การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID

นายวิศรุต กิมชัยวงศ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

DATA CLASSIFICATION BY CHAID ALGORITHM

Mr. Wissarut Kimchaiwong

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 CHAID
โดย	นายวิศรุต กิมชัยวงศ์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา)

.....กรรมการ

(อาจารย์ ดร.อนันตณัฐ กันต์ธัญญรัตน์)

.....กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร.มนต์ทิพย์ เทียนสุวรรณ)

วิศรุต กิมชัยวงศ์ : การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID (DATA CLASSIFICATION BY CHAID ALGORITHM) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.สุพล ดุรงค์วัฒนา, 84 หน้า.

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID สำหรับข้อมูลระหว่างตัวแปร 2 ตัวแปรที่มีการแจกแจงแบบพหุนามและอยู่ในตารางการณ์จรสองทาง โดยพิจารณาความสามารถในการควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 การแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูลเป็นเกณฑ์ในการพิจารณาว่าอัลกอริทึมมีประสิทธิภาพในการจำแนกกลุ่มได้ดีหรือไม่ โดยข้อมูลที่ใช้ในการศึกษาจะจำลองภายใต้จำนวนกลุ่มของตัวแปร 2, 3, 4 และ 5, ขนาดข้อมูลเท่ากับ 200, 400 และ 1,200, ระดับความสัมพันธ์ของข้อมูลเท่ากับ 0, 0.05, 0.1 และ 0.3 และ ระดับนัยสำคัญเท่ากับ 0.05 และ 0.1 และสามารถสรุปผลการศึกษาได้ดังนี้

1) อัลกอริทึม CHAID สามารถควบคุมความน่าจะเป็นของความผิดประเภทที่ 1 ได้ทุกกรณีเมื่อระดับความสัมพันธ์ของข้อมูลเท่ากับ 0

2) เมื่อพิจารณาที่ระดับความสัมพันธ์ของข้อมูลและระดับนัยสำคัญเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น อำนาจการทดสอบและการแยกจะมีแนวโน้มเพิ่มขึ้น ส่วนการรวมมีแนวโน้มลดลง

3) เมื่อพิจารณาที่ระดับความสัมพันธ์ของข้อมูลและขนาดข้อมูลเท่ากัน เมื่อระดับนัยสำคัญเพิ่มขึ้น อำนาจการทดสอบและการแยกจะมีแนวโน้มเพิ่มขึ้น ส่วนการรวมมีแนวโน้มลดลง

4) เมื่อพิจารณาที่ขนาดข้อมูลและระดับนัยสำคัญเท่ากัน เมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น อำนาจการทดสอบ การแยก และร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูลจะมีแนวโน้มเพิ่มขึ้น ส่วนการรวมมีแนวโน้มลดลง

นอกจากนี้อำนาจการทดสอบมีแนวโน้มลดลงเมื่อความแตกต่างระหว่างแถวกับหลักเพิ่มขึ้น และร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูลมีแนวโน้มลดลงเมื่อจำนวนกลุ่มของตัวแปรตามเพิ่มขึ้น

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2558

5681590126 : MAJOR STATISTICS

KEYWORDS: DATA CLASSIFICATION / CHAID ALGORITHM / CHI-SQUARE TEST

WISSARUT KIMCHAIWONG: DATA CLASSIFICATION BY CHAID ALGORITHM. ADVISOR: ASSOC. PROF. SUPOL DURONGWATANA, Ph.D., 84 pp.

The purpose of this paper is to study the classification process of CHAID (Chi-Square Automatic Interaction Detection) algorithm for bivariate multinomial distribution in two way contingency table. Their capacity of controlling probability of type I error, splitting, merging, power of the test and classification rate are used as the measure how good the algorithm for its classification. The data are simulated under several situations. Each situation depends upon the numbers of levels in variable are 2, 3, 4 and 5, the sample size of each set of data are 200, 400, and 1,200, the strength of the relationship between the variables are 0, 0.05, 0.1 and 0.3 and lastly the levels of significant is used with 0.05 and 0.1. The results of this paper can be concluded as below.

1) CHAID algorithm can control probability of type I error in all cases when the strength of the relationship between the variables is 0.

2) If the strength of the relationship between the variables and the significant levels are equal when the number of sample size increases, then power of the test and the splitting tend to increase and the merging tends to decrease.

3) If the strength of the relationship between the variables and the number of sample size are equal when the significant levels increases, then power of the test and the splitting tend to increase and the merging tends to decrease.

4) If the number of sample size and the significant levels are equal when the strength of the relationship between the variables increases, then power of the test, the splitting and the classification rate tend to increase and the merging tends to decrease.

Also, power of the test tends to decrease when the difference of rows and column increase and the classification rate tends to decrease when the number of levels in dependent variable increases.

Department: Statistics Student's Signature .....

Field of Study: Statistics Advisor's Signature .....

Academic Year: 2015

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลือของ รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ คำปรึกษา ตลอดจนช่วยเหลือแก้ไขข้อบกพร่องต่างๆ เป็นอย่างดีเพื่อปรับปรุงแก้ไขวิทยานิพนธ์ จนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์ ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงและสำนึกในพระคุณเป็นอย่างยิ่ง

ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. อนุภาพ สมบูรณ์สวัสดี ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. อนันตณัฐ กันต์ธัญญรัตน์ กรรมการสอบวิทยานิพนธ์ และรองศาสตราจารย์ ดร. มนต์ทิพย์ เทียนสุวรรณ กรรมการภายนอกสอบวิทยานิพนธ์ ที่กรุณาตรวจสอบและให้คำชี้แนะเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้เสร็จสมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ให้โอกาสทางการศึกษาและอบรมสั่งสอนความรู้ให้แก่ผู้วิจัย จนกระทั่งสำเร็จการศึกษา

สุดท้ายนี้ผู้วิจัยใคร่ขอกราบขอบพระคุณครอบครัว ที่เป็นกำลังใจและสนับสนุนด้านการศึกษาแก่ผู้วิจัยเสมอมาจนกระทั่งสำเร็จการศึกษา ตลอดจนเพื่อนๆ ทุกคนที่คอยช่วยเหลือ ให้คำปรึกษาและเป็นกำลังใจให้ด้วยดีมาโดยตลอด

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิจัยและศึกษาข้อมูลทางด้านธุรกิจหรือการตลาดไม่ว่าจะเป็นข้อมูลส่วนบุคคล พฤติกรรมการใช้จ่าย หรือทัศนคติในตัวผลิตภัณฑ์และบริการของผู้บริโภคที่ใช้ในการวิเคราะห์ข้อมูลล้วนมีบทบาทสำคัญในการนำเสนอผลิตภัณฑ์และบริการใหม่ๆ ออกสู่ตลาด เนื่องจากในปัจจุบันมีการแข่งขันที่สูงขึ้น และพฤติกรรมของผู้บริโภคที่มีการเปลี่ยนแปลงอยู่ตลอดเวลาด้วยความก้าวหน้าทางสภาวะเศรษฐกิจ สังคม และด้านต่างๆ เพื่อตอบสนองต่อกลุ่มลูกค้าให้พึงพอใจมากสุด อย่างไรก็ตามก็ไม่สามารถนำเสนอผลิตภัณฑ์หรือบริการที่เหมาะสมหรือตอบสนองความต้องการของผู้บริโภคได้ทั้งหมดทุกกลุ่ม เนื่องจากความต้องการ ความสามารถ และพฤติกรรมผู้บริโภคมีความแตกต่างกัน ดังนั้นงานวิจัยนี้มุ่งเน้นไปที่การจำแนกกลุ่มข้อมูล เพื่อระบุกลุ่มลูกค้าหรือกลุ่มผู้บริโภคที่เป็นเป้าหมายสำคัญ โดยจำแนกประเภทกลุ่มผู้บริโภคที่มีอยู่ทั้งหมดออกเป็นกลุ่มๆ โดยจัดให้บุคคล ความต้องการ ความสามารถ หรือพฤติกรรมที่คล้ายคลึงกันเป็นกลุ่มเดียวกันตามเกณฑ์อัลกอริทึม หลังจากนั้นจึงกำหนดกลุ่มเป้าหมายและตำแหน่งผลิตภัณฑ์ในตลาด เพื่อนำมาวิเคราะห์และตัดสินใจวางแผนทางการตลาดที่มีประสิทธิภาพที่จะทำให้กลุ่มผู้บริโภคหรือผู้สนใจหันมาบริโภคสินค้าและบริการ อัลกอริทึมที่ใช้ในการจำแนกกลุ่มข้อมูลมีหลากหลาย

อัลกอริทึม THAID นำเสนอโดย (Messenger & Mandell, 1972) เป็นอัลกอริทึมที่ใช้ในการวิเคราะห์ในการจำแนกกลุ่มข้อมูล มีข้อจำกัดโดยสามารถจำแนกข้อมูลแค่แบบทวิ โดยใช้ค่าสถิติ Theta เป็นเงื่อนไขในการแบ่งกลุ่ม

อัลกอริทึม CART นำเสนอโดย (Breiman, Friedman, Olshen, & Stone, 1984) เป็นอัลกอริทึมที่ใช้ในการวิเคราะห์ในการจำแนกกลุ่มข้อมูล มีข้อจำกัดโดยสามารถจำแนกข้อมูลแค่แบบทวิ อีกทั้งมีความซับซ้อนในการจำแนกกลุ่มข้อมูล กล่าวคือ ก่อนและหลังการจำแนกกลุ่มข้อมูล ต้องทำการพิจารณาถึงความเป็นไปได้ของการจำแนกกลุ่มข้อมูลนั้นคือกระบวนการเล็มข้อมูล ซึ่งอัลกอริทึม CART เหมาะกับการวิเคราะห์การจำแนกกลุ่มข้อมูลที่มีลักษณะข้อมูลตัวแปรปริมาณต่อเนื่อง

อัลกอริทึม CHAID นำเสนอโดย (Kass, 1980) เป็นอัลกอริทึมที่นิยมและรู้จักอย่างกว้างขวางที่ได้รับการพัฒนาจากอัลกอริทึม THAID เป็นสถิตินอนพาราเมตริก ที่ใช้ในการวิเคราะห์ในการจำแนกกลุ่มข้อมูล สามารถจำแนกข้อมูลแบบพหุภาค (Multi-branch tree) โดยอัลกอริทึม CHAID ที่มี

ข้อมูลตัวแปรตามเป็นตัวแปรเชิงคุณภาพ จะใช้ค่าตัวสถิติทดสอบเพียร์สันไคสแควร์ (Pearson chi-square statistic) เป็นเงื่อนไขในการจำแนกกลุ่ม โดยวิธีการจะคัดเลือกตัวแปรพร้อมทั้งพิจารณาเกณฑ์การรวมและเกณฑ์การหยุดควบคู่กันเป็นขั้นตอน ซึ่งลักษณะข้อมูลตัวแปรอิสระที่ใช้กับอัลกอริทึม CHAID เมื่อตัวแปรอิสระเป็นตัวแปรเชิงปริมาณต่อเนื่อง ต้องทำการแปลงเป็นตัวแปรอิสระเชิงคุณภาพก่อนที่จะนำมาวิเคราะห์ข้อมูล

ขั้นตอนในการจำแนกกลุ่มข้อมูลเพื่อจะได้มาซึ่งผลลัพธ์ที่มีประสิทธิภาพ ก่อนนำไปใช้จำแนกข้อมูล ซึ่งจะต้องผ่านขั้นตอนกระบวนการที่หลากหลาย ฉะนั้นงานวิจัยชิ้นนี้ผู้วิจัยจึงเลือกใช้และศึกษาอัลกอริทึม CHAID ซึ่งเป็นอัลกอริทึมที่เหมาะสมในการจำแนกกลุ่มข้อมูลและใช้ในการวิเคราะห์ข้อมูลเชิงคุณภาพตามขอบเขตของงานวิจัย และสามารถจำแนกกลุ่มข้อมูลแบบพหุภาค ซึ่งเหมาะสมที่จะใช้วิเคราะห์ข้อมูล และเพื่อหาเกณฑ์การตัดสินใจและกลุ่มของข้อมูลที่มีผลต่อการบริโภคสินค้าและบริการตามตลาดเป้าหมายที่กำหนด โดยมีงานวิจัยที่นำอัลกอริทึม CHAID นี้มาใช้ในการจำแนกกลุ่มข้อมูลทางธุรกิจและการตลาด

งานวิจัยของ (Chris, Jyun-Cheng, & David, 2002) เป็นงานวิจัยการจำแนกกลุ่มข้อมูลที่เกี่ยวข้องกับการจัดการข้อมูลความสัมพันธ์ของลูกค้าโดยเลือกใช้อัลกอริทึม CHAID และเทคนิคโครงข่ายประสาทเทียม (Neural networks) ซึ่งทำการศึกษาความสัมพันธ์ของวัฏจักรของกลุ่มลูกค้ากับข้อมูลทางธุรกิจ โดยวัฏจักรกลุ่มลูกค้า 4 กลุ่ม ได้แก่ กลุ่มลูกค้าคาดหวัง (Prospects customers), กลุ่มลูกค้าตอบสนอง (Responders customers), กลุ่มลูกค้ากระตือรือร้น (Active customers) และกลุ่มลูกค้าทางการ (Former customers) ในการวิเคราะห์ทางธุรกิจและการตลาด นอกจากนี้ได้ทำการเปรียบเทียบการทำงานของอัลกอริทึม CHAID และเทคนิคโครงข่ายประสาทเทียมด้วย โดยโครงข่ายประสาทเทียมเหมาะกับการพยากรณ์ค่าของข้อมูล แต่ไม่เหมาะกับการวิเคราะห์และอธิบายการจำแนกกลุ่มข้อมูล เนื่องจากแสดงข้อมูลคุณลักษณะที่เป็นโครงข่ายที่ซับซ้อน ทำให้เข้าใจผลลัพธ์การจำแนกกลุ่มข้อมูลได้ยาก เมื่อเทียบกับอัลกอริทึม CHAID

ดังนั้นเพื่อเป็นการนำข้อมูลมาจำแนกกลุ่มข้อมูลโดยอัลกอริทึมให้ได้ผลลัพธ์ที่มีประสิทธิภาพสูงสุด เข้าใจและมาใช้วิเคราะห์ทางการตลาดที่ง่าย ผู้วิจัยเล็งเห็นว่า ควรมีการศึกษา ทดลอง และพัฒนาความเป็นไปได้ในการแก้ปัญหาในขั้นตอนการจำแนกกลุ่มข้อมูลโดยเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ระหว่างตัวแปร 2 ตัวแปรที่มีการแจกแจงพหุนามและอยู่ในตารางการณ์จรสองทางเป็นแนวทางในการสร้างกระบวนการและวิธีการจำแนกกลุ่มข้อมูลที่มีประสิทธิภาพ นำไปสู่การวิเคราะห์ข้อมูลต่อไป

1.2 วัตถุประสงค์ของการวิจัย

ในงานวิจัยนี้ประกอบด้วยวัตถุประสงค์ดังต่อไปนี้

1. เพื่อศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ระหว่างตัวแปร 2 ตัวแปรที่มีการแจกแจงพหุนามและอยู่ในตารางการณ์จรสองทาง
2. เพื่อสร้างกระบวนการและวิธีขั้นตอนที่มีประสิทธิภาพในการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID มาวิเคราะห์กลุ่มข้อมูล

1.3 ข้อตกลงเบื้องต้น

ในงานวิจัยชิ้นนี้มีข้อตกลงเบื้องต้นดังต่อไปนี้

1. ข้อมูลที่นำมาวิเคราะห์เพื่อการศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูล เป็นข้อมูลที่สร้างขึ้นโดยใช้วิธีการจำลองข้อมูล (simulation) ที่เป็นตัวแปรเชิงคุณภาพ 2 ตัวแปรที่มีการแจกแจงพหุนามและอยู่ในตารางการณ์จรสองทาง โดยตัวแปรแรกเป็นตัวแปรอิสระ X ที่มี I กลุ่ม และตัวแปรที่สองเป็นตัวแปรตาม Y ที่มี J กลุ่ม
2. ข้อมูลตัวแปรทั้งข้อมูลจำลองและข้อมูลที่สนใจที่นำมาใช้วิเคราะห์เป็นตัวแปรเชิงคุณภาพ
3. ข้อมูลจำลองและข้อมูลที่สนใจนำมาใช้วิเคราะห์ไม่มีค่าสูญหาย (missing value) และข้อมูลจำลองในทุกเซลล์ของตารางการณ์จรไม่เป็นศูนย์

1.4 ขอบเขตของการวิจัย

ในงานวิจัยนี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

1. ศึกษาและเปรียบเทียบการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ภายใต้จำนวนประเภทของตัวแปร ขนาดตัวอย่าง ระดับความสัมพันธ์ของข้อมูล และระดับนัยสำคัญ ตามสถานการณ์ต่างๆ ดังต่อไปนี้
   1.1 ศึกษาภายใต้ตัวแปรอิสระ X ที่มีจำนวน $2 \le I \le 5$ กลุ่ม และ ตัวแปรตาม Y ที่มีจำนวน $2 \le J \le 5$ กลุ่ม
   1.2 ศึกษาภายใต้ขนาดตัวอย่าง 200 400 และ 1,200 ในแต่ละตารางการณ์จรตามสถานการณ์
   1.3 ศึกษาภายใต้ระดับความสัมพันธ์ของข้อมูล ตามตารางการณ์จร X x Y ดังนี้

- ศึกษาตารางการณ์จรขนาด 2x2, 2x3, 2x4, 2x5, 3x2, 3x3, 3x4, 3x5, 4x2, 4x3, 4x4, 4x5, 5x2, 5x3, 5x4 และ 5x5 ที่ตัวแปรทั้งสองไม่มีความสัมพันธ์กัน ในแต่ละตารางการณ์จรตามสถานการณ์
- ศึกษาตารางการณ์จรขนาด 2x2, 2x3, 2x4, 2x5, 3x2, 3x3, 3x4, 3x5, 4x2, 4x3, 4x4, 4x5, 5x2, 5x3, 5x4 และ 5x5 ที่ตัวแปรทั้งสองมีความสัมพันธ์กัน ที่ระดับความสัมพันธ์ (τ) เท่ากับ 0.05 0.1 และ 0.3 สำหรับแต่ละตารางการณ์จรตามสถานการณ์

1.4 ศึกษาภายใต้ระดับนัยสำคัญของการจำแนกข้อมูล ที่ระดับ 0.05 และ 0.1

2. ในการศึกษาครั้งนี้จะจำลองข้อมูลตามขอบเขตงานวิจัยข้างต้น ซึ่งผู้วิจัยจะประมวลผลโดยใช้โปรแกรม R เวอร์ชั่น 3.2.3 โดยวิธีดังกล่าวจะกำหนดการจำลองซ้ำของข้อมูลแต่ละกรณีไว้ที่จำนวน 1,000 รอบ เพื่อหาตัววัดประสิทธิภาพที่เหมาะสม

1.5 เกณฑ์ที่ใช้ในการตัดสินใจ

ในการศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID จะพิจารณาการแยก คือจำนวนครั้งในการปฏิเสธสมมติฐานว่างของการทดสอบความเป็นอิสระของการทดสอบไคสแควร์ และการรวม คือจำนวนครั้งในการไม่ปฏิเสธสมมติฐานของการทดสอบเอกพันธ์ของการทดสอบไคสแควร์ โดยจำนวนครั้งในการรวมขึ้นอยู่กับจำนวนกลุ่มของตัวแปรอิสระ กล่าวคือ ถ้าจำนวนกลุ่มของตัวแปรอิสระเท่ากับ I กลุ่ม แล้วจำนวนครั้งในการรวมสูงสุดเท่ากับ I-2 ครั้ง และจะดำเนินการดังนี้

1. ภายใต้สถานการณ์ที่ข้อมูลตัวแปรทั้งสองไม่มีความสัมพันธ์กัน จะพิจารณาการแยก การรวมของการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID และจะพิจารณาความสามารถในการควบคุมความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก โดยพิจารณาการทดสอบสมมติฐานเกี่ยวกับค่าสัดส่วนประชากร โดยใช้ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 คือ สัดส่วนของการปฏิเสธสมมติฐานว่าง โดยที่สมมติฐานว่างเป็นจริงน้อยกว่าหรือเท่ากับที่ระดับนัยสำคัญกำหนด แล้วแต่ละสถานการณ์ในการจำแนกกลุ่มข้อมูลมีความสามารถในการควบคุมความผิดพลาดประเภทที่ 1

2. ภายใต้สถานการณ์ที่ข้อมูลตัวแปรทั้งสองมีความสัมพันธ์กัน จะพิจารณาการแยก การรวมของการการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID และพิจารณาค่าอำนาจการทดสอบหรืออำนาจการตัดสินใจถูกต้อง ในขั้นตอนการแยกของแต่ละสถานการณ์นั้น คือ สัดส่วนของการปฏิเสธสมมติฐานว่าง โดยที่สมมติฐานว่างไม่จริง และพิจารณาค่าร้อยละความถูกต้องเฉลี่ยของการจำแนก

