การเปรียบเทียบประสิทธิภาพการทำนายผลการแปลงข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล

Performance Comparison of Transformation Methods in Data Mining Classification Technique

ปิยวรรณ นิลถนอม*, ธนพร มาลัย และ สายชล สินสมบูรณ์ทอง

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Piyawan Ninthanom*, Thanaporn Malai and Saichon Sinsomboonthong

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang

Received: November 10, 2020; Accepted: January 18, 2021

บทคัดย่อ

การแปลงข้อมูลเป็นส่วนหนึ่งในกระบวนการเตรียมข้อมูลก่อนทำเหมืองข้อมูล งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการแปลงข้อมูล 5 วิธี คือ ไม่แปลงข้อมูล การทำให้เป็นปรกติน้อยที่สุด-มากที่สุด การทำให้เป็นมาตรฐานแซด การแปลงข้อมูลให้เป็นเลขทศนิยม และการแปลงข้อมูลโดยค่ามัธยฐาน โดยวิธีการจำแนก 3 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม และวิธีนาอีฟเบส์ ว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความแม่น ทำการแบ่งข้อมูลในอัตราส่วน 70 และ 30 ตามลำดับ ในข้อมูลส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบ ร้อยละ 70 และข้อมูลส่วนที่ 2 ข้อมูลทดสอบ นำไปทดสอบตัวแบบ ร้อยละ 30 โดยการกำหนดตัวสร้างเลขสุ่มเทียม เป็น 10, 20, 30, 40 และ 50 มีข้อมูลที่นำมาแปลงในการศึกษา 6 ชุด และใช้เกณฑ์การแบ่งประเภทของข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มที่ข้อมูลแตกต่างกันน้อย ได้แก่ คุณภาพไวน์ขาว การเป็นโรคเบาหวานของชนเผ่าไพม่า การตรวจกระดูกแกนกลางของร่างกาย และกลุ่มที่ข้อมูลแตกต่างกันมาก ได้แก่ การเป็นโรคตับของคนอินเดีย ชั่วโมงการทำงานของแม่บ้าน และอะโวคาโด โดยใช้โปรแกรมอาร์ จากการเปรียบเทียบข้อมูลทั้งหมด 4 ใน 6 ชุด ให้ผลไปในทิศทางเดียวกันโดย วิธีที่ให้ค่าความแม่นสูงสุดคือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการแปลงข้อมูลให้เป็นเลขทศนิยม รองลงมาคือวิธีโครงข่ายประสามเทียม โดยการแปลงข้อมูลให้เป็นเลขทศนิยม ในงานวิจัยครั้งนี้อาจเป็นประโยชน์โดยตรงต่อผู้ที่มีความสนใจในการทำเหมืองข้อมูลสำหรับข้อมูลที่มีขนาดใหญ่

คำสำคัญ : วิธีการทำให้เป็นปรกติน้อยที่สุด-มากที่สุด วิธีการทำให้เป็นมาตรฐานแซด วิธีการแปลงข้อมูลให้เป็นเลขทศนิยม วิธีการแปลงข้อมูลโดยค่ามัธยฐาน วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีการนาอีฟเบส์

**Corresponding author: 59051217@kmitl.ac.th*

Abstract

Transformation is a data preparation process for data mining. The main objective of this research was to compare five transformation methods in terms of classification accuracy that the transformed data provided. Those methods were the following: No transformation, Min-Max Normalization, Z-Score Standardization, Decimal Scaling, and Median Method. Three classification methods K-Nearest Neighbor, Artificial Neural Network, and Naïve Bayes were used to evaluate the transformation methods. Each of these datasets was divided into two groups at a ratio of 70:30. The first group was a training data set; the second group was a testing data set. The range of tested random seed parameter was from 10, 20, 30, 40 to 50. Six datasets were datasets of White Wine Quality, Pima Indians Diabetes, and Vertebral Column of which data were not much different and datasets of Indian Liver Patient, Working Hours, and Avocado of which data were much different. All algorithms and procedures were implemented in R programming language. On 4 out of 6 tested datasets, transformation by Decimal Scaling and classification by K-Nearest Neighbor were the best combination, followed by transformation by Decimal Scaling and classification by Artificial Neural Network. Our findings may directly benefit those who are interested in efficiently mining some big data.

Keywords: Transformation, Min- Max Normalization, Z- Score Standardization, Decimal Scaling, Median Method, K-Nearest Neighbor, Artificial Neural Network, Naïve Bayes

1. บทนำ

ในปัจจุบันความก้าวหน้าทางเทคโนโลยีสารสนเทศ ทำให้มีการจัดเก็บข้อมูลเป็นจำนวนมาก อย่างไรก็ตามการใช้ข้อมูลโดยส่วนใหญ่ยังอยู่ในลักษณะการดึงข้อมูลจากฐานข้อมูลมาใช้งาน ความรู้ที่ได้จากการวิเคราะห์ข้อมูลเช่นนี้สามารถนำไปใช้ประโยชน์ในการดำเนินงานและการตัดสินใจขององค์กรได้อย่างมาก การทำเหมืองข้อมูลสามารถทำได้หลายรูปแบบ ทั้งนี้ขึ้นอยู่กับจุดประสงค์ของการทำเหมืองข้อมูล ในที่นี้จะขอกล่าวถึงเพียงหัวข้อเดียวคือ เป็นการสร้างตัวแบบการจำแนกประเภทข้อมูลจากข้อมูลที่มีการจำแนกประเภทแล้ว เพื่อใช้ตัวแบบนั้นในการจำแนกข้อมูลใหม่ที่ไม่ทราบประเภท การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการอย่างหนึ่งในการดึงเอาความรู้มาจากชุดข้อมูลต่างๆ มากมาย เพื่อนำความรู้ที่ได้เหล่านั้นไปใช้ให้เป็นประโยชน์ในการตัดสินใจ ความรู้จากข้อมูลที่ได้เหล่านั้นอาจนำมาสร้างการพยากรณ์หรือสร้างเป็นตัวแบบสำหรับการจำแนกหรือแสดงให้เห็นถึงความสัมพันธ์ระหว่างหน่วยต่างๆ ซึ่งการทำเหมืองข้อมูลนั้นสามารถนำไปประยุกต์ใช้ได้กับหลายหน่วยงาน ไม่ว่าจะเป็นทางการเงิน การประกันภัย การแพทย์ และอื่นๆ อีกมากมาย ปัจจุบันมีการทำงานวิจัยหรือการสำรวจสิ่งต่างๆ ที่น่าสนใจเกิดขึ้นมากมาย และในขั้นตอนการทำงานวิจัยเหล่านั้น ผู้วิจัยมักนำระเบียบหรือวิธีการทางสถิติมาใช้เพื่อการวิเคราะห์ข้อมูลและสรุปผลสำหรับงานวิจัยเหล่านั้น เพื่อนำมาแก้ไขหรือพัฒนาในด้านอื่นๆ ต่อไป ซึ่งการที่จะได้มาด้วยข้อมูลต่างๆนั้น ข้อมูลที่รวบรวมมาอาจมีความไม่เป็นระเบียบ มักเกิดปัญหาข้อมูลแต่ละตัวแปรมีค่าแตกต่างกันตั้งแต่แตกต่างกันน้อย แตกต่างกันปาน

