

## บทความวิจัย (Research Article)

# การเปรียบเทียบตัวแบบการจำแนกการเกิดโรคหลอดเลือดสมองของผู้สูงอายุ: กรณีศึกษาโรงพยาบาลสมเด็จพระปิ่นเกล้า

พรทิพย์ เดชพิชัย<sup>1\*</sup>, ธัญพิชชา สัตตบรรณ<sup>1</sup>, รัตนา เมฆวัน<sup>1</sup> และ อภิษฐา อรัญยะपाल<sup>1</sup>

## A comparison of classification models for the stroke among elderly: A case study of Somdech Phra Pinklao Hospital

Porntip Dechpichai<sup>1\*</sup>, Thunpitcha Sattabun<sup>1</sup>, Rattana Mekwan<sup>1</sup> and Apittha Arunyapal<sup>1</sup>

<sup>1</sup> Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi,  
Thung Khru, Bangkok, 10140

\* Corresponding author : porntip.dec@kmutt.ac.th

**Naresuan Phayao J. 2023;16(1): 56-70.**

*Received: 1 January 2022; Revised: 16 Jun 2022; Accepted: 16 November 2022*

### บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการจำแนกและศึกษาปัจจัยที่เกี่ยวข้องกับการเป็นโรคหลอดเลือดสมองในผู้สูงอายุ โรงพยาบาลสมเด็จพระปิ่นเกล้า โดยใช้ข้อมูลเวชระเบียนของผู้สูงอายุที่มีอายุตั้งแต่ 60 ปีขึ้นไป และเข้ารับการรักษาในโรงพยาบาลในปีพ.ศ. 2561 จำนวน 28,928 คน และเนื่องจากข้อมูลไม่สมดุลจึงใช้เทคนิคสังเคราะห์ข้อมูลเพิ่ม (SMOTE) เพิ่มข้อมูลให้มีสัดส่วนใกล้เคียงกัน แล้วแบ่งข้อมูลเป็น 2 ชุด ชุดที่ 1 (80%) สำหรับสร้างตัวแบบด้วยวิธีการการถดถอยลอจิสติกทวิภาค (glm) และต้นไม้ตัดสินใจแบบ ID3, CART, J48, CTREE และ C5.0 ร่วมกับ Bootstrap Aggregating (Bagging) และชุดที่ 2 (20%) สำหรับประเมินประสิทธิภาพของตัวแบบ ผลการวิจัยพบว่า อัตราความชุกของผู้ที่เป็นโรคหลอดเลือดสมองเป็น 5.50% (95% CI 5.24% - 5.76%) และตัวแบบที่มีประสิทธิภาพดีที่สุด คือ ตัวแบบต้นไม้ตัดสินใจแบบ C5.0 โดยมีค่าความแม่นยำ ร้อยละ 95.31 ค่าความไว ร้อยละ 94.48 ค่าความจำเพาะ ร้อยละ 96.12 ค่าทำนายผลบวก ร้อยละ 95.93 และค่าทำนายผลลบ ร้อยละ 94.73 และจากตัวแบบต้นไม้ตัดสินใจแบบ C5.0 พบว่าปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคหลอดเลือดสมองของผู้สูงอายุเรียงตามลำดับความสำคัญคือ ภาวะหลอดเลือดสมองครั้งคราว อายุ ภาวะโลหิตจาง โรคลมชัก การสูบบุหรี่ ภาวะแข็งตัวของเลือดผิดปกติและเลือดออก การบาดเจ็บที่ศีรษะ กลุ่มโรคหัวใจ โรคเบาหวาน การดื่มแอลกอฮอล์ โรคไต การมีอุปกรณ์ฝังและปลูกถ่ายของหัวใจและหลอดเลือด เพศ โรคความดันโลหิต โรคเบาหวาน ค่าดัชนีมวลกาย กลุ่มโรคของหลอดเลือดแดง หลอดเลือดแดงย่อย และหลอดเลือดฝอย และโรคกล้ามเนื้อหัวใจตายที่ปอด ส่วนปัจจัยที่ไม่ได้ใช้ในการจำแนกผู้ป่วยโรคหลอดเลือดสมอง คือ กลุ่มโรคอ้วนและภาวะโภชนาการเกิน และภาวะเผาผลาญไขมันผิดปกติ