กลุ่มข้อมูล คือ ค่าเฉลี่ยของร้อยละความถูกต้องใน 1,000 รอบ โดยร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูล (Classification Rate) จะใช้วัดความถูกต้องของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม ซึ่งหาได้จาก ซึ่งหาได้จาก

ร้อยละความถูกต้อง (CR) = จำนวนความถูกต้องในการจำแนกกลุ่มข้อมูล / จำนวนข้อมูลทั้งหมด

3. ในข้อมูลที่สนใจนำมาจำแนกกลุ่มข้อมูล เกณฑ์ที่ใช้ในการตัดสินใจว่ากลุ่มของข้อมูลใดที่มีประสิทธิภาพหรือเป็นกลุ่มที่มีผลต่อตลาดเป้าหมาย (กลุ่มของตัวแปรตามที่กำหนดเป็นเป้าหมาย) ที่กำหนดไว้มากสุด หลังจากผ่านขั้นตอนการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID โดยจะพิจารณาจากค่าร้อยละเกนดัชนี (Gain index percentage) คือ อัตราส่วนร้อยละของกลุ่มที่เป็นเป้าหมายในตัวแปรตามที่กำหนดไว้ในโหนดที่ไม่สามารถแยกได้อีก เทียบกับร้อยละของกลุ่มที่เป็นเป้าหมายในตัวแปรตามที่กำหนดไว้ในกลุ่มตัวอย่างทั้งหมด ถ้าค่ามากสุดที่มากกว่า 100% แล้วจะเป็นกลุ่มที่ส่งผลต่อตลาดที่กำหนดเป็นเป้าหมายมากสุด

1.6 คำจำกัดความของงานวิจัย

1. การจำแนกกลุ่มข้อมูล (Data Classification) คือกระบวนการสร้างตัวแบบจัดการกับข้อมูลเพื่อทำนายกลุ่มของข้อมูลใหม่ โดยกลุ่มที่ได้จากการจำแนกกลุ่มข้อมูลที่อยู่กลุ่มเดียวกันจะมีลักษณะข้อมูลที่เหมือนหรือคล้ายคลึงกัน
2. ข้อมูลเชิงคุณภาพ (Qualitative data) คือ ข้อมูลที่ไม่สามารถระบุค่าได้ว่ามากหรือน้อย เป็นข้อมูลที่แสดงจำนวนหรือความถี่ของแต่ละกลุ่มหรือประเภทของข้อมูลเชิงคุณภาพ
3. อัลกอริทึม CHAID คือกระบวนการที่ใช้วิเคราะห์และแก้ปัญหาการจำแนกกลุ่มข้อมูล ที่นิยมในการจำแนกกลุ่มข้อมูลที่มีลักษณะเป็นข้อมูลเชิงคุณภาพ จะใช้ค่าตัวสถิติทดสอบไคสแควร์ในการจำแนกกลุ่ม โดยวิธีการจะคัดเลือกตัวแปรพร้อมทั้งพิจารณาเกณฑ์การรวมและเกณฑ์การหยุดควบคู่กันเป็นขั้นตอน จนกระทั่งเสร็จสิ้นการจำแนกกลุ่ม
4. การคัดเลือกตัวแปร (Variable Selection) คือกระบวนการเลือกตัวแปรที่มีความจำเป็นเพื่อให้ได้การจำแนกกลุ่มข้อมูลที่ดีที่สุด
5. การทดสอบไคสแควร์ (Chi-Square Test) คือกระบวนการทดสอบเพื่อเปรียบเทียบข้อมูลที่อาจอยู่ในรูปของสัดส่วนหรือความถี่ซึ่งจำแนกออกเป็นกลุ่มได้ โดยทดสอบความเป็นอิสระ (Test of Independence) เพื่อทดสอบความสัมพันธ์ระหว่าง 2 ตัวแปร และทดสอบความเป็นเอกพันธ์ (Test of Homogeneity) เพื่อทดสอบความคล้ายคลึงของตัวแปร เพื่อใช้ในการจำแนกกลุ่มข้อมูล

6. เกณฑ์การรวม (Merging Rules) คือกฎเกณฑ์ในการวิเคราะห์ข้อมูลและเป็นแนวทางในการดำเนินการรวมกลุ่มตัวแปรของข้อมูลเพื่อจำแนกกลุ่มข้อมูล
7. เกณฑ์การหยุด (Stopping Rules) คือกฎเกณฑ์ในการวิเคราะห์ข้อมูลและเป็นแนวทางการดำเนินการหยุดการแยกโหนดในการตัดสินใจเพื่อจำแนกกลุ่มข้อมูล

1.7 วิธีการดำเนินงานวิจัย

ในงานวิจัยชิ้นนี้มีวิธีการดำเนินงานแบ่งออกเป็น 8 ขั้นตอน ดังต่อไปนี้

1. ศึกษาตัวแบบ วิธีการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID และทฤษฎีที่เกี่ยวข้อง
2. กำหนดการจำลองข้อมูลและข้อมูลที่สนใจ
   2.1 จำลองข้อมูลตามสถานการณ์ต่างๆ ที่ได้กำหนดไว้ในขอบเขตการวิจัย
   2.2 จำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ในแต่ละสถานการณ์ของข้อมูลจำลอง โดยคำนวณ การแยก การรวม ค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 อำนาจการทดสอบในขั้นตอนการแยก และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม
   2.3 กำหนดค่าเริ่มต้นต่างๆ ในการจำแนกกลุ่มข้อมูลที่สนใจโดยอัลกอริทึม CHAID
      - กำหนดตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพมาใช้วิเคราะห์การจำแนกกลุ่มข้อมูล
      - กำหนดกลุ่มที่เป็นเป้าหมายในตัวแปรตาม
      - กำหนดค่าความลึกมากสุด พร้อมทั้งกำหนดขนาดของโหนดต่ำสุด
      - กำหนดระดับนัยสำคัญในการจำแนกกลุ่มข้อมูล
3. เปรียบเทียบผลลัพธ์และวัดประสิทธิภาพที่ได้จากการจำแนกกลุ่มข้อมูลจำลองโดยอัลกอริทึม CHAID โดยพิจารณาสรุปผลการวิจัยในแต่ละสถานการณ์ ตามเกณฑ์ที่ใช้ในการตัดสินใจ
4. นำผลการวิจัยการจำแนกกลุ่มข้อมูลจำลองมาประยุกต์ใช้กับการจำแนกกลุ่มข้อมูลที่สนใจศึกษาโดยอัลกอริทึม CHAID และพิจารณาการจำแนกกลุ่มข้อมูลจากค่าร้อยละเกนดัชนี
5. สรุปผลการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID
6. ทดสอบประสิทธิภาพ และแก้ไขข้อผิดพลาด
7. วิเคราะห์และสรุปผลการศึกษา

1.8 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัยชิ้นนี้ ประกอบไปด้วย

1. เพื่อให้ผู้วิจัยและผู้สนใจเข้าใจถึงกระบวนการทำงานการจำแนกกลุ่มข้อมูลและเงื่อนไขต่างๆ ของอัลกอริทึม CHAID มากขึ้น
2. เพื่อให้ผู้วิจัยและผู้สนใจสามารถนำอัลกอริทึม CHAID ที่ได้ทำการศึกษามาประยุกต์ใช้ในการไปจำแนกกลุ่มข้อมูล

บทที่ 2
ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

การวิจัยครั้งนี้เป็นการศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID กับข้อมูลที่จำลองด้วยตัวแบบ 2 ตัวแปรที่มีการแจกแจงพหุนาม (Bivariate Multinomial Distribution) ที่อยู่ในรูปของตารางการณ์จร X x Y ขนาดตาราง I x J โดยเปรียบเทียบประสิทธิภาพของอัลกอริทึม CHAID ตามสถานการณ์ต่างๆ และนำผลลัพธ์ที่ได้จากการจำแนกกลุ่มข้อมูลจำลองมาศึกษาการจำแนกกลุ่มข้อมูลที่สนใจเพื่อให้ได้การจำแนกกลุ่มข้อมูลที่ดีและมีประสิทธิภาพ โดยมีทฤษฎีที่เกี่ยวข้องและเกณฑ์ในการวัดประสิทธิภาพ ดังต่อไปนี้

2.1 ข้อมูลตารางการณ์จรสองทาง

ข้อมูลที่นำมาทดสอบจะเป็นข้อมูลที่จำแนกสองทาง จึงมีลักษณะเป็นตารางการณ์จรสองทาง (contingency table) โดยให้ X เป็นตัวแปรอิสระเชิงคุณภาพที่มี I กลุ่ม (แบ่งตามแถวนอน) และ Y เป็นตัวแปรตามเชิงคุณภาพที่มี J กลุ่ม (แบ่งตามหลักตั้ง) สามารถจำแนกได้ถึงขนาด I x J ตามตารางการณ์จร X x Y โดยข้อมูลที่นำมาวิเคราะห์อยู่ในรูปความถี่ดังนี้

ตารางที่ 1 ตารางการณ์จรสองทางของตัวแปรเชิงคุณภาพที่มี 2 ตัวแปร

(X) ตัวแปรอิสระ	(Y) ตัวแปรตาม				รวม
	Y_1	Y_2	...	Y_j	
X_1	o_{11}	o_{12}	...	o_{1j}	$n_{1.}$
X_2	o_{21}	o_{22}	...	o_{2j}	$n_{2.}$
X_3	o_{31}	o_{32}	...	o_{3j}	$n_{3.}$
⋮	⋮	⋮	...	⋮	⋮
X_i	o_{i1}	o_{i2}	...	o_{ij}	$n_{i.}$
รวม	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	n

โดยที่ o_{ij} เป็นความถี่ของค่าสังเกตใน X_i และ Y_j

$n_{i.}$ เป็นผลรวมจำนวนความถี่ของค่าสังเกตใน X_i

$n_{.j}$ เป็นผลรวมจำนวนความถี่ของค่าสังเกตใน Y_j

2.2 การแจกแจงแบบพหุนาม

คุณลักษณะทางข้อมูล 2 ลักษณะที่สนใจศึกษา คือ คุณลักษณะ X สามารถแบ่งออกได้ i กลุ่ม คือ $X_1,X_2,\ldots, X_i$ และคุณลักษณะ Y สามารถแบ่งออกได้ j กลุ่ม คือ $Y_1,Y_2,\ldots, Y_j$ ถ้าการทดลองที่ทำซ้ำๆ กัน n ครั้งอย่างเป็นอิสระซึ่งกันและกัน ปรากฏผลการทดลองดังนี้ คือ X_1 และ Y_1 จะเกิดขึ้น o_{11} ครั้ง X_1 และ Y_2 เกิดขึ้น o_{12} ครั้ง X_i และ Y_j เกิดขึ้น o_{ij} ครั้ง ด้วยความน่าจะเป็น $p_{11}, p_{12},\ldots, p_{ij}$ (i=1,2,...,I และ j=1,2,...,J) ตามลำดับ มีฟังก์ชันการแจกแจงความน่าจะเป็นร่วม คือ

$$f(o_{11}, o_{12}, \ldots, o_{ij}: p_{11}, p_{12,\ldots,} p_{ij}) = \binom{n}{o_{11}, o_{12}, \ldots, o_{ij}} \prod_{ij} p_{ij}^{o_{ij}}$$

โดยที่ o_{ij} คือความถี่ของค่าสังเกตของตัวแปรอิสระ X กลุ่มที่ i และตัวแปรตาม Y กลุ่มที่ j ซึ่งมีการแจกแจงพหุนาม

p_{ij} คือความน่าจะเป็นของค่าสังเกตของตัวแปรอิสระ X กลุ่มที่ i และตัวแปรตาม Y กลุ่มที่ j

n คือขนาดความถี่ของข้อมูล ที่ $\sum_i \sum_j x_{ij} = n$

2.3 ความเป็นอิสระต่อกันของตัวแปรในตารางการณ์จรสองทาง

โดยกำหนดให้ X เป็นตัวแปรอิสระเชิงคุณภาพที่มี I กลุ่ม และ Y เป็นตัวแปรตามเชิงคุณภาพที่มี J กลุ่ม พบว่าเราสามารถใช้การแจกแจงแบบมีเงื่อนไขของ X เมื่อกำหนด Y มาช่วยอธิบายความเกี่ยวข้องหรือสอดคล้องกันได้ โดย

$$p_{j|i} = p_{ij}/p_{i.} \quad \text{ทุกๆ i และ j}$$

ซึ่งตัวแปรทั้งสองจะเป็นอิสระต่อกันหรือไม่มีความสัมพันธ์ เมื่อ

$$p_{ij} = p_{i.}\, p_{.j} \quad \text{i=1,2,...,I , j=1,2,...,J}$$

2.4 การทดสอบไคสแควร์

การทดสอบไคสแควร์เป็นการทดสอบสมมติฐานเพื่อเปรียบเทียบข้อมูลที่อยู่ในรูปของความถี่ ซึ่งได้มาจากตัวแปรเชิงคุณภาพหรือแบ่งประเภทโดยสามารถนำไปจัดลงในตารางการณ์จรได้ โดยการทดสอบไคสแควร์ที่ใช้ในงานวิจัยนี้ มีการทดสอบหลัก 2 ประเภท คือ การทดสอบความเป็นเอกพันธ์ของข้อมูล และ การทดสอบความสัมพันธ์หรือการทดสอบความเป็นอิสระระหว่างตัวแปร

2.4.1 การทดสอบความเป็นเอกพันธ์ของข้อมูล

การทดสอบความเป็นเอกพันธ์ของข้อมูล (Test of Homogeneity) ใช้ในการทดสอบความแตกต่างระหว่างกลุ่มข้อมูล 2 กลุ่ม ที่ข้อมูลจัดลงในตารางการณ์จรสองทาง เพื่อใช้ในการจำแนกกลุ่มข้อมูลว่ามีลักษณะคล้ายคลึงกันหรือไม่ โดยสมมติฐานเพื่อการทดสอบคือ

H_0: กลุ่มข้อมูล 2 กลุ่มไม่แตกต่างกันหรือมีความคล้ายคลึงกัน

H_1: กลุ่มข้อมูล 2 กลุ่มแตกต่างกันหรือไม่มีความคล้ายคลึงกัน

สถิติทดสอบที่ใช้ คือตัวสถิติเพียร์สันไคสแควร์ คือ $\chi_c^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ โดยที่ $e_{ij} = \frac{n_{i.}n_{.j}}{n}$

โดยที่ o_{ij} คือความถี่สังเกตกลุ่มที่ i ของตัวแปรตัวที่ 1 เมื่อ i=1,2,...,I และกลุ่มที่ j ของตัวแปร ตัวที่ 2 เมื่อ j=1,2,...,J

e_{ij} คือความถี่คาดหวังของกลุ่มที่ i ของตัวแปรตัวที่ 1 เมื่อ i=1,2,...,I และกลุ่มที่ j ของตัวแปรตัวที่ 2 เมื่อ j=1,2,...,J

ให้ค่า p-value คือ $P(\chi^2 > \chi_c^2)$ ซึ่งมีการแจกแจงแบบไคสแควร์ที่มีองศาอิสระที่ (I-1)(J-1)
เขตปฏิเสธ จะปฏิเสธสมมติฐานว่าง ถ้าค่า χ_c^2 ที่คำนวณได้มากกว่าค่า χ^2 ด้วยองศาอิสระ (I-1)(J-1) ณ ระดับนัยสำคัญที่กำหนด หรือค่า p-value ที่คำนวณได้น้อยกว่าค่าระดับนัยสำคัญที่กำหนด หมายความว่า กลุ่มข้อมูล 2 กลุ่มแตกต่างกันหรือไม่มีความคล้ายคลึงกัน

2.4.2 การทดสอบความเป็นอิสระระหว่างตัวแปร

การทดสอบความเป็นอิสระระหว่างตัวแปร (Test of Independence) หรือเรียกอีกการทดสอบหนึ่งว่าเป็นการทดสอบความสัมพันธ์ระหว่างตัวแปร (Test of Association) 2 ลักษณะจากตารางการณ์จร เป็นการทดสอบไคสแควร์เพื่อศึกษาว่าตัวแปรทั้งสองมีความสัมพันธ์หรือเป็นอิสระซึ่งกันและกันหรือไม่ โดยสมมติฐานเพื่อการทดสอบคือ

H_0: ตัวแปรทั้งสองเป็นอิสระต่อกัน หรือ $H_0: p_{ij} = p_{i.}\,p_{.j}$

H_1: ตัวแปรทั้งสองไม่เป็นอิสระต่อกัน $H_1: p_{ij} \neq p_{i.}\,p_{.j}$

โดยที่ $p_{i.} = \sum_{j=1}^{J} p_{ij}$ และ $p_{.j} = \sum_{j=1}^{I} p_{ij}$

สถิติทดสอบที่ใช้ คือตัวสถิติเพียร์สันไคสแควร์ คือ $\chi_c^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ โดยที่ $e_{ij} = \frac{n_{i.}n_{.j}}{n}$

เขตปฏิเสธ จะปฏิเสธสมมติฐานว่าง ถ้าค่า χ_c^2 ที่คำนวณได้มากกว่าค่า χ^2 ด้วยองศาอิสระ (I-1)(J-1) ณ ระดับนัยสำคัญที่กำหนด หรือค่า p-value ที่คำนวณได้น้อยกว่าค่าระดับนัยสำคัญที่กำหนด หมายความว่า ตัวแปรทั้งสองไม่เป็นอิสระกัน หรือตัวแปรทั้งสองมีความสัมพันธ์กัน

2.4.3 ข้อจำกัดในการใช้สถิติทดสอบไคสแควร์

(กัลยา วานิชย์บัญชา, 2553) ทำการสรุปข้อจำกัดดังกล่าวดังนี้

1. ความถี่คาดหวังไม่ควรต่ำกว่า 5 ในแต่ละเซลล์
2. ถ้า I=2 , J=2 หรือตารางการณ์จรสองทางที่มีขนาด 2x2 องศาอิสระเป็น (I-1)(J-1)=1 จึงต้องปรับค่าสถิติทดสอบ χ^2 ที่ $\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(|o_{ij}-e_{ij}|-0.5)^2}{e_{ij}}$ แต่ถ้าขนาดตัวอย่าง n≥50 จะไม่ต้องปรับค่า χ^2

2.5 อัลกอริทึม CHAID

ในงานวิจัยชิ้นนี้ที่มีข้อมูลเชิงคุณภาพ จะใช้อัลกอริทึม CHAID เพื่อจำแนกกลุ่มข้อมูลและแบ่งกลุ่มข้อมูลอย่างมีประสิทธิภาพ ซึ่งงานวิจัยของ (Alkhasawneh, Ngah, Tay, Mat Isa, & Al-Batah, 2014) ได้ศึกษาขั้นตอนการทำงานอัลกอริทึม CHAID โดยขั้นตอนการทำงานเริ่มต้นจากขั้นตอนการรวม การแยก และการหยุด ตามลำดับ ซึ่งในงานวิจัยนี้จะใช้อัลกอริทึม CHAID ในการจำแนกกลุ่มข้อมูลโดยเริ่มต้นจากขั้นตอนการแยก การรวม และการหยุด

<u>ขั้นตอนการทำงานอัลกอริทึม CHAID</u>

อัลกอริทึม (CHAID) เป็นอัลกอริทึมที่ใช้ในการจำแนกกลุ่มข้อมูลตามคุณลักษณะของข้อมูลและหาความสัมพันธ์ระหว่างตัวแปร ซึ่งงานวิจัยชิ้นนี้จะศึกษาโดยคัดเลือกตัวแปรอิสระ X ที่มีความสัมพันธ์กันสูงกับลักษณะข้อมูลของตัวแปรตาม Y และหาการรวมกลุ่มของตัวแปรอิสระ X ที่ไม่มีนัยสำคัญ ใช้สร้างการจำแนกกลุ่มข้อมูลซึ่งแสดงอยู่ในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” ซึ่งง่ายต่อการจำแนกกลุ่มเพื่อใช้ในการวิเคราะห์ข้อมูล โดยขั้นตอนการทำงานหลักของอัลกอริทึม CHAID ที่ใช้ในงานวิจัยนี้ คือ ขั้นตอนการแยก ขั้นตอนการรวม และขั้นตอนการหยุด

2.5.1 ขั้นตอนการแยก

ในขั้นตอนการแยก (Splitting) จะพิจารณาตัวแปรอิสระทุกตัว โดยจะคัดเลือกตัวแปรอิสระเพื่อใช้แยกโหนดนั้น โดยพิจารณาการเลือกตัวแปรอิสระเหล่านั้นจะเทียบค่า p-value ที่คำนวณได้ ขั้นตอนการแยกประกอบด้วยขั้นตอนการทำงานดังต่อไปนี้

