การทดแทนข้อมูลสูญหายโดยใช้วิธี เค-เพื่อนบ้านใกล้ที่สุด ในเอ็กเซลวีบีเอ
Missing Values Imputation Using K-Nearest Neighbor in Excel VBA

สาโรช หว่างนุ่ม
sarot_wa@rmutto.ac.th
คณะบริหารธุรกิจและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก

บทคัดย่อ

การวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อศึกษาแนวทางการทดแทนข้อมูลสูญหายโดยวิธี เค-เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) ในเอ็กเซลวีบีเอ (Excel VBA) และเพื่อพัฒนาแนวทางการทดแทนข้อมูลสูญหายโดยใช้เอ็กเซลวีบีเอ เป็นงานวิจัยเชิงทดลอง โดยนำข้อมูล จำนวนประชากรจากการทะเบียนราษฎร ที่มีอายุระหว่าง 1-100 ปี ของจังหวัดในภาคตะวันออก ที่ได้มาจากเว็บไซต์ของสำนักงานสถิติแห่งชาติ จำนวน 100 ชุด มาเพื่อศึกษาแนวทางการทดแทนข้อมูลสูญหาย จากข้อมูลที่สมบูรณ์ นำมาทำให้สูญหาย 5%, 10%, และ 15% แล้วนำข้อมูลที่สูญหายนั้น ไปทำการทดแทนข้อมูลสูญหายโดยใช้วิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ย วิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่ามัธยฐาน วิธีการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธีการทดแทนข้อมูลสูญหายด้วยมัธยฐาน และเปรียบเทียบประสิทธิภาพโดยใช้ MAE, MSE, และ MAPE ผลการศึกษา พบว่า วิธีการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธีการทดแทนข้อมูลสูญหายด้วยมัธยฐาน มีประสิทธิภาพน้อยกว่า วิธี เค-เพื่อนบ้านใกล้ที่สุด ทั้ง 2 แบบ

คำสำคัญ ข้อมูลสูญหาย, วิธี เค-เพื่อนบ้านที่ใกล้ที่สุด, เอ็กเซลวีบีเอ

Abstract

This research aims to study the method of replacing lost data by K-Nearest Neighbor in Excel VBA and to develop an approach to replace lost data using Excel VBA. It is an experimental research by using data from the population from the civil registration, who are between 1-100 years of age in the eastern provinces which obtained from the website of the National Statistical Office of 100 records to study the ways to replace lost data. From complete data have to adopt to loss of 5%, 10%, and 15% and bring the lost data to imputation using the K-Nearest Neighbor Mean method, K-Nearest Neighbor Median method, Mean Imputation method, and Median Imputation method. The efficiency of each method was compared using MAE, MSE, and MAPE. From this paper found that Mean Imputation method and Median Imputation method less efficient than the two K-Nearest Neighbor methods.

Keywords Missing Value, K-Nearest Neighbor, Excel VBA

บทนำ

ความอยู่รอดขององค์กร ขึ้นอยู่กับการตัดสินใจของผู้บริหาร การที่ผู้บริหารจะสามารถตัดสินใจใดๆ ให้เหมาะสมนั้น นอกจากต้องอาศัยวิจารณญาณของผู้บริหารแล้ว ยังต้องอาศัยข้อมูลหรือสารสนเทศที่ดี สารสนเทศที่ดีมาจากระบบสนับสนุนการตัดสินใจที่ดี ที่ให้สารสนเทศที่รองรับการตัดสินใจของผู้บริหารนั้นได้ การมีสารสนเทศที่ดี ที่ถูกต้อง ทันสมัย เป็นปัจจุบัน และรองรับการตัดสินใจของผู้บริหารได้นั้น ขึ้นอยู่กับข้อมูลที่นำมาวิเคราะห์ หากข้อมูลที่นำมาวิเคราะห์นั้นมีความถูกต้อง เป็นปัจจุบัน และมีความสมบูรณ์ ย่อมส่งผลให้เกิดสารสนเทศที่ดี แต่หากข้อมูลที่ได้มาไม่สมบูรณ์ ย่อมไม่สามารถที่จะนำไปวิเคราะห์ให้เป็นสารสนเทศที่ดีได้ ปัญหาที่เกิดจากความไม่สมบูรณ์ของข้อมูลนี้นับเป็นปัญหาสำคัญ ที่อาจทำให้ไม่สามารถนำข้อมูลที่ได้ไปวิเคราะห์ให้เกิดสารสนเทศที่ดีได้

ในงานวิจัยก็เช่นกัน หากข้อมูลที่ได้มาจากการเก็บรวบรวมนั้นไม่สมบูรณ์ เกิดการขาดหายหรือการสูญหาย (Missing) ย่อมส่งผลกระทบต่อการวิเคราะห์ข้อมูลและผลลัพธ์ที่ได้ อาจทำให้ต้องทิ้งข้อมูลบางชุดไป เนื่องจากความไม่สมบูรณ์ ทั้ง ๆ ที่กว่าจะได้ข้อมูลนั้นมา เป็นไปด้วยความยากลำบาก ต้องใช้เวลาและความอุตสาหะของผู้ทำวิจัยเป็นอย่างมาก และในบางครั้ง การที่ผู้วิจัยจะไปเก็บข้อมูลนั้นซ้ำอาจจะไม่สามารถกระทำได้ เนื่องจากติดขัดด้วยปัจจัยหลายประการที่ทำให้เกิดความไม่สะดวก

