บทที่ 3
วิธีการดำเนินการวิจัย

3.1 เครื่องมือและอุปกรณ์

1. ฮาร์ดแวร์

   คอมพิวเตอร์ส่วนบุคคล CPU Intel Pentium 4 2.93GHz, 256 DDR SDRAM, HD 40 GB

2. ซอฟต์แวร์

   Weka version 3.6.2 สำหรับการสร้างแบบจำลองต้นไม้การตัดสินใจ

3. ระบบฐานข้อมูล

   MySQL version 5.0.45 สำหรับจัดเก็บข้อมูลลงฐานข้อมูล

4. ระบบปฏิบัติการ

   Microsoft Window 7

3.2 ข้อมูลสำหรับทำการวิจัย (Data set)

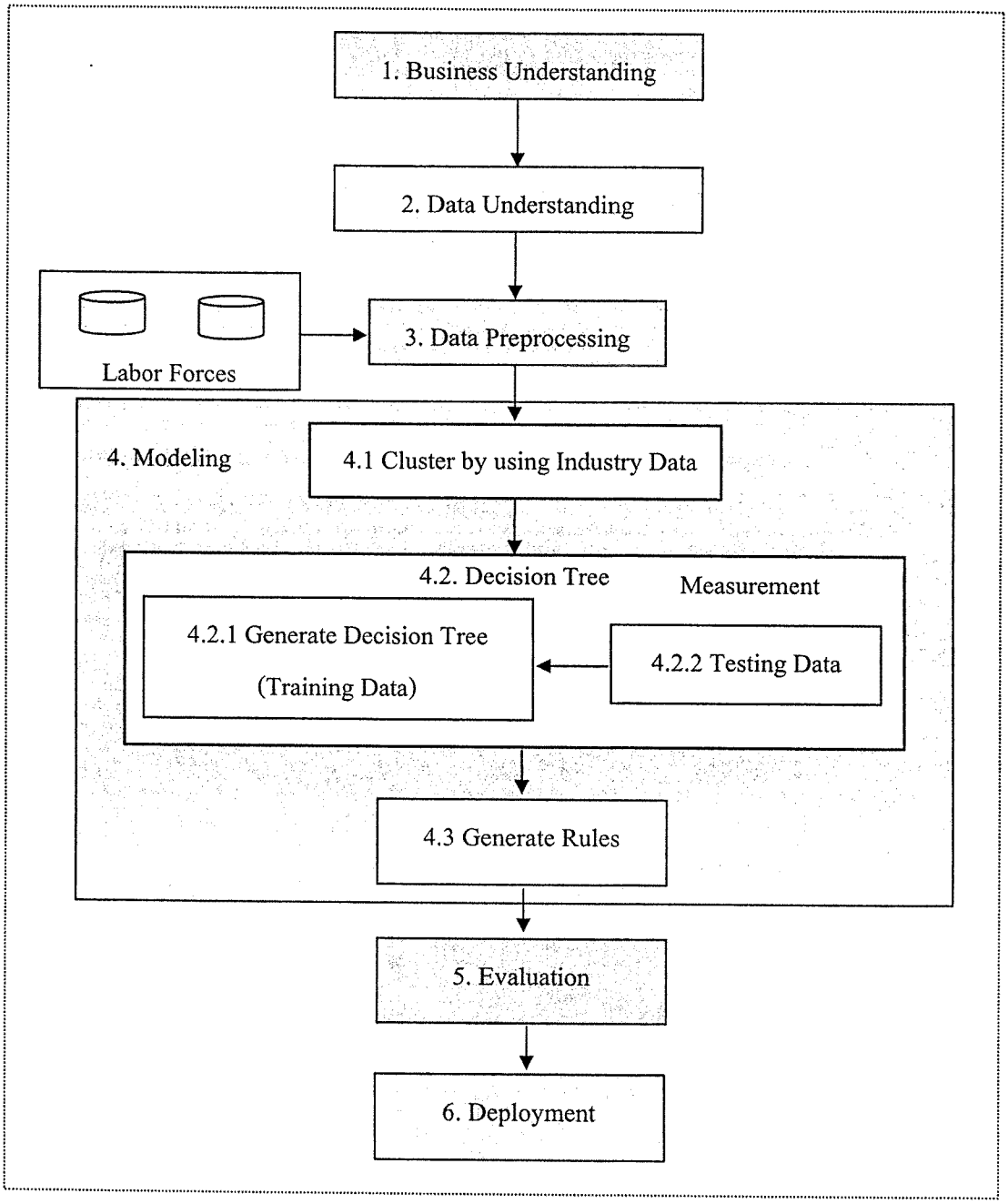
ข้อมูลสำหรับทำการวิจัยนี้ คือข้อมูลการสำรวจด้านการทำงานของประชากร โดยเก็บรวบรวมข้อมูลเกี่ยวกับข้อมูลพื้นฐานของประชากรทั่วประเทศ ได้แก่ สถานภาพการทำงาน ความพร้อมในการทำงาน การสมัครงาน ความต้องการทำงานเพิ่มเติมจากงานประจำ และรายได้ค่าตอบแทนในรูปแบบต่างๆ ซึ่งจัดเก็บเป็นแบบรายไตรมาส ข้อมูลสำหรับทำการวิจัยนี้ได้จากศูนย์วิจัยมหาวิทยาลัยชิคาโก-มหาวิทยาลัยหอการค้าไทย ที่จัดเก็บข้อมูลดังกล่าวในฐานข้อมูลปฐมภูมิ ซึ่งได้มาจากสำนักงานสถิติแห่งชาติ โดยใช้ข้อมูลในปี พ.ศ. 2552 จำนวน 193,176 รายการ โดยแสดงตัวอย่างข้อมูลการสำรวจด้านการทำงานของประชากรดังตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างตารางข้อมูลการสำรวจด้านการทำงานของประชากร

