การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม MODIFIED REGRESSION TREE

นางสาวพรพิมล อุดมมาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

DATA CLASSIFICATION BY MODIFIED REGRESSION TREE ALGORITHM

Miss Pornpimon Udommalai

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Statistics

Department of Statistics

Faculty of Commerce and Accountancy

Chulalongkorn University

Academic Year 2015

หัวข้อวิทยานิพนธ์	การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม MODIFIED REGRESSION TREE
โดย	นางสาวพรพิมล อุดมมาลัย
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(อาจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา)

.....กรรมการ

(อาจารย์ ดร.ณัตติฤดี เจริญรักษ์)

.....กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร.อรุณี กำลัง)

พรพิมล อุดมมาลัย : การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม MODIFIED REGRESSION TREE (DATA CLASSIFICATION BY MODIFIED REGRESSION TREE ALGORITHM) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.สุพล ดุรงค์วัฒนา, 38 หน้า.

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการทำงานของการจำแนกกลุ่มข้อมูลโดยใช้อัลกอริทึม MODIFIED REGRESSION TREE (MRT) ซึ่งอัลกอริทึมนี้ได้ถูกประยุกต์มาจากการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Regression Analysis) และการวิเคราะห์การถดถอยเชิงเส้นพหุ (Multiple Regression Analysis) จะทำการจำลองข้อมูลในแต่ละกรณีโดยใช้โปรแกรม R

ภายใต้ขนาดตัวอย่างจำนวน 200, 600 และ 1,800 จำนวน ตัวแปรอิสระจำนวน 2, 3 และ 4 ตัวแปร และค่าความแปรปรวนของความคลาดเคลื่อนมีขนาด 500, 10,000 และ 40,000 โดยที่มีระดับนัยสำคัญคือ 0.05 และ 0.10 อัลกอริทึมนี้มีกระบวนการคล้ายกับการคัดเลือกแบบไปข้างหน้าและมีขั้นตอนการทำงาน 2 ขั้นตอนคือการคัดเลือกตัวแปรอิสระและการแยก จะคัดเลือกตัวแปรอิสระที่มีค่า p-value น้อยที่สุดจากตัวแปรอิสระทั้งหมด จากนั้นนำมาเปรียบเทียบกับระดับนัยสำคัญที่กำหนดถ้าค่าของ p-value ของตัวแปรอิสระมีค่าน้อยกว่าก็จะนำตัวแปรอิสระตัวนั้นเข้ามาจำแนกกลุ่มโดยใช้ค่าเฉลี่ยเลขคณิตแต่ถ้าค่า p-value ของตัวแปรอิสระมีค่ามากกว่าจะหยุดกระบวนการคัดเลือกตัวแปรอิสระตัวถัดมาภายในกลุ่มนั้นๆ จนกว่าจะไม่มีตัวแปรอิสระใดที่ทำการจำแนกได้แล้วจึงจะหยุดกระบวนการ จากนั้นจะทำการวัดประสิทธิภาพโดยวัดร้อยละความถูกต้อง

จากการศึกษาพบว่าขนาดตัวอย่าง ระดับนัยสำคัญ และจำนวนของตัวแปรอิสระต่างก็ส่งผลให้ร้อยละความถูกต้องมีค่าเพิ่มขึ้นหรือไม่ก็ลดลง ร้อยละความถูกต้องมีแนวโน้มเพิ่มมากขึ้นเมื่อกำหนดขนาดตัวอย่างให้มีจำนวนมากขึ้น แต่ร้อยละความถูกต้องมีแนวโน้มลดลงเมื่อเพิ่มระดับนัยสำคัญและจำนวนของตัวแปรอิสระ ส่วนค่าความแปรปรวนของความคลาดเคลื่อนนั้นไม่ส่งผลต่อร้อยละความถูกต้อง

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2558

5681563226 : MAJOR STATISTICS
KEYWORDS: DATA CLASSIFICATION, MODIFIED REGRESSION TREE ALGORITHM

PORNPIMON UDOMMALAI: DATA CLASSIFICATION BY MODIFIED REGRESSION TREE ALGORITHM. ADVISOR: ASSOC. PROF.SUPOL DURONGWATANA, Ph.D., 38 pp.

This research is aimed at studying the algorithm of classification named as MODIFIED REGRESSION TREE (MRT). The algorithm can be applied for either simple regression model or multiple regression model. The data are simulated under several situations by R free program. Each situation of simulated data depends upon the sample size of each set of data, the number of independent variables, the variance of random error in the regression model, and lastly the level of significance. The algorithm MRT has its procedure almost like the forward selection. There are 2 steps in this algorithm. Those are independent variable selection and splitting steps. These 2 steps combine as one hierarchy of the algorithm. Each independent variable is selected using the least p-value of the simple regression F-test. When the least p-value of the selected independent variable shows the statistical significance to be selected, then the arithmetic mean of that independent variables is used to binary split the data into 2 groups otherwise the algorithm will be stopped. In each of splitting group, the next hierarchy for the rest of independent variables will be classified separately and independently and so on until there is no independent variable to classify or the algorithm is stopped. In the study, the percentage of correct classification is used as the measure how good the algorithm.

The results of the study show that when the number of sample size increases, the percentage of correct classification also increases; when the significance level increases, the percentage of correct classification decreases; when the number of independent variables increases, then the percentage of correct classification decreases; and when the value of variance for random error increases, then the percentage of correct classification is indifferent.

Department: Statistics Student's Signature .....
Field of Study: Statistics Advisor's Signature .....
Academic Year: 2015

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ไม่อาจเสร็จสมบูรณ์ขึ้นมาได้ หากปราศจากความกรุณาและความเอาใจใส่จาก รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา ที่รับเป็นอาจารย์ที่ปรึกษาของผู้วิจัย ท่านได้ให้ข้อมูลและคำแนะนำต่างๆ เพื่อเป็นแนวทางการวางโครงร่าง เขียนเนื้อหาและการวิเคราะห์ข้อมูล ทั้งนี้ท่านยังสละเวลาอันมีค่าเพื่อตรวจสอบความถูกต้องของวิทยานิพนธ์ ซึ่งเป็นประโยชน์ต่อผู้วิจัยเป็นอย่างมาก

ผู้วิจัยขอกราบขอบพระคุณ อาจารย์ ดร.อัครินทร์ ไพบูลย์พานิช ประธานกรรมการวิทยานิพนธ์ อาจารย์ ดร.ณัตติฤดี เจริญรักษ์ และอาจารย์ ดร.อรุณี กำลัง คณะกรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านอาจารย์ทั้งสามท่านได้สละเวลาเพื่อสอบ ตรวจสอบและให้คำแนะนำเพื่อนแก้ไขในส่วนต่างๆ ของวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์มากยิ่งขึ้น นอกจากนี้ผู้วิจัยขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยและคณาจารย์ในระดับปริญญาตรี ที่ให้โอกาสและอบรมสั่งสอนให้ความรู้ทางด้านวิชาการแก่ผู้วิจัยจนสำเร็จการศึกษาในครั้งนี้

สุดท้ายนี้ขอกราบขอบพระคุณบิดามารดาที่คอยให้กำลังใจและความห่วงใยมาโดยตลอด และขอขอบคุณน้องๆ ที่ให้ความช่วยเหลือและคำแนะนำตลอดระยะเวลาที่ศึกษาและจัดทำวิทยานิพนธ์แก่ผู้วิจัยด้วยความเต็มใจเสมอมา

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการวิเคราะห์ทางสถิติที่เกี่ยวข้องกับการสร้างตัวแบบทางคณิตศาสตร์ เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable) ซึ่งมักแทนด้วย y และตัวแปรอิสระ (Independent Variable) ซึ่งมักแทนด้วย x มีวัตถุประสงค์เพื่อการอธิบายความสัมพันธ์ของตัวแปรอิสระที่มีต่อตัวแปรตาม และเพื่อการทำนายและสถิติอนุมานอื่นๆ

การศึกษาความสัมพันธ์ระหว่างตัวแปรและการวิเคราะห์การถดถอย ได้ถูกนำมาประยุกต์ใช้มากมายในหลากหลายสาขา เช่น การศึกษาความสัมพันธ์ระหว่างจำนวนเงินที่ใช้ในการโฆษณาสินค้ากับจำนวนเงินที่ขายสินค้าได้ใน 20 ปีที่ผ่านมาของบริษัทแห่งหนึ่ง ความสัมพันธ์ระหว่างร้อยละของผู้ไปใช้สิทธิลงคะแนนเสียงเลือกตั้งกับปัจจัยทางเศรษฐกิจและสังคมต่างๆ เช่น อายุ ระดับการศึกษาและรายได้เฉลี่ย การศึกษาความสัมพันธ์ระหว่างผลสัมฤทธิ์ของการทำงานกับคะแนนสัมภาษณ์เมื่อมาสมัครเข้าทำงานของฝ่ายพัฒนาและบริหารบุคลาการของบริษัทธุรกิจแห่งหนึ่ง

ปัจจุบันการประกอบธุรกิจของผู้ประกอบการไม่ว่าขนาดเล็ก ขนาดกลาง หรือขนาดใหญ่นั้นย่อมมีการแข่งขันซึ่งกันและกันเพื่อตอบสนองความต้องการของลูกค้า แต่ผู้ประกอบการจะทำอย่างไรจึงจะสามารถเข้าใจและเข้าถึงลูกค้าให้ได้มากที่สุด แม้ในทางปฏิบัตินั้นเราก็ไม่สามารถที่จะเข้าถึงลูกค้าได้ทั้งหมดเพราะลูกค้าแต่ละคนมีทัศนคติและพฤติกรรมในการอุปโภคบริโภคที่แตกต่างกันไป ทางผู้วิจัยจึงเล็งเห็นความสำคัญของการจำแนกกลุ่มลูกค้า โดยจะนำลูกค้าที่มีคุณลักษณะใกล้เคียงกันมารวมกลุ่มไว้ด้วยกัน เมื่อทำการจำแนกกลุ่มลูกค้าได้แล้วอาจทำให้ผู้ประกอบการวางแผนงานทางธุรกิจได้ง่ายขึ้นเพื่อเสนอสินค้าหรือบริการที่ตรงต่อความต้องการของลูกค้าได้ ดังนั้นงานวิจัยฉบับนี้จะทำการศึกษาวิธีการจำแนกกลุ่มข้อมูล (Data Classification) โดยจะใช้การวิเคราะห์การถดถอยซึ่งเป็นวิธีการวิเคราะห์ทางสถิติที่ผู้วิจัยสนใจ ทางผู้วิจัยจึงได้นำเสนอวิธีการที่ใช้ในการจำแนกกลุ่มข้อมูลที่มีชื่อว่า “อัลกอริทึม MODIFIED REGRESSION TREE”

อัลกอริทึม MODIFIED REGRESSION TREE นี้ ผู้วิจัยทำการประยุกต์มาจากอัลกอริทึม CART (Classification and Regression Trees Algorithm) ซึ่งถูกคิดค้นโดย (Breiman, Freidman, Olshen, & Stone, 1984) โดยที่อัลกอริทึม CART จะทำการจำแนกกลุ่มข้อมูลภายใต้เงื่อนไข 2 แบบคือ

1. ในกรณีที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ และตัวแปรอิสระเป็นได้ทั้งตัวแปรเชิงปริมาณและคุณภาพ จะเรียกการจำแนกกลุ่มข้อมูลลักษณะนี้ว่า “Classification Trees Algorithm”

2. ในกรณีที่ตัวแปรตามและตัวแปรอิสระเป็นตัวแปรเชิงปริมาณ จะเรียกการจำแนกกลุ่มข้อมูลลักษณะนี้ว่า “Regression Trees Algorithm”

จากนั้นจะทำการสร้างแผนภาพต้นไม้โดยการแบ่งแบบทวิ (Binary Trees) แล้วทำการจำแนกกลุ่มข้อมูลไปเรื่อยๆ จนถึงกลุ่มสุดท้ายที่ไม่สามารถทำการจำแนกได้อีกหรือเรียกว่า “กลุ่มปลายทาง” ในส่วนของอัลกอริทึม MODIFIED REGRESSION TREE นั้น จะทำการสร้างแผนภาพต้นไม้โดยการแบ่งแบบทวิเช่นเดียวกันแต่จะทำการแบ่งโดยใช้ตัวสถิติทดสอบ F จากการวิเคราะห์การถดถอย ในส่วนข้อมูลที่จะนำมาศึกษานั้นจะต้องเป็นข้อมูลเชิงปริมาณทั้งตัวแปรตามและตัวแปรอิสระ ผู้วิจัยจึงเล็งเห็นความสำคัญของการสร้างแบบจำลองเพื่อช่วยตัดสินใจในการแบ่งกลุ่มข้อมูลและนำผลลัพธ์ที่ได้ไปทดลอง พัฒนาและวางแผนการแบ่งกลุ่มข้อมูลที่ดีในอนาคต

1.2 วัตถุประสงค์

งานวิจัยฉบับนี้มีจุดประสงค์เพื่อศึกษากระบวนการทำงานและเงื่อนไขในการจำแนกกลุ่มข้อมูลโดยใช้อัลกอริทึม MODIFIED REGRESSION TREE ซึ่งประยุกต์มาจากการวิเคราะห์การถดถอย เพื่อทำการสร้างขั้นตอนของการจำแนกข้อมูลให้มีประสิทธิภาพและแม่นยำมากที่สุด

1.3 ข้อตกลงเบื้องต้น

ตัวแบบที่ทำการศึกษาคือ ตัวแบบความถดถอยพหุเชิงเส้น (Multiple Regression Model) ในรูป

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

เมื่อ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรตาม มีขนาด $n \times 1$

$\mathbf{X}$ เป็นเมทริกซ์ของตัวแปรอิสระ มีขนาด $n \times p$

$\boldsymbol{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ของตัวแบบ มีขนาด $p \times 1$

