

การเปรียบเทียบประสิทธิภาพการแทนค่าข้อมูลสูญหาย กับการจำแนกกลุ่ม 4 วิธี

Efficiency Comparison in Missing Value Replacement with Four Classification Methods

สายชล สิ้นสมบุญทอง*

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Saichon Sinsomboonthong*

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

Chalongkrung Road, Ladkrabang, Bangkok 10520

Received: February 4, 2020; Accepted: March 11, 2020

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการแทนค่าข้อมูลสูญหาย 5 วิธี คือ ค่าเฉลี่ยอนุกรม ค่าเฉลี่ยของค่าใกล้เคียง ค่ามัธยฐานของค่าใกล้เคียง การประมาณค่าในช่วงเชิงเส้น และ แนวโน้มเชิงเส้น แล้วนำข้อมูลที่แทนค่าสูญหายทั้ง 5 วิธี ข้างต้นมาจำแนกกลุ่ม 4 วิธี คือ วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีนาอีฟเบส์ และวิธีการถดถอยลอจิสติกทวิภาคว่าวิธีใดมีประสิทธิภาพในการจำแนกกลุ่มดีที่สุด พิจารณาจากค่าความแม่นยำ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย โดยแบ่งข้อมูลในอัตราส่วน 70, 20 และ 10 ตามลำดับ ข้อมูลส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบ ร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำข้อมูลไปประเมินความผิดพลาดของตัวแบบ ร้อยละ 20 และ ข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบ ร้อยละ 10 โดยการกำหนดตัวสร้างเลขสุ่มเทียมเป็น 10, 20, 30, 40 และ 50 มีข้อมูลสูญหายในการศึกษา 3 ชุด คือ ชุดข้อมูลโรคหัวใจ ชุดข้อมูลประสิทธิภาพในการสอบของนักเรียน และชุดข้อมูลการซื้อของในวันศุกร์สัปดาห์สุดท้ายของเดือนพฤศจิกายน โดยใช้โปรแกรม WEKA การเปรียบเทียบชุดข้อมูลที่มีค่าสูญหายต่ำ โรคหัวใจ วิธีการจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีต้นไม้ตัดสินใจ แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง ชุดข้อมูลที่มีค่าสูญหายปานกลาง ประสิทธิภาพในการสอบของนักเรียน วิธีการจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีการถดถอยลอจิสติกทวิภาค แทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าในช่วงเชิงเส้น และชุดข้อมูลที่มีค่าสูญหายสูง การซื้อของในวันศุกร์สัปดาห์สุดท้ายของเดือนพฤศจิกายน วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีนาอีฟเบส์ แทนค่าข้อมูลสูญหายด้วยวิธีค่ามัธยฐานของค่าใกล้เคียง

คำสำคัญ : วิธีต้นไม้ตัดสินใจ; วิธีโครงข่ายประสาทเทียม; วิธีนาอีฟเบส์; วิธีการถดถอยลอจิสติกทวิภาค

Abstract

We compared the missing value replacement efficiency of five methods, i.e. series mean, mean of nearby points, median of nearby points, linear interpolation, and linear trend at point using four classification methods, including decision tree, artificial neural network, naive Bayes and binary logistic regression. Metrics were accuracy, mean square error and mean absolute error. The data sets were heart disease, students' performance in exams and black Friday. Each of these data sets was divided into three proportions in the ratio of 70 : 20 : 10. By using the data part 1, training data are used to create a model 70 percentages. For the data part 2, validation data are used to evaluate an error as a model 20 percentages, and the data part 3, testing data are used to test a model 10 percentages using the random seeds of 10, 20, 30, 40 and 50 by WEKA program. When we compared the heart disease data set, the best classification method was the decision tree in missing value replacement with mean of nearby points. For the students' performance in exams data sets, the best classification method was the binary logistic regression in missing value replacement with linear interpolation. For the black Friday data sets, the best method was the naive Bayes in missing value replacement with median of nearby points.

Keywords: decision tree; artificial neural network; naive Bayes; binary logistic regression

1. คำนำ

การดำเนินการกับข้อมูลสูญหายพบว่าการเกิดข้อมูลสูญหายเป็นสิ่งที่พบได้เป็นปกติในการเก็บรวบรวมข้อมูล ซึ่งอาจเกิดจากกระบวนการวิจัย เช่น การกำหนดประเด็นการวิจัยที่กว้างเกินไป การใช้ข้อคำถามที่ไม่เหมาะสม บางกรณีข้อมูลสูญหายเกิดจากผู้ให้ข้อมูลเอง เช่น ผู้ให้ข้อมูลไม่เต็มใจให้ข้อมูล จึงทำให้ข้อมูลชุดนั้นเกิดข้อมูลสูญหาย ข้อมูลสูญหายอาจเกิดจากความคลาดเคลื่อนในการเก็บรวบรวมข้อมูลและการบันทึกข้อมูล (สายชล, 2560) เมื่อเกิดข้อมูลสูญหายในการวิจัย นักวิจัยควรดำเนินการกับข้อมูลสูญหาย ไม่ควรปล่อยให้ข้อมูลสูญหายหรือไม่ควรตัดข้อมูลสูญหายทิ้งไป (วุฒิ, 2558) การตัดข้อมูลสูญหายออกจากการวิเคราะห์จะมีผลทำให้จำนวนข้อมูลลดน้อยลง แต่ในบางครั้งข้อมูลอาจมีจำนวนไม่มากนัก จึงไม่สามารถตัดข้อมูลออก ทำให้ต้องมีการประมาณค่าข้อมูลสูญ

หาย จึงมีการแทนค่าข้อมูลสูญหายและจำแนกข้อมูล เพื่อให้ได้วิธีที่เหมาะสมกับข้อมูลที่มีความแตกต่างกันไป ดังนั้นผู้ทำวิจัยจึงเลือกวิธีการแทนค่าข้อมูลสูญหายและจำแนกให้เหมาะสมกับข้อมูล (วรฤทธิ์, 2552)