REG	AREA	MONTH	YEAR	SEX	AGE	MARITAL	OCCUP	INDUS	TOTAL_HR	WAGE_TYPE	SALARY
1	1	01	52	1	47	2	8324	6023	50	4	8800
1	1	03	52	1	62	2	6113	0140	48	4	6500
1	1	03	52	2	57	2	6113	0140	48	4	6500
2	1	01	52	2	45	1	9132	8511	24	4	5300
2	1	01	52	1	38	1	7213	5020	56	4	7000
3	2	01	52	1	40	1	6121	0121	42	4	4000
3	1	01	52	1	41	2	5161	7523	35	4	11700
2	1	01	52	2	20	4	8283	3210	48	2	5278
2	1	02	52	2	31	1	3415	2424	48	4	25000
2	1	02	52	2	41	2	2230	8511	50	4	20000
2	1	02	52	2	47	1	2412	3210	48	4	17000
2	2	02	52	1	60	2	7124	4520	70	2	7800
2	2	02	52	1	33	5	8322	6711	40	4	10000
2	2	02	52	2	34	2	7412	1544	48	2	4680
2	2	02	52	1	42	6	7124	4520	48	2	7020

3.3 ขั้นตอนการทดลอง

ขั้นตอนการทดลองในงานวิจัยนี้ใช้เทคนิคของการทำเหมืองข้อมูล คือการจัดกลุ่มข้อมูลตามประเภทอุตสาหกรรม จากนั้นจะใช้การจำแนกประเภทข้อมูลเพื่อสร้างแบบจำลองที่ใช้จำแนกข้อมูล โดยใช้ทฤษฎีแบบจำลองต้นไม้การตัดสินใจ ผลลัพธ์ที่ได้จะถูกแปลงเป็นกฎความสัมพันธ์ เพื่อใช้ในการศึกษาปัจจัยที่มีผลต่อรายได้ของประชากร ดังรูปที่ 3.1



รูปที่ 3.1 แสดงภาพโครงสร้างการทำงานของงานวิจัย

1. การทำความเข้าใจธุรกิจ (Business Understanding)