$\boldsymbol{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อนเชิงสุ่ม มีขนาด $n \times 1$

โดยมีข้อสมมติ $\boldsymbol{\varepsilon}$ มีการแจกแจงปกติ มีค่าเฉลี่ย $\mathbf{0}$ และความแปรปรวน $\sigma^2\mathbf{I}_n$ นั่นคือ

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$

ดังนั้น $E(\mathbf{y} \mid \mathbf{X}) = \boldsymbol{\mu}_{y|x} = \mathbf{X}\boldsymbol{\beta}$

1.4 ขอบเขตของการวิจัย

ในการศึกษานี้ผู้วิจัยจะทำการศึกษาภายใต้เงื่อนไขดังต่อไปนี้

1. ศึกษาภายใต้ขนาดตัวอย่าง (n) กับจำนวนตัวแปรอิสระ (p) โดยจะพิจารณาเป็นกรณีต่างๆ ดังนี้

<u>กรณีที่ 1</u> ขนาดตัวอย่างเท่ากับ 200 ($n=200$) จะแยกพิจารณาเป็น

- ขนาดตัวอย่างเท่ากับ 200 กับจำนวนตัวแปรอิสระ 2 ตัว
- ขนาดตัวอย่างเท่ากับ 200 กับจำนวนตัวแปรอิสระ 3 ตัว
- ขนาดตัวอย่างเท่ากับ 200 กับจำนวนตัวแปรอิสระ 4 ตัว

<u>กรณีที่ 2</u> ขนาดตัวอย่างเท่ากับ 600 ($n=600$) จะแยกพิจารณาเป็น

- ขนาดตัวอย่างเท่ากับ 600 กับจำนวนตัวแปรอิสระ 2 ตัว
- ขนาดตัวอย่างเท่ากับ 600 กับจำนวนตัวแปรอิสระ 3 ตัว
- ขนาดตัวอย่างเท่ากับ 600 กับจำนวนตัวแปรอิสระ 4 ตัว

<u>กรณีที่ 3</u> ขนาดตัวอย่างเท่ากับ 1,800 ($n=1800$) จะแยกพิจารณาเป็น

- ขนาดตัวอย่างเท่ากับ 1,800 กับจำนวนตัวแปรอิสระ 2 ตัว
- ขนาดตัวอย่างเท่ากับ 1,800 กับจำนวนตัวแปรอิสระ 3 ตัว
- ขนาดตัวอย่างเท่ากับ 1,800 กับจำนวนตัวแปรอิสระ 4 ตัว

2. ศึกษาภายใต้ความคลาดเคลื่อน (ε_i) โดยการแจกแจงของความคลาดเคลื่อนนั้นมีการแจกแจงแบบปกตินั่นคือ ค่าเฉลี่ยของความคลาดเคลื่อนมีค่าเท่ากับ 0 ($E(\varepsilon_i)=0$) ส่วนค่าของความแปรปรวน (σ^2) จะถูกกำหนดให้มีค่าเท่ากับ 500 10,000 และ 40,000 ($\sigma^2(\varepsilon_i)=500,\ 10000$ และ 40000)

3. ศึกษาภายใต้ค่าสัมประสิทธิ์ (β) ในแต่ละตัว โดยจะพิจารณาเป็นกรณีต่างๆ ดังนี้

<u>กรณีที่ 1</u> จำนวนตัวแปรอิสระ 2 ตัว ค่าสัมประสิทธิ์ (β) ในแต่ละตัวจะมีค่าดังนี้

1.1) $\beta_0=100$ $\beta_1=0$ และ $\beta_2=-100$

1.2) $\beta_0=100$ $\beta_1=100$ และ $\beta_2=0$

<u>กรณีที่ 2</u> จำนวนตัวแปรอิสระ 3 ตัว ค่าสัมประสิทธิ์ (β) ในแต่ละตัวจะมีค่าดังนี้

2.1) $\beta_0=100$ $\beta_1=0$ $\beta_2=-100$ และ $\beta_3=-100$

2.2) $\beta_0=100$ $\beta_1=100$ $\beta_2=0$ และ $\beta_3=-100$

2.3) $\beta_0=100$ $\beta_1=100$ $\beta_2=-100$ และ $\beta_3=0$

2.4) $\beta_0=100$ $\beta_1=0$ $\beta_2=0$ และ $\beta_3=-100$

2.5) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ และ $\beta_3 = 0$

2.6) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ และ $\beta_3 = 0$

<u>กรณีที่ 3</u> จำนวนตัวแปรอิสระ 4 ตัว ค่าสัมประสิทธิ์ (β) ในแต่ละตัวจะมีค่าดังนี้

3.1) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.2) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.3) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.4) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 0$

3.5) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.6) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.7) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 0$

3.8) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.9) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 0$

3.10) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 0$

3.11) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.12) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.13) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.14) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 0$

4. ศึกษาภายใต้ระดับนัยสำคัญ (α) 2 ระดับคือ 0.05 และ 0.10

5. ศึกษาภายใต้การจำลองข้อมูลโดยให้มีสถานการณ์ตามในแต่ละกรณีดังที่กล่าวมาข้างต้น โดยจะทำการจำลองในสถานการณ์ทั้งหมดจำนวน 1,000 รอบ

1.5 คำจำกัดความที่ใช้ในงานวิจัย

การจำแนกกลุ่มข้อมูล (Data Classification)

คือกระบวนการสร้างตัวแบบจำลองเพื่อจำแนกประเภทของข้อมูล โดยมีวัตถุประสงค์เพื่อทำนายกลุ่มของข้อมูลใหม่ โดยข้อมูลที่มีความคล้ายคลึงกันหรือเหมือนกันจะถูกจำแนกให้อยู่ในกลุ่มเดียวกัน

การคัดเลือกตัวแปร (Variable Selection)

คือกระบวนการคัดเลือกตัวแปรอิสระที่เหมาะสมที่สุด เพื่อใช้ในการจำแนกกลุ่มของข้อมูล เพื่อให้ได้กลุ่มของข้อมูลที่ดีที่สุด

อัลกอริทึม MODIFIED REGRESSION TREE

คือกระบวนการที่ใช้วิเคราะห์กลุ่มของข้อมูล จากนั้นจะทำการจำแนกและแก้ปัญหาการจำแนกกลุ่มของข้อมูล ในส่วนของข้อมูลของตัวแปรตามและตัวแปรอิสระต้องเป็นข้อมูลเชิงปริมาณทั้งคู่ จากนั้นจะใช้ตัวแบบความถดถอยพหุเชิงเส้นมาวิเคราะห์เพื่อทำการจำแนกกลุ่มของข้อมูลโดยทำการคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดแล้วจะทำการสร้างแผนภาพต้นไม้แบบทวิโดยใช้ค่าเฉลี่ยของตัวแปรนั้น พร้อมทั้งพิจารณาเกณฑ์การหยุดควบคู่ไปด้วยในแต่ละขั้นตอน จนกระทั่งสิ้นสุดการทำงาน

การวิเคราะห์การถดถอย (Regression Analysis)

คือกระบวนการศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณ โดยจะทำการทดสอบตัวสถิติ F เพื่อพิจารณาค่า $p-value$ ที่มีนัยสำคัญและมีค่าน้อยที่สุด เพื่อคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากที่สุด

เกณฑ์การหยุด (Stopping Rules)

คือเกณฑ์ที่ใช้ในการดำเนินการหยุดการจำแนกกลุ่มของข้อมูลเพื่อไม่ให้กลุ่มปลายทางของการจำแนกมีมากเกินไป

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

$$\text{ร้อยละความถูกต้องของการจำแนก} = \frac{\text{ผลรวมของการจำแนกด้วยตัวแปรอิสระที่ถูกต้อง}}{1{,}000} \times 100\%$$

1.7 วิธีดำเนินงานวิจัย

1. ศึกษาตัวแบบและทฤษฎีของตัวแบบ พร้อมทั้งกระบวนการทำงานในการจำแนกกลุ่มข้อมูลโดยใช้อัลกอริทึม MODIFIED REGRESSION TREE
2. กำหนดและทำการจำลองข้อมูล
   2.1 จำลองข้อมูลของตัวแปรอิสระที่มีการแจกแจงแบบปกติ (Normal Distribution) ภายใต้ตัวแบบความถดถอยพหุเชิงเส้น ในรูป

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

เมื่อ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรตาม มีขนาด $n \times 1$

$\mathbf{X}$ เป็นเมทริกซ์ของตัวแปรอิสระ มีขนาด $n \times p$

$\boldsymbol{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ของตัวแบบ มีขนาด $p \times 1$

$\boldsymbol{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อนเชิงสุ่ม มีขนาด $n \times 1$

โดยมีข้อสมมติ $\boldsymbol{\varepsilon}$ มีการแจกแจงปกติ มีค่าเฉลี่ย $\mathbf{0}$ และความแปรปรวน $\sigma^2\mathbf{I}_n$ นั่นคือ

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$

2.2 กำหนดค่าสัมประสิทธิ์ตามกรณีที่กำหนดไว้ข้างต้นจากนั้นนำข้อมูลที่จำลองได้พร้อมทั้งค่าสัมประสิทธิ์ไปใส่ในตัวแบบ

3. กำหนดค่าระดับนัยสำคัญ (α) เพื่อเอาไว้ทดสอบกับตัวสถิติทดสอบ F ในอัลกอริทึม MODIFIED REGRESSION TREE ในการคัดเลือกตัวแปรอิสระ

4. วิเคราะห์ผลและแก้ปัญหาของแบบจำลองจากการจำแนกกลุ่มข้อมูล

5. สรุปผลของการวิจัยในแต่ละกรณีที่ศึกษา

1.8 ประโยชน์ที่คาดว่าจะได้รับ

เป็นแนวทางหนึ่งในการจำแนกกลุ่มข้อมูลเพื่อให้ผู้ที่สนใจได้นำอัลกอริทึม MODIFIED REGRESSION TREE ไปประยุกต์ใช้ในด้านต่างๆ ตัวอย่างเช่น ผู้ประกอบการนำไปวางแผนทางการตลาดโดยวัตถุประสงค์คือจำแนกกลุ่มลูกค้าเพื่อตอบสนองความต้องการของลูกค้าส่วนใหญ่ได้ดียิ่งขึ้น ฯลฯ

บทที่ 2
ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

ในงานวิจัยฉบับนี้ทำการออกแบบการทดลองเพื่อศึกษาวิธีการจำแนกกลุ่มของข้อมูล (Data Classification) โดยใช้อัลกอริทึม MODIFIED REGRESSION TREE โดยอัลกอริทึมนี้ได้ถูกประยุกต์มาจากการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Regression Analysis) และการวิเคราะห์การถดถอยเชิงเส้นพหุ (Multiple Regression Analysis) ซึ่งในงานวิจัยฉบับนี้มีทฤษฎีที่เกี่ยวข้องดังนี้

2.1 การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย

2.1.1 ตัวแบบถดถอยเชิงเส้นอย่างง่ายและข้อสมมติฐาน

การวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายเป็นการสร้างตัวแบบทางคณิตศาสตร์ เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรอิสระ x เพียงตัวเดียวกับตัวแปรตาม y การศึกษาความสัมพันธ์ระหว่างตัวแปรทั้ง 2 ตัว อาจเขียนแผนภูมิการกระจายเพื่อศึกษาลักษณะความสัมพันธ์ของ x และ y ถ้าความสัมพันธ์ระหว่างตัวแปรอิสระ x และตัวแปรตาม y เป็นแบบเชิงเส้นตัวแบบถดถอยอาจเขียนได้เป็น $y = \beta_0 + \beta_1 x_1 + \varepsilon$ (กัลยา วาณิชย์บัญชา, 2545) หรืออาจเขียนในรูปของค่าสังเกตจากหน่วยตัวอย่างได้เป็น

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ ; \ i = 1, 2, ..., n$$

เมื่อ y_i คือค่าสังเกตของตัวแปรตามจากหน่วยวิเคราะห์ที่ i

x_i คือค่าสังเกตของตัวแปรอิสระจากหน่วยวิเคราะห์ที่ i

β_0 และ β_1 คือพารามิเตอร์ของตัวแบบ

และ ε_i คือความคลาดเคลื่อนเชิงสุ่มในหน่วยวิเคราะห์ที่ i

ข้อสมมติพื้นฐานของตัวแบบถดถอยเชิงเส้นอย่างง่าย

1. ε_i มีการแจกแจงปกติ
2. ε_i มีค่าเฉลี่ยเท่ากับศูนย์
3. ε_i มีความแปรปรวนเท่ากับ σ^2
4. $E(\varepsilon_i \varepsilon_j) = 0$ เมื่อ $i \neq j$

หรืออาจสรุปได้ว่า $\varepsilon_i \sim N(0, \sigma^2)$

2.1.2 การทดสอบสมมติฐานโดยใช้ตารางวิเคราะห์ความแปรปรวน

กำหนดตัวแบบ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ ; \ i = 1,2,...,n$ เมื่อ b_0 และ b_1 เป็นตัวประมาณค่ากำลังสองน้อยที่สุดของพารามิเตอร์ β_0 และ β_1 ตามลำดับ และ $\hat{y}_i = b_0 + b_1 x_i$ เป็นสมการถดถอย แล้วจะได้ว่า

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

ด้วยองศาเสรีเท่ากับ $n-1$

เมื่อ y_i คือค่าสังเกตของหน่วยวิเคราะห์ที่ i

$\hat{y}_i$ คือค่าทำนายของหน่วยวิเคราะห์ที่ i

และ $\bar{y}$ คือค่าเฉลี่ยของค่าสังเกต y

n คือจำนวนค่าสังเกตทั้งหมด

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ คือผลบวกกำลังสองของส่วนเหลือ เขียนแทนด้วย SSE และมีองศาเสรีเท่ากับ $n-2$

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ คือผลบวกกำลังสองทั้งหมด เขียนแทนด้วย SST และมีองศาเสรีเท่ากับ $n-1$