**คำสำคัญ:** โรคหลอดเลือดสมอง, การถดถอยลอจิสติกทวิภาค, ต้นไม้ตัดสินใจ, ข้อมูลไม่สมดุล

<sup>1</sup> ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี เขตทุ่งครุ กรุงเทพมหานคร 10140

## Abstract

The objective of this paper is to compare classification modes and study factors associated for stroke prediction in the elderly in Somdech Phra Pinklao Hospital, Bangkok, Thailand. The personal medical records of elderly patients who are over 60-year-old and visit the hospital in 2018, total 28,928 patients have been collected and preprocessed. Because of imbalance data, over-sampling technique is used to increase smaller group size. Then they have been partitioned into two groups. The former (80%) is used to construct models, which are the stepwise binary logistic regression (glm) and decision tree models (ID3, CART, J48, CTREE and C5.0) with Bootstrap Aggregating (Bagging). While the latter (20%) is used to evaluate the accuracy of the model. The result shows that the prevalence rate of stroke patients is 5.50% (95% CI 5.24% -5.76%). The most effective model is the C5.0 decision tree model with the accuracy of 95.31 percent, sensitivity of 94.48 percent, specificity of 96.12 percent, the positive prediction value of 95.93 percent and the negative prediction value of 94.73 percent. Using the C5.0 decision tree model, the important risk factors effecting on the stroke of the elderly by order are Transient ischemic attack, Age, Anemia, Epilepsy, Smoking, Clotting disorder and bleeding, Head injury, Heart disease, Cancer, Drinking alcohol, Kidney disease, The presence of implants and implants for the heart and blood vessels, Sex, Hypertension, Diabetes, Body mass index, Disorders of arteries, arterioles and capillaries, and Pulmonary embolism. While Overweight, obesity & hypernutrition and Metabolic disorder are not included the model to classify the stroke of the elderly.

**Keywords:** Stroke, Binary logistic regression, Decision trees, Imbalanced data

## บทนำ

การสำรวจสถิติประชากรในประเทศไทยปี พ.ศ. 2563 พบว่า มีผู้สูงอายุจำนวน 11,627,130 คน จากจำนวนประชากรทั้งสิ้น 66,186,727 คน คิดเป็น ร้อยละ 17.57 (<http://www.dop.go.th/th/knownside/1/1/335>) แสดงให้เห็นว่าประเทศไทยกำลังอยู่ใน สถานการณ์สังคมสูงวัย (aged society) การเพิ่มขึ้น ของประชากรผู้สูงอายุสามารถนำไปสู่ปัญหาค่าใช้จ่าย ของรัฐและครอบครัวในการดูแลสุขภาพของผู้สูงอายุ เนื่องจากการเปลี่ยนแปลงทางสภาพร่างกายของ ผู้สูงอายุทำให้ประสิทธิภาพการทำงานของระบบต่าง ๆ ลดลง ก่อให้เกิดปัญหาสุขภาพ รวมทั้งเกิดความ เปราะบางด้านจิตใจ รู้สึกว่าชีวิตไม่มีคุณค่า และมีความ พึงพอใจในชีวิตลดลงมากกว่าคนวัยอื่น

นอกจากนี้การเปลี่ยนแปลงวิถีชีวิตและ พัฒนาการทางวิทยาศาสตร์การแพทย์ในปัจจุบันได้ ส่งผลให้ภาวะการเจ็บป่วยของผู้สูงอายุเปลี่ยนจากการ เจ็บป่วยแบบเฉียบพลันเป็นการเจ็บป่วยแบบเรื้อรัง จากรายงานพบว่าโรคเรื้อรังที่ทำให้ผู้สูงอายุไทย