ปัจจุบันนี้สภาพเศรษฐกิจของประเทศไทยอยู่ในสภาวะที่มีการแข่งขันสูง ซึ่งมีผลกระทบต่อชีวิตความเป็นอยู่ของประชากร โดยเฉพาะอย่างยิ่งรายได้ซึ่งเป็นปัจจัยสำคัญต่อการดำรงชีวิตของประชากร ในงานวิจัยนี้จะทำการรวบรวมข้อมูลเพื่อจัดกลุ่มข้อมูลและสร้างแบบจำลอง รวมถึงสร้างกฎความสัมพันธ์เพื่อหาปัจจัยที่มีผลต่อรายได้ของประชาการในประเทศไทย

2. การทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลที่ใช้ในงานวิจัยนี้คือข้อมูลการสำรวจด้านการทำงานของประชากร ซึ่งได้จากศูนย์วิจัยมหาวิทยาลัยชิคาโก-มหาวิทยาลัยหอการค้าไทย ที่จัดเก็บข้อมูลดังกล่าวในฐานข้อมูลปฐมภูมิ ซึ่งได้มาจากสำนักงานสถิติแห่งชาติ ตั้งแต่เดือนมกราคมถึงเดือนธันวาคม พ.ศ. 2552 จัดเก็บข้อมูลเป็นรายไตรมาส

3. การเตรียมข้อมูล (Data Preprocessing)

เป็นขั้นตอนนี้จะทำการรวบรวมข้อมูลข้อมูลการสำรวจด้านการทำงานของประชากร ในระยะเวลา 1 ปี (พ.ศ. 2552) มาจัดเก็บลงฐานข้อมูล MySQL จากนั้นคัดเลือกข้อมูลที่ต้องการนำมาใช้ประโยชน์ โดยจะคัดเลือกเฉพาะรายการที่มีรายได้ทั้งที่เป็นตัวเงินและไม่เป็นตัวเงินเท่านั้น จากนั้นจะทำการทำความสะอาดข้อมูล ทำการตัดข้อมูลที่ไม่ครบถ้วนทิ้ง และแปลงข้อมูลให้อยู่ในรูปที่สามารถนำไปใช้ประมวลผลได้ เช่น ข้อมูลอายุ ที่ได้มาจากสำนักงานสถิติจะเป็นจำนวนจริงตั้งแต่ 1-99 ทางคณะผู้วิจัยต้องทำการแปลงอายุให้เป็นช่วง ซึ่งแบ่งออกเป็น 4 กลุ่มคือ กลุ่มที่ 1 คือ อายุ 15-24 ปี กลุ่มที่ 2 คือ อายุ 25-34 ปี กลุ่มที่ 3 คือ อายุ 35-49 ปี และกลุ่มที่ 4 คือ อายุมากกว่าหรือเท่ากับ 50 ปี ข้อมูลที่นำมาใช้ในงานวิจัยนี้ มีจำนวนทั้งหมด 193,176 รายการ

4. การพัฒนาตัวแบบ (Modeling)