กลาง และแตกต่างกันมาก หรืออาจยังมีความเหลื่อมล้ำของหน่วยระหว่างข้อมูลอยู่ก็เป็นได้ หากนำข้อมูลเหล่านั้นมาวิเคราะห์จะทำให้ผลการวิเคราะห์ข้อมูลเกิดการคลาดเคลื่อนไปจากความเป็นจริง ส่งผลให้ไม่เป็นไปตามข้อสันนิษฐานเบื้องต้น (Presumption) และทำให้ไม่สามารถนำข้อมูลไปใช้ประโยชน์ได้อย่างสูงสุด ซึ่งวิธีหนึ่งที่จะนำมาแก้ไขปัญหาในส่วนนี้ คือ การแปลงข้อมูล (Data Transformation) โดยเป็นการใช้วิธีต่างๆ ทางคณิตศาสตร์ที่ไม่ยุ่งยากหรือซับซ้อนเกินไป เพื่อนำมาปรับข้อมูลที่ได้เก็บรวบรวมมาให้อยู่ในรูปแบบขนาดใหม่ เพื่อให้มีความเป็นมาตรฐานเดียวกัน

จากงานวิจัยที่เกี่ยวข้อง (Amit and Achin, 2017) ได้ศึกษาการเปรียบเทียบประสิทธิภาพการแปลงข้อมูล 2 เทคนิค คือ การทำให้เป็นปรกติน้อยที่สุด-มากที่สุด (Min-Max Normalization) กับการทำให้เป็นมาตรฐานแซด (Z-Score Standardization) และใช้เทคนิคจำแนกด้วยวิธีเพื่อนบ้าน ใกล้สุด k ตัว (K Nearest-Neighbor) ที่มีค่า k ต่างกัน โดยการทำโปรแกรมอาร์ และวัดผลด้วยค่าความแม่น (Accuracy) ของการทำนายผลลัพท์ที่ได้คือค่าความแม่นของการทำนายโดยเฉลี่ยเป็น 88.0925% สำหรับเทคนิคการทำให้เป็นปรกติน้อยที่สุด-มากที่สุด และ 78.5675% สำหรับเทคนิคการทำให้เป็นมาตรฐานแซด Patro and Sahu (2016) ได้ทำการศึกษางานวิจัยที่เกี่ยวกับการแปลงข้อมูลด้วย วิธีการทำให้เป็นปรกติน้อยที่สุด-มากที่สุด วิธีการทำให้เป็นมาตรฐานแซด และการแปลงข้อมูลให้เป็นเลขทศนิยม (Decimal Scaling) จากที่ได้ทำการศึกษาพบว่าการแปลงข้อมูลนั้นเป็นไปได้ดี สำหรับทุกๆงานวิจัย ดังนั้นเขาจึงจะเสนอแผนการทำข้อมูลให้เป็นปรกติมาตรฐานในการวิจัยดำเนินงานแขนงอื่นอีกด้วย

ดังนั้นการทำวิจัยครั้งนี้จัดทำขึ้นมาเพื่อศึกษาเกี่ยวกับการแปลงข้อมูลด้วยวิธีต่างๆ ทั้ง 5 วิธี คือไม่แปลงข้อมูล (No Transformation Data) การทำให้เป็นปรกติน้อยที่สุด-มากที่สุด การทำให้เป็นมาตรฐานแซด การแปลงข้อมูลให้เป็นเลขทศนิยม การแปลงข้อมูลโดยค่ามัธยฐาน (Median Method) และศึกษาการจำแนกข้อมูลด้วยวิธีต่างๆ ทั้ง 3 วิธีคือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีนาอีฟเบส์ เพื่อที่จะเปรียบเทียบประสิทธิภาพวิธีการแปลงผลข้อมูลทั้ง 5 วิธี และวิธีการจำแนกอีก 3 วิธี ว่าวิธีใดมีค่าความแม่นที่สูงกว่ากันโดยใช้ชุดข้อมูลต่างๆ 6 ชุด วิเคราะห์ผลโดยใช้โปรแกรมอาร์

2. วิธีการ

2.1 เครื่องมือที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยครั้งนี้คือ โปรแกรมอาร์

2.2 การเก็บรวบรวมข้อมูล การแปลงข้อมูล การแบ่งข้อมูล การศึกษาขั้นตอนวิธีและการเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

2.2.1 การเก็บรวบรวมข้อมูล ทำการศึกษาข้อมูล จากเว็บไซต์ UCI.com Kaggle.com และmldata.com จำนวน 6 ชุด คือ