$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ คือผลบวกกำลังสองของการถดถอย เขียนแทนด้วย SSR และมีองศาเสรีเท่ากับ 1

การทดสอบสมมติฐานโดยใช้ตารางวิเคราะห์ความแปรปรวน

การทดสอบสมมติฐาน $H_0 : \beta_1 = 0$ เทียบกับ $H_1 : \beta_1 \neq 0$ อาจใช้การทดสอบ F ได้โดยการสร้างตารางวิเคราะห์ความแปรปรวน

เนื่องจาก $\frac{(n-2)s^2}{\sigma^2}$ มีการแจกแจงไคกำลังสองด้วยองศาเสรี $(n-2)$ และ $\frac{SSR}{E(SSR)}$ มีการแจกแจงไคกำลังสองด้วยองศาเสรี 1 ตัวแปรสุ่ม s^2 และ SSR เป็นอิสระแก่กัน ดังนั้น ภายใต้ $H_0 : \beta_1 = 0$

$$\frac{\frac{SSR}{E(SSR)}}{\frac{(n-2)s^2}{\sigma^2(n-2)}}$$ มีการแจกแจง F ด้วยองศาเสรี $1, \ n-2$

$$E(SSR) = E(MSR) = \sigma^2$$

ดังนั้นจึงทำให้ ตัวสถิติการทดสอบ $F = \frac{MSR}{s^2}$ มีการแจกแจง F ด้วยองศาเสรี $1, \ n-2$

การทดสอบ $H_0 : \beta_1 = 0$ เทียบกับ $H_1 : \beta_1 \neq 0$ จะทำการทดสอบโดยใช้อัตราส่วนระหว่างการถดถอยกำลังสองเฉลี่ยกับส่วนเหลือกำลังสองเฉลี่ย เท่ากับ

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{(n-2)}}$$

$$F = \frac{MSR}{MSE}$$

นั่นคือ จะปฏิเสธ $H_0 : \beta_1 = 0$ ที่ระดับนัยสำคัญ α เมื่อค่าเอฟที่คำนวณได้มีค่ามากกว่าค่าเอฟที่เปิดจากตารางที่องศาเสรีเท่ากับ 1 และ $n-2$ และระดับนัยสำคัญเท่ากับ α

ถ้าค่าของ $\hat{y}_i$ มีค่าเท่ากับ $\bar{y}$ หมายความว่า สมการถดถอยประมาณค่าของ y ได้เท่ากับ $\bar{y}$ ไม่ว่า x จะมีค่าเท่าใดก็ตาม แสดงว่าสมการถดถอยไม่มีประโยชน์ในการประมาณ เนื่องจากตัวแปรอิสระ x ไม่สามารถช่วยในการประมาณค่าของ y ได้ ซึ่งในกรณีนี้จะทำให้ค่าผลบวกกำลังสองของการถดถอยมีค่าเป็นศูนย์และทำให้ $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$ ซึ่งจะทำให้ F มีค่าเป็นศูนย์ สรุปได้ว่า ไม่มีหลักฐานเพียงพอที่จะปฏิเสธสมมติฐานว่าง แต่หากค่าประมาณ $\hat{y}_i$ มีค่าใกล้เคียงกับค่า y_i ที่เป็นค่าที่เกิดขึ้นจริง ทำให้พจน์ที่เป็นตัวส่วนของค่า F มีค่าเป็นศูนย์หรือใกล้เคียงกับศูนย์ ทำให้ค่า F มีค่าสูงมาก เมื่อนำค่า F ที่ได้ไปเปรียบเทียบกับค่า F ที่เปิดจากตารางที่องศาเสรีเท่ากับ 1 และ $n-2$ ที่ระดับนัยสำคัญ α ค่า F ที่คำนวณได้จะมีค่าสูงกว่า นำพาไปสู่การปฏิเสธสมมติฐานว่าง นั่นคือ β_1 มีค่าไม่เท่ากับศูนย์ แสดงว่าตัวแปรอิสระ x สามารถใช้ในการอธิบายตัวแปรตาม y ได้นั่นเอง

2.2 การวิเคราะห์การถดถอยเชิงเส้นพหุ

การวิเคราะห์การถดถอย ตัวแปรตามที่นำมาใช้ในวิเคราะห์ต้องมีมาตรวัดแบบช่วงหรือแบบอัตราส่วน และต้องมีลักษณะเป็นตัวแปรแบบต่อเนื่อง สำหรับตัวแปรอิสระอาจเป็นตัวแปรต่อเนื่องหรือไม่ต่อเนื่อง มีมาตรวัดในระดับใดก็ได้ แต่จะต้องไม่มีสหสัมพันธ์เชิงเส้นในระหว่างตัวแปรอิสระด้วยกันสูงมากนัก มิฉะนั้นจะก่อให้เกิดปัญหาพหุสัมพันธ์เชิงเส้น (Multicollinearity) (สุพล ดุรงค์วัฒนา, 2558)

สมการของตัวแบบการถดถอยเขียนได้ดังนี้

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

$$y = \beta_0 + \sum_{j=1}^{k}\beta_j x_j + \varepsilon \qquad (2.1)$$

เมื่อ y คือตัวแปรสุ่ม

$\beta_0, \beta_1, ..., \beta_k$ คือพารามิเตอร์ของตัวแบบ

$x_1, x_2, ..., x_k$ คือค่าสังเกตของตัวแปรอิสระถือว่าเป็นตัวแปรทางคณิตศาสตร์ และตัวแปรอิสระเหล่านี้ต้องไม่มีสหสัมพันธ์เชิงเส้นที่สมบูรณ์ต่อกัน $r_{x_i x_j} \neq \pm 1$ โดยที่ $i, j = 1, 2, ..., k$

ε คือความคลาดเคลื่อนเชิงสุ่ม

ตัวแบบการถดถอยเชิงเส้นอาจเขียนในรูปของตัวอย่างที่ได้จากการสังเกตของตัวแปรตาม y และตัวแปรอิสระ $x_1, x_2, ..., x_k$ ได้ดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \ ; \ i = 1, 2, ..., n$$

$$y_i = \sum_{j=0}^{k} \beta_j x_{ij} + \varepsilon_i \ ; \ x_{i0} = 1$$

เมื่อ y_i คือค่าสังเกตของตัวแปรตามของหน่วยที่ i

x_{ij} คือค่าสังเกตของตัวแปรอิสระที่ j ของหน่วยที่ i

$\beta_0, \beta_1, ..., \beta_k$ คือพารามิเตอร์ของตัวแบบ

ε_i คือความคลาดเคลื่อนเชิงสุ่มของหน่วยที่ i มีการแจกแจงปกติ มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวน (σ^2) คงตัว

ข้อสมมติพื้นฐานเกี่ยวกับ ε_i ในการวิเคราะห์การถดถอย

1. ε_i มีการแจกแจงปกติ
2. ε_i มีค่าคาดหมาย (expected value) เป็น 0 นั่นคือ $E(\varepsilon_i) = 0$
3. ε_i มีภาวะความแปรปรวนเอกพันธุ์ (homoscedasticity) นั่นคือ $E(\varepsilon_i^2) = \sigma^2$ ซึ่งหมายถึงความแปรปรวนไม่เปลี่ยนแปลงตลอดพิสัยของตัวแปรอิสระ ไม่ว่าค่าของตัวแปรอิสระจะมีค่ามากหรือน้อยก็ตาม
4. ε_i และ ε_j เป็นอิสระต่อกันทำให้มีความแปรปรวนร่วม (covariance) เป็น 0 นั่นคือ $E(\varepsilon_i \varepsilon_j) = 0$ สำหรับ $i \neq j$

การเขียนตัวแบบแสดงความสัมพันธ์ระหว่างตัวแปรตาม y กับตัวแปรอิสระ $x_1, x_2, ..., x_k$ ควรเขียนขึ้นโดยมีทฤษฎีรองรับว่า ตัวแปรอิสระใดบ้างที่มีอิทธิพลต่อตัวแปรตาม y และหากไม่มีทฤษฎีรองรับอาจต้องใช้วิธีการเขียนแผนภูมิการกระจายระหว่างตัวแปรตาม y และตัวแปรอิสระ x ต่างๆ แต่ละตัว แต่ความสัมพันธ์อาจมองยากเนื่องจากแผนภูมิการกระจายเขียนได้ระหว่างตัวแปร 2 ตัว ซึ่งความสัมพันธ์ระหว่างตัวแปรตาม y กับตัวแปรอิสระ $x_1, x_2, ..., x_k$ ที่เขียน ไม่ได้นำตัวแปรอิสระอื่นๆ เข้ามาพิจารณาร่วมด้วย ความสัมพันธ์ระหว่างตัวแปรตาม y และตัวแปรอิสระ $x_1, x_2, ..., x_k$ อาจเปลี่ยนรูปแบบหรือระดับความสัมพันธ์เมื่อมีตัวแปรอิสระตัวอื่นๆ เข้ามาเกี่ยวข้อง ทำให้มีความยากลำบากในการเขียนตัวแบบความสัมพันธ์ระหว่างตัวแปรตาม y และตัวแปรอิสระ

$x_1, x_2, ..., x_k$ อย่างไรก็ตาม แม้ความสัมพันธ์ที่แท้จริงระหว่างตัวแปรเหล่านี้อาจจะไม่ใช่เชิงเส้น แต่ในพิสัยหนึ่งของ x_j ความสัมพันธ์อาจประมาณได้ด้วยความสัมพันธ์เชิงเส้นดังในสมการ (2.3) หากเกิดความไม่แน่ใจว่าตัวแบบที่กำหนดนั้นเหมาะสมกับข้อมูลที่นำมาวิเคราะห์หรือไม่ก็สามารถตรวจสอบได้ในภายหลัง โดยใช้เกณฑ์การคัดเลือกตัวแบบและการตรวจสอบส่วนเหลือที่ได้จากสมการถดถอยที่ประมาณพารามิเตอร์ภายใต้ตัวแบบที่กำหนดมาพิจารณา

ตัวแบบการถดถอยในรูปของเมทริกซ์

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2.2)$$

เมื่อ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรตาม มีขนาด $n \times 1$

$\mathbf{X}$ เป็นเมทริกซ์ของตัวแปรอิสระ มีขนาด $n \times p$

$\boldsymbol{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ของตัวแบบ มีขนาด $p \times 1$

$\boldsymbol{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อนเชิงสุ่ม มีขนาด $n \times 1$

โดยมีข้อสมมติ $\boldsymbol{\varepsilon}$ มีการแจกแจงปกติ มีค่าเฉลี่ย $\mathbf{0}$ และความแปรปรวน $\sigma^2\mathbf{I}_n$ นั่นคือ

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$

ดังนั้น $E(\mathbf{y} \mid \mathbf{X}) = \boldsymbol{\mu}_{y|x} = \mathbf{X}\boldsymbol{\beta}$

2.3 การทดสอบเอฟบางส่วน (Partial F-test)

ในการวิเคราะห์การถดถอย มักจะพบคำถามว่า พจน์บางพจน์ควรอยู่ในตัวแบบหรือไม่ เช่น ข้อมูลชุดหนึ่งมีตัวแปรอิสระจำนวน k ตัว และตัวแปรตาม y และมีตัวแบบการถดถอยอยู่ 2 ตัวแบบคือ

ตัวแบบที่ 1 $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + ... + \beta_q z_q + \beta_{q+1} z_{q+1} + ... + \beta_k z_k + \varepsilon$

ตัวแบบที่ 2 $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + ... + \beta_q z_q + \varepsilon$

และ $z_1, z_2, ..., z_k$ เป็นฟังก์ชันที่ทราบรูปแบบของ $x_1, x_2, ..., x_k$ โดย q มีค่าน้อยกว่า k คำถามคือ พจน์ $\beta_{q+1} z_{q+1} + ... + \beta_k z_k$ ที่เพิ่มเข้าไปในตัวแบบที่ 1 มีความเหมาะสมหรือไม่ที่จะอยู่ในตัวแบบ คำตอบต่อคำถามนี้อาจค้นพบได้เมื่อพิจารณาผลบวกกำลังสองการถดถอยอันเนื่องจากพจน์ดังกล่าว

ให้ $b_0(1), b_1(1), ..., b_k(1)$ เป็นตัวประมาณค่ากำลังสองน้อยที่สุดของพารามิเตอร์ $\beta_0(1), \beta_1(1), ..., \beta_k(1)$ ในตัวแบบที่ 1 และให้ S_1 คือผลบวกกำลังสองของการถดถอยของตัวแบบที่ 1 นั่นคือ $S_1 = SSR(\beta_0(1), \beta_1(1), ..., \beta_k(1))$ ให้ s^2 เป็นกำลังสองเฉลี่ยของส่วนเหลือของตัวแบบที่ 1 ซึ่งเป็นตัวประมาณ σ^2 ในทำนองเดียวกันสมมติว่าตัวแบบที่ 2 ตัวประมาณค่ากำลังสองน้อยที่สุดของพารามิเตอร์ $\beta_0(2), \beta_1(2), ..., \beta_q(2)$ คือ $b_0(2), b_1(2), ..., b_q(2)$ และให้ S_2 คือผลบวกกำลังสองของการถดถอยของตัวแบบที่ 2 นั่นคือ $S_2 = SSR(\beta_0(2), \beta_1(2), ..., \beta_q(2))$ ดังนั้น $S_1 - S_2$ คือ

ผลบวกกำลังสองของการถดถอยที่มีสาระมาจากพจน์ $\beta_{q+1}z_{q+1}+...+\beta_k z_k$ ซึ่งเพิ่มเข้าไปในตัวแบบที่ 2 ให้เป็นตัวแบบที่ 1 ผลบวกกำลังสองของการถดถอยที่เพิ่มขึ้นมีองศาเสรีเท่ากับ $k-q$

