Research Article

Received: July 4, 2019; Accepted: July 28, 2019

# การลดจำนวนกลุ่มในการจำแนกแบบหลายกลุ่มเป็นสองกลุ่มสำหรับการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน

# Reducing Multiclass to Binary for Classification Techniques of Re-Hospitalization of Diabetes Patients

วิชญ์วิสิฐ เกษรสิทธิ์*, จิราวัลย์ จิตรถเวช และวิชิต หล่อจีระชุณห์กุล

*คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์*

*ถนนเสรีไทย แขวงคลองจั่น เขตบางกะปิ กรุงเทพมหานคร 10240*

Witwisit Kesornsit*, Jirawan Jitthavech and Vichit Lorchirachoonkul

*Graduate School of Applied Statistics, National Institute of Development Administration,*

*Serithai Road, Khlong Chan, Bangkapi, Bangkok 10240*

**บทคัดย่อ**

การวิจัยครั้งนี้เป็นการวิจัยเพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภทของการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานแบบหลายกลุ่ม (multiclass) หรืออเนกนาม (multinomial) และแบบสองกลุ่มหรือทวิภาค (binary) จำนวน 2 กรณี โดยใช้เทคนิคการวิเคราะห์การถดถอยลอจิสติกและต้นไม้การตัดสินใจ ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลประวัติการรักษาพยาบาลของผู้ป่วยโรคเบาหวานจาก Clinical Care at 130 US Hospitals and Integrated Delivery Networks ตัวแปรเป้าหมายในการจำแนกประกอบด้วยประเภทการนัดหมายให้กลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน จำนวน 3 กลุ่ม คือ ไม่กลับมารักษาซ้ำหรือไม่มีภาวะโรค กลับมารักษาซ้ำภายใน 30 วัน และกลับมารักษาซ้ำมากกว่า 30 วัน ผลการวิจัยพบว่าประสิทธิภาพของการจำแนกประเภทโดยใช้เทคนิคต้นไม้การตัดสินใจแบบทวิภาค จำนวน 2 กรณี มีประสิทธิภาพสูงสุด

**คำสำคัญ :** การถดถอยลอจิสติก; ต้นไม้การตัดสินใจ; การจำแนกประเภทอเนกนาม; การจำแนกประเภทหลายกลุ่ม; การจำแนกประเภททวิภาค; ผู้ป่วยโรคเบาหวาน

**Abstract**

This research intends to compare the classification performance of re-hospitalization of diabetes patients using multiclass or multinomial classification and 2 cases of binary classification in logistic regression and decision tree techniques. The data used in the study are diabetes patients

***ผู้รับผิดชอบบทความ :** witwisit.kes@gmail.com **doi:** 10.14456/tstj.2020.4

from Clinical Care at 130 US Hospitals and Integrated Delivery Networks. The patients were divided into 3 groups, i.e. not re-hospitalization, less than 30 days of re-hospitalization and more than 30 days of re-hospitalization. By comparing the classification techniques, it can be concluded that the classification by decision tree technique using 2 cases of binary classification yields the best result.

**Keywords:** logistic regression; decision tree; multinomial classification; multiclass classification; binary classification; diabetes patient

## 1. บทนำ

โรคเบาหวานเป็นโรคไม่ติดต่อเรื้อรังชนิดหนึ่งที่เป็นสาเหตุหลักของการเสียชีวิตของผู้ป่วยและมีแนวโน้มของผู้ป่วยเพิ่มขึ้นในทุก ๆ ปี จากการศึกษาความชุกของเบาหวานในกลุ่มผู้ใหญ่อายุระหว่าง 20-79 ปี พบว่าจากในปี พ.ศ. 2553-2573 ความชุกของเบาหวานมีแนวโน้มเพิ่มสูงขึ้นจากร้อยละ 6.4 เป็นร้อยละ 7.7 โดยเฉพาะในประเทศที่กำลังพัฒนา เช่น ประเทศไทย [1] การที่ผู้ที่มีความเสี่ยงต่อการเป็นโรคเบาหวาน (pre-diabetes) พัฒนาไปเป็นโรคเบาหวานจะส่งผลกระทบโดยตรงต่อผู้ป่วย ครอบครัว และเศรษฐกิจโดยรวมของประเทศ โดยเฉพาะอย่างยิ่งผลกระทบต่องบประมาณของประเทศที่ใช้ในเรื่องของการรักษาพยาบาล จากภาวะดังกล่าวจะส่งผลให้โรงพยาบาลต้องแบกรับค่าใช้จ่ายในการรักษาพยาบาลเพิ่มขึ้น การศึกษาผลการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานพบว่าปัจจัยดังกล่าวเป็นองค์ประกอบสำคัญที่มีผลต่อการเพิ่มค่าใช้จ่ายทางการแพทย์ (medical expenditure) และคุณภาพของการรักษาผู้ป่วย (quality of care) ที่มีแนวโน้มของอาการที่แย่ลง [2]

สำหรับการเข้ารับการรักษาพยาบาลของผู้ป่วยโรคเบาหวานทางโรงพยาบาลคาดหวังให้ผู้ป่วยมีการกลับมารักษาซ้ำในโรงพยาบาลตามความเหมาะสมกับลักษณะของอาการ ซึ่งจะสามารถช่วยลดค่าใช้จ่ายของโรงพยาบาลในเรื่องการบริหารการจัดบุคลากรทางการแพทย์และสิ่งอำนวยความสะดวกในการให้บริการผู้ป่วย รวมทั้งสามารถช่วยลดค่าใช้จ่ายของภาครัฐในการดูแลรักษาผู้ป่วยตามสิทธิการรักษา [3] การนำปัจจัยเสี่ยงต่อโรคกับกระบวนการรักษาที่แพทย์จัดให้กับผู้ป่วยมาใช้ในการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลจะเป็นกระบวนการที่สำคัญในการบริหารจัดการเรื่องค่าใช้จ่ายของโรงพยาบาล รวมทั้งสามารถใช้ในการวางแผนการรักษาและวางแผนการบริหารจัดการภายในโรงพยาบาลให้เหมาะสม เช่น การบริหารจัดการปริมาณยาในคลัง จำนวนคนไข้ที่เข้ามารับการรักษาในแต่ละวัน นอกจากนี้ยังเป็นกระบวนการในการลดผลกระทบต่อผู้ป่วยทั้งด้านร่างกาย จิตใจ อารมณ์ สังคม และเศรษฐกิจในการเข้ารับการรักษาในโรงพยาบาล [4]