2.2.1.1 ข้อมูลคุณภาพไวน์ขาว (White Wine Quality) มีจำนวนข้อมูลทั้งหมด 1,500 ค่า แบ่งข้อมูลทั้งหมดออกเป็นข้อมูลเรียนรู้จำนวน 1,050 ค่า และข้อมูลทดสอบจำนวน 450 ค่า ตัวแปรอิสระประกอบด้วย (X_1) ค่าความเป็นกรด (X_2) ค่าความเป็นกรดระเหย (X_3) ค่ากรดซิตริก (X_4) ค่าน้ำตาลรีดิวซ์ (X_5) ค่าคลอไรด์ (X_6) ค่าซัลเฟอร์ไดออกไซด์ (X_7) ค่าซัลเฟอร์ไดออกไซด์ทั้งหมด (X_8) ค่าความหนาแน่น (X_9) ค่าความเป็น

กรด (X_{10}) ค่าเกลือของกรดซัลฟิวริก (X_{11}) ค่าแอลกอฮอล์ (X_{12}) คะแนนคุณภาพ และตัวแปรตาม คือ (Y) คุณภาพไวน์ขาว (Yes คือ คุณภาพดี และ No คือ คุณภาพไม่ดี) โดยมีตัวแปรที่ค่า c.v. ออกนอกเกณฑ์ 1-5 ตัวแปร (Cortez *et al*, 2009)

2.2.1.2 ข้อมูลการเป็นโรคเบาหวานของชนเผ่าไพมา (Pima Indians Diabetes) มีจำนวนข้อมูลทั้งหมด 768 ค่า แบ่งข้อมูลทั้งหมดออกเป็นข้อมูลเรียนรู้จำนวน 537 ค่า และข้อมูลทดสอบจำนวน 231 ค่า ตัวแปรอิสระประกอบด้วย (X_1) จำนวนครั้งที่ตั้งครรภ์ (X_2) ความเข้มข้นของกลูโคสในช่องปาก (X_3) ความดันโลหิต (X_4) ความหนาของไขมันใต้ผิวหนัง (X_5) ความเข้มข้นของอินซูลินในร่างกายหลังฉีดเซรั่มเข้าไป 2 ชม. (X_6) ดัชนีมวลกาย (X_7) ความเสี่ยงของโรคเบาหวานที่มาจากพันธุกรรม (X_8) อายุ และตัวแปรตาม คือ (Y) แทน ผลลัพท์ (Yes แทน เป็นโรคเบาหวาน และNo แทน ไม่เป็นโรคเบาหวาน) โดยมีตัวแปรที่ค่า c.v. ออกนอกเกณฑ์ 1-5 ตัวแปร (Smith *et al*, 1988)

2.2.1.3 ข้อมูลกระดูกแกนกลางของร่างกาย (Vertebral Column) มีจำนวนข้อมูลทั้งหมด 310 ค่า แบ่งข้อมูลทั้งหมดออกเป็นข้อมูลเรียนรู้จำนวน 217 ค่า และข้อมูลทดสอบจำนวน 93 ค่า ตัวแปรอิสระประกอบด้วย (X_1) อุ้งเชิงกราน (X_2) กระดูกเชิงกรานเอียง (X_3) ภาวะกระดูกสันหลังแอ่น (X_4) กระดูกสันหลังส่วนกระเบนเหน็บ (X_5) รัศมีเชิงกราน (X_6) การเคลื่อนของกระดูก และตัวแปรตาม คือ (Y) ความปกติของกระดูกแกนกลาง (Normal แทน กระดูกปกติ และ Abnormal แทน กระดูกไม่ปกติ) โดยมีตัวแปรที่ค่า c.v. ออกนอกเกณฑ์ 1-5 ตัวแปร (Mota, 2011)

2.2.1.4 ข้อมูลการเป็นโรคตับของอินเดีย (Indian Liver Patient) มีจำนวนทั้งหมด 575 ค่า แบ่งข้อมูลทั้งหมดออกเป็นข้อมูลเรียนรู้จำนวน 402 ค่า และข้อมูลทดสอบจำนวน 173 ค่า ตัวแปรอิสระประกอบด้วย (X_1) อายุของผู้ป่วย (X_2) ค่าบิลิรูบินทั้งหมด (X_3) ค่าบิลิรูบิน (X_4) ค่าอัลคาไลน์ ฟอสฟาเตส (X_5) ค่าเอนไซม์อะลานีนอะมิโนทรานเฟเรส (X_6) ค่าแอสปาร์เทต อะมีโนทรานสเฟอเรส (X_7) ปริมาณโปรตีนรวมในกระแสเลือด (X_8) โปรตีนชนิดหนึ่ง (X_9) อัตราส่วนของอัลบูมินและโกลบูลิน และตัวแปรตาม คือ (Y) การตรวจพบโรคตับ (Yes แทน เป็นโรคตับ และ No แทน ไม่เป็นโรคตับ) โดยมีตัวแปรที่ค่า c.v. ออกนอกเกณฑ์ 6-10 ตัวแปร (Ramana, 2012)

2.2.1.5 ข้อมูลชั่วโมงการทำงาน (Working Hours) มีจำนวนทั้งหมด 956 ค่า แบ่งข้อมูลทั้งหมดออกเป็นข้อมูลเรียนรู้จำนวน 669 ค่า และข้อมูลทดสอบจำนวน 287 ค่า ตัวแปรอิสระประกอบด้วย (X_1) เวลาการทำงาน (X_2) รายได้ (X_3) อายุ (X_4) ปีที่ได้รับการศึกษา (X_5) จำนวนลูก (X_6) อัตราการว่างงานในพื้นที่ และตัวแปรตาม คือ (Y) บ้านพักอาศัย (1 แทน บ้านเช่า และ 2 แทน บ้านตัวเอง) โดยมีตัวแปรที่ค่า c.v. ออกนอกเกณฑ์ 6-10 ตัวแปร (Myoung, 1995)

