บทความวิจัย (Research Article)

การเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่ม เมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล

Clustering Efficiency Comparison of Outliers Data in Data Mining

ณัฐวรรณ ผลจันทร์*, ปาริฉัตร ใจมีธรรม และสายชล สินสมบูรณ์ทอง

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Natthawan Phonchan*, Parichar Jaimeetham and Saichon Sinsomboonthong

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Road, Ladkrabang, Bangkok 10520

Received: May 10, 2020; Accepted: June 14, 2020

บทคัดย่อ

งานวิจัยนี้มีนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่มแบบเป็นขั้นตอนและวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอนของข้อมูลที่มีทั้งข้อมูลเชิงปริมาณและเชิงคุณภาพที่มีค่านอกเกณฑ์ 5 ชุด โดยวิธีการจัดกลุ่มแบบเป็นขั้นตอนใช้การจัดกลุ่มเชื่อมโยงแบบเดี่ยว การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์ และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ย และใช้วิธีวัดระยะห่าง 3 แบบ คือ ระยะห่างยูคลิเดียน ระยะห่างแมนฮัทตัน และระยะห่างเชบีเชฟ วิธีการจัดกลุ่มแบบไม่เป็นขั้นตอนใช้วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม และใช้วิธีวัดระยะห่าง 2 แบบ คือ ระยะห่างยูคลิเดียนและระยะห่างแมนฮัทตัน พิจารณาจากค่าความแม่นโดยใช้โปรแกรม WEKA สำหรับการค้นคว้าและศึกษาค่านอกเกณฑ์ได้ใช้ข้อมูลมีข้อมูล 5 ชุด คือ โรคหัวใจเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ร้อยละ 1.39 มะเร็งเต้านมเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ร้อยละ 2.28 โรคหัวใจและหลอดเลือดเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ร้อยละ 3.43 โรคเบาหวานเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ร้อยละ 4.02 และการทำประกันสุขภาพเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ร้อยละ 5.53 โดยใช้โปรแกรม SPSS ในการตรวจหาค่านอกเกณฑ์ ผลการเปรียบเทียบวิธีการจัดกลุ่มแบบเป็นขั้นตอนชุดข้อมูลหัวใจและหลอดเลือด โรคเบาหวาน และการประกันสุขภาพ พบว่าวิธีการจัดกลุ่มเชื่อมโยงแบบเดี่ยวให้ค่าเฉลี่ยของค่าความแม่นสูงสุด ส่วนชุดข้อมูลโรคหัวใจและมะเร็งเต้านมพบว่าวิธีการจัดกลุ่มเชื่อมโยงแบบเฉลี่ยให้ค่าเฉลี่ยของค่าความแม่นสูงสุด และการศึกษาวิธีวัดระยะห่างชุดข้อมูลโรคหัวใจ โรคหัวใจและหลอดเลือด และโรคเบาหวาน พบว่าวิธีวัดระยะห่างแมนฮัทตันให้ค่าเฉลี่ยของค่าความแม่นสูงสุด ส่วนข้อมูลมะเร็งเต้านม วิธีวัดระยะห่างยูคลิเดียนให้ค่าเฉลี่ยของค่าความแม่นสูงสุด และข้อมูลการทำประกันสุขภาพ วิธีวัดระยะห่างเชบีเชฟให้ค่าเฉลี่ยของค่าความแม่นสูงสุด การเปรียบเทียบวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอนโดยใช้วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม ชุดข้อมูลมะเร็งเต้า

**Corresponding author: 58051180@kmitl.ac.th*

DOI: 10.14456/tjst.2020.64

นม โรคหัวใจและหลอดเลือด และการทำประกันสุขภาพ พบว่าวิธีวัดระยะห่างยูคลิเดียนให้ค่าความแม่นสูงสุด ส่วนข้อมูลโรคหัวใจและโรคเบาหวานวิธีวัดระยะห่างแมนฮัทตันให้ค่าเฉลี่ยของค่าความแม่นสูงสุด

คำสำคัญ : การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย; การจัดกลุ่มเชื่อมโยงแบบเดี่ยว; การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์; ค่านอกเกณฑ์; ระยะห่างเชบีเชฟ; ระยะห่างแมนฮัทตัน; ระยะห่างยูคลิเดียน; วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม

Abstract

Our research objective was to evaluate an efficacy of different types of hierarchical and non-hierarchical clustering methods on five well-known data sets with different qualities and quantities of outliers. Each of the three types of the hierarchical clustering method adopted the different linkage criteria. i.e. single-linkage, complete-linkage, or average-linkage clustering. Each type could use any of three different metrics: Euclidean, Manhattan, or Chebyshev Distances. The non-hierarchical clustering method performed k-means clustering analysis employing one of two metrics: Euclidean or Manhattan distances. All data sets were pre-processed with WEKA software and their outliers detected with SPSS software. The five data sets were a heart disease data set (with 1.39 % outliers), a breast cancer (2.28 %), a cardiovascular disease (3.43 %), a diabetes (4.02 %), and an insurance claim (5.53 %) data set by SPSS software for outlier detection. The two clustering methods were run on the five data sets, and their clustering accuracy values were evaluated. A type of hierarchical and non-hierarchical clustering methods was chosen as the most efficacy for a particular data set type for that respective method according to its clustering accuracy. For hierarchical clustering method, the most efficacy clustering type for cardiovascular disease, diabetes, and insurance claim data sets was the single-linkage clustering type; the most efficacy type for heart disease and breast cancer data sets was the average-linkage clustering type; the most efficacy metric for heart disease, cardiovascular disease, and diabetes data sets was Manhattan distance; the most efficacy metric for breast cancer data set was Euclidean distance; the most efficacy metric for insurance claim data set was Chebyshev distance. For non-hierarchical clustering method performed k-means clustering analysis, the most efficacy metric for breast cancer, cardiovascular disease, and insurance claim data sets was Euclidean distance; the most efficacy metric for heart disease and diabetes data sets was Manhattan distance.

