การเปรียบเทียบประสิทธิภาพการทำนายผลการจำแนกกรณีข้อมูลสูญหายด้วยเทคนิคการทำเหมืองข้อมูล

Efficiency Comparison in Classification of Data Mining Techniques with Missing Data

จิตกานต์ จันทราช, มนทิราลัย ชัยมงคล*, รัตนชัย แซ่โง้ว,
สายทิพย์ พลอยสัมฤทธิ์ และสายชล สินสมบูรณ์ทอง
ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520
Jittakan Jantarach, Montiralai Chaimongkhol*, Rattanachai Saengow,
Saithip Ploysamrit and Saichon Sinsomboonthong
Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang,
Chalongkrung Road, Ladkrabang, Bangkok 10520

Received: May 10, 2019; Accepted: June 28, 2019

บทคัดย่อ

งานวิจัยนี้มีนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนก 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยการค้นคว้าและศึกษาการแทนค่าข้อมูลสูญหายจากข้อมูล 3 ชุด คือ โรคตับของรัฐอานธรประเทศ ประเทศอินเดีย รายได้และรายจ่ายของครอบครัวฟิลิปปินส์ และการตลาดของธนาคาร ในกรณีที่มีการแทนค่าข้อมูลสูญหาย 5 วิธี คือ วิธีค่าเฉลี่ย วิธีค่าเฉลี่ยของค่าใกล้เคียง วิธีค่ามัธยฐานของค่าใกล้เคียง วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น และวิธีแนวโน้มเชิงเส้น โดยใช้โปรแกรม SPSS ในการแทนค่าข้อมูลสูญหายว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความถูกต้องและค่าความแม่นยำของการทำนายที่สูงกว่า ค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำกว่า โดยแบ่งข้อมูลเป็นชุดข้อมูลเรียนรู้ ชุดข้อมูลตรวจสอบความถูกต้อง และชุดข้อมูลทดสอบ ในอัตราส่วน 70, 20 และ 10 ตามลำดับ การเปรียบเทียบข้อมูลโรคตับของรัฐอานธรประเทศ ประเทศอินเดีย มีข้อมูลสูญหายร้อยละ 1.89 เป็นชุดข้อมูลที่มีค่าสูญหายต่ำ วิธีจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม และประมาณค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง เนื่องจากให้ค่าความแม่นยำเฉลี่ยสูงสุดและค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ข้อมูลรายได้และรายจ่ายของครอบครัวฟิลิปปินส์ มีข้อมูลสูญหายร้อยละ 4.21 เป็นชุดข้อมูลที่มีค่าสูญหายปานกลาง วิธีจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม และประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น และข้อมูลการตลาดของธนาคาร มีข้อมูลสูญหายร้อยละ 9.72 เป็นชุดข้อมูลที่มีค่าสูญ

**Corresponding author: montiralai.ple@gmail.com*
DOI: 10.14456/tjst.2020.2

หายสูง วิธีจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม และประมาณค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย

คำสำคัญ : ข้อมูลสูญหาย; วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีต้นไม้ตัดสินใจ; วิธีโครงข่ายประสาทเทียม; วิธีซัพพอร์ตเวกเตอร์แมชชีน

Abstract

The objective of this research was to compare the efficiencies of four classification methods: K-nearest neighbor, decision tree, artificial neural network and support vector machine, on three datasets with some missing data. The tested datasets, i.e. a dataset of incidents of liver disease in Andhra Pradesh, India, a dataset of annual incomes and expenditures of Filipino families, and a dataset of issued and non-issued credit cards by a bank data points were constructed to replace the missing data by five replacement methods: series mean, mean of nearby points, median of nearby points, linear interpolation and linear trend at a point, offered in SPSS software program. The metrics that indicated the efficiency of a classification method were the prediction accuracy and the mean squared error of classification. Each dataset was divided into three subsets: a learning set, a validation set and a test set, at a ratio of 70 : 20 : 10. For the classification of the dataset of incidents of liver disease in Andhra Pradesh, it had missing data 1.89 percentages and had the least amount of missing data. The most accurate outcomes were from the highest mean of precision for the outcomes and the lowest mean of mean squared error were from the artificial neural network method with missing data replaced by the mean of nearby points method. For the classification of the dataset of annual incomes and expenses of Filipino families, it had missing data 4.21 percentages and had a moderate amount of missing data. The most accurate outcomes were from the artificial neural network method with missing data replaced by the linear interpolation method. For the classification of the dataset of issued and non-issued credit cards by a bank, it had missing data 9.72 percentages and had the highest amount of missing data. The most accurate outcomes were from the artificial neural network method with missing data replaced by the series mean method.

Keywords: missing data; K-nearest neighbor; decision tree; artificial neural network; support vector machine

1. คำนำ

ปัจจุบันการทำงานวิจัยหรือการสำรวจสิ่งที่สนใจต่าง ๆ มักเก็บรวบรวมข้อมูล แล้วนำมาวิเคราะห์และประมวลผลให้ได้ข้อสรุป เพื่อนำมาตอบข้อสงสัยหรือแก้ไขปัญหางานวิจัยหรือการสำรวจนั้น ๆ การเก็บรวบรวมข้อมูลจากการสำรวจมักเกิดปัญหาข้อมูลสูญหาย (missing data) ในลักษณะการไม่ตอบสำหรับหน่วยตัวอย่างบางหน่วยตัวอย่าง (unit nonresponse) หรือการสูญหายของข้อมูลที่เกิดจากการไม่ตอบเฉพาะบางคำถาม (item

nonresponse) หากนำข้อมูลที่มีข้อมูลสูญหายนั้นมาวิเคราะห์จะทำให้ผลการวิเคราะห์ข้อมูลเกิดความคลาดเคลื่อนไปจากความเป็นจริง การกระจายของข้อมูลและค่าเฉลี่ยของข้อมูลไม่ดีส่งผลให้ไม่เป็นไปตามข้อกำหนดเบื้องต้น (assumption) และทำให้ไม่สามารถนำข้อมูลไปใช้ประโยชน์อย่างสูงสุด โดยทั่วไปมักแก้ไขปัญหานี้ด้วยการตัดข้อมูลสูญหายออกจากการวิเคราะห์ ซึ่งมีผลทำให้จำนวนข้อมูลลดน้อยลง แต่ในบางครั้งข้อมูลอาจมีจำนวนไม่มากนัก จึงไม่สามารถตัดข้อมูลออก ทำให้ต้องมีการประมาณค่าข้อมูลสูญหาย เราจึงมีการแทนค่าข้อมูลสูญหายและจำแนกข้อมูล เพื่อให้ได้วิธีที่เหมาะสมกับข้อมูลที่มีความแตกแต่งกันไป ดังนั้นผู้วิจัยจึงเลือกวิธีการแทนค่าข้อมูลสูญหายและจำแนกให้เหมาะสมกับข้อมูล (วราฤทธิ์, 2552) การแทนค่าสูญหายในโปรแกรม SPSS ในงานวิจัยนี้ประกอบด้วย 5 วิธี ได้แก่ วิธีค่าเฉลี่ย (series mean) วิธีค่าเฉลี่ยของค่าใกล้เคียง (mean of nearby points) วิธีมัธยฐานของค่าใกล้เคียง (median of nearby points) วิธีประมาณค่าระหว่างจุดแบบเชิงเส้น (linear interpolation) และวิธีแนวโน้มเชิงเส้น (linear trend at point) ซึ่งแต่ละวิธีนั้นมีวิธีการคำนวณและรูปแบบการใช้งานที่ต่างกันตามลักษณะของข้อมูล โดยวิธีค่าเฉลี่ยเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีค่าใกล้เคียงกันทั้งชุด วิธีค่าเฉลี่ยของค่าใกล้เคียงเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีค่าใกล้เคียงกันระหว่างค่าสูญหาย วิธีมัธยฐานของค่าใกล้เคียงเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีที่มีค่าสูงหรือต่ำต่างกันมากที่อยู่ระหว่างค่าสูญหาย วิธีประมาณค่าระหว่างจุดแบบเชิงเส้นเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีค่าใกล้เคียงกันระหว่างค่าสูญหายและข้อมูลสูญหายอยู่ติดกันมีความสัมพันธ์กับข้อมูลที่อยู่ระหว่างกลุ่มค่าสูญหายเป็นแบบถ่วงน้ำหนัก และวิธีแนวโน้มเชิงเส้นเหมาะสมกับการประมาณค่าข้อมูลสูญหายที่ข้อมูลมีความสัมพันธ์กันแบบเชิงเส้นระหว่างตัวแปรของค่าที่ไม่สูญหายกับตำแหน่งของค่าที่ไม่สูญหาย (ฐณัฐ, 2559)