สูญเสียปีสุขภาวะ (Disability adjusted life year) เป็น อันดับ 1 คิดเป็นร้อยละ 11.2 คือ โรคหลอดเลือดสมอง หรือ Stroke หรือ Cerebrovascular disease (CVD) [1] และจากข้อมูลย้อนหลังพบว่าจำนวนผู้ป่วยโรค หลอดเลือดสมอง ตั้งแต่ปี พ.ศ. 2559-2561 มีแนวโน้ม สูงขึ้น และจำนวนผู้เสียชีวิตจากโรคหลอดเลือดสมองปี ละประมาณ 30,000 คน ซึ่งมีอัตราการตายสูงที่สุดใน ระหว่าง 5 โรคไม่ติดต่อหลัก (โรคความดันโลหิตสูง โรคเบาหวาน โรคหัวใจขาดเลือด และโรคหลอดเลือด อักเสบ ถุงลมโป่งพอง) [2] และโรคหลอดเลือดสมองยัง เป็นสาเหตุ หลักการเสียชีวิตของคนไทยเป็นอันดับที่ 2 รองจากโรคมะเร็ง โดยในปี พ.ศ. 2558 และ พ.ศ. 2562 มีผู้เสียชีวิตจากโรคนี้ คิดเป็นร้อยละ 43.3 และ 53.0 ตามลำดับ [3] ซึ่งสอดคล้องกับองค์การอนามัยโลก (World Stroke Organization) ที่รายงานว่า โรคหลอดเลือด สมองเป็นสาเหตุสำคัญของการเสียชีวิตเป็น อันดับ 2 ของโลก โดยพบว่าทั่วโลกมีผู้ป่วยโรคหลอดเลือด สมองจำนวน 80 ล้านคน พิจารณาจากโรคหลอดเลือด สมองจำนวน 50 ล้านคน และเสียชีวิตประมาณ 5.5 ล้านคน นั่นคือเฉลี่ยทุก ๆ 6 วินาที จะมีคนตาย

ด้วยโรคหลอดเลือดสมองอย่างน้อย 1 คน และในแต่ละปีคาดว่าจะมีผู้ป่วยใหม่ประมาณ 13.7 ล้านคน และจากการประมาณความเสี่ยงของการเกิดโรคหลอดเลือดสมองในประชากรโลกปี พ.ศ. 2562 พบว่า ทุก ๆ 4 คน จะป่วยด้วยโรคหลอดเลือดสมอง 1 คน นั่นคือประมาณร้อยละ 80 ของประชากรโลกที่มีความเสี่ยงสามารถป้องกันโรคนี้ได้ และเนื่องจากการเปลี่ยนแปลงสภาพแวดล้อมและเทคโนโลยีต่าง ๆ ในทุกวันนี้ อาจส่งผลต่อการเปลี่ยนแปลงปัจจัยเสี่ยงต่าง ๆ ที่จะส่งผลต่อการเกิดโรคหลอดเลือดสมองได้

ตัวแบบที่มักจะนิยมใช้สำหรับการศึกษาความสัมพันธ์ของปัจจัยต่าง ๆ กับตัวแปรตามที่น่าสนใจในงานวิจัยด้านวิทยาศาสตร์สุขภาพและสาธารณสุข ซึ่งส่วนใหญ่เป็นข้อมูลมาตามาตรานามบัญญัติที่มักใช้สถิติการแจกแจง นั่นคือ ตัวแบบการถดถอยลอจิสติกทวิภาค [4] และ เทคนิคต้นไม้ตัดสินใจ [5], [6], [7] และเนื่องจากข้อมูลผู้ป่วยสูงอายุที่เป็นและไม่เป็นโรคหลอดเลือดสมองมีจำนวนในแต่ละกลุ่มแตกต่างกันมาก ถือว่าเป็นข้อมูลไม่สมดุล โดยข้อมูลกลุ่มน้อย (Minority class) คือข้อมูลที่มีจำนวนน้อยในชุดข้อมูล และข้อมูลกลุ่มมาก (Majority class) คือข้อมูลที่มีจำนวนมากในชุดข้อมูล ถ้านำข้อมูลประเภทนี้ไปสร้างตัวแบบจะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาดได้ เพราะตัวแบบมักจะทำนายข้อมูลกลุ่มน้อยให้เป็นข้อมูลกลุ่มมาก ซึ่งเป็นการจำแนกข้อมูลผิดกลุ่ม (Misclassification) คือ จำแนกข้อมูลกลุ่มมากได้ถูกต้อง ในขณะที่จำแนกข้อมูลกลุ่มน้อยไม่ถูกต้อง แต่ความถูกต้องของการทำนายมีค่าสูง [8] ดังนั้นจึงควรจะมีการนำเทคนิคการสังเคราะห์ข้อมูลเพิ่มด้วยวิธีสุ่มมาช่วยปรับสมดุลของข้อมูล เพื่อเพิ่มประสิทธิภาพให้กับตัวแบบ [7], [9], [10], [11], [12]

ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อ

- 1) เปรียบเทียบตัวแบบการถดถอยลอจิสติกทวิภาค และต้นไม้ตัดสินใจร่วมกับ Bootstrap Aggregating (Bagging) และเทคนิคการสุ่มเกิน (Over-sampling) ในการปรับข้อมูลให้เหมาะสม สำหรับการจำแนกการเกิดโรคหลอดเลือดสมองของผู้สูงอายุ และ 2) เพื่อศึกษาปัจจัยที่เกี่ยวข้องกับการเกิดโรคหลอดเลือดสมองของผู้สูงอายุที่เข้ารับบริการที่โรงพยาบาลสมเด็จพระปิ่นเกล้าจากตัวแบบที่เลือก เพื่อใช้เป็นข้อมูลสำหรับองค์กรและหน่วยงานต่าง ๆ ที่อาจจะนำปัจจัยที่ศึกษาไปใช้ในการตรวจคัดกรองผู้สูงอายุที่มีความเสี่ยงได้อย่างแม่นยำ และสามารถที่จะนำไปใช้เป็นแนวทางพัฒนาระบบบริการทางด้านโรคหลอดเลือดสมอง (Stroke service plan) เพื่อเพิ่มการเข้าถึงบริการที่มีคุณภาพและเพื่อลดอัตราการตาย ลดภาวะแทรกซ้อน การกลับเป็นซ้ำ และลดความพิการในผู้สูงอายุสำหรับโรคหลอดเลือดสมองได้อย่างมีประสิทธิภาพ โดยโครงการวิจัยได้ผ่านการรับรองจริยธรรมการวิจัยจากคณะกรรมการพิจารณาจริยธรรมวิจัยในมนุษย์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี หมายเลขการรับรอง KMUTT-IRB-2020/0114/054

## วิธีดำเนินการวิจัย

### ข้อมูลและแหล่งข้อมูล

ข้อมูลที่ใช้ศึกษาครั้งนี้เป็นข้อมูลจากเวชระเบียนผู้ป่วยที่มีอายุตั้งแต่ 60 ปีขึ้นไป ที่เข้ารับบริการที่โรงพยาบาลสมเด็จพระปิ่นเกล้าระหว่างวันที่ 1 มกราคม - 31 ธันวาคม 2561 ผู้วิจัยใช้ Microsoft Excel ในการเตรียมและทำความสะอาดข้อมูลที่ไม่สมบูรณ์และมีค่าสูญหาย จึงได้ข้อมูลจำนวน 28,928 ระเบียน (คน) ที่มีความถูกต้องครบถ้วนและอยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ โดยมีตัวแปรตามและปัจจัยที่ต้องการศึกษาซึ่งโรงพยาบาลได้บันทึกไว้ในเวชระเบียนของผู้ป่วยทั้งหมด 20 ปัจจัย ดังตารางที่ 1

ตารางที่ 1 รายละเอียดตัวแปรตามและคุณลักษณะที่นำมาศึกษา

| ตัวแปร | คุณลักษณะ                                                | มาตรวัดข้อมูล | ค่าที่เป็นไปได้                    |
|--------|----------------------------------------------------------|---------------|------------------------------------|
| sex    | เพศ                                                      | นามบัญญัติ    | 1: ชาย, 2: หญิง                    |
| age    | อายุ                                                     | อัตราส่วน     |                                    |
| BMI    | ค่าดัชนีมวลกาย                                           | อันตรภาค      |                                    |
| cigar  | การสูบบุหรี่                                             | นามบัญญัติ    | 0: ไม่สูบ, 1: สูบ, 2: บุหรี่มือสอง |
| drink  | การดื่มแอลกอฮอล์                                         | นามบัญญัติ    | 0: ไม่ดื่ม, 1: ดื่ม                |
| x1     | กลุ่มโรคมะเร็ง                                           | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x2     | ภาวะโลหิตจาง                                             | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x3     | ภาวะแข็งตัวของเลือดผิดปกติ และเลือดออก                   | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x4     | โรคเบาหวาน                                               | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x5     | กลุ่มโรคอ้วนและภาวะโภชนาการเกิน                          | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x6     | ภาวะเผาผลาญไขมันผิดปกติ                                  | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x7     | โรคลมชัก                                                 | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x8     | ภาวะหลอดเลือดสมองครั้งคราว                               | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x9     | กลุ่มโรคหัวใจ                                            | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x10    | โรคความดันโลหิต                                          | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x11    | โรคลิ้นเลือดอุดตันที่ปอด                                 | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x12    | กลุ่มโรคของหลอดเลือดแดง หลอดเลือดแดงย่อย และหลอดเลือดฝอย | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x13    | โรคไต                                                    | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x14    | การบาดเจ็บที่ศีรษะ                                       | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| x15    | การมีอุปกรณ์ฝังและปลุกถ่ายของหัวใจและหลอดเลือด           | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |
| y      | การเป็นโรคหลอดเลือดสมอง                                  | นามบัญญัติ    | 0: ไม่เป็น, 1: เป็น                |