Keywords: average-linkage clustering; single-linkage clustering; complete-linkage clustering; outlier; Chebyshev distance; Manhattan distance; Euclidean distance; k-mean clustering

1. คำนำ

ปัจจุบันการทำงานวิจัยมักจะต้องเก็บรวบรวมข้อมูลแล้วนำมาวิเคราะห์และประมวลผลให้ได้ข้อสรุป เพื่อนำมาตอบข้อสงสัยหรือแก้ไขปัญหางานวิจัยนั้น ๆ และเพื่อนำข้อมูลไปใช้ประโยชน์ได้สูงสุด บางครั้งพบว่าข้อมูลมีค่ามากเกินไปหรือน้อย

เกินไปแฝงอยู่เรียกว่าค่านอกเกณฑ์ (outlier) เป็นค่าที่อยู่ปลายสุด ซึ่งตกอยู่ใกล้กับขีดจำกัดของพิสัยข้อมูล การหาค่านอกเกณฑ์มีความสำคัญเนื่องจากแสดงค่าความคลาดเคลื่อนในข้อมูล ถ้าเรานำข้อมูลที่มีค่านอกเกณฑ์ไปวิเคราะห์จะส่งผลให้เกิดความคลาดเคลื่อนของผลลัพธ์ข้อมูลที่ได้ การกระจายของข้อมูลและค่าเฉลี่ยของข้อมูลไม่ดี ส่งผลให้ไม่เป็นไปตามข้อกำหนดเบื้องต้น (assumption) ทำให้ไม่สามารถนำข้อมูลไปใช้ประโยชน์อย่างสูงสุด สำหรับสาเหตุที่ทำให้เกิดค่านอกเกณฑ์ เช่น ความคลาดเคลื่อนจากการแปรผันข้อมูลที่เก็บรวบรวมมาซึ่งเป็นความคลาดเคลื่อนที่ไม่สามารถควบคุม ความคลาดเคลื่อนที่เกิดจากเครื่องมือที่ใช้วัดมีคุณภาพต่ำทำให้เกิดค่านอกเกณฑ์ ความคลาดเคลื่อนที่เกิดจากการบันทึกข้อมูลจากการปฏิบัติโดยไม่ตรวจสอบให้ถี่ถ้วน เพื่อให้ได้ผลการวิเคราะห์ที่เชื่อถือได้ การตรวจสอบหาค่านอกเกณฑ์จึงเป็นสิ่งที่สำคัญก่อนการนำข้อมูลไปวิเคราะห์ เพื่อให้ได้ข้อมูลที่ไม่มีความผิดปกติและสามารถแก้ไขได้ บางครั้งข้อมูลอาจไม่สามารถใช้งานได้เต็มประสิทธิภาพหรือตรงตามความต้องการ จึงมีการจัดกลุ่มของข้อมูลเพื่อให้ได้วิธีที่มีความเหมาะสมกับข้อมูลที่มีความแตกต่างกันไป ดังนั้นผู้วิจัยจึงต้องเลือกวิธีการจัดกลุ่มให้เหมาะสมกับข้อมูล (พนิดา และคณะ, 2560)

วีระยุทธ และพยุง (2557) ศึกษางานวิจัยเกี่ยวกับการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล งานวิจัยนี้มุ่งเน้นผลประสิทธิภาพการจัดกลุ่มข้อมูล วัดประสิทธิภาพด้วยค่าความแม่น พบว่าขั้นตอนวิธีการจัดกลุ่มบนปริภูมิย่อยที่ทำงานร่วมกับการจัดกลุ่มแบบเป็นขั้นตอน ให้ค่าความแม่นในภาพรวมสูงที่สุดที่ระดับร้อยละ 94 รองลงมาคือ ขั้นตอนวิธีการจัดกลุ่มบนปริภูมิย่อยที่ทำงานร่วมกับการจัดกลุ่มแบบเฉลี่ย k กลุ่ม ให้ค่าความแม่นในภาพรวมที่ระดับร้อยละ 83

อุมาพร และคณะ (2561) ศึกษางานเกี่ยวกับการเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่มข้อมูล 2 วิธี คือ การจัดกลุ่มแบบเป็นขั้นตอนและการจัดกลุ่มแบบเฉลี่ย k กลุ่ม วัดประสิทธิภาพของวิธีการจัดกลุ่มข้อมูลด้วยความแม่น พบว่าวิธีการจัดกลุ่มแบบเป็นขั้นตอนมีประสิทธิภาพดีกว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม

จิรวรรณ และนัท (2557) ศึกษาวิจัยเกี่ยวกับการเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่มข้อมูล 4 วิธี ได้แก่ วิธีการจัดกลุ่มแบบวอร์ด (Ward linkage) วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม วิธีการจัดกลุ่มแบบฟัซซี่ซีมีน และวิธีการจัดกลุ่มแบบขั้นตอนวิธี expectation maximization โดยวัดประสิทธิภาพของการจัดกลุ่มข้อมูลด้วย 2 วิธี คือ การวัดค่าความแตกต่างของข้อมูลภายในกลุ่มและการวัดค่าความต่างของข้อมูลระหว่างกลุ่ม พบว่าเมื่อจำนวนตัวแปรศึกษาเท่ากับ 2 วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม และวิธีการจัดกลุ่มแบบฟัซซี่ซีมีนมีประสิทธิภาพของการจัดกลุ่มใกล้เคียงกัน ถ้าจำนวนตัวแปรศึกษาเท่ากับ 3 วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม มีประสิทธิภาพของการจัดกลุ่มดีที่สุด

Bhatt และคณะ (2016) ศึกษาวิจัยเกี่ยวกับการจัดกลุ่มโดยใช้การจัดกลุ่มแบบเฉลี่ย k กลุ่ม และมี 5 วิธีการเปรียบเทียบ คือ ไม่ใช้การวัดระยะห่าง ใช้การวัดระยะห่าง คือ ยูคลิเดียน แมนฮัทตัน เชบีเชฟ และมิงโคฟสกี พบว่าวิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม คำนวณระยะห่างด้วยเชบีเชฟ ให้ผลการสร้างกลุ่มข้อมูลที่มีประสิทธิภาพดีที่สุด

Frederic และ Serge (2019) ศึกษาขั้นตอนวิธีการจัดกลุ่มแบบเป็นขั้นตอนและการพัฒนาเกณฑ์การเชื่อมโยงแบบเดี่ยวเพื่อจัดการกับข้อมูลรบกวน โดยใช้วิธีเปรียบเทียบ 9 วิธี คือ การจัดกลุ่มเชื่อมโยงแบบเดี่ยว (single linkage) การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์ (complete linkage) การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย (average linkage) การจัด

กลุ่มจุดศูนย์กลาง (centroid) การจัดกลุ่มแบบวอร์ด การจัดกลุ่มแบบฮาวดอร์ฟ (Hausdorff linkage) วิธีจัดกลุ่มเพื่อนบ้านใกล้สุด (full with neighborhood) การเชื่อมโยงแบบเดี่ยวเพื่อนบ้านใกล้สุด (single with neighborhood) และการเชื่อมโยงแบบเดี่ยวเมื่อมีข้อมูลรบกวน (single link with noise) พบว่าวิธีการจัดกลุ่มเชื่อมโยงแบบเดี่ยว การเชื่อมโยงแบบเดี่ยวเพื่อนบ้านใกล้สุด และการเชื่อมโยงแบบเดี่ยวเมื่อมีข้อมูลรบกวนให้ผลดีกว่าอีก 6 วิธี จากนั้นนำวิธีการจัดกลุ่มเชื่อมโยงแบบเดี่ยว การเชื่อมโยงแบบเดี่ยวเพื่อนบ้านใกล้สุด และวิธีการเชื่อมโยงแบบเดี่ยวเมื่อมีข้อมูลรบกวนมาเปรียบเทียบโดยเพิ่มข้อมูลรบกวน พบว่าวิธีการเชื่อมโยงแบบเดี่ยวเมื่อมีข้อมูลรบกวนให้ผลลัพธ์ที่ดีที่สุด