1. เลือกตัวแปรอิสระที่มีค่า p-value ที่น้อยที่สุด (มีนัยสำคัญมากที่สุด)

2. สำหรับ ค่า p-value ที่น้อยที่สุด
   - ถ้าค่า p-value นั้นมีค่าน้อยกว่าหรือเท่ากับค่าระดับนัยสำคัญในขั้นตอนการแยกที่ผู้ใช้กำหนด แล้วแยกโหนดโดยใช้ตัวแปรอิสระนั้น
   - ถ้าค่า p-value นั้นมีมากกว่าค่าระดับนัยสำคัญในขั้นตอนการแยกที่ผู้ใช้กำหนด แล้วโหนดนั้นจะแยกไม่ได้และจะพิจารณาเป็นโหนดปลายทาง

2.5.2 ขั้นตอนการรวม

ในขั้นตอนการรวม (Merging) ของอัลกอริทึม CHAID จะทำการลดจำนวนกลุ่มของตัวแปรอิสระ X จากขั้นตอนการแยกโดยจะพิจารณาการรวมกลุ่มของตัวแปรอิสระ X ที่ไม่มีนัยสำคัญจนไม่สามารถรวมได้อีกตามขั้นตอน ขั้นตอนการรวมมี 5 ขั้นตอนดังต่อไปนี้

1. ถ้าตัวแปรอิสระ X มี 2 ประเภท ไปขั้นตอนการทำงานที่ 5
2. ถ้าตัวแปรอิสระ X มีอย่างน้อย 3 ประเภทขึ้นไป พิจารณาคู่กลุ่มของตัวแปรอิสระ หาคู่ที่มีค่านัยสำคัญน้อยที่สุด (คล้ายคลึงกันที่สุด) โดยคู่ที่ใกล้เคียงกันที่สุดคือคู่ที่มีตัวสถิติคำนวณที่ให้ค่า p-value มากที่สุดตามลักษณะตัวแปรตาม Y
3. สำหรับคู่ที่มีค่า p-value มากที่สุด
   - ถ้ามีค่า p-value มากกว่าค่าระดับนัยสำคัญในขั้นตอนการรวมที่ผู้ใช้กำหนด ให้คู่นี้รวมกันเป็นกลุ่มร่วม 1 ตัว จะได้เซตใหม่ของกลุ่มของตัวแปรอิสระ X
   - ถ้ามีค่า p-value น้อยกว่าหรือเท่ากับค่าระดับนัยสำคัญในขั้นตอนการรวมที่ผู้ใช้กำหนด แล้วไปที่ขั้นตอนการทำงานที่ 5
4. ไปที่ขั้นตอน 1
5. แยกตัวแปรอิสระ X นั้นด้วยจำนวนกลุ่มของตัวแปรอิสระ X ที่เหลืออยู่

2.5.3 ขั้นตอนการหยุด

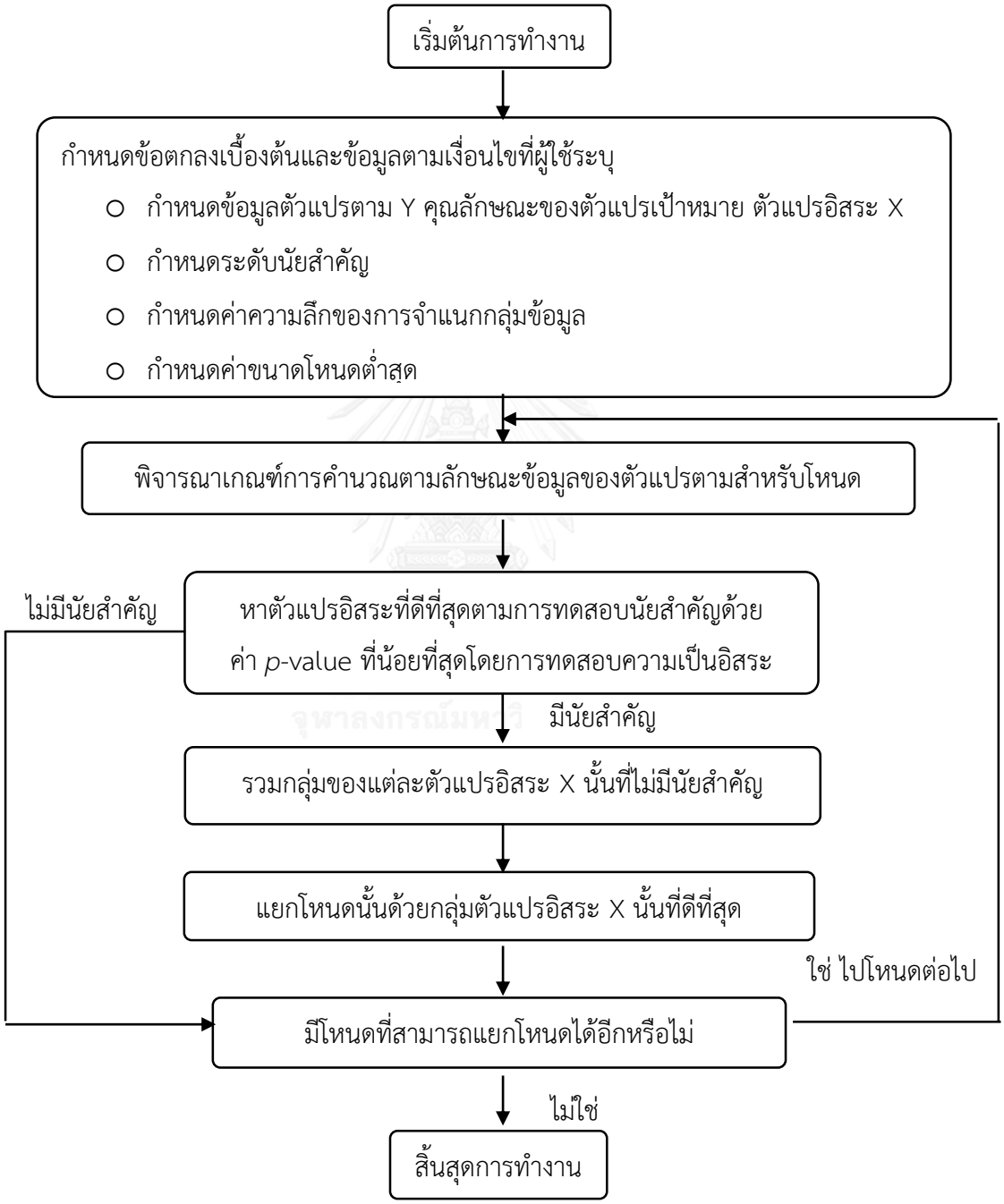
ในขั้นตอนการหยุด (Stopping) จะนำมาใช้ตรวจสอบว่าเมื่อไรกระบวนการแผนภาพการตัดสินใจถึงจะหยุดการสร้างและเมื่อไรควรหยุดการแยกโหนด โดยเกณฑ์การหยุดมีข้อกำหนดไว้ดังนี้

1. ถ้าความลึกของการจำแนกกลุ่มข้อมูลผ่านค่าจำกัดความลึกมากที่สุดที่ผู้ใช้กำหนดแล้วกระบวนการจำแนกกลุ่มข้อมูลจะหยุด
2. ถ้าขนาดของโหนดนั้นมีค่าน้อยกว่าค่าขนาดของโหนดต่ำสุดที่ผู้ใช้กำหนดแล้วโหนดนั้นจะหยุดการแยก

โดยผู้ใช้สามารถกำหนดระดับนัยสำคัญในขั้นตอนการแยกและการรวมที่แตกต่างกันได้ และสามารถกำหนดค่าความลึกของการจำแนกกลุ่มข้อมูลและขนาดโหนดต่ำสุดได้ตามความเหมาะสม ซึ่ง

ผู้วิจัยกำหนดระดับนัยสำคัญในทั้งสองขั้นตอนที่เท่ากัน ขั้นตอนการทำงานการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม สามารถเขียนแผนผังขั้นตอนการทำงานของอัลกอริทึม CHAID เพื่อให้ได้มาซึ่งการจำแนกกลุ่มข้อมูลที่มีประสิทธิภาพ ได้ดังแผนผังที่ 1

แผนผังที่ 1 ขั้นตอนกระบวนการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID



2.6 ความผิดพลาดในการทดสอบสมมติฐานทางสถิติ

การทดสอบสมมติฐานทางสถิติเพื่อตรวจสอบว่าสิ่งที่ผู้วิจัยคาดหวังไว้เป็นจริงหรือไม่ ซึ่งการทดสอบมักจะมีความผิดพลาดหรือความคลาดเคลื่อนในการสรุปเกิดขึ้นเสมอ โดยความผิดพลาดแบ่งเป็น 2 ชนิด คือ

1. ความผิดพลาดประเภทที่ 1 (Type I Error)

เป็นความผิดพลาดที่เกิดขึ้นเนื่องจากการปฏิเสธสมมติฐานว่าง โดยที่ความเป็นจริงสมมติฐานว่างเป็นจริง

α = P (ปฏิเสธ H_0 โดยที่ H_0 เป็นจริง) หรือ P (ปฏิเสธ H_0 | H_0 เป็นจริง)

หรือ α คือโอกาสที่จะสรุปผิด โดยสรุปว่าสมมติฐานว่างไม่เป็นจริง โดยที่ความเป็นจริงสมมติฐานว่างเป็นจริง และจะเรียกความผิดพลาดชนิดนี้ว่า ระดับนัยสำคัญ (Level of significance)

2. ความผิดพลาดประเภทที่ 2 (Type II Error)

เป็นความผิดพลาดที่เกิดขึ้นเนื่องจากการไม่ปฏิเสธสมมติฐานว่าง โดยที่ความเป็นจริงสมมติฐานว่างเป็นเท็จ

β = P (ไม่ปฏิเสธ H_0 โดยที่ H_0 ไม่เป็นจริง) หรือ P (ไม่ปฏิเสธ H_0 | H_0 ไม่เป็นจริง)

หรือ β คือโอกาสที่จะสรุปผิด โดยสรุปว่าสมมติฐานว่างเป็นจริง โดยที่ความเป็นจริงสมมติฐานว่างเป็นเท็จ

ตารางที่ 2 ความผิดพลาดในการทดสอบ

สรุปผลการทดสอบ	ความเป็นจริง	
	H_0 เป็นจริง	H_0 เป็นเท็จ
ไม่ปฏิเสธ H_0	1- α	ความผิดพลาดประเภทที่ 2 (β)
ปฏิเสธ H_0	ความผิดพลาดประเภทที่ 1 (α)	1- β

โดยที่ **อำนาจการทดสอบ** หรือ **อำนาจการตัดสินใจถูกต้อง** = 1- β = P (ปฏิเสธ H_0 | H_0 ไม่เป็นจริง) หมายถึง โอกาสที่จะสรุปผิด โดยสรุปว่าสมมติฐานว่างไม่เป็นจริง โดยที่ความเป็นจริงสมมติฐานว่างเป็นเท็จ

2.7 การวัดประสิทธิภาพการจำแนกกลุ่มข้อมูล

ในการศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID จะพิจารณาการแยก คือจำนวนครั้งในการปฏิเสธสมมติฐานว่างของการทดสอบความเป็นอิสระของการทดสอบไคสแควร์ และการรวม คือจำนวนครั้งในการไม่ปฏิเสธสมมติฐานของการทดสอบเอกพันธ์ของการทดสอบไคสแควร์ โดยจำนวนครั้งในการรวมขึ้นอยู่กับจำนวนกลุ่มของตัวแปรอิสระ กล่าวคือ ถ้าจำนวนกลุ่มของตัวแปรอิสระเท่ากับ I กลุ่ม แล้วจำนวนครั้งในการรวมสูงสุดเท่ากับ I-2 ครั้ง และจะพิจารณาค่าต่างๆ เป็นเกณฑ์ในการตัดสินใจ ดังนี้

1. ความสามารถในการควบคุมความผิดพลาดประเภทที่ 1

(ศศิธร เจษฎาฐิติกุล, 2545) พิจารณาการทดสอบเกี่ยวกับค่าสัดส่วนประชากรที่เป็นการทดสอบทางขวา เพื่อใช้ประกอบการพิจารณาความสามารถในการควบคุมความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยกของอัลกอริทึม CHAID โดยใช้ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 คือสัดส่วนของการปฏิเสธสมมติฐานว่าง โดยที่สมมติฐานว่างเป็นจริงจากการทดลองจำแนกกลุ่มข้อมูลในแต่ละสถานการณ์ เป็นตัวควบคุมความผิดพลาดประเภทที่ 1 ที่ระดับนัยสำคัญของการทดสอบ โดยสมมติฐานเพื่อการทดสอบคือ

$$H_0: \alpha \leq \alpha_0$$

$$H_1: \alpha > \alpha_0$$

สถิติทดสอบที่ใช้ $Z = \frac{\hat{\alpha}-\alpha_0}{\sqrt{\frac{\alpha_0(1-\alpha_0)}{N}}}$

เขตปฏิเสธ จะปฏิเสธสมมติฐานว่าง $Z > Z_\alpha$

โดยที่ N เป็นจำนวนรอบของการทดลอง

n_r เป็นจำนวนทั้งหมดที่ของความผิดพลาดประเภทที่ 1

$\hat{\alpha}$ เป็นค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 จากการจำแนกกลุ่มข้อมูลที่มีค่าเท่ากับ n_r/N

α_0 เป็นระดับนัยสำคัญในการทดสอบที่กำหนดในการวิจัยครั้งนี้ ที่ต้องการควบคุม

โดยถ้า $Z \leq Z_\alpha$ จะไม่ปฏิเสธ H_0 หมายความว่า สามารถควบคุมความผิดพลาดประเภทที่ 1

ถ้าค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 จากการจำแนกกลุ่มข้อมูลตกอยู่ในช่วงของการไม่ปฏิเสธ H_0 ดังกล่าว แล้วแต่ละสถานการณ์และกรณีในการจำแนกกลุ่มข้อมูลมีความสามารถควบคุมความผิดพลาดประเภทที่ 1

2. การคำนวณค่าร้อยละความถูกต้องของการจำแนกข้อมูลจากเมทริกซ์ความคลาดเคลื่อน

เมทริกซ์ความคลาดเคลื่อน (Confusion matrix) เป็นตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวนคอลัมภ์และเท่ากับจำนวนกลุ่มของตัวแปรตาม Y ที่แสดงจำนวนความถูกต้องและไม่ถูกต้องในการจำแนกกลุ่มข้อมูล โดยร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูล (Classification Rate) จะใช้วัดความถูกต้องของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID ซึ่งหาได้จาก

ร้อยละความถูกต้อง (CR) = จำนวนความถูกต้องในการจำแนกกลุ่มข้อมูล / จำนวนข้อมูลทั้งหมด

3. การหากลุ่มเป้าหมายที่ดีที่สุดจากการจำแนกกลุ่มข้อมูล

ในงานวิจัยชิ้นนี้จะทำการศึกษาการจำแนกกลุ่มข้อมูลที่สนใจด้วยข้อมูลเชิงคุณภาพ เป็นข้อมูลที่ได้มาจากทาง (Norusis, 2011) เป็นข้อมูลที่ธนาคารเก็บข้อมูลสำคัญของลูกค้า 2,464 คน ที่กู้ยืมจากธนาคารที่ประกอบด้วยข้อมูลส่วนบุคคลและข้อมูลประวัติสินเชื่อการชำระคืนเงินกู้ของธนาคาร โดยเกณฑ์ที่ใช้ในการตัดสินใจว่ากลุ่มของข้อมูลใดที่มีประสิทธิภาพหรือเป็นกลุ่มที่มีผลต่อตลาดเป้าหมาย (กลุ่มของตัวแปรตามที่กำหนดเป็นเป้าหมาย) ที่กำหนดไว้มากสุด หลังจากผ่านขั้นตอนการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID โดยจะพิจารณาจากค่าร้อยละเกนดัชนี ที่ (Chen, 2003)นำมาระบุกลุ่มตลาดที่เป็นเป้าหมายสำคัญ โดยมีรายละเอียดดังนี้

ค่าร้อยละเกนดัชนี (Gain index percentage) คือ อัตราส่วนร้อยละของกลุ่มที่เป็นเป้าหมายในตัวแปรตามที่กำหนดไว้ในโหนดที่ไม่สามารถแยกได้อีก เทียบกับร้อยละของกลุ่มที่เป็นเป้าหมายในตัวแปรตามที่กำหนดไว้ในกลุ่มตัวอย่างทั้งหมด ถ้าค่ามากสุดที่มากกว่า 100% แล้วจะเป็นกลุ่มที่ส่งผลต่อตลาดที่กำหนดเป็นเป้าหมายมากสุด

บทที่ 3
วิธีดำเนินการศึกษา

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ระหว่างตัวแปร 2 ตัวแปรที่มีการแจกแจงพหุนามและอยู่ในตารางการณ์จรสองทาง โดยการศึกษานี้จะเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลจากข้อมูลจำลองในแต่ละสถานการณ์ที่ใช้ในการจำแนกกลุ่มข้อมูล โดยทำการวิเคราะห์ข้อมูลทั้งหมดโดยใช้โปรแกรม R เวอร์ชัน 3.2.3 ภายใต้วิธีการดำเนินงานดังนี้

3.1 ขั้นตอนในการดำเนินการศึกษา

1. ศึกษาตัวแบบ วิธีการจำแนกกลุ่มข้อมูลและทฤษฎีที่เกี่ยวข้อง
2. กำหนดการจำลองข้อมูลและข้อมูลที่สนใจ
   2.1 จำลองข้อมูลตามสถานการณ์ต่างๆ ที่ได้กำหนดไว้ในขอบเขตการวิจัย
   2.2 จำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ในแต่ละสถานการณ์ของข้อมูลจำลอง ภายใต้สถานการณ์ที่ข้อมูลตัวแปรทั้งสองไม่มีความสัมพันธ์กัน คำนวณค่าการแยก การรวม และค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก และภายใต้สถานการณ์ที่ข้อมูลตัวแปรทั้งสองมีความสัมพันธ์กัน คำนวณค่าการแยก การรวม อำนาจการทดสอบในขั้นตอนการแยก และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID
   2.3 กำหนดค่าเริ่มต้นต่างๆ ในการจำแนกกลุ่มข้อมูลที่สนใจโดยอัลกอริทึม CHAID
      - กำหนดตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพมาใช้วิเคราะห์การจำแนกกลุ่มข้อมูล
      - กำหนดกลุ่มที่เป็นเป้าหมายในตัวแปรตามคือ กำหนดเป้าหมายยอดแย่ (Bad) เพื่อวิเคราะห์หากลุ่มข้อมูลที่อาจจะไม่มาชำระเงินที่กู้ยืมตามกำหนด
      - กำหนดค่าความลึกมากสุดที่ 3 พร้อมทั้งกำหนดขนาดของโหนดต่ำสุดที่ 200
      - กำหนดระดับนัยสำคัญในการจำแนกกลุ่มข้อมูลที่ระดับ 0.05
3. เปรียบเทียบผลลัพธ์และวัดประสิทธิภาพที่ได้จากการจำแนกกลุ่มข้อมูลจำลองโดยอัลกอริทึม CHAID โดยพิจารณาการแยก การรวม และความสามารถในการควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก สำหรับกรณีข้อมูลตัวแปรทั้งสองไม่มีความสัมพันธ์กัน และพิจารณาการแยก การรวม อำนาจการทดสอบในขั้นตอนการแยกและ

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID สำหรับกรณีข้อมูลตัวแปรทั้งสองมีความสัมพันธ์กัน มาสรุปผลการวิจัยในแต่ละสถานการณ์

4. นำผลการวิจัยการจำแนกกลุ่มข้อมูลจำลองมาประยุกต์ใช้กับการจำแนกกลุ่มข้อมูลที่สนใจโดยอัลกอริทึม CHAID และพิจารณาการจำแนกกลุ่มข้อมูลจากค่าร้อยละเกนดัชนี
5. สรุปผลการจำแนกกลุ่มข้อมูลที่สนใจโดยอัลกอริทึม CHAID
6. ทดสอบประสิทธิภาพ และแก้ไขข้อผิดพลาด
7. วิเคราะห์และสรุปผลการศึกษา

3.2 การสร้างข้อมูลจำลองที่มีการแจกแจงแบบพหุนาม

การสร้างข้อมูลตามสถานการณ์ต่างๆ ที่มีการแจกแจงแบบพหุนามที่กำหนดไว้ในขอบเขตการวิจัยนั้นจะใช้โปรแกรม R เวอร์ชั่น 3.2.3 โดยขั้นตอนในการสร้างมีดังนี้ (ศศิธร เจษฎาฐิติกุล, 2545)