การนำเสนอวิธีการแทนค่าสูญหายในโปรแกรม SPSS โดยแทนค่าข้อมูลสูญหาย (replace missing value) มีวิธีต่าง ๆ ประกอบด้วย 5 วิธี ได้แก่ วิธีค่าเฉลี่ยอนุกรม (series mean) วิธีค่าเฉลี่ยของค่าใกล้เคียง (mean of nearby point) วิธีมัธยฐานของค่าใกล้เคียง (median of nearby point) วิธีประมาณค่าในช่วงเชิงเส้น (linear interpolation) และวิธีแนวโน้มเชิงเส้น (linear trend at point) ซึ่งแต่ละวิธีจะมีวิธีการ การคำนวณ และรูปแบบการใช้งานที่ต่างกันตามลักษณะของข้อมูล โดยที่วิธีค่าเฉลี่ยอนุกรมเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีค่าใกล้เคียงกันทั้งหมด วิธี

ค่าเฉลี่ยของค่าใกล้เคียงเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีค่าใกล้เคียงกันระหว่างค่าสูญหาย วิธีมัธยฐานของค่าใกล้เคียงเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีค่าสูงหรือต่ำต่างกันมากที่อยู่ระหว่างค่าสูญหาย วิธีประมาณค่าในช่วงเชิงเส้นเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีค่าใกล้เคียงกันระหว่างค่าสูญหาย และข้อมูลสูญหายอยู่ติดกันมีความสัมพันธ์กับข้อมูลที่อยู่ระหว่างกลุ่มค่าสูญหายเป็นแบบถ่วงน้ำหนัก และวิธีแนวโน้มเชิงเส้นเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีความสัมพันธ์กันแบบเชิงเส้นระหว่างตัวแปรของค่าที่ไม่สูญหายกับตำแหน่งของค่าที่ไม่สูญหาย (ฐนัญ, 2559)

การศึกษางานวิจัยในการแทนค่าสูญหายด้วยวิธีค่าเฉลี่ย ค่ามัธยฐาน และค่าใกล้เคียงที่สุด (closest value) เปรียบเทียบกัน โดยศึกษาปัจจัยที่ทำให้เกิดค่าสูญหายที่เกิดจากความผิดพลาดที่เกิดจากมนุษย์ ฮาร์ดแวร์ หรืออื่น ๆ ซึ่งต้องประมาณค่าสูญหายก่อนที่จะนำชุดข้อมูลไปทำเหมืองข้อมูลด้วยการประมาณค่าสูญหายทั้ง 3 วิธี แล้วนำชุดข้อมูลที่ถูกประมาณค่าข้อมูลสูญหายแล้วนำไปทำเหมืองข้อมูล เพื่อวัดประสิทธิภาพของการประมาณค่าว่าวิธีใดมีประสิทธิภาพดีที่สุด โดยพิจารณาจากค่าความแม่นยำ (accuracy) ที่สูงกว่า โดยการจำแนกด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว จากงานวิจัยพบว่าวิธีค่าใกล้เคียงที่สุดมีค่าความแม่นยำสูงสุด (Rahman and Davis, 2013) ส่วนในการจัดการค่าข้อมูลสูญหายในชุดข้อมูลจากงานวิจัยที่มีจำนวนข้อมูลมากที่มีข้อมูลสูญหาย โดยการประมาณค่าข้อมูลสูญหายด้วยวิธีค่ากลาง (common value) มีทั้งหมด 2 วิธี ได้แก่ วิธีค่าเฉลี่ยและวิธีค่ามัธยฐานกับวิธีค่าใกล้เคียงที่สุด แล้วนำชุดข้อมูลที่ถูกประมาณค่าข้อมูลสูญหายแล้วไปทำเหมืองข้อมูลเพื่อวัดประสิทธิภาพในการจำแนกกลุ่มว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยการจำแนกมี

ทั้งหมด 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม และวิธีกฎความสัมพันธ์ โดยพิจารณาจากค่าความแม่นยำที่สูงกว่า งานวิจัยพบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียมและวิธีกฎความสัมพันธ์ให้ค่าความแม่นยำสูงสุด เมื่อใช้วิธีค่าใกล้เคียงที่สุดในการประมาณค่าข้อมูลสูญหาย (Kaiser, 2014)

การประเมินความถูกต้องของวิธีการประมาณค่าเพื่อการแทนค่าข้อมูลสูญหายสำหรับข้อมูลอนุกรมเวลาโดยใช้โปรแกรมสำเร็จรูป SPSS และ MINITAB ใช้วิธีการแทนค่าข้อมูลสูญหาย 5 วิธี คือ ค่าเฉลี่ยอนุกรม ค่าเฉลี่ยของค่าใกล้เคียง ค่ามัธยฐานของค่าใกล้เคียง การประมาณค่าในช่วงเชิงเส้นและแนวโน้มเชิงเส้นที่ทำให้ค่าความคลาดเคลื่อนในการทำนายน้อยที่สุด โดยพิจารณาจากค่าคลาดเคลื่อนเปอร์เซ็นต์สัมบูรณ์เฉลี่ย (mean absolute percentage error, MAPE) ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (mean absolute deviation, MAD) ส่วนเบี่ยงเบนกำลังสองเฉลี่ย (mean square deviation, MSD) พบว่าเมื่อข้อมูลมีการแจกแจงสมมาตร เช่น การแจกแจงปกติ สำหรับตัวอย่างมีขนาดใหญ่ ($n > 50$) ควรใช้แนวโน้มเชิงเส้นหรือการประมาณค่าในช่วงเชิงเส้น แต่ถ้าข้อมูลมีการแจกแจงที่ สำหรับ $n < 500$ ควรใช้ค่าเฉลี่ยอนุกรมหรือค่าเฉลี่ยของค่าใกล้เคียง และสำหรับ $n > 500$ ควรใช้การประมาณค่าในช่วงเชิงเส้น เมื่อข้อมูลมีการแจกแจงไม่สมมาตร เช่น การแจกแจงล็อกปกติ สำหรับ $n < 500$ ควรใช้ค่าเฉลี่ยอนุกรมหรือค่าเฉลี่ยของค่าใกล้เคียง และสำหรับ $n > 500$ ควรใช้การประมาณค่าในช่วงเชิงเส้นหรือค่ามัธยฐานของค่าใกล้เคียง (Mahmoud, 2010) และการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหาย 5 วิธี คือ วิธีค่าเฉลี่ยอนุกรม วิธีค่าเฉลี่ยของค่าใกล้เคียง วิธีค่ามัธยฐานของค่าใกล้เคียง วิธีการประมาณค่าในช่วงเชิงเส้นและวิธีแนวโน้มเชิงเส้น โดยวิธีการ