ดังนั้นงานวิจัยนี้จึงศึกษาการจัดกลุ่ม (clustering) ด้วย 2 วิธี คือ วิธีการจัดกลุ่มแบบเป็นขั้นตอน (hierarchical clustering) และวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน (nonhierarchical clustering) ซึ่งวิธีการจัดกลุ่มแบบเป็นขั้นตอน ประกอบด้วย 3 วิธี คือ การจัดกลุ่มเชื่อมโยงแบบเดี่ยว (single-linkage clustering) การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์ (complete-linkage clustering) และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ย (average-linkage clustering) ใช้วิธีวัดระยะห่าง 3 แบบ คือ ระยะห่างยูคลิเดียน (Euclidean distance) ระยะห่างแมนฮัทตัน (Manhattan distance) และระยะห่างเชบีเชฟ (Chebyshev distance) และวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอนใช้วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม (k-means clustering หรือ simple k-means) ใช้วิธีวัดระยะห่าง 2 แบบ คือ ระยะห่างยูคลิเดียนและระยะห่างแมนฮัทตัน เพื่อต้องการเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่ม ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูล โดยใช้วิธีการจัดกลุ่มด้วยโปรแกรม WEKA เพื่อเปรียบเทียบว่าวิธีการจัดกลุ่มใดให้ค่าความแม่น (accuracy) สูงกว่า

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยครั้งนี้ คือ WEKA เวอร์ชั่น 3.9 และ SPSS เวอร์ชั่น 25

2.2 การเก็บรวบรวมข้อมูล การหาค่านอกเกณฑ์ การศึกษาขั้นตอนวิธี และการเปรียบเทียบประสิทธิภาพของวิธีการจัดกลุ่ม

2.2.1 การเก็บรวบรวมข้อมูล ค้นหาและศึกษาข้อมูลที่มีค่านอกเกณฑ์จากเว็บไซต์ Kaggle ซึ่งข้อมูลส่วนมากมีค่านอกเกณฑ์อยู่ระหว่างร้อยละ 1.00-5.00 โดยมีข้อมูล 5 ชุด

(1) โรคหัวใจ (heart disease) จำนวนข้อมูลทั้งหมด 1,025 ค่า พบค่านอกเกณฑ์คิดเป็นร้อยละ 1.39 (David, 1988) ประกอบด้วยตัวแปรจำนวน 13 ตัว ได้แก่ อายุ (X_1) เพศ (X_2) ระดับอาการเจ็บหน้าอก (X_3) ความดันโลหิต (X_4) ระดับของคอเลสเตอรอลในเลือด (X_5) ระดับน้ำตาลในเลือดมากกว่า 120 มิลลิกรัมต่อเดซิลิตร (X_6) ผลการตรวจคลื่นแม่เหล็กไฟฟ้า (X_7) อัตราการเต้นของหัวใจสูงสุด (X_8) โรคหลอดเลือดหัวใจตีบที่เกิดจากการออกกำลังกาย (X_9) ภาวะซึมเศร้าที่เกิดจากการออกกำลังกาย (X_{10}) ค่าความชันสูงสุดในการออกกำลังกาย (X_{11}) สีของการตรวจหัวใจโดยใช้เครื่อง fluoroscopy (X_{12}) การเต้นของหัวใจ (X_{13})

(2) มะเร็งเต้านม (breast cancer) จำนวนข้อมูลทั้งหมด 569 ค่า พบจำนวนค่านอกเกณฑ์คิดเป็นร้อยละ 2.28 (Merishna, 2018) ประกอบด้วยตัวแปรจำนวน 5 ตัว ได้แก่ รัศมีของเนื้องอก (X_1) พื้นผิวของเนื้องอก (X_2) เส้นรอบวงของเนื้องอก (X_3) ขนาดพื้นที่ของเนื้องอก (X_4) ความเรียบเนียนของเนื้องอก (X_5)

(3) โรคหัวใจและหลอดเลือด (cardiovascular disease) จำนวนข้อมูลทั้งหมด 1,000 ค่า พบจำนวนค่านอกเกณฑ์คิดเป็นร้อยละ 3.43 (Svetlana, 2019) ประกอบด้วยตัวแปรจำนวน 11

ตัว ได้แก่ อายุ (X_1) เพศ (X_2) ความสูง (X_3) น้ำหนัก (X_4) ความดันของเลือดสูงสุดขณะหัวใจห้องล่างบีบตัว (X_5) ความดันของเลือดต่ำสุดขณะหัวใจห้องล่างคลายตัว (X_6) ระดับคอเลสเตอรอล (X_7) ระดับน้ำตาล (X_8) สูบบุหรี่ (X_9) การดื่มแอลกอฮอล์ (X_{10}) การออกกำลังกาย (X_{11})

(4) โรคเบาหวาน (diabetes) จำนวนข้อมูลทั้งหมด 1,050 ค่า พบจำนวนค่านอกเกณฑ์เท่ากับ 169 ค่า คิดเป็นร้อยละ 4.02 (Harry, 2017) ประกอบด้วยตัวแปรจำนวน 8 ตัว ได้แก่ จำนวนครั้งที่ตั้งครรภ์ (X_1) ความเข้มข้นของกลูโคสในกระแสเลือด (X_2) ความดันโลหิต (X_3) ความหนาของกล้ามเนื้อแขน (X_4) ความเข้มข้นของอินซูลินในร่างกาย (X_5) ดัชนีมวลกาย (X_6) ความเสี่ยงของโรคเบาหวานที่มาจากทางพันธุกรรม (X_7) อายุ (X_8)

(5) การทำประกันสุขภาพ (insurance claim) จำนวนข้อมูลทั้งหมด 1,338 ค่า พบจำนวนค่านอกเกณฑ์คิดเป็นร้อยละ 5.53 (Eason, 2018) ประกอบด้วยตัวแปรจำนวน 7 ตัว ได้แก่ โดย อายุของผู้ถือกรมธรรม์ (X_1) เพศของผู้ถือกรมธรรม์ (X_2) ค่าดัชนีมวลกายให้ความเข้าใจเกี่ยวกับร่างกายน้ำหนักที่ค่อนข้างสูงหรือต่ำเมื่อเทียบกับความสูง (X_3) จำนวนลูกของผู้ถือกรมธรรม์ (X_4) สถานะการสูบบุหรี่ของผู้ถือกรมธรรม์ (X_5) เขตที่อยู่อาศัยของผู้ถือกรมธรรม์ในสหรัฐอเมริกา (X_6) ค่ารักษาพยาบาลส่วนบุคคลที่เรียกเก็บโดยประกันสุขภาพ (X_7)