ถ้า $\beta_{q+1}=\beta_{q+2}=\beta_{q+3}=...=\beta_k=0$ แล้ว $E((S_1-S_2)/(k-q))=\sigma^2$ และถ้า ε_i มีการแจกแจงปกติ จะได้ว่าผลต่างของ S_1-S_2 จะมีการแจกแจงไคกำลังสองด้วยองศาเสรี $k-q$ และเป็นอิสระกับ s^2 ด้วย ดังนั้นในการทดสอบ

$$H_0:\beta_{q+1}=\beta_{q+2}=...=\beta_k=0$$

ตัวสถิติที่ใช้ในการทดสอบคือ $F=\dfrac{(S_1-S_2)/(k-q)}{s^2}$ ซึ่งมีการแจกแจงเอฟด้วยองศาเสรี $k-q$ และ ν เมื่อ ν คือองศาเสรีของส่วนเหลือของตัวแบบที่ 1 และอาจจะใช้สัญลักษณ์ $SSR(\beta_{q+1},\beta_{q+2},...,\beta_k\mid\beta_0,\beta_1,...,\beta_q)$ แทน S_1-S_2 ซึ่งเป็นผลบวกกำลังสองของการถดถอยที่มีสาระมาจากตัวแปรอิสระ $x_{q+1},x_{q+2},...,x_k$ โดยอาศัยหลักการเดียวกัน สามารถหาผลบวกกำลังสองของพจน์ต่างๆ ได้ เช่น $SSR(\beta_0)$, $SSR(\beta_1\mid\beta_0)$, $SSR(\beta_2\mid\beta_1,\beta_0)$, ..., $SSR(\beta_k\mid\beta_{k-1},...,\beta_1,\beta_0)$ ซึ่งแต่ละพจน์จะมีองศาเสรีเท่ากับ 1 และเป็นอิสระกับ s^2 ด้วย ผลบวกกำลังสองของพจน์ต่างๆ เหล่านี้เรียกว่าผลบวกกำลังสองเชิงลำดับ

ถ้ามีพจน์ในตัวแบบการถดถอยหลายๆ พจน์ อาจพิจารณาการเข้าของพจน์ต่างๆ ในตัวแบบตามลำดับที่ต้องการ ซึ่งหมายถึงการนำเอาตัวแปรอิสระที่สอดคล้องกับพารามิเตอร์เข้าไปในตัวแบบนั่นเอง โดยทั่วไปให้ตัวแปรอิสระ x_j เข้าไปในตัวแบบหลังจากตัวแปรอิสระ $x_1,...,x_{j-1},x_{j+1},...,x_k$ อยู่ในตัวแบบแล้ว จะได้ว่า $SSR(\beta_j\mid\beta_0,\beta_1,...,\beta_{j-1},\beta_{j+1},...,\beta_k)$, $j=1,2,...,k$ เป็นส่วนของผลบวกกำลังสองของการถดถอยที่อธิบายตัวแปรตาม y ด้วยตัวแปรอิสระ x_j เมื่อตัวแปรอิสระอื่นๆ รวมทั้งพจน์ค่าคงตัวอยู่ในตัวแบบ ผลบวกกำลังสองของการถดถอยดังกล่าวมีองศาเสรีเท่ากับ 1 การทดสอบนัยสำคัญของการอธิบาย y เมื่อมีการเพิ่มตัวแปรอิสระ x_j ในตัวแบบจะใช้การทดสอบแบบเอฟดังนี้

$$F=\frac{SSR(\beta_j\mid\beta_0,\beta_1,...,\beta_{j-1},\beta_{j+1},...,\beta_k)}{s^2}$$

ด้วยองศาเสรีเท่ากับ 1 และ $n-k-1$ การทดสอบเอฟในลักษณะนี้เรียกว่า การทดสอบเอฟบางส่วน (Partial F-test) ซึ่งมีประโยชน์ในการพิจารณาเพิ่มหรือตัดพจน์ในตัวแบบการถดถอย อาจเป็นไปได้ว่า การพิจารณาสมการถดถอย ตัวแปรอิสระ x_j มีอิทธิพลต่อตัวแปรตาม y มาก เมื่อ x_j อยู่ในสมการ แต่เมื่อมีตัวแปรอิสระตัวอื่นเข้ามาในสมการอาจทำให้อิทธิพลของตัวแปรอิสระ x_j ที่มีต่อตัวแปรตาม y ลดลงเหลือเพียงเล็กน้อยก็ได้ ทั้งนี้เพราะตัวแปรอิสระ x_j มีความสัมพันธ์กับตัวแปรอิสระอื่นที่เข้ามาในตัวแบบ หรือในทางกลับกัน ตัวแปรอิสระ x_j อาจไม่มีอิทธิพลต่อตัวแปรตาม y หรือมีน้อย แต่เมื่อนำตัวแปรอื่นเข้ามาในตัวแบบ จะทำให้ตัวแปรอิสระ x_j มีอิทธิพลต่อตัวแปรตาม

y เพิ่มขึ้น เนื่องจากตัวแปรอิสระที่นำเข้ามาใหม่กับตัวแปรอิสระ x_j มีความสัมพันธ์กับตัวแปรตาม y เชิงพหุต่อกัน

2.4 การเลือกสมการถดถอยที่เหมาะสม

การทำนายค่าของตัวแปรตามมักจะดีขึ้นถ้าเลือกตัวแปรอิสระที่เหมาะสมได้จำนวนหนึ่ง แต่ในทางปฏิบัติอาจไม่สามารถใช้ตัวแปรอิสระที่เลือกมาได้ทุกตัว เนื่องจากความยากลำบากและค่าใช้จ่ายในการเก็บข้อมูล ฉะนั้นผู้ศึกษาต้องเลือกสมการถดถอยที่เหมาะสมที่สุด ในขอบเขตของความสามารถในการหาข้อมูลและงบประมาณที่มีอยู่ ถึงแม้ว่างบประมาณมีไม่จำกัด และไม่มีความยากลำบากในการเก็บรวบรวมข้อมูลก็ตาม แต่การเลือกสมการถดถอยที่เหมาะสมที่สุด ก็ยังมีความสำคัญต่อการวิเคราะห์การถดถอย เพราะอาจเป็นไปได้ว่าตัวแปรอิสระบางตัวไม่ควรอยู่ในสมการ เนื่องจากตัวแปรอิสระนั้นๆ มีความสัมพันธ์กับตัวแปรอิสระตัวอื่นๆ ที่อยู่ในสมการถดถอยอยู่แล้ว ซึ่งถ้ารวมตัวแปรอิสระนั้นเข้าไปในสมการด้วย ทำให้สมการถดถอยที่ใช้ในการประมาณค่าคาดหมายของตัวแปรตามไม่ดีเท่าที่ควร นอกจากนั้นการทำนายค่าของตัวแปรตามในอนาคต ณ เวลา t ต้องทำนายค่าของตัวแปรอิสระทุกตัวในสมการที่ใช้ในการอธิบายตัวแปรตาม ณ เวลา t หากมีการทำนายตัวแปรอิสระจำนวนมากทำให้ความคลาดเคลื่อนในการทำนายตัวแปรตามสูงขึ้น วิธีการเลือกสมการถดถอยที่เหมาะสมที่สุดหรือดีที่สุดภายใต้ข้อจำกัดบางประการนั้นมี 2 วัตถุประสงค์หลักคือ ประการที่หนึ่ง ต้องการสมการถดถอยที่มีตัวแปรอิสระจำนวนมากที่สุดเท่าที่จะมากได้ เพราะสามารถให้ข้อสนเทศเกี่ยวกับตัวแปรตามได้มากกว่าสมการที่มีตัวแปรอิสระจำนวนน้อยกว่า ประการที่สอง ต้องการเลือกสมการถดถอยที่มีตัวแปรอิสระน้อยที่สุดเท่าที่จะน้อยได้ เพราะความแปรปรวนของค่าทำนายเพิ่มขึ้นเมื่อมีจำนวนตัวแปรอิสระเพิ่มขึ้น นอกจากนี้ค่าใช้จ่ายในการเก็บข้อมูลเพื่อการทำนายจะสูงกว่าการใช้สมการทำนายที่มีตัวแปรอิสระน้อยกว่า และด้วยเหตุผลที่ได้อธิบายแล้วข้างต้น การเลือกตัวแปรตามจึงได้ผสมผสานระหว่างวัตถุประสงค์ทั้ง 2 ข้อให้พบกันครึ่งทาง สรุปได้ว่าวิธีการเลือกสมการถดถอยที่ดีที่สุดหมายถึง การเลือกตัวแปรอิสระเข้าสู่สมการเพื่อใช้ในการอธิบายตัวแปรตาม ซึ่งขึ้นอยู่กับเกณฑ์ที่ใช้ในการคัดเลือกตัวแปร การเลือกสมการที่เหมาะสมนั้นอาจจะมีหลายสมการ วิธีการเลือกตัวแปรอิสระที่นิยมใช้กันอยู่มีหลายวิธี แต่ละวิธีมีข้อดีข้อเสียแตกต่างกัน การเลือกสมการถดถอยที่เหมาะสมแต่ละวิธีของข้อมูลชุดเดียวกันไม่จำเป็นต้องให้ผลลัพธ์เหมือนกัน แต่ในหลายกรณีแต่ละวิธีอาจให้ผลลัพธ์เหมือนกัน ในที่นี่จะนำเสนอวิธีการเลือกตัวแปรอิสระ 5 วิธี ได้แก่

1. การพิจารณาสมการถดถอยที่เป็นไปได้ทั้งหมด (All possible Regression)
2. การคัดเลือกแบบไปข้างหน้า (Forward Selection)
3. การกำจัดแบบถอยหลัง (Backward Elimination)

4. การถดถอยแบบขั้นบันได (Stepwise Regression)
5. การเลือกแบบผสมโดยใช้วิธีที่กล่าวถึงข้างต้น (variations on the previous methods)

ในงานวิจัยฉบับนี้ผู้วิจัยจะทำการคัดเลือกตัวแปรอิสระโดยใช้วิธีการคัดเลือกแบบไปข้างหน้า (Forward Selection)

การคัดเลือกแบบไปข้างหน้า (Forward Selection)

วิธีการคัดเลือกตัวแปรอิสระแบบไปข้างหน้าเป็นการคัดเลือกตัวแปรอิสระเข้าไปในตัวแบบครั้งละ 1 ตัว การนำตัวแปรอิสระเข้าสู่ตัวแบบใช้ความสัมพันธ์ของตัวแปรอิสระกับตัวแปรตามเป็นเกณฑ์ และในแต่ละขั้นตอนของการนำตัวแปรอิสระเข้าสู่สมการจะมีการทดสอบตัวแปรอิสระดังกล่าวหากการทดสอบพารามิเตอร์ของตัวแบบการถดถอยแตกต่างจากศูนย์อย่างมีนัยสำคัญ ตัวแปรอิสระดังกล่าวจะถูกนำเข้าสู่สมการถดถอยอย่างถาวร และทำการพิจารณาตัวแปรอิสระที่เหลือเพื่อจะนำเข้าสู่สมการต่อไป จนกระทั่งตัวแปรอิสระที่เลือกเข้ามาถูกทดสอบและพบว่าไม่มีนัยสำคัญ กระบวนการจะหยุดลงซึ่งสรุปได้เป็นขั้นตอนดังนี้

1. ทดสอบนัยสำคัญ $H_0 : \beta_i = 0$ เทียบกับ $H_1 : \beta_i \neq 0$ ซึ่งจะให้ค่าเอฟสูงสุดในบรรดาตัวแปรอิสระทั้งหมด ถ้าการทดสอบมีนัยสำคัญให้ดำเนินการนำตัวแปรอิสระตัวนั้นเข้ามา. ถ้าการทดสอบไม่มีนัยสำคัญแสดงว่าการถดถอยมีแต่เฉพาะพจน์ของค่าคงตัว แล้วกระบวนการในการคัดเลือกตัวแปรอิสระจะหยุด
2. พิจารณาการทดสอบเอฟบางส่วนของตัวแปรอิสระ x_i ที่เข้ามาใหม่ โดยทดสอบสมมติฐาน $H_0 : \beta_i = 0$ เทียบกับ $H_1 : \beta_i \neq 0$ โดยใช้สถิติเอฟ
   ถ้าค่าเอฟบางส่วนสูงกว่า $F_{\alpha,1,n-m-2}$ ปฏิเสธสมมติฐานว่าง แสดงว่าตัวแปรอิสระ x_i มีความจำเป็นต้องนำเข้ามาในสมการ เมื่อ m เป็นจำนวนตัวแปรอิสระในสมการในแต่ละขั้นตอน และ n เป็นจำนวนค่าสังเกตทั้งหมด
3. กลับไปดำเนินการตามข้อ 1 และ 2 โดยถือว่าสมการได้รวมตัวแปรอิสระไว้ในสมการ 1,2,3,... ตามลำดับ ขั้นตอนต่างๆ จะดำเนินการต่อไปจนกระทั่งการทดสอบเอฟบางส่วนของตัวแปรอิสระที่เข้ามาใหม่มีค่าน้อยกว่า $F_{\alpha,1,n-m-2}$ กระบวนการคัดเลือกตัวแปรอิสระจึงหยุด สมการที่คัดเลือกได้ประกอบด้วยตัวแปรอิสระทุกตัวที่อยู่ในสมการไม่รวมตัวแปรอิสระที่ไม่สามารถปฏิเสธสมมติฐานว่าง

วิธีเลือกตัวแปรอิสระแบบไปข้างหน้าดีกว่าการพิจารณาสมการถดถอยที่เป็นไปได้ทั้งหมดและวิธีการกำจัดตัวแปรอิสระแบบถอยหลัง ในแง่ของการประหยัดเวลาในการคำนวณโดยที่ไม่ต้องพิจารณาตัวแปรอิสระทั้งหมดโดยไม่จำเป็น วิธีนี้มีจุดอ่อน เนื่องจากไม่ได้มีการพิจารณาบทบาทของ