จำแนก 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีนว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความแม่นยำและค่าความเที่ยงของการทำนายที่สูงกว่า ค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำกว่า การเปรียบเทียบข้อมูลโรคตับของรัฐบาลนครประเทศ ประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียมโดยแทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง ให้ค่าความเที่ยงสูงสุดและค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ส่วนวิธีต้นไม้ตัดสินใจโดยแทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียงก็เป็นวิธีที่มีประสิทธิภาพสูง เนื่องจากให้ค่าความแม่นยำสูงสุด ข้อมูลรายได้และรายจ่ายของครอบครัวฟิลิปปินส์ วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียมโดยแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าในช่วงเชิงเส้นให้ค่าความแม่นยำและค่าความเที่ยงสูงสุดและให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ส่วนข้อมูลการตลาดของธนาคาร วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียมโดยแทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยอนุกรม ให้ค่าความแม่นยำและค่าความเที่ยงสูงสุด และให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (จิตกานต์ และคณะ, 2561)

งานวิจัยนี้ศึกษาวิธีการแทนค่าสูญหายของข้อมูลด้วยวิธีต่าง ๆ 5 วิธีคือค่าเฉลี่ยอนุกรม ค่าเฉลี่ยของค่าใกล้เคียง ค่ามัธยฐานของค่าใกล้เคียง การประมาณค่าในช่วงเชิงเส้น และแนวโน้มเชิงเส้น แล้วนำข้อมูลที่แทนค่าสูญหายทั้ง 5 วิธี ข้างต้นมาจำแนกกลุ่ม (classification) ด้วยวิธีต่าง ๆ 4 วิธี คือ วิธีโครงข่ายประสาทเทียม (artificial neural network) เนื่องจากมีประสิทธิภาพสูงสุด ส่วนวิธีต้นไม้ตัดสินใจ (decision tree) ก็เป็นวิธีที่มีประสิทธิภาพสูงตามงานวิจัยของ จิตกานต์ และคณะ (2561) ส่วนวิธีนาอิวเบส (Naïve Bayes) และวิธีการ

ถดถอยลอจิสติกทวิภาค (binary logistic regression) ยังไม่มีการศึกษามากนัก ผู้วิจัยจึงสนใจเลือก 2 วิธีนี้มาศึกษาด้วย เนื่องจากผู้วิจัยได้ทดลองกับข้อมูลบางชุดพบว่าประสิทธิภาพสูง เพื่อต้องการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มข้อมูลทั้ง 4 วิธี ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูล โดยใช้วิธีการสุ่มตัวอย่างด้วยโปรแกรม WEKA เพื่อเปรียบเทียบว่าวิธีการจำแนกกลุ่มใดให้ค่าความแม่นยำที่มีค่าสูงที่สุด ส่วนค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (mean absolute error, MAE) ที่มีค่าต่ำที่สุด

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยครั้งนี้ คือ SPSS เวอร์ชัน 25 และโปรแกรม WEKA เวอร์ชัน 3.9.2

2.2 การเก็บรวบรวมข้อมูล การแทนค่าข้อมูลสูญหาย การแบ่งข้อมูล การศึกษาขั้นตอนวิธี และการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม

2.2.1 การเก็บรวบรวมข้อมูล ศึกษาข้อมูลสูญหาย จากเว็บไซต์ UCI จำนวน 1 ชุด และ Kaggle จำนวน 2 ชุด คือ

(1) โรคหัวใจ (heart disease) มีจำนวนข้อมูลทั้งหมด 567 ค่า พบค่าสูญหายจำนวน 21 ค่า คิดเป็นร้อยละ 3.70 เป็นชุดข้อมูลที่มีลักษณะค่าสูญหายต่ำ (Ramana, 2012) ตัวแปรอิสระประกอบด้วยอายุ (หน่วย : ปี) เพศของผู้ป่วย (male คือ เพศชาย, female คือ เพศหญิง) อาการเจ็บหน้าอก (0 คือ ไม่เจ็บ, 1 คือ เจ็บน้อย, 2 คือ เจ็บมาก, 3 คือ เจ็บปวดขั้นรุนแรง) ความดันโลหิต (หน่วย : มิลลิเมตรปรอท) ปริมาณคอเลสเตอรอล (หน่วย : มิลลิกรัมต่อเดซิลิตร)

ปริมาณน้ำตาลในเลือด (0 คือ น้อยกว่า 120 มิลลิกรัมต่อเดซิลิตร, 1 คือ มากกว่า 10 มิลลิกรัมต่อเดซิลิตร) การวัดคลื่นการเต้นของหัวใจ (0 คือ ปกติ, 1 คือ ไม่ปกติ) อัตราการเต้นของหัวใจสูงสุด (หน่วย : ครั้งต่อวินาที) โรคหลอดเลือดหัวใจตีบ (0 คือ ไม่เป็นโรคหลอดเลือดหัวใจตีบ, 1 คือ เป็นหลอดเลือดหัวใจตีบ) ส่วนตัวแปรตามที่ใช้ในการจำแนกกลุ่มแทนความเสี่ยงในการเป็นโรคหัวใจ (Yes คือ เป็นโรคหัวใจ, No คือ ไม่เป็นโรคหัวใจ)