1. กำหนดการแจกแจงความน่าจะเป็นของตารางการณ์จรขนาด X x Y มีลักษณะดังตารางที่ 3

ตารางที่ 3 ตารางแจกแจงความน่าจะเป็นของแต่ละเซลล์ในตารางการณ์จรขนาด X x Y

(X) ตัวแปรอิสระ	(Y) ตัวแปรตาม				รวม
	Y_1	Y_2	...	Y_j	
X_1	p_{11}	p_{12}	...	p_{1j}	$p_{1.}$
X_2	p_{21}	p_{22}	...	p_{2j}	$p_{2.}$
X_3	p_{31}	p_{32}	...	p_{3j}	$p_{3.}$
⋮	⋮	⋮	...	⋮	⋮
X_i	p_{i1}	p_{i2}	...	p_{ij}	$p_{i.}$
รวม	$p_{.1}$	$p_{.2}$	...	$p_{.j}$	1

โดยที่ p_{ij} เป็นความน่าจะเป็นของค่าสังเกตใน X_i และ Y_j

$p_{i.}$ เป็นผลรวมความน่าจะเป็นของค่าสังเกตใน X_i หรือ $p_{i.} = \sum_{j=1}^{J} p_{ij}$

$p_{.j}$ เป็นผลรวมความน่าจะเป็นของค่าสังเกตใน Y_j หรือ $p_{.j} = \sum_{i=1}^{I} p_{ij}$

2. กำหนดความน่าจะเป็นส่วนริม (marginal probability) ของแถวและหลักจะได้ $p_{1.}, p_{2.},..., p_{i.}$ และ $p_{.1}, p_{.2},..., p_{.j}$

เมื่อ $p_{i.}$ คือ ความน่าจะเป็นส่วนริมของแถวที่ i เมื่อ i=1,...,I

และ $p_{.j}$ คือ ความน่าจะเป็นส่วนริมของแถวที่ j เมื่อ j=1,...,J

3. คำนวณค่าความน่าจะเป็นร่วม (p_{ij}) ของแต่ละเซลล์ในตารางการณ์จร

   3.1 กรณีตัวแปรทั้งสองไม่มีความสัมพันธ์กัน คำนวณค่าความน่าจะเป็นร่วม (p_{ij}) ได้โดย

   $p_{ij} = p_{i.}\, p_{.j}$

   3.2 กรณีตัวแปรทั้งสองมีความสัมพันธ์กัน โดยการคำนวณค่าความน่าจะเป็นร่วม (p_{ij}) จะพิจารณาจากค่า Goodman and Kruskal's tau (τ) ซึ่งคำนวณได้จากสูตร

   $\tau = \frac{\sum_i \sum_j {p_{ij}^2}/{p_{i.}} - \sum_j p_{.j}^2}{1 - \sum_j p_{.j}^2}$ ข้อมูลจะมีความสัมพันธ์กันมากขึ้นถ้า τ มีค่ามากขึ้น

4. สุ่มข้อมูลตามที่ได้กำหนด โดยใช้ฟังก์ชันของ runif(n,0,1) ซึ่งจะสร้างเลขสุ่มที่แจกแจงสม่ำเสมอในช่วงของ [0,1] และพิจารณาตัวเลขที่สุ่มได้ สมมติให้เป็น x_k

   เมื่อ $0 \le x_k \le p_{11}$ ค่าของข้อมูลจะอยู่ในเซลล์ (1,1)

   เมื่อ $p_{11} < x_k \le p_{11} + p_{12}$ ค่าของข้อมูลจะอยู่ในเซลล์ (1,2)

   เมื่อ $p_{11} + p_{12} < x_k \le p_{11} + p_{12} + p_{13}$ ค่าของข้อมูลจะอยู่ในเซลล์ (1,3)

   ⋮

   เมื่อ $p_{11}+\ldots+ p_{i,j-2} < x_k \le p_{11}+\ldots+ p_{i,j-1}$ ค่าของข้อมูลจะอยู่ในเซลล์ (i,j-1)

   เมื่อ $p_{11}+\ldots+ p_{i,j-1} < x_k \le p_{11}+\ldots+ p_{i,j}$ ค่าของข้อมูลจะอยู่ในเซลล์ (i,j)

   และจะสุ่มข้อมูลจนมีข้อมูลในตารางการณ์จรครบตามขนาดข้อมูลที่ได้กำหนดไว้

<u>ตัวอย่างที่ 1</u> การสร้างเลขสุ่มที่มีการแจกแจงพหุนามสองตัวแปรเมื่อตัวแปรทั้งสองไม่มีความสัมพันธ์กันสำหรับตารางการณ์จรขนาด 2x2 จะอยู่ในรูปตารางที่ 4

ตารางที่ 4 ตารางแจกแจงความน่าจะเป็นของแต่ละเซลล์ในตารางการณ์จรขนาด 2x2 เมื่อตัวแปรทั้งสองไม่มีความสัมพันธ์กัน

(X) ตัวแปรอิสระ	(Y) ตัวแปรตาม		รวม
	Y_1	Y_2	
X_1	p_{11}	p_{12}	$p_{1.}$
X_2	p_{21}	p_{22}	$p_{2.}$
รวม	$p_{.1}$	$p_{.2}$	1

โดยที่ $p_{1.} = p_{11} + p_{12}$, $p_{2.} = p_{21} + p_{22}$

$p_{.1} = p_{11} + p_{21}$, $p_{.2} = p_{12} + p_{22}$

เมื่อ $p_{11} + p_{12} + p_{21} + p_{22} = 1$

2. กำหนดค่าความน่าจะเป็นส่วนริมของแถวและหลัก

กำหนดให้ค่า $p_{1.} = 0.67$, $p_{2.} = 0.33$, $p_{.1} = 0.53$ และ $p_{.2} = 0.47$

3. คำนวณค่าความน่าจะเป็น (p_{ij}) ภายใต้สถานการณ์ที่ตัวแปรทั้งสองไม่มีความสัมพันธ์กัน โดย $p_{ij} = p_{i.} p_{.j}$ เมื่อ i=1,2, j=1,2

ดังนั้นจะได้ค่า $p_{11} = 0.3551$, $p_{12} = 0.3149$, $p_{21} = 0.1749$ และ $p_{22} = 0.1551$

ซึ่งค่าความน่าจะเป็น (p_{ij}) ของตารางการณ์จรจะอยู่ในรูปตารางที่ 5

ตารางที่ 5 ตัวอย่างตารางแจกแจงค่าความน่าจะเป็นของแต่ละเซลล์ในตารางการณ์จรขนาด 2x2 เมื่อตัวแปรทั้งสองไม่มีความสัมพันธ์กัน

(X) ตัวแปรอิสระ	(Y) ตัวแปรตาม		รวม
	Y_1	Y_2	
X_1	0.3551	0.3149	0.67
X_2	0.1749	0.1551	0.33
รวม	0.53	0.47	1

4. สุ่มข้อมูล x_k ขึ้นมาที่มีการแจกแจงสม่ำเสมอในช่วง [0,1]

เมื่อ $0 \le x_k \le 0.3551$ ค่า x_k อยู่ในเซลล์ (1,1)

เมื่อ $0.3551 < x_k \le 0.3551 + 0.3149 = 0.67$ ค่า x_k อยู่ในเซลล์ (1,2)

เมื่อ $0.67 < x_k \le 0.67 + 0.1749 = 0.8449$ ค่า x_k อยู่ในเซลล์ (2,1)

เมื่อ $0.8449 < x_k \le 0.8449 + 0.1551 = 1$ ค่า x_k อยู่ในเซลล์ (2,2)

เมื่อ x_k อยู่ในเซลล์ใดจะนับความถี่ในเซลล์ของตารางนั้นเพิ่มขึ้น 1 จากนั้นกลับไปสุ่ม x_k ใหม่ และสุ่มข้อมูลจนข้อมูลในตารางการณ์จรครบตามขนาดข้อมูลที่กำหนด

<u>ตัวอย่างที่ 2</u> การสร้างเลขสุ่มที่มีการแจกแจงพหุนามสองตัวแปรเมื่อตัวแปรทั้งสองมีความสัมพันธ์กัน สำหรับตารางการณ์จรขนาด 2x2 จะอยู่ในรูปตารางที่ 6

ตารางที่ 6 ตารางแจกแจงความน่าจะเป็นของแต่ละเซลล์ในตารางการณ์จรขนาด 2x2 เมื่อตัวแปรทั้งสองมีความสัมพันธ์กัน

(X) ตัวแปรอิสระ	(Y) ตัวแปรตาม		รวม
	Y_1	Y_2	
X_1	p_{11}	p_{12}	$p_{1.}$
X_2	p_{21}	p_{22}	$p_{2.}$
รวม	$p_{.1}$	$p_{.2}$	1

โดยที่ $p_{1.} = p_{11} + p_{12}$, $p_{2.} = p_{21} + p_{22}$

$p_{.1} = p_{11} + p_{21}$, $p_{.2} = p_{12} + p_{22}$

เมื่อ $p_{11} + p_{12} + p_{21} + p_{22} = 1$

2. กำหนดค่าความน่าจะเป็นส่วนริมของแถวและหลัก

กำหนดให้ค่า $p_{1.} = 0.36$, $p_{2.} = 0.64$, $p_{.1} = 0.35$ และ $p_{.2} = 0.65$

3. คำนวณค่าความน่าจะเป็น (p_{ij}) ภายใต้สถานการณ์ที่ตัวแปรทั้งสองมีความสัมพันธ์กันที่ระดับความสัมพันธ์ของข้อมูล (τ) เท่ากับ 0.3 จาก $\tau = \frac{\sum_i \sum_j {p_{ij}^2}/{p_{i.}} - \sum_j p_{.j}^2}{1 - \sum_j p_{.j}^2}$

ดังนั้นจะได้ค่า $p_{11} = 0.251$, $p_{12} = 0.109$, $p_{21} = 0.099$ และ $p_{22} = 0.541$

ซึ่งค่าความน่าจะเป็น (p_{ij}) ของตารางการณ์จรจะอยู่ในรูปตารางที่ 7

ตารางที่ 7 ตัวอย่างตารางแจกแจงค่าความน่าจะเป็นของแต่ละเซลล์ในตารางการณ์จรขนาด 2x2 เมื่อตัวแปรทั้งสองมีความสัมพันธ์กันที่ระดับ 0.3

(X) ตัวแปรอิสระ	(Y) ตัวแปรตาม		รวม
	Y_1	Y_2	
X_1	0.251	0.109	0.36
X_2	0.099	0.541	0.64
รวม	0.35	0.65	1

4. สุ่มข้อมูล x_k ขึ้นมาที่มีการแจกแจงสม่ำเสมอในช่วง [0,1]

เมื่อ $0 \le x_k \le 0.251$ ค่า x_k อยู่ในเซลล์ (1,1)

เมื่อ $0.251 < x_k \le 0.251 + 0.109 = 0.36$ ค่า x_k อยู่ในเซลล์ (1,2)

เมื่อ $0.36 < x_k \le 0.36 + 0.099 = 0.459$ ค่า x_k อยู่ในเซลล์ (2,1)

เมื่อ $0.459 < x_k \le 0.459 + 0.541 = 1$ ค่า x_k อยู่ในเซลล์ (2,2)

เมื่อ x_k อยู่ในเซลล์ใดจะนับความถี่ในเซลล์ของตารางนั้นเพิ่มขึ้น 1 จากนั้นกลับไปสุ่ม x_k ใหม่ และสุ่มข้อมูลจนข้อมูลในตารางการณ์จรครบตามขนาดข้อมูลที่กำหนด

3.3 ขั้นตอนการทำงานของโปรแกรมในการหาตัววัดประสิทธิภาพ

เมื่อได้ตัวอย่างสุ่มที่มีการแจกแจงพหุนาม จำนวนกลุ่มของตัวแปร และขนาดตัวอย่างตามขอบเขตการวิจัยที่กำหนดไว้แล้ว จะนำข้อมูลที่ได้ไปวิเคราะห์การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID เพื่อหาตัววัดประสิทธิภาพในการจำแนกกลุ่มข้อมูล ประกอบด้วยการแยก การรวม ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก อำนาจการทดสอบในขั้นตอนการแยก และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID

เมื่อตัวแปรทั้งสองไม่มีความสัมพันธ์กัน จะพิจารณาการแยก การรวม และความสามารถในการควบคุมความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยกของอัลกอริทึม CHAID เพื่อเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ในแต่ละสถานการณ์ ขั้นตอนในการคำนวณค่าดังกล่าวมีดังนี้

1. สุ่มตัวอย่างตามสถานการณ์ที่กำหนด โดยให้ตัวแปรทั้งสองไม่มีความสัมพันธ์กัน
2. ใช้อัลกอริทึม CHAID ในการจำแนกกลุ่มข้อมูล โดยในขั้นตอนการแยกจะคำนวณค่า p-value ของตัวสถิติทดสอบเพียร์สันไคสแควร์ แล้วเทียบกับระดับนัยสำคัญ เพื่อตรวจสอบการแยกตัวแปรอิสระ X ทำซ้ำๆกันเป็นจำนวน 1,000 รอบ
   2.1 ถ้าในแต่ละรอบ ค่า p-value น้อยกว่าหรือเท่ากับระดับนัยสำคัญ แล้วให้นับมีการแยกเพิ่ม 1 ครั้ง และมีการปฏิเสธสมมติฐานว่างเพิ่ม 1 ครั้ง แล้วไปขั้นตอนที่ 3
   2.2 ถ้าในแต่ละรอบ ค่า p-value มากกว่าระดับนัยสำคัญ แล้วไม่มีการแยกและไปรันข้อมูลรอบต่อไป

3. ในขั้นตอนการรวมจะคำนวณค่า p-value ของตัวสถิติทดสอบเพียร์สันไคสแควร์ แล้วเทียบกับระดับนัยสำคัญ เพื่อตรวจสอบการรวมกลุ่มตัวแปรอิสระ X จนไม่สามารถรวมได้อีก
   3.1 ถ้าค่า p-value น้อยกว่าหรือเท่ากับระดับนัยสำคัญ แล้วจะไม่มีการรวม แล้วไปรันข้อมูลรอบต่อไป
   3.2 ถ้าค่า p-value มากกว่าระดับนัยสำคัญ แล้วให้นับว่ามีการรวมเพิ่ม 1 ครั้ง ทำเช่นนี้ไปเรื่อยๆ จนกว่ากลุ่มของ X เหลือ 2 กลุ่ม หรือจนกว่า p-value น้อยกว่าหรือเท่ากับระดับนัยสำคัญ และนับจำนวนครั้งที่รวมสูงสุด ถ้ามีการรวมแค่ครั้งเดียว ให้นับมีการรวมครั้งที่ 1 1 ครั้ง ถ้ามีการรวมถึงสองครั้ง ให้นับมีการรวมครั้งที่ 2 1 ครั้ง ถ้ามีการรวมถึงสามครั้ง ให้นับมีการรวมครั้งที่ 3 1 ครั้ง และไปรันข้อมูลรอบต่อไป
4. เมื่อรันข้อมูลครบจำนวน 1,000 รอบแล้ว คำนวนหาค่าการแยก การรวม และค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก

เมื่อตัวแปรทั้งสองไม่มีความสัมพันธ์กัน จะพิจารณาการแยก การรวม อำนาจการทดสอบในขั้นตอนการแยก และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID เพื่อเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลในแต่ละสถานการณ์ ขั้นตอนในการคำนวณค่าดังกล่าวมีดังนี้

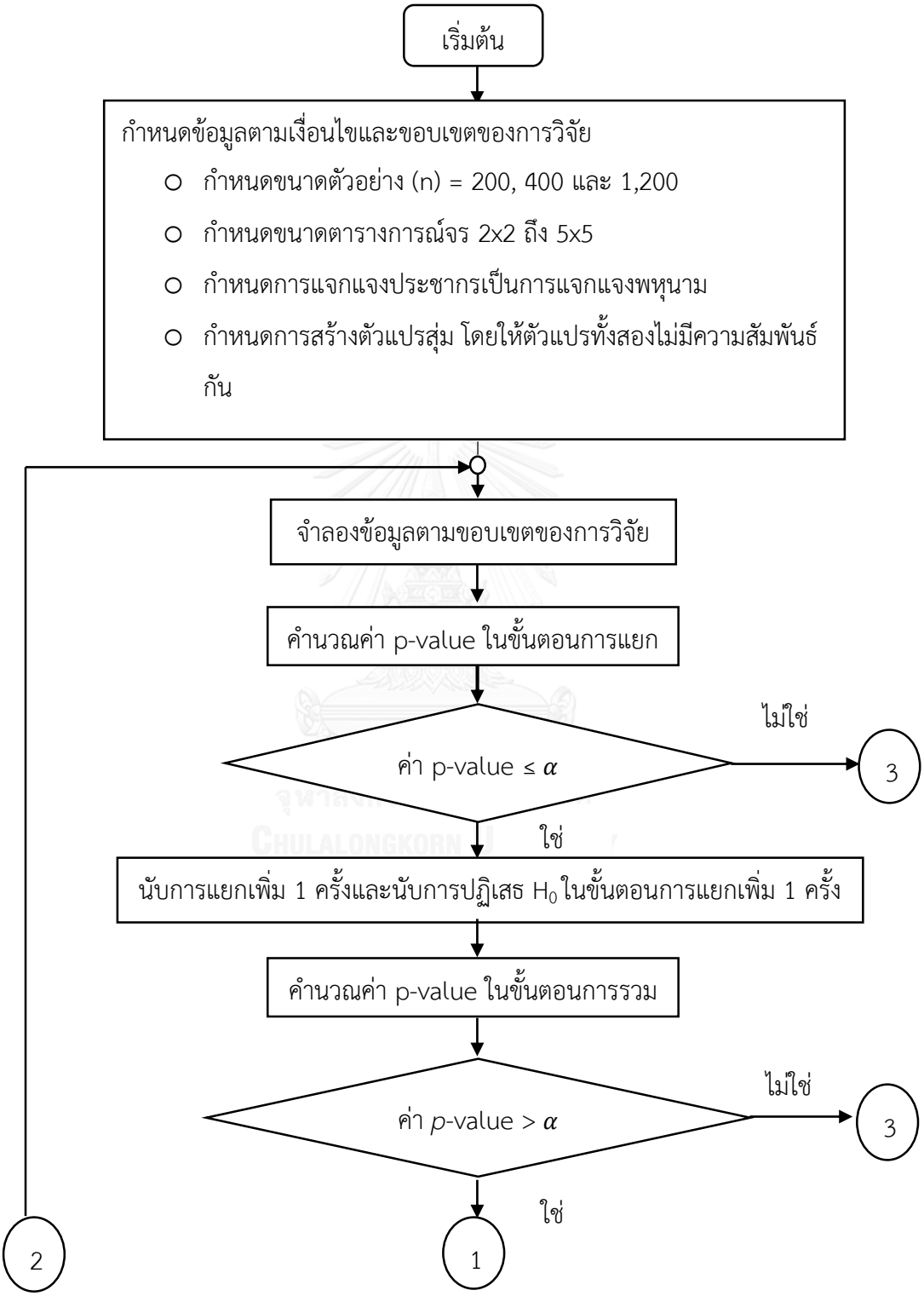
1. สุ่มตัวอย่างตามสถานการณ์ที่กำหนด โดยให้ตัวแปรทั้งสองมีความสัมพันธ์กัน
2. ใช้อัลกอริทึม CHAID ในการจำแนกกลุ่มข้อมูล โดยในขั้นตอนการแยกจะคำนวณค่า p-value ของตัวสถิติทดสอบเพียร์สันไคสแควร์ แล้วเทียบกับระดับนัยสำคัญ เพื่อตรวจสอบการแยกตัวแปรอิสระ X ทำซ้ำๆกันเป็นจำนวน 1,000 รอบ
   2.1 ถ้าในแต่ละรอบ ค่า p-value น้อยกว่าหรือเท่ากับระดับนัยสำคัญ แล้วให้นับมีการแยกเพิ่ม 1 ครั้ง และมีการปฏิเสธสมมติฐานว่างเพิ่ม 1 ครั้ง แล้วไปขั้นตอนที่ 3
   2.2 ถ้าในแต่ละรอบ ค่า p-value มากกว่าระดับนัยสำคัญ แล้วไม่มีการแยก และไปรันข้อมูลรอบต่อไป
3. ในขั้นตอนการรวมจะคำนวณค่า p-value ของตัวสถิติทดสอบเพียร์สันไคสแควร์ แล้วเทียบกับระดับนัยสำคัญ เพื่อตรวจสอบการรวมกลุ่มตัวแปรอิสระ X จนไม่สามารถรวมได้อีก