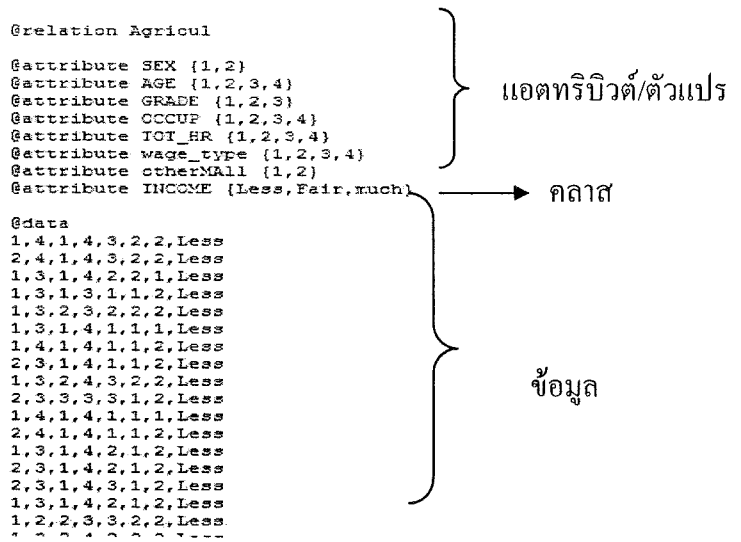
4.1 จัดกลุ่มข้อมูลตามประเภทอุตสาหกรรม

ในขั้นตอนนี้จะทำการจัดกลุ่มข้อมูลการสำรวจด้านการทำงานของประชากรที่ได้จากขั้นตอนที่ 1 โดยจะแบ่งข้อมูลออกเป็น 4 กลุ่ม ตามประเภทอุตสาหกรรม คือ (1) การเกษตรและประมง (2) การผลิต (3) การค้าและการบริการ และ (4) อื่น ๆ จากนั้นจะนำข้อมูลทั้ง 4 กลุ่ม นี้ไปใช้ในกระบวนการสร้างแบบจำลองต้นไม้การตัดสินใจในขั้นตอนที่ 4.2 ต่อไป

4.2 ขั้นตอนการสร้างแบบจำลองต้นไม้การตัดสินใจ

งานวิจัยนี้เลือกใช้การสร้างแบบจำลองต้นไม้การตัดสินใจ เนื่องจากง่ายต่อการทำความเข้าใจ สามารถนำมาสร้างเป็นกฎ (Rule) แสดงความสัมพันธ์ได้ง่าย รวมถึงสามารถทำนายหรือจำแนกประเภทของข้อมูลได้ทั้งข้อมูลที่เป็นตัวเลข (Numeric) และเป็นข้อมูลที่เป็นชนิดของกลุ่ม (Category) โดยขั้นตอนนี้สามารถแบ่งเป็น 2 ขั้นตอนย่อยดังนี้

4.2.1 ขั้นตอนการสร้างแบบจำลองโดยการเรียนรู้จากข้อมูลที่กำหนดคลาสไว้แล้ว ซึ่งถูกเรียกว่าข้อมูลที่ใช้ในการเรียนรู้หรือข้อมูลฝึก โดยแบบจำลองที่ได้แสดงในรูปของต้นไม้การตัดสินใจ ในขั้นตอนนี้จะใช้อัลกอริทึม C4.5 ในการสร้างแบบจำลองต้นไม้ตัดสิน โดยจะนำข้อมูลทั้ง 4 กลุ่มที่ได้จากขั้นตอนที่ 3 มาแบ่งออกเป็น 2 ส่วน ส่วนแรกมีทั้งสิ้นร้อยละ 80 จะถูกใช้ในการเรียนรู้ที่เหลืออีกร้อยละ 20 จะถูกใช้ในขั้นตอนการประเมินความถูกต้องต่อไป โดยในที่นี้จะใช้ช่วงของรายได้ (INCOME) เป็นคลาสเพื่อสร้างแบบจำลอง และก่อนที่จะนำข้อมูลเข้าสู่โปรแกรม Weka เพื่อใช้ในการสร้างแบบจำลองนั้น ข้อมูลจะต้องถูกแปลงให้อยู่ในรูปแบบ .arff ดังรูปที่ 3.2



รูปที่ 3.2 แสดงตัวอย่างไฟล์ .arff

พารามิเตอร์ที่ใช้ในการสร้างต้นไม้การตัดสินใจในโปรแกรม Weka

- Confidence Factor (CF) คือ ค่าความเชื่อมั่นใช้สำหรับการตัดเล็มต้นไม้ตัดสินใจ (Pruning) แบบ Error Base Pruning (EBP) ค่าจะอยู่ระหว่าง 0.00-1.00 ค่า CF ที่สูงจะทำให้ได้แบบจำลองต้นไม้ที่มีลักษณะเฉพาะ (Specific) เพราะมีการตัดเล็มต้นไม้ตัดสินใจ และหากค่า CF มีค่าน้อย ต้นไม้จะมีขนาดเล็กและมีลักษณะทั่วไป (Generalize) ในโปรแกรม Weka จะกำหนดไว้ เท่ากับ 0.25 ในงานวิจัยนี้จะทำการปรับค่า CF ทั้งหมด 4 ค่า คือ 0.01, 0.1, 0.25 และ 0.5
- Minimum Number of Instance per Leaf (M) จำนวนข้อมูลที่น้อยที่สุดต่อ 1 Leaf ซึ่งกำหนดให้เป็นค่าจำนวนเต็ม โดยมีค่ามากกว่าหรือเท่ากับ 0 และกำหนดค่า Default เท่ากับ 2