(2) ประสิทธิภาพในการสอบของนักเรียน (student performance in exam) มีจำนวนข้อมูลทั้งหมด 889 ค่า พบค่าสูญหายจำนวน 41 ค่า คิดเป็นร้อยละ 4.61 เป็นชุดข้อมูลที่มีลักษณะค่าสูญหายปานกลาง (The Philippine Statistics Authority, 2013) ตัวแปรอิสระประกอบด้วยเพศของนักเรียน (male คือ เพศชาย, female คือ เพศหญิง) สถานะการศึกษา (1 คือ กำลังศึกษาในระดับมัธยมศึกษาตอนปลาย, 2 คือ จบการศึกษาในระดับมัธยมศึกษาตอนปลาย, 3 คือ กำลังศึกษาในระดับปริญญาตรี, 4 คือ จบการศึกษาในระดับปริญญาตรี, 5 คือ จบการศึกษาในระดับปริญญาโท, 6 คือ จบการศึกษาในระดับปริญญาเอก) การพักกลางวัน (0 คือ พักในเวลาปกติ, 1 คือ พักตามความสะดวก) คะแนนการสอบด้านการคำนวณ (หน่วย : คะแนน) คะแนนการสอบด้านการอ่าน (หน่วย : คะแนน) คะแนนการสอบด้านการเขียน (หน่วย : คะแนน) ส่วนตัวแปรตามที่ใช้ในการจำแนกกลุ่มแทนความพร้อมในการสอบ (Yes คือ มีความพร้อมในการสอบ, No คือ ไม่มีความพร้อมในการสอบ)

(3) การซื้อของในวันศุกร์สีดำที่สุดท้ายของเดือนพฤศจิกายน (Black Friday) มีจำนวนข้อมูลทั้งหมด 1071 ค่า พบค่าสูญหายจำนวน 93 ค่า คิดเป็นร้อยละ 8.49 เป็นชุดข้อมูลที่มีลักษณะค่าสูญหายสูง (Portuguese Banking

Institution, 2012) ตัวแปรอิสระประกอบด้วยรหัสสินค้า เพศของผู้ซื้อสินค้า (male คือ เพศชาย, female คือ เพศหญิง) ช่วงอายุ (0 คือ 0-17, 1 คือ 18-25, 2 คือ 26-35, 3 คือ 36-45, 4 คือ 46-50, 5 คือ 51-55, 6 คือ 56 เป็นต้นไป) ช่วงจำนวนครั้งในการเดินทางเข้าเมืองของลูกค้าต่อปี (0 คือ ไม่เคย, 1 คือ 1-3 ครั้ง, 2 คือ มากกว่า 4 ครั้ง) ปริมาณเงินที่ใช้จ่าย (หน่วย : ดอลลาร์) ส่วนตัวแปรตามที่ใช้ในการจำแนกกลุ่มแทนลูกค้าที่จดทะเบียนสมรส (Yes คือ จดทะเบียนสมรส, No คือ ไม่ได้จดทะเบียนสมรส)

2.2.2 การแทนค่าข้อมูลสูญหาย

(1) ค่าเฉลี่ยอนุกรม คือ การแทนที่ค่าสูญหายด้วยค่าเฉลี่ยทั้งชุดข้อมูล (ฐนัฐ, 2559)

$$\hat{X} = \frac{\sum_{i=1}^k X_i}{k}$$

โดยที่ X_i คือข้อมูลที่ไม่สูญหาย ; k คือ จำนวนข้อมูลที่ไม่สูญหาย; $\hat{X}$ คือ ค่าประมาณข้อมูลที่สูญหายที่ถูกประมาณด้วยค่าเฉลี่ยของข้อมูลที่ไม่สูญหาย

(2) ค่าเฉลี่ยของค่าใกล้เคียง คือ การแทนที่ค่าสูญหายด้วยค่าเฉลี่ยของค่าที่อยู่ใกล้เคียงช่วงของจุดใกล้เคียงคือค่าที่อยู่ด้านบนและด้านล่างของค่าที่สูญหายไปจะใช้ในการคำนวณค่าเฉลี่ย (ฐนัฐ, 2559)

$$\hat{X}_p = \frac{L+U}{2}$$

โดยที่ $\hat{X}_p$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p ด้วยค่าเฉลี่ยรอบจุด; L คือ ค่าเฉลี่ยของจุดจำนวน a จุดที่มีตำแหน่งต่ำกว่าจุด p ; U คือ ค่าเฉลี่ยของจุดจำนวน a จุดที่มีตำแหน่งสูงกว่าจุด p

ข้อจำกัดของโปรแกรม SPSS เวอร์ชัน 25 คือ ต้องมีข้อมูลที่ไม่สูญหายอยู่ต่ำกว่าจุด p เป็นจำนวน a จุด และสูงกว่าจุด p เป็นจำนวน a จุด

(3) ค่ามัธยฐานของค่าใกล้เคียง คือ การแทนที่ค่าสูญหายด้วยค่ามัธยฐานของค่าที่อยู่ใกล้เคียง ช่วงของจุดใกล้เคียงคือค่าที่อยู่ด้านบนและด้านล่างของค่าที่ขาดหายไปจะใช้ในการคำนวณค่ามัธยฐาน (ฐนัฐ, 2559)

$$\hat{X}_p = \frac{T_a + T_{a+1}}{2}$$

โดยที่ $\hat{X}_p$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p ด้วยค่ามัธยฐานรอบจุด; T_a คือ ค่าที่อยู่ก่อนตำแหน่งของมัธยฐาน; T_{a+1} คือ ค่าที่อยู่หลังตำแหน่งของมัธยฐาน

(4) การประมาณค่าในช่วงเชิงเส้น คือ การแทนที่ค่าสูญหายด้วยค่าในช่วงเชิงเส้นโดยแทนที่ข้อมูลสูญหายด้วยค่าในช่วงเชิงเส้น หากข้อมูลชุดแรกหรือชุดสุดท้ายมีค่าที่หายไป ค่าที่หายไปจะไม่ถูกแทนที่ (ฐนัฐ, 2559) แบ่งเป็น 2 กรณี คือ

(4.1) ค่าสูญหายไม่อยู่ติดกัน

$$\hat{X}_p = \frac{X_{p-1} + X_{p+1}}{2}$$

โดยที่ $\hat{X}_p$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p ด้วยค่าในช่วงเชิงเส้น; X_{p-1} คือ ข้อมูลที่อยู่ก่อนข้อมูลสูญหายที่ตำแหน่ง p; X_{p+1} คือ ข้อมูลที่อยู่หลังข้อมูลสูญหายที่ตำแหน่ง p

(4.2) ค่าสูญหายอยู่ติดกัน

$$\hat{X}_{p+r-1} = \frac{X_{p-1} + r X_{p+r}}{r+1}$$

โดยที่ $\hat{X}_{p+r-1}$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p+r-1 ด้วยค่าในช่วงเชิงเส้น; r คือ จำนวนข้อมูลสูญหายที่อยู่ติดกัน; X_{p-1} คือ ข้อมูลที่อยู่ก่อนข้อมูลสูญหายตัวแรก; X_{p+r} คือ ข้อมูลที่มีตำแหน่งอยู่หลังข้อมูลสูญหายตัวสุดท้าย

