงานโดดเด่นCluster 0 มีความมั่นคงในอาชีพและประสบการณ์ แต่ควรเพิ่มคุณภาพผลงานตีพิมพ์เพื่อสอดคล้องกับเกณฑ์ ก.พ.อ.Cluster 2 เป็นกลุ่มที่ยังต้องการพัฒนาทั้งด้านประสบการณ์และผลงาน

3.1.6 ขั้นตอนการประเมินผล (Evaluation)

โดยใช้วิธีการวิเคราะห์ข้อมูลด้วยเทคนิคการคาดคะเนในรูปแบบการจัดกลุ่มข้อมูลและการทำนายผล ได้ทำการทดสอบประสิทธิภาพของโมเดลโดยเครื่องมือ **Cross-validation test** เพื่อให้มั่นใจในความถูกต้องและความน่าเชื่อถือของผลลัพธ์ที่ได้การแบ่งข้อมูล 90:10 โดยแบ่งข้อมูลสำหรับการเรียนรู้(Training Data) 90% และข้อมูลสำหรับการ

ทดสอบ (Testing Data) 10% (Vrigazova, 2021)โดยใช้วิธีการประเมินผลประสิทธิภาพของโมเดลด้วยตาราง Confusion Matrix (Sripaoray and Sinsomboonthong, 2017)



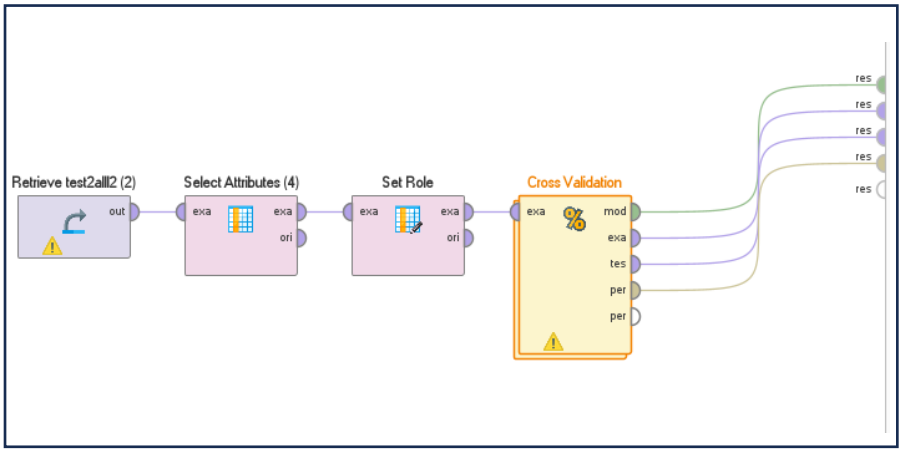
ภาพที่ 3.10 ตาราง Confusion Matrix

โดยที่

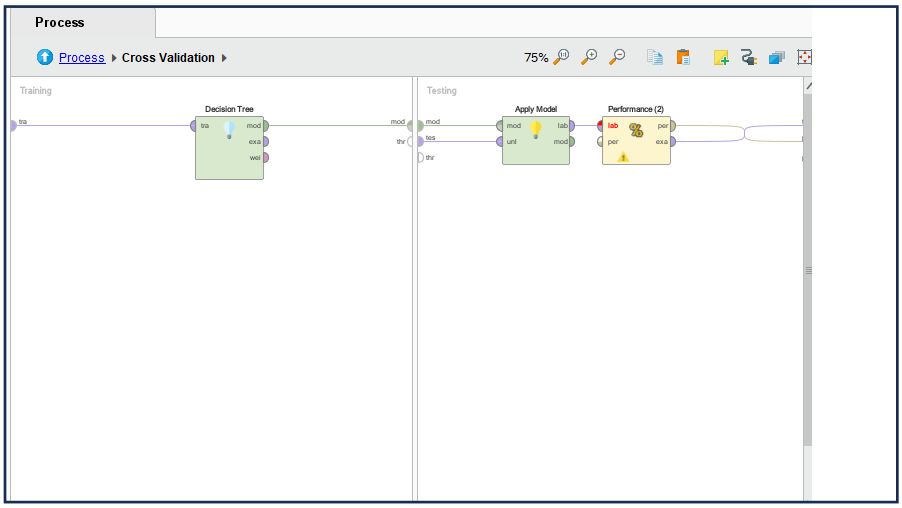
- True Positive (TP)= สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริง ในกรณี ทำนายว่าจริง และสิ่งที่เกิดขึ้น ก็คือ จริง
- True Negative (TN)= สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้น ก็คือ ไม่จริง
- False Positive (FP)= สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง
- False Negative (FN)= สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง
- โดย TP,TN,FP,FN ในตารางจะแทนด้วยค่าความถี่เราสามารถใช้ Confusion Matrix มาคำนวน การประเมินประสิทธิภาพของการทำนายด้วยModel ของเรา ในรูปแบบค่าต่าง ๆ ได้หลายค่า ได้แก่