ตารางที่ 3.2 แสดงชุดทดสอบในการสร้างแบบจำลองต้นไม้การตัดสินใจ

ชุดทดสอบ	Confidence Factor
T1	0.01
T2	0.1
T3	0.25
T4	0.5

โดยที่ ชุดทดสอบ T1 คือ ชุดทดสอบที่มีการกำหนดค่า Confidence Factor เท่ากับ 0.01
ชุดทดสอบ T2 คือ ชุดทดสอบที่มีการกำหนดค่า Confidence Factor เท่ากับ 0.1
ชุดทดสอบ T3 คือ ชุดทดสอบที่มีการกำหนดค่า Confidence Factor เท่ากับ 0.25
ชุดทดสอบ T4 คือ ชุดทดสอบที่มีการกำหนดค่า Confidence Factor เท่ากับ 0.5

4.2.2 ขั้นตอนการประเมินความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ ซึ่งเป็นข้อมูลที่เหลืออีกร้อยละ 20 จากขั้นตอน 4.2.1 ข้อมูลเหล่านี้ถูกนำไปใช้เพื่อทดสอบความถูกต้องของแบบจำลองที่ได้ จากขั้นตอนที่ 4.2.1 การทดสอบประสิทธิภาพในการสร้างแบบจำลองต้นไม้การตัดสินใจจะทำการวัดประสิทธิภาพจาก

- ขนาดของต้นไม้ จำนวนโหนด และโหนดใบ
- ค่าความถูกต้องในการทำนายผล (Correctly Classified Instances) คือค่าเปอร์เซ็นต์ของความถูกต้องในการทำนายผลแบบจำลอง โดยการใช้ข้อมูลทดสอบ สามารถคำนวณได้สมการ (3)

$$\text{Correction} = \frac{\text{จำนวนข้อมูลทดสอบที่จำแนกได้ถูกต้อง}}{\text{จำนวนข้อมูลทดสอบทั้งหมด}} \qquad (3)$$

- ค่าความระลึก (Recall) คำนวณจากค่าของข้อมูลที่ผลลัพธ์ถูกต้องโดยพิจารณาจากข้อมูลของผลลัพธ์เดียวกัน สามารถคำนวณได้สมการ (4)

$$\text{Recall} = \frac{\text{จำนวนข้อมูลที่จำแนกได้ถูกต้องของกลุ่มนั้น}}{\text{จำนวนข้อมูลที่อยู่ในกลุ่มนั้นจริง}} \qquad (4)$$

- ค่าความแม่นยำ (Precision) คำนวณจากค่าของข้อมูลที่มีผลลัพธ์ถูกต้องโดยพิจารณาจากจำนวนข้อมูลทั้งหมดที่ถูกจำแนกที่มีผลลัพธ์เดียวกัน สามารถคำนวณได้สมการ (5)

$$\text{Precision} = \frac{\text{จำนวนข้อมูลที่จำแนกได้ถูกต้องของกลุ่มนั้น}}{\text{จำนวนข้อมูลที่ถูกจำแนกว่าเป็นกลุ่มนั้นทั้งหมด}} \quad (5)$$

- ค่าความเหวี่ยง (F-Measure) คำนวณจากค่าความแม่นยำ (p) และค่าความระลึก (r) สามารถคำนวณได้สมการ (6)

$$F(r,p) = \frac{2pr}{p+r} \quad (6)$$

4.3 สร้างกฎความสัมพันธ์ (Generate Rules)

ขั้นตอนนี้จะทำการสร้างกฎความสัมพันธ์ซึ่งได้จากแบบจำลองต้นไม้การตัดสินใจเพื่อหาตัวแปรที่มีความสัมพันธ์กัน และระบุปัจจัยที่มีผลต่อรายได้ของประชากร

5. การแปลผลและการประเมินผล (Evaluation) ขั้นตอนนี้เป็นการประเมินประสิทธิภาพของผลลัพธ์จากโมเดลวิเคราะห์ข้อมูลว่าครอบคลุมและสามารถตอบเป้าหมายและวัตถุประสงค์ที่ตั้งไว้ในขั้นตอนแรกหรือไม่