(5) แนวโน้มเชิงเส้น คือ การแทนที่ข้อมูลสูญหายด้วยสมการประมาณค่าที่เป็นสมการเส้นตรงที่เกิดจากข้อมูลที่ไม่สูญหาย (สายชล, 2560) สมการประมาณค่า คือ

$$\hat{Y}_{X^*} = a + bX^*$$

โดยที่ X^* คือ ตำแหน่งของข้อมูลที่สูญหาย; $\hat{Y}_{X^*}$ คือ ค่าประมาณของข้อมูลที่สูญหายตำแหน่งที่ X^* ; a คือ ค่ากำหนดตำแหน่งของเส้นตรง มีค่าเป็นค่าคงที่; b คือ ความชันของสมการเส้นตรง มีค่าเป็นค่าคงที่

2.2.3 การแบ่งข้อมูล

แบ่งชุดข้อมูลโดยโปรแกรม WEKA เวอร์ชัน 3.9.2 สุ่มจำนวน 5 รอบ โดยการกำหนดตัวสร้างเลขสุ่มเทียม (random seed) เป็น 10, 20, 30, 40 และ 50 ในอัตราส่วน 70:20:10 ส่วนที่ 1 ข้อมูลเรียนรู้ (training data) นำไปสร้างตัวแบบ (model) ร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง (validation data) นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (testing data) นำไปทดสอบตัวแบบ ร้อยละ 10 (Shams, 2014)

2.2.4 การศึกษาขั้นตอนวิธี

(1) วิธีต้นไม้ตัดสินใจ เป็นวิธีการเรียนรู้ของเครื่องที่นิยมใช้มากที่สุดแบบหนึ่ง (Mitchell, 1997) โดยการจำแนก (classification) ข้อมูลออกเป็นกลุ่ม (class) ต่าง ๆ โดยใช้คุณลักษณะ (attribute) ของข้อมูลในการจำแนกว่าคุณลักษณะใดของข้อมูลที่เป็นตัวกำหนดการจำแนกและคุณลักษณะแต่ละตัวของข้อมูลมีการวัดความสำคัญอย่างไร ต้นไม้ตัดสินใจประกอบด้วยโหนดภายใน (internal node) กิ่ง (branch, link) และโหนดใบ (leaf node) ใช้ขั้นตอนวิธี J48 (C4.5) (Kijisirikul, 2004)

(2) วิธีโครงข่ายประสาทเทียม ใช้ขั้นตอนวิธีชนิดเพอร์เซปตรอนหลายชั้น โดยกำหนดค่าอัตราการเรียนรู้เป็น 0.1 ค่าโมเมนตัมเป็น 0.9 จำนวนรอบการสอน 20,000 รอบ การวิจัยครั้งนี้ใช้ขั้นตอนวิธีของวิธีโครงข่ายประสาทเทียมชนิดเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน 1 ชั้น การ

เชื่อมโยงกันระหว่างเซลล์ประสาทโดยทั่วไปนิยมใช้การเชื่อมโยงแบบแพร่ย้อนกลับ (back-propagation) ซึ่งเป็นขั้นตอนที่ใช้สอนโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น โดยแบบจำลองโครงข่ายประสาทเทียมมีการเชื่อมโยงกันเป็นโครงข่ายแบบเป็นชั้น ๆ โครงข่ายชนิดนี้มีการเชื่อมโยงกัน 3 ชั้น ประกอบด้วยชั้นข้อมูลเข้า (input layer) ถัดมาเป็นชั้นซ่อน (hidden layer) และชั้นสุดท้าย คือ ชั้นข้อมูลออก (output layer) (Berson and Smith, 1997) โดยส่วนประกอบที่ถูกบรรจุอยู่ในเซลล์ประสาทแต่ละตัวประกอบด้วย 2 ฟังก์ชันย่อย คือ ฟังก์ชันผลรวม (summation function) และฟังก์ชันกระตุ้น (activation function)

(2.1) ฟังก์ชันผลรวม ทำหน้าที่ในการคำนวณผลรวมของข้อมูลที่ได้จากชั้นข้อมูลเข้า ซึ่งคำนวณได้ดังสมการ (1) (Hagan *et al.*, 1996) โดยกำหนดให้ตัวแปร x คือ ค่าข้อมูลเข้าตัวที่ i ; ตัวแปร w คือ ค่าน้ำหนักถ่วงของข้อมูลเข้าตัวที่ i ; ตัวแปร g คือ ข้อมูลออกจากฟังก์ชันผลรวม; ตัวแปร z คือ จำนวนเซลล์ประสาทของข้อมูลเข้า; ตัวแปร β คือ ค่าความเอนเอียง (bias)

$$g = \sum_{i=1}^z x_i w_i + \beta \quad (1)$$

(2.2) ฟังก์ชันกระตุ้น ทำหน้าที่ปรับเปลี่ยนค่าของข้อมูลที่ได้จากฟังก์ชันผลรวมให้อยู่ในช่วงที่ต้องการ ฟังก์ชันกระตุ้นที่นิยม ได้แก่ ฟังก์ชันเชิงเส้น (linear function) ฟังก์ชันซิกมอยด์ (sigmoid function) และฟังก์ชันไฮเพอร์โบลิคแทนเจนต์ (hyperbolic tangent function) (ธนาวุฒิ, 2552)

(3) วิธีนออีฟเบส เป็นขั้นตอนวิธีที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (class) โดยมีคำตอบ 2 คำตอบ (สายชล, 2560)

(4) วิธีการถดถอยลอจิสติกทวิภาค (binary logistic regression method) เป็นการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มี 2 ค่า ส่วนตัวแปรอิสระอาจเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพหรืออาจมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ (กัลยา, 2552) จากนั้นนำข้อมูลที่แบ่งออกเป็น 3 ส่วน มาวิเคราะห์โดยใช้โปรแกรม WEKA ซึ่งวิเคราะห์จากวิธีการจำแนกกลุ่มทั้ง 4 วิธี ข้างต้น