ตัวแปรอิสระที่รวมอยู่ในสมการถดถอยก่อนหน้านี้ เมื่อมีตัวแปรอิสระตัวใหม่เข้ามาในสมการ บทบาทในที่นี้หมายถึงความสามารถของตัวแปรอิสระเดิมในการทำนายตัวแปรตามซึ่งอาจจะเปลี่ยนแปลงได้เมื่อมีตัวแปรอิสระตัวใหม่เข้าไปในสมการถดถอย เนื่องจากความสัมพันธ์ระหว่างตัวแปรอิสระที่นำเข้ามาในสมการกับตัวแปรอิสระที่มีอยู่เดิมในสมการ จุดอ่อนที่กล่าวมานี้สามารถแก้ไขได้ด้วยการเลือกตัวแปรอิสระเข้าสู่สมการถดถอยโดยวิธีการถดถอยแบบขั้นบันไดได้ (จิราวัลย์ จิตรถเวช, 2558)

2.5 เกณฑ์ที่ใช้ในการตัดสินใจ

- กรณีที่มีตัวแปรอิสระ 2 ตัวและตัวแปรตาม 1 ตัว จำนวนกลุ่มปลายทางที่เกิดขึ้นได้มากที่สุดนั้นมีค่าเท่ากับ 2^p = 4 กลุ่มปลายทาง จากกรณีที่ศึกษานั้นจะกำหนดให้สัมประสิทธิ์การถดถอยของตัวแปรอิสระตัวใดตัวหนึ่งมีค่าเท่ากับศูนย์ โดยการวัดร้อยละความถูกต้องเพื่อตรวจสอบประสิทธิภาพของอัลกอริทึมนั้นจะทำการนับจำนวนครั้งที่ในรอบนั้นๆ มีการแยกเพียงแค่ 0 หรือ 2 กลุ่มปลายทาง จากการทำซ้ำจำนวน 1,000 รอบ
- กรณีที่มีตัวแปรอิสระ 3 ตัวและตัวแปรตาม 1 ตัว จำนวนกลุ่มปลายทางที่เกิดขึ้นได้มากที่สุดนั้นมีค่าเท่ากับ 2^p = 8 กลุ่มปลายทาง ในงานวิจัยฉบับนี้จะแบ่งกรณีศึกษาเป็น 2 กรณีดังนี้
    I. กำหนดสัมประสิทธิ์การถดถอยของตัวแปรอิสระตัวใดตัวหนึ่งมีค่าเท่ากับศูนย์ ตัวแปรอิสระที่เหลืออีกสองตัวนั้นก็ควรจะเข้าไปอยู่ในกระบวนการจำแนก จากนั้นจะวัดร้อยละความถูกต้องเพื่อตรวจสอบประสิทธิภาพของอัลกอริทึมจากการนับจำนวนครั้งที่ในรอบนั้นๆ มีการแยกตั้งแต่ 0, 2, 3 หรือ 4 กลุ่มปลายทาง จากการทำซ้ำจำนวน 1,000 รอบ
    II. กำหนดสัมประสิทธิ์การถดถอยของตัวแปรอิสระมีค่าเป็นศูนย์จำนวน 2 ตัวแปร ตัวแปรอิสระที่เหลืออีกตัวหนึ่งนั้นก็ควรจะเข้าไปอยู่ในกระบวนการจำแนก จากนั้นจะวัดร้อยละความถูกต้องเพื่อตรวจสอบประสิทธิภาพของอัลกอริทึมจากการนับจำนวนครั้งที่ในรอบนั้นๆ มีการแยกตั้งแต่ 0 หรือ 2 กลุ่มปลายทาง จากการทำซ้ำจำนวน 1,000 รอบ
- ในกรณีที่มีตัวแปรอิสระ 4 ตัวและตัวแปรตาม 1 ตัว จำนวนกลุ่มปลายทางที่เกิดขึ้นได้มากที่สุดนั้นมีค่าเท่ากับ 2^p = 16 กลุ่มปลายทาง ในงานวิจัยฉบับนี้จะแบ่งกรณีศึกษาเป็น 3 กรณีดังนี้

I. กำหนดสัมประสิทธิ์การถดถอยของตัวแปรอิสระตัวใดตัวหนึ่งมีค่าเท่ากับศูนย์ ตัวแปรอิสระที่เหลืออีกสามตัวนั้นก็ควรจะเข้าไปอยู่ในกระบวนการจำแนก จากนั้นจะวัดร้อยละความถูกต้องเพื่อตรวจสอบประสิทธิภาพของอัลกอริทึมจากการนับจำนวนครั้งที่ในรอบนั้นๆ มีการแยกตั้งแต่ 0, 2, 3, 4, 5, 6, 7 หรือ 8 กลุ่มปลายทาง จากการทำซ้ำจำนวน 1,000 รอบ

II. กำหนดสัมประสิทธิ์การถดถอยของตัวแปรอิสระมีค่าเป็นศูนย์จำนวน 2 ตัวแปร ตัวแปรอิสระที่เหลืออีกสองตัวนั้นก็ควรจะเข้าไปอยู่ในกระบวนการจำแนก จากนั้นจะวัดร้อยละความถูกต้องเพื่อตรวจสอบประสิทธิภาพของอัลกอริทึมจากการนับจำนวนครั้งที่ในรอบนั้นๆ มีการแยกตั้งแต่ 0, 2, 3 หรือ 4 กลุ่มปลายทาง จากการทำซ้ำจำนวน 1,000 รอบ

III. กำหนดสัมประสิทธิ์การถดถอยของตัวแปรอิสระมีค่าเป็นศูนย์จำนวน 3 ตัวแปร ตัวแปรอิสระที่เหลืออีกหนึ่งตัวนั้นก็ควรจะเข้าไปอยู่ในกระบวนการจำแนก จากนั้นจะวัดร้อยละความถูกต้องเพื่อตรวจสอบประสิทธิภาพของอัลกอริทึมจากการนับจำนวนครั้งที่ในรอบนั้นๆ มีการแยกตั้งแต่ 0 หรือ 2 กลุ่มปลายทาง จากการทำซ้ำจำนวน 1,000 รอบ

2.6 อัลกอริทึม MODIFIED REGRESSION TREE

ในงานวิจัยฉบับนี้จะสนใจการจำแนกกลุ่มข้อมูลเพื่อนำไปประยุกต์ใช้กับข้อมูลทางการตลาดในการจำแนกกลุ่มลูกค้า โดยที่อัลกอริทึม MODIFIED REGRESSION TREE นี้ถูกดัดแปลงมาจากอัลกอริทึม CART เนื่องจากวัตถุประสงค์ของอัลกอริทึม CART นั้นก็ใช้จำแนกกลุ่มข้อมูล แต่อัลกอริทึม CART มีความซับซ้อนและมีขั้นตอนมากกว่าอัลกอริทึม MODIFIED REGRESSION TREE ผู้วิจัยจึงใช้การวิเคราะห์การถดถอยเชิงเส้นมาหาความสัมพันธ์ของตัวแปรอิสระและตัวแปรตามเพื่อใช้ในการจัดกลุ่มข้อมูลโดยที่ตัวแปรอิสระและตัวแปรตามนั้นต้องเป็นตัวแปรเชิงปริมาณทั้งคู่ จะมีขั้นตอนการทำงานดังนี้

1. คัดเลือกตัวแปรอิสระตัวที่มีค่าของตัวสถิติทดสอบ F ในการวิเคราะห์การถดถอยเชิงเส้นมากที่สุดหรือพิจารณาตัวแปรอิสระที่มีค่า $p-value$ ของตัวสถิติทดสอบ F น้อยที่สุด
2. เปรียบเทียบค่า $p-value$ ของตัวสถิติทดสอบ F ว่ามีค่าน้อยกว่าระดับนัยสำคัญหรือไม่ ถ้ามีค่าน้อยกว่าระดับนัยสำคัญให้นำตัวแปรอิสระนั้นมาใช้ในการจำแนกกลุ่มข้อมูล ถ้ามีค่ามากกว่าระดับนัยสำคัญก็จะหยุดกระบวนการจำแนกข้อมูล

3. เมื่อนำตัวแปรอิสระที่ได้รับการคัดเลือกเข้ามาจะทำการจำแนกกลุ่มข้อมูลโดยพิจารณาจากค่าเฉลี่ยเลขคณิตของข้อมูลในตัวแปรอิสระนั้นๆ เป็นเกณฑ์ในการจำแนก ในกรณีที่มีข้อมูลบางตัวนั้นมีค่าเท่ากับค่าเฉลี่ยเลขคณิตที่คำนวณได้ จะนำข้อมูลชุดนั้นไปรวมไว้ในกลุ่มของข้อมูลที่มีค่าน้อยกว่าค่าเฉลี่ยเลขคณิต
4. ทำการคัดเลือกตัวแปรอิสระตัวถัดมาโดยจะพิจารณาเฉพาะชุดของข้อมูลในกลุ่มนั้นๆ กลับไปข้อ 1. ข้อ 2. และ ข้อ 3.
5. จำแนกกลุ่มข้อมูลเช่นนี้ไปเรื่อยๆ และจะหยุดกระบวนการจำแนกกลุ่มข้อมูลก็ต่อเมื่อค่า $p-value$ ของตัวแปรอิสระของชุดข้อมูลนั้นๆ มีค่ามากกว่าระดับนัยสำคัญที่กำหนด

บทที่ 3
วิธีดำเนินการวิจัย

จุดประสงค์ของงานวิจัยฉบับนี้คือการจำแนกกลุ่มข้อมูล (Data Classification) ผู้วิจัยจะใช้หลักการของการวิเคราะห์การถดถอยมาประยุกต์ใช้ในอัลกอริทึม MODIFIED REGRESSION TREE โดยที่ตัวแปรตามและตัวแปรอิสระต้องเป็นตัวแปรเชิงปริมาณต่อมาจะทำการจำลองข้อมูลที่มีการแจกแจงแบบปกติ (Normal Distribution) ตามกรณีต่างๆ ที่กำหนดไว้ จากนั้นทำการกำหนดระดับนัยสำคัญเพื่อใช้ในการคัดเลือกตัวแปรอิสระที่จะนำมาจำแนกกลุ่มข้อมูล ผู้วิจัยจะทำการจำลองและวิเคราะห์ข้อมูลทั้งหมดโดยใช้โปรแกรม R เวอร์ชั่น 3.3.0 ภายใต้ขอบเขตและกรณีต่างๆ โดยมีวิธีการดำเนินการดังต่อไปนี้นี้

3.1 ขั้นตอนในการดำเนินการวิจัย

ในงานวิจัยฉบับนี้มีขั้นตอนในการดำเนินการวิจัยดังต่อไปนี้

3.1.1 ศึกษาตัวแบบและทฤษฎีของตัวแบบ พร้อมทั้งกระบวนการทำงานในการจำแนกกลุ่มข้อมูลโดยใช้อัลกอริทึม MODIFIED REGRESSION TREE

3.1.2 กำหนดและทำการจำลองข้อมูล

- จำลองข้อมูลของตัวแปรอิสระที่มีการแจกแจงแบบปกติ (Normal Distribution) ภายใต้ตัวแบบความถดถอยพหุเชิงเส้น ในรูป

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

เมื่อ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรตาม มีขนาด $n \times 1$

$\mathbf{X}$ เป็นเมทริกซ์ของตัวแปรอิสระ มีขนาด $n \times p$

$\boldsymbol{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ของตัวแบบ มีขนาด $p \times 1$

$\boldsymbol{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อนเชิงสุ่ม มีขนาด $n \times 1$

โดยมีข้อสมมติ $\boldsymbol{\varepsilon}$ มีการแจกแจงปกติ มีค่าเฉลี่ย $\mathbf{0}$ และความแปรปรวน $\sigma^2\mathbf{I}_n$ นั่นคือ $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$

- ศึกษาภายใต้ความคลาดเคลื่อน (ε_i) โดยการแจกแจงของความคลาดเคลื่อนนั้นมีการแจกแจงแบบปกตินั่นคือ ค่าเฉลี่ยของความคลาดเคลื่อนมีค่าเท่ากับ 0 ($E(\varepsilon_i) = 0$) ส่วนค่าของความแปรปรวน (σ^2) จะถูกกำหนดให้มีค่าเท่ากับ 500 10,000 และ 40,000 ($\sigma^2(\varepsilon_i) = 500,\ 10000,\ 40000$)

3.1.3 ศึกษาภายใต้ขนาดตัวอย่าง (n) กับจำนวนตัวแปรอิสระ (p) โดยจะพิจารณาเป็นกรณีต่างๆ ดังนี้

<u>กรณีที่ 1</u> ขนาดตัวอย่างเท่ากับ 200 ($n=200$) จะแยกพิจารณาเป็น

1.1) ขนาดตัวอย่างเท่ากับ 200 กับจำนวนตัวแปรอิสระ 2 ตัว

1.2) ขนาดตัวอย่างเท่ากับ 200 กับจำนวนตัวแปรอิสระ 3 ตัว

1.3) ขนาดตัวอย่างเท่ากับ 200 กับจำนวนตัวแปรอิสระ 4 ตัว

<u>กรณีที่ 2</u> ขนาดตัวอย่างเท่ากับ 600 ($n=600$) จะแยกพิจารณาเป็น

2.1) ขนาดตัวอย่างเท่ากับ 600 กับจำนวนตัวแปรอิสระ 2 ตัว

2.2) ขนาดตัวอย่างเท่ากับ 600 กับจำนวนตัวแปรอิสระ 3 ตัว

2.3) ขนาดตัวอย่างเท่ากับ 600 กับจำนวนตัวแปรอิสระ 4 ตัว

<u>กรณีที่ 3</u> ขนาดตัวอย่างเท่ากับ 1,800 ($n=1800$) จะแยกพิจารณาเป็น

3.1) ขนาดตัวอย่างเท่ากับ 1,800 กับจำนวนตัวแปรอิสระ 2 ตัว

3.2) ขนาดตัวอย่างเท่ากับ 1,800 กับจำนวนตัวแปรอิสระ 3 ตัว

3.3) ขนาดตัวอย่างเท่ากับ 1,800 กับจำนวนตัวแปรอิสระ 4 ตัว

3.1.4 กำหนดค่าสัมประสิทธิ์การถดถอยตามกรณีต่างๆ ดังนี้

<u>กรณีที่ 1</u> จำนวนตัวแปรอิสระ 2 ตัว โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์เพียงหนึ่งตัวในตัวแบบ จะแยกเป็นกรณีศึกษาทั้งหมด 2 กรณี ดังนี้