การวิจัยครั้งนี้ผู้วิจัยใช้ข้อมูลผู้ป่วยโรคเบาหวานที่มีการแบ่งกลุ่มการรักษาผู้ป่วยโรคเบาหวานเป็น 3 กลุ่ม คือ กลุ่มที่ไม่กลับมารักษาซ้ำ กลุ่มที่กลับมารักษาซ้ำภายใน 30 วัน และกลุ่มที่กลับมารักษาซ้ำมากกว่า 30 วัน โดยนำปัจจัยเสี่ยงของผู้ป่วยกับกระบวนการรักษาที่จัดให้กับผู้ป่วยมาใช้ในการจำแนกการกลับมารักษาซ้ำในโรงพยาบาล โดยใช้เทคนิคการจำแนกในการสร้างตัวแบบเพื่อจำแนกการมารักษาซ้ำของผู้ป่วยในอนาคต จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากตัวอย่างข้อมูลที่เรียกว่าข้อมูลชุดฝึกฝน (training data set) โดยปัจจุบันเทคนิคการจำแนกมีผู้เสนอแนวคิดและพัฒนาอัลกอริทึมขึ้นเป็น

จำนวนมาก นักวิจัยได้ทดสอบและเปรียบเทียบความสามารถของเทคนิคแต่ละประเภทเพื่อค้นหาวิธีการที่ดีที่สุด ผลการทดสอบส่วนใหญ่ไม่สามารถระบุได้ชัดเจนว่าวิธีการใดเป็นวิธีที่ดีที่สุดในข้อมูลทุกประเภทเพราะข้อมูลแต่ละประเภทมีความเฉพาะตัวที่แตกต่างกัน [5-9] ซึ่งผู้วิจัยได้ศึกษาเปรียบเทียบประสิทธิภาพของวิธีการจำแนกประเภทจำนวน 2 วิธี คือ การวิเคราะห์การถดถอยลอจิสติกและต้นไม้การตัดสินใจ

นอกจากนี้ข้อมูลประวัติการักษาของผู้ป่วยโรคเบาหวานที่ใช้ในการวิจัยครั้งนี้มีจำนวนประเภทการนัดหมายให้กลับมารักษาซ้ำในโรงพยาบาล 3 กลุ่ม คือ ไม่กลับมารักษาซ้ำ กลับมารักษาซ้ำภายใน 30 วัน และกลับมารักษาซ้ำมากกว่า 30 วัน ดังนั้นจึงได้ศึกษาเปรียบเทียบประสิทธิภาพของการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลโดยใช้ตัวแบบการถดถอยลอจิสติกอเนกนาม และแบบทวิภาค สำหรับเทคนิคการจำแนกทั้ง 2 ประเภท [10,11] การกำหนดจำนวนกลุ่มในการจำแนกประเภทจะมีผลต่อประสิทธิภาพของการจำแนก เช่น ความถูกต้องและความไว รวมทั้งมีผลต่อความซับซ้อนในการแปลความหมายผลการจำแนก เช่น การแปลความผลบวกเท็จและผลลบเท็จ

## 2. เทคนิคการจำแนกประเภท

การจำแนกประเภทเป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่เป็นกระบวนการสร้างตัวแบบเพื่อการทำนายการจำแนกค่าข้อมูล (predictive modeling) ให้อยู่ในกลุ่มที่ถูกต้อง โดยใช้ข้อมูลจากตัวอย่างที่เรียกว่าชุดข้อมูลฝึกฝน (training data set) และตรวจสอบตัวแบบที่สร้างขึ้นโดยใช้ชุดข้อมูลตรวจสอบ (validation data set) เมื่อได้สมการจำแนกที่ผ่านเกณฑ์การประเมินจะนำสมการนี้ไปทำนายเพื่อจำแนกประเภทหน่วยตัวอย่างที่เข้ามาใหม่ในอนาคต ซึ่งลักษณะดังกล่าวเรียกว่าการเรียนรู้แบบมีผู้แนะนำการสอน (supervised learning) [12, 13] สำหรับเทคนิคการจำแนกได้มีนักวิจัยคิดค้นและพัฒนาขึ้นมาหลายวิธี ซึ่งเทคนิคการจำแนกแต่ละวิธีมีความสามารถในการจำแนกแตกต่างกัน สำหรับการวิจัยครั้งนี้เลือกใช้เทคนิคการวิเคราะห์การถดถอยลอจิสติกและต้นไม้การตัดสินใจ

### 2.1 การถดถอยลอจิสติก

2.1.1 การถดถอยลอจิสติกทวิภาค

การวิเคราะห์การถดถอยลอจิสติกทวิภาค (binary logistic regression) เป็นการวิเคราะห์การถดถอยที่ตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่า 2 ค่า โดยมีวัตถุประสงค์เพื่อสร้างสมการถดถอยลอจิสติกเพื่อจำแนกตัวแปรตามออกเป็น 2 กลุ่ม โดยอาศัยสารสนเทศจากตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม ซึ่งตัวแบบของการถดถอยลอจิสติกคือ [14,15]

$$P\left[Y_i=1|\,\mathbf{x}_i\right]=\frac{e^{g(\mathbf{x}_i)}}{1+e^{g(\mathbf{x}_i)}}=\pi\left(\mathbf{x}_i\right) \quad (1)$$