2.2.5 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม

นำผลการวิเคราะห์ของแต่ละวิธีทั้ง 4 วิธี มาเปรียบเทียบประสิทธิภาพโดยพิจารณาจากเมทริกซ์ความสับสน (confusion matrix) ซึ่งเป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในตารางเมทริกซ์ความสับสนซึ่งจะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของขั้นตอนวิธีดังกล่าวที่ 1

(1) ค่าความแม่นยำในการทำนาย คือ การแสดงการวัดที่ได้มีความแม่นยำในรูปอัตราส่วนโดยคิดเป็นร้อยละ (สุรวัชร และสายชล, 2560)

Accuracy = (จำนวนข้อมูลที่จำแนกถูกว่าเป็นบวกและลบ ÷ จำนวนข้อมูลทั้งหมด) × 100 %

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

(2) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยความเอนเอียงและความแปรปรวน (พนิดา และคณะ, 2560)

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

โดยที่ y_i แทน ค่าจริง; $\hat{y}_i$ แทน ค่าทำนาย; n แทน จำนวนข้อมูลของกลุ่มตัวอย่าง

(3) ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MA) คือ ค่าวัดความถูกต้องของการทำนายที่วัดจากค่าความคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน MAE มีหน่วยวัดหน่วยเดียวกับค่าสังเกต (สายชล, 2560)

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

หลักเกณฑ์ในการพิจารณาการเปรียบเทียบประสิทธิภาพของการแทนที่ข้อมูลสูญหายกับ

ตารางที่ 1 เมทริกซ์ความสับสน

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

โดยที่ 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

3. ผลการวิจัย

3.1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่ม

งานวิจัยนี้ใช้การจำแนกกลุ่มด้วยวิธีการทำเหมืองข้อมูล โดยนำชุดข้อมูลที่ค้นคว้า 3 ชุด มาวิเคราะห์ข้อมูล ซึ่งสุมแบ่งข้อมูลเป็น 3 ส่วน คือ ส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบร้อยละ 10 และได้นำมาเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่ม โดยพิจารณาจากค่าความแม่นยำ ค่าคลาดเคลื่อนกำลัง

การจำแนกกลุ่มทั้ง 4 วิธี จะพิจารณาค่าความแม่นยำ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยไปพร้อม ๆ กัน โดยพิจารณาค่าความแม่นยำที่มีค่าสูงที่สุด ส่วนค่าคลาดเคลื่อนกำลังสองเฉลี่ยและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยที่มีค่าต่ำที่สุด โดยวิธีการจำแนกกลุ่มทั้ง 4 วิธี ที่มีค่าทั้ง 3 ค่าดังกล่าว ดีที่สุด 2 ใน 3 ค่า จะถือว่ามีประสิทธิภาพดีที่สุด

สองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ซึ่งวิธีที่ใช้ในการทดสอบครั้งนี้มี 4 วิธี คือ วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีนาอีฟเบส และวิธีการถดถอยลอจิสติกทวิภาค ผลการวิเคราะห์ข้อมูลจากตารางที่ 2 ซึ่งเป็นกรณีที่มีข้อมูลที่มีค่าสูญหายต่ำพบว่าวิธีต้นไม้ตัดสินใจ แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง ให้ค่าความถูกต้องสูงสุดคือ ร้อยละ 67.3684 และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.3691 ส่วนวิธีนาอีฟเบส แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2563 ส่วนตารางที่ 3 ซึ่งเป็นกรณีที่ข้อมูลที่มีค่าสูญหายปานกลาง พบว่าวิธีการถดถอยลอจิสติกทวิภาค แทนค่า

ข้อมูลสูญหายด้วยวิธีการประมาณค่าในช่วงเชิงเส้น ให้ค่าความแม่นยำสูงสุด คือ ร้อยละ 84.9438 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.1113 และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.1956 และจากตารางที่ 4 ซึ่งเป็นกรณีที่มีข้อมูลที่มีค่าสูญหายสูง พบว่าวิธีนาอ์ฟเบส แทนค่าข้อมูลสูญหายด้วยวิธีค่ามัธยฐานของค่าใกล้เคียงให้ค่าความแม่นยำสูงสุด คือ ร้อยละ 67.4074 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2163 ส่วนวิธีโครงข่ายประสาทเทียม แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียงให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.3854

4. สรุปผลการวิจัย

การวิจัย เรื่อง การแทนที่ข้อมูลสูญหาย 5 วิธี คือ วิธีค่าเฉลี่ยอนุกรม วิธีค่าเฉลี่ยของค่าใกล้เคียง วิธีค่ามัธยฐานของค่าใกล้เคียง วิธีการประมาณค่าในช่วงเชิงเส้น และวิธีแนวโน้มเชิงเส้น แล้วนำข้อมูลที่แทนค่าสูญหายทั้ง 5 วิธี ข้างต้นมาจำแนกกลุ่ม 4 วิธี คือ วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีนาอ์ฟเบส และวิธีการถดถอยลอจิสติกทวิภาค ว่าวิธีใดมีประสิทธิภาพในการจำแนกกลุ่มที่ดีที่สุด โดยพิจารณาจากค่าความแม่นยำ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ใช้ข้อมูลสูญหาย 3 ชุด โดยชุดข้อมูลที่มีค่าสูญหายต่ำคือ โรคหัวใจ วิธีการจำแนกที่มีประสิทธิภาพสูงสุดคือ วิธีต้นไม้ตัดสินใจ แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง ให้ค่าความถูกต้องสูงสุด

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มสำหรับข้อมูลโรคหัวใจ (ข้อมูลชุดที่ 1)