วริษฐา และอนุภาพ (2556) เปรียบเทียบวิธีการประมาณข้อมูลสูญหายสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ เมื่อตัวแปรตามและตัวแปรอิสระมีข้อมูลสูญหาย โดยใช้วิธี K-nearest neighbor (KNN) วิธี expectation-maximization algorithm (EM) และวิธี predictive mean matching (PMM) ซึ่งข้อมูลที่ใช้ในการศึกษาได้จากการจำลองและมีสัดส่วนของการสูญหาย 3 ระดับ คือ 10, 20 และ 30 % โดยเปรียบเทียบประสิทธิภาพของแต่ละวิธีด้วยค่าเฉลี่ย ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (average mean square error, AMSE) ระหว่างค่าพยากรณ์กับค่าจริง พบว่าในกรณีสัดส่วนการสูญหาย 10 % การประมาณค่าสูญหายของข้อมูลด้วยวิธี EM จะให้ประสิทธิภาพดีกว่าวิธีอื่น และในกรณีสัดส่วนการสูญหาย 20 และ 30 % การประมาณค่าสูญหายของข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว จะให้ประสิทธิภาพดีที่สุด

ณัฐภัทร และคณะ (2555) เปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลเพื่อแทนค่าสูญหาย ได้แก่ นาอีฟเบส์ (Naive Bayes) เพื่อนบ้านใกล้สุด k ตัว (k-nearet neighbor) การถดถอยเชิงเส้น (linear regression) ต้นไม้ตัดสินใจ (decision tree) และฐานกฎ (rule based) โดยนำมาใช้กับชุดข้อมูลจากฐานข้อมูล UCI ที่มีความแตกต่างกันอย่างชัดเจน ได้แก่ ข้อมูลการจำแนกประเภทเห็ดเป็นข้อมูลชนิดไม่ต่อเนื่องทุกตัว ข้อมูลการจำแนกประเภทแก้ว เป็นข้อมูลชนิดต่อเนื่อง และข้อมูลมาตรวัดความสมดุล เป็นข้อมูลชนิดไม่ต่อเนื่องแบบเรียงลำดับ โดยใช้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (mean absolute error, MAE) และรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (root

mean squre error, RMSE) ผลที่ได้จากการวิจัยพบว่าข้อมูลชนิดไม่ต่อเนื่องจะแทนค่าสูญหายได้ดีด้วยวิธีต้นไม้ตัดสินใจ แต่เทคนิคนี้จะคลาดเคลื่อนน้อยก็ต่อเมื่อมีกฎจำนวนมากในการแทนค่าสูญหาย ข้อมูลชนิดตัวเลขต่อเนื่องจะแทนค่าสูญหายได้ดีด้วยเพื่อนบ้านใกล้สุด k ตัว และข้อมูลชนิดไม่ต่อเนื่องแบบเรียงลำดับจะแทนค่าสูญหายได้ดีด้วยวิธีนาอีฟเบส์

Kaiser (2014) ศึกษาเกี่ยวกับการจัดการค่าข้อมูลสูญหายในชุดข้อมูลจากงานวิจัยที่มีจำนวนข้อมูลมากที่มีข้อมูลสูญสูหาย มีสาเหตุความผิดพลาดมาจากการป้อนข้อมูลไม่ถูกต้อง และการวัดค่าข้อมูลของอุปกรณ์ไม่ถูกต้อง อันเป็นเหตุให้สูญเสียประสิทธิภาพในการวิเคราะห์ข้อมูล และมีความแตกต่างระหว่างข้อมูลที่มีค่าสูญหายกับข้อมูลที่สมบูรณ์มากจนเกินไป ด้วยเหตุนี้ผู้วิจัยจึงประมาณค่าข้อมูลสูญหายด้วยวิธีค่ากลาง (common value) มีทั้งหมด 2 วิธี ได้แก่ วิธีค่าเฉลี่ย (mean value) และวิธีค่ามัธยฐาน (median value) และวิธีค่าใกล้เคียงที่สุด (closest value) แล้วนำชุดข้อมูลที่ได้ประมาณค่าข้อมูลสูญหายแล้วไปทำเหมืองข้อมูลเพื่อวัดประสิทธิภาพในการจำแนกว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด จาก 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว (K-nearest neighbor) วิธีโครงข่ายประสาทเทียม (neural networks) และกฎความสัมพันธ์ (association rule) โดยพิจารณาจากค่าความถูกต้อง (accuracy) ที่สูงกว่า การวิจัยพบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม และกฎความสัมพันธ์ ให้ค่าความถูกต้องสูงสุด เมื่อใช้วิธีค่าใกล้เคียงที่สุด ในการประมาณค่าขอมูลสูญหาย ในกรณีแถว (row) ของตัวแปรของข้อมูลมีค่าสูญหายหลายสดมภ์ (column) สามารถจัดการข้อมูลสูญหายด้วยวิธีการตัดแถวออก จะทำให้ค่าความถูกต้องเพิ่มขึ้น ซึ่งมีประสิทธิภาพมากกว่าการประมาณค่าข้อมูล

Blomberg และ Ruiz (2013) ศึกษาเกี่ยวกับการประเมินอิทธิพลของข้อมูลสูญหายในขั้นตอนวิธีการจำแนกเพื่อประยุกต์ใช้ในการทำเหมืองข้อมูล โดยใช้ชุดข้อมูลที่ตัวแปรอิสระเป็นแบบต่อเนื่องในฐานข้อมูล UCI จำนวน 10 ชุดข้อมูล และควมคุมระดับข้อมูลสูญหาย พบว่าประสิทธิภาพในการจำแนกลดลงหลังจากแทนค่าสูญหายในชุดข้อมูลที่ทดสอบ วิธีนาอีฟเบส์ได้รับอิทธิพลโดยข้อมูลสูญหายน้อยที่สุด รองลงมา คือ วิธีซัพพอร์ตเวกเตอร์แมชชีน ส่วนวิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้ขั้นตอนวิธี IBK ได้รับอิทธิพลมากที่สุดโดยพิจารณาจากค่าความถูกต้อง (accuracy)