2.2.2 การหาค่านอกเกณฑ์ โดยค่านอกเกณฑ์หมายถึงค่าที่อยู่ปลายสุดซึ่งตกอยู่ใกล้กับขีดจำกัดของพิสัยข้อมูล การหาค่านอกเกณฑ์มีความสำคัญเนื่องจากค่านอกเกณฑ์อาจแสดงความคลาดเคลื่อนในข้อมูล (สายชล, 2560) ค่านอกเกณฑ์ คือ ค่าสังเกตที่เบี่ยงเบนไปจากค่าสังเกตอื่นมากจนทำให้สงสัยว่าค่าสังเกตนั้นได้มาจากวิธีอื่น ลักษณะของค่านอกเกณฑ์มี 2 ลักษณะ (Hawkins, 1980) คือ ค่าสังเกตที่มีค่าสูงหรือต่ำมาก เรียกค่านอกเกณฑ์นี้ว่า discordant observation และค่าสังเกตที่มีลักษณะการแจกแจงแตกต่างจากลักษณะการแจกแจงของประชากรที่สนใจศึกษา เรียกค่านอกเกณฑ์นี้ว่า contaminate observation (Beckman and Cook, 1983) โดยค่านอกเกณฑ์ระดับต่ำมีค่าร้อยละ 1.00-1.66 ระดับปานกลางร้อยละ 1.67-3.33 และระดับสูงร้อยละ 3.34-5.00 การตรวจสอบค่านอกเกณฑ์ของข้อมูลด้วยโปรแกรม SPSS ซึ่งจะใช้ข้อมูลและวิธีการเดียวกับตารางการแจกแจงปรกติมาตรฐาน คือ การใช้ Z-score ตรวจสอบหาค่านอกเกณฑ์

ขั้นตอนการหาค่านอกเกณฑ์โดยใช้โปรแกรม SPSS มีขั้นตอนการคลิกโปรแกรมดังนี้

(1) เปิดไฟล์ชุดข้อมูล โดยคลิก File → Open → Data → เลือกไฟล์ชุดข้อมูล

(2) ขั้นตอนหาค่านอกเกณฑ์ เลือก Analyze → Descriptive Statistics → Explore

(3) เลือกตัวแปรเชิงปริมาณใส่ในช่อง Dependent List

(4) คลิก Statistics เลือก Outliers → Continue → OK

2.2.3 การศึกษาขั้นตอนวิธี

(1) การจัดกลุ่ม เป็นการเรียนรู้แบบไม่มีผู้สอน (unsupevised learing) โดยรวมข้อมูลที่มีลักษณะใกล้เคียงกันออกเป็นกลุ่มและนำกลุ่มข้อมูลที่ได้ไปวิเคราะห์ มีวิธีการหลายวิธีในการวัดความใกล้เคียงกันของชุดข้อมูล (Galit *et al*., 2007) ข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีลักษณะที่เหมือนกันหรือคล้ายกัน ส่วนข้อมูลที่อยู่ต่างกลุ่มกันจะมีลักษณะที่แตกต่างกัน (กัลยา, 2544)

(1.1) การจัดกลุ่มแบบเป็นขั้นตอน เป็นวิธีที่นิยมใช้กันมากในการแบ่งกลุ่มที่มีจำนวนกรณีต่ำกว่า 200 กรณี โดยไม่จำเป็นต้องทราบจำนวนกลุ่มหรือทราบว่าตัวแปรใดหรือกรณีใดอยู่กลุ่มใดมาก่อน โดยวิธีการจัดกลุ่มแบบเป็นขั้นตอนที่

นิยม ได้แก่

(ก) การจัดกลุ่มเชื่อมโยงแบบเดี่ยว เป็นการรวมหน่วยในกลุ่ม 2 กลุ่ม เข้าด้วยกัน โดยใช้ระยะห่างที่น้อยที่สุดระหว่าง 2 กลุ่ม (จันทร์จิรา, 2558) สมการที่ใช้ในการคำนวณการจัดกลุ่มเชื่อมโยงแบบเดี่ยว

$$d_{ij} = \min(d_{13}, d_{14}, d_{15}, d_{23}, d_{24}, d_{25})$$

โดยที่ d_{ij} คือ ระยะห่างที่น้อยที่สุดของกลุ่ม i และกลุ่ม j

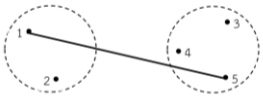
ระยะห่างระหว่างกลุ่ม คือ d_{24}

(Johnson and Wichern, 2007)

(ข) การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์ เป็นการเป็นการรวมหน่วยในกลุ่ม 2 กลุ่ม เข้าด้วยกัน โดยระยะห่างระหว่างกลุ่มจะเป็นระยะห่างที่ไกลที่สุดระหว่าง 2 กลุ่ม (ธรา และจิติมนต์, 2557) สมการที่ใช้ในการคำนวณการจัดกลุ่มเชื่อมโยงแบบสมบูรณ์

$$d_{ij} = \max(d_{13}, d_{14}, d_{15}, d_{23}, d_{24}, d_{25})$$

โดยที่ d_{ij} คือ ระยะห่างที่มากที่สุดของกลุ่ม i และกลุ่ม j



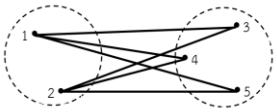
ระยะห่างระหว่างกลุ่ม คือ d_{15}

(Johnson and Wichern, 2007)

(ค) การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย เป็นการรวมกลุ่มโดยใช้ค่าเฉลี่ยระหว่าง 2 กลุ่ม โดยการหาระยะห่างระหว่างหน่วยต่าง ๆ ทุกคู่ที่อยู่ใน 2 กลุ่ม จากนั้นจะรวมกลุ่มที่มีค่าเฉลี่ยน้อยที่สุดไว้ด้วยกัน ในกรณีที่ภายในกลุ่มมีเพียง 1 ค่า จะให้ค่านั้นเป็นค่าเฉลี่ยของกลุ่ม สมการที่ใช้ในการคำนวณการจัดกลุ่มเชื่อมโยงแบบเฉลี่ย

$$d_{ij} = \frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{n}$$

โดยที่ d_{ij} คือ ระยะห่างเฉลี่ยของทุกคู่ระหว่างกลุ่ม i และกลุ่ม j; n คือ จำนวนคู่ในการหาระยะห่างทั้งหมด



ระยะห่างระหว่างกลุ่ม คือ $\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$

(Johnson and Wichern, 2007)

(1.2) การจัดกลุ่มแบบไม่เป็นขั้นตอน วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม เป็นวิธีที่นิยมใช้เพราะระยะเวลาในการคำนวณน้อยกว่าวิธีการจัดกลุ่มข้อมูลแบบอื่น (Jiawei *et al*., 2006) วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่มนิยมใช้เมื่อมีจำนวนกรณีมากกว่า 200 กรณี โดยจะต้องกำหนดจำนวนกลุ่มที่ต้องการ เช่น กำหนดให้มี k กลุ่ม

(2) ฟังก์ชันระยะห่าง คือ ฟังก์ชันค่าจริง d ที่ทำให้จุดโคออร์ดิเนท x, y และ z เป็นไปตามคุณสมบัติต่อไปนี้