เมื่อ $0<\pi(\mathbf{x}_i)<1,\ g(\mathbf{x}_i)=\sum_{k=0}^{p}\beta_k x_{ik},\ -\infty<g(\mathbf{x}_i)<\infty,$ $i=1,2,...,n$ โดยที่ $\beta_k$ เป็นพารามิเตอร์ของตัวแบบ $k=0,1,2,...,p;\ x_{i0}=1$ และ $\pi\left(\mathbf{x}_i\right)$ เรียกว่าการแปลงรูปลอจิสติก และ $g\left(\mathbf{x}_i\right)$ เรียกว่าลอจิต โดยมีข้อสมมุติของตัวแบบการถดถอยลอจิสติกทวิภาค ดังนี้ คือ (1) $y_i$ มีการแจกแจงแบบแบร์นูลลี $y_i\in\{0,1\};i=1,2,...,n$ (2) $Y_i;\ i=1,2,...,n$ เป็นอิสระแก่กัน และ (3) ตัวแปรอิสระไม่มีความสัมพันธ์เชิงเส้นระหว่างกัน $x_{ik};\ i=1,2,...,n;\ k=1,2,...,p$ และ $p$ เป็นจำนวนตัวแปรอิสระในตัวแบบ จากข้อกำหนดจะได้ว่า $E\left[Y_i\right]=1P\left[Y_i=1\right]+0P\left[Y_i=0\right]=P\left[Y_i=1\right]$

ดังนั้น $P\left[Y_i=1|\,\mathbf{x}_i\right]=\frac{e^{g(\mathbf{x}_i)}}{1+e^{g(\mathbf{x}_i)}}$ (2)

$$1-P\left[Y_i=1|\,\mathbf{x}_i\right]=\frac{1}{1+e^{g(\mathbf{x}_i)}} \quad (3)$$

สมการที่ (2) หารด้วยสมการที่ (3) จะได้

$\frac{P[Y_i=1|\mathbf{x}_i]}{1-P[Y_i=1|\mathbf{x}_i]}=e^{g(\mathbf{x}_i)}$ เรียกว่าออดส์ (oods) เมื่อใส่ลอการิทึมทั้งสองข้างจะได้

$ln\left[\frac{P[Y_i=1|\mathbf{x}_i]}{1-P[Y_i=1|\mathbf{x}_i]}\right]=g(\mathbf{x}_i)$ หรือ

$ln\left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right]=g(\mathbf{x}_i)$ เรียกว่าล็อกออดส์ (log odds) หรือลอจิต (logit) [14]

การประมาณพารามิเตอร์ของตัวแบบการถดถอยลอจิสติกที่นิยมใช้คือวิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood estimation) ตัวประมาณค่าที่ได้ทำให้ฟังก์ชันภาวะน่าจะเป็นสูงสุด ซึ่งจากฟังก์ชันภาวะน่าจะเป็นหาอนุพันธ์บางส่วนของฟังก์ชันภาวะน่าจะเป็นเทียบกับตัวพารามิเตอร์ที่ต้องการประมาณค่าเทียบอนุพันธ์บางส่วนที่ได้ให้เท่ากับ 0 สมมุติว่ามีตัวอย่างสุ่ม (random sample) ขนาด $n$ ของ $y_i, i=1,2,...,n$ เมื่อ $y_i$ มีการแจกแจงแบบแบร์นูลลี และความน่าจะเป็นของ $y_i=1$ เท่ากับ $\pi(\mathbf{x}_i)$ จะได้ว่า $\zeta(\mathbf{x}_i)=\pi(\mathbf{x}_i)^{y_i}\left[1-\pi(\mathbf{x}_i)\right]^{1-y_i}$ และฟังก์ชันความน่าจะเป็นคือ

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\zeta(\mathbf{x}_i) \quad (4)$$

$$lnL(\boldsymbol{\beta})=\sum_{i=1}^{n}\left[y_i ln\pi(\mathbf{x}_i)+(1-y_i)ln(1-\pi(\mathbf{x}_i))\right] \quad (5)$$

เมื่อ $\boldsymbol{\beta}'=(\beta_0,\beta_1,...,\beta_k)$ และการหาค่าค่าของ $\boldsymbol{\beta}$ ที่ทำให้ $L(\boldsymbol{\beta})$ มีค่าสูงสุดจะเหมือนกับการหาค่าของ $lnL(\boldsymbol{\beta})$ มีค่าสูงสุดดังนั้นจึงต้องใช้วิธีการหาอนุพันธ์บางส่วนเทียบกับ $\beta_0,\beta_1,...,\beta_k$ คำนวณค่าที่ $\beta_j=b_j;\ j=0,1,2,...,\mathrm{k}$ แล้วกำหนดค่าให้เท่ากับ 0

$$\left.\frac{\partial lnL(\boldsymbol{\beta})}{\partial\beta_0}\right|_{\beta_0=b_0}=\sum_{i=1}^{n}\left[y_i-\pi(\mathbf{x}_i)\right]=0 \quad (6)$$

$$\left.\frac{\partial lnL(\boldsymbol{\beta})}{\partial\beta_j}\right|_{\beta_j=b_j}=\sum_{i=1}^{n}x_{ij}\left[y_i-\pi(\mathbf{x}_i)\right]=0;\ j=1,2,...,k \quad (7)$$

จากสมการที่ (6) และ (7) เรียกว่าภาวะน่าจะเป็นเป็นสมการไม่เชิงเส้น (non-linear) ในพจน์ของ $\beta_j;\ j=0,1,2,...,\mathrm{k}$ ดังนั้นการหาค่า $b_j$ จึงไม่สามารถแก้สมการตามปกติได้จะต้องใช้วิธีการแก้สมการแบบไม่เชิงเส้น [14]

2.1.2 การถดถอยลอจิสติกอเนกนาม

การวิเคราะห์การถดถอยลอจิสติกอเนกนาม (multinomial logistic regression) เป็นการศึกษาอิทธิพลของตัวแปรอิสระที่มีผลต่อตัวแปรตามทำให้ทราบว่าตัวแปรอิสระตัวใดบ้างที่มีความสัมพันธ์กับตัวแปรตามและสร้างสมการจำแนกโดยใช้ตัวแปรอิสระเป็นตัวจำแนกตัวแปรตามซึ่งตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ค่า [14] เมื่อตัวแปรตาม $Y$ เป็นตัวแปรจำแนกประเภทหรือกลุ่มของผลลัพธ์มีจำนวน $J+1$ กลุ่ม คือ $Y=\{0,1,...,J\}$ จะใช้ฟังก์ชันลอจิตจำนวน $J$ ฟังก์ชันและจะกำหนดให้ $Y=0$ เป็นกลุ่มอ้างอิง (reference group) หรือผลลัพธ์พื้นฐาน (baseline outcome) เพื่อสร้างฟังก์ชันลอจิตในการเปรียบเทียบกับประเภทหรือกลุ่มอื่น ๆ อีก $J$ กลุ่ม ในการพัฒนาตัวแบบสมมุติว่ามีตัวแปรอิสระจำนวน $p$ ตัวและพจน์คงตัว (constant term) ดังนั้นเมทริกซ์ $\mathbf{X}$ มีมิติเท่ากับ $p+1$ สามารถเขียนฟังก์ชันลอจิตได้ดังนี้ [11]

$$g_j(\mathbf{x}_i)=ln\left[\frac{\mathrm{P}(Y_{ij}=j|\mathbf{x}_i}{\mathrm{P}(Y_{ij}=0|\mathbf{x}_i}\right]=\beta_{j0}+\sum_{k=1}^{p}\beta_{jk}x_{ki} \quad (8)$$