2.2.1.5 ข้อมูลอะโวคาโด (Avocado) มีจำนวนทั้งหมด 1,149 ค่า แบ่งข้อมูลทั้งหมดออกเป็นข้อมูลเรียนรู้จำนวน 804 ค่า และข้อมูลทดสอบจำนวน 345 ค่า ตัวแปรอิสระประกอบด้วย (X_1) จำนวนอะโวคาโดที่จำหน่ายทั้งหมด (X_2) จำนวนทั้งหมดของอะโวคาโดทั้งหมดที่มีขายของ PLU 4046 (X_3) จำนวนทั้งหมดของอะโวคาโดทั้งหมดที่มีขายของ PLU 4225 (X_4) จำนวนทั้งหมดของอะโวคาโดทั้งหมดที่มีขายของ PLU 4770 (X_5) จำนวนทั้งหมดของอะโวคาโดทั้งหมดที่มีขายของ (X_6) ถุงใบเล็กใส่อะโว (X_7) ถุงใบใหญ่ใส่อะโว (X_8) ถุงใบขนาดใหญ่ใส่อะโว และ

ตัวแปรตาม คือ (Y) ประเภทของอะโวคาโด (conventional แทน อะโวคาโดที่ปลูกทั่วไป และ organic อะโวคาโดที่ปลูกแบบออร์แกนิค) โดยมีตัวแปรที่ค่า c.v. ออกนอกเกณฑ์ 6-10 ตัวแปร (Millenials, 2018)

2.2.2 การแปลงข้อมูล

2.2.2.1 การทำให้เป็นปรกติน้อยที่สุด-มากที่สุด (Min-Max Normalization) คือ การทำข้อมูลให้อยู่ในรูปคะแนนปรกติมาตรฐานน้อยที่สุด-มากที่สุด เป็นการแปลงข้อมูลที่เป็นตัวเลข (Numerical Data) ให้มีค่าอยู่ในช่วงที่เป็นมาตรฐานเดียวกันคือ 0 ถึง 1 (สุรพงศ์, 2559)

$$X^{*} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

โดยที่ X^{*} คือ ค่าที่จะทำให้เป็นปรกติมาตรฐานของตัวแปร X

X_{max} คือ ค่าสูงสุดของตัวแปร X

X_{min} คือ ค่าต่ำสุดของตัวแปร X

2.2.2.2 การทำให้เป็นมาตรฐานแซด (Z-Score Standardization) คือ การทำให้เป็นมาตรฐานเป็นวิธีที่ใช้มากในการวิเคราะห์ทางสถิติ โดยเฉพาะอย่างยิ่งการทำให้เป็นคะแนนแซด เป็นการคำนวณหาสัดส่วนระหว่างผลต่างของค่าตัวแปรและค่าเฉลี่ยของตัวแปรเมื่อเทียบกับค่าของส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ของค่าตัวแปรจะมีค่าอยู่ในช่วง -4 ถึง 4 (สุรพงศ์, 2559)

$$X^{*} = \frac{X - \bar{X}}{SD(X)}$$

โดยที่ X^{*} คือ ค่าที่จะทำให้เป็นมาตรฐานแซดของตัวแปร X

$\bar{X}$ คือ ค่าเฉลี่ยของตัวแปร X

SD(X) คือ ค่าส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปร X

2.2.2.3 การแปลงข้อมูลให้เป็นเลขทศนิยม (Decimal Scaling) คือ การแปลงข้อมูลให้เป็นเลขทศนิยม ตำแหน่งทศนิยมจะกำหนดโดยค่าสัมบูรณ์ที่มีค่ามากที่สุด

$$X^{*} = \frac{X}{10^{j}}$$

โดยที่ X^{*} คือ ค่าที่จะทำเป็นปรกติมาตรฐานของตัวแปร X

j คือ จำนวนหน่วยที่มากกว่า X หน่วย

2.2.2.4 การแปลงข้อมูลโดยใช้ค่ามัธยฐาน (Median Method) คือ การแปลงข้อมูลโดยการใช้ค่ากลางของข้อมูลเป็นตัวหารข้อมูลแต่ละตัวในแต่ละแถว

$$X^{*} = \frac{X}{MED\ (X)}$$

โดยที่ X^{*} คือ ค่าที่จะทำเป็นปรกติมาตรฐานของตัวแปร X

MED (X) คือ ค่ามัธยฐานหรือค่ากลางของข้อมูลแต่ละแถว

2.2.3 การแบ่งข้อมูล

2.2.3.1 ทำการแบ่งชุดข้อมูลโดยโปรแกรมอาร์ ทำการสุ่มจำนวน 5 รอบ โดยการกำหนดตัวสร้างเลขสุ่มเทียมเป็น 10, 20, 30, 40 และ 50 ในอัตราส่วน 70:30 ส่วนที่ 1 ข้อมูลเรียนรู้ (Training Data) นำไปสร้างตัวแบบร้อยละ 70 และข้อมูลส่วนที่ 2 ข้อมูลทดสอบ (Testing Data) นำไปทดสอบตัวแบบร้อยละ 30 (Shams, 2014)

2.2.4 การศึกษาขั้นตอนวิธี

2.2.4.1 วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbor) เป็นวิธีการที่ได้รับความนิยมอย่างมาก เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพ ซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย เช่น งานทางด้านการจำแนก รวมถึงงานทางด้านการแทนที่ข้อมูลที่สูญหาย ใช้

ขั้นตอนวิธี IBk ซึ่งมีวิธีการดำเนินการดังนี้ (Troyanskaya, 2001)

i) กำหนดค่า k เพื่อใช้พิจารณาสมาชิกที่อยู่ใกล้กันมากที่สุด เช่น k=3 คือจะพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่อยู่ใกล้กับจุดที่ต้องการทำนาย

ii) คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่นๆ ทุกตัวด้วยระยะห่างยูคลิเดียน (Euclidian Distance) จากสมการ (1) ดังนี้

$$D_{Euclidian}(x_i, y_i) = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \qquad (1)$$