(2.1) $d(x, y) \geq 0$ และ $d(x, y) = 0$ ก็ต่อเมื่อ $x = y$

(2.2) $d(x, y) = d(y, x)$

(2.3) $d(x, z) \leq d(x, y) + d(y, z)$

คุณสมบัติข้อที่ 1 ระยะห่างไม่เป็นลบและระยะห่างจะเป็น 0 สำหรับจุดโคออร์ดิเนท (ได้แก่ พล็อตการกระจาย) ที่เหมือนกัน คุณสมบัติข้อที่ 2 แสดงถึงการสลับเปลี่ยน ตัวอย่าง เช่น ระยะห่างจากนิวยอร์กถึงลอสแองเจอรีสเหมือนกับระยะห่างจากลอสแองเจอรีสถึงนิวยอร์ก ส่วนคุณสมบัติข้อที่ 3 เป็นอสมการสามเหลี่ยม (triangle inequality) ซึ่งจุดที่ 3 ไม่เคยสั้นกว่าระยะห่างระหว่าง 2 จุดแรก (สายชล, 2560)

(ก) ระยะห่างยูคลิเดียน มีสูตรการคำนวณดังนี้

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)}$$

(ข) ระยะห่างแมนฮัทตัน มีสูตรการคำนวณดังนี้

$$d_{Manhattan}(x, y) = \sum_i |x_i - y_i|$$

(ค) ระยะห่างเชบีเชฟ มีสูตรการคำนวณดังนี้ (Arwa and Heba, 2019)

$$d_{Chebyshev}(x, y) = \max\left(|x_i - x_j|, |y_i - y_j|\right)$$

สำหรับตัวแปรเชิงกลุ่ม จะนิยามฟังก์ชันความแตกต่าง (different function) สำหรับเปรียบเทียบค่าคุณลักษณะที่ i ของระเบียงคู่หนึ่ง ๆ ดังนี้ (สายชล, 2560)

$$\text{Different}(x_i, y_i) = \begin{cases} 0 & \text{ถ้า } x_i = y_i \\ 1 & \text{กรณีอื่น ๆ} \end{cases}$$

2.2.4 การเปรียบเทียบประสิทธิภาพของวิธีการจัดกลุ่ม

ตารางที่ 1 เมทริกซ์ความสับสน

		ผลลัพธ์จากสมการหรือการทดสอบ	
		คำตอบเป็นบวก	คำตอบที่เป็นลบ
ผลลัพธ์ที่เกิดขึ้นจริง	คำตอบเป็นบวก	TP (true positive)	FN (false negative)
	คำตอบเป็นลบ	FP (false positive)	TN (true negative)

บวกจริง (true positive, TP) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นบวก; ลบจริง (true negative, TN) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นลบ; บวกเท็จ (false positive, FP) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นลบ; ลบเท็จ (false negative, FN) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นบวก

นำผลการวิเคราะห์ของแต่ละวิธีมาเปรียบเทียบประสิทธิภาพโดยพิจารณาจากเมทริกซ์ความสับสน (confusion matrix) ซึ่งเป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในตารางเมทริกซ์ความสับสน ช่วยให้ง่ายต่อการมองเห็นค่าทำนายของขั้นตอนวิธีดังตารางที่ 1

ค่าความแม่น คือ การแสดงการวัดที่ได้มีความแม่นในรูปอัตราส่วน (สายชล, 2560)

Accuracy = จำนวนข้อมูลที่จำแนกถูกว่าคำตอบเป็นบวกและลบ x 100 %

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\ \%$$

3. ผลการวิจัย

3.1 ผลการเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่ม

งานวิจัยนี้จัดกลุ่มข้อมูลโดยใช้วิธีการทำเหมืองข้อมูล นำชุดข้อมูลที่ค้นคว้าจำนวน 5 ชุด มาวิเคราะห์ข้อมูลและเปรียบเทียบประสิทธิภาพการจัดกลุ่มโดยพิจารณาจากค่าความแม่น ซึ่งวิธีที่ใช้ในการทดสอบครั้งนี้ คือ วิธีการจัดกลุ่มแบบเป็นขั้นตอนและวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน

3.1.1 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจัดกลุ่มแบบเป็นขั้นตอน

ตารางที่ 2 ข้อมูลโรคหัวใจ พบว่าการจัดกลุ่มเชื่อมโยงแบบเดี่ยวโดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 51.8374 การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์โดยใช้วิธีวัดระยะห่างยูคลิเดียนให้ค่าความแม่นสูงสุด คือ ร้อยละ 51.4472 และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ยโดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 62.2439

ตารางที่ 3 ข้อมูลมะเร็งเต้านม พบว่าการจัดกลุ่มเชื่อมโยงแบบเดี่ยวโดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 63.1127 การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์โดยใช้วิธีวัดระยะห่างยูคลิเดียนให้ค่าความแม่นสูงสุด คือ ร้อยละ 64.9482 และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ย

โดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่น

สูงสุด คือ ร้อยละ 78.1097

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบเป็นขั้นตอน ชุดข้อมูลโรคหัวใจ จำนวนข้อมูลทั้งหมด 1,025 ค่านอกเกณฑ์ 57 ค่า คิดเป็นร้อยละ 1.39

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น		
	วิธีวัดระยะห่าง		
	ยูคลิเดียน	แมนฮัทตัน	เชบีเชฟ
การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	48.9106	**51.8374**	42.8184
การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	**51.4472**	45.5284	46.9378
การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	55.2846	**62.2439**	41.5501

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบเป็นขั้นตอน ชุดข้อมูลมะเร็งเต้านม จำนวนข้อมูลทั้งหมด 569 ค่านอกเกณฑ์ 65 ค่า คิดเป็นร้อยละ 2.28

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น		
	วิธีวัดระยะห่าง		
	ยูคลิเดียน	แมนฮัทตัน	เชบีเชฟ
การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	62.7417	**63.1127**	62.7026
การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	**64.9482**	56.4148	56.6491
การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	77.5434	**78.1097**	70.6307

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบเป็นขั้นตอน ชุดข้อมูลโรคหัวใจและหลอดเลือด จำนวนข้อมูลทั้งหมด 1,000 ค่านอกเกณฑ์ 137 ค่า คิดเป็นร้อยละ 3.43

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น		
	วิธีวัดระยะห่าง		
	ยูคลิเดียน	แมนฮัทตัน	เชบีเชฟ
การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	**52.0111**	51.9111	40.2556
การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	38.8111	40.4844	**48.3444**
การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	48.1667	**49.6444**	40.2556

ตารางที่ 5 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบเป็นขั้นตอน ชุดข้อมูลโรคเบาหวาน จำนวนข้อมูลทั้งหมด 1,050 ค่านอกเกณฑ์ 84 ค่า คิดเป็นร้อยละ 4.02

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น		
	วิธีวัดระยะห่าง		
	ยูคลิเดียน	แมนฮัทตัน	เชบีเชฟ
การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	66.9736	66.9418	**67.1111**

การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	40.4233	**45.4498**	39.9788
การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	67.0264	66.5714	**67.3651**