6. การนำไปใช้ (Deployment) เป็นการนำผลลัพธ์ที่ได้จากการงานวิจัยไปใช้ประโยชน์ต่อไป

3.4 ตัวแปรที่ใช้ในงานวิจัย

การสร้างตัวแบบสำหรับการจำแนกประเภทข้อมูลการสำรวจด้านการทำงานของประชากรนั้นใช้ตัวแปร 14 ตัวแปรดังต่อไปนี้

ตารางที่ 3.3 ตัวแปรทั้งหมดที่จะใช้ในการสร้างแบบจำลองการจำแนกประเภทข้อมูลการสำรวจด้านการทำงานของประชากร

ข้อมูล	ชื่อตัวแปร
ภาค	REG
เขต	AREA
เพศ	SEX
อายุ	AGE
สถานะภาพสมรส	MARITAL
ระดับการศึกษาสูงสุด	EDUCATION
อาชีพ	OCCUP
ประเภทอุตสาหกรรม	INDUSTRY
ประเภทการจ้างงาน	WAGE_TYPE
จำนวนชั่วโมงการทำงานทั้งสิ้น	TOTAL_HR
สถานะภาพการทำงาน	STATUS
ขนาดกิจการของที่ทำงาน	SIZE
รายได้อื่น ๆ	OTHER_INCOME
ช่วงของรายได้	INCOME

- ภาค (REG) แบ่งออกเป็น 5 กลุ่ม คือ
  - กลุ่มที่ 1 คือ กรุงเทพมหานคร
  - กลุ่มที่ 2 คือ ภาคกลาง
  - กลุ่มที่ 3 คือ ภาคเหนือ
  - กลุ่มที่ 4 คือ ภาคตะวันออกเฉียงเหนือ
  - กลุ่มที่ 5 คือ ภาคใต้
- เขต (AREA) แบ่งออกเป็น 2 กลุ่ม คือ
  - กลุ่มที่ 1 คือ เทศบาล
  - กลุ่มที่ 2 คือ นอกเขตเทศบาล
- เพศ (SEX) แบ่งออกเป็น 2 กลุ่ม คือ
  - กลุ่มที่ 1 คือ เพศชาย
  - กลุ่มที่ 2 คือ เพศหญิง

- อายุ (AGE) แบ่งออกเป็น 4 กลุ่ม คือ
  - กลุ่มที่ 1 คือ อายุ 15-24 ปี
  - กลุ่มที่ 2 คือ อายุ 25-34 ปี
  - กลุ่มที่ 3 คือ อายุ 35-49 ปี
  - กลุ่มที่ 4 คือ อายุมากกว่าหรือเท่ากับ 50 ปี
- สถานะภาพสมรส (MARITAL) แบ่งออกเป็น 3 กลุ่ม คือ
  - กลุ่มที่ 1 คือ โสด
  - กลุ่มที่ 2 คือ สมรส
  - กลุ่มที่ 3 คือ อื่น ๆ (แยกกันอยู่ หย่า)
- ระดับการศึกษาสูงสุด (EDUCATION) แบ่งออกเป็น 3 กลุ่ม คือ
  - กลุ่มที่ 1 คือ ต่ำกว่ามัธยมศึกษา
  - กลุ่มที่ 2 คือ ระดับมัธยมศึกษา
  - กลุ่มที่ 3 คือ ระดับอุดมศึกษา
- อาชีพ (OCCUP) แบ่งออกเป็น 4 กลุ่ม คือ
  - กลุ่มที่ 1 คือ ผู้บัญญัติกฎหมาย ข้าราชการระดับอาวุโส และผู้จัดการ
  - กลุ่มที่ 2 คือ ผู้ประกอบวิชาชีพด้านต่างๆ ช่างเทคนิคสาขาต่างๆ และผู้ประกอบวิชาชีพอื่นๆ ที่เกี่ยวข้อง
  - กลุ่มที่ 3 คือ เสมียน พนักงานบริการ พนักงานขายในร้านค้าและตลาด อาชีพขั้นพื้นฐานต่างๆ
  - กลุ่มที่ 4 คือ ผู้ปฏิบัติงานที่มีฝีมือในด้านการเกษตรและการประมง ผู้ปฏิบัติงานในธุรกิจด้านความสามารถทางฝีมือและธุรกิจอื่นๆ ที่เกี่ยวข้อง ผู้ปฏิบัติการเครื่องจักรโรงงานและเครื่องจักร และผู้ปฏิบัติงานด้านการประกอบ
- ประเภทอุตสาหกรรม (INDUSTRY) แบ่งออกเป็น 4 กลุ่ม คือ
  - กลุ่มที่ 1 คือ เกษตร และ ประมง
  - กลุ่มที่ 2 คือ การผลิต
  - กลุ่มที่ 3 คือ การค้า และการบริการ
  - กลุ่มที่ 4 คือ อื่น ๆ (ข้าราชการ การศึกษา องค์กรระหว่างประเทศ เป็นต้น)
- ประเภทการจ้างงาน (WAGE_TYPE) แบ่งออกเป็น 2 กลุ่ม
  - กลุ่มที่ 1 คือ รายชั่วโมง รายวัน หรือรายสัปดาห์
  - กลุ่มที่ 2 คือ รายเดือน