โดยที่ $D_{Euclidian}(x_i, y_i)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง y_i

k คือ คุณลักษณะทั้งหมดของตัวอย่าง

iii) เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ

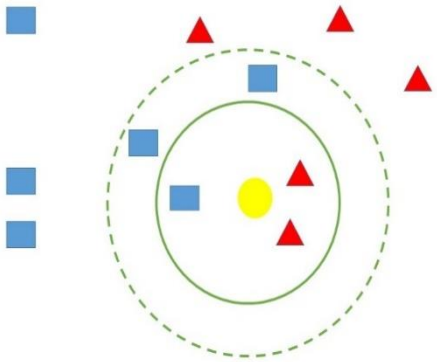

Figure 1 k-nearest neighbors = 1

จากFig.1 กำหนดให้จุดที่ต้องการทำนาย คือ วงกลม ควรจัดกลุ่มให้จุดที่ต้องการทำนายไปอยู่ในกลุ่มแรกของสี่เหลี่ยม หรือกลุ่มสองของสามเหลี่ยม

ถ้า k = 3 แล้ว วงกลมจะอยู่ในกลุ่มสอง เพราะมีสี่เหลี่ยม 1 รูป และสามเหลี่ยม 2 รูป อยู่ในวงกลมใน

ถ้า k = 5 แล้ว วงกลมจะอยู่ในกลุ่มแรก เพราะมีสี่เหลี่ยม 3 รูป และสามเหลี่ยม 2 รูป อยู่ในวงกลมนอก

IBk เป็นฟังก์ชันหลัก ซึ่งเป็นพื้นฐานของวิธีเพื่อนบ้านใกล้สุด k ตัว อย่างไรก็ตามขั้นตอนวิธี IBk ยังสามารถกำหนดน้ำหนัก ระยะห่างและทางเลือก (Option) เพื่อกำหนดค่า k โดยใช้การตรวจสอบไขว้ (Cross Validation) (สุรวัชร และสายชล, 2560) ซึ่งในที่นี้กำหนดตามโปรแกรม (Default) ให้ k = 1

2.2.4.2 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) เครือข่ายใยประสาทเป็นเทคโนโลยีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence: AI) เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของเครือข่ายใยประสาทเป็นวิธีที่ให้เครื่องได้เรียนรู้จากตัวอย่างต้นแบบแล้วฝึกให้ระบบได้รู้จักคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของเครือข่ายใยประสาทประกอบด้วยโหนดสำหรับข้อมูลเข้า-ข้อมูลออก (Input-output Data) และการประมวลผลกระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ ชั้นข้อมูลเข้า (Input Layer) ชั้นข้อมูลออก (Output Layer) และ ชั้นซ่อน (Hidden Layers) การประมวลผลของเครือข่ายใยประสาทจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ในชั้นเหล่านี้ ตัวอย่างเครือข่ายใยประสาท ดังแสดงใน Fig.2

โครงข่ายประสามเทียมอย่างง่าย (Simple Example of a Neural Network) ส่วนใหญ่จะประกอบด้วย 3 ชั้น คือ ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออกหรือชั้นผลลัพธ์ โครงข่ายแบบเชื่อมต่อกันอย่างสมบูรณ์เป็นโครงข่ายประสามเทียมที่โหนดทุกโหนดในชั้นที่กำหนดเชื่อมต่อกันทุกโหนดกับชั้นถัดไปแม้จะไม่ได้เชื่อมต่อกับโหนด

อื่นใดในชั้นเดียวกัน การเชื่อมต่อกันระหว่างโหนดมีการถ่วงน้ำหนักที่สัมพันธ์กันในขณะเริ่มต้น การจัดน้ำหนักถูกจัดอย่างสุ่มโดยมีค่าอยู่ระหว่าง 0 ถึง 1 โดยทั่วไปจำนวนโหนดข้อมูลเข้าขึ้นอยู่กับจำนวนและชนิดของคุณลักษณะในชุดข้อมูล จำนวนของชั้นซ่อนและจำนวนของโหนดในชั้นซ่อนขึ้นอยู่กับผู้ใช้งานเป็นผู้กำหนด จำนวนของโหนดในชั้นข้อมูลออกอาจจะมีมากกว่า 1 โหนด ซึ่งขึ้นอยู่กับงานในการจำแนกกลุ่ม ดังสมการ (2)

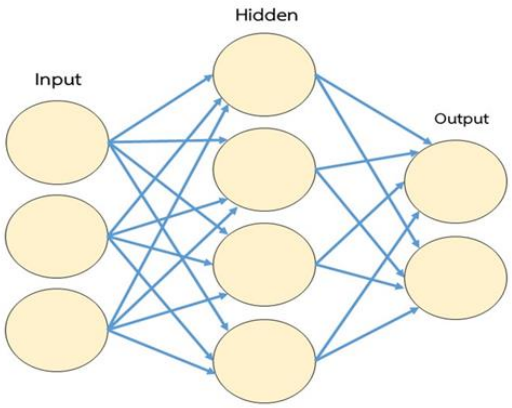

Figure 2 Functioning of the neural network

$W_{ij}x_{ij}= W_{0_j}x_{0_j}+W_{1_j}x_{1_j}+\ldots+W_{ij}x_{ij}$ (2)

โดยที่ X_{ij} คือ ข้อมูลเข้าที่ I ไปยังโหนดที่ j

W_{ij} คือ น้ำหนักถ่วงที่สัมพันธ์กับข้อมูลเข้าที่ i ไปยังโหนด j และมีข้อมูลเข้าจำนวน I+1 ไปยังโหนด j