ตารางที่ 6 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบเป็นขั้นตอน ชุดข้อมูลการทำประกันสุขภาพ จำนวนข้อมูลทั้งหมด 1,338 ค่านอกเกณฑ์ 148 ค่า คิดเป็นร้อยละ 5.53

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น		
	วิธีวัดระยะห่าง		
	ยูคลิเดียน	แมนฮัทตัน	เชบีเชฟ
การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	**43.0660**	**43.0660**	40.6660
การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	33.6240	33.0261	**44.9344**
การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	39.1048	**42.6756**	38.1997

ตารางที่ 4 ข้อมูลโรคหัวใจและหลอดเลือด พบว่าการจัดกลุ่มเชื่อมโยงแบบเดี่ยวโดยใช้วิธีวัดระยะห่างยูคลิเดียนให้ค่าความแม่นสูงสุด คือ ร้อยละ 52.0111 การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์โดยใช้วิธีวัดระยะห่างเชบีเชฟให้ค่าความแม่นสูงสุด คือ ร้อยละ 48.3444 และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ยโดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 49.6444

ตารางที่ 5 ข้อมูลโรคเบาหวาน พบว่าการจัดกลุ่มเชื่อมโยงแบบเดี่ยวโดยใช้วิธีวัดระยะห่างเชบีเชฟให้ค่าความแม่นสูงสุด คือ ร้อยละ 67.1111 การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์โดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 45.4498 และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ยโดยใช้วิธีวัดระยะห่างเชบีเชฟให้ค่าความแม่นสูงสุด คือ ร้อยละ 67.3651

ตารางที่ 6 ข้อมูลการประกันสุขภาพ พบว่าการจัดกลุ่มเชื่อมโยงแบบเดี่ยวโดยใช้วิธีวัดระยะห่างยูคลิเดียนและแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 43.0660 การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์โดยใช้วิธีวัดระยะห่างเชบีเชฟให้ค่าความแม่นสูงสุด คือ ร้อยละ 44.9344 และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ยโดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 42.6756

การนำวิธีการจัดกลุ่มเชื่อมโยงทั้ง 3 วิธี มาหาค่าเฉลี่ยอีกครั้งหนึ่ง และวิธีวัดระยะห่างทั้ง 3 วิธี มาหาค่าเฉลี่ยอีกครั้งหนึ่ง ได้ดังตารางที่ 7 พบว่าชุดข้อมูลโรคหัวใจ การจัดกลุ่มเชื่อมโยงแบบเฉลี่ยให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 53.0262 และวิธีวัดระยะห่างโดยวิธีแมนฮัทตันให้ค่าเฉี่ยของค่าความแม่นสูงสุ คือ ร้อยละ 53.2032 ชุดข้อมูลมะเร็งเต้านม การจัดกลุ่มเชื่อมโยงแบบเฉลี่ยให้ค่าเฉี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 75.4279 และวิธีวัดระยะห่างโดยวิธียูคลิเดียนให้ค่าเฉี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 68.4111 ชุดข้อมูลโรคหัวใจและหลอดเลือด การจัดกลุ่มเชื่อมโยงแบบเดี่ยวให้ค่าเฉี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 48.0593 และวิธีวัดระยะห่างโดยวิธีแมนฮัทตันให้ค่าเฉี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 47.3466 ชุดข้อมูลโรคเบาหวาน การจัดกลุ่มเชื่อมโยงแบบเดี่ยวให้ค่าเฉี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 67.0088 และวิธีวัดระยะห่างโดยวิธีแมน

ฮัทตันให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 59.6543 ชุดข้อมูลการทำประกันสุขภาพ การจัดกลุ่มเชื่อมโยงแบบเดี่ยวให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 43.2660 และวิธีวัดระยะห่างโดยวิธีเชบีเชฟให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 41.2667

ตารางที่ 7 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของของวิธีการจัดกลุ่มแบบเป็นขั้นตอน ข้อมูลทั้งหมด 5 ชุด

ชุดข้อมูล	วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น			ค่าเฉลี่ย
		วิธีวัดระยะห่าง			
		ยูคลิเดียน	แมนฮัทตัน	เชบีเชฟ	
โรคหัวใจ	การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	48.9106	51.8374	42.8184	47.8555
	การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	51.4472	45.5284	46.9378	47.9711
	การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	55.2846	62.2439	41.5501	**53.0262**
	ค่าเฉลี่ย	51.8808	**53.2032**	43.7688	
มะเร็งเต้านม	การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	62.7417	63.1127	62.7026	62.8523
	การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	64.9482	56.4148	56.6491	59.3374
	การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	77.5434	78.1097	70.6307	**75.4279**
	ค่าเฉลี่ย	**68.4111**	65.8791	63.3275	
โรคหัวใจและหลอดเลือด	การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	52.0111	51.9111	40.2556	**48.0593**
	การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	38.8111	40.4844	48.3444	42.5466
	การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	48.1667	49.6444	40.2556	46.0222
	ค่าเฉลี่ย	46.3296	**47.3466**	42.9519	
โรคเบาหวาน	การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	66.9736	66.9418	67.1111	**67.0088**
	การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	40.4233	45.4498	39.9788	41.9506
	การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	67.0264	66.5714	67.3651	66.9876
	ค่าเฉลี่ย	58.1411	**59.6543**	58.1517	
การทำประกันสุขภาพ	การจัดกลุ่มเชื่อมโยงแบบเดี่ยว	43.0660	43.0660	40.6660	**43.2660**
	การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์	33.6240	33.0261	44.9344	37.1948
	การจัดกลุ่มเชื่อมโยงแบบเฉลี่ย	39.1048	42.6756	38.1997	39.9934
	ค่าเฉลี่ย	38.5983	40.5892	**41.2667**	

3.1.2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน

ตารางที่ 8 ข้อมูลโรคหัวใจ พบว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม โดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 46.6558

ตารางที่ 9 ข้อมูลมะเร็งเต้านม พบว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม โดยใช้วิธีวัดระยะห่างยูคลิเดียนให้ค่าความแม่นสูงสุด คือ ร้อยละ 46.2997

ตารางที่ 10 ข้อมูลโรคหัวใจและหลอดเลือด พบว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม

โดยใช้วิธีวัดระยะห่างยูคลิเดียนให้ค่าความแม่นสูงสุด คือ ร้อยละ 34.7111

ตารางที่ 11 ข้อมูลโรคเบาหวาน พบว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม โดยใช้วิธีวัดระยะห่างแมนฮัทตันให้ค่าความแม่นสูงสุด คือ ร้อยละ 34.8360

ตารางที่ 12 ข้อมูลการประกันสุขภาพ พบว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม โดยใช้วิธีวัดระยะห่างยูคลิเดียนให้ค่าความแม่นสูงสุด คือ ร้อยละ 32.6607

การนำวิธีวัดระยะห่างทั้ง 2 วิธี มาหาค่าเฉลี่ยอีกครั้งหนึ่งได้ดังตารางที่ 13 พบว่าวิธีการจัดกลุ่มแบบเป็นไม่ขั้นตอน วิธีวัดระยะห่างยูคลิเดียนให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 38.2162