วิธีการแทนค่าข้อมูลสูญหาย	ค่าเฉลี่ยของค่าความแม่นยำ	ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกกลุ่มด้วยวิธีต้นไม้ตัดสินใจ			
วิธีค่าเฉลี่ยอนุกรม	58.2456	0.3489	0.4347
วิธีค่าเฉลี่ยของค่าใกล้เคียง	67.3684	0.2827	0.3691
วิธีค่ามัธยฐานของค่าใกล้เคียง	61.4035	0.3211	0.4038
วิธีการประมาณค่าในช่วงเชิงเส้น	54.3860	0.3914	0.4742
วิธีแนวโน้มเชิงเส้น	51.2281	0.4153	0.5019
การจำแนกกลุ่มด้วยวิธีโครงข่ายประสาทเทียม			
วิธีค่าเฉลี่ยอนุกรม	56.8421	0.3737	0.4433
วิธีค่าเฉลี่ยของค่าใกล้เคียง	54.3860	0.3960	0.4635
วิธีค่ามัธยฐานของค่าใกล้เคียง	55.7895	0.3732	0.4406
วิธีการประมาณค่าในช่วงเชิงเส้น	51.9298	0.4156	0.4804
วิธีแนวโน้มเชิงเส้น	50.0000	0.4221	0.4996
การจำแนกกลุ่มด้วยวิธีนาอ์ฟเบส			
วิธีค่าเฉลี่ยอนุกรม	56.1314	0.2800	0.4711
วิธีค่าเฉลี่ยของค่าใกล้เคียง	60.9926	0.2563	0.4413
วิธีค่ามัธยฐานของค่าใกล้เคียง	54.3860	0.2807	0.4702

วิธีการประมาณค่าในช่วงเชิงเส้น	55.0877	0.2780	0.4750
วิธีแนวโน้มเชิงเส้น	55.7895	0.2789	0.4674

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม สำหรับข้อมูลประสิทธิภาพในการสอบของนักเรียน (ข้อมูลชุดที่ 2)

วิธีการแทนค่าข้อมูลสูญหาย	ค่าเฉลี่ยของค่าความแม่นยำ	ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกกลุ่มด้วยวิธีต้นไม้ตัดสินใจ			
วิธีค่าเฉลี่ยอนุกรม	63.1461	0.2647	0.4207
วิธีค่าเฉลี่ยของค่าใกล้เคียง	69.2135	0.2429	0.3815
วิธีค่ามัธยฐานของค่าใกล้เคียง	66.5168	0.2622	0.3864
วิธีการประมาณค่าในช่วงเชิงเส้น	61.3483	0.2892	0.4287
วิธีแนวโน้มเชิงเส้น	56.1798	0.3090	0.4607
การจำแนกกลุ่มด้วยวิธีโครงข่ายประสาทเทียม			
วิธีค่าเฉลี่ยอนุกรม	69.6629	0.2251	0.3408
วิธีค่าเฉลี่ยของค่าใกล้เคียง	69.6629	0.2224	0.3459
วิธีค่ามัธยฐานของค่าใกล้เคียง	68.0899	0.2367	0.3530
วิธีการประมาณค่าในช่วงเชิงเส้น	67.8652	0.2463	0.3741
วิธีแนวโน้มเชิงเส้น	71.6854	0.2143	0.2589
การจำแนกกลุ่มด้วยวิธีนาอ์ฟเบส			
วิธีค่าเฉลี่ยอนุกรม	70.1123	0.2100	0.3809
วิธีค่าเฉลี่ยของค่าใกล้เคียง	64.9438	0.2256	0.4119
วิธีค่ามัธยฐานของค่าใกล้เคียง	69.8878	0.2007	0.3655
วิธีการประมาณค่าในช่วงเชิงเส้น	84.9438	0.1113	0.1956
วิธีแนวโน้มเชิงเส้น	59.7753	0.2467	0.4491

และให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด ชุดข้อมูลที่มีค่าสูญหายปานกลาง คือ ประสิทธิภาพในการสอบของนักเรียน วิธีการจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีการถดถอยลอจิสติก ทวิภาค แทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าในช่วงเชิงเส้น ให้ค่าความแม่นยำสูงสุด ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด และชุดข้อมูลที่มีค่าสูญหายสูง คือ การซื้อของในวันศุกร์สัปดาห์สุดท้ายของเดือนพฤศจิกายน วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีนาอ์ฟเบส แทนค่าข้อมูลสูญหายด้วย

วิธีค่ามัธยฐานของค่าใกล้เคียงให้ค่าความแม่นยำสูงสุดและค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด

5. ข้อเสนอแนะ

5.1 การแทนค่าข้อมูลสูญหายสามารถใช้วิธีการอื่น ๆ ได้แก่ ขั้นตอนวิธีการคาดหมาย-สูงสุด (expectation-maximization algorithm) วิธีที่น่าจะเป็นใช้ข้อมูลเต็มรูปแบบ (full information maximum-likelihood method) วิธีทดแทนพหุคูณ (multiple imputation method) เป็นต้น

5.2 การแทนค่าข้อมูลสูญหายสามารถใช้โปรแกรมอื่น ๆ ได้แก่ SAS, R Commander, Excel, Mplus เป็นต้น

5.3 การนำไปใช้ประโยชน์ เป็นแนวทางในการศึกษาวิธีการแทนค่าข้อมูลสูญหายกับการจำแนกกลุ่มได้

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม สำหรับข้อมูลการซื้อของในวันศุกร์ สัปดาห์สุดท้ายของเดือนพฤศจิกายน (ข้อมูลชุดที่ 3)

วิธีการแทนค่าข้อมูลสูญหาย	ค่าเฉลี่ยของค่าความแม่นยำ	ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกกลุ่มด้วยวิธีต้นไม้ตัดสินใจ			
วิธีค่าเฉลี่ยอนุกรม	56.1111	0.2542	0.4744
วิธีค่าเฉลี่ยของค่าใกล้เคียง	59.4444	0.2413	0.4531
วิธีค่ามัธยฐานของค่าใกล้เคียง	57.5926	0.2489	0.4631
วิธีการประมาณค่าในช่วงเชิงเส้น	59.2593	0.2445	0.4551
วิธีแนวโน้มเชิงเส้น	56.4815	0.2507	0.4731
การจำแนกกลุ่มด้วยวิธีโครงข่ายประสาทเทียม			
วิธีค่าเฉลี่ยอนุกรม	61.6667	0.2645	0.4073
วิธีค่าเฉลี่ยของค่าใกล้เคียง	65.1852	0.2543	0.3854
วิธีค่ามัธยฐานของค่าใกล้เคียง	63.7037	0.2628	0.4011
วิธีการประมาณค่าในช่วงเชิงเส้น	61.8519	0.2720	0.4126
วิธีแนวโน้มเชิงเส้น	60.7407	0.2694	0.4114
การจำแนกกลุ่มด้วยวิธีนาอ์ฟเบส			
วิธีค่าเฉลี่ยอนุกรม	63.7037	0.2289	0.4238
วิธีค่าเฉลี่ยของค่าใกล้เคียง	64.8148	0.2172	0.4137
วิธีค่ามัธยฐานของค่าใกล้เคียง	67.4074	0.2163	0.4169
วิธีการประมาณค่าในช่วงเชิงเส้น	66.1111	0.2179	0.4201
วิธีแนวโน้มเชิงเส้น	62.7778	0.2282	0.4280
การจำแนกกลุ่มด้วยวิธีการถดถอยลอจิสติกทวิภาค			
วิธีค่าเฉลี่ยอนุกรม	60.9259	0.3710	0.3944
วิธีค่าเฉลี่ยของค่าใกล้เคียง	61.2963	0.3654	0.3891
วิธีค่ามัธยฐานของค่าใกล้เคียง	62.0370	0.3689	0.3924
วิธีการประมาณค่าในช่วงเชิงเส้น	61.4815	0.3696	0.3912
วิธีแนวโน้มเชิงเส้น	60.9259	0.3724	0.3983