2.2.4.3 นาอีฟเบส์ (Naïve Bayes) จะใช้วิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น โดยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อน รูปแบบการหาความสัมพันธ์ที่ไม่ซับซ้อนได้ผลลัพธ์ดี ดังสมการ (3)

$$P(C|A) = \frac{P(A|C)\,x\,P(C)}{P(A)} \quad (3)$$

จากสมการเบส์ (Bayes) อธิบายว่าถ้าต้องการทำนายคำตอบ (Class) C เมื่อทราบคุณลักษณะ (Attribute) สามารถคำนวณจากความน่าจะเป็นของคุณลักษณะ A ที่มีคำตอบ C ในข้อมูลเรียนรู้ และค่าความน่าจะเป็นของคุณลักษณะ A และ C มีสมการ 3 ส่วนดังนี้

$P(C|A)$ คือ ค่าความน่าจะเป็นที่ข้อมูลเรียนรู้ที่มีคุณลักษณะ A จะมีคำตอบ C

$P(A|C)$ คือ ค่าความน่าจะเป็นที่ข้อมูลเรียนรู้ที่มีคำตอบ C และมีคุณลักษณะ A โดยที่ $A = A_1 \cap A_2 \ldots \cap A_M$ โดยที่ M คือจำนวนคุณลักษณะในข้อมูลเรียนรู้

$P(C)$ คือ คือ ความน่าจะเป็นของคำตอบ C

$P(A)$ คือ คือ ความน่าจะเป็นของคำตอบ A

แต่การที่คุณลักษณะ $A = A_1 \cap A_2 \ldots \cap A_M$ ที่เกิดขึ้นในข้อมูลเรียนรู้อาจจะมีจำนวนน้อยมากหรือไม่มีรูปแบบของคุณลักษณะแบบนี้เกิดขึ้นเลย ดังนั้นจึงได้ใช้หลักการที่ว่าแต่ละคุณลักษณะแบบนี้เกิดขึ้นเลย ดังนั้นจึงได้ใช้หลักการที่ว่าแต่ละคุณลักษณะเป็นอิสระกัน ทำให้สามารถเปลี่ยนสมการ $P(A|C)$ ได้เป็น

$P(A|C)=P(A_1|C) \times P(A_2|C) \times \ldots \times P(A_M|C)$

หลังจากนั้นนำข้อมูลที่แบ่งออกเป็น 2 ส่วนมาทำการวิเคราะห์โดยใช้โปรแกรมอาร์ ซึ่งทำการวิเคราะห์จากวิธีการจำแนกทั้ง 3 วิธี ข้างต้น

2.2.5 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

นำผลการวิเคราะห์ของแต่ละวิธีทั้ง 3 วิธี มาเปรียบเทียบประสิทธิภาพโดยพิจารณาจากเมทริกซ์ความสับสน (Confusion Matrix) ซึ่งเป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในตารางเมทริกซ์ความสับสนซึ่งจะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของขั้นตอนวิธีดังTable.*1*

2. 2. 5. 1 ค่าความแม่น (Accuracy) ในการทำนาย คือการแสดงการวัดที่ได้มีความแม่น ในรูปอัตราส่วนโดยคิดเป็นร้อยละ Accuracy = จำนวนข้อมูลที่จำแนกถูกว่าคำตอบ

$$\text{Accuracy} = \frac{\text{เป็นบวกและลบ}}{\text{จำนวนข้อมูลทั้งหมด}} \times 100\%$$

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

Table 1 Confusion matrix

		Predicted condition	
		Condition Positive	Condition Negative
Actual condition	Condition Positive	TP (True Positive)	FN (False Negative)
	Condition Negative	FP (False Positive)	TN (True Negative)

Note : A true positive is an outcome where the model correctly predicts the positive class.
A true negative is an outcome where the model correctly predicts the negative class.
A false positive is an outcome where the model incorrectly predicts the positive class.
A false negative is an outcome where the model incorrectly predicts the negative class.

3. ผลการวิจัยและวิจารณ์ผล

3.1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนก

งานวิจัยครั้งนี้ใช้การจำแนกโดยใช้วิธีการทำเหมืองข้อมูล โดยนำชุดข้อมูลจำนวน 6 ชุด มาทำการวิเคราะห์ข้อมูล ซึ่งจะสุ่มแบ่งข้อมูลออกเป็น 2 ส่วน คือ ส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบร้อยละ 70 และข้อมูลส่วนที่ 2 ข้อมูลทดสอบ นำไปทดสอบตัวแบบร้อยละ 30 และผู้วิจัยได้นำมาเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกโดยพิจารณาจากค่าความแม่น ซึ่งวิธีที่ใช้ในการทดสอบครั้งนี้มี 3 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม และวิธีนาอีฟเบส์ ผลการวิเคราะห์ข้อมูลจากข้อมูลชุดที่ 1 ซึ่งเป็นกรณีที่มีตัวแปรที่ค่าออกนอกเกณฑ์ต่ำ พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมให้ค่าความแม่นสูงสุดคือ ร้อยละ 100 และ วิธีโครงข่ายประสาทเทียมด้วยวิธีการทำให้เป็นปรกติน้อยที่สุด-มากที่สุด วิธีการทำให้เป็นมาตรฐานแซด วิธีการแปลงข้อมูลให้เป็นเลขทศนิยม และวิธีการแปลงข้อมูลโดยค่ามัธยฐาน ให้ค่าความแม่นสูงสุดคือร้อยละ 100 และวิธีนาอีฟเบส์ด้วยวิธีการทำให้เป็นปรกติน้อยที่สุด-มากที่สุดให้ค่าความแม่นสูงสุดคือร้อยละ 99.23767 ข้อมูลชุดที่ 2 ซึ่งเป็นกรณีที่มีตัวแปรที่ค่าออกนอกเกณฑ์ต่ำ พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการทำให้เป็นปรกติน้อยที่สุด-มากที่สุดให้ค่าความแม่นสูงสุดคือร้อยละ 69.7836 วิธีโครงข่ายประสาทเทียมด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมให้ค่าความแม่นสูงสุดคือร้อยละ 77.2824 น้อยที่สุด-มากที่สุดสูงสุดคือร้อยละ 77.31446