F1 score F1–Score เป็นค่าเฉลี่ยแบบ harmonic mean ระหว่าง precision และ recall จุดประสงค์ของการสร้าง F1 ขึ้นมา คือ เพื่อเป็น single metric ที่วัดความสามารถของโมเดล

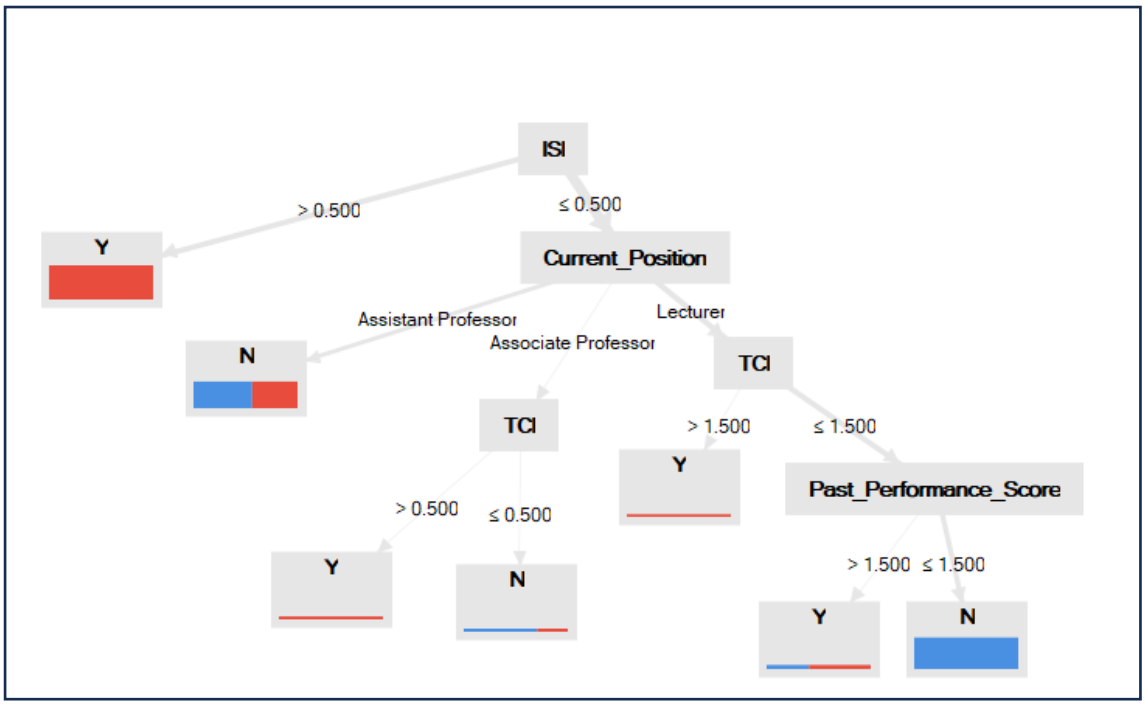
F1 = 2 x (Precision x Recall)/(Precision + Recall)