จากปัญหาความไม่สมบูรณ์ของข้อมูลนี้ ผู้วิจัยมองเห็นความสำคัญของการนำคอมพิวเตอร์เข้ามาช่วยในการแก้ไขปัญหา ซึ่งปัญหาที่ผู้วิจัยพบบ่อย ๆ คือ ปัญหาเรื่องข้อมูลงานวิจัยที่จะนำมาวิเคราะห์ ไม่สมบูรณ์ หรือเกิดค่าสูญหาย (Missing Values) ในบางตัวแปร จากจุดนี้ ทำให้ผู้วิจัยมีความคิดที่จะนำเสนอแนวทางในการเติมเต็มข้อมูลที่สูญหายไป และเพื่อความสะดวกสำหรับผู้ใช้ข้อมูล ผู้วิจัยได้เลือกพัฒนาการเติมเต็มข้อมูลที่สูญหายไป โดยเลือกวิธี K-Nearest Neighbor ใน Microsoft Excel VBA เพื่อให้ผู้ทำวิจัยหรือนักวิจัยสามารถนำไปใช้แก้ไขปัญหาค่าสูญหายของข้อมูลในงานวิจัยที่ดำเนินอยู่นั้นได้ รวมถึงสามารถใช้เป็นแนวทางประกอบการตัดสินใจเลือกวิธีการแทนที่ข้อมูลสูญหายแนวทางหนึ่งได้

พัชนา สุวรรณแสน (2562) ได้ศึกษาเรื่อง การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์ (KNN) พบว่า วิธีการเคเนียร์เรสเนเบอร์ เป็นวิธีการที่น่าสนใจในการนำมาใช้ประมาณค่าสูญหาย ปัจจัยสำคัญแนวทางหนึ่ง คือ การพิจารณาค่า k ที่เหมาะสม ซึ่งมีหลากหลายงานวิจัยที่เกี่ยวข้อง และนอกจากนี้ ยังมีการพัฒนาปรับปรุงวิธีการนี้ในหลายรูปแบบ ไม่ว่าจะเป็นการใช้วิธีเลือกคุณลักษณะเข้ามาร่วมกับการประมาณค่าสูญหายด้วยวิธี KNN เช่น วิธี Sequential KNN (SKNN) วิธี KNN Feature Selection (KNNFS) เป็นต้น

ปูเป้ สุดศิลา, อำไพ ทองธีรภาพ และบุญอ้อม โฉมที (2561) ได้ศึกษาเรื่อง การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระ ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม โดยมีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหาย 4 วิธี ได้แก่ วิธี Mean Imputation (Mean) วิธี Multiple Imputation (MI) วิธี K-Nearest Neighbor (KNN) และวิธี Weight Locally Linear Reconstruction

(WLLR) เกณฑ์ในการเปรียบเทียบ คือ ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) พบว่า ที่ทุกร้อยละการสูญหายที่กำหนด เมื่อขนาดตัวอย่างเท่ากับ 60, 100, และ 300 วิธี Mean มีประสิทธิภาพดีที่สุด เมื่อขนาดตัวอย่างเท่ากับ 500 วิธี MI มีประสิทธิภาพที่ดีที่สุด นอกจากนี้พบว่า ค่า EMSE เพิ่มขึ้น เมื่อร้อยละการสูญหายเพิ่มขึ้น และลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

พงศกร ธีรรัศมี (2558) ได้ศึกษาเรื่อง วิธีการหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์ โดยมีวัตถุประสงค์เพื่อแนะนำเกี่ยวกับค่า เค ที่เหมาะสมในการจำแนกข้อมูลให้มีค่าความแม่นที่สูง ผลการศึกษาพบว่า จากข้อมูลทดลองทั้ง 5 ข้อมูล ได้แก่ ข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคเบาหวาน ข้อมูลโรคหอบหืด และข้อมูลโรคไทรอยด์ วิเคราะห์ด้วยอัลกอริทึมจากงานวิจัย และมีการใช้การวัดการกระจายข้อมูลแบบมาร์เดีย ซึ่งจะใช้ค่า Mardia's of Multivariate Skew เมื่อพิจารณาค่าเฉลี่ย Mardia's of Multivariate Skew ตามอัลกอริทึมจากงานวิจัย จะให้ผลลัพธ์ คือ ข้อมูลโรคหัวใจ เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 74 ซึ่งเป็นค่าที่สูงที่สุดในตาราง ข้อมูลโรคมะเร็งเต้านม เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 99 ซึ่งเป็นค่าที่สูงที่สุดในตาราง ข้อมูลโรคเบาหวาน เมื่อใช้ค่า เค เป็น 10% จากจำนวนข้อมูล และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 77 ซึ่งเป็นค่าที่สูงที่สุดในตาราง ข้อมูลโรคหอบหืด เมื่อใช้ค่า เค เป็น 10% จากจำนวนข้อมูล และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 81 ซึ่งเป็นค่าที่สูงที่สุดในตาราง ข้อมูลโรคไทรอยด์ เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 93 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาแนวทางการทดแทนข้อมูลสูญหายโดยวิธี เค-เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) ในเอ็กเซลวีบีเอ (Excel VBA)
2. เพื่อพัฒนาแนวทางการทดแทนข้อมูลสูญหายโดยใช้เอ็กเซลวีบีเอ (Excel VBA)