6. กิตติกรรมประกาศ

ขอขอบคุณ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่

สนับสนุนให้ทุนวิจัย เรื่อง การเปรียบเทียบประสิทธิภาพการแทนค่าข้อมูลสูญหายกับการจำแนกกลุ่ม 4 วิธี

7. รายการอ้างอิง

- กัลยา วานิชย์บัญชา, 2552, การวิเคราะห์ข้อมูลหลายตัวแปร, บริษัทธรรมสาร จำกัด, กรุงเทพฯ.
- จิตกานต์ จันทราช, มนทิราลัย ชัยมงคล, รัตนชัย แซ่โจ้ว, สายทิพย์ พลอยสัมฤทธิ์ และสายชล สินสมบูรณ์ทอง, 2563, การเปรียบเทียบประสิทธิภาพการดำเนินงานผลการจำแนกกรณีข้อมูลสูญหายด้วยเทคนิคการทำเหมืองข้อมูล, Thai J. Sci. Technol. 9(1): 1-15.
- ฐณัฐ วงศ์สายเชื้อ, 2559, Replace Missing Value – การแทนค่าสูญหายในโปรแกรม SPSS, แหล่งที่มา : https://www.youtube.com/watch?v=WzaeJ_HAqtk, 15 ธันวาคม 2561.
- ธนาวุฒิ ประกอบผล, 2552, โครงข่ายประสาทเทียม, ว.มฉก.วิชาการ 12(24): 73-87.
- พนิดา สมบัติมาก, ภัสสร จันท์หอม, ศุภกร รัศมี และโอฬาร รุ่งมณีธรรมคุณ, 2560, การเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล, ปัญหาพิเศษปริญญาตรี, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.
- วรฤทธิ์ พานิชกิจโกศลกุล, 2552, การจำลองแบบมอนติคาร์โลสำหรับประมาณค่าความแปรปรวนของการแจกแจงอินเวอร์เกาส์เซียนเมื่อข้อมูลมีค่าสูญหาย, ว.การวิจัยกาสะลองคำ 3(1): 14-23.
- วุฒิ สุขเจริญ, 2558, การดำเนินการกับข้อมูลขาดหาย, ว.ร่วมฤกษ์ มหาวิทยาลัยเกริก 33(2): 11-32.
- สายชล สินสมบูรณ์ทอง, 2560, การทำเหมืองข้อมูล เล่ม 1 : การค้นหาคำรู้จากข้อมูล, พิมพ์ครั้งที่ 2, จามจุรีโปรดักส์ จำกัด, กรุงเทพฯ.
- สายชล สินสมบูรณ์ทอง, 2560, สถิติเบื้องต้น, พิมพ์ครั้งที่ 11, จามจุรีโปรดักส์ จำกัด, กรุงเทพฯ.
- สุวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง, 2560, การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย, ว.วิทยาศาสตร์และเทคโนโลยี 25(5): 839-853.
- Berson, A. and Smith, S.J., 1997, Data Warehousing: Data Mining and OLAP, McGraw-Hill, Inc., New York.
- Hagan, M., Demuth, H. and Beale, M., 1996, Neural Network Design, Martin T. Hagan, Oklahoma.
- Kaiser, J., 2014, Dealing with missing values in data, J. Syst. Integrat. 5: 42-51.
- Kijsirikul, B., 2004, Data Mining Algorithms, The Final Report on the Joint Government and Private Sectors, Chulalongkorn University, Bangkok.
- Mahmoud, M.M., 2010, Evaluation of Accuracy of the Estimation Methods for Replacing Missing Values for Time Series Variables Using the Statistical Packages Software SPSS and MINITAB, Department of Statistics and Mathematics, Faculty of Commerce., Tanta University, Al Gharbiyah.
- Mitchell, T.M., 1997, Machine Learning, McGraw-Hill, Inc., New York.
- Portuguese Banking Institution, 2012, Black Friday, Available Source: https://www.kaggle.com/mehdidag/black-friday?fbclid=IwAR0GIK0EJwAyDzb0iD06X7U7Ev-65nfgM_

- v04ZL7li5sXBEEUiR_CtgVg, February 10, 2019.
- Rahman, M. M. and Davis, D. N. , 2013, Addressing the class imbalance problem in medical datasets, *Int. J. Mach. Learn. Comput.* 3: 224-228.
- Ramana, B. V. , 2012, Liver Disease of Andhra Pradesh India Data Set, Available Source: [https://archive.ics.uci.edu/ml/datasets/ILPD+\(Indian+Liver+Patient+Dataset\)](https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)), February 10, 2019.
- Shams, R. , 2014, Creating Training, Validation and Test Sets (Data Preprocessing) , Available Source: <https://www.youtube.com/watch?v=uiDFa7iY9yo>, January 22, 2019.
- The Philippine Statistics Authority (PSA), 2013, Students Performance in Exams, Available Source: https://www.kaggle.com/spscientist/students-performance-in-exams?fbclid=IwAR1Cr-8mrbJzAswNqUn2yRd_nM_30XCkGD9zKZI74xMxL_vesBryqfbwQ_U, February 10, 2019.