### การวิเคราะห์ข้อมูล

งานวิจัยนี้วิเคราะห์ข้อมูลโดยใช้โปรแกรม R และ RStudio [13] โดยมีรายละเอียดในการวิเคราะห์ดังต่อไปนี้

1. การวิเคราะห์ข้อมูลเบื้องต้นของผู้สูงอายุ ได้แก่ อายุ เพศ ค่าดัชนีมวลกาย ประวัติการสูบบุหรี่ ประวัติการดื่มแอลกอฮอล์ ประวัติการเป็นโรคต่าง ๆ ด้วยสถิติพื้นฐาน เช่น จำนวน ร้อยละ ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐาน โดยจำแนกตามผู้ป่วยที่เป็นและไม่เป็นโรคหลอดเลือดสมอง

2. การแก้ปัญหาข้อมูลไม่สมดุล โดยใช้วิธีสุ่มเกิน (Over-sampling) เนื่องจากในการศึกษาครั้งนี้จำนวนข้อมูลของผู้ที่เป็นโรคหลอดเลือดสมอง มีจำนวน

1,590 คน จากจำนวนผู้เข้ารับบริการทั้งหมด 28,928 คน จะเห็นได้ว่าข้อมูลของผู้ที่เป็นโรคหลอดเลือดสมองมีสัดส่วนน้อยกว่าผู้ที่ไม่เป็นโรคหลอดเลือดสมองค่อนข้างมาก ซึ่งถือว่าเป็นลักษณะของข้อมูลไม่สมดุล เมื่อทำการจำแนกจะทำให้มีการเรียนรู้แต่ข้อมูลกลุ่มมาก ผลที่ได้ก็จะจำแนกไปในข้อมูลกลุ่มมาก ซึ่งเทคนิคสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling TEchnique: SMOTE) จะเป็นเทคนิคหนึ่งที่จะนิยมใช้ในการแก้ปัญหาการจำแนกข้อมูลไม่สมดุล โดยจะเพิ่มจำนวนข้อมูลประเภทที่มีข้อมูลน้อยให้เพิ่มปริมาณข้อมูลใกล้เคียงกับประเภทที่มีมากที่สุดเพื่อให้ข้อมูลมีความสมดุลมากขึ้น โดยสุ่มค่าข้อมูลที่อยู่ในกลุ่มข้อมูลน้อยมา 1 ค่า แล้วเลือกค่าข้อมูลใกล้เคียงอีกจำนวน K

ค่า (K-nearest neighbor) แล้วหาค่าระยะห่าง Euclidean distance) ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียง หลังจากนั้นก็สร้างข้อมูลเทียมระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูล

$$x_{new} = x_1 + distance(\hat{x}_u, x_1) \times \delta \quad (1)$$

โดยที่  $x_{new}$  คือ ข้อมูลใหม่  $x_1$  คือ ข้อมูลที่สุ่มในครั้งแรก  $\hat{x}_u$  คือ ข้อมูลที่สุ่มมาเพิ่ม K ค่า และ  $\delta$  คือ ค่าที่สุ่มตั้งแต่ 0-1 งานวิจัยนี้ใช้ชุดคำสั่ง “SMOTE” ใน Package “DMwR” ของ R-project [14]