Table 2 Comparison of the efficiency of different classification methods (Data set 1-3)

Data Transformation Methods	Accuracy		
	Data set 1	Data set 2	Data set 3
k-nearest Neighbors			
No Transformation Data	75.0666	66.6666	84.0860
Min-Max Normalization	95.2	69.7836	79.3548
Z-Score Standardization	93.4222	68.4849	81.9354
Decimal Scaling	100	69.4337	100
Median Method	83.7777	65.6277	79.7849
Artificial Neural Network			
No Transformation Data	92.6186	68.3070	67.0217
Min-Max Normalization	100	77.0998	84.0798
Z-Score Standardization	100	77.0043	83.4525
Decimal Scaling	100	77.2824	83.9367
Median Method	100	76.9123	84.0798
Naïve Bayes			
No Transformation Data	99.0118	74.7735	77.3144
Min-Max Normalization	99.2376	75.2761	75.7143
Z-Score Standardization	98.8446	75.7745	74.5348
Decimal Scaling	99.0235	64.4787	75.0115
Median Method	98.7066	74.4413	76.5910

และวิธีนาอีฟเบส์ด้วยวิธีการทำให้เป็นมาตรฐานแซดให้ค่าความแม่นสูงสุดคือร้อยละ 75.7745 ข้อมูลชุดที่ 3 ซึ่งเป็นกรณีที่มีตัวแปรที่ค่าออกนอกเกณฑ์ต่ำ พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมให้ค่าความแม่นสูงสุดคือร้อยละ 100 วิธีโครงข่ายประสาทเทียมด้วยวิธีการทำให้เป็นปรกติวิธีการแปลงข้อมูลโดยค่ามัธยฐานให้ค่าความแม่นสูงสุดคือร้อยละ 84.07986 และวิธีนาอีฟเบส์ด้วยวิธีไม่แปลงข้อมูลให้ค่าความแม่น ข้อมูลชุดที่ 4 ซึ่งเป็นกรณีที่มีตัวแปรที่ค่าออกนอกเกณฑ์สูง พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วย วิธีการแปลงข้อมูลโดยค่ามัธยฐานให้ค่าความแม่นสูงสุดคือร้อยละ 65.7584 วิธีโครงข่ายประสาทเทียมด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมให้ค่าความแม่นสูงสุดคือร้อยละ 70.2721 และวิธีนาอีฟเบส์ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมให้ค่าความแม่นสูงสุดคือร้อยละ 71.1385 ข้อมูลชุดที่ 5 ซึ่งเป็นกรณีที่มีตัวแปรที่ค่าออกนอกเกณฑ์สูง พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมให้ค่าความแม่นสูงสุดคือร้อยละ 100 วิธีโครงข่ายประสาทเทียมด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมให้ค่าความแม่นสูงสุดคือร้อยละ 78.0755 และวิธีนาอีฟเบส์ด้วยวิธีไม่แปลงข้อมูลให้ค่าความแม่นสูงสุดคือ

ร้อยละ 79.69319 และจากข้อมูลชุดที่ 6 ซึ่งเป็นกรณีที่มีตัวแปรที่ค่าออกนอกเกณฑ์สูง พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีไม่แปลงข้อมูลให้ค่าความแม่นสูงสุดคือร้อยละ 100 วิธีโครงข่ายประสาทเทียมด้วยวิธีการแปลงข้อมูลโดยค่ามัธยฐานให้ค่าความแม่นสูงสุดคือร้อยละ 99.5133 และวิธีนาอีฟเบส์ด้วยวิธีไม่แปลงข้อมูล วิธีการแปลงข้อมูลโดยค่ามัธยฐานให้ ค่าความแม่นสูงสุดคือร้อยละ 90.9456

Table 3 Comparison of the efficiency of different classification methods (Data set 4-6)

Data Transformation Methods	Accuracy		
	Data set 4	Data set 5	Data set 6
k-nearest Neighbors			
No Transformation Data	64.9711	72.6132	100
Min-Max Normalization	63.0058	70.9407	99.5942
Z-Score Standardization	64.1619	71.6376	99.7101
Decimal Scaling	61.3341	100	85.3333
Median Method	65.7584	72.8223	99.4782
Artificial Neural Network			
No Transformation Data	42.0796	74.9277	97.7313
Min-Max Normalization	69.5860	78.2005	99.2074
Z-Score Standardization	65.7435	78.1346	99.2691
Decimal Scaling	70.2721	78.0755	99.3303
Median Method	60.6806	77.9938	99.5133
Naïve Bayes			
No Transformation Data	61.5759	79.6931	90.9456
Min-Max Normalization	59.5770	79.6227	90.4500
Z-Score Standardization	63.2700	79.5518	90.0744
Decimal Scaling	71.1385	78.8154	89.4525
Median Method	63.9079	79.6222	90.9456

4. สรุป

4.1 สรุปผลการวิจัย

งานวิจัยนี้ได้ทำการศึกษาประสิทธิภาพในการแปลงข้อมูล 5 วิธี คือ ไม่แปลงข้อมูล การทำให้เป็นปรกติน้อยที่สุด-มากที่สุด การทำให้เป็นมาตรฐานแซด การแปลงข้อมูลให้เป็นเลขทศนิยมและ การแปลงข้อมูลโดยค่ามัธยฐาน ด้วยวิธีการจำแนก 3 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบส์ และ วิธีโครงข่ายประสาทเทียม ว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความแม่น โดยใช้ชุดข้อมูล 6 ชุด ชุดข้อมูลที่มีค่าตัวแปรออกนอกเกณฑ์ต่ำจำนวน 3 ชุด ข้อมูลชุด