ขอบเขตการวิจัย

1. ขอบเขตประชากร งานวิจัยนี้ใช้ข้อมูลที่นำมาศึกษา คือ จำนวนประชากรจากการทะเบียนราษฎร จำแนกตามอายุ เพศ ภาค และจังหวัด ปี พ.ศ. 2563 จากเว็บไซต์ของสำนักงานสถิติแห่งชาติ เท่านั้น
2. ขอบเขตตัวแปร เป็นการนำข้อมูล จำนวนประชากรจากการทะเบียนราษฎร จำแนกตามอายุ เพศ ภาค และจังหวัด ปี พ.ศ. 2563 จากเว็บไซต์ของสำนักงานสถิติแห่งชาติ มาใช้ในการศึกษา โดยข้อมูลที่ผู้วิจัยได้มานี้ เป็นข้อมูลจำนวนประชากรที่มีอายุไม่ถึง 1 ปี จนถึงมีอายุมากกว่า 100 ปี จากข้อมูลที่ได้ ผู้วิจัยพบว่าจำนวนประชากรที่มีอายุไม่ถึง 1 ปี และจำนวนประชากรที่มีอายุมากกว่า 100 ปี มีจำนวนแตกต่างจากกลุ่มอายุ 1-100 ปี (Outliers) ผู้วิจัยจึงตัดข้อมูลจำนวนประชากร 2 กลุ่มอายุนี้ออก ดังนั้นข้อมูลที่นำมาศึกษาคงเหลือแต่ข้อมูลจำนวนประชากรแต่ละช่วงอายุ ตั้งแต่ 1 ปี จนถึง 100 ปี รวมเป็นจำนวนข้อมูลทั้งสิ้น 100 ชุด เท่านั้น

3. ขอบเขตเวลา ระยะเวลาที่ดำเนินการวิจัย คือ 1 เมษายน 2564 – 20 มิถุนายน 2564

วิธีดำเนินการวิจัย

1. ระเบียบวิธีวิจัย งานวิจัยชิ้นนี้เป็นการวิจัยเชิงทดลอง โดยนำข้อมูลสมบูรณ์ที่คัดเลือก มาทำให้มีข้อมูลสูญหาย 5%, 10%, และ 15% เพื่อทดสอบการทดแทนข้อมูลสูญหายเปรียบเทียบ 4 วิธีการ โดยใช้โปรแกรมเอ็กเซลวีบีเอ

2. ขั้นตอนการวิจัย

2.1 เริ่มจากการศึกษาค้นคว้า การค้นหา และเก็บรวบรวมข้อมูล และคัดเลือกข้อมูลสมบูรณ์ที่จะนำมาศึกษาการทดแทนข้อมูลสูญหาย เมื่อได้ข้อมูลที่ต้องการแล้ว นำข้อมูลที่คัดเลือกนั้นมาทำให้มีข้อมูลสูญหาย 5%, 10%, และ 15%

2.2 พัฒนาโปรแกรมเอ็กเซลวีบีเอ เพื่อทำการทดแทนข้อมูลสูญหายวิธี เค-เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor : KNN) วิธีการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean Imputation) และวิธีการทดแทนข้อมูลสูญหายด้วยค่ามัธยฐาน (Median Imputation) ในส่วนของการทดแทนข้อมูลสูญหายโดยวิธี เค-เพื่อนบ้านใกล้ที่สุด ผู้วิจัยได้แยกย่อยเป็น 2 แบบ คือ แบบที่ใช้ค่าเฉลี่ย และแบบที่ใช้มัธยฐาน โดยมีวิธีการวิเคราะห์แตกต่างกันในรายละเอียด คือ เมื่อคำนวณหาระยะห่างโดยใช้ ระยะทางแบบยูคลิด (Euclidean Distance) เสร็จแล้ว วิธีที่ใช้ค่าเฉลี่ย จะเลือกค่าที่มีระยะทางที่น้อยที่สุดจำนวน K ค่า มาคำนวณหาค่าเฉลี่ย แล้วนำผลหรือค่าเฉลี่ยที่ได้มาเป็นค่าทดแทนข้อมูลสูญหาย ส่วนวิธีที่ใช้ค่ามัธยฐาน ก็จะเลือกค่าที่มีระยะทางที่น้อยที่สุดจำนวน K ค่า มาคำนวณหาค่ามัธยฐาน จากนั้น นำผลลัพธ์ที่ได้มาเป็นค่าทดแทนข้อมูลสูญหาย ในงานวิจัยชิ้นนี้ จะใช้ค่า K ในการศึกษา 3 ค่า คือ K = 1, 3, และ 5