งานวิจัยนี้ศึกษาวิธีการแทนค่าสูญหายของข้อมูลด้วยวิธีต่าง ๆ 5 วิธี คือ ค่าเฉลี่ย (series mean) ค่าเฉลี่ยของค่าใกล้เคียง (mean of nearby points) ค่ามัธยฐานของค่าใกล้เคียง (median of nearby points) การประมาณค่าระหว่างจุดแบบเชิงเส้น (linear interpolation) และแนวโน้มเชิงเส้น (linear trend at point) แล้วจำแนก (classification) ด้วย 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว (k-nearest neighbor) วิธีต้นไม้ตัดสินใจ (decision tree) วิธีโครงข่ายประสาทเทียม (artificial neural network) และวิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) เพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนกทั้ง 4 วิธี ในกรณีที่มีการแทนค่าข้อมูลสูญหาย 5 วิธี ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูล โดยใช้วิธีการสุ่มตัวอย่างด้วยโปรแกรม WEKA เพื่อเปรียบเทียบว่าวิธีการจำแนกใดที่ให้ค่าความถูกต้อง (accuracy) และค่าความแม่นยำ (precision) ที่สูงกว่า ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) ที่ต่ำกว่า

2. การแทนค่าข้อมูลสูญหาย

ข้อมูลสูญหาย หมายถึง ค่าสังเกตที่ต้องการทราบค่า แต่ไม่สามารถทราบค่า โดยที่ค่านั้นควรจะ

สามารถทราบค่า หากมีวิธีการที่ใช้ในการรวบรวมข้อมูลหรือวัดค่าที่มีประสิทธิภาพดีขึ้นหรือมีความเหมาะสมมากขึ้น ซึ่งเรียกว่าเป็นค่าสูญหาย โดยมีวิธีการแทนค่าสูญหาย ดังนี้

2.1 ค่าเฉลี่ย คือ การแทนที่ค่าสูญหายด้วยค่าเฉลี่ยทั้งชุดข้อมูล (ฐณัฐ, 2559)

$$\hat{x} = \frac{\sum_{i=1}^{k} x_i}{k}$$

โดยที่ x_i คือ ข้อมูลที่ไม่สูญหาย; k คือ จำนวนข้อมูลที่ไม่สูญหาย; $\hat{x}$ คือ ค่าประมาณข้อมูลที่สูญหายด้วยค่าเฉลี่ยของข้อมูลที่ไม่สูญหาย

2.2 ค่าเฉลี่ยของค่าใกล้เคียง คือ การแทนที่ค่าสูญหายด้วยค่าเฉลี่ยของค่าที่อยู่ใกล้เคียง ช่วงของจุดใกล้เคียง คือ ค่าที่อยู่ด้านบนและด้านล่างของค่าที่สูญหายไปจะใช้ในการคำนวณค่าเฉลี่ย (ฐณัฐ, 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 เวอร์ชั่น 25 จะต้องมีข้อมูลที่ไม่สูญหายอยู่ต่ำกว่าจุด p เป็นจำนวน a จุด และสูงกว่าจุด p เป็นจำนวน a จุด

2.3 ค่ามัธยฐานของค่าใกล้เคียง คือ การแทนที่ค่าสูญหายด้วยค่ามัธยฐานของค่าที่อยู่ใกล้เคียง ช่วงของจุดใกล้เคียง คือ ค่าที่อยู่ด้านบนและด้านล่างของค่าที่สูญหายไปจะใช้ในการคำนวณค่ามัธยฐาน (ฐณัฐ, 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

2.4 การประมาณค่าระหว่างจุดแบบเชิงเส้น คือ การแทนที่ค่าสูญหายด้วยค่าระหว่างจุดแบบเชิงเส้น โดยแทนที่ข้อมูลสูญหายด้วยค่าระหว่างจุดแบบเชิงเส้น ถ้าข้อมูลชุดแรกหรือชุดสุดท้ายมีค่าที่สูญหายไป ค่าที่สูญหายไปจะไม่ถูกแทนที่ (ฐณัฐ, 2559) แบ่งเป็น 2 กรณี คือ

2.4.1 ค่าสูญหายไม่อยู่ติดกัน

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

โดยที่ $\hat{x}_p$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p ด้วยค่าระหว่างจุดแบบเชิงเส้น; x_{p-1} คือ ข้อมูลที่ตำแหน่ง p-1; x_{p+1} คือ ข้อมูลที่ตำแหน่ง p+1

2.4.2 ค่าสูญหายอยู่ติดกัน

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

โดยที่ $\hat{x}_{p+r-1}$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p+r-1 ด้วยค่าระหว่างจุดแบบเชิงเส้น; r คือ จำนวนข้อมูลสูญหายที่อยู่ติดกัน; x_{p-1} คือ ข้อมูลที่อยู่ก่อนข้อมูลสูญหายตัวแรก; x_{p+r} คือ ข้อมูลที่มีตำแหน่งอยู่หลังข้อมูลสูญหายตัวสุดท้าย (หมายเหตุ : ตำแหน่ง p เป็นค่าสูญหายตัวแรกของกลุ่มข้อมูลสูญหายที่ติดกัน r ตัว) ข้อจำกัดของโปรแกรม SPSS คือ ข้อมูลที่สูญหายจะต้องไม่อยู่ในตำแหน่งแรกหรือตำแหน่งสุดท้ายของข้อมูล

2.5 แนวโน้มเชิงเส้น คือ การแทนที่ข้อมูลสูญหายด้วยสมการประมาณค่าที่เป็นสมการเส้นตรงที่เกิดจากข้อมูลที่ไม่สูญหาย (ฐณัฐ, 2559)

สมการประมาณค่า $\hat{y}_{x^*} = a + bX^*$

โดยที่ X^* คือ ตำแหน่งของข้อมูลที่สูญหาย; $\hat{y}_{x^*}$ คือ ค่าประมาณของข้อมูลที่สูญหาย ตำแหน่งที่ X^* ;

a คือ ค่ากำหนดตำแหน่งของเส้นตรง มีค่าเป็นค่าคงที่; b คือ ความชันของสมการเส้นตรง มีค่าเป็นค่าคงที่