1.1) $\beta_0=100$ $\beta_1=0$ และ $\beta_2=-100$

1.2) $\beta_0=100$ $\beta_1=100$ และ $\beta_2=0$

<u>กรณีที่ 2</u> จำนวนตัวแปรอิสระ 3 ตัว โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์ภายใต้เงื่อนไขดังนี้

กรณีที่ 2.1) กำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์เพียงหนึ่งตัวในตัวแบบ จะแยกเป็นกรณีศึกษาทั้งหมด 3 กรณี ดังนี้

2.1.1) $\beta_0=100$ $\beta_1=0$ $\beta_2=-100$ และ $\beta_3=-100$

2.1.2) $\beta_0=100$ $\beta_1=100$ $\beta_2=0$ และ $\beta_3=-100$

2.1.3) $\beta_0=100$ $\beta_1=100$ $\beta_2=-100$ และ $\beta_3=0$

กรณีที่ 2.2) กำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์จำนวน 2 ตัวในตัวแบบ จะแยกเป็นกรณีศึกษาทั้งหมด 3 กรณี ดังนี้

2.2.1) $\beta_0=100$ $\beta_1=0$ $\beta_2=0$ และ $\beta_3=-100$

2.2.2) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ และ $\beta_3 = 0$

2.2.3) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ และ $\beta_3 = 0$

<u>กรณีที่ 3</u> จำนวนตัวแปรอิสระ 4 ตัว โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์ภายใต้เงื่อนไขดังนี้

กรณีที่ 3.1) กำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์เพียงหนึ่งตัวในตัวแบบ จะแยกเป็นกรณีศึกษาทั้งหมด 4 กรณี ดังนี้

3.1.1) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.1.2) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.1.3) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.1.4) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 0$

กรณีที่ 3.2) กำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์จำนวน 2 ตัวในตัวแบบ จะแยกเป็นกรณีศึกษาทั้งหมด 6 กรณี ดังนี้

3.2.1) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 100$

3.2.2) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.2.3) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = -100$ และ $\beta_4 = 0$

3.2.4) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.2.5) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 0$

3.2.6) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 0$

กรณีที่ 3.3) กำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์จำนวน 3 ตัวในตัวแบบ จะแยกเป็นกรณีศึกษาทั้งหมด 4 กรณี ดังนี้

3.3.1) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = 0$ $\beta_3 = 0$ และ $\beta_4 = 100$

3.3.2) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = 0$ $\beta_3 = -100$ และ $\beta_4 = 0$

3.3.3) $\beta_0 = 100$ $\beta_1 = 0$ $\beta_2 = -100$ $\beta_3 = 0$ และ $\beta_4 = 0$

3.3.4) $\beta_0 = 100$ $\beta_1 = 100$ $\beta_2 = 0$ $\beta_3 = 0$ และ $\beta_4 = 0$

3.1.5 นำข้อมูลของตัวแปรอิสระที่เกิดจากการจำลองพร้อมทั้งค่าสัมประสิทธิ์ไปใส่ในตัวแบบความถดถอยพหุเชิงเส้น

3.1.6 กำหนดค่าระดับนัยสำคัญ (α) เพื่อเอาไว้ทดสอบกับตัวสถิติทดสอบ F ในอัลกอริทึม MODIFIED REGRESSION TREE ในการคัดเลือกตัวแปรอิสระ

3.1.7 ทำการคัดเลือกตัวแปรอิสระตัวที่มีค่าของตัวสถิติทดสอบ F มากที่สุดหรือพิจารณาตัวแปรอิสระที่มีค่า $p-value$ ของตัวสถิติทดสอบ F น้อยที่สุด

3.1.8 นำตัวแปรอิสระที่ได้จาก 3.1.7 มาเปรียบเทียบกับระดับนัยสำคัญ ถ้าค่า $p-value$

ของตัวสถิติทดสอบ F มีค่าน้อยกว่าค่าระดับนัยสำคัญจะใช้ตัวแปรอิสระตัวนั้นในการจำแนกข้อมูล

3.1.9 ทำการจำแนกข้อมูลโดยสร้างกลุ่มแรกจากการหาค่าเฉลี่ยเลขคณิต (mean) ของตัวแปรอิสระที่ได้จาก 3.1.8

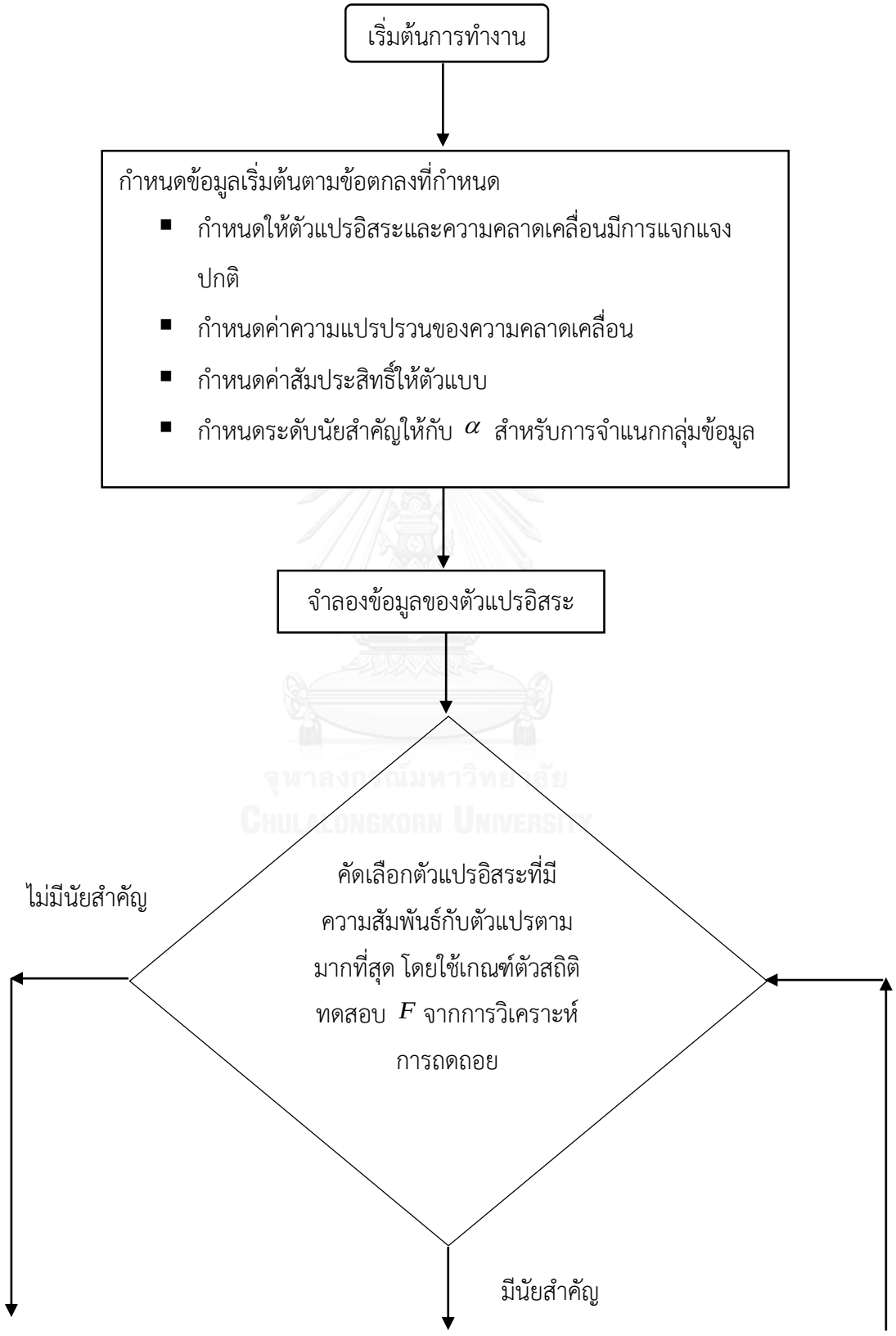
3.1.10 ทำการสร้างกลุ่มต่อๆ ไป โดยพิจารณาเฉพาะชุดข้อมูลที่เกี่ยวข้องภายในกลุ่มนั้นๆ กลับไปที่ 3.1.7

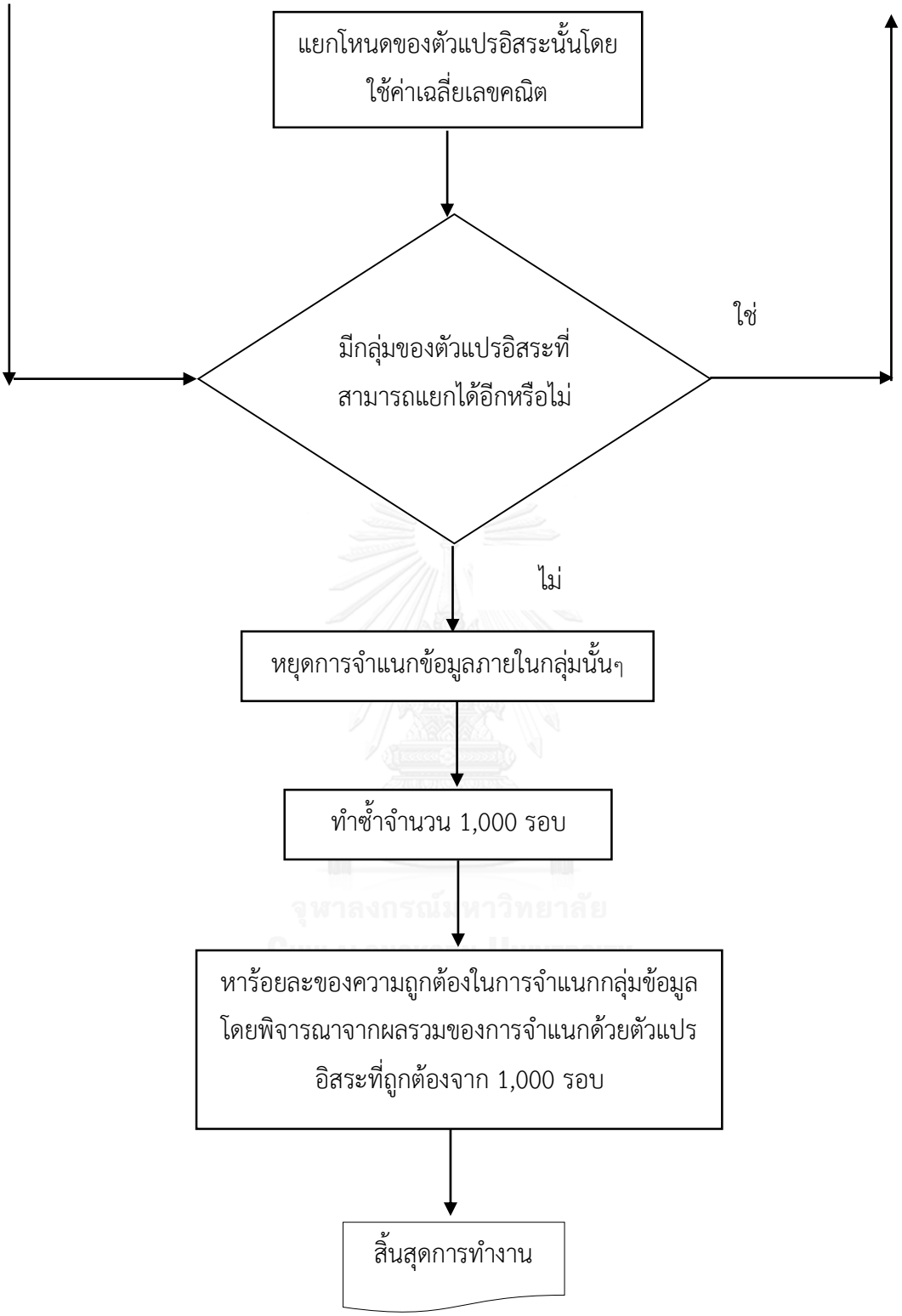
3.1.11 จะทำการหยุดกระบวนการจำแนกกลุ่มข้อมูลก็ต่อเมื่อค่า $p-value$ ของตัวสถิติทดสอบ F มีค่ามากกว่าค่าระดับนัยสำคัญ

3.1.12 หาค่าเฉลี่ยเลขคณิตของตัวแปรอิสระและตัวแปรตามที่ได้จากกลุ่มปลายทางนั้น เพื่อสรุปผลของการจำแนกกลุ่มของข้อมูล

3.1.13 ทำการทดลองซ้ำจำนวน 1,000 รอบ เพื่อหาร้อยละความถูกต้องของอัลกอริทึม

3.2 ขั้นตอนการทำงานของโปรแกรม




แยกโหนดของตัวแปรอิสระนั้นโดย
ใช้ค่าเฉลี่ยเลขคณิต
มีกลุ่มของตัวแปรอิสระที่
สามารถแยกได้อีกหรือไม่
ใช่
ไม่
หยุดการจำแนกข้อมูลภายในกลุ่มนั้นๆ
ทำซ้ำจำนวน 1,000 รอบ
หาร้อยละของความถูกต้องในการจำแนกกลุ่มข้อมูล
โดยพิจารณาจากผลรวมของการจำแนกด้วยตัวแปร
อิสระที่ถูกต้องจาก 1,000 รอบ
สิ้นสุดการทำงาน