2.3 นำผลการทดแทนข้อมูลสูญหายมาเปรียบเทียบกัน โดยใช้ MAE (Mean Absolute Error) MSE (Mean Squared Error) และ MAPE (Mean Absolute Percentage Error) แล้วสรุปผล

3. การเก็บรวบรวมข้อมูล ข้อมูลที่นำมาศึกษาเป็นข้อมูลแบบทุติภูมิ ผู้วิจัยไม่ได้เป็นผู้ทำการเก็บรวบรวมข้อมูลด้วยตนเอง แต่ได้จากการค้นหา และคัดเลือกข้อมูลมาจากเว็บไซต์ต่าง และได้เลือกข้อมูลจำนวนประชากรจากการทะเบียนราษฎร จำแนกตามอายุ เพศ ภาค และจังหวัด ปี พ.ศ. 2563 ของสำนักงานสถิติแห่งชาติ มาใช้ในการศึกษา (สำนักงานสถิติแห่งชาติ, 2563)

4. การวิเคราะห์ข้อมูล ในงานวิจัยชิ้นนี้ มีวิธีการวิเคราะห์ข้อมูล ดังนี้

4.1 ระยะทางแบบยูคลิด (Euclidean Distance) คำนวณระยะห่างระหว่างจุด 2 จุด (ในที่นี้คือ จุด A และ จุด B) ซึ่งก็คือ คำนวณระยะห่างของข้อมูลแต่ละชุด มีสูตร ดังนี้ (Arjun Panesar, 2019)

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

เมื่อ a_i คือ ค่าที่ข้อมูลชุด A ตัวที่ i (หรือคอลัมน์ที่ i)

b_i คือ ค่าที่ข้อมูลชุด B ตัวที่ i (หรือคอลัมน์ที่ i)

4.2 ค่าเฉลี่ย (Mean) หรือค่าเฉลี่ยเลขคณิต สามารถคำนวณได้จากสูตร ดังนี้ (ธีระพงษ์ กระการดี, 2564)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

เมื่อ x_i คือ ค่าของข้อมูลตัวที่ i
n คือ จำนวนข้อมูล

4.3 ค่ามัธยฐาน (Median) การหามัธยฐานของข้อมูล ดังนี้ (ธีระพงษ์ กระการดี, 2564)

4.3.1 เรียงข้อมูลจากน้อยไปมาก

4.3.2 หาตำแหน่งของมัธยฐาน จากข้อมูลตำแหน่งที่ $\frac{n+1}{2}$

เมื่อ n = จำนวนข้อมูลทั้งหมดที่จะหามัธยฐาน

4.4 ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) เป็นค่าวัดความถูกต้องของการทำนาย ที่วัดจากค่าความคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน มีหน่วยวัดหน่วยเดียวกับค่าสังเกต (สายชล สินสมบูรณ์ทอง, 2563)

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

เมื่อ y_i คือ ค่าจริงของข้อมูล
$\hat{y}_i$ คือ ค่าที่ได้จากการทดแทนข้อมูลสูญหาย
n คือ จำนวนข้อมูลของกลุ่มตัวอย่าง

4.5 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error : MSE) เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยความเอนเอียงและความแปรปรวน มีสูตรคำนวณดังนี้ (สายชล สินสมบูรณ์ทอง, 2563)

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

เมื่อ y_i คือ ค่าจริงของข้อมูล
$\hat{y}_i$ คือ ค่าที่ได้จากการทดแทนข้อมูลสูญหาย
n คือ จำนวนข้อมูลของกลุ่มตัวอย่าง

4.6 ค่าสัมบูรณ์ของเปอร์เซ็นต์ของความคลาดเคลื่อน (Mean Absolute Percentage Error : MAPE) มีสูตรคำนวณ ดังนี้ (อาวีพร ปานทอง, 2562)

$$MAPE = 100 * \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

เมื่อ y_i คือ ค่าจริงของข้อมูล
$\hat{y}_i$ คือ ค่าที่ได้จากการทดแทนข้อมูลสูญหาย

ผลการวิจัย

จากการพัฒนาโปรแกรมเอ็กเซลวีบีเอ เพื่อทดแทนข้อมูลสูญหายทั้ง 4 วิธีการ ได้แก่ การทดแทนข้อมูลสูญหายโดยวิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ยในการทดแทนข้อมูลสูญหาย (KNN-Mean) วิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่ามัธยฐานในการทดแทนข้อมูลสูญหาย (KNN-Median) วิธีการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean Imputation) และวิธีการทดแทนข้อมูลสูญหายด้วยค่ามัธยฐาน (Median