3. วิธีการจำแนก

3.1 วิธีเพื่อนบ้านใกล้สุด k ตัว

วิธีเพื่อนบ้านใกล้สุด k ตัว เป็นการหาระยะห่างระหว่างแต่ละตัวแปรในข้อมูล ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลเชิงตัวเลขและสามารถใช้กับตัวแปรไม่ต่อเนื่อง สามารถจัดในลักษณะพิเศษเพิ่มขึ้น เช่น สี สามารถใช้วิธีเพื่อนบ้านใกล้สุด k ตัว วัดความแตกต่างระหว่างสีน้ำเงินกับสีเขียว หลังจากนั้นต้องมีวิธีในการรวมค่าระยะห่างของคุณลักษณะ (attribute) ทุกค่า เพื่อคำนวณระยะห่างระหว่างเงื่อนไขหรือกรณีต่าง ๆ จากนั้นเลือกชุดของเงื่อนไขที่ใช้จัดกลุ่ม (class) มาเป็นฐานสำหรับการจัดกลุ่มในเงื่อนไขใหม่ ๆ จะตัดสินใจได้ว่าขอบเขตของจุดข้างเคียงที่ควรเป็นนั้นควรมีขนาดใหญ่ และอาจมีการตัดสินใจได้ด้วยการนับจำนวนจุดข้างเคียง โดยขั้นตอนวิธีเพื่อนบ้านใกล้สุด k ตัว มีขั้นตอนโดยสรุปดังนี้ (นิรนาม, 2561)

3.1.1 กำหนดขนาดของ k (ควรกำหนดให้เป็นเลขคี่) เช่น k = 3 คือ จะพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่อยู่ใกล้กับจุดที่ต้องการทำนาย

3.1.2 คำนวณระยะห่างของข้อมูลที่ต้องการทำนายกับกลุ่มข้อมูลตัวอย่าง โดยใช้ระยะห่างยูคลิเดียน (Euclidean distance) ดังสมการ

$$D_{Euclidian}(X_i, Y_i) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

โดยที่ $D_{Euclidian}(X_i, Y_i)$ คือ ระยะห่างระหว่าง X_i กับตัวอย่าง Y_i

3.1.3 จัดเรียงลำดับของระยะห่างและเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการทำนายตามจำนวน k ที่กำหนดไว้

3.1.4 พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่มไหนที่ใกล้จุดที่ต้องการทำนายเป็นจำนวนมากที่สุด

3.1.5 กำหนดกลุ่มให้กับจุดที่ต้องการทำนาย กลุ่มที่ใกล้จุดที่ต้องการทำนายมากที่สุด

3.2 วิธีต้นไม้ตัดสินใจ

วิธีต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ของเครื่องที่นิยมใช้มากที่สุดแบบหนึ่ง โดยการจำแนกข้อมูลออกเป็นกลุ่มต่าง ๆ ด้วยคุณลักษณะของข้อมูลในการจำแนกว่าคุณลักษณะใดของข้อมูลที่เป็นตัวกำหนดการจำแนก และคุณลักษณะแต่ละตัวของข้อมูลมีการวัดความสำคัญอย่างไร ต้นไม้ตัดสินใจประกอบด้วยโหนดภายใน (internal node) กิ่ง (branch, link) และโหนดใบ (leaf node) ใช้ขั้นตอนวิธี J48 (C4.5) (Mitchell, 1997) ขั้นตอนการสร้างต้นไม้ตัดสินใจมีดังนี้ (รุจิรา, 2554)

3.2.1 ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึก (training set)

3.2.2 ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้วให้โหนดนั้นเป็นใบ และตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3.2.3 ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่จะต้องวัดผลกำไร (gain) ของแต่ละคุณลักษณะ เพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือกคุณลักษณะที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ดีที่สุด โดยคุณลักษณะที่มีผลกำไรมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

3.2.4 กิ่งของต้นไม้สร้างขึ้นจากค่าต่าง ๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะแบ่งออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

3.2.5 วนซ้ำเพื่อหาคุณลักษณะที่มีผลกำไรมากที่สุดสำหรับข้อมูลที่แบ่งแยกออกมาแต่ละกิ่ง เพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดตัดสินใจ

ต่อไป โดยที่คุณลักษณะที่เลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับต่อ ๆ ไป

3.2.6 วนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อย ๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งข้างบนนี้เป็นจริง

3.3 วิธีโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมสร้างขึ้นโดยการจำลองลักษณะการประมวลผลของสมองมนุษย์ด้วยแบบจำลองทางคณิตศาสตร์ (mathematics model) ประกอบด้วยส่วนของการประมวลผลที่เรียกว่านิวรอน (neuron) ซึ่งทุก ๆ นิวรอนสามารถมีข้อมูลเข้า (input data) ได้หลายค่า แต่ข้อมูลออก (output data) มีได้เพียงค่าเดียว และทุก ๆ ข้อมูลออกจะเชื่อมโยงไปยังข้อมูลเข้าของนิวรอนอื่น ๆ ภายในโครงข่าย สำหรับการเชื่อมโยงกันภายในระหว่างนิวรอน ทุก ๆ ข้อมูลเข้าจะมีค่าถ่วงน้ำหนักเป็นตัวกำหนดกำลังของการเชื่อมโยง ภายในนิวรอนจะมีฟังก์ชันกำหนดสัญญาณออกที่เรียกว่าฟังก์ชันถ่ายโอน (transfer function) (ภัททิรา และวิทยา, 2557) โดยใช้ขั้นตอนวิธีชนิดเพอร์เซปตรอนแบบหลายชั้น ซึ่งเป็นโครงข่ายประสาทเทียมแบบ MLP มีโครงสร้างเป็นแบบหลาย ๆ ชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบมีผู้สอน (supervise) และใช้ขั้นตอนการส่งค่าย้อนกลับ (back propagation) สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อย คือ การส่งผ่านไปข้างหน้า (forward pass) และการส่งผ่านย้อนกลับ (backward pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่ง จนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (error correction) คือ ผลต่างของผลตอบที่แท้จริง (actual response) กับผลตอบเป้าหมาย (target response) เกิดเป็นสัญญาณผิดพลาด (error signal) ซึ่งสัญญาณผิดพลาดนี้จะส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ และค่าน้ำหนักของการเชื่อมต่อจะปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย (นิรนาม, 2560) และกำหนดค่าอัตราการเรียนรู้เป็น 0.1 ค่าโมเมนตัมเป็น 0.9 (Berson and Smith, 2001) จำนวนรอบการสอน 20,000 รอบ และชั้นซ่อน 1 ชั้น

3.4 วิธีซัพพอร์ทเวกเตอร์แมชชีน

หลักการของวิธีการนี้ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลเป็น 2 ส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งขอบเขตข้อมูล 2 กลุ่ม ออกจากกัน โดยมีวัตถุประสงค์ที่จะพยายามลดความผิดพลาดจากการทำนาย (minimize error) พร้อมกับเพิ่มระยะแยกแยะให้มากที่สุด (maximize margin) ซึ่งต่างจากเทคนิคโดยทั่วไป เช่น โครงข่ายประสาทเทียม (artificial neural network, ANN) ที่มุ่งเพียงทำให้ความผิดพลาดจากการทำนายให้ต่ำที่สุดเพียงอย่างเดียว โดยจะใช้ฟังก์ชันแมพ (map function) ข้อมูลจากพื้นที่นำเข้า (input space) ไปยังพื้นที่คุณลักษณะ (feature space) และสร้างฟังก์ชันวัดความคล้ายที่เรียกว่าฟังก์ชันเคอร์เนล (Kernel function) บนพื้นที่คุณลักษณะ เหมาะใช้สำหรับข้อมูลที่มีลักษณะมิติของข้อมูลที่มีปริมาณมาก (พรพล และคณะ, 2553)

3.5 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

3.5.1 ค่าความถูกต้อง คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน (สุรวัชร และสายชล, 2560) (ตารางที่ 1)