ภาพที่ 3.11 Process การประเมินผลด้วย โดยเครื่องมือ **Cross-validation**



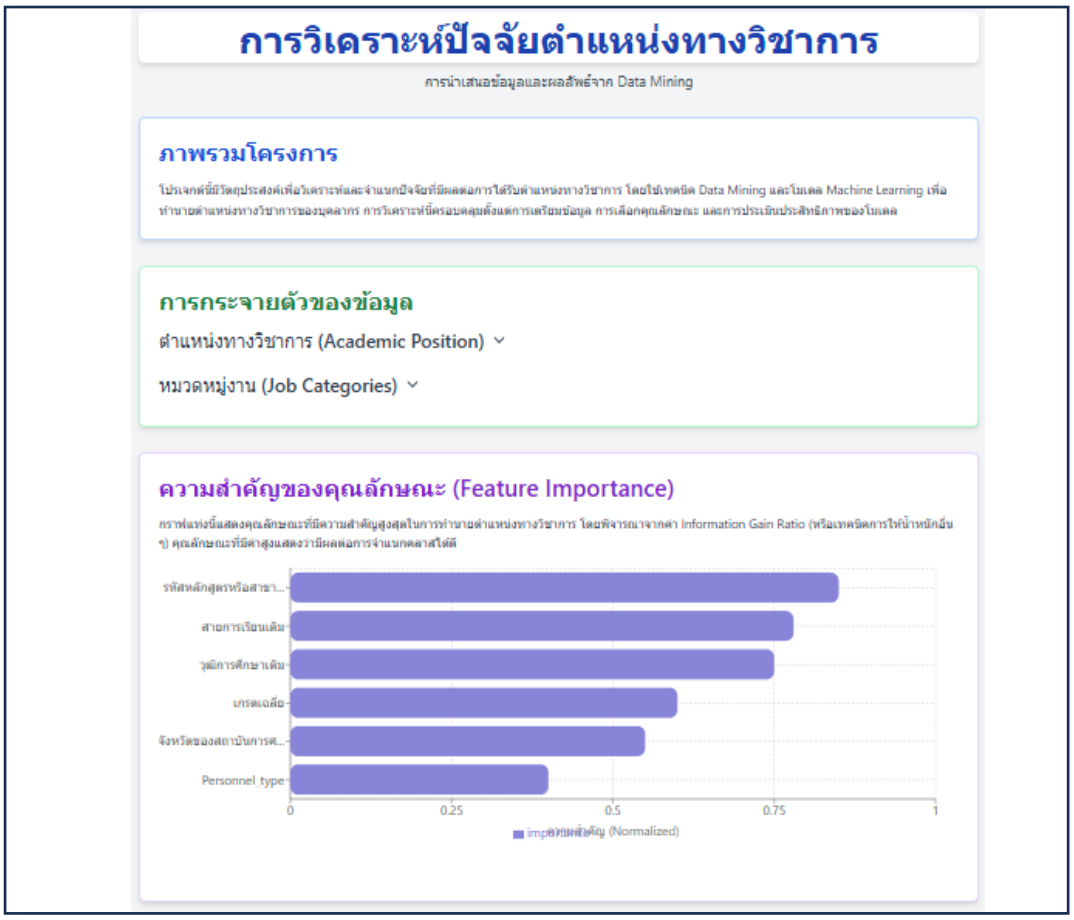
ภาพที่ 3.12 Process การพัฒนาแบบจำลองการจำแนกประเภทข้อมูล



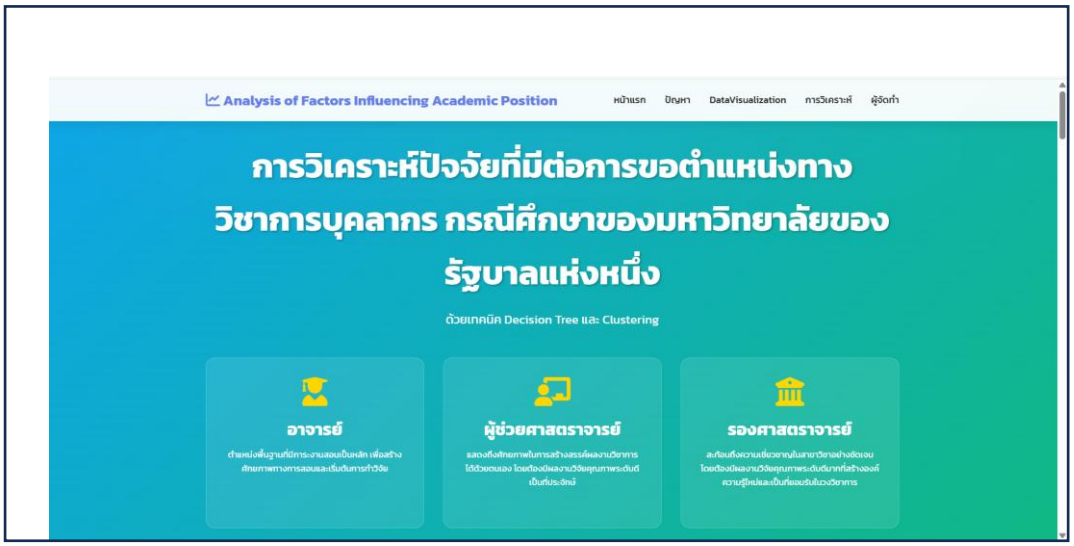
ภาพที่ 3.13 โมเดล Tree

3.1.6 ขั้นตอนการเผยแพร่ผลการวิเคราะห์ (Deployment)