Imputation) เมื่อทำการประมวลผลชุดข้อมูล จำนวนประชากรจากการทะเบียน ที่มีอายุระหว่าง 1-100 ปี ของจังหวัดในภาคตะวันออก ปี พ.ศ. 2563 จำนวน 100 ชุด โดยการนำข้อมูลชุดสมบูรณ์ มาทำให้เกิดการสูญหาย 5%, 10%, และ 15% จากนั้นนำชุดข้อมูลที่สูญหายนี้มาทำการประมวลผลด้วยโปรแกรมเอ็กเซลวีบีเอ แล้วทดสอบประสิทธิภาพการทดแทนข้อมูลสูญหาย ด้วย MAE, MSE, และ MAPE มีผลการทดสอบประสิทธิภาพการทดแทนข้อมูลสูญหาย ดังแสดงในตารางที่ 1

ตารางที่ 1 เปรียบเทียบผลการทดแทนข้อมูลสูญหาย 5%, 10%, และ 15% โดยใช้ MAE MSE และ MAPE

ข้อมูลสูญหาย	วิธีการทดแทนข้อมูลสูญหาย	ตัววัดประสิทธิภาพ		
		MAE	MSE	MAPE
5%	KNN-Mean เมื่อ K=1	293.25	131,910.00	4.40
	KNN-Mean เมื่อ K=3	290.33	134,314.18	11.97
	KNN-Mean เมื่อ K=5	192.39	99,732.29	19.74
	KNN-Median เมื่อ K=1	293.25	131,910.00	4.40
	KNN-Median เมื่อ K=3	394.25	200,924.80	13.44
	KNN-Median เมื่อ K=5	353.25	220,499.80	18.21
	Mean Imputation	10,205.00	104,601,533.49	1,719.50
	Median Imputation	10,983.50	130,301,957.20	2,125.04
10%	KNN-Mean เมื่อ K=1	407.50	385,891.10	5.17
	KNN-Mean เมื่อ K=3	302.23	143,734.14	7.54
	KNN-Mean เมื่อ K=5	254.42	89,167.87	10.79
	KNN-Median เมื่อ K=1	407.50	385,891.10	5.17
	KNN-Median เมื่อ K=3	375.80	217,061.80	9.39
	KNN-Median เมื่อ K=5	383.50	197,713.30	11.60
	Mean Imputation	8,677.54	94,105,511.99	980.73
	Median Imputation	8,374.40	111,675,920.60	1,210.50
15%	KNN-Mean เมื่อ K=1	404.53	323,370.67	4.04
	KNN-Mean เมื่อ K=3	268.15	116,681.01	5.36
	KNN-Mean เมื่อ K=5	278.33	102,165.69	7.67
	KNN-Median เมื่อ K=1	404.53	323,370.67	4.04
	KNN-Median เมื่อ K=3	380.60	207,855.40	6.84
	KNN-Median เมื่อ K=5	404.73	215,098.07	8.39
	Mean Imputation	8,257.53	83,163,230.91	647.49
	Median Imputation	6,794.53	79,152,488.93	784.90

จากตารางที่ 1 พบว่า เมื่อทำการทดแทนข้อมูลสูญหายด้วยวิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ย (KNN-Mean) วิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่ามัธยฐาน (KNN-Median) วิธีทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean Imputation) และวิธีการทดแทนข้อมูลสูญหายด้วยค่ามัธยฐาน (Median Imputation) ในภาพรวมพบว่า ไม่ว่าจะมีข้อมูลสูญหาย 5%, 10%, หรือ 15% ก็ตาม วิธีการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธีการทดแทนข้อมูลสูญหายด้วยค่ามัธยฐาน มีประสิทธิภาพน้อยที่สุด และเมื่อพิจารณาเปรียบเทียบระหว่างวิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ย กับวิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่ามัธยฐาน พบว่า วิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ย มีประสิทธิภาพใกล้เคียงกับวิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่ามัธยฐาน แต่วิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ย จะมีประสิทธิภาพดีกว่าเล็กน้อย และวิธีการทั้ง 2 แบบนี้ จะมีประสิทธิภาพในการทดแทนข้อมูลสูญหายดีขึ้น เมื่อค่า K เพิ่มขึ้น และเมื่อพิจารณาในส่วนของการเพิ่มขึ้นของค่าสูญหาย พบว่า การทดแทนข้อมูลสูญหายโดยวิธี เค-เพื่อนบ้านใกล้ที่สุดทั้ง 2 แบบ กรณีที่ข้อมูลที่สูญหาย 5% จะมีประสิทธิภาพดีกว่า 10% เล็กน้อย และการทดแทนข้อมูลสูญหาย กรณีข้อมูลสูญหาย 10% มีประสิทธิภาพใกล้เคียงกับ กรณีข้อมูลสูญหาย 15%