Accuracy = จำนวนข้อมูลที่จำแนกถูกว่าเป็นคำตอบบวกและลบ ÷ จำนวนข้อมูลทั้งหมด

$$= \frac{TP + TN}{TP + TN + FP + FN}$$

ตารางที่ 1 เมทริกซ์ความสับสน (confusion matrix) เป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในรูปตารางเมทริกซ์ ซึ่งจะช่วยให้ง่ายต่อการทำนายค่าขั้นตอนวิธี

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

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

3.5.2 ค่าความแม่นยำ (precision) คือ การแสดงค่าเดิมเมื่อวัดซ้ำหลาย ๆ ครั้งในรูปอัตราส่วน (สายชล, 2560)

Precision = จำนวนข้อมูลที่จำแนกถูกว่าเป็นคำตอบบวก ÷ จำนวนข้อมูลที่ทำนายได้ในคำตอบบวก

$$= \frac{TP}{TP + FP}$$

3.5.3 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยทั้งความเอนเอียงและความแปรปรวน (สุรวัชร และสายชล, 2560)

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2$$

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

โดยที่ Y_i แทนค่าจริง; $\hat{Y}_i$ แทนค่าทำนาย; n แทนจำนวนข้อมูลของกลุ่มตัวอย่าง

4. วิธีการดำเนินการวิจัย

4.1 การเก็บรวบรวมข้อมูล

ค้นหาและศึกษาข้อมูลที่มีค่าสูญหายจากเว็บไซต์ UCI และ Kaggle ได้ข้อมูล 3 ชุด คือ

4.1.1 โรคตับของรัฐอานธรประเทศประเทศอินเดีย (liver disease of Andhra Pradesh, India) จำนวนข้อมูลทั้งหมด 583 ค่า พบค่าสูญหายจำนวน 11 ค่า คิดเป็นร้อยละ 1.89 เป็นชุดข้อมูลที่มีค่าสูญหายต่ำ (Ramana, 2012) ตัวแปรอิสระมีจำนวน 10 ตัว คือ อายุ (X_1) เพศ (X_2) บิลิรูบินทั้งหมด (X_3) บิลิรูบินชนิดละลายน้ำ (X_4) อัลคาไลน์ฟอสฟาเตส (X_5) อะลามีนอะมิโนทรานฟาเรส (X_6) อะสเปอเทสอะมิโนทรานฟาเรส (X_7) ปริมาณโปรตีนรวมในกระแสเลือด (X_8) อัลบูมินในเลือด (X_9) สัดส่วนอัลบูมินและโกลบูลิน (X_{10}) และตัวแปรตามคือ ความเสี่ยงการเป็นโรคตับ (Yes คือ เป็นโรคตับ และ No คือ ไม่เป็นโรคตับ)

4.1.2 รายได้และรายจ่ายของครอบครัวฟิลิปปินส์ (Filipino family income and expenditure) จำนวนข้อมูลทั้งหมด 736 ค่า พบค่าสูญหายจำนวน 31 ค่า คิดเป็นร้อยละ 4.21 เป็นชุดข้อมูลที่มีค่าสูญหายปานกลาง (The Philippine Statistics Authority, 2013) ตัวแปรอิสระมีจำนวน 25 ตัว คือ เพศของหัวหน้าครอบครัว (X_1) อายุของหัวหน้าครอบครัว (X_2) สถานภาพสมรสของหัวหน้าครอบครัว (X_3) ประเภทของครอบครัว (X_4) แหล่งรายได้หลัก (X_5) ครัวเรือนเกษตร (X_6) จำนวนสมาชิกในครอบครัวทั้งหมด (X_7) จำนวนสมาชิกที่มีอายุต่ำกว่า 5 ปี (X_8) จำนวนสมาชิกที่มีอายุระหว่าง 5-17 ปี (X_9) จำนวนลูกจ้าง (X_{10}) รายได้ครอบครัวรวม (X_{11}) ค่าใช้จ่ายอาหาร (X_{12}) ค่าใช้จ่ายของร้านอาหารและโรงแรม (X_{13}) ค่าเครื่องดื่ม

แอลกอฮอล์ (X_{14}) ค่ายาสูบ (X_{15}) ค่าเสื้อผ้าและค่าใช้จ่ายอื่น ๆ (X_{16}) ค่าใช้จ่ายที่อยู่อาศัยและค่าน้ำ (X_{17}) ค่าเช่าบ้าน (X_{18}) ค่ารักษาพยาบาล (X_{19}) ค่าขนส่ง (X_{20}) ค่าใช้จ่ายด้านการสื่อสาร (X_{21}) ค่าใช้จ่ายด้านการศึกษา (X_{22}) ค่าใช้จ่ายสินค้าและบริการเบ็ดเตล็ด (X_{23}) ค่าใช้จ่ายโอกาสพิเศษ (X_{24}) ค่าใช้จ่ายในการเพาะปลูกพืชและสวน (X_{25}) และตัวแปรตาม คือ เงินเก็บ (Yes คือ มีเงินเก็บ และ No คือ ไม่มีเงินเก็บ)

4.1.3 การตลาดของธนาคาร (issued and non-issued credit cards by a bank) จำนวนข้อมูลทั้งหมด 854 ค่า พบค่าสูญหายจำนวน 83 ค่า คิดเป็นร้อยละ 9.72 เป็นชุดข้อมูลที่มีค่าสูญหายสูง (Portuguese Banking Institution, 2012) ตัวแปรอิสระมีจำนวน 12 ตัว คือ อายุ (X_1) การทำงาน (X_2) สถานภาพสมรส (X_3) การศึกษา (X_4) มีเครดิตผิดนัด (X_5) มีสินเชื่อที่อยู่อาศัย (X_6) มีสินเชื่อส่วนบุคคล (X_7) การติดต่อสื่อสาร (X_8) เดือนที่ติดต่อล่าสุดของปี (X_9) ระยะเวลาการติดต่อล่าสุด (X_{10}) จำนวนผู้ติดต่อที่ดำเนินการในโครงการนี้พร้อมลูกค้ารายนี้ (X_{11}) ผลจากโครงการการตลาดก่อนหน้า (X_{12}) และตัวแปรตาม คือ ลูกค้าจะสมัครรับฝากเงินในระยะยาว (Yes คือ สมัคร และ No คือ ไม่สมัคร)

4.2 ขั้นตอนในการดำเนินงานวิจัย

4.2.1 ขั้นตอนการแทนค่าข้อมูลสูญหายด้วยโปรแกรม SPSS

นำชุดข้อมูลที่รวบรวมได้มาแทนค่าสูญหาย โดยใช้โปรแกรม SPSS เวอร์ชั่น 25

(1) เปิดไฟล์ชุดข้อมูลนามสกุล .CSV เข้าสู่โปรแกรม SPSS เวอร์ชั่น 25 โดยคลิก File ⟶ Open ⟶ Data ⟶ เลือกไฟล์ชุดข้อมูล .CSV