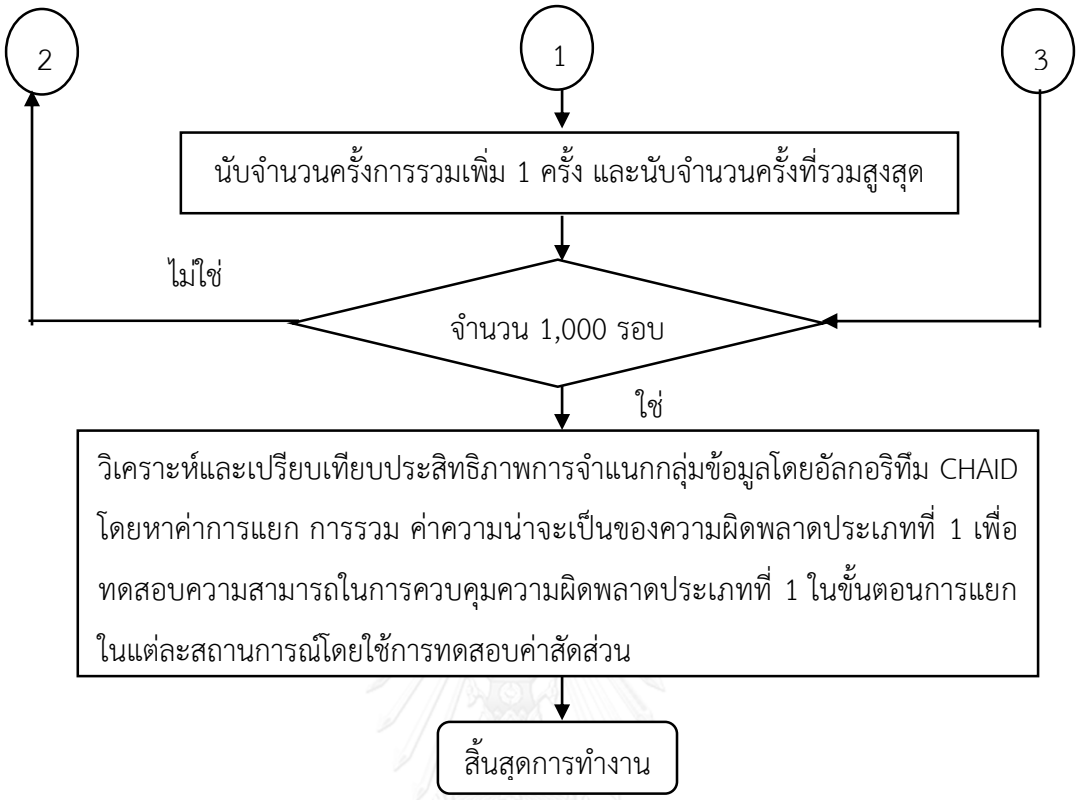
3.1 ถ้าค่า p-value น้อยกว่าหรือเท่ากับระดับนัยสำคัญ แล้วจะไม่มีการรวม แล้วไปขั้นตอนที่ 4

3.2 ถ้าค่า p-value มากกว่าระดับนัยสำคัญ แล้วให้นับว่ามีการรวมเพิ่ม 1 ครั้ง ทำเช่นนี้ไปเรื่อยๆ จนกว่ากลุ่มของ X เหลือ 2 กลุ่ม หรือจนกว่า p-value น้อยกว่าหรือเท่ากับระดับนัยสำคัญ และนับจำนวนครั้งที่รวมสูงสุด ถ้ามีการรวมแค่ครั้งเดียว ให้นับมีการรวมครั้งที่ 1 1 ครั้ง ถ้ามีการรวมถึงสองครั้ง ให้นับมีการการรวมครั้งที่ 2 1 ครั้ง ถ้ามีการรวมถึงสามครั้ง ให้นับมีการการรวมครั้งที่ 3 1 ครั้ง และไปขั้นตอนที่ 4

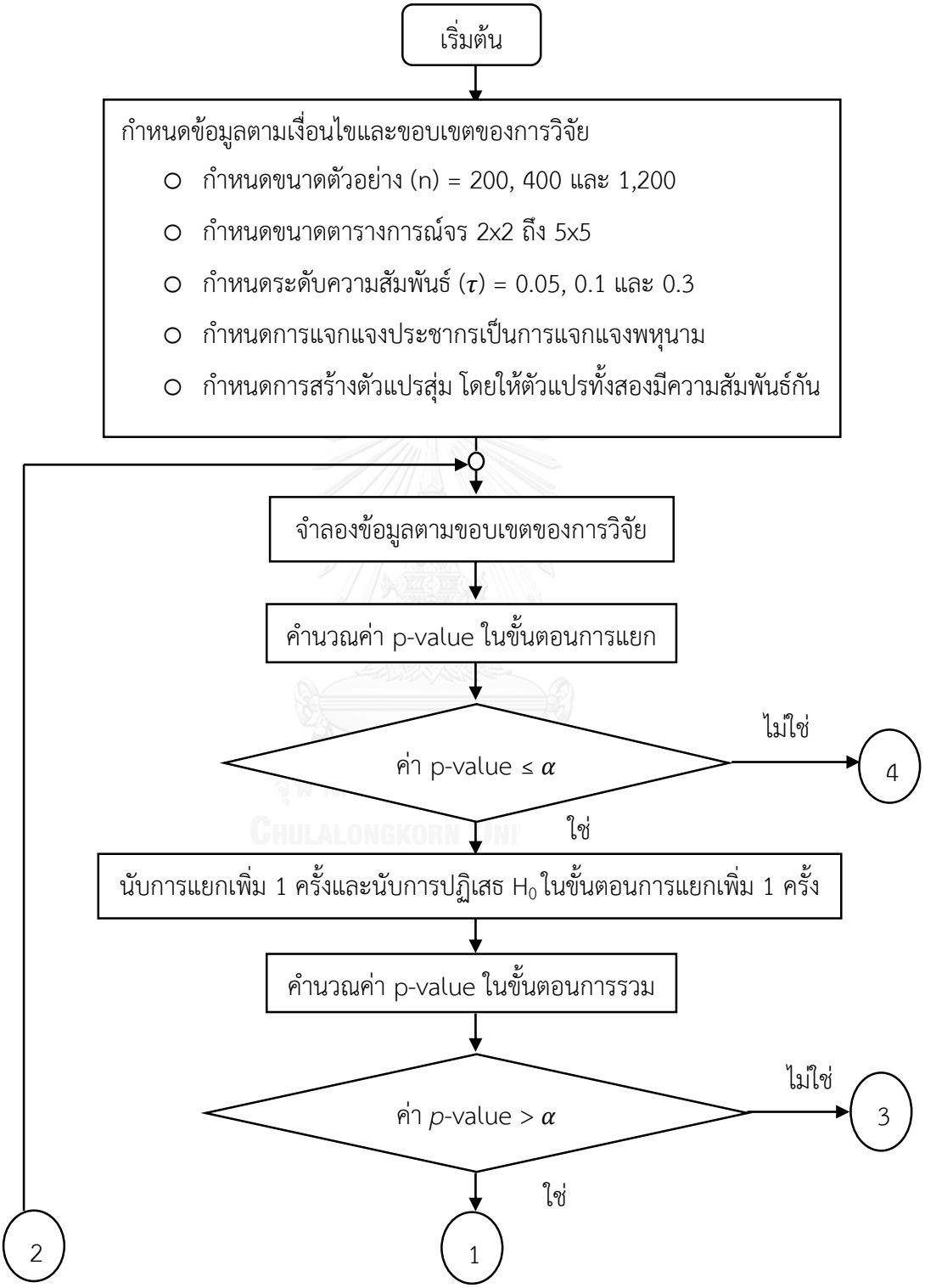
4. คำนวณร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูล แล้วไปรันข้อมูลรอบต่อไป
5. เมื่อรันข้อมูลครบจำนวน 1,000 รอบแล้ว คำนวนหาค่าการแยก การรวม อำนาจการทดสอบของแต่ละสถานการณ์ในขั้นตอนการแยก ซึ่งคำนวณได้จากจำนวนครั้งของการปฏิเสธสมมติฐานว่างในขั้นตอนการแยกหารด้วย 1,000 ส่วนค่าร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID คำนวณได้จากค่าเฉลี่ยของร้อยละความถูกต้องใน 1,000 รอบ

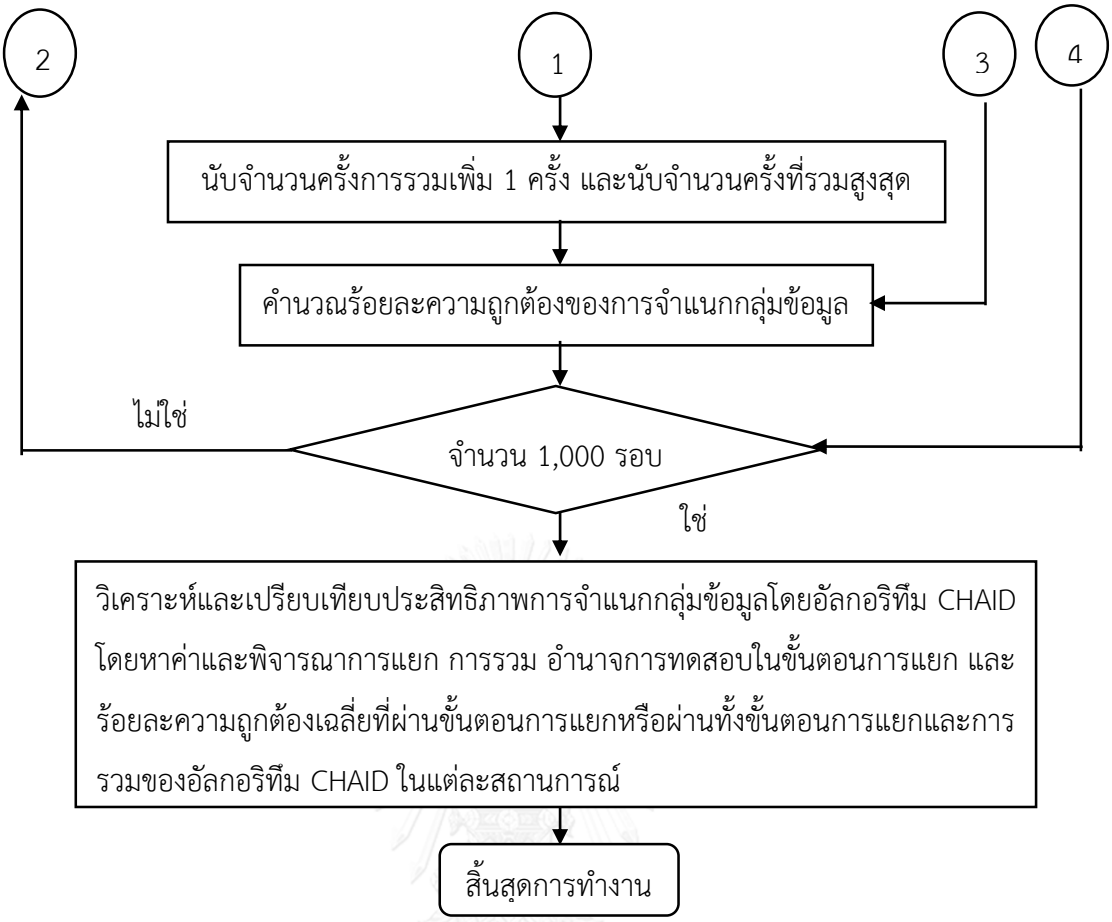
แผนผังที่ 2 ขั้นตอนการทำงานของโปรแกรมในการหาการแยก การรวม และความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ของการจำแนกกลุ่มข้อมูล





แผนผังที่ 3 ขั้นตอนการทำงานของโปรแกรมในการหาการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล





บทที่ 4
ผลการวิจัย

งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อเพื่อศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ระหว่างตัวแปร 2 ตัวแปร ที่มีการแจกแจงพหุนาม โดยตัวแปรแรกเป็นตัวแปรอิสระ X ที่มี I กลุ่ม และตัวแปรที่สองเป็นตัวแปรตาม Y ที่มี J กลุ่ม และอยู่ในตารางการณ์จรสองทาง XxY ที่มีขนาด IxJ แต่ละสถานการณ์ โดยพิจารณาการแยก การรวม ความสามารถการควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก อำนาจการทดสอบในขั้นตอนการแยก และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID ณ ขนาดตัวอย่าง ระดับความสัมพันธ์ของข้อมูลและระดับนัยสำคัญต่างๆ และนำผลวิจัยจากข้อมูลจำลองมาศึกษาจำแนกกลุ่มข้อมูลที่สนใจเพื่อหาผลลัพธ์การจำแนกกลุ่มข้อมูลที่มีประสิทธิภาพและระบุกลุ่มเป้าหมายของข้อมูลที่ดีที่สุด โดยพิจารณาค่าร้อยละเกนดัชนี

อักษรย่อและสัญลักษณ์ต่างๆ ที่ปรากฏในการเสนอผลการวิจัยทั้งในตารางและข้อความต่างๆ แทนความหมายดังต่อไปนี้

n	แทน ขนาดตัวอย่าง
α	แทน ระดับนัยสำคัญที่กำหนด
τ	แทน ระดับความสัมพันธ์
Type I	แทน ค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1
$\overline{\mathrm{CR}}$	แทน ค่าร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล

สำหรับงานวิจัยชิ้นนี้จะนำเสนอผลการวิจัยและผลการเปรียบเทียบโดยแบ่งออกเป็น 3 ส่วน คือ ในส่วนที่ 1 จะพิจารณาและเปรียบเทียบการแยก การรวม และความสามารถการควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก ส่วนที่ 2 พิจารณาและเปรียบเทียบการแยก การรวม อำนาจการทดสอบในขั้นตอนการแยก และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID และส่วนที่ 3 พิจารณาและเปรียบเทียบผลลัพธ์การจำแนกกลุ่มข้อมูลที่สนใจทางการตลาดและกลุ่มเป้าหมายที่ดีที่สุด

4.1 เปรียบเทียบตัววัดประสิทธิภาพกรณีตัวแปรทั้งสองไม่มีความสัมพันธ์กัน

โดยจะนำเสนอค่าการแยก การรวม ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ที่ระดับนัยสำคัญ (α) 0.05 และ 0.1 สำหรับการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ตามขนาดตารางและขนาดตัวอย่างของข้อมูลที่ไม่มีระดับความสัมพันธ์ ดังนี้

- ตารางการณ์จรขนาด 2x2, 2x3, 2x4, 2x5, 3x2, 3x3, 3x4, 3x5, 4x2, 4x3, 4x4, 4x5, 5x2, 5x3, 5x4 และ 5x5
- ขนาดตัวอย่าง 200 400 และ 1,200 ในแต่ละตารางการณ์จร

ความสามารถในการควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1

ในการวิจัยนี้จะทำการหาความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ที่ระดับนัยสำคัญ 0.05 และ 0.1 โดยทำการจำลองข้อมูลทั้งหมด 1,000 ครั้ง ในแต่ละสถานการณ์ และจะพิจารณาว่าสามารถควบคุมความผิดพลาดประเภทที่ 1 ได้หรือไม่ จากความสามารถการควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ที่เคยกล่าวมา ดังนั้น

ที่ระดับนัยสำคัญ 0.05 สถานการณ์ของการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ว่าสามารถควบคุมความผิดพลาดประเภทที่ 1 เมื่อ Type I ≤ 0.061

และที่ระดับนัยสำคัญ 0.1 สถานการณ์ของการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ว่าสามารถควบคุมความผิดพลาดประเภทที่ 1 เมื่อ Type I ≤ 0.112

ในผลการวิจัยในส่วนที่ 1 ผู้วิจัยจะนำเสนอในรูปแบบตารางที่ 8, ตารางที่ 9, ตารางที่ 10 และตารางที่ 11 ซึ่งมีรายละเอียดของค่าการแยก การรวมของการจำแนกกลุ่มข้อมูล ซึ่งจำนวนครั้งในการรวมขึ้นอยู่กับจำนวนกลุ่มของตัวแปรอิสระ ถ้าจำนวนกลุ่มของตัวแปรอิสระเท่ากับ I กลุ่ม แล้วจำนวนครั้งในการรวมสูงสุดเท่ากับ I-2 ครั้ง กล่าวคือ ถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 2 จะไม่มีการรวม ถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 3 จะมีการรวมมากสุด 1 ครั้ง ถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 4 จะมีการรวมมากสุด 2 ครั้ง และถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 5 จะมีการรวมมากสุด 3 ครั้ง พร้อมทั้งค่าความน่าจะเป็นของความผิดพลาดประเภทที่ 1 (Type I) ในขั้นตอนการแยกดังต่อไปนี้

ตารางที่ 8 แสดงการแยก การรวม ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 สำหรับตามขนาดตาราง ขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05

ขนาดตาราง	ขนาดตัวอย่าง	ตัววัดประสิทธิภาพ					
		การแยก	การรวม				Type I
			1	2	3	รวม	
2x2	200	33	-	-	-	-	0.033
	400	34	-	-	-	-	0.034
	1,200	39	-	-	-	-	0.039
2x3	200	45	-	-	-	-	0.045
	400	41	-	-	-	-	0.041
	1,200	44	-	-	-	-	0.044
2x4	200	43	-	-	-	-	0.043
	400	53	-	-	-	-	0.053
	1,200	50	-	-	-	-	0.050
2x5	200	46	-	-	-	-	0.046
	400	54	-	-	-	-	0.054
	1,200	52	-	-	-	-	0.052
3x2	200	48	48	-	-	48	0.048
	400	47	47	-	-	47	0.047
	1,200	51	51	-	-	51	0.051
3x3	200	33	33	-	-	33	0.033
	400	53	53	-	-	53	0.053
	1,200	54	54	-	-	54	0.054
3x4	200	52	52	-	-	52	0.052
	400	44	44	-	-	44	0.044
	1,200	50	49	-	-	49	0.050
3x5	200	51	51	-	-	51	0.051
	400	49	48	-	-	48	0.049
	1,200	43	43	-	-	43	0.043

ตารางที่ 9 แสดงการแยก การรวม ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 สำหรับตามขนาดตาราง ขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05

ขนาดตาราง	ขนาดตัวอย่าง	ตัววัดประสิทธิภาพ					
		การแยก	การรวม				Type I
			1	2	3	รวม	
4x2	200	51	0	51	-	51	0.051
	400	36	1	35	-	36	0.036
	1,200	45	1	44	-	45	0.045
4x3	200	26	2	24	-	26	0.026
	400	55	2	53	-	55	0.055
	1,200	55	3	52	-	55	0.055
4x4	200	48	2	46	-	48	0.048
	400	53	4	49	-	53	0.053
	1,200	40	3	37	-	40	0.040
4x5	200	36	1	35	-	36	0.036
	400	37	3	34	-	37	0.037
	1,200	49	3	46	-	49	0.049
5x2	200	51	0	0	51	51	0.051
	400	44	0	2	42	44	0.044
	1,200	53	0	3	50	53	0.053
5x3	200	38	0	4	34	38	0.038
	400	48	0	9	39	48	0.048
	1,200	45	0	7	38	45	0.045
5x4	200	36	0	9	27	36	0.036
	400	54	0	9	45	54	0.054
	1,200	57	0	11	46	57	0.057
5x5	200	28	0	5	23	28	0.028
	400	40	0	6	34	40	0.040
	1,200	56	0	8	48	56	0.056

ตารางที่ 10 แสดงการแยก การรวม ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 สำหรับตามขนาดตาราง ขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.1

ขนาดตาราง	ขนาดตัวอย่าง	ตัววัดประสิทธิภาพ					
		การแยก	การรวม				Type I
			1	2	3	รวม	
2x2	200	73	-	-	-	-	0.073
	400	75	-	-	-	-	0.075
	1,200	86	-	-	-	-	0.086
2x3	200	93	-	-	-	-	0.093
	400	91	-	-	-	-	0.091
	1,200	83	-	-	-	-	0.083
2x4	200	94	-	-	-	-	0.094
	400	107	-	-	-	-	0.107
	1,200	93	-	-	-	-	0.093
2x5	200	96	-	-	-	-	0.096
	400	103	-	-	-	-	0.103
	1,200	109	-	-	-	-	0.109
3x2	200	102	102	-	-	102	0.102
	400	101	101	-	-	101	0.101
	1,200	85	85	-	-	85	0.085
3x3	200	98	93	-	-	93	0.098
	400	100	97	-	-	97	0.100
	1,200	102	97	-	-	97	0.102
3x4	200	88	86	-	-	86	0.088
	400	108	108	-	-	108	0.108
	1,200	108	105	-	-	105	0.105
3x5	200	69	67	-	-	67	0.069
	400	77	73	-	-	73	0.077
	1,200	109	107	-	-	107	0.109

ตารางที่ 11 แสดงการแยก การรวม ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 สำหรับตามขนาดตาราง ขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.1