### 3. การสร้างตัวแบบในการจำแนกผู้ป่วยโรคหลอดเลือดสมอง

สำหรับงานวิจัยด้านวิทยาศาสตร์สุขภาพและสาธารณสุขนั้น พบว่า ตัวแบบการถดถอยลอจิสติกทวิภาค และ ต้นไม้ตัดสินใจ เป็นตัวแบบที่มักจะนิยมใช้สำหรับการศึกษาความสัมพันธ์ของปัจจัยต่าง ๆ กับตัวแปรตามที่เป็นข้อมูลมาตราชานามบัญญัติที่มี 2 ค่า [4], [5], [6], [7] ดังนั้นงานวิจัยนี้จึงสร้างตัวแบบการถดถอยลอจิสติกทวิภาค (Binary logistic regression analysis: glm) และต้นไม้ตัดสินใจร่วมกับเทคนิค Bootstrap Aggregating (Bagging) เพื่อศึกษาปัจจัยที่ส่งผลต่อการเกิดโรคหลอดเลือดสมอง

$$\log \left( \frac{P(Y=1)}{P(Y=0)} \right) = b_0 + b_1x_1 + \dots + b_px_p \quad (2)$$

โดยที่  $b_i$  คือค่าสัมประสิทธิ์การถดถอยลอจิสติกของตัวแปรอิสระแต่ละตัว ซึ่งประมาณค่าจากชุดข้อมูลฝึกฝนโดยใช้วิธีภาวะน่าจะเป็นสูงสุด

ใกล้เคียงตัวที่ให้ค่าระยะทางที่น้อยที่สุดดังสมการที่ (1) [9], [10], [12]

### 3.1 ตัวแบบการถดถอยลอจิสติกทวิภาค

ตัวแบบนี้เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เมื่อตัวแปรตาม (Y) เป็นตัวแปรเชิงคุณภาพที่มี 2 ค่า (Dichotomous variable) คือ เกิดเหตุการณ์ที่สนใจ ( $Y = 1$ ) ด้วยความน่าจะเป็น  $p$  และไม่เกิดเหตุการณ์ที่สนใจ ( $Y = 0$ ) ด้วยความน่าจะเป็น  $1 - p$  และเนื่องจากตัวแปรตามเป็นตัวแปรเชิงคุณภาพทำให้ความสัมพันธ์กับตัวแปรอิสระ ( $x$ ) ไม่อยู่ในรูปเชิงเส้น (Nonlinear) แต่จะอยู่ในรูปคล้ายตัว S จึงปรับให้ความสัมพันธ์ให้อยู่ในรูปเชิงเส้นในรูปแบบ log ของ odd ratio ซึ่งคือ log ของอัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ( $P(Y = 1)$ ) กับโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ ( $P(Y = 0)$ ) เรียกว่าฟังก์ชันผลตอบสนองลอจิสติก (Logistic response function) [15] ดังนี้

(Maximum likelihood) ที่มีฟังก์ชันภาวะน่าจะเป็น (Likelihood function) ดังสมการที่ (3)

$$L(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^n [p_i]^{y_i} [1 - p_i]^{1-y_i} \quad (3)$$

การเลือกตัวแปรอิสระจะคัดเลือกแบบ Forward Stepwise : Likelihood ratio เริ่มจากมีเฉพาะค่าคงที่  $\beta_0$  อยู่ในสมการ ขั้นตอนต่อไปจะเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดและมีความสำคัญทางสถิติ เข้ามาในสมการครั้งละ 1 ตัว และจะทดสอบว่าตัวแปรอิสระที่อยู่ในสมการแล้วควรจะตัดออกจากสมการหรือไม่ โดยพิจารณาค่าภาวะน่าจะเป็น (Likelihood value) โดยวัดค่า  $-2LL$  ( $-2 \log$  Likelihood) ค่า  $-2LL$  ที่น้อยกว่าแสดงว่าสมการถดถอยลอจิสติกมีความเหมาะสมมากกว่า นั่นคือ ถ้าค่า  $-2LL$  ลดลง

แสดงว่าตัวแปรอิสระนั้นควรจะคงอยู่ในสมการ และทำซ้ำ ๆ จนกระทั่งไม่สามารถเลือกตัวแปรอิสระตัวใดเข้ามาในสมการอีก [16]