(2) เลือก Transform ⟶ Replace Missing Values ⟶ เลือกตัวแปรที่ต้องการแทนค่า

การเลือกช่อง Method มีให้เลือก 5 วิธี ได้แก่ ค่าเฉลี่ย (series mean) ค่าเฉลี่ยของค่าใกล้เคียง (mean of nearby points) ค่ามัธยฐานของค่าใกล้เคียง (median of nearby points) การประมาณค่าระหว่างจุดแบบเชิงเส้น (linear interop-lation) และแนวโน้มเชิงเส้น (linear trend at point)

4.2.2 วิธีการแบ่งข้อมูล

แบ่งชุดข้อมูลด้วยโปรแกรม WEKA เวอร์ชั่น 3.9.2 สุ่มจำนวน 5 รอบ โดยกำหนดตัวสร้างเลขสุ่มเทียมเป็น 10, 20, 30, 40 และ 50 ในอัตราส่วน 70 : 20 : 10 ส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบร้อยละ 70 ข้อมูล ส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบร้อยละ 10 (Shams, 2014) ดังแสดงในตารางที่ 2 โดยที่ผลการวิเคราะห์จะนำข้อมูลส่วนที่ 3 ข้อมูลทดสอบไปใช้ในการวิเคราะห์ข้อมูล

ตารางที่ 2 ผลการแบ่งข้อมูลของชุดข้อมูลทั้ง 3 ชุด

ชุดข้อมูล	จำนวนข้อมูลทั้งหมด	ข้อมูลเรียนรู้ ร้อยละ 70	ข้อมูลตรวจสอบความถูกต้อง ร้อยละ 20	ข้อมูลทดสอบ ร้อยละ 10
โรคตับของรัฐอานธรประเทศ ประเทศอินเดีย	583	408	116	59
รายได้และรายจ่ายของครอบครัวฟิลิปปินส์	736	515	147	74
การตลาดของธนาคาร	854	597	171	86

ตารางที่ 2 ข้อมูลแต่ละชุดแบ่งเป็นข้อมูลเรียนรู้ร้อยละ 70 ข้อมูลตรวจสอบความถูกต้องร้อยละ 20 และข้อมูลทดสอบร้อยละ 10

ข้อมูลโรคตับของรัฐอานธรประเทศ ประเทศอินเดีย มีจำนวนข้อมูลทั้ง 583 ค่า แบ่งข้อมูลทั้งหมดเป็นข้อมูลเรียนรู้จำนวน 408 ค่า ข้อมูลตรวจสอบความถูกต้องจำนวน 116 ค่า และข้อมูลทดสอบจำนวน 59 ค่า

ข้อมูลรายได้และรายจ่ายของครอบครัวฟิลิปปินส์ มีจำนวนข้อมูลทั้ง 736 ค่า แบ่งข้อมูลทั้งหมดเป็นข้อมูลเรียนรู้จำนวน 515 ค่า ข้อมูลตรวจสอบความถูกต้องจำนวน 147 ค่า และข้อมูลทดสอบจำนวน 74 ค่า

ข้อมูลการตลาดของธนาคาร มีจำนวนข้อมูลทั้ง 854 ค่า แบ่งข้อมูลทั้งหมดเป็นข้อมูลเรียนรู้จำนวน 597 ค่า ข้อมูลตรวจสอบความถูกต้องจำนวน 171 ค่า และข้อมูลทดสอบจำนวน 86 ค่า

5. ผลการวิจัย

ตารางที่ 3 ผลการจำแนกเพื่อเปรียบเทียบประสิทธิภาพ โดยการกำหนดตัวสร้างเลขสุ่มเทียม (random seed) 10, 20, 30, 40 และ 50 ของชุดข้อมูลโรคตับของรัฐอานธรประเทศ ประเทศอินเดีย

วิธีการจำแนก	วิธีการแทนค่าข้อมูลสูญหาย	ค่าความถูกต้องเฉลี่ย	ค่าความแม่นยำเฉลี่ย	ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
เพื่อนบ้านใกล้สุด k ตัว	วิธีค่าเฉลี่ย	65.0847	0.7647	0.33678
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	66.4407	0.7702	0.32374
	วิธีค่ามัธยฐานของค่าใกล้เคียง	58.3051	0.6753	0.40214
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	64.8237	0.7617	0.33352
	วิธีแนวโน้มเชิงเส้น	65.4068	0.7567	0.33355
ต้นไม้ตัดสินใจ	วิธีค่าเฉลี่ย	67.1187	0.7681	0.28222
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	**71.1864**	0.7764	0.25049
	วิธีค่ามัธยฐานของค่าใกล้เคียง	66.7797	0.7629	0.28352
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	67.4576	0.7393	0.27772
	วิธีแนวโน้มเชิงเส้น	67.7966	0.7565	0.28923
โครงข่ายประสาทเทียม	วิธีค่าเฉลี่ย	70.1695	0.7838	0.24214
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	69.4915	**0.8305**	**0.23793**
	วิธีค่ามัธยฐานของค่าใกล้เคียง	66.7797	0.7752	0.25663
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	67.7966	0.7960	0.25115
	วิธีแนวโน้มเชิงเส้น	60.6787	0.6828	0.28895
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีค่าเฉลี่ย	58.6441	0.6008	0.41354
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	63.0506	0.6567	0.36950
	วิธีค่ามัธยฐานของค่าใกล้เคียง	60.6243	0.5824	0.39320
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	67.7966	0.6965	0.32203
	วิธีแนวโน้มเชิงเส้น	61.3559	0.6161	0.38645

ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนก โดยการแทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย วิธีค่าเฉลี่ยของค่าใกล้เคียง วิธีค่ามัธยฐานของค่าใกล้เคียง วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น และวิธีแนวโน้มเชิงเส้น

5.1 ชุดข้อมูลโรคตับของรัฐอานธรประเทศ ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่าสูญหายต่ำ จำนวนข้อมูลทั้งหมด 583 ค่า วิเคราะห์ข้อมูลด้วยข้อมูลทดสอบร้อยละ 10 จำนวน 59 ค่า

ตารางที่ 3 วิธีต้นไม้ตัดสินใจ แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ ร้อยละ 71.1864 และวิธีโครงข่ายประสาทเทียม แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง ให้ค่าความแม่นยำเฉลี่ยสูงสุด คือ 0.8305 และค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.23793

ตารางที่ 4 ผลการจำแนกเพื่อเปรียบเทียบประสิทธิภาพ โดยการกำหนดตัวสร้างเลขสุ่มเทียม 10, 20, 30, 40 และ 50 ของชุดข้อมูลรายได้และรายจ่ายของครอบครัวฟิลิปปินส์