ขนาดตาราง	ขนาดตัวอย่าง	ตัววัดประสิทธิภาพ					
		การแยก	การรวม				Type I
			1	2	3	รวม	
4x2	200	80	0	80	-	80	0.080
	400	91	1	90	-	91	0.091
	1,200	81	1	80	-	81	0.081
4x3	200	98	12	86	-	98	0.098
	400	93	9	84	-	93	0.093
	1,200	110	11	99	-	110	0.110
4x4	200	81	10	71	-	81	0.081
	400	94	15	79	-	94	0.094
	1,200	95	17	77	-	94	0.095
4x5	200	56	5	51	-	56	0.056
	400	95	15	80	-	95	0.095
	1,200	97	15	81	-	96	0.097
5x2	200	88	0	4	84	88	0.088
	400	84	0	5	79	84	0.084
	1,200	102	0	5	97	102	0.102
5x3	200	103	0	30	73	103	0.103
	400	70	0	20	50	70	0.070
	1,200	99	0	30	69	99	0.099
5x4	200	94	0	24	70	94	0.094
	400	80	0	27	53	80	0.080
	1,200	101	0	39	62	101	0.101
5x5	200	84	0	29	55	84	0.084
	400	90	0	26	64	90	0.090
	1,200	110	0	41	69	110	0.110

จากตารางที่ 8, ตารางที่ 9, ตารางที่ 10 และตารางที่ 11 สามารถสรุปผลดังนี้

ณ ข้อมูลตัวแปรทั้งสองไม่มีความสัมพันธ์กัน การแยกจะมีค่าเพิ่มขึ้นเมื่อ α เพิ่มขึ้นที่ขนาดตัวอย่างคงที่ และการรวมจะมีค่าใกล้เคียงกับการแยก กล่าวคือ เมื่อมีการแยกจะมีการรวมทุกครั้ง และมีแนวโน้มจะทำการรวมมากสุดถึงครั้งที่ I-2 เนื่องจากเมื่อข้อมูลตัวแปรทั้งสองไม่มีความสัมพันธ์กัน ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าลดลง ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าเพิ่มขึ้น ทำให้จำนวนการแยกลดลง และการรวมเพิ่มขึ้นในการจำแนกกลุ่มข้อมูล

และทุกสถานการณ์ในการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID สามารถควบคุมความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยกได้ เนื่องจาก ที่ระดับนัยสำคัญ 0.05 ค่า Type I ของทุกสถานการณ์มีค่า ≤ 0.061 และ ที่ระดับนัยสำคัญ 0.1 ค่า Type I ของทุกสถานการณ์มีค่า ≤ 0.112

4.2 เปรียบเทียบตัววัดประสิทธิภาพกรณีตัวแปรทั้งสองมีความสัมพันธ์กัน

ในส่วนที่ 2 ของงานวิจัยนี้จะนำเสนอค่าการแยก การรวม อำนาจการทดสอบในขั้นตอนการแยก และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID ที่ระดับนัยสำคัญ (α) 0.05 และ 0.1 สำหรับการจำแนกกลุ่มข้อมูลตามขนาดตาราง ขนาดตัวอย่าง และระดับความสัมพันธ์ของข้อมูล ดังนี้

- ตารางการณ์จรขนาด 2x2, 2x3, 2x4, 2x5, 3x2, 3x3, 3x4, 3x5, 4x2, 4x3, 4x4, 4x5, 5x2, 5x3, 5x4 และ 5x5
- ขนาดตัวอย่าง 200 400 และ 1,200 ในแต่ละตารางการณ์จร
- ระดับความสัมพันธ์ของข้อมูลเท่ากับ 0.05 0.1 และ 0.3 สำหรับแต่ละตารางการณ์จร

ในผลการวิจัยในส่วนที่ 2 ผู้วิจัยจะนำเสนอในรูปแบบตารางที่ 12 – 27 ซึ่งมีรายละเอียดของค่าการแยก การรวมของการจำแนกกลุ่มข้อมูล ซึ่งจำนวนครั้งในการรวมขึ้นอยู่กับจำนวนกลุ่มของตัวแปรอิสระ ถ้าจำนวนกลุ่มของตัวแปรอิสระเท่ากับ I กลุ่ม แล้วจำนวนครั้งในการรวมสูงสุดเท่ากับ I-2 ครั้ง กล่าวคือ ถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 2 จะไม่มีการรวม ถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 3 จะมีการรวมมากสุด 1 ครั้ง ถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 4 จะมีการรวมมากสุด 2 ครั้ง และถ้ากลุ่มของตัวแปรอิสระ X เท่ากับ 5 จะมีการรวมมากสุด 3 ครั้ง พร้อมอำนาจการทดสอบในขั้นตอนกาแยกและร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล ($\overline{\text{CR}}$) ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID ดังต่อไปนี้

ตารางที่ 12 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 2x2 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	883	-	-	-	-	0.883	66.5810
		400	989	-	-	-	-	0.989	65.8640
		1,200	1,000	-	-	-	-	1	65.4403
	0.1	200	998	-	-	-	-	0.998	65.6503
		400	1,000	-	-	-	-	1	65.3298
		1,200	1,000	-	-	-	-	1	65.0504
	0.3	200	1,000	-	-	-	-	1	79.2840
		400	1,000	-	-	-	-	1	79.2448
		1,200	1,000	-	-	-	-	1	79.1809
0.1	0.05	200	900	-	-	-	-	0.900	66.5250
		400	999	-	-	-	-	0.999	65.9555
		1,200	1,000	-	-	-	-	1	65.4327
	0.1	200	1,000	-	-	-	-	1	65.6630
		400	1,000	-	-	-	-	1	65.2550
		1,200	1,000	-	-	-	-	1	64.9725
	0.3	200	1,000	-	-	-	-	1	79.1856
		400	1,000	-	-	-	-	1	79.1110
		1,200	1,000	-	-	-	-	1	79.2353

จากตารางที่ 12 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 และจะมีการแยกที่ลดลงตามขนาดตัวอย่างและระดับนัยสำคัญที่ลดลง ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 และจะมีอำนาจการทดสอบที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 13 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 2x3 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	950	-	-	-	-	0.950	58.3168
		400	997	-	-	-	-	0.997	57.9321
		1,200	1,000	-	-	-	-	1	57.9970
	0.1	200	999	-	-	-	-	0.999	62.9820
		400	1,000	-	-	-	-	1	62.9065
		1,200	1,000	-	-	-	-	1	63.0337
	0.3	200	1,000	-	-	-	-	1	74.9435
		400	1,000	-	-	-	-	1	75.0145
		1,200	1,000	-	-	-	-	1	75.0163
0.1	0.05	200	962	-	-	-	-	0.962	58.2656
		400	999	-	-	-	-	0.999	58.0688
		1,200	1,000	-	-	-	-	1	57.9707
	0.1	200	1,000	-	-	-	-	1	62.9660
		400	1,000	-	-	-	-	1	62.9793
		1,200	1,000	-	-	-	-	1	62.9368
	0.3	200	1,000	-	-	-	-	1	74.7440
		400	1,000	-	-	-	-	1	74.9653
		1,200	1,000	-	-	-	-	1	75.0047

จากตารางที่ 13 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 และจะมีการแยกที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 และจะมีอำนาจการทดสอบที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 14 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 2x4 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	996	-	-	-	-	0.996	43.3007
		400	1,000	-	-	-	-	1	42.3420
		1,200	1,000	-	-	-	-	1	41.5650
	0.1	200	1,000	-	-	-	-	1	49.9925
		400	1,000	-	-	-	-	1	49.9988
		1,200	1,000	-	-	-	-	1	50.0679
	0.3	200	1,000	-	-	-	-	1	62.0270
		400	1,000	-	-	-	-	1	61.6010
		1,200	1,000	-	-	-	-	1	61.1105
0.1	0.05	200	1000	-	-	-	-	1	43.3485
		400	1,000	-	-	-	-	1	42.3363
		1,200	1,000	-	-	-	-	1	41.4603
	0.1	200	1,000	-	-	-	-	1	49.9045
		400	1,000	-	-	-	-	1	49.9400
		1,200	1,000	-	-	-	-	1	49.9713
	0.3	200	1,000	-	-	-	-	1	61.789
		400	1,000	-	-	-	-	1	61.6338
		1,200	1,000	-	-	-	-	1	61.1292

จากตารางที่ 14 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันทุกระดับความสัมพันธ์ ขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 996 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

อำนาจการทดสอบมีค่าเท่ากันทุกระดับความสัมพันธ์ ขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.996 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ จากมากไปหาน้อยตามลำดับ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 15 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 2x5 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	989	-	-	-	-	0.989	56.2517
		400	1,000	-	-	-	-	1	56.0533
		1,200	1,000	-	-	-	-	1	55.5962
	0.1	200	1,000	-	-	-	-	1	59.0940
		400	1,000	-	-	-	-	1	59.0795
		1,200	1,000	-	-	-	-	1	58.9590
	0.3	200	1,000	-	-	-	-	1	70.7540
		400	1,000	-	-	-	-	1	71.0343
		1,200	1,000	-	-	-	-	1	71.0410
0.1	0.05	200	991	-	-	-	-	0.991	56.4652
		400	1,000	-	-	-	-	1	55.9605
		1,200	1,000	-	-	-	-	1	55.6504
	0.1	200	1,000	-	-	-	-	1	59.1515
		400	1,000	-	-	-	-	1	58.9568
		1,200	1,000	-	-	-	-	1	58.9589
	0.3	200	1,000	-	-	-	-	1	70.7060
		400	1,000	-	-	-	-	1	70.9010
		1,200	1,000	-	-	-	-	1	70.9793

จากตารางที่ 15 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 989 และที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 991 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.989 และ ที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.991 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 16 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 3x2 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	852	839	-	-	839	0.852	59.36678
		400	991	950	-	-	950	0.991	58.4760
		1,200	1,000	931	-	-	931	1	58.1605
	0.1	200	992	948	-	-	948	0.992	65.3664
		400	1,000	916	-	-	916	1	65.1880
		1,200	1,000	754	-	-	754	1	65.0230
	0.3	200	1,000	654	-	-	654	1	75.7680
		400	1,000	329	-	-	329	1	75.9715
		1,200	1,000	18	-	-	18	1	76.0230
0.1	0.05	200	991	850	-	-	850	0.991	59.4325
		400	994	886	-	-	886	0.994	58.8011
		1,200	1,000	857	-	-	857	1	58.2379
	0.1	200	996	907	-	-	907	1	65.5768
		400	1,000	838	-	-	838	1	65.1503
		1,200	1,000	645	-	-	645	1	65.0198
	0.3	200	1,000	542	-	-	542	1	75.9010
		400	1,000	224	-	-	224	1	76.0118
		1,200	1,000	3	-	-	3	1	75.9903

จากตารางที่ 16 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีการแยกที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมจะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมมีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมแปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีอำนาจการทดสอบที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 17 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 3x3 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	983	627	-	-	627	0.983	45.1673
		400	1,000	159	-	-	159	1	44.6640
		1,200	1,000	0	-	-	0	1	44.2637
	0.1	200	1,000	248	-	-	248	1	51.2550
		400	1,000	24	-	-	24	1	51.2265
		1,200	1,000	0	-	-	0	1	50.9691
	0.3	200	1,000	241	-	-	241	1	65.4515
		400	1,000	22	-	-	22	1	65.8183
		1,200	1,000	0	-	-	0	1	66.0898
0.1	0.05	200	994	445	-	-	445	0.994	45.3692
		400	1,000	93	-	-	93	1	44.6988
		1,200	1,000	0	-	-	0	1	44.2695
	0.1	200	1,000	134	-	-	134	1	51.2910
		400	1,000	11	-	-	11	1	51.0493
		1,200	1,000	0	-	-	0	1	51.1113
	0.3	200	1,000	154	-	-	154	1	65.6030
		400	1,000	11	-	-	11	1	66.0753
		1,200	1,000	0	-	-	0	1	66.0045

จากตารางที่ 17 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 983 และที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 994 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมจะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมมีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมแปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.983 และ ที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.994 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 18 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 3x4 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	1,000	285	-	-	285	1	55.2010
		400	1,000	33	-	-	33	1	55.5280
		1,200	1,000	0	-	-	0	1	55.1075
	0.1	200	1,000	66	-	-	66	1	60.0315
		400	1,000	0	-	-	0	1	59.9975
		1,200	1,000	0	-	-	0	1	60.0208
	0.3	200	1,000	0	-	-	0	1	66.4510
		400	1,000	0	-	-	0	1	66.5743
		1,200	1,000	0	-	-	0	1	66.1464
0.1	0.05	200	1,000	150	-	-	150	1	55.6400
		400	1,000	13	-	-	13	1	55.3693
		1,200	1,000	0	-	-	0	1	55.2221
	0.1	200	1,000	23	-	-	23	1	59.8390
		400	1,000	0	-	-	0	1	60.1280
		1,200	1,000	0	-	-	0	1	59.9633
	0.3	200	1,000	0	-	-	0	1	66.4395
		400	1,000	0	-	-	0	1	66.4573
		1,200	1,000	0	-	-	0	1	66.1081

จากตารางที่ 18 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันทุกระดับความสัมพันธ์และขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1

การรวมจะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมมีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมแปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันทุกระดับความสัมพันธ์และขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 19 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 3x5 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	994	392	-	-	392	0.994	42.0981
		400	1,000	27	-	-	27	1	42.0210
		1,200	1,000	0	-	-	0	1	42.0218
	0.1	200	1,000	828	-	-	828	1	45.0580
		400	1,000	631	-	-	631	1	45.2418
		1,200	1,000	108	-	-	108	1	45.0448
	0.3	200	1,000	0	-	-	0	1	66.3300
		400	1,000	0	-	-	0	1	66.9130
		1,200	1,000	0	-	-	0	1	67.0028
0.1	0.05	200	994	246	-	-	246	0.994	42.4336
		400	1,000	13	-	-	13	1	42.2465
		1,200	1,000	0	-	-	0	1	42.0787
	0.1	200	1,000	730	-	-	730	1	45.2665
		400	1,000	470	-	-	470	1	45.1500
		1,200	1,000	65	-	-	65	1	45.0538
	0.3	200	1,000	0	-	-	0	1	66.2285
		400	1,000	0	-	-	0	1	67.0145
		1,200	1,000	0	-	-	0	1	66.9372

จากตารางที่ 19 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 994 และที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 994 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมจะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมมีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมแปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.994 และ ที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.994 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 20 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 4x2 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	722	40	682	-	722	0.722	78.2749
		400	971	449	522	-	971	0.971	78.1437
		1,200	1,000	793	15	-	808	1	78.0451
	0.1	200	957	32	925	-	957	0.957	78.8135
		400	999	247	747	-	994	0.999	78.9715
		1,200	1,000	643	199	-	842	1	78.9821
	0.3	200	1,000	643	357	-	1,000	1	84.0825
		400	1,000	851	24	-	875	1	84.0283
		1,200	1,000	100	0	-	100	1	84.0753
0.1	0.05	200	812	149	663	-	812	0.812	78.1386
		400	983	648	320	-	968	1	78.0468
		1,200	1,000	648	3	-	651	1	78.0555
	0.1	200	976	103	873	-	976	0.976	78.9800
		400	1,000	397	569	-	966	1	79.0448
		1,200	1,000	619	104	-	723	1	79.0705
	0.3	200	1,000	756	226	-	982	1	84.3425
		400	1,000	720	11	-	731	1	84.0950
		1,200	1,000	40	0	-	40	1	84.0178

จากตารางที่ 20 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีการแยกที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 2 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 2 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่ 2 หรือการรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีอำนาจการทดสอบที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 21 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 4x3 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	926	476	411	-	887	0.926	52.4622
		400	1,000	467	37	-	504	1	52.7510
		1,200	1,000	2	0	-	2	1	52.9682
	0.1	200	999	869	86	-	955	0.999	56.3669
		400	1,000	779	0	-	779	1	56.3905
		1,200	1,000	349	0	-	349	1	56.4198
	0.3	200	1,000	91	3	-	94	1	69.3245
		400	1,000	0	0	-	0	1	69.4838
		1,200	1,000	0	0	-	0	1	69.1835
0.1	0.05	200	974	601	222	-	823	0.974	52.8137
		400	1,000	288	14	-	302	1	53.0123
		1,200	1,000	0	0	-	0	1	53.0276
	0.1	200	1,000	824	22	-	846	1	56.5550
		400	1,000	639	2	-	641	1	56.6305
		1,200	1,000	236	0	-	236	1	56.4039
	0.3	200	1,000	46	0	-	46	1	69.3915
		400	1,000	0	0	-	0	1	69.6078
		1,200	1,000	0	0	-	0	1	69.2579

จากตารางที่ 21 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีการแยกที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้นหรือขนาดเพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 2 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 2 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่ 2 หรือจำนวนการรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีอำนาจการทดสอบที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 22 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 4x4 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	864	113	751	-	864	0.864	59.3808
		400	1,000	347	637	-	984	1	59.6630
		1,200	1,000	586	126	-	712	1	59.2571
	0.1	200	1,000	245	3	-	248	1	60.7390
		400	1,000	15	0	-	15	1	60.9270
		1,200	1,000	0	0	-	0	1	61.0413
	0.3	200	1,000	270	24	-	294	1	71.1605
		400	1,000	12	0	-	12	1	71.9275
		1,200	1,000	0	0	-	0	1	71.9554
0.1	0.05	200	931	239	683	-	922	0.931	59.5252
		400	1,000	480	461	-	941	1	59.5498
		1,200	1,000	512	51	-	563	1	59.2262
	0.1	200	1,000	163	0	-	163	1	60.6965
		400	1,000	9	0	-	9	1	60.9518
		1,200	1,000	0	0	-	0	1	61.0549
	0.3	200	1,000	181	7	-	188	1	71.1540
		400	1,000	6	0	-	6	1	71.9313
		1,200	1,000	0	0	-	0	1	71.9808

จากตารางที่ 22 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 864 และที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 931 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 2 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 2 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง จากมากไปน้อย

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.864 และ ที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.931 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 23 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 4x5 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	987	823	160	-	983	0.987	51.4119
		400	1,000	910	3	-	913	1	51.9253
		1,200	1,000	359	0	-	359	1	51.9909
	0.1	200	1,000	746	254	-	1,000	1	52.8925
		400	1,000	980	5	-	985	1	53.8663
		1,200	1,000	951	0	-	951	1	53.9531
	0.3	200	1,000	748	0	-	748	1	61.9040
		400	1,000	301	0	-	301	1	62.9433
		1,200	1,000	4	0	-	4	1	62.9563
0.1	0.05	200	992	881	96	-	977	0.992	51.0494
		400	1,000	834	0	-	834	1	51.9235
		1,200	1,000	252	0	-	252	1	51.9843
	0.1	200	1,000	847	148	-	995	1	53.2240
		400	1,000	941	4	-	945	1	53.9483
		1,200	1,000	902	0	-	920	1	53.9687
	0.3	200	1,000	620	0	-	620	1	62.0560
		400	1,000	182	0	-	182	1	63.0080
		1,200	1,000	1	0	-	1	1	63.0316

จากตารางที่ 23 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 987 และที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 992 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 2 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 2 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง จากมากไปน้อย

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.987 และ ที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.992 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 24 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 5x2 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	704	0	99	605	704	0.704	75.9737
		400	975	1	378	596	975	0.975	75.9259
		1,200	1,000	51	858	91	1,000	1	75.9698
	0.1	200	988	0	166	822	988	0.988	76.1108
		400	1,000	2	425	573	1,000	1	76.0240
		1,200	1,000	44	773	183	1,000	1	75.9460
	0.3	200	1,000	17	918	65	1,000	1	82.7355
		400	1,000	154	843	0	997	1	82.9470
		1,200	1,000	494	472	0	966	1	82.9805
0.1	0.05	200	809	0	217	592	809	0.809	76.0717
		400	992	4	570	418	992	0.992	76.0149
		1,200	1,000	124	827	47	998	1	76.0110
	0.1	200	999	0	331	668	999	0.999	76.1436
		400	1,000	10	549	441	1,000	1	76.0903
		1,200	1,000	101	789	110	1,000	1	76.0447
	0.3	200	1,000	57	910	33	1,000	1	82.7890
		400	1,000	267	725	0	992	1	83.0200
		1,200	1,000	635	296	0	9321	1	82.9915

จากตารางที่ 24 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีการแยกที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 3 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 3 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีอำนาจการทดสอบที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 25 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 5x3 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	860	1	146	713	860	0.860	51.3116
		400	999	8	296	695	999	0.999	50.5253
		1,200	1,000	101	710	185	996	1	50.3081
	0.1	200	1,000	388	544	45	977	1	51.0325
		400	1,000	633	77	0	710	1	50.7613
		1,200	1,000	74	0	0	74	1	50.5008
	0.3	200	1,000	848	119	0	967	1	65.2890
		400	1,000	799	1	0	800	1	65.3608
		1,200	1,000	453	0	0	453	1	65.2507
0.1	0.05	200	924	15	296	613	924	0.924	51.2257
		400	998	27	444	527	998	0.998	50.5606
		1,200	1,000	255	636	105	996	1	50.3565
	0.1	200	1,000	592	327	11	930	1	51.2275
		400	1,000	509	25	0	534	1	50.8388
		1,200	1,000	33	0	0	33	1	50.5223
	0.3	200	1,000	838	71	0	909	1	65.4765
		400	1,000	688	1	0	689	1	65.4973
		1,200	1,000	319	0	0	319	1	65.2646