บทที่ 4
ผลการวิจัย

งานวิจัยฉบับนี้มีวัตถุประสงค์เปรียบเทียบประสิทธิภาพของการจำแนกกลุ่มข้อมูลของอัลกอริทึม MODIFIED REGRESSION TREE ระหว่างตัวแปรตามจำนวน 1 ตัวกับตัวแปรอิสระจำนวน 2 ตัว 3 ตัวและ 4 ตัว เริ่มจากการจำลองข้อมูลเท่ากับขนาดตัวอย่าง พร้อมทั้งกำหนดระดับนัยสำคัญและค่าความแปรปรวนของความคลาดเคลื่อนภายใต้กรณีที่ศึกษา จากนั้นจะวัดประสิทธิภาพของอัลกอริทึมโดยพิจารณาจากร้อยละของความถูกต้อง และนำเสนอผลของการวิจัยในรูปแบบตารางดังต่อไปนี้

4.1 ผลการวิจัย

ตารางที่ 1 แสดงค่าร้อยละของความถูกต้องในการจำแนกข้อมูลภายใต้ตัวแบบถดถอย $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยในตัวแบบมีค่าเท่ากับศูนย์เพียงหนึ่งตัวในตัวแบบ จะแยกเป็นกรณีศึกษาได้ทั้งหมด 2 กรณี

กรณีศึกษา	ขนาดตัวอย่าง	ระดับนัยสำคัญ 0.05			ระดับนัยสำคัญ 0.1		
		ความแปรปรวน (σ_ε^2)			ความแปรปรวน (σ_ε^2)		
		500	10000	40000	500	10000	40000
1) $\beta_1 = 0$	200	89.1%	89%	89.6%	81.2%	81.8%	81.5%
	600	90.1%	90.2%	90%	81.9%	82.2%	82.3%
	1800	90.9%	91.1%	91.3%	82.3%	82.4%	82.7%
2) $\beta_2 = 0$	200	89%	88.1%	89.7%	80.1%	79.1%	80.8%
	600	90.2%	90.7%	90%	80.9%	80.2%	82.3%
	1800	90.9%	91%	90.2%	81.5%	81.4%	82.4%

จากตารางที่ 4.1 สรุปได้ว่าเมื่อเพิ่มขนาดตัวอย่างจะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มเพิ่มขึ้น แต่ในทางกลับกันถ้าเพิ่มระดับนัยสำคัญก็จะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มลดลง ส่วนค่าความแปรปรวนของค่าความคลาดเคลื่อนเชิงสุ่มนั้นจะมีค่าน้อยหรือมากก็ไม่ส่งผลให้ร้อยละของความถูกต้องเพิ่มขึ้นหรือลดลงแต่อย่างใด

ตารางที่ 2 แสดงค่าร้อยละของความถูกต้องในการจำแนกข้อมูลภายใต้ตัวแบบถดถอย $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับศูนย์จำนวน 1 ตัว แยกเป็นกรณีศึกษาได้ทั้งหมด 3 กรณี

กรณีศึกษา	ขนาดตัวอย่าง	ระดับนัยสำคัญ 0.05			ระดับนัยสำคัญ 0.1		
		ความแปรปรวน (σ_ε^2)			ความแปรปรวน (σ_ε^2)		
		500	10000	40000	500	10000	40000
1) $\beta_1 = 0$	200	80.5%	79.8%	80.8%	80.1%	79.2%	79%
	600	81.1%	80.3%	80.9%	80.2%	80.1%	80%
	1800	82.3%	81.9%	81.5%	81%	81%	80.1%
2) $\beta_2 = 0$	200	80.2%	80%	79.7%	79.7%	79.3%	79.5%
	600	80.7%	82%	81.5%	79.9%	81%	80.2%
	1800	81.3%	83.7%	82.6%	80.7%	81.8%	81.6%
3) $\beta_3 = 0$	200	79.9%	80%	80.1%	79.5%	79.6%	79.6%
	600	80.2%	81.6%	81.6%	79.9%	81.3%	80%
	1800	81.5%	83%	82.3%	80.6%	82.1%	81%

จากตารางที่ 4.2 สรุปได้ว่าเมื่อเพิ่มขนาดตัวอย่างจะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มเพิ่มขึ้น แต่ในทางกลับกันถ้าเพิ่มระดับนัยสำคัญก็จะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มลดลง ส่วนค่าความแปรปรวนของค่าความคลาดเคลื่อนเชิงสุ่มนั้นจะมีค่าน้อยหรือมากก็ไม่ส่งผลให้ร้อยละของความถูกต้องเพิ่มขึ้นหรือลดลงแต่อย่างใด

ตารางที่ 3 แสดงค่าร้อยละของความถูกต้องในการจำแนกข้อมูลภายใต้ตัวแบบถดถอย $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับศูนย์จำนวนสองตัวในตัวแบบถดถอย แยกเป็นกรณีศึกษาได้ทั้งหมด 3 กรณี

กรณีศึกษา	ขนาดตัวอย่าง	ระดับนัยสำคัญ 0.05			ระดับนัยสำคัญ 0.1		
		ความแปรปรวน (σ_ε^2)			ความแปรปรวน (σ_ε^2)		
		500	10000	40000	500	10000	40000
1) $\beta_1 = \beta_2 = 0$	200	80.5%	80.2%	80.9%	78.9%	78.7%	79%
	600	80.9%	80.5%	81.1%	79.5%	79%	80.4%
	1800	81.7%	82.4%	82.7%	80.8%	80.1%	81.6%
2) $\beta_1 = \beta_3 = 0$	200	80.1%	79.9%	80%	79.1%	77.9%	78%
	600	80.9%	80.9%	80.5%	79.5%	78.9%	79.5%
	1800	81.5%	82%	81.9%	80.2%	80.1%	79.9%
3) $\beta_2 = \beta_3 = 0$	200	79.9%	80.4%	79.8%	79.1%	79.1%	78.9%
	600	80.5%	81.9%	80.8%	79.5%	79.9%	79.2%
	1800	81.6%	82.2%	81.6%	80.4%	81.1%	80.1%

จากตารางที่ 4.3 สรุปได้ว่าเมื่อเพิ่มขนาดตัวอย่างจะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มเพิ่มขึ้น แต่ในทางกลับกันถ้าเพิ่มระดับนัยสำคัญก็จะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มลดลง ส่วนค่าความแปรปรวนของค่าความคลาดเคลื่อนเชิงสุ่มนั้นจะมีค่าน้อยหรือมากก็ไม่ส่งผลให้ร้อยละของความถูกต้องเพิ่มขึ้นหรือลดลงแต่อย่างใด

ตารางที่ 4 แสดงค่าร้อยละของความถูกต้องในการจำแนกข้อมูลภายใต้ตัวแบบถดถอย $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$ โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับศูนย์เพียงตัวเดียวในตัวแบบถดถอย แยกเป็นกรณีศึกษาได้ทั้งหมด 4 กรณี

กรณีศึกษา	ขนาดตัวอย่าง	ระดับนัยสำคัญ 0.05			ระดับนัยสำคัญ 0.1		
		ความแปรปรวน (σ_ε^2)			ความแปรปรวน (σ_ε^2)		
		500	10000	40000	500	10000	40000
1) $\beta_1 = 0$	200	63.5%	63.1%	63.9%	62.5%	61.3%	62.7%
	600	64%	64.8%	66.8%	62.9%	62.7%	64.5%
	1800	64.7%	69.4%	67.1%	63.4%	67.5%	65.9%
2) $\beta_2 = 0$	200	63.7%	64.3%	63.1%	62.8%	62.9%	63%
	600	64.2%	65.5%	64%	63.1%	63.4%	63%
	1800	65.6%	67.9%	65.7%	63.9%	64.7%	62.9%
3) $\beta_3 = 0$	200	64.5%	64.7%	63.5%	63.3%	62.5%	62.1%
	600	64.9%	65.1%	64.7%	63.9%	63.7%	64.1%
	1800	66.2%	68.5%	67%	64.6%	65.2%	65.8%
4) $\beta_4 = 0$	200	63.9%	63.3%	64%	62.7%	62.5%	62.2%
	600	64.5%	64.5%	64.9%	63.8%	63.1%	63.2%
	1800	65.7%	67.2%	68%	64.9%	64%	65%

จากตารางที่ 4.4 สรุปได้ว่าเมื่อเพิ่มขนาดตัวอย่างจะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มเพิ่มขึ้น แต่ในทางกลับกันถ้าเพิ่มระดับนัยสำคัญก็จะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มลดลง ส่วนค่าความแปรปรวนของค่าความคลาดเคลื่อนเชิงสุ่มนั้นจะมีค่าน้อยหรือมากก็ไม่ส่งผลให้ร้อยละของความถูกต้องเพิ่มขึ้นหรือลดลงแต่อย่างใด

ตารางที่ 5 แสดงค่าร้อยละของความถูกต้องในการจำแนกข้อมูลภายใต้ตัวแบบถดถอย $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$ โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับศูนย์จำนวนสองตัวในตัวแบบถดถอย แยกเป็นกรณีศึกษาได้ทั้งหมด 6 กรณี

กรณีศึกษา	ขนาดตัวอย่าง	ระดับนัยสำคัญ 0.05			ระดับนัยสำคัญ 0.1		
		ความแปรปรวน (σ_ε^2)			ความแปรปรวน (σ_ε^2)		
		500	10000	40000	500	10000	40000
1) $\beta_1 = \beta_2 = 0$	200	63.9%	64.3%	65.5%	63.4%	63%	63.5%
	600	64.4%	65.5%	65.2%	63.8%	64.7%	63.7%
	1800	66.1%	67.7%	66.9%	64.9%	65.2%	64%
2) $\beta_1 = \beta_3 = 0$	200	62.8%	64.5%	65.1%	62%	63.2%	64.2%
	600	63.7%	65.6%	66.7%	62.8%	64.1%	65.4%
	1800	65.2%	68.1%	67.2%	63.8%	66.6%	67.1%
3) $\beta_1 = \beta_4 = 0$	200	64.3%	64.9%	66.2%	63.1%	63.1%	64.9%
	600	65.2%	66.1%	66.9%	64.2%	64.5%	65.4%
	1800	66.1%	67%	67.4%	65%	66.2%	65.9%
4) $\beta_2 = \beta_3 = 0$	200	64.8%	64.7%	64.9%	62.7%	63.1%	64.1%
	600	65%	65.2%	66.1%	63.2%	64.2%	65.6%
	1800	65.9%	67%	66.8%	64.6%	65.7%	66.2%
5) $\beta_2 = \beta_4 = 0$	200	63.4%	63.9%	63.8%	62.1%	62.7%	62.9%
	600	64.5%	65.1%	64.7%	62.9%	63.8%	63.1%
	1800	64.9%	66.9%	66.2%	63.2%	64.9%	65.6%
6) $\beta_3 = \beta_4 = 0$	200	64.8%	64.9%	65.1%	63.7%	63.2%	64.2%
	600	65.2%	66.3%	67.3%	64.2%	64%	65.6%
	1800	66.4%	67.2%	68.4%	64.8%	66.1%	66.2%

จากตารางที่ 4.5 สรุปได้ว่าเมื่อเพิ่มขนาดตัวอย่างจะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มเพิ่มขึ้น แต่ในทางกลับกันถ้าเพิ่มระดับนัยสำคัญก็จะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มลดลง ส่วนค่าความแปรปรวนของค่าความคลาดเคลื่อนเชิงสุ่มนั้นจะมีค่าน้อยหรือมากก็ไม่ส่งผลให้ร้อยละของความถูกต้องเพิ่มขึ้นหรือลดลงแต่อย่างใด

ตารางที่ 6 แสดงค่าร้อยละของความถูกต้องในการจำแนกข้อมูลภายใต้ตัวแบบถดถอย $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$ โดยจะกำหนดให้ค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับศูนย์จำนวนสามตัวในตัวแบบถดถอย แยกเป็นกรณีศึกษาได้ทั้งหมด 4 กรณี

กรณีศึกษา	ขนาดตัวอย่าง	ระดับนัยสำคัญ 0.05			ระดับนัยสำคัญ 0.1		
		ความแปรปรวน (σ_ε^2)			ความแปรปรวน (σ_ε^2)		
		500	10000	40000	500	10000	40000
1) $\beta_1 = \beta_2 = \beta_3 = 0$	200	71.8%	72.1%	71.9%	69.2%	69.9%	68.7%
	600	72.5%	73.5%	73.6%	70.1%	71.2%	71.2%
	1800	74.2%	75.2%	75.8%	71.3%	72.6%	73.9%
2) $\beta_1 = \beta_2 = \beta_4 = 0$	200	71.6%	72.9%	72.2%	70.2%	70.2%	69.7%
	600	72.6%	73.7%	73%	70.9%	70.9%	70.5%
	1800	74.6%	76.1%	75.4%	71.1%	73%	72.7%
3) $\beta_1 = \beta_3 = \beta_4 = 0$	200	71.3%	71.8%	72.3%	69.9%	69.2%	70%
	600	72.9%	72.9%	74%	70.6%	70.5%	71.3%
	1800	74.7%	75.9%	76.6%	72.5%	72.1%	72.6%
4) $\beta_2 = \beta_3 = \beta_4 = 0$	200	72.1%	72.6%	73.9%	71.5%	69.7%	70.4%
	600	73.5%	73.7%	75.8%	72.3%	70.9%	72.6%
	1800	74.1%	76%	76.9%	73.2%	73.2%	73.8%

จากตารางที่ 4.6 สรุปได้ว่าเมื่อเพิ่มขนาดตัวอย่างจะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มเพิ่มขึ้น แต่ในทางกลับกันถ้าเพิ่มระดับนัยสำคัญก็จะทำให้ค่าร้อยละของความถูกต้องในการจำแนกข้อมูลมีแนวโน้มลดลง ส่วนค่าความแปรปรวนของค่าความคลาดเคลื่อนเชิงสุ่มนั้นจะมีค่าน้อยหรือมากก็ไม่ส่งผลให้ร้อยละของความถูกต้องเพิ่มขึ้นหรือลดลงแต่อย่างใด