อภิปรายผลการวิจัย

ผลจากการศึกษาพบว่า การทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย และการทดแทนข้อมูลสูญหายด้วยค่ามัธยฐาน ไม่เหมาะสมที่จะนำมาใช้ทดแทนข้อมูลสูญหายได้จริง ซึ่งสอดคล้องกับงานการศึกษาของ ดร. อานนท์ ศักดิ์วรวิชญ์ (อานนท์ ศักดิ์วรวิชญ์, 2559) ที่กล่าวถึงเรื่อง การทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย ถือเป็นกรณีที่เลวร้ายที่สุดในการวิเคราะห์ข้อมูลสูญหาย เนื่องจากการแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย เป็นการแทนที่ข้อมูลด้วยค่าเพียงค่าเดียว ถึงแม้ว่าผลโดยรวมของข้อมูลจะได้ค่าของข้อมูลโดยรวมคงเดิม แต่ธรรมชาติของข้อมูลจริงๆ ไม่ได้มีค่าเดียวเสมอไป มีค่ามากน้อยเกิดขึ้นได้ การแทนที่ข้อมูลด้วยค่าเพียงค่าเดียวนี้ จะส่งผลต่อสถิติทดสอบตัวอื่นได้ อย่างเช่น ค่าของความแปรปรวนจะลดลงไปเป็นอย่างมาก วิธีการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ยนี้ (รวมถึงค่ามัธยฐาน) อาจจะเป็นตัวเลือกที่มีอยู่ในโปรแกรมสำเร็จรูปหลายๆ โปรแกรม เนื่องจากความสะดวกในการใช้ แต่อาจจะไม่เหมาะสม

ในส่วนของการทดแทนข้อมูลสูญหายโดยใช้วิธี เค-เพื่อนบ้านที่ใกล้ที่สุด จากการทดลองเปลี่ยนค่าเฉลี่ย K ค่า มาเป็นค่ามัธยฐาน K ค่า และประมวลผลหลายๆ รอบกับข้อมูล ผู้วิจัยพบว่า ให้ผลลัพธ์ในการการทดแทนข้อมูลสูญหายได้ดีระดับหนึ่ง อาจจะไม่ดีในทุกค่า แต่ถือว่าดีกว่าการปล่อยให้ข้อมูลสูญหาย หรือดีกว่าการแทนที่ด้วยค่าเพียงค่าเดียว แบบค่าเฉลี่ย หรือมัธยฐาน ซึ่งสอดคล้องการงานการศึกษาของ พัชนา สุวรรณแสน (พัชนา สุวรรณแสน, 2562) ที่ศึกษาเรื่อง การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์ (KNN) ได้กล่าวถึงว่า วิธีการเคเนียร์เรสเนเบอร์ เป็นวิธีการที่น่าสนใจในการนำมาใช้ประมาณค่าสูญหาย ปัจจัยสำคัญแนวทางหนึ่ง คือ การพิจารณาค่า k ที่เหมาะสม (ในเอ็กเซลวีบีเอ ที่ผู้วิจัยได้พัฒนาขึ้นมานี้ ตั้งค่าเบื้องต้นไว้ที่ K = 3 ผู้ใช้สามารถเปลี่ยนค่า K ที่ต้องการก่อนการประมวลผลได้) เพื่อให้เห็นภาพ ผู้วิจัยขอนำเสนอผลการทดแทนข้อมูลสูญหายจากเอ็กเซลวีบีเอ เมื่อเทียบกับข้อมูลจริง ในกรณีข้อมูลสูญหาย 5% และ K = 1, 3, และ 5 โดยวิธี เค-เพื่อนบ้านใกล้ที่สุด ทั้งแบบที่ใช้ค่าเฉลี่ย (KNN-Mean) และแบบที่ใช้ค่ามัธยฐาน (KNN-Median)

เทียบกับการทดแทนข้อมูลด้วยค่าเฉลี่ย (Mean Imputation) และการทดแทนข้อมูลด้วยค่ามัธยฐาน (Median Imputation) ดังแสดงในตารางที่ 2

ตารางที่ 2 ตัวอย่างข้อมูลที่ได้จากการทดแทนข้อมูลสูญหาย 5% ด้วยวิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ย K ค่า แบบที่ใช้ค่ามัธยฐาน K ค่า เมื่อ K = 1, 3, และ 5 เทียบกับการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย และการทดแทนข้อมูลสูญหายด้วยค่ามัธยฐาน จากเอ็กเซลวีบีเอ

ข้อมูลจริง	KNN-Mean			KNN-Median			Mean Imputation	Median Imputation
	K=1	K =3	K =5	K =1	K =3	K =5		
18,437	18,388	18,567	19,015	18,388	18,388	19,083	15,410	19,043
27,929	28,601	27,951	27,819	28,601	27,718	27,718	15,410	19,043
17,486	17,034	18,228	17,773	17,034	18,388	18,184	15,410	19,043
4,411	4,430	4,721	4,620	4,430	4,778	4,778	15,410	19,043
184	214	272	347	214	281	321	15,410	19,043