จากตารางที่ 25 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีการแยกที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 3 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 3 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.3 ทุกขนาดตัวอย่าง คือ 1 และ ณ ระดับนัยสำคัญ 0.05 และ 0.1 จะมีอำนาจการทดสอบที่ลดลงตาม ขนาดตัวอย่างที่และระดับนัยสำคัญที่ลดลง ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 26 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 5x4 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	989	40	520	429	989	0.989	45.5460
		400	1,000	378	558	58	994	1	45.3065
		1,200	1,000	833	83	0	916	1	44.7723
	0.1	200	1,000	22	530	448	1,000	1	50.5700
		400	1,000	274	598	112	984	1	50.9015
		1,200	1,000	331	54	0	385	1	51.0163
	0.3	200	1,000	751	241	0	992	1	66.6130
		400	1,000	868	4	0	872	1	67.8350
		1,200	1,000	438	0	0	438	1	67.9980
0.1	0.05	200	994	114	610	269	993	0.994	45.5468
		400	1,000	499	451	30	980	1	45.4513
		1,200	1,000	759	42	0	801	1	44.7844
	0.1	200	1,000	78	659	261	998	1	51.0780
		400	1,000	428	475	39	942	1	51.0598
		1,200	1,000	214	15	0	229	1	51.0415
	0.3	200	1,000	832	126	0	958	1	66.8235
		400	1,000	730	1	0	731	1	67.9835
		1,200	1,000	298	0	0	298	1	67.9994

จากตารางที่ 26 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 989 และที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 994 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 3 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 3 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.989 และ ที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.994 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

ตารางที่ 27 แสดงการแยก การรวม อำนาจการทดสอบ และร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูล สำหรับตามขนาดตาราง 5x5 ตามขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

α	τ	n	ตัววัดประสิทธิภาพ						
			การแยก	การรวม				อำนาจทดสอบ	$\overline{\text{CR}}$
				1	2	3	รวม		
0.05	0.05	200	995	45	783	167	995	0.995	50.8337
		400	1,000	439	553	7	999	1	51.8903
		1,200	1,000	895	34	0	929	1	52.0577
	0.1	200	1,000	229	672	99	1,000	1	53.2765
		400	1,000	879	112	0	991	1	55.1193
		1,200	1,000	910	0	0	910	1	55.0008
	0.3	200	1,000	635	75	0	710	1	62.3000
		400	1,000	146	0	0	146	1	63.8475
		1,200	1,000	0	0	0	0	1	63.998
0.1	0.05	200	998	123	778	97	998	0.998	50.8502
		400	1,000	603	381	2	986	1	51.8450
		1,200	1,000	838	15	0	853	1	51.9732
	0.1	200	1,000	392	563	45	1,000	1	53.7730
		400	1,000	895	63	0	958	1	55.0693
		1,200	1,000	828	0	0	828	1	54.9788
	0.3	200	1,000	489	34	0	523	1	62.3745
		400	1,000	75	0	0	75	1	63.8313
		1,200	1,000	0	0	0	0	1	64.0478

จากตารางที่ 27 สามารถสรุปผลดังนี้

การแยกมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1,000 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 995 และที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีการแยกที่ลดลงคือ 998 ดังนั้น การแยกแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น การแยกก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น

การรวมครั้งที่ 3 จะมีค่าลดลงเมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น และถ้าพิจารณาที่ระดับความสัมพันธ์เดียวกัน ณ ระดับนัยสำคัญ 0.05 และ 0.1 ทุกขนาดตัวอย่าง แล้วการรวมครั้งที่ 3 มีแนวโน้มลดลงเมื่อระดับนัยสำคัญและขนาดตัวอย่างเพิ่มขึ้น ดังนั้น การรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ ระดับนัยสำคัญ และขนาดตัวอย่าง

อำนาจการทดสอบมีค่าเท่ากันที่ระดับความสัมพันธ์ 0.1 และ 0.3 ทุกขนาดตัวอย่าง คือ 1 ณ ระดับนัยสำคัญ 0.05 และ 0.1 ยกเว้นที่ระดับนัยสำคัญ 0.05 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.995 และ ที่ระดับนัยสำคัญ 0.1 ระดับความสัมพันธ์ 0.05 และขนาดตัวอย่าง 200 จะมีอำนาจการทดสอบคือ 0.998 ดังนั้น อำนาจการทดสอบแปรผันตามระดับความสัมพันธ์ข้อมูล ขนาดตัวอย่าง และระดับนัยสำคัญ เมื่อระดับความสัมพันธ์เพิ่มขึ้น อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 เพิ่มขึ้นในขั้นตอนการแยก

ร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลมีค่าใกล้เคียงกันที่ระดับความสัมพันธ์เดียวกัน ทุกขนาดตัวอย่างและระดับนัยสำคัญ และร้อยละความถูกต้องเฉลี่ยจะมีค่าลดลงเมื่อระดับความสัมพันธ์ลดลง ดังนั้น ร้อยละความถูกต้องเฉลี่ยแปรผันตามระดับความสัมพันธ์ข้อมูล เมื่อระดับความสัมพันธ์เพิ่มขึ้น ร้อยละความถูกต้องเฉลี่ยก็เพิ่มขึ้น

4.3 เปรียบเทียบผลลัพธ์การจำแนกกลุ่มข้อมูลที่สนใจ

ข้อมูลที่เลือกมาศึกษาในการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID เป็นข้อมูลที่ธนาคารเก็บข้อมูลสำคัญของลูกค้า 2,464 คน ที่กู้ยืมเงินจากธนาคารที่ประกอบด้วยข้อมูลส่วนบุคคลและข้อมูลประวัติสินเชื่อการชำระคืนเงินกู้ของธนาคาร จากนั้นเลือกตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพมาทำการศึกษา ซึ่งข้อมูลที่นำมาศึกษาประกอบไปด้วยข้อมูลตัวแปรดังต่อไปนี้

1. ตัวแปรกลุ่มรายได้ กำหนดให้เป็น $\mathbf{X_1}$ หรือ $\mathbf{X}_{\text{รายได้}}$

โดยที่ $\begin{cases} \mathbf{X_{11}} & \text{เป็นกลุ่มรายได้ต่ำ} \\ \mathbf{X_{12}} & \text{เป็นกลุ่มรายได้ปานกลาง} \\ \mathbf{X_{13}} & \text{เป็นกลุ่มรายได้สูง} \end{cases}$

2. ตัวแปรจำนวนบัตรเครดิต กำหนดให้เป็น $\mathbf{X_2}$ หรือ $\mathbf{X}_{\text{บัตรเครดิต}}$

โดยที่ $\begin{cases} \mathbf{X_{21}} & \text{มีจำนวนบัตรเครดิตน้อยกว่าห้าใบ} \\ \mathbf{X_{22}} & \text{มีจำนวนบัตรเครดิตห้าใบขึ้นไป} \end{cases}$

3. ตัวแปรระดับการศึกษา กำหนดให้เป็น $\mathbf{X_3}$ หรือ $\mathbf{X}_{\text{การศึกษา}}$

โดยที่ $\begin{cases} \mathbf{X_{31}} & \text{ระดับมัธยมศึกษา} \\ \mathbf{X_{32}} & \text{ระดับอุดมศึกษา} \end{cases}$

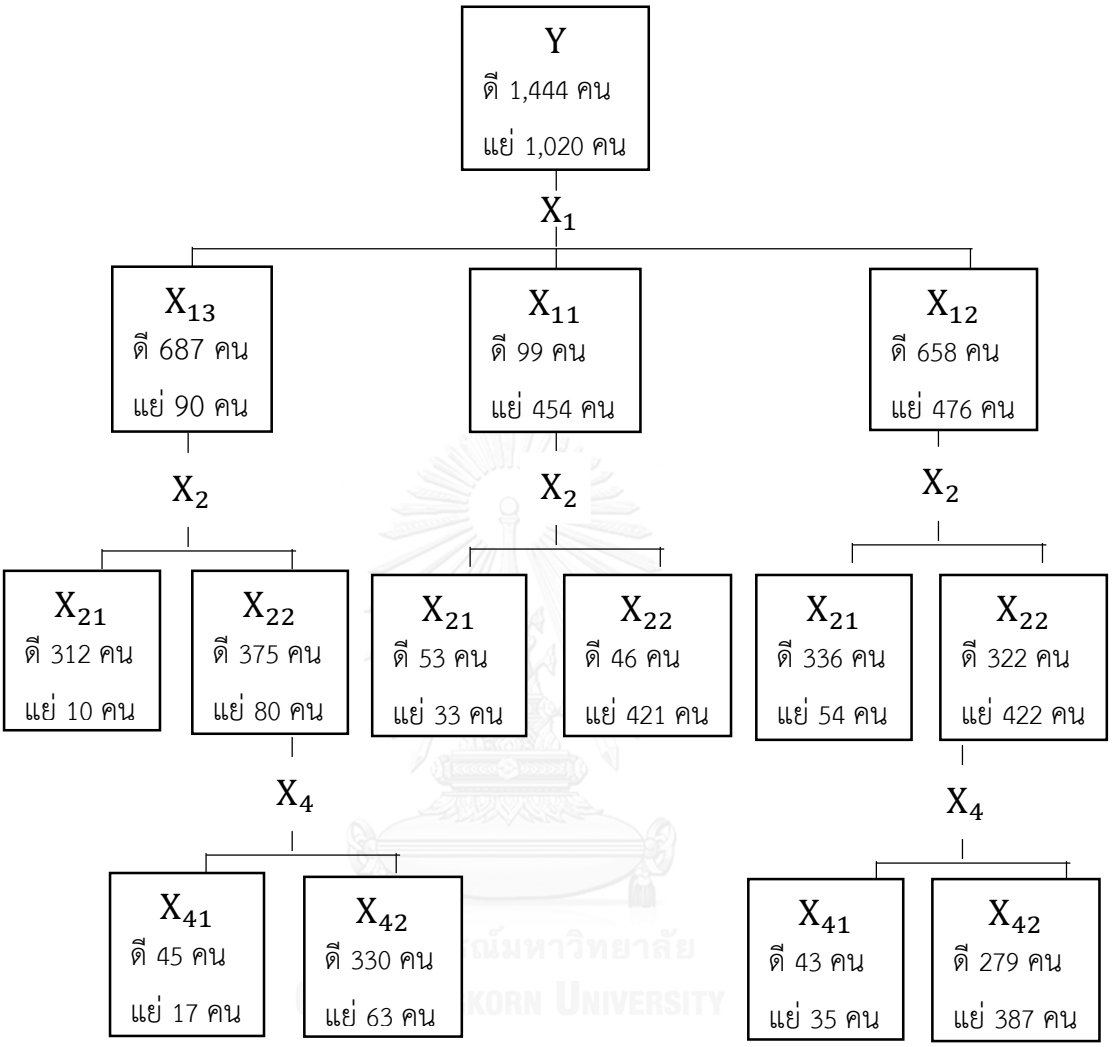
4. ตัวแปรรีไฟแนนซ์รถยนต์ กำหนดให้เป็น $\mathbf{X_4}$ หรือ $\mathbf{X}_{\text{รถยนต์}}$

โดยที่ $\begin{cases} \mathbf{X_{41}} & \text{ไม่มีหรือมีการรีไฟแนนซ์หนึ่ง} \\ \mathbf{X_{42}} & \text{มีการรีไฟแนนซ์มากกว่าสอง} \end{cases}$

5. ตัวแปรเป้าหมายคือตัวแปรระดับความน่าเชื่อถือ กำหนดให้เป็น $\mathbf{Y}$ โดยที่มีประเภทของตัวแปรตามคือ ระดับความน่าเชื่อถือดีเยี่ยม และ ระดับความน่าเชื่อถือยอดแย่

ซึ่งประเภทตัวแปรเป้าหมายที่จะศึกษาคือ ประเภทระดับความน่าเชื่อถือยอดแย่ กล่าวคือจะพิจารณาว่ากลุ่มใดเป็นกลุ่มอาจจะไม่มาชำระเงินที่กู้ยืมตามกำหนดมากสุด เพื่อนำไปใช้วิเคราะห์ทางการตลาดต่อไป โดยทำการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ที่ กำหนดค่าระดับนัยสำคัญที่ 0.05 ค่าความลึกที่ระดับ 3 และขนาดโหนดต่ำสุดที่ขนาด 200 ซึ่งจำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 1,444 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 1,020 ได้ผลลัพธ์การจำแนกกลุ่มข้อมูล ดังแผนผังที่ 4 ต่อไปนี้

แผนผังที่ 4 ผลลัพธ์การจำแนกกลุ่มข้อมูลที่สนใจ



ซึ่งผลลัพธ์การจำแนกกลุ่มข้อมูลที่สนใจโดยอัลกอริทึม CHAID สามารถอธิบายผลได้ดังนี้

- กลุ่ม 1: กลุ่มรายได้สูง และมีจำนวนเครดิตการ์ดน้อยกว่า 5 ใบ ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 312 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 10 คน โดยให้ค่าร้อยละเกนดัชนี 7.5022%
- กลุ่ม 2: กลุ่มรายได้สูง มีจำนวนเครดิตการ์ด 5 ใบขึ้นไป และมีการรีไฟแนนซ์รถยนต์น้อยกว่า 1 ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 45 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 17 คน โดย ให้ค่าร้อยละเกนดัชนี 66.2367%
- กลุ่ม 3: กลุ่มรายได้สูง มีจำนวนเครดิตการ์ด 5 ใบขึ้นไป และมีการรีไฟแนนซ์รถยนต์มากกว่า 2 ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 330 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 63 คน โดยให้ค่าร้อยละเกนดัชนี 38.7247%
- กลุ่ม 4: กลุ่มรายได้ต่ำ และมีจำนวนเครดิตการ์ดน้อยกว่า 5 ใบ ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 53 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 33 คน โดยให้ค่าร้อยละเกนดัชนี 92.6950%
- กลุ่ม 5: กลุ่มรายได้ต่ำ และมีจำนวนเครดิตการ์ด 5 ใบขึ้นไป ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 26 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 421 คน โดยให้ค่าร้อยละเกนดัชนี 217.7739%
- กลุ่ม 6: กลุ่มรายได้ปานกลาง และมีจำนวนเครดิตการ์ดน้อยกว่า 5 ใบ ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 336 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 54 คน โดยให้ค่าร้อยละเกนดัชนี 33.4481%
- กลุ่ม 7: กลุ่มรายได้ปานกลาง มีจำนวนเครดิตการ์ด 5 ใบขึ้นไป และมีการรีไฟแนนซ์รถยนต์น้อยกว่า 1 ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 43 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 35 คน โดยให้ค่าร้อยละเกนดัชนี 108.3962%
- กลุ่ม 8: กลุ่มรายได้ปานกลาง มีจำนวนเครดิตการ์ด 5 ใบขึ้นไป และมีการรีไฟแนนซ์รถยนต์มากกว่า 2 ให้ผลลัพธ์จำนวนลูกค้าที่ระดับความน่าเชื่อดีเยี่ยม คือ 279 คน และจำนวนลูกค้าที่ระดับความน่าเชื่อยอดแย่ 387 คน โดยให้ค่าร้อยละเกนดัชนี 140.3710%

ซึ่งร้อยละความถูกต้องของการจำแนกข้อมูลอยู่ที่ 78.2062% และเมื่อพิจารณาจากค่าร้อยละเกนดัชนีแล้ว กลุ่มลูกค้าที่อาจจะไม่มาชำระเงินที่กู้ยืมตามกำหนดมากสุด คือ กลุ่ม 5, กลุ่ม 8, กลุ่ม 7 ตามลำดับ ซึ่งเป็นกลุ่มเป้าหมายที่สามารถนำไปใช้วิเคราะห์ทางการตลาดต่อไป

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของอัลกอริทึม CHAID ระหว่างตัวแปร 2 ตัวแปร ที่มีการแจกแจงพหุนาม โดยตัวแปรแรกเป็นตัวแปรอิสระ X ที่มี I กลุ่ม และตัวแปรที่สองเป็นตัวแปรตาม Y ที่มี J กลุ่ม และอยู่ในตารางการณ์จรสองทาง XxY ที่มีขนาด IxJ แต่ละสถานการณ์ โดยจะพิจารณาแยกตามขนาดตัวอย่างเป็น 200, 400 และ 1,200 ระดับความสัมพันธ์ของข้อมูลเป็น 0.05, 0.1 และ 0.3 และระดับนัยสำคัญ 0.05 และ 0.1 โดยมีเกณฑ์ที่ใช้ในการพิจารณาเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อมูลของแต่ละสถานการณ์จากค่าการแยก การรวม ความน่าจะเป็นของความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก อำนาจการทดสอบในขั้นตอนการแยกและร้อยละความถูกต้องเฉลี่ยของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID โดยสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

1. การจำแนกกลุ่มข้อมูลกรณีตัวแปรทั้งสองไม่มีความสัมพันธ์กันโดยอัลกอริทึม CHAID

การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID มีขั้นตอนหลักในการจำแนกข้อมูล คือ ขั้นตอนการแยกและขั้นตอนการรวม พร้อมขั้นตอนการหยุดไว้ประยุกต์ใช้ในการจำแนกกลุ่มข้อมูล เมื่อพิจารณาการจำแนกกลุ่มข้อมูลในขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID การแยกจะมีค่าน้อยและมีค่าใกล้เคียงกับการรวมด้วย กล่าวคือเมื่อมีการแยกจะมีการรวมของข้อมูลทุกครั้ง เนื่องจากข้อมูลตัวแปรทั้งสองไม่มีความสัมพันธ์กัน ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าลดลง ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าเพิ่มขึ้น ทำให้การแยกลดลง และการรวมเพิ่มขึ้น

และการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID สามารถควบคุมความผิดพลาดประเภทที่ 1 ในขั้นตอนการแยก ได้ในทุกตาราง XxY ขนาด IxJ และขนาดตัวอย่าง ที่ระดับนัยสำคัญ 0.05 และ 0.1

2. การจำแนกกลุ่มข้อมูลกรณีตัวแปรทั้งสองมีความสัมพันธ์กันโดยอัลกอริทึม CHAID

จากผลการวิจัยข้างต้น สามารถสรุปรายละเอียดการแยก การรวม อำนาจการทดสอบในขั้นตอนการแยกและร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้ง

ขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID เมื่อพิจารณาตามระดับความสัมพันธ์ ขนาดตัวอย่างและระดับนัยสำคัญของการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ได้ดังต่อไปนี้

การแยกและการรวมของการจำแนกกลุ่มข้อมูล

- <u>ระดับความสัมพันธ์ของข้อมูล</u> เมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น ที่ขนาดตัวอย่างและระดับนัยสำคัญของข้อมูลมีค่าเท่ากัน การแยกเพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การแยกเพิ่มขึ้น ส่วนการรวมและการรวมครั้งที่มากสุด (I-2) ก็ลดลง เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้การรวมและการรวมครั้งที่มากสุด (I-2) ก็ลดลง
- <u>ขนาดตัวอย่าง</u> เมื่อขนาดตัวอย่างเพิ่มขึ้น ที่ระดับความสัมพันธ์และระดับนัยสำคัญของข้อมูลมีค่าเท่ากัน การแยกก็เพิ่มขึ้น ส่วนการรวมและการรวมครั้งที่มากสุด (I-2) ลดลง เนื่องจากขนาดตัวอย่างเพิ่มขึ้นจึงทำให้สามารถอธิบายประชากรและกลุ่มข้อมูลได้ชัดเจนขึ้น กล่าวคือ ค่าของตัวสถิติทดสอบ χ^2 มีโอกาสค่าเพิ่มขึ้น
- <u>ระดับนัยสำคัญ</u> เมื่อระดับนัยสำคัญเพิ่มขึ้น ที่ระดับความสัมพันธ์และขนาดตัวอย่างของข้อมูลมีค่าเท่ากัน การแยกก็เพิ่มขึ้น ส่วนการรวมและการรวมครั้งที่มากสุด (I-2) ลดลง เนื่องจากระดับนัยสำคัญเพิ่มขึ้น โอกาสที่จะปฏิเสธสมมติฐานว่างของการทดสอบความเป็นอิสระเพิ่มขึ้น ทำให้การแยกเพิ่มขึ้น และโอกาสที่จะปฏิเสธสมมติฐานว่างของการทดสอบความเป็นเอกพันธ์ลดลง ทำให้การแยกเพิ่มขึ้น

ซึ่งการแยกแปรผันตามระดับความสัมพันธ์ ขนาดตัวอย่างและระดับนัยสำคัญของข้อมูล ส่วนการรวมและการรวมครั้งที่มากสุด (I-2) แปรผกผันกับระดับความสัมพันธ์ของข้อมูล ขนาดตัวอย่างและระดับนัยสำคัญของข้อมูล