ตารางที่ 8 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน ชุดข้อมูลโรคหัวใจ จำนวนข้อมูลทั้งหมด 1,025 ค่านอกเกณฑ์ 57 ค่า คิดเป็นร้อยละ 1.39

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น	
	วิธีวัดระยะห่างยูคลิเดียน	วิธีวัดระยะห่างแมนฮัทตัน
การจัดกลุ่มแบบเฉลี่ย k กลุ่ม	44.5420	**46.6558**

ตารางที่ 9 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน ชุดข้อมูลมะเร็งเต้านม จำนวนข้อมูลทั้งหมด 569 ค่านอกเกณฑ์ 65 ค่า คิดเป็นร้อยละ 2.28

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น	
	วิธีวัดระยะห่างยูคลิเดียน	วิธีวัดระยะห่างแมนฮัทตัน
การจัดกลุ่มแบบเฉลี่ย k กลุ่ม	**46.2997**	43.5071

ตารางที่ 10 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน ชุดข้อมูลโรคหัวใจและหลอดเลือด จำนวนข้อมูลทั้งหมด 1,000 ค่านอกเกณฑ์ 137 ค่า คิดเป็นร้อยละ 3.43

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น	
	วิธีวัดระยะห่างยูคลิเดียน	วิธีวัดระยะห่างแมนฮัทตัน
การจัดกลุ่มแบบเฉลี่ย k กลุ่ม	**34.7111**	32.4778

ตารางที่ 11 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน ชุดข้อมูลโรคเบาหวาน จำนวนข้อมูลทั้งหมด 1,050 ค่านอกเกณฑ์ 84 ค่า คิดเป็นร้อยละ 4.02

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น	
	วิธีวัดระยะห่างยูคลิเดียน	วิธีวัดระยะห่างแมนฮัทตัน
การจัดกลุ่มแบบเฉลี่ย k กลุ่ม	32.8677	**34.8360**

ตารางที่ 12 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน ชุดข้อมูลการทำประกันสุขภาพ จำนวนข้อมูลทั้งหมด 1,338 ค่านอกเกณฑ์ 148 ค่า คิดเป็นร้อยละ 5.53

วิธีการจัดกลุ่ม	ค่าเฉลี่ยของค่าความแม่น	
	วิธีวัดระยะห่างยูคลิเดียน	วิธีวัดระยะห่างแมนฮัทตัน
การจัดกลุ่มแบบเฉลี่ย k กลุ่ม	**32.6607**	31.8635

ตารางที่ 13 ผลการเปรียบเทียบประสิทธิภาพค่าเฉลี่ยของค่าความแม่นของวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอนด้วยวิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม ข้อมูลทั้งหมด 5 ชุด

ชุดข้อมูล	ค่าเฉลี่ยของค่าความแม่น	
	วิธีวัดระยะห่าง	
	ยูคลิเดียน	แมนฮัทตัน
โรคหัวใจ	44.5420	46.6558
มะเร็งเต้านม	46.2997	43.5071
โรคหัวใจและหลอดเลือด	34.7111	32.4778
โรคเบาหวาน	32.8677	34.8360
การทำประกันสุขภาพ	32.6607	31.8635
ค่าเฉลี่ย	**38.2162**	37.8680

4. สรุปผลการวิจัย

งานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพในการจัดกลุ่มด้วยวิธีการจัดกลุ่มแบบเป็นขั้นตอนใช้การจัดกลุ่มเชื่อมโยงแบบเดี่ยว การจัดกลุ่มเชื่อมโยงแบบสมบูรณ์ และการจัดกลุ่มเชื่อมโยงแบบเฉลี่ย และใช้วิธีวัดระยะห่าง 3 แบบ คือ ระยะห่างยูคลิเดียน ระยะห่างแมนฮัทตัน และระยะห่างเชบีเชฟ ส่วนวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอนใช้วิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม และใช้วิธีวัดระยะห่าง 2 แบบ คือ ระยะห่างยูคลิเดียน และระยะห่างแมนฮัทตัน โดยพิจารณาจากค่าความแม่นและใช้ข้อมูลที่มีค่านอกเกณฑ์ 5 ชุด

สำหรับวิธีการจัดกลุ่มแบบเป็นขั้นตอนข้อมูลหัวใจและหลอดเลือด โรคเบาหวาน และการประกันสุขภาพ วิธีการจัดกลุ่มเชื่อมโยงแบบเดี่ยวให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 48.0593, 67.0088 และ 43.2660 ตามลำดับ ส่วนข้อมูลโรคหัวใจและมะเร็งเต้านม วิธีการจัดกลุ่มเชื่อมโยงแบบเฉลี่ยให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 53.0262 และ 75.4279 ตามลำดับ ชุดข้อมูลโรคหัวใจ โรคหัวใจและหลอดเลือด และโรคเบาหวาน วิธีวัดระยะห่างแมนฮัทตันให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 53.2032, 47.3466 และ 59.6543 ตามลำดับ ส่วนข้อมูลมะเร็งเต้านม วิธีวัดระยะห่างยูคลิเดียนให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 68.4111 และข้อมูลการประกันสุขภาพ วิธีวัดระยะห่างเชบีเชฟให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 41.2667 ดังนั้นวิธีการจัดกลุ่มเชื่อมโยงแบบเดี่ยวและวิธีวัดระยะห่างแบบแมนฮัทตันให้ค่าเฉลี่ยของค่าความแม่นสูงสุด

ส่วนวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอนข้อมูลมะเร็งเต้านม โรคหัวใจและหลอดเลือด และการทำประกันสุขภาพ วิธีวัดระยะห่างยูคลิเดียนให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 46.2997, 34.7113 และ 32.6607 ตามลำดับ ส่วนข้อมูลโรคหัวใจและโรคเบาหวาน วิธีวัดระยะห่างแมนฮัทตันให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 46.6688 และ 34.8860 ตามลำดับ ดังนั้นวิธีจัดกลุ่มแบบเฉลี่ย k กลุ่ม ใช้วิธีวัด

ระยะห่างยูคลิเดียนให้ค่าเฉลี่ยของค่าความแม่นสูงสุด

5. อภิปรายผล

วิธีการจัดกลุ่มแบบเป็นขั้นตอนชุดข้อมูลโรคหัวใจและหลอดเลือด โรคเบาหวาน และการทำประกันสุขภาพ วิธีการจัดกลุ่มเชื่อมโยงแบบเดี่ยวมีประสิทธิภาพดีที่สุดเพราะว่ามีค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 48.0593, 67.0088 และ 43.2660 ตามลำดับ สอดคล้องกับงานวิจัยของ Frederic และ Serge (2019) ที่พบว่าวิธีการเชื่อมโยงแบบเดี่ยวเมื่อมีข้อมูลรบกวนให้ผลลัพธ์ที่ดีที่สุด และวิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน ชุดข้อมูลมะเร็งเต้านม โรคหัวใจและหลอดเลือด และการทำประกันสุขภาพ วิธีวัดระยะห่างยูคลิเดียนมีประสิทธิภาพดีที่สุดเพราะว่าให้ค่าเฉลี่ยของค่าความแม่นสูงสุด คือ ร้อยละ 46.2997, 34.7113 และ 32.6607 ตามลำดับ ซึ่งให้ผลขัดแย้งกับงานวิจัยของ Bhatt และคณะ (2016) ที่พบว่าวิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม วัดระยะห่างเชบีเชฟมีประสิทธิภาพดีที่สุด เนื่องจากโปรแกรม WEKA ไม่มีวิธีวัดระยะห่างเชบีเชฟ จึงให้ผลไม่สอดคล้องกัน