เมื่อ $j=1,2,...,J;\ i=1,2,...,n;\ k=1,2,...,p$ และ $g_0(\mathbf{x}_i)=0$ และความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) โดยที่กำหนดเงื่อนไขของเวกเตอร์ตัวแปรร่วม (covariate vector) ของแต่ละประเภทของผลลัพธ์ดังนี้

$$\pi_j(\mathbf{x}_i)=P(Y_{ij}=j|\mathbf{x}_i)=\frac{e^{g_j(\mathbf{x}_i)}}{\sum_{j=0}^{J}e^{g_j(\mathbf{x}_i)}};\ j=0,1,2,...,J \quad (9)$$

ดังนั้นฟังก์ชันภาวะน่าจะเป็นแบบมีเงื่อนไข (conditional likelihood function) สำหรับค่าสังเกตที่เป็นอิสระกัน $n$ ตัว คือ

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\left[\pi_0(\mathbf{x}_i)^{y_{0i}}\pi_1(\mathbf{x}_i)^{y_{1i}}\pi_2(\mathbf{x}_i)^{y_{2i}}\ldots\pi_J(\mathbf{x}_i)^{y_{ji}}\right] \quad (10)$$

จากนั้นการแก้สมการโดยการ take log และใช้สมบัติ $\sum_{j=0}^{J} Y_{ij}=1$ ของแต่ละ $i$ จะได้ฟังก์ชันล็อกการิธึมของความเป็นไปได้ (log-likelihood function) คือ

$$L(\boldsymbol{\beta})=\sum_{i=1}^{n}\sum_{j=1}^{J} y_{ji}\left(g_j(\mathbf{x}_i)-ln\left(\sum_{j=0}^{J} e^{g_j(\mathbf{x}_i)}\right)\right. \quad (11)$$

ดังนั้นจะได้สมการความเป็นไปได้ (likelihood) โดยการทำอนุพันธ์บางส่วนอันดับที่ 1 ของ $L(\boldsymbol{\beta})$ เทียบกับพารามิเตอร์ที่มีจำนวน $2J$ ตัวที่ไม่ทราบค่าและกำหนดให้ $\pi_{ji}=\pi_j(\mathbf{x}_i)$ จะได้สมการทั่วไปดังนี้

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}}=\sum_{i=1}^{n} x_{ki}\left(y_{ji}-\pi_{ji}\right) \quad (12)$$

โดยที่ $j=1,2,\ldots,J$ และ $k=0,1,2,\ldots,p$ และ $x_{0i}=1$ สำหรับตัวประมาณค่าภาวะน่าจะเป็นสูงสุด ($\hat{\boldsymbol{\beta}}$) หาค่าได้โดยการกำหนดสมการมีค่าเท่ากับ 0 แล้วแก้สมการหาค่า $\hat{\boldsymbol{\beta}}$ และสำหรับเมทริกซ์ของอนุพันธ์บางส่วนอันดับที่ 2 ต้องการข้อมูลจากตัวประมาณค่าของเมทริกซ์ความแปรปรวนของตัวประมาณค่าภาวะน่าจะเป็นสูงสุด จะได้รูปทั่วไปของอนุพันธ์บางส่วนอันดับที่ 2 ของสมการที่ (13) และ (14) ดังนี้

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jk}\partial \beta_{jk'}}=-\sum_{i=1}^{n} x_{k'i}x_{ki}\pi_{ji}(1-\pi_{ji}) \quad (13)$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jk}\partial \beta_{j'k'}}=\sum_{i=1}^{n} x_{k'i}x_{ki}\pi_{ji}(\pi_{j'i}) \quad (14)$$

สำหรับ $j$ และ $j'=1,2,\ldots,J$ และ $k$ และ $k'=0,1,2,\ldots,p$ จากสมาชิกของเมทริกซ์ค่าสังเกตอินฟอเมชัน (observed information matrix) $\mathbf{I}(\hat{\boldsymbol{\beta}})$ ที่มีมิติ $2(p+1)\times 2(p+1)$ ซึ่งจะมีสมาชิกทุกตัวเป็นลบดังสมการที่ (13) และสมการที่ (14) ใช้ในการประมาณค่า $\hat{\boldsymbol{\beta}}$ ตัวประมาณค่าของเมทริกซ์ความแปรปรวน (covariance matrix) ของตัวประมาณค่าแบบวิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood estimator) คืออินเวอร์สของสมาชิกของเมทริกซ์ค่าสังเกตอินฟอเมชันดังนี้

$$\hat{Var}(\hat{\boldsymbol{\beta}})=\mathbf{I}(\hat{\boldsymbol{\beta}})^{-1} \quad (15)$$

โดยที่ $Y_i$ คือ ตัวแปรตามโดยมีตัวแปรอิสระที่ใช้ในการอธิบายตัวแปลตาม คือ $x_{i1},x_{i2},\ldots,x_{ip}$ ซึ่งจะมีตัวแปรทำนาย $p$ ตัว และ $\beta_{00},\beta_{01},\ldots,\beta_{0p}$ เป็นพารามิเตอร์ของตัวแบบในกลุ่มที่ 0 ซึ่งเป็นกลุ่มอ้างอิงและ $i=1,2,\ldots,n$