- จำนวนชั่วโมงการทำงานทั้งสิ้น (TOT_HR) แบ่งออกเป็น 3 กลุ่ม
  - กลุ่มที่ 1 คือ 1-34 ชั่วโมงต่อสัปดาห์
  - กลุ่มที่ 2 คือ 35-45 ชั่วโมงต่อสัปดาห์
  - กลุ่มที่ 3 คือ มากกว่า 45 ชั่วโมงต่อสัปดาห์
- สถานะภาพการทำงาน (STATUS) แบ่งออกเป็น 3 กลุ่ม
  - กลุ่มที่ 1 คือ ลูกจ้างรัฐบาล
  - กลุ่มที่ 2 คือ ลูกจ้างรัฐวิสาหกิจ
  - กลุ่มที่ 3 คือ ลูกจ้างเอกชน
- ขนาดกิจการของที่ทำงาน (SIZE) แบ่งออกเป็น 4 กลุ่ม
  - กลุ่มที่ 1 คือ 1-19 คน
  - กลุ่มที่ 2 คือ 20-99 คน
  - กลุ่มที่ 3 คือ ตั้งแต่ 100 คน ขึ้นไป
  - กลุ่มที่ 4 คือ ไม่ได้ระบุ
- รายได้อื่นๆ (OTHER_INCOME) แบ่งออกเป็น 2 กลุ่ม
  - กลุ่มที่ 1 คือ มีรายได้อื่นๆ ประกอบด้วยรายได้ที่มาจาก อาหาร เสื้อผ้าหรือเครื่องแต่งกายอื่น ๆ ที่อยู่อาศัย และอื่น ๆ
  - กลุ่มที่ 2 คือ ไม่มีรายได้อื่นๆ
- ช่วงของรายได้ (INCOME) แบ่งออกเป็น 3 กลุ่ม
  - กลุ่มที่ 1 คือ รายได้น้อย (น้อยกว่า 15,000 บาท ต่อเดือน)
  - กลุ่มที่ 2 คือ รายได้ปานกลาง (15,000-30,000 บาท ต่อเดือน)
  - กลุ่มที่ 3 คือ รายได้มาก (มากกว่า 30,000 บาท ต่อเดือน)

*หมายเหตุ - การแบ่งกลุ่มตัวแปรในงานวิจัยนี้ใช้หลักเกณฑ์ตามสำนักงานสถิติแห่งชาติ เช่น ภาค เขต อายุ เพศ ประเภทการจ้างงาน จำนวนชั่วโมงการทำงาน ขนาดกิจการ รายได้

- ตัวแปรที่ทางคณะผู้วิจัยนำข้อมูลของสำนักงานสถิติแห่งชาติมาปรับใช้ เพื่อให้เหมาะสมกับงานวิจัย ได้แก่ อาชีพ ประเภทอุตสาหกรรม ระดับการศึกษา สถานะภาพสมรส