การตรวจสอบความเหมาะสมของตัวแบบซึ่งมีสมมติฐานทางเลือกคือตัวแบบเหมาะสม โดยใช้สถิติทดสอบอัตราส่วนภาวะน่าจะเป็น (Likelihood ratio test statistics: LR) ซึ่งเปรียบเทียบค่าภาวะน่าจะเป็นของตัวแบบเริ่มต้น ( $L(0)$ ) กับ ค่าภาวะน่าจะเป็นของตัวแบบสุดท้าย ( $L(\beta)$ ) ดังสมการที่ (4) และ

ทดสอบนัยสำคัญของสัมประสิทธิ์ตัวแปรอิสระแต่ละตัว โดยใช้ตัวสถิติวัตต์ (Wald statistic) ดังสมการที่ (5)

$$LR = -2 \log \frac{L(0)}{L(\beta)} \sim \chi^2_{df=p} \quad (4)$$

$$W = \left( \frac{b_i}{SE_{b_i}} \right)^2 \sim \chi^2_{df=1} \quad (5)$$

การตรวจสอบข้อสมมติเบื้องต้นว่าตัวแปรอิสระไม่ควรจะมีปัญหาความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity) งานวิจัยนี้ใช้สถิติในการวิเคราะห์ความสัมพันธ์ 4 ค่า ดังตารางที่ 2 โดยพิจารณาตามระดับการวัดของตัวแปรที่นำมาหาความสัมพันธ์ และ

จะใช้เกณฑ์ของ Stevens [17] ในการสรุปความสัมพันธ์ คือ ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเกิน 0.80 แสดงว่าตัวแปรอิสระคู่หนึ่งมีความสัมพันธ์กันสูงซึ่งจะส่งผลให้เกิดปัญหาความสัมพันธ์เชิงเส้นแบบพหุได้

ตารางที่ 2 สถิติที่ใช้ในการวิเคราะห์ความสัมพันธ์จำแนกตามลักษณะตัวแปร

| สถิติ                | ลักษณะตัวแปร                                                                                           | สูตร                                                                                                                           |
|----------------------|--------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Phi correlation      | ตัวแปรมาตราส่วนบัญญัติ<br>ที่ตัวแปรทั้งสองตัวที่มี 2 กลุ่ม                                             | $\hat{\phi} = \frac{O_{21}O_{12} - O_{11}O_{22}}{\sqrt{(O_{11} + O_{12})(O_{21} + O_{22})(O_{11} + O_{21})(O_{12} + O_{22})}}$ |
| Cramer's correlation | ตัวแปรมาตราส่วนบัญญัติ<br>ที่ตัวแปรทั้งสองตัวที่มีมากกว่า 2 กลุ่ม                                      | $V = \sqrt{\frac{\chi^2}{n(k-1)}}$                                                                                             |
| Pearson correlation  | ตัวแปรมาตราส่วนอัตราส่วน<br>หรืออัตราส่วนทั้งสองตัว ( $x, y$ )                                         | $r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$                                    |
| Biserial correlation | ตัวแปรมาตราส่วนบัญญัติ ( $x$ ) แบ่งได้<br>2 กลุ่ม กับตัวแปรมาตราส่วนอัตราส่วน<br>หรืออัตราส่วน ( $y$ ) | $r_{bi} = \frac{(\bar{y}_1 - \bar{y}_0)pq}{s_y}$                                                                               |

เมื่อ  $O_{ij}$  คือจำนวนค่าสังเกตของตัวแปรที่  $i$  และตัวแปรที่  $j$ ,  $\chi^2$  คือสถิติทดสอบไคสแควร์สำหรับตารางสองทาง  $n$  คือ จำนวนตัวอย่าง  $k = \min(r, c)$ ,  $r$  และ  $c$  คือ จำนวนกลุ่มของตัวแปรที่ 1 และ 2,  $x$  และ  $y$  คือ ตัวแปรที่ 1 และ 2,  $\bar{y}_1$  และ  $\bar{y}_0$  คือ ค่าเฉลี่ยของข้อมูลระหว่างตัวแปร  $y$  จากกลุ่มตัวแปร  $x = 1$  และ  $x = 0$ ,  $p$  และ  $q$  คือ สัดส่วนของคนที่อยู่ในกลุ่มตัวแปร  $x = 1$  และ  $x = 0$ ,  $s_y$  คือส่วนเบี่ยงเบนมาตรฐานของข้อมูลจากตัวแปร  $y$  ทั้งหมด