**2.2 ต้นไม้การตัดสินใจ**

ต้นไม้การตัดสินใจ (decision tree) เป็นกฎการตัดสินใจเพื่อหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาวิเคราะห์เพื่อสร้างกฎการตัดสินใจในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้แนะนำการสอน ส่วนประกอบของต้นไม้การตัดสินใจประกอบด้วย (1) โหนด (node) คือ สมบัติต่าง ๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใดซึ่งโหนดที่อยู่สูงสุดเรียกว่าโหนดราก (root node) (2) กิ่ง (branch) คือ สมบัติของโหนดที่แตกออกมาโดยจำนวนของกิ่งจะเท่ากับสมบัติของโหนด และ (3) ใบ (leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูลซึ่งโหนดที่อยู่ล่างสุดเรียกว่าโหนดใบ (leaf node) ซึ่งสามารถแสดงส่วนประกอบของต้นไม้ตัดสินใจดังรูปที่ 1 การสร้างโมเดลด้วยวิธีต้นไม้การตัดสินใจจะคัดเลือกตัวแปรที่มีความสัมพันธ์กับกลุ่มมากที่สุดขึ้นมาเป็นโหนดบนสุดของต้นไม้ หลังจากนั้นก็จะหาตัวแปรที่มีความสัมพันธ์ถัดไปเรื่อย ๆ เพื่อสร้างกิ่งและโหนดใบต่อไป ในการหาความสัมพันธ์ของตัวแปรนี้ใช้ตัววัดที่เรียกว่าอัตราส่วนของค่าเกน (information gain, IG) [16]

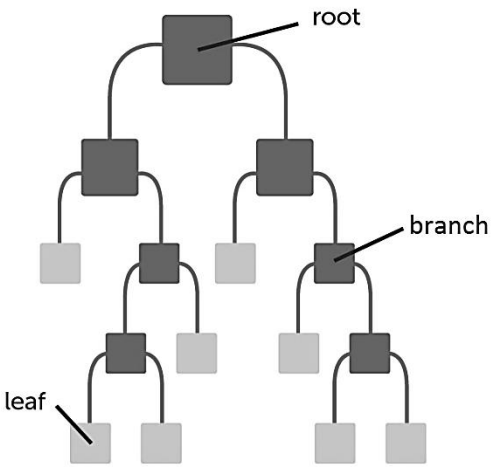


**Figure 1** The components of a decision tree: The Nodes represent the possible attributes associated with an event. The first node is called root and represents the attribute with largest information gain; Branches represent the attributes values; and Leaves represent the classes. [17]

## 3. วิธีดำเนินการวิจัย

ใช้ข้อมูลจาก 130 โรงพยาบาล ในประเทศสหรัฐอเมริกา จำนวน 101,766 ระเบียนของ Clinical Care at 130 US Hospitals and Integrated Delivery Networks [18] ซึ่งเก็บอยู่ที่ UCI Machine Learning Repository (http://archive.ics.uci.edu/ml; Irvine, CA: University of California, School of Information and Computer Science) ตัวแปรที่ใช้ในการวิจัยจำนวน 33 ตัวแปร ประกอบด้วย (1) เพศ จำนวน 2 ระดับ คือ ชาย และหญิง (2) ช่วงอายุจำนวน 4 ระดับ คือ 0-60 ปี 61-70 ปี 71-80 ปี และ 80 ปีขึ้นไป (3) ประเภทของการเข้ารับการรักษาในโรงพยาบาล จำนวน 3 ระดับ คือ ฉุกเฉิน/เร่งด่วน เพิ่งปรากฏว่ามีแนวโน้มการเป็นโรค และประเภทอื่น ๆ (4) ประเภทของการออกจากโรงพยาบาลจำนวน 3 ระดับ คือ กลับบ้าน เข้ารับการรักษาในโรงพยาบาลเดิม และโอนไปยังสถานพยาบาลอื่น ๆ (5) ประเภทของแหล่งหรือแผนกที่เข้ารับการรักษาจำนวน 3 ระดับ คือ ห้องฉุกเฉิน โรงพยาบาลหรือคลินิกอื่น และประเภทอื่น ๆ (6) จำนวนวันที่รักษาตัวในโรงพยาบาล (7) จำนวนของการสั่งตรวจทางห้องปฏิบัติการ (8) จำนวนของการสั่งตรวจรักษานอกเหนือจากการตรวจทางห้องปฏิบัติการ (9) จำนวนยาที่ใช้ในการรักษา (10) จำนวนของการเข้ารับการรักษาเป็นผู้ป่วยนอก (11) จำนวนของการเข้ารับการรักษาเป็นผู้ป่วยฉุกเฉิน (12) จำนวนของการเข้ารับการรักษาเป็นผู้ป่วยใน (13) จำนวนโรคของผู้ป่วยที่ถูกวินิจฉัยเข้ามาในระบบ (14) ผลการทดสอบซีรัมกลูโคส (glucose serum) จำนวน 4 ระดับ คือ ไม่ได้ทดสอบ ปกติ มากกว่าสองร้อย และมากกว่าสามร้อย (15) ผลการทดสอบ A1C จำนวน 4 ระดับ คือ ไม่ได้ทดสอบ ปกติ มากกว่าเจ็ด และมากกว่าแปด (16) การเปลี่ยนแปลงในการใช้ยาเบาหวานทั้งปริมาณหรือชื่อทั่วไปจำนวน 2 ระดับ คือ ไม่มีการเปลี่ยนแปลง และมีการเปลี่ยนแปลง (17) สถานะการใช้ยารักษาโรคเบาหวาน จำนวน 2 ระดับ คือ ใช้ และไม่ใช้ (18-32) การเปลี่ยนแปลงปริมาณยาที่ใช้ต่อหนึ่งครั้ง จำนวน 15 ชนิด ได้แก่ metformin, repaglinide, nateglinide, insulin เป็นต้น จำนวน 3 ระดับ คือ ไม่ได้ใช้ยาในการรักษา (no) ไม่มีการเปลี่ยนแปลงปริมาณยาที่ใช้ต่อหนึ่งครั้ง (steady) และมีการเปลี่ยนแปลงปริมาณยาที่ใช้ต่อหนึ่งครั้ง (up, down) และ (33) สถานะการนัดกลับมารักษาซ้ำ จำนวน 3 ระดับ คือ ไม่นัดกลับมารักษาซ้ำ นัดกลับมารักษาซ้ำภายใน 30 วัน และนัดกลับมารักษาซ้ำมากกว่า 30 วัน และโปรแกรมที่ใช้ในการวิเคราะห์ข้อมูลการวิจัย คือ SAS (statistical analysis system) โดยใช้ SAS 9.4 และ SAS Enterprise Miner 14.2