วิธีการจำแนก	วิธีการแทนค่าข้อมูลสูญหาย	ค่าความถูกต้องเฉลี่ย	ค่าความแม่นยำเฉลี่ย	ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
เพื่อนบ้านใกล้สุด k ตัว	วิธีค่าเฉลี่ย	59.7297	0.5228	0.38909
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	67.2973	0.6410	0.317687
	วิธีค่ามัธยฐานของค่าใกล้เคียง	57.5676	0.5233	0.41213
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	58.1081	0.4955	0.40689
	วิธีแนวโน้มเชิงเส้น	49.3892	0.7070	0.29938
ต้นไม้ตัดสินใจ	วิธีค่าเฉลี่ย	67.8378	0.5960	0.29896
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	65.4054	0.5742	0.31270
	วิธีค่ามัธยฐานของค่าใกล้เคียง	65.9460	0.5780	0.31018
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	65.6757	0.5765	0.30808
	วิธีแนวโน้มเชิงเส้น	60.2703	0.5203	0.34916
โครงข่ายประสาทเทียม	วิธีค่าเฉลี่ย	72.9730	0.6869	0.23806
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	62.1622	0.5876	0.31074
	วิธีค่ามัธยฐานของค่าใกล้เคียง	70.3998	0.6482	0.25979
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	77.5676	0.8014	0.18318
	วิธีแนวโน้มเชิงเส้น	66.4865	0.6501	0.27434
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีค่าเฉลี่ย	69.2792	0.6579	0.30807
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	69.1892	0.6880	0.30809
	วิธีค่ามัธยฐานของค่าใกล้เคียง	64.3243	0.6151	0.35672
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	61.3514	0.5783	0.38647
	วิธีแนวโน้มเชิงเส้น	64.0541	0.5997	0.35944

ตารางที่ 5 ผลการจำแนกเพื่อเปรียบเทียบประสิทธิภาพ โดยการกำหนดตัวสร้างเลขสุ่มเทียม 10, 20, 30, 40 และ 50 ของชุดข้อมูลการตลาดของธนาคาร

วิธีการจำแนก	วิธีการแทนค่าข้อมูลสูญหาย	ค่าความถูกต้องเฉลี่ย	ค่าความแม่นยำเฉลี่ย	ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
เพื่อนบ้านใกล้สุด k ตัว	วิธีค่าเฉลี่ย	84.6512	0.9038	0.14978
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	82.0930	0.8916	0.17472
	วิธีค่ามัธยฐานของค่าใกล้เคียง	57.6419	0.8352	0.22004
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	77.4419	0.8331	0.22003
	วิธีแนวโน้มเชิงเส้น	77.4419	0.8518	0.22004
ต้นไม้ตัดสินใจ	วิธีค่าเฉลี่ย	86.7442	0.8974	0.11373
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	82.3256	0.8725	0.14922
	วิธีค่ามัธยฐานของค่าใกล้เคียง	81.6279	0.8574	0.16694
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	83.4884	0.8804	0.13935
	วิธีแนวโน้มเชิงเส้น	82.3256	0.8566	0.15260
โครงข่ายประสาทเทียม	วิธีค่าเฉลี่ย	88.6047	0.9440	0.09475
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	80.2326	0.8970	0.15023
	วิธีค่ามัธยฐานของค่าใกล้เคียง	82.0930	0.8932	0.15456
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	77.9070	0.8590	0.18237
	วิธีแนวโน้มเชิงเส้น	84.1861	0.9098	0.13524
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีค่าเฉลี่ย	70	0.7236	0.29998
	วิธีค่าเฉลี่ยของค่าใกล้เคียง	65.1163	0.6651	0.34883
	วิธีค่ามัธยฐานของค่าใกล้เคียง	66.7442	0.6618	0.33257
	วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น	61.8605	0.6289	0.38139
	วิธีแนวโน้มเชิงเส้น	60.9302	0.6104	0.39068

5.2 ชุดข้อมูลรายได้และรายจ่ายของครอบครัวฟิลิปปินส์ ซึ่งเป็นชุดข้อมูลที่มีค่าสูญหายปานกลาง จำนวนข้อมูลทั้งหมด 736 ค่า วิเคราะห์ข้อมูลด้วยข้อมูลทดสอบร้อยละ 10 จำนวน 74 ค่า

ตารางที่ 4 วิธีโครงข่ายประสาทเทียมแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ ร้อยละ 77.5676 ค่าความแม่นยำเฉลี่ยสูงสุด คือ 0.8014 และค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.18318

5.3 ชุดข้อมูลการตลาดของธนาคาร ซึ่งเป็นชุดข้อมูลที่มีค่าสูญหายสูง จำนวนข้อมูลทั้งหมด 854 ค่า วิเคราะห์ข้อมูลด้วยข้อมูลทดสอบร้อยละ 10 จำนวน 86 ค่า

ตารางที่ 5 วิธีโครงข่ายประสาทเทียมแทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย ให้ค่าความ

ถูกต้องเฉลี่ยสูงสุด คือ ร้อยละ 88.6047 ค่าความแม่นยำเฉลี่ยสูงสุด คือ 0.9440 และค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.09475

6. สรุปผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อต้องการเปรียบเทียบประสิทธิภาพวิธีการจำแนก 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด kตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน ในกรณีที่มีการแทนค่าข้อมูลสูญหาย 5 วิธี คือ วิธีค่าเฉลี่ย วิธีค่าเฉลี่ยของค่าใกล้เคียง วิธีค่ามัธยฐานของค่าใกล้เคียง วิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น และวิธีแนวโน้มเชิงเส้น ว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความถูกต้องและค่าความแม่นยำของการทำนายที่สูงกว่า ค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำกว่า

การค้นคว้าและศึกษาในการหาค่าข้อมูลสูญหายได้ข้อมูล 3 ชุด คือ โรคตับของรัฐอานธรประเทศ ประเทศอินเดีย เป็นชุดข้อมูลที่มีค่าสูญหายต่ำ วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยแทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง เนื่องจากให้ค่าความแม่นยำเฉลี่ยสูงสุดและค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด รายได้และรายจ่ายของครอบครัวฟิลิปปินส์ เป็นชุดข้อมูลที่มีค่าสูญหายปานกลาง วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น และการตลาดของธนาคาร เป็นชุดข้อมูลที่มีค่าสูญหายสูง วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยแทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย

7. อภิปรายผล

การสรุปผลงานวิจัยครั้งนี้ โรคตับของรัฐอานธรประเทศ ประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ยของค่าใกล้เคียง มีประสิทธิภาพในการจำแนกดีที่สุด เพราะว่ามีค่าความแม่นยำสูงที่สุด และค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด ให้ผลไม่สอดคล้องกับ Rahman และคณะ (2016) เรื่อง ปัจจัยที่ทำให้เกิดค่าสูญหายจากความผิดพลาดที่เกิดจากมนุษย์ ฮาร์ดแวร์ และอื่น ๆ พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว แทนค่าข้อมูลสูญหายด้วยวิธีค่าใกล้เคียงที่สุด (closest value) ให้ประสิทธิภาพดีที่สุด เพราะว่ามีค่าความถูกต้องสูงที่สุด เนื่องจากในงานวิจัยครั้งนี้ไม่ได้ศึกษาวิธีแทนค่าข้อมูลสูญหายด้วยวิธีค่าใกล้เคียงที่สุด ทำให้ผลไม่สอดคล้องกัน รายได้และรายจ่ายของครอบครัวฟิลิปปินส์ วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม แทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าระหว่างจุดแบบเชิงเส้น มีประสิทธิภาพในการจำแนกดีที่สุด เพราะว่ามีค่าความถูกต้องและค่าความแม่นยำสูงที่สุด และค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำสุด ให้ผลสอดคล้องกับ Kaiser (2014) เรื่อง การจัดการค่าข้อมูลสูญหายจากงานวิจัยที่มีจำนวนข้อมูลมากและมีข้อมูลสูญหาย โดยวิธีโครงข่ายประสาทเทียมเมื่อแทนค่าข้อมูลสูญหายด้วยวิธีค่าใกล้เคียงที่สุดมีประสิทธิภาพการจำแนกที่ดีที่สุด ให้ค่าความถูกต้องสูงที่สุด เพราะว่าให้ผลการจำแนกประสิทธิภาพดีที่สุดเหมือนกัน และการตลาดของธนาคาร วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม แทนค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย มีประสิทธิภาพในการจำแนกดีที่สุด เพราะว่ามีค่าความถูกต้องและค่าความแม่นยำสูงที่สุด ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด ให้ผลสอดคล้อง Hartini (2017) เรื่อง กระบวนการจำแนกการทำเหมืองข้อมูลเมื่อเกิดข้อมูลสูญหายจากการอ่านข้อมูลของเครื่องมือวัดและการบันทึกข้อมูล วิธีโครงข่ายประสาทเทียม โดยการประมาณค่าข้อมูล