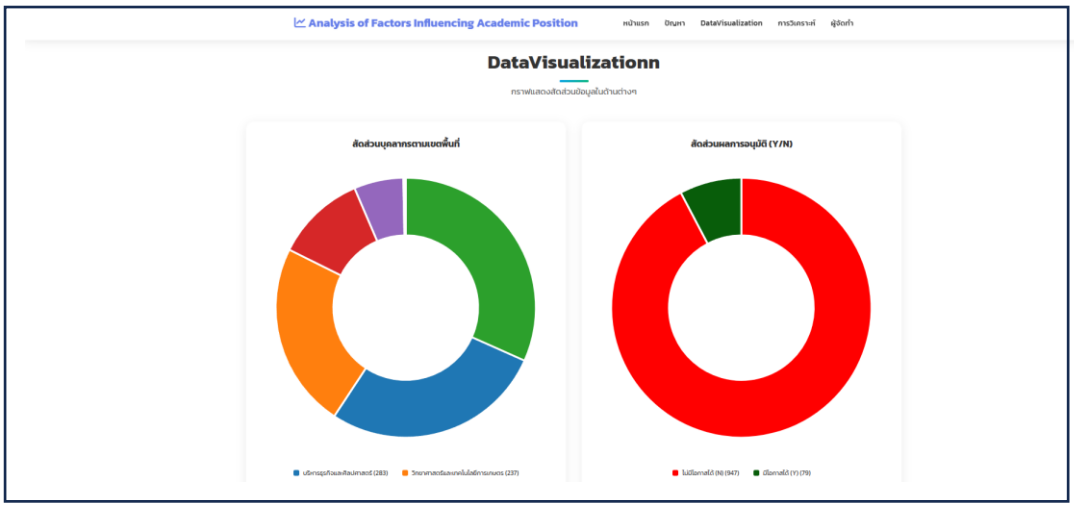
ข้อมูลและผลการวิเคราะห์ถูกนำเสนอผ่านเว็บไซต์ที่พัฒนาขึ้น โดยใช้วิธีการนำโมเดล Decision Tree ที่จัดเก็บในรูปแบบไฟล์ .json มาบูรณาการเข้ากับภาษา HTML เพื่อสร้างเป็นเว็บสำหรับการนำเสนอผลการวิจัย



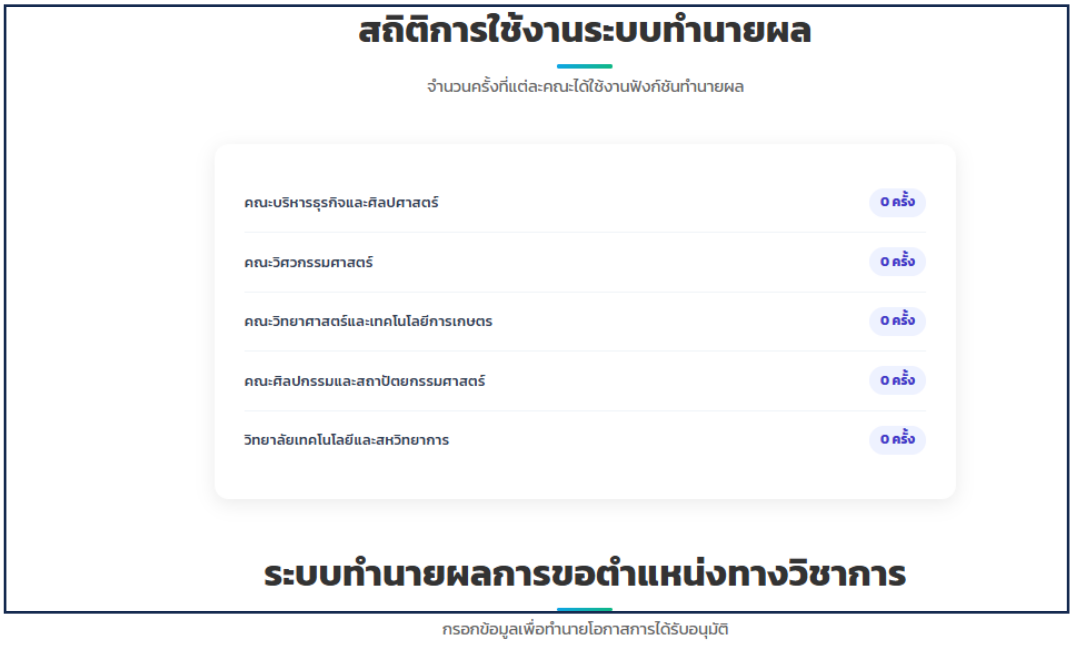
ภาพที่ 3.14 ต้นแบบเว็บนำเสนอ visualization