อำนาจการทดสอบของการจำแนกกลุ่มข้อมูล

- <u>ระดับความสัมพันธ์ของข้อมูล</u> เมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น ที่ขนาดตัวอย่างและระดับนัยสำคัญของข้อมูลมีค่าเท่ากัน อำนาจการทดสอบเพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงทำให้ค่าสถิติทดสอบ

χ^2 มีค่าเพิ่มขึ้น ส่งผลให้ค่า p-value ของตัวสถิติ χ^2 สำหรับขั้นตอนการจำแนกข้อมูลของอัลกอริทึม CHAID มีค่าลดลง ทำให้โอกาสที่ปฏิเสธ H_0 ในขั้นตอนการแยกเพิ่มขึ้น

- ขนาดตัวอย่าง เมื่อขนาดตัวอย่างเพิ่มขึ้น ที่ระดับความสัมพันธ์และระดับนัยสำคัญของข้อมูลมีค่าเท่ากัน อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากตัวอย่างเพิ่มขึ้นจึงทำให้สามารถอธิบายประชากรและกลุ่มข้อมูลได้ชัดเจนขึ้น
- ระดับนัยสำคัญ เมื่อระดับนัยสำคัญเพิ่มขึ้น ที่ระดับความสัมพันธ์และขนาดตัวอย่างของข้อมูลมีค่าเท่ากัน อำนาจการทดสอบก็เพิ่มขึ้น เนื่องจากระดับนัยสำคัญเพิ่มขึ้น โอกาสที่จะปฏิเสธสมมติฐานว่างในขั้นตอนการแยกเพิ่มขึ้น คือ ความน่าจะเป็นของของผิดประเภทที่ 1 เพิ่มขึ้น แล้วความน่าจะเป็นของของผิดประเภทที่ 2 ลดลง ทำให้อำนาจการทดสอบในขั้นตอนการแยกก็เพิ่มขึ้น

ซึ่งอำนาจการทดสอบหรืออำนาจการตัดสินใจที่ถูกต้องในการขั้นตอนการแยกของการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID แปรผันตามระดับความสัมพันธ์ ขนาดตัวอย่างและระดับนัยสำคัญของข้อมูล นอกจากนี้อำนาจการทดสอบมีแนวโน้มลดลงเมื่อความแตกต่างระหว่างแถวกับหลักเพิ่มขึ้น นั้นคือ ความแตกต่างระหว่างจำนวนกลุ่มของตัวแปรอิสระ X และจำนวนกลุ่มของตัวแปรตาม Y

ร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูล

- ระดับความสัมพันธ์ของข้อมูล เมื่อระดับความสัมพันธ์ของข้อมูลเพิ่มขึ้น ที่ขนาดตัวอย่างและระดับนัยสำคัญของข้อมูลมีค่าเท่ากัน ร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูลที่ผ่านขั้นตอนการแยกหรือผ่านทั้งขั้นตอนการแยกและการรวมของอัลกอริทึม CHAID เพิ่มขึ้น เนื่องจากระดับความสัมพันธ์เพิ่มขึ้น ค่าผลต่างของค่าสังเกตกับค่าคาดหวังจะลดลงของตัวสถิติทดสอบ χ^2 ทำให้สามารถอธิบายความสัมพันธ์ระหว่างตัวแปร Y กับตัวแปร X ได้ดีและชัดเจนขึ้น

ซึ่งร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูล จะพิจารณาจากขั้นตอนการแยกและขั้นตอนการรวมของอัลกอริทึม CHAID ซึ่งร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูลแปรผันตามระดับความสัมพันธ์ของข้อมูล แต่เมื่อพิจารณาขนาดตัวอย่างและระดับนัยสำคัญจะส่งผลต่อน้อยต่อร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูล และเมื่อพิจารณาจำนวนกลุ่มของตัวแปรตาม Y ถ้ามีค่าเพิ่มขึ้น ร้อยละความถูกต้องมีแนวโน้มที่ลดลง เนื่องจากจำนวนกลุ่มของตัวแปรตาม Y มีค่าเพิ่มขึ้น

จะมีแนวโน้มในการลดจำนวนความถูกต้องในการจำแนกกลุ่มข้อมูลลง ทำให้ร้อยละความถูกต้องของการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID มีแนวโน้มที่ลดลง

3. การจำแนกกลุ่มข้อมูลที่สนใจโดยอัลกอริทึม CHAID

ซึ่งจากศึกษาผลสรุปดังที่ได้กล่าวมา ที่ข้อมูลตัวอย่าง 200 จำนวน อำนาจการทดสอบและการแยกของการจำแนกกลุ่มข้อมูลมีค่าต่ำ และข้อมูลที่มีจำนวนน้อยจะไม่สามารถอธิบายกลุ่มข้อมูลหลังจากผ่านขั้นตอนการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ได้ชัดเจน กล่าวคือ ในการจำแนกกลุ่มข้อมูล เมื่อมีข้อมูลที่น้อย การแยกข้อมูลเพื่อหาความสัมพันธ์ระหว่างตัวแปรตาม Y กับตัวแปรอิสระ X ก็แยกได้น้อย และการรวมในขั้นตอนการรวมจะรวมกลุ่มของตัวแปรอิสระ X มากเกินไปที่จะสามารถอธิบายการจำแนกข้อมูลได้อย่างมีประสิทธิภาพ ดังนั้นผู้วิจัยจึงกำหนดขนาดโหนดขั้นต่ำในการแยกโหนดไว้ที่ 200 เพื่อไม่ให้มีการแยกโหนดได้อีก เพื่อไม่ให้ส่งผลกระทบในการอธิบายข้อมูล ซึ่งร้อยละความถูกต้องในการจำแนกกลุ่มข้อมูลอยู่ที่ 78.2062 และได้กลุ่มข้อมูลเป้าหมายที่ดีที่สามารถอธิบายทางการตลาดได้ เนื่องจากการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID เป็นการจำแนกกลุ่มข้อมูลแบบลำดับชั้นจากบนลงล่างและหาความสัมพันธ์ของกลุ่มข้อมูลตามลำดับขั้นซึ่งเหมาะสมในการวิเคราะห์และอธิบายผลลัพธ์ที่เข้าใจง่าย และสามารถกำหนดระดับนัยสำคัญในขั้นตอนการแยกและการรวมที่แตกต่างกันได้ และสามารถกำหนดค่าความลึกของการจำแนกกลุ่มข้อมูลและขนาดโหนดต่ำสุดในขั้นตอนการแยกเพื่อประยุกต์ใช้ในการจำแนกกลุ่มข้อมูลตามความเหมาะสม

5.2 ข้อเสนอแนะ

จากงานวิจัยชิ้นนี้ผู้ที่สนใจสามารถนำไปศึกษาต่อได้อีกในเรื่องดังต่อไปนี้

1. ขอบเขตในการวิจัย ในเรื่องของขนาดตัวอย่าง จำนวนกลุ่มของตัวแปรตาม Y จำนวนกลุ่มของตัวแปรอิสระ X ระดับความสัมพันธ์ของข้อมูล และระดับนัยสำคัญ อาจจะมีการเพิ่มหรือลดค่าเหล่านั้นให้มีความหลากหลายมากขึ้นได้
2. กรณีที่ข้อมูลเป็นข้อมูลเชิงปริมาณต่อเนื่อง อาจใช้อัลกอริทึม CART ในการจำแนกกลุ่มข้อมูล
3. กรณีที่ข้อมูลอาจมีค่าสูญหาย (missing value)

รายการอ้างอิง

ภาษาไทย

กัลยา วานิชย์บัญชา. (2553). หลักสถิติ. กรุงเทพมหานคร: ธรรมสาร.

ศศิธร เจษฎาฐิติกุล. (2545). การเปรียบเทียบวิธีการทดสอบความเป็นอิสระระหว่างตัวแปร 2 ตัวแปรที่มีการแจกแจงพหุนาม. วิทยานิพนธ์ สถิติศาสตรมหาบัณฑิต สาขาวิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาต่างประเทศ

Alkhasawneh, M. S., Ngah, U. K., Tay, L. T., Mat Isa, N. A., & Al-Batah, M. S. (2014). Modeling and testing landslide hazard using decision tree. *Journal of Applied Mathematics, 2014*, 1-9. doi:10.1155/2014/929768

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. I. (1984). *Classification and regression trees*. California, CA, USA: CRC Press.

Chen, J. S. (2003). Market segmentation by tourists' sentiments. *Annals of Tourism Research, 30*(1), 178-193. doi:10.1016/s0160-7383(02)00046-4

Chris, R., Jyun-Cheng, W., & David, C. Y. (2002). Data mining techniques for customer relationship management. *Technology in Society, 24*, 483-502.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119-127.

Messenger, R. C., & Mandell, L. M. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of American Statistical Society, 67*, 768-772.

Norusis, M. J. (2011). IBM SPSS decision tree 20 (3 ed.): IBM Corporation.

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลโดยโปรแกรมแกรม R

1. ตัวอย่างกรณีการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ของข้อมูล ตัวแปร 2 ตัวแปร ที่มีการแจกแจงพหุนาม โดยตัวแปรแรกเป็นตัวแปรอิสระ X ที่มี 2 กลุ่ม และตัวแปรที่สองเป็นตัวแปรตาม Y ที่มี 2 กลุ่ม โดยมีขนาดตารางเป็น 2x2 ขนาดตัวอย่าง 200 ที่ระดับนัยสำคัญ (α) 0.05 เมื่อตัวแปรทั้งสองไม่มีความสัมพันธ์กัน เพื่อใช้วิเคราะห์หาตัววัดประสิทธิภาพ ใน 1,000 รอบ

```
library(DescTools)
alphamerge<-0.05
alphasplit<-0.05

#################################
####### Case n=200, round=1,000 #####
#################################

n..<-200
loopround<-1000

#################################
####### Simulation Data X and Y #####
#################################

p_i.<-c(0.67,0.33)
p_.j<-c(0.53,0.47)
p<-outer(p_i.,p_.j)

dataobs<-list()
length(dataobs)<-loopround

for(r in 1:loopround){
k<-runif(n..,0,1)
o11<-ifelse(k<=p[1,1],1,0)
o12<-ifelse(p[1,1]<k & k<=p[1,1]+p[1,2],1,0)
o21<-ifelse(p[1,1]+p[1,2]<k & k<=p[1,1]+p[1,2]+p[2,1],1,0)
o22<-ifelse(p[1,1]+p[1,2]+p[2,1]<k & k<=p[1,1]+p[1,2]+p[2,1]+p[2,2],1,0)
obs11<-sum(o11)
obs12<-sum(o12)
obs21<-sum(o21)
obs22<-sum(o22)
```

```
obsmatrix<-cbind(c(obs11,obs21),c(obs12,obs22))

while(any(obsmatrix==0)){
k<-runif(n..,0,1)
o11<-ifelse(k<=p[1,1],1,0)
o12<-ifelse(p[1,1]<k & k<=p[1,1]+p[1,2],1,0)
o21<-ifelse(p[1,1]+p[1,2]<k & k<=p[1,1]+p[1,2]+p[2,1],1,0)
o22<-ifelse(p[1,1]+p[1,2]+p[2,1]<k & k<=p[1,1]+p[1,2]+p[2,1]+p[2,2],1,0)
obs11<-sum(o11)
obs12<-sum(o12)
obs21<-sum(o21)
obs22<-sum(o22)
obsmatrix<-cbind(c(obs11,obs21),c(obs12,obs22))}
dataobs[[r]]<-obsmatrix}

transdata<-function(matr,datax=1:nrow(matr),datay=1:ncol(matr)){
x<-rep(rep_len(datax,length(matr)),as.numeric(matr))
y<-rep(rep(datay,each=nrow(matr)),as.numeric(matr))
return(data.frame(y,x))}

dataobs<-lapply(dataobs,transdata)

#################################
######### CHAID Algorithm #########
#################################

algorithmCHAID<-function(data){
pvaluesplit<-c()
for(i in 1:(ncol(data)-1))
pvaluesplit[i]<-chisq.test(table(data[,1],data[,i+1]))$p.value
names(pvaluesplit)<-names(data)[-1]

if(any(pvaluesplit<=alphasplit)){
pvaluemerge<-c()
pvaluesplitmin<-which.min(pvaluesplit)+1
cat<-as.character(data[,pvaluesplitmin])
ncat<-levels(factor(cat))
pair<-CombSet(ncat,2)
for(i in 1:nrow(pair))
pvaluemerge[i]<-chisq.test(table(data[,1][cat%in%pair[i,]],cat[cat%in%pair[i,]]))$p.value
names(pvaluemerge)<-paste(names(data)[pvaluesplitmin],apply(pair,1,toString),sep=":")
```

```
while(any(pvaluemerge>alphamerge)&length(ncat)>2){
most<-pair[which.max(pvaluemerge),]
cat<-ifelse(cat%in%most,toString(most),cat)
ncat<-levels(factor(cat))
pair<-CombSet(ncat,2)
pvaluemerge<-c()
for(i in 1:nrow(pair))
pvaluemerge[i]<-chisq.test(table(data[,1][cat%in%pair[i,]],cat[cat%in%pair[i,]]))$p.value
names(pvaluemerge)<-paste(names(data)[pvaluesplitmin],apply(pair,1,toString),sep=":")}

datarelate<-split(data[,-pvaluesplitmin],cat)
names(datarelate)<-paste(names(data)[pvaluesplitmin],names(datarelate))
return(datarelate)
}else {
return(data)}}

DATACHAID<-lapply(dataobs,algorithmCHAID)
nsplit<-c()
nmerge<-c()

for(r in 1:loopround){
if(class(DATACHAID[[r]])=="list"){
nsplit[r]<-T
nmerge[r]<-nlevels(factor(dataobs[[r]][,2]))-length(DATACHAID[[r]])


}  else { nsplit[r]<-F
nmerge[r]<-NA}}

#################################
############# Result #############
#################################

table(nsplit)
table(nsplit)/loopround
nnmerge<-table(factor(nmerge,levels=0:3))
nnmerge
nnmerge/loopround
```

2. ตัวอย่างกรณีการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม CHAID ของข้อมูล ตัวแปร 2 ตัวแปร ที่มีการแจกแจงพหุนาม โดยตัวแปรแรกเป็นตัวแปรอิสระ X ที่มี 2 กลุ่ม และตัวแปรที่สองเป็นตัวแปรตาม Y ที่มี 2 กลุ่ม โดยมีขนาดตารางเป็น 2x2 ขนาดตัวอย่าง 200 ที่ระดับนัยสำคัญ (α) 0.05 เมื่อตัวแปรทั้งสองมีความสัมพันธ์กัน ที่ระดับความสัมพันธ์ของข้อมูล 0.3 เพื่อใช้วิเคราะห์หาตัววัดประสิทธิภาพ ใน 1,000 รอบ

```
library(DescTools)
alphamerge<-0.05
alphasplit<-0.05

#################################
####### Case n=200, round=1,000 #####
#################################

n..<-200
looround<-1000
#################################
####### Simulation Data X and Y #####
#################################

p_i.<-c(0.36,0.64)
p_.j<-c(0.35,0.65)
p<-c(0.251,0.109,0.099,0.541)
p<-matrix(p,2,byrow=T)

dataobs<-list()
length(dataobs)<-looround

for(r in 1:looround){
k<-runif(n..,0,1)
o11<-ifelse(k<=p[1,1],1,0)
o12<-ifelse(p[1,1]<k & k<=p[1,1]+p[1,2],1,0)
o21<-ifelse(p[1,1]+p[1,2]<k & k<=p[1,1]+p[1,2]+p[2,1],1,0)
o22<-ifelse(p[1,1]+p[1,2]+p[2,1]<k & k<=p[1,1]+p[1,2]+p[2,1]+p[2,2],1,0)
obs11<-sum(o11)
obs12<-sum(o12)
obs21<-sum(o21)
obs22<-sum(o22)
obsmatrix<-cbind(c(obs11,obs21),c(obs12,obs22))
```

```
while(any(obsmatrix==0)){
k<-runif(n..,0,1)
o11<-ifelse(k<=p[1,1],1,0)
o12<-ifelse(p[1,1]<k & k<=p[1,1]+p[1,2],1,0)
o21<-ifelse(p[1,1]+p[1,2]<k & k<=p[1,1]+p[1,2]+p[2,1],1,0)
o22<-ifelse(p[1,1]+p[1,2]+p[2,1]<k & k<=p[1,1]+p[1,2]+p[2,1]+p[2,2],1,0)
obs11<-sum(o11)
obs12<-sum(o12)
obs21<-sum(o21)
obs22<-sum(o22)
obsmatrix<-cbind(c(obs11,obs21),c(obs12,obs22))}
dataobs[[r]]<-obsmatrix}

transdata<-function(matr,datax=1:nrow(matr),datay=1:ncol(matr)){
x<-rep(rep_len(datax,length(matr)),as.numeric(matr))
y<-rep(rep(datay,each=nrow(matr)),as.numeric(matr))
return(data.frame(y,x))}

dataobs<-lapply(dataobs,transdata)

#################################
######### CHAID Algorithm #########
#################################

algorithmCHAID<-function(data){
pvaluesplit<-c()
for(i in 1:(ncol(data)-1))
pvaluesplit[i]<-chisq.test(table(data[,1],data[,i+1]))$p.value
names(pvaluesplit)<-names(data)[-1]

if(any(pvaluesplit<=alphasplit)){
pvaluemerge<-c()
pvaluesplitmin<-which.min(pvaluesplit)+1
cat<-as.character(data[,pvaluesplitmin])
ncat<-levels(factor(cat))
pair<-CombSet(ncat,2)
for(i in 1:nrow(pair))
pvaluemerge[i]<-chisq.test(table(data[,1][cat%in%pair[i,]],cat[cat%in%pair[i,]]))$p.value
names(pvaluemerge)<-paste(names(data)[pvaluesplitmin],apply(pair,1,toString),sep=":")

while(any(pvaluemerge>alphamerge)&length(ncat)>2){
```

```
most<-pair[which.max(pvaluemerge),]
cat<-ifelse(cat%in%most,toString(most),cat)
ncat<-levels(factor(cat))
pair<-CombSet(ncat,2)
pvaluemerge<-c()
for(i in 1:nrow(pair))
pvaluemerge[i]<-chisq.test(table(data[,1][cat%in%pair[i,]],cat[cat%in%pair[i,]]))$p.value
names(pvaluemerge)<-paste(names(data)[pvaluesplitmin],apply(pair,1,toString),sep=":")}

datarelate<-split(data[,-pvaluesplitmin],cat)
names(datarelate)<-paste(names(data)[pvaluesplitmin],names(datarelate))
return(datarelate)
}else {
return(data)}}

DATACHAID<-lapply(dataobs,algorithmCHAID)
nsplit<-c()
nmerge<-c()

classmanage<-function(x){
termnode<-list()
length(termnode)<-length(x)
nameclass<-sapply(x,class)
names(termnode)<-names(x)
return(termnode)}

resvalue<-list()
length(resvalue)<-looproround

for(r in 1:looproround){
if(class(DATACHAID[[r]])=="list"){
nsplit[r]<-T
nmerge[r]<-nlevels(factor(dataobs[[r]][,2]))-length(DATACHAID[[r]])

termnode<-classmanage(DATACHAID[[r]])
groupy<-c()

for(i in 1:length(DATACHAID[[r]])){
if(NCOL(DATACHAID[[r]][[i]])>1){
termnode[[i]]<-table(DATACHAID[[r]][[i]][,1])
groupy<-rbind(groupy,data.frame(y=DATACHAID[[r]][[i]][,1],term=names(DATACHAID[[r]])[i]))
```

```
} else{ termnode[[i]]<-table(DATACHAID[[r]][[i]])
groupy<-rbind(groupy,data.frame(y=DATACHAID[[r]][[i]],term=names(DATACHAID[[r]])[i]))}}

predict<-tapply(groupy$y,groupy$term,function(x){
classgroupy<-table(x)
names(classgroupy)[which.max(classgroupy)]})
groupy$predict<-factor(rep(predict,table(groupy$term)),levels=levels(factor(groupy$y)))

confmat<-table(groupy$y,groupy$predict)
diagonal<-diag(confmat)
confmat<-cbind(confmat,diagonal/table(groupy$y)*100)
confmat<-rbind(confmat,c(prop.table(table(groupy$predict)),sum(diagonal)/nrow(groupy))*100)

resvalue[[r]]<-confmat
}  else { nsplit[r]<-F
nmerge[r]<-NA}}

#################################
############# Result #############
#################################

table(nsplit)
table(nsplit)/loopround
nnmerge<-table(factor(nmerge,levels=0:3))
nnmerge
nnmerge/looproung
mean(sapply(resvalue[sapply(resvalue,class)=="matrix"],function(x)x[nrow(x),ncol(x)]))
```

ประวัติผู้เขียนวิทยานิพนธ์

นายวิศรุต กิมชัยวงศ์ เกิดวันพุธที่ 12 กันยายน พ.ศ. 2533 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2556