จากตารางที่ 2 จะเห็นได้ว่า การทดแทนข้อมูลสูญหายโดยวิธี เค-เพื่อนบ้านใกล้ที่สุด แบบที่ใช้ค่าเฉลี่ยและแบบที่ใช้ค่ามัธยฐาน เมื่อ K = 1 จะมีค่าเดียวกัน เนื่องจากมีเพียง 1 ค่า (ที่ K = 1 ค่าเฉลี่ย = ค่ามัธยฐาน) จะแตกต่างกันเมื่อ K > 1 (ตามตารางที่ 2 จะเห็นได้ที่ K = 3 และ K = 5) จะเห็นได้ว่า เมื่อค่า K เพิ่มมากขึ้น ค่าที่ได้จากการประมวลผล ยิ่งห่างจากค่าตามข้อมูลจริง แต่แบบที่ใช้ค่าเฉลี่ย ยังคงมีความแตกต่างจากข้อมูลจริงไม่มากเท่าแบบที่ใช้ค่ามัธยฐาน สำหรับเรื่องนี้ผู้วิจัยมีความเห็นว่า เมื่อเวลานำไปใช้งานกับข้อมูลจริงชุดอื่นๆ ผู้ใช้โปรแกรมต้องปรับค่า K หลายๆ ค่า เพื่อหาค่า K ที่เหมาะสมต่อไป และเมื่อพิจารณาการทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย และการทดแทนข้อมูลสูญหายด้วยมัธยฐาน จะเห็นว่า ทั้ง 2 วิธีการ ใช้ค่าเดียวในการทดแทนข้อมูลสูญหาย เมื่อพิจารณาเห็นได้ว่า ที่ข้อมูลจริง 184 การทดแทนข้อมูลสูญหายด้วยค่าเฉลี่ย และการทดแทนข้อมูลสูญหายด้วยค่ามัธยฐาน จะมีค่าแตกต่างจากข้อมูลจริงในระดับที่สูงมาก ดังนั้น การทดแทนข้อมูลสูญหายด้วยทั้ง 2 วิธีนี้ จึงดูไม่เหมาะสม หากข้อมูลมีความแตกต่างกันมาก

ข้อเสนอแนะ

1. ข้อเสนอแนะเพื่อการนำผลการวิจัยไปใช้งาน

ในการทดแทนข้อมูลสูญหาย โดยวิธี เค-เพื่อนบ้านใกล้ที่สุด นี้ สิ่งที่ควรให้ความสำคัญ คือ ค่า K ที่เหมาะสม จากการทดลองประมวลผลหลายๆ รอบ และจากการศึกษางานวิจัย พบว่า เป็นเรื่องยากที่จะระบุค่า K ที่เหมาะสม แบบตายตัวลงไป เนื่องจากข้อมูลแต่ละประเภท แต่ละเรื่อง ล้วนมีลักษณะหรือธรรมชาติของข้อมูลที่แตกต่างกัน เพื่อให้ได้ค่า K ที่เหมาะสม ควรมีการประมวลโปรแกรมหลายๆ รอบ โดยแต่ละรอบมีค่า K ที่แตกต่างกันไป เพื่อให้การทดแทนข้อมูลสูญหายมีความถูกต้องมากขึ้น

2. ข้อเสนอแนะเพื่อการวิจัยต่อไป

การทดแทนข้อมูลสูญหาย มีหลายวิธี หลายแนวทาง ผู้สนใจสามารถนำแต่ละวิธีการเหล่านี้มาผสมผสานกัน เป็นวิธีการใหม่ในการทดแทนข้อมูลสูญหายได้ และนอกจากนี้ โปรแกรมสำเร็จรูปที่ช่วยเหลือผู้ใช้ข้อมูลในการทดแทนข้อมูลสูญหาย ส่วนใหญ่มีวิธีการทดแทนข้อมูลสูญหายให้เลือกใช้ได้อยู่จำกัด และในการใช้งานยังมีข้อจำกัด สำหรับผู้สนใจในการทำวิจัยที่เกี่ยวข้องกับการทดแทนข้อมูลสูญหาย การศึกษาจากงานการศึกษา งานวิจัย ที่มีผู้ศึกษา มีผู้จัดทำมาก่อน สามารถใช้เป็นแนวทางในการทำวิจัยเรื่องการทดแทนข้อมูลสูญหายได้เป็นอย่างดี บางงานการศึกษาอาจจะผ่านกาลเวลาล่วงเลยมาพอสมควร แต่ไม่ได้เป็นงานที่ล้าสมัยเสมอไป จากการค้นคว้าของผู้วิจัย พบว่า มีงานวิจัยที่เกี่ยวกับกับข้อมูลสูญหายทั้งที่เป็นภาษาไทย และที่เป็นภาษาอังกฤษอยู่พอสมควร มีหลายงานน่าสนใจ แต่ไม่มีการนำมาใช้ ยกตัวอย่างเช่น งานวิจัยของ รศ.ดร. มนตรี พิริยะกุล ซึ่งเป็นงานที่เป็นแรงบันดาลใจเริ่มต้นนี้ของผู้วิจัย (มนตรี พิริยะกุล, 2548) ไม่มีผู้นำไปใช้งานต่อ ซึ่งเป็นที่น่าเสียดาย สำหรับผู้สนใจ หรือผู้อยากทำวิจัยเกี่ยวกับข้อมูลสูญหาย ควรค้นคว้าจากงานวิจัย งานการศึกษาที่มีผู้จัดทำมาแล้ว งานเหล่านี้สามารถใช้เป็นแนวทางในการลดความผิดพลาดของข้อมูลสูญหายได้เป็นอย่างดี