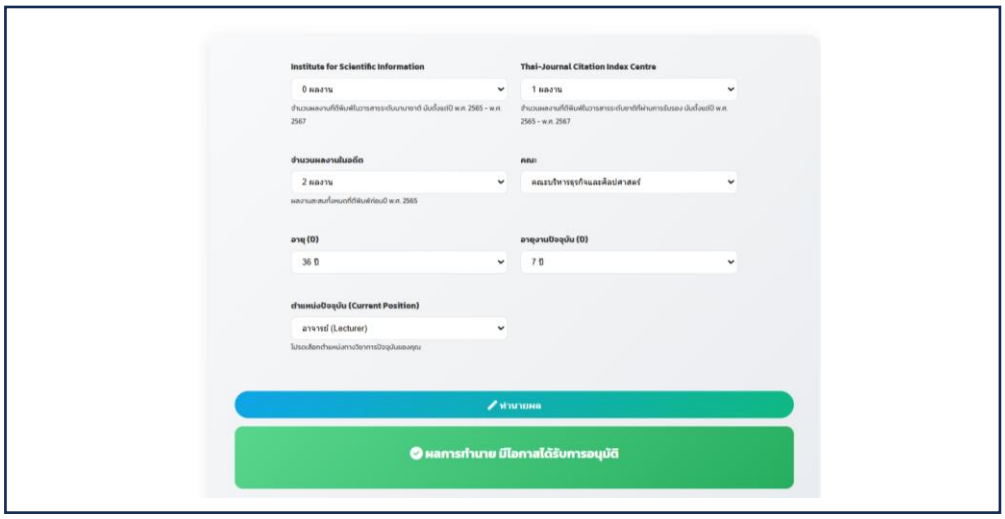
ภาพที่ 3.15 หน้าแรกของเว็บไซต์



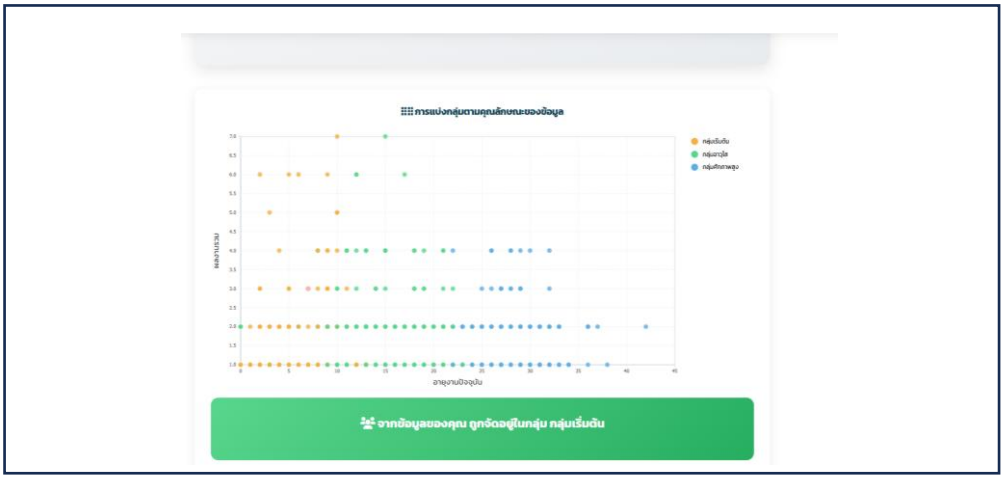
ภาพที่ 3.16 Data visualization



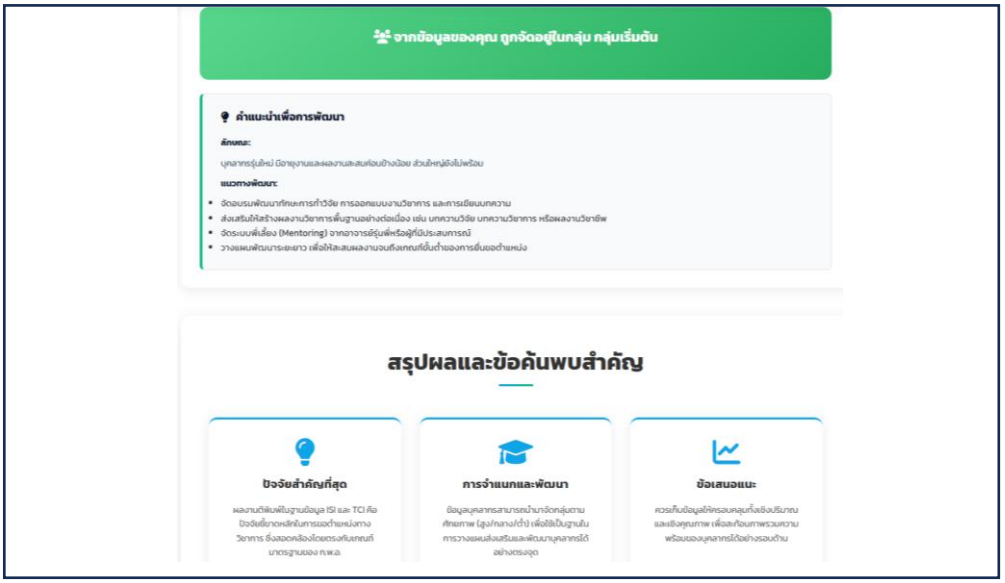
ภาพที่ 3.17 สถิติการใช้งานระบบทำนายผล



ภาพที่ 3.18 ฟอร์มกรอกค่าเพื่อทำนายผลและผลลัพธ์



ภาพที่ 3.19 ผลลัพธ์ของการจัดกลุ่มของข้อมูล



ภาพที่ 3.20 คำแนะนำในการพัฒนาตัวเอง

ภาพที่ 3.21 ผู้จัดทำ

การพัฒนาเว็บไซต์เพื่อวิเคราะห์ปัจจัยจากข้อมูลการวิจัย มีประโยชน์สำคัญในด้านการนำผลการวิเคราะห์ไปใช้ประโยชน์จริง เนื่องจากเว็บไซต์ที่พัฒนาขึ้นช่วยให้ผู้ใช้งานสามารถกรอกข้อมูลของตนเองหรือบุคลากรที่สนใจ และได้รับผลการทำนายหรือวิเคราะห์ได้อย่างสะดวกรวดเร็ว โดยไม่จำเป็นต้องใช้ซอฟต์แวร์เชิงสถิติที่ซับซ้อน อีกทั้งยังทำให้โมเดลที่ได้จากงานวิจัยสามารถนำไปใช้ในเชิงปฏิบัติได้จริง และช่วยลดช่องว่างระหว่างผลการวิจัยกับการใช้งานในชีวิตจริง