สูญหายด้วยวิธี imputation มีประสิทธิภาพในการจำแนกมากกว่าวิธีเพื่อนบ้านใกล้สุด k ตัว และวิธีนาอีฟเบส์ เพราะมีค่าความถูกต้องและค่าความแม่นยำสูงที่สุด เพราะว่าให้ผลการจำแนกประสิทธิภาพดีที่สุดเหมือนกัน

8. ข้อเสนอแนะ

8.1 เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น ดังนั้นผู้วิจัยอาจวิเคราะห์ข้อมูลด้วยวิธีอื่น ๆ ได้แก่ วิธีฐานกฎ (rule based) วิธีนาอีฟเบส์ (Naive Bayes) วิธีลาดลงสโตแคสติก (Stochastic gradient descent) วิธีเบส์เนท (Bayes Net) วิธีการถดถอยลอจิสติกทวิภาค (binary logistic regression) และวิธีเพอร์เซปตรอนให้คะแนน (voted perceptron)

8.2 การแทนค่าข้อมูลสูญหายยังมีวิธีอื่น ๆ ที่สามารถแทนค่าข้อมูลสูญหาย เพื่อให้ได้ผลการแทนค่าข้อมูลสูญหายที่มีความสมบูรณ์มากขึ้น ผู้วิจัยอาจแทนค่าข้อมูลสูญหายด้วยวิธีอื่น ๆ ได้แก่ วิธี expectation-maximization algorithm วิธี full information maximum-likelihood และวิธี multiple imputation

9. รายการอ้างอิง

ฐณัฐ วงศ์สายเชื้อ, 2559, Replace Missing Value – การแทนค่าสูญหายในโปรแกรม SPSS, แหล่งที่มา : https://www.youtube.com/watch?v=WzaeJ_HAqtk, 25 ตุลาคม 2561.

ณัฐภัทร แก้วรัตนภัทร์, ปรีดาวรรณ เกษมธีการุณ และชนินทร์ มโนชญากร, 2555, การเปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลเพื่อแมนค่าสูญหาย, น. 561-567, การประชุมวิชาการระดับประเทศด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 8, สาขาวิชาการจัดการสารสนเทศ มหาวิทยาลัยราชภัฏสวนสุนันทา, กรุงเทพฯ.

นรุตม์ บุตรพลอย, 2553, การประยุกต์ Soft Computing และ k-Nearest Neighbor เพื่อใช้ประมาณค่าสูญหายของข้อมูล, น. 25-29, การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 3, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

นิรนาม, ขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้สุด k ตัว, 2561, แหล่งที่มา : https://th.wikipedia.org/wiki/, 11 ตุลาคม 2561.

พรพล ธรรมรงค์รัตน์, ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์ประสิทธิ์, 2553, การจำแนกประเภทเว็บเพจโดยใช้ค่าความถี่เอกสารและซัพพอร์ตเวกเตอร์แมชชีน, น. 55-61, การประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 12, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, กรุงเทพฯ.

ภัททิรา ล้อมเล็ก และวิทยา ยงเจริญ, 2557, การประยุกต์ใช้โครงข่ายประสาทเทียมสำหรับการทำนายสมรรถนะเครื่องทำความเย็นแบบดูดกลืน, ว.วิจัยพลังงาน 11(2): 67-78.

รุจิรา ธรรมสมบัติ, 2554, ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเตอร์เน็ตมือถือโดยใช้ต้นไม้ตัดสินใจ, รายงานวิจัย, สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์, กรุงเทพฯ.

วราฤทธิ์ พานิชกิจโกศลกุล, 2552, การจำลองแบบมอนติคาร์โลสำหรับประมาณค่าความแปรปรวนของการแจกแจงอินเวอร์เกาส์เซียนเมื่อข้อมูลมีค่าสูญหาย, ว.การวิจัยกาสะลองคำ 3(1): 14-23.

วริษฐา กณิกนันต์ และอนุภาพ สมบูรณ์สวัสดี, 2556, การเปรียบเทียบวิธีการประมาณสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหาย

แบบนอนอิกนอร์เรเบิล, น. 43-49, การประชุมหาดใหญ่วิชาการ ครั้งที่ 4 เรื่อง การวิจัยเพื่อพัฒนาสังคมไทย, มหาวิทยาลัยหาดใหญ่, สงขลา.

วุฒิ สุขเจริญ, 2558, การดำเนินการกับข้อมูลขาดหาย, ว.ร่มพฤกษ์ มหาวิทยาลัยเกริก 33(2): 11-32.

สายชล สินสมบูรณ์ทอง, 2560, การทำเหมืองข้อมูล เล่ม 1 การค้นหาความรู้จากข้อมูล, พิมพ์ครั้งที่ 2, จามจุรีโปรดักส์, กรุงเทพฯ.

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง, 2560, การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย, ว. วิทยาศาสตร์และเทคโนโลยี 25(5): 839-853.

Berson, A. and Smith, S. J., 2001, Data Warehousing, Data Mining and OLAP, McGraw-Hill, Boston.

Blomberg, L. C. and Ruiz, D. D. A., 2013, Evaluating the Influence of Missing Data on Classification Algorithms in Data Mining Application, Pomtificia Universidade Catolica do Rio Grande do Sul.

Hartini, E., 2017, Classification of missing values handling method during data mining, Sigma Epsilon 21(2): 49-60.

Kaiser, J., 2014, Dealing with missing values in data, J. Syst. Integr. 5(1): 42-51.

Mitchell, T. M., 1997, Machine Learning, McGraw-Hill, New York.

Portuguese Banking Institution, 2012, Bank Marketing Data Set, Available Source : https://archive.ics.uci.edu/ml/datasets/Bank+Marketing, November 12, 2018.

Rahman, S., Waqas, I., Imran, M.J. and Rehan, A., 2016, Treatment of missing values in data mining, J. Comp. Sci. Syst. Biol. 9(2): 51-53.

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), October 25, 2018.

Shams, R., 2014, Creating Training, Validation and Test Sets (Data Preprocessing), Available Source: https://www.youtube.com/watch?v=uiDFa7iY9yo, November 13, 2018.

The Philippine Statistics Authority (PSA), 2013, Filipino Family Income and Expenditure, Available Source: https://www.kaggle.com/grosvenpaul/family-income-and-expenditure/data, October 28, 2018.