6. ข้อเสนอแนะ

6.1 การวิเคราะห์ข้อมูลด้วยโปรแกรม WEKA ด้วยวิธีการจัดกลุ่มแบบเฉลี่ย k กลุ่ม วิธีการวัดระยะห่างมีเงื่อนไขที่สามารถใช้ได้เพียงวิธีวัดระยะห่างยูคลิเดียนและแมนฮัทตันเท่านั้น อาจมีโปรแกรมอื่นที่หาวิธีวัดระยะห่างได้มากกว่านี้

6.2 เพื่อให้ได้ข้อสรุปของการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น ดังนั้นอาจวิเคราะห์ข้อมูลด้วยวิธีการจัดกลุ่มประเภทอื่น ๆ เช่น วิธีการจัดกลุ่มแบบวอร์ด วิธีการจัดกลุ่มแบบฟัซซี่มีน

6.3 การศึกษาวิธีการหาค่านอกเกณฑ์มีโปรแกรมอื่น ๆ ที่สามารถหาค่านอกเกณฑ์ ซึ่งอาจมีประสิทธิภาพที่ดีกว่า เช่น Excel, NCSS

7. การนำไปใช้ประโยชน์

สามารถนำผลการเปรียบเทียบประสิทธิภาพที่ได้ไปใช้เป็นแนวทางในการเลือกวิธีการจัดกลุ่มที่เหมาะสมที่สุดและช่วยลดข้อผิดพลาดในการเลือกใช้วิธีการจัดกลุ่ม

8. รายการอ้างอิง

กัลยา วานิชย์บัญชา, 2544, การวิเคราะห์สถิติ : สถิติเพื่อการตัดสินใจ, บริษัทธรรมสาร, กรุงเทพฯ.

จันทร์จิรา พิลาแดง, 2558, การจัดกลุ่มแบบสองด้านโดยขั้นตอนวิธีเชิงพันธุกรรมเพื่อแบ่งกลุ่มระดับความเข้มแข็งของครอบครัวไทย, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยธรรมศาสตร์, ปทุมธานี

จิรวรรณ ไพบูลย์วรชาติ และนัท กุลวานิช, 2557, การเปรียบเทียบวิธีการจัดกลุ่มสำหรับข้อมูลที่มีการแจกแจงปกติแบบผสม, น. 311-326, การประชุมสัมมนาทางวิชาการ มทร.ตะวันออก มรภ.กลุ่มศรีอยุธยา และราชนครินทร์วิชาการและวิจัย, สถาบันวิจัยและพัฒนา มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก, ชลบุรี.

ธรา อั่งสกุล และจิติมนต์ อั่งสกุล, 2557, การพัฒนาระบบส่วนบุคคลสำหรับแนะนำสถานที่ท่องเที่ยวในประเทศไทยเพื่อสร้างแรงจูงใจให้กับนักท่องเที่ยวต่างชาติ, สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา.

พนิดา สมบัติมาก, ภัสสร จันทร์หอม, ศุภกร รัศมี และโอพาร รุ่งมณีธรรมคุณ, 2560, การเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล, ปัญหาพิเศษปริญญาตรี, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

วีระยุทธ พิมพาภรณ์ และพยุง มีสัจ, 2557, การเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลโดยวิธีการเลือกลักษณะสำคัญแบบพลวัตเพื่อเพิ่มประสิทธิภาพของอัลกอริทึมการจัดกลุ่มบนปริภูมิย่อย, ว.เทคโนโลยีสารสนเทศ 10(2): 43-51.

สายชล สินสมบูรณ์ทอง, 2560, การทำเหมืองข้อมูล, บริษัทจามจุรีโปรดักท์, กรุงเทพฯ.

อุมาพร ยกกำพล, อัชฌาณัท รัตนเลิศนุสรณ์ และอุไรวรรณ เจริญกีรติกุล, 2561, การเปรียบเทียบประสิทธิภาพของการจัดกลุ่มข้อมูลวิธีกำรแบบลำดับขั้นและวิธีการเคมีนสำหรับข้อมูลผสมเชิงกลุ่มกับเชิงตัวเลข, น. 1-10, การประชุมวิชาการสถิติประยุกต์และเทคโนโลยีสารสนเทศระดับชาติ ประจำปี พ.ศ. 2561, คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์, กรุงเทพฯ.

Arwa, A. and Heba, K., 2019, An energy-efficient gossiping protocol for wireless sensor networks using Chebyshev distance, Sci. Direct 151: 1066-1071.

Beckman, R.J. and Cook, R. D., 1983, Outlier … … …. S, J. Technometrics 25: 119-149.

Bhatt, V. Dhakar, M and Chaurasia, B.K., 2016, Filtered clustering based on local outlier factor in data mining, Database Theor. Appl. 9: 275-282.

David, L., 1988, Heart Disease Dataset, Available Source: https://www.kaggle.com/johnsmith88/ heart- disease- dataset, December 12, 2019.

Eason, E., 2018, Sample Insurance Claim Prediction Dataset, Available Source: https://www.kaggle.com/easonlai/sample-insurance- claim- prediction- dataset, December 17, 2019.

Frederic, R. and Serge, G., 2019, A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise, Expert Syst. Appl. 128: 96-108.

Galit, S., Nitin, R.N. and Peter, C.B., 2007, Data Mining for Business Intelligence, John Wiley and Sons, New Jersey.

Harry, C., 2017, Diabetes Dataset, Available Source: https://www.kaggle.com/fmendes/diabetes- from- dat263x- lab01Diabetes from DAT263x Lab01 โรคเบาหวานจาก DAT263x Lab01, December 24, 2019.

Hawkins, D.M., 1980, Identification of Outliers, Springer Science and Business Media, Berlin.

Jiawei, H., Michline, K. and Jian, P., 2006, Data Mining Concepts and Techniques**,** Elsevier, Waltham.

Johnson, R.A. and Wichern, D.W., 2007, Applied Multivariate Statistical Analysis**,** Pearson Prentice Hall, New Jersey.

Merishna, S.S., 2018, Breast Cancer Prediction Dataset, Available Source: https://www.kaggle.com/merishnasuwal/breast- cancer- prediction- dataset, November 27, 2019.

Svetlana, U., 2019, Cardiovascular Disease Dataset, Available Source: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset/version/1, November 20, 2019.