4.2 ตัวอย่างการใช้อัลกอริทึม MODIFIED REGRESSION TREE กับข้อมูลจริง

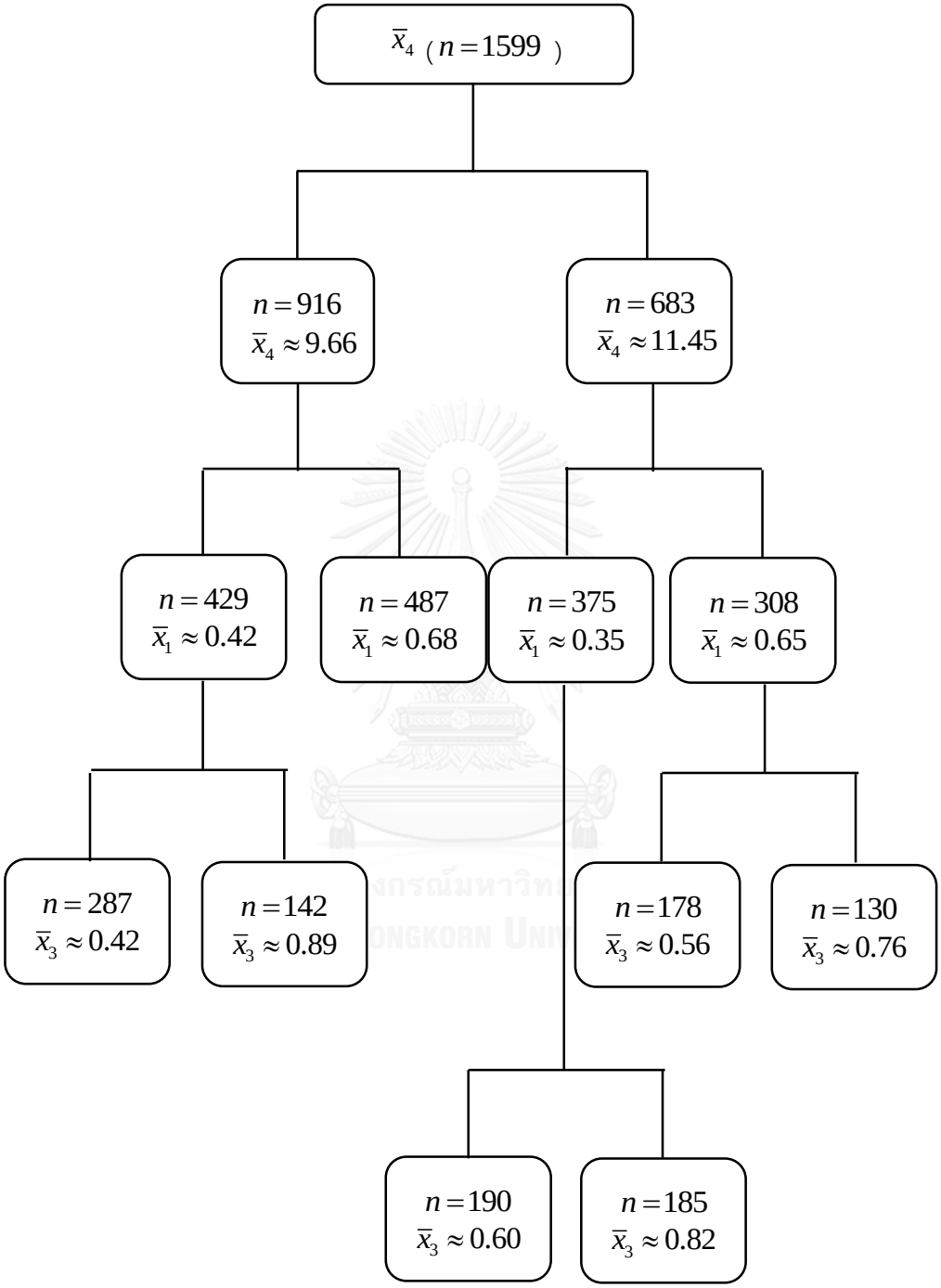
ข้อมูลจริงที่นำมาศึกษานั้นเป็นข้อมูลทางการตลาดซึ่งนำมาจากบริษัทจำหน่ายเครื่องดื่มแอลกอฮอล์ประเภทไวน์ ประกอบไปด้วยตัวแปรอิสระ 4 ตัว และตัวแปรตาม 1 ตัว (*Wine Quality*, 2009) โดยที่แต่ละตัวแปรมีรายละเอียดดังนี้

- ตัวแปรอิสระ x_1 คือ ปริมาณร้อยละของกรดระเหยง่าย (volatile acidity)
  x_2 คือ ปริมาณร้อยละของกรดซิตริก (citric acid)
  x_3 คือ ปริมาณร้อยละของซัลเฟต (sulfates)
  x_4 คือ ปริมาณร้อยละของแอลกอฮอล์ (alcohol)
- ตัวแปรตาม y คือ คะแนนความพึงพอใจของลูกค้า โดยมีคะแนนเต็มจำนวน 10 คะแนน

เมื่อทำการจำแนกข้อมูลโดยใช้อัลกอริทึม MODIFIED REGRESSION TREE แล้วจะได้ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าดังนี้

1. ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าของกลุ่มปลายทางที่ 1 คือ $\bar{y}_1 \approx 5.20$
2. ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าของกลุ่มปลายทางที่ 2 คือ $\bar{y}_2 \approx 5.37$
3. ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าของกลุ่มปลายทางที่ 3 คือ $\bar{y}_3 \approx 5.79$
4. ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าของกลุ่มปลายทางที่ 4 คือ $\bar{y}_4 \approx 6.08$
5. ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าของกลุ่มปลายทางที่ 5 คือ $\bar{y}_5 \approx 6.40$
6. ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าของกลุ่มปลายทางที่ 6 คือ $\bar{y}_6 \approx 5.50$
7. ค่าเฉลี่ยของคะแนนความพึงพอใจของลูกค้าของกลุ่มปลายทางที่ 7 คือ $\bar{y}_7 \approx 6.14$

แผนภาพต้นไม้แบบทวิของการจำแนกข้อมูลจากอัลกอริทึม MODIFIED REGRESSION TREE



- สรุปผลของการจำแนกในกลุ่มปลายทางที่ 1 ได้รับคะแนนความพึงพอใจจากลูกค้าเฉลี่ยโดยประมาณ 5.2 คะแนน ประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.68 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.61 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 9.66
- สรุปผลของการจำแนกในกลุ่มปลายทางที่ 2 ได้รับคะแนนความพึงพอใจจากลูกค้าเฉลี่ยโดยประมาณ 5.37 คะแนน ประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.43 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.57 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 9.62
- สรุปผลของการจำแนกในกลุ่มปลายทางที่ 3 ได้รับคะแนนความพึงพอใจจากลูกค้าเฉลี่ยโดยประมาณ 5.79 คะแนน ประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.38 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.88 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 9.71
- สรุปผลของการจำแนกในกลุ่มปลายทางที่ 4 ได้รับคะแนนความพึงพอใจจากลูกค้าเฉลี่ยโดยประมาณ 6.08 คะแนน ประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.36 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.60 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 11.56
- สรุปผลของการจำแนกในกลุ่มปลายทางที่ 5 ได้รับคะแนนความพึงพอใจจากลูกค้าเฉลี่ยโดยประมาณ 6.4 คะแนน ประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.34 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.82 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 11.50
- สรุปผลของการจำแนกในกลุ่มปลายทางที่ 6 ได้รับคะแนนความพึงพอใจจากลูกค้าเฉลี่ยโดยประมาณ 5.5 คะแนน ประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.69 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.56 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 11.33
- สรุปผลของการจำแนกในกลุ่มปลายทางที่ 7 ได้รับคะแนนความพึงพอใจจากลูกค้าเฉลี่ยโดยประมาณ 6.14 คะแนน ประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.60 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.76 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 11.39

จากผลลัพธ์ของการจำแนกกลุ่มข้อมูลจริงที่นำมาศึกษานั้นในกลุ่มปลายทางที่ 5 ได้รับคะแนนความพึงพอใจจากลูกค้าโดยเฉลี่ยมากที่สุด ในกรณีที่เป็นผู้ผลิตไวน์ส่วนผสมในการผลิตไวน์นั้น

ควรประกอบด้วยกรดระเหยง่ายเฉลี่ยประมาณร้อยละ 0.34 ซัลเฟตเฉลี่ยประมาณร้อยละ 0.82 และแอลกอฮอล์เฉลี่ยประมาณร้อยละ 11.50

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อทำการศึกษาและเปรียบเทียบร้อยละของความถูกต้องของอัลกอริทึม MODIFIED REGRESSION TREE เพื่อหาข้อสรุปว่าตัวแบบถดถอยใด ภายใต้เงื่อนไขของขนาดตัวอย่าง ค่าความแปรปรวนของความคลาดเคลื่อน และระดับนัยสำคัญที่เท่าใด ที่มีร้อยละความถูกต้องของการจำแนกดีกว่ากันโดยพิจารณาจากเกณฑ์ที่ใช้ในการตัดสินใจ

5.1 สรุปผลการวิจัย

จากผลการวิจัย สามารถสรุปผลได้ดังนี้

1) เมื่อเพิ่มขนาดของตัวอย่างในทุกๆ กรณีที่ศึกษาของตัวแบบที่มีความถดถอยที่มีตัวแปรอิสระอยู่ในตัวแบบกี่ตัวก็ตาม จะส่งผลให้ร้อยละของความถูกต้องนั้นมีค่าเพิ่มขึ้นด้วย

2) เมื่อเพิ่มระดับนัยสำคัญในกรณีที่ศึกษาของตัวแบบที่มีความถดถอยที่มีตัวแปรอิสระอยู่ในตัวแบบกี่ตัวก็ตาม จะส่งผลให้ร้อยละของความถูกต้องของระดับนัยสำคัญที่มีค่าเท่ากับ 0.05 มีค่ามากกว่าระดับนัยสำคัญที่มีค่า เนื่องจากจะทำการเปรียบเทียบค่า $p-value$ ในการคัดเลือกตัวแปรอิสระแต่ละตัวกับระดับนัยสำคัญ ดังนั้นยิ่งระดับนัยสำคัญมีค่ามาก โอกาสของตัวแปรอิสระที่ไม่ควรอยู่ในกระบวนการจำแนกข้อมูลอาจเข้ามาในกระบวนการได้ จึงส่งผลให้มีจำนวนโหนดปลายทางที่ไม่ได้ต้องการเพิ่มมากขึ้น

3) เมื่อเพิ่มตัวแปรอิสระในตัวแบบความถดถอยภายใต้ตัวแบบถดถอยที่มีตัวแปรอิสระจำนวน 2 ตัวแปร 3 ตัวแปร และ 4 ตัวแปร ส่งผลให้ร้อยละความถูกต้องนั้นลดลง

4) ค่าความแปรปรวนของค่าความคลาดเคลื่อนไม่ส่งผลต่อร้อยละความถูกต้อง

5) อัลกอริทึม MODIFIED REGRESSION TREE ไม่ค่อยเหมาะกับข้อมูลที่มีขนาดตัวอย่างต่ำกว่า 200 จำนวน เนื่องจากในบางครั้งของการจำแนกกลุ่มข้อมูลเพื่อหาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามนั้นมีโอกาสที่จะทำการจำแนกได้น้อยหรือไม่สามารถแยกได้เลย ทำให้จำนวนตัวอย่างในแต่ละกลุ่มปลายทางอาจมีค่าน้อยเกินไป ไม่เหมาะสมกับการจำแนกกลุ่มข้อมูล

6) ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันมาก (Multicollinearity) อาจทำให้กลุ่มปลายทางของการจำแนกข้อมูลมีจำนวนน้อย

5.2 ข้อเสนอแนะ

จากงานวิจัยชิ้นนี้ผู้ที่สนใจอาจไปทำการศึกษาต่อในกรณีดังต่อไปนี้

1. ทำการเพิ่มจำนวนของตัวแปรอิสระ หรืออาจลดระดับนัยสำคัญแล้วสังเกตแนวโน้มร้อยละความถูกต้อง
2. ใช้ค่ามัธยฐานเป็นเกณฑ์ในการจำแนกแทนค่าเฉลี่ยเลขคณิต
3. ปรับเปลี่ยนค่าพารามิเตอร์ของตัวแบบ จากนั้นสังเกตแนวโน้มร้อยละความถูกต้อง

รายการอ้างอิง

ภาษาไทย

กัลยา วาณิชย์บัญชา. (2545). หลักสถิติ. กรุงเทพมหานคร: ธรรมสาร.

จิราวัลย์ จิตรถเวช. (2558). การวิเคราะห์การถดถอย (*Regression Analysis)*. กรุงเทพมหานคร: สำนักกิจการโรงพิมพ์ องค์การสงเคราะห์ทหารผ่านศึก.

สุพล ดุรงค์วัฒนา. (2558). *Regression Models : Analytics-based Approach*. กรุงเทพมหานคร: แดเน็กซ์อินเตอร์คอร์ปอเรชั่น.

ภาษาต่างประเทศ

Breiman, L., Freidman, J. H., Olshen, R. A., & Stone, C. I. (1984). *Classification and Regression Trees*. California: Wadsworth.

Wine Quality. (2009). Retrieved from: http://archive.ics.uci.edu/ml/

ภาคผนวก

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวพรพิมล อุดมมาลัย เกิดเมื่อวันเสาร์ที่ 28 มกราคม พ.ศ. 2532 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท. บ.) สาขาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2554 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท. ม.) สาขาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2556