นอกจากนี้ จากการศึกษาค้นคว้าของผู้วิจัย ในเรื่องข้อมูลสูญหาย มีอีกประเด็นหนึ่งที่ผู้วิจัย สนใจและคาดว่าสามารถใช้เป็นแนวทางในการทำวิจัยเรื่องต่อไปได้ คือ จำนวนของข้อมูลสูญหายที่เกิดขึ้น ควรจะมีได้สูงสุดไม่เกินสัดส่วนเท่าไร ถ้าเทียบกับจำนวนข้อมูลทั้งหมดที่มี ในบางงานวิจัย ที่ผู้วิจัยได้ค้นคว้า พบว่า มีการทดสอบข้อมูลสูญหายในระดับที่สูงกว่าร้อยละ 25 ผู้วิจัยมีความเห็นว่า เริ่มมีการสูญหายในระดับที่สูงพอสมควร (เพื่อให้เข้าใจตรงกัน ผู้วิจัยขอยกตัวอย่างเปรียบเทียบ ข้อความใน 1 หน้ากระดาษ สมมติว่ามี 1,000 ตัวอักษร หากมีข้อมูลสูญหาย ร้อยละ 25 หมายความว่า เนื้อความหายไป 1 ใน 4 ของหน้ากระดาษ การทดแทนเนื้อความให้สมบูรณ์อาจจะยืนยันความถูกต้องของผลลัพธ์ได้ยาก) แต่ยังไม่มีงานวิจัยชิ้นใดที่แสดงให้เห็นว่า ไม่สามารถทำการทดแทนข้อมูลสูญหายนั้นได้ เพราะถ้าข้อมูลมีค่าใกล้เคียงกัน ก็สามารถใช้แนวทางการทดแทนข้อมูลสูญหายนั้นได้ ดีกว่าต้องทิ้งข้อมูลเหล่านั้นไป

เอกสารอ้างอิง

ธีระพงษ์ กระการดี. (2564). **การวัดแนวโน้มเข้าสู่ส่วนกลาง**. สืบค้นเมื่อ เมษายน 10, 2564, จาก http://www.stvc.ac.th/elearning/stat/csu2.html

ปูเป้ สุดศิลา, อำไพ ทองธีรภาพ และบุญอ้อม โฉมที (2561). **การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระ ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม**. ใน เอกสารประกอบการประชุมวิชาการและนำเสนอผลงานวิชาการระดับชาติ UTCC Academic Day ครั้งที่ 2, มิถุนายน 8, 2561. 1713-1727. กรุงเทพฯ: มหาวิทยาลัยหอการค้าไทย.

พงศกร ธีรรัศมี (2558). **วิธีการหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์.** ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

พัชนา สุวรรณแสน (2562). **การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์.** วารสารวิจัยวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา. 4(1), 1-9.

มนตรี พิริยะกุล (2548). **ตัวแบบการทดแทนข้อมูลในการวิจัยทางสังคมศาสตร์ : การจำลองแบบกรณีตัวแบบอย่างง่าย**. ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง.

สายชล สินสมบูรณ์ทอง. (2563). **การเปรียบเทียบประสิทธิภาพการแทนค่าข้อมูลสูญหายกับการจำแนกกลุ่ม 4 วิธี**. Thai Journal of Science and Technology 9(5), 593-594.

สำนักงานสถิติแห่งชาติ. (2563). **จำนวนประชากรจากการทะเบียน จำแนกตามอายุ เพศ ภาค และจังหวัด ปี พ.ศ. 2563**. สืบค้นเมื่อ เมษายน 30, 2564, จาก http://statbbi.nso.go.th/staticreport/Page/sector/TH/report/sector_01_11101_TH_.xlsx

อานนท์ ศักดิ์วรวิชญ์ (2559). **Mean imputation is the worst method for missing data analysis.** สืบค้นเมื่อ พฤษภาคม 14, 2564. จาก https://businessanalyticsnida.wordpress.com/2016/09/16/mean-imputation-is-the-worst-method-for-missing-data-analysis/

อาวีพร ปานทอง. (2562). **การเปรียบเทียบประสิทธิภาพของวิธีจัดการข้อมูลสูญหายภายใต้สภาวะการสูญหาย ที่ไม่ใช่อย่างสุ่มสาหรับการประมาณค่าพารามิเตอร์ ของแบบสอบถามที่ให้คะแนนหลายค่า**. วารสารวิชาการวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครสวรรค์. 11(14), 18.

Arjun Panesar. (2019). **Machine Learning and AI for Healthcare : Big Data for Improved Health Outcomes** (1st ed.). New York: Apress Media.