ข้อมูลทั้งหมดได้สุ่มแบ่งข้อมูลเป็น 2 ชุด คือ ชุดข้อมูลฝึกฝน และชุดข้อมูลตรวจสอบ นำชุดข้อมูลฝึกฝนมาสร้างสมการหรือกฎการตัดสินใจโดยวิธีการถดถอยลอจิสติกและต้นไม้การตัดสินใจที่มีตัวแปรตามเป็นตัวแปรจำแนกประเภทแบบพหุที่มี 3 กลุ่ม และแบบทวิภาคที่มี 2 กลุ่ม ซึ่งแบ่งเป็นกรณีการวิเคราะห์ดังนี้

กรณีที่ 1 จำแนกตัวแปรตามเป็นแบบพหุ 3 กลุ่ม โดยกลุ่มที่ 1 คือ กลุ่มผู้ป่วยที่ไม่ต้องกลับมารักษาซ้ำในโรงพยาบาลหรือผู้ป่วยที่ไม่มีภาวะโรค กลุ่มที่ 2 คือ กลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน และกลุ่มที่ 3 คือ กลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน

กรณีที่ 2 จำแนกตัวแปรตามเป็นแบบทวิภาค โดยกลุ่มที่ 1 คือ กลุ่มผู้ป่วยที่ไม่ต้องกลับมารักษาซ้ำในโรงพยาบาลหรือผู้ป่วยที่ไม่มีภาวะโรค และกลุ่มที่ 2 คือ กลุ่มผู้ป่วยที่ต้องกลับมารักษาซ้ำในโรงพยาบาล

กรณีที่ 3 จำแนกตัวแปรตามเป็นแบบทวิภาค จากข้อมูลในกลุ่มที่ 2 ของกรณีที่ 2 โดยกลุ่มที่ 1 คือ กลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน และกลุ่มที่ 2 คือ กลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน

## 4. ผลการดำเนินการวิจัย

ข้อมูลผู้ป่วยโรคเบาหวานที่ใช้ในการวิจัยประกอบด้วยกลุ่มเป้าหมาย จำนวน 3 กลุ่ม โดยมีอัตราส่วนของข้อมูลระหว่างกลุ่มไม่กลับมารักษาซ้ำในโรงพยาบาล กลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน และกลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน เป็นร้อยละ 53.92 : 11.16 : 34.92 ตามลำดับ เนื่องด้วยการจำแนกประเภทโดยกลุ่มเป้าหมาย จำนวน 3 กลุ่ม จะมีผลต่อประสิทธิภาพของการจำแนกเช่นความถูกต้องและความไว รวมทั้งมีผลต่อความซับซ้อนในการแปลความหมายผลการจำแนก เช่น การแปลความผลบวกเท็จและผลลบเท็จ [19] ผู้วิจัยจึงได้ออกแบบการวิจัยโดยทำการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานเป็น 2 กรณี แบ่งเป็นกรณีละ 2 กลุ่ม โดยกรณีที่ 1 แบ่งเป็นไม่กลับมารักษาซ้ำในโรงพยาบาล และกลับมารักษาซ้ำในโรงพยาบาลมีอัตราส่วนของข้อมูลเป็นร้อยละ 53.92 : 46.08 ตามลำดับ และกรณีที่ 2 แบ่ง เป็นกลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน และกลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน มีอัตราส่วนของข้อมูลเป็น 24.22 : 75.78 ตามลำดับ แสดงดังตารางที่ 1

**Table 1** The number and percentage of re-hospitalization in diabetes patients

| Target group | Not re-hospitalization | Less than 30 days of re-hospitalization | More than 30 days of re-hospitalization |
|---|---|---|---|
| Case 1 | 54,850 (53.92) | 11,356 (11.16) | 35,528 (34.92) |
| Target group | Not re-hospitalization | Re-hospitalization | |
| Case 2 | 54,850 (53.92) | 46,884 (46.08) | |
| Target group | - | Less than 30 days of re-hospitalization | More than 30 days of re-hospitalization |
| Case 3 | - | 11,356 (24.22) | 35,528 (75.78) |

Values in parentheses are the percentage value.

**Table 2** Classification of re-hospitalization of diabetes patients in case 1

| Classification techniques | False positive | False negative | Sensitivity | | Accuracy | Variable selection | ROC | Lift |
|---|---|---|---|---|---|---|---|---|
| | | | ≤ 30 | >30 | | | | |
| Case 1 | | | | | | | | |
| Logistic regression | 0.16 | 34.95 | 1.06 | 24.00 | 57.09 | 4 | 0.61 | 1.41 |
| Decision tree | 0.19 | 34.96 | 1.41 | 36.20 | 57.60 | 16 | 0.62 | 1.44 |

Lift value is the highest value from the lift chart.

**Table 3** Classification of re-hospitalization of diabetes patients in case 2 and 3

| Classification techniques | False positive | False negative | Sensitivity | Accuracy | Variable selection | ROC | Lift |
|---|---|---|---|---|---|---|---|
| Case 2 | | | | | | | |
| Logistic regression | 9.52 | 28.98 | 64.65 | 61.51 | 10 | 0.65 | 1.50 |
| Decision tree | 12.43 | 25.22 | 62.99 | 62.35 | 14 | 0.66 | 1.52 |
| Case 3 | | | | | | | |
| Logistic regression | 24.19 | 0.40 | 75.62 | 75.41 | 4 | 0.55 | 1.03 |
| Decision tree | 24.18 | 0.41 | 75.62 | 75.42 | 7 | 0.59 | 1.08 |

Lift value is the highest value from the lift chart.