นอกจากนี้ เว็บไซต์ยังเป็นเครื่องมือที่ช่วยให้การนำเสนอผลลัพธ์มีความโปร่งใสและตรวจสอบได้ เนื่องจากสามารถแสดงเงื่อนไขของการตัดสินใจตามโมเดล Decision Tree ได้อย่างชัดเจน ผู้ใช้จึงสามารถเข้าใจปัจจัยที่มีผลต่อการตัดสินใจหรือการทำนายผลลัพธ์ได้โดยตรง อีกทั้งยังเป็นประโยชน์ต่อผู้บริหารหรือผู้กำหนดนโยบาย ที่สามารถนำข้อมูลที่ได้ไปใช้ในการวางแผนพัฒนาบุคลากร การสนับสนุนการอบรม และการกำหนดกลยุทธ์เพื่อยกระดับคุณภาพงานวิชาการขององค์กร

ในขณะเดียวกัน เว็บไซต์ดังกล่าวยังทำหน้าที่เป็นแหล่งเรียนรู้ให้แก่บุคลากร นักศึกษา และผู้สนใจทั่วไป ในการทำความเข้าใจความสัมพันธ์ของปัจจัยต่าง ๆ ที่มีผลต่อความพร้อมในการขอตำแหน่งทางวิชาการ ส่งผลให้เกิดการขยายผลของงานวิจัยไปสู่การเรียนรู้และการพัฒนาทั้งในระดับบุคคลและองค์กรอย่างต่อเนื่อง