ที่ 1 คุณภาพไวน์ขาว วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยม และวิธีโครงข่ายประสาทเทียม ด้วยวิธีการทำให้เป็นปรกติน้อยที่สุด-มากที่สุด วิธีการทำให้เป็นมาตรฐานแซด วิธีการแปลงข้อมูลให้เป็นเลขทศนิยม และวิธีการแปลงข้อมูลโดยค่ามัธยฐาน ข้อมูลชุดที่ 2 การเป็นโรคเบาหวานของชนเผ่าไพม่า วิธีที่มีประสิทธิภาพสูงสุดคือวิธีโครงข่ายประสาทเทียมด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยม ข้อมูลชุดที่ 3 การตรวจกระดูกแกนกลางของร่างกาย วิธีที่มีประสิทธิภาพสูงสุดคือวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยม ข้อมูลชุดที่ 4 การเป็นโรคตับของคนอินเดีย วิธีที่มีประสิทธิภาพสูงสุดคือวิธีนาอีฟเบส์ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยม ข้อมูลชุดที่ 5 ชั่วโมงการทำงานของแม่บ้าน วิธีที่มีประสิทธิภาพสูงสุดคือวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยม ข้อมูลชุดที่ 6 อะโวคาโด วิธีที่มีประสิทธิภาพสูงสุดคือวิธีเพื่อนบ้านใกล้สุด k ตัว โดยไม่แปลงข้อมูล

4.2 อภิปรายผล

จากการทำวิจัยครั้งนี้ ชุดข้อมูลที่มีค่านอกเกณฑ์น้อย คือ ข้อมูลคุณภาพไวน์ขาว ข้อมูลการเป็นโรคเบาหวานของชนเผ่าไพม่า และข้อมูลการตรวจกระดูกแกนกลางของร่างกาย โดยข้อมูลส่วนใหญ่ วิธีที่มีประสิทธิภาพสูงสุดคือวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยม และชุดข้อมูลที่มีค่านอกเกณฑ์มาก คือ ข้อมูลการเป็นโรคตับของคนอินเดีย ข้อมูลชั่วโมงการทำงานของแม่บ้าน และข้อมูลอะโวคาโด โดยข้อมูลส่วนใหญ่ วิธีที่มีประสิทธิภาพสูงสุดคือวิธีเพื่อนบ้านใกล้สุด k ตัว ด้วยวิธีการแปลงข้อมูลให้เป็นเลขทศนิยมเช่นกัน ให้ผลสอดคล้องกับ Patro and Sahu (2016) ได้ทำการศึกษางานวิจัยที่เกี่ยวกับการแปลงข้อมูลด้วยวิธีการทำให้เป็นปรกติน้อยที่สุด-มากที่สุด วิธีการทำให้เป็นมาตรฐานแซด และการแปลงข้อมูลให้เป็นเลขทศนิยม จากที่ได้ทำการศึกษาพบว่าการแปลงข้อมูลนั้นเป็นไปได้ดีสำหรับทุกๆงานวิจัย ดังนั้นจึงจะเสนอแผนการทำข้อมูลให้เป็นมาตรฐานในการวิจัยดำเนินงานแขนงอื่นอีกด้วย

4.3 ข้อเสนอแนะ

4.3.1 เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น ดังนั้นผู้วิจัยอาจวิเคราะห์ข้อมูลด้วยวิธีอื่นๆ ได้แก่ วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วิธีฐานกฎ (Rule-Based) วิธีลาดลงสโตแคสติก (Stochastic Gradient Descent) วิธีเบส์เนท (Bayes Net) วิธีการถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) และวิธีเพอร์เซปตรอนโหวตคะแนน (Voted Perceptron)

4.3.2 การแปลงข้อมูลยังมีวิธีอื่นๆ ที่สามารถแปลงค่าข้อมูลได้ เพื่อให้ได้ผลการแปลงข้อมูลที่มีความสมบูรณ์มากขึ้น ผู้วิจัยอาจจะทำการแปลงข้อมูลด้วยวิธีอื่นๆ เช่น วิธีการทำให้เป็นปรกติแบบซิกมอยด์ (Sigmoid Normalization) และวิธีการทำให้เป็นปรกติแบบสดมภ์เชิงสถิติ (Statistical Column Normalization) เป็นต้น

4.3.3 การศึกษาวิธีการแปลงข้อมูลและการจำแนกข้อมูลยังมีโปรแกรมอื่นๆ ที่สามารถแปลงข้อมูลและจำแนกข้อมูล เช่น SAS และ Python เป็นต้น

4.4 การนำไปใช้ประโยชน์

สามารถนำผลการเปรียบเทียบประสิทธิภาพที่ได้ไปใช้เป็นแนวทางในการเลือกวิธีการแปลงข้อมูลและวิธีการจำแนกที่เหมาะสมที่สุด

5. References

Amit, P. and Achin, J., 2017, Comparative Analysis of KNN Algorithm using Various Normalization Techniques, International Computer Network and Information Security 11: 36-42.

Ramana, B. V., 2012, Indian Liver Patient, Available Source: https: / / www. mldata. io/ dataset-details/ indian_liver_patient/ December 26, 2019.

Mota, H. D., 2011, Vertebral Column Data Set, Available Source: https://www.kaggle.com/caesarlupum/vertebralcolumndataset, January 25, 2020.

Justin K., 2018, Avocado Prices Data Set, Available Source: https://www.kaggle.com/neuromusic/avocado-prices, December 20, 2019.

Patro, S. K. and Sahu, K. K., 2017, Normalization: A Preprocessing Stage, Department of CSE & IT, VSSUT, Burla, Odisha, India.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., 2009, Wine Quality Data Set, Available Source: archive.ics.uci.edu/ml/datasets/Wine+Quality, December 8, 2019.

Shams, R., 2014, Creating Training, Validation and Test Sets Data Preprocessing, Available Source: www. youtube. com/ uiDFa7iY9yo, January 13, 2020.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B, 2001, Missing Values Estimation Methods for DNA Microarrays Bioinformatics, 17(1): 520-525.