เมื่อพิจารณาประสิทธิภาพในการจำแนกประเภทของการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานที่ตัวแปรตามมีจำนวน 3 กลุ่ม พบว่าประสิทธิภาพของการจำแนกประเภทโดยใช้เทคนิคต้นไม้การตัดสินใจจะมีประสิทธิภาพสูงสุดโดยให้ค่าความถูกต้อง (accuracy) เท่ากับร้อยละ 57.60 ซึ่งมีค่ามากที่สุด ค่าความไวในการจำแนกประเภท (sensitivity) ทั้งข้อมูลที่อยู่ในกลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน และกลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน เท่ากับร้อยละ 1.41 และ 36.20 ตามลำดับ รวมทั้งมีค่า ROC index เท่ากับ 0.62 และค่า lift เท่ากับ 1.44 ซึ่งมีค่ามากที่สุด ซึ่งเมื่อพิจารณาค่าความไวในการจำแนกกลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน จะคำนวณโดยใช้ผลรวมของกลุ่มผู้ป่วยที่ไม่กลับมารักษาซ้ำในโรงพยาบาล และกลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน เป็นกลุ่มอ้างอิง และค่าความไวในการจำแนกกลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน จะคำนวณโดยใช้ผลรวมของกลุ่มผู้ป่วยที่ไม่กลับมารักษาซ้ำในโรงพยาบาล และกลุ่มผู้ป่วยที่กลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน เป็นกลุ่มอ้างอิง ซึ่งในกรณีดังกล่าวจะทำให้ผู้วิจัยไม่สามารถทำการแปลความหมายค่าความไวในการจำแนกกลุ่มอ้างอิง ซึ่งค่าผลบวกเท็จและผลลบเท็จก็เป็นไปตามเหตุผลดังกล่าวเช่นเดียวกัน แสดงผลการวิเคราะห์ดังตารางที่ 2

เมื่อพิจารณาประสิทธิภาพในการจำแนกประเภทของการกลับมารักษาซ้ำในโรงพยาบาลของ

ผู้ป่วยโรคเบาหวานจำนวน 2 กรณีแบ่งเป็นกรณีละ 2 กลุ่ม โดยขั้นตอนในการจำแนกจะจำแนกในกรณีที่ 2 เป็นขั้นตอนแรก หากผลการจำแนกปรากฏว่าเป็นผู้ป่วยที่ไม่กลับมารักษาซ้ำในโรงพยาบาลจะสามารถตัดสินใจตามผลการจำแนกโดยทันที แต่หากผลการจำแนกในกรณีที่ 2 ปรากฏว่าเป็นกลับมารักษาซ้ำในโรงพยาบาลจะต้องจำแนกประเภทต่อในกรณีที่ 3 โดยจำแนกประเภทเป็นกลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน และกลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน ซึ่งผลการวิจัยพบว่าประสิทธิภาพของการจำแนกประเภทโดยใช้เทคนิคต้นไม้การตัดสินใจจะมีประสิทธิภาพสูงสุดในทั้ง 2 กรณี โดยให้ค่าความถูกต้องเท่ากับร้อยละ 62.35 และ 75.42 ตามลำดับ ซึ่งมีค่ามากที่สุด ความไวในการจำแนกประเภทเท่ากับร้อยละ 62.99 และ 75.62 ตามลำดับ รวมทั้งมีค่า ROC index เท่ากับ 0.66 และ 0.59 ตามลำดับ และค่า lift เท่ากับ 1.52 และ 1.08 ตามลำดับ ซึ่งมีค่ามากที่สุด แสดงดังตารางที่ 3

## 5. สรุปผลการวิจัย

การศึกษาเปรียบเทียบประสิทธิภาพในการจำแนกประเภทของการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานแบบอเนกนามหรือหลายกลุ่มและแบบทวิภาค ซึ่งจำแนกเป็น 2 กรณี โดยใช้เทคนิคการวิเคราะห์การถดถอยลอจิสติกและต้นไม้การตัดสินใจ พบว่าประสิทธิภาพของการจำแนกประเภทโดยใช้เทคนิคต้นไม้การตัดสินใจแบบทวิภาคที่จำแนกเป็น 2 กรณี มีประสิทธิภาพสูงสุด สอดคล้องกับงานวิจัยของ Jelinek (2013) เรื่อง Decision trees and multi-level ensemble classifiers for neurological diagnostics [20] และงานวิจัยของ Bihis (2015) เรื่อง A generalized flow for multi-class and binary classification tasks: An Azure ML approach [21] ซึ่งการจำแนกประเภทของการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานที่ตัวแปรตามมีจำนวน 3 กลุ่ม คือ ไม่กลับมารักษาซ้ำในโรงพยาบาล กลับมารักษาซ้ำในโรงพยาบาลภายใน 30 วัน และกลับมารักษาซ้ำในโรงพยาบาลมากกว่า 30 วัน ส่งผลให้แปลความหมายของค่าในเมทริกซ์ได้ยากและเกิดความสับสน (confusion matrix) เนื่องจากตัววัดดังกล่าวนั้นคำนวณภายใต้เงื่อนไขแบบทวิภาค เมื่อมีจำนวนกลุ่มมากกว่า 2 กลุ่ม ตัววัดจะกำหนดกลุ่มที่สนใจขึ้นมาหนึ่งกลุ่ม และจำนวนข้อมูลในกลุ่มที่เหลือซึ่งรวมเป็นกลุ่มเดียวกันนั้นกำหนดเป็นกลุ่มอ้างอิง ตัวอย่าง เช่น กรณีกลุ่มที่สนใจเป็นกลุ่มที่ 2 การคำนวณค่าโดยกำหนดให้จำนวนผลการจำแนกในกลุ่มที่ 2 เป็นกลุ่มที่สนใจ และจำนวนผลการจำแนกในกลุ่มที่เหลือที่จะรวมเป็นกลุ่มเดียวกันและกำหนดเป็นกลุ่มอ้างอิง ดังนั้นในการอ่านค่าจากตัววัดดังกล่าวจึงทำให้ยากต่อการตีความผลของการจำแนกผิดพลาด [22]

ดังนั้นการเลือกใช้การจำแนกประเภทข้อมูลแบบสองกลุ่มหรือทวิภาค จึงเป็นเทคนิคการจำแนกที่ทำให้แปลความหมายของค่าในเมทริกซ์สับสนได้ง่าย เนื่องจากกลุ่มของตัวแปรตามมีจำนวน 2 กลุ่ม มีการคำนวณค่าของตัววัดดังกล่าวโดยการกำหนดกลุ่มใดกลุ่มหนึ่งเป็นกลุ่มอ้างอิง ทำให้การอ่านค่าจากตัววัดสามารถใช้เป็นตัวบ่งชี้หรือเกณฑ์การประเมินประสิทธิภาพของตัวแบบหรือผลการตรวจวินิจฉัยอย่างเหมาะสม โดยเฉพาะอย่างยิ่งงานวิจัยทางการแพทย์ที่นิยมใช้ค่าจากเมทริกซ์ความสับสนในการแปลผลการวินิจฉัยโรคหรือการวินิจฉัยภาวะผิดปกติต่าง ๆ ด้วยการตรวจทางห้องปฏิบัติการ การรายงานผลตรวจวิธีการหนึ่ง คือ รายงานว่าผลตรวจเป็นบวกหรือผลตรวจเป็นลบ โดยผลการตรวจย่อมมีโอกาสเกิดผลบวกเท็จ (false positive) หรือผลลบเท็จ (false negative) ขึ้นได้เสมอ การแปลความผลจึงต้องระมัดระวังเป็น

อย่างยิ่ง และต้องอาศัยความรู้ทางสถิติและทฤษฎีความน่าจะเป็น ค่าผลบวกเท็จหรือผลลบเท็จจึงเป็นตัวชี้วัดที่มีความสำคัญในการพิจารณาผลการวิจัยหรือผลการวินิจฉัยโรค ซึ่งจะมีผลต่อประสิทธิภาพการรักษาที่ถูกต้องและเหมาะสมอีกด้วย [23,24]

## 6. References


[1] Shaw, J.E., Sicree, R.F. and Zimmet, P.Z., 2010, Global estimates of the prevalence of diabetes for 2010 and 2030, Diabetes Res. Clin. Pract. 87: 4-14.

[2] Dungan, K.M., 2012, The effect of diabetes on hospital readmissions, J. Diabetes. Sci. Technol. 6: 1045-1052.

[3] Bradley, E.H., Curry, L., Horwitz, L.I., Sipsma, H., Wang, Y., Walsh, M.N., Goldmann, D., White, N., Piña, I.L. and Krumholz, H.M., 2013, Hospital strategies associated with 30-day readmission rates for patients with heart failure, Circ. Cardiovasc. Qual. Outcomes 6: 444-450.

[4] Choowattanapakorn, T., Karuna, R., Konghan, S. and Tangmettajittakun, D., 2016, Factors predicting quality of life in older people with diabetes in Thailand, Songklanakarin J. Sci. Technol. 38: 707-714.

[5] Anitha, J. and Pethalakshmi, A., 2017, Comparison of classification algorithms in diabetic dataset, The 5 th National Conference on Computational Methods, Communication Techniques and Informatics Department of Computer Science & Applications, The Gandhigram Rural Institute-Deemed University, Tamil Nadu.

[6] Chen, H., Zhang, J., Xu, Y., Chen, B. and Zhang, K., 2012, Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans, J. Med. Eng. Technol. 39: 11503-11509.

[7] Dogan, N. and Tanrikulu, Z., 2013, A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness, Inf. Technol. Manage. 14: 105-124.

[8] Eftekhar, B., Mohammad, K., Ardebili, H. E., Ghodsi, M. and Ketabchi, E., 2005, Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data, BMC Med. Inform. Decis. Mak. 5: 1-8.

[9] Thangaraju, P., Deepa, B. and Karthikeyan, T., 2014, Comparison of data mining techniques for forecasting diabetes mellitus, IJARCSEE 3: 7674-7677.

[10] Srinivas, M., Bharath, R., Rajalakshmi, P. and Mohan, C.K., 2015, Multi-level classification: A generic classification method for medical datasets, The 17th International Conference on E-health Networking, Application & Services (HealthCom).

[11] Hosmer, Jr.D.W., Lemeshow, S. and Sturdivant, R.X., 2013, Applied Logistic

Regression, John Wiley & Sons, New York, 398 p.

[12] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013, An Introduction to Statistical Learning with Applications in R, Springer, 426 p.

[13] Cichosz, P., 2015, Data mining algorithms: Explained using R, Chichester, West Sussex, Wiley, Chichester, West Sussex, Wiley, 683 p.

[14] Jitthavech, J., 2015, Regression Analysis, Bangkok, National Institute of Development Administration, 433 p. (in Thai)

[15] Maiprasert, D., 2013, Prediction of breast cancer stage using multinomial logistic regression and artificial neural network, Ph.D. Thesis, Rangsit University, Bangkok, 215 p. (in Thai)

[16] Pacharawongsakda, E., 2014, An Introduction to Data Mining Techniques. Bangkok, Asia Digital Press Co., Ltd., 124 p. (in Thai)

[17] Grisanti, J., Decision Trees: An Overview. Available Source: https://www.aunalytics.com/2015/01/30/decision-trees-an-overview, June 9, 2015.

[18] Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J. and Clore, J.N., 2014, Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records, Biomed. Res. Int. 2014: 1-11.

[19] Cotha, N.K.P. and Sokolova, M., 2015, Multi-Labeled Classification of Demographic Attributes of Patients: A Case Study of Diabetics Patients, pp. 1 - 16, Cornall University Library, New York.

[20] Jelinek, H., Abawajy, J., Kelarev, A., Chowdhury, M. and Stranieri, A., 2013, Decision trees and multi-level ensemble classifiers for neurological diagnostics, AIMS Med. Sc. 1: 1-12.

[21] Bihis, M. and Roychowdhury, S., 2015, A generalized flow for multi-class and binary classification tasks: An Azure ML approach, pp. 1728-1737, The 2015 IEEE International Conference on Big Data.

[22] Beleites, C., Salzer, R. and Sergo, V., 2013, Validation of soft classification models using partial class memberships: An extended concept of sensitivity & co. applied to grading of astrocytoma tissues, Chemometr. Intell. Lab. Syst. 122: 12-22.

[23] Ebrahimi, N., 2008, Simultaneous control of false positives and false negatives in multiple hypotheses testing, J. Multivar. Anal. 99: 437-450.

[24] Zhao, C., Crothers, B.A., Tabatabai, Z.L., Li, Z., Ghofrani, M., Souers, R.J., Husain, M., Fan, F., Shen, R. and Ocal, I.T., 2017, False-negative interpretation of adenocarcinoma *in situ* in the college of American Pathologists Gynecologic PAP Education Program, Arch. Pathol. Lab. Med. 141: 666-670.