### 3.2 การวิเคราะห์ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเป็นวิธีการที่นิยมใช้มากที่สุดแบบหนึ่ง ซึ่งเป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ การสร้างต้นไม้ตัดสินใจแต่ละตัวแบบจะมีลักษณะคล้ายกันคือ เริ่มต้นทำการคัดเลือกคุณลักษณะที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโนดรากของต้นไม้ หลังจากนั้นจะทำ

การแตกกิ่งคุณลักษณะออกไปเรื่อย ๆ จนสามารถแบ่งข้อมูลออกเป็นคลาสได้ชัดเจน

กำหนดให้  $D$  แทน ชุดข้อมูลฝึกฝน (Training set),  $m$  แทน หมวดหมู่ของข้อมูลที่เป็นไปได้  $p_i = \frac{|C_i D|}{|D|}$  คือ ค่าความน่าจะเป็นที่ข้อมูลหนึ่ง ๆ จะมีหมวดหมู่ของข้อมูลเป็นหมวดหมู่  $C_i$  ซึ่ง  $|D|$  และ  $|C_i, D|$  คือ จำนวนข้อมูลที่อยู่ในชุดข้อมูล  $D$  และ  $C_i, D$  เมื่อ  $C_i, D$  คือ เซตข้อมูลที่อยู่ในหมวดหมู่  $C_i$  ( $i = 1, 2, \dots, m$ ),  $x$  แทนคุณลักษณะที่ถูกเลือกให้เป็นตัวจำแนกข้อมูล  $v$  แทนจำนวนค่าที่เป็นไปได้ของคุณลักษณะ  $x$  และ  $|D_j|$  แทน จำนวนข้อมูลที่มีค่า  $x = j$  เกณฑ์การเลือกคุณลักษณะ มีดังนี้

1. ค่าเกนความรู้ (Information gain: Gain) เป็นค่าที่บอกระดับความสามารถของการจำแนกคลาสของคุณลักษณะ หน่วยของการวัดเป็น bits ดังสมการที่ (6)-(8) โดยจะพิจารณาเลือกจากค่ามากที่สุด

$$\text{info}(D) = -\sum_{i=1}^m p_i \log_2(p_i) \quad (6)$$

$$\text{info}_x(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times \text{info}(D_j) \quad (7)$$

$$\text{Gain}(x) = \text{info}(D) - \text{info}_x(D) \quad (8)$$

2. ค่าอัตราส่วนเกน (*GainRatio*) เป็นตัววัดการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยที่พัฒนามาจากค่าเกนความรู้ โดยเมื่อใช้ค่าเกนความรู้ในการแบ่งชุดข้อมูลจะทำให้เกิดความเอนเอียงเกิดขึ้นได้เมื่อคุณลักษณะที่กำลังพิจารณามีค่าที่เกิดขึ้นเป็นจำนวนมาก ตัวชี้วัดนี้จึงได้พยายามที่จะลดทอนความเอนเอียงลง โดยประยุกต์การทำนอร์มัลไลซ์ค่าเกน

ความรู้ด้วยการใช้ค่า “Split information” ซึ่งจะแสดงถึงปริมาณข้อมูลที่ถูกพิจารณาโดยการแบ่งข้อมูลในชุดข้อมูล  $D$  ออกเป็น  $v$  ชุดข้อมูลย่อยตามค่าในคุณลักษณะ  $x$  เพื่อใช้หาค่าอัตราส่วนเกน ดังสมการที่ (9)-(10) โดยจะพิจารณาเลือกจากค่ามากที่สุดเป็นคุณลักษณะที่จะนำมาใช้ในการแบ่งกลุ่ม

$$\text{splitinfo}_x(D) = -\sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \left( \frac{|D_j|}{|D|} \right) \quad (9)$$

$$\text{GainRatio}(x) = \frac{\text{Gain}(A)}{\text{splitinfo}_x(D)} \quad (10)$$

3. ดัชนีจีนิ (*Gini index*) จะพิจารณาความไม่บริสุทธิ์ของชุดข้อมูล  $D$  ที่มีเซตของข้อมูลที่มีหมวดหมู่ของข้อมูลไม่เหมือนกันอยู่ ดังสมการที่ (11) ในการพิจารณาคุณลักษณะ  $x$  จะแบ่งชุดข้อมูลเป็น 2 ชุดข้อมูลย่อยเท่านั้น ( $D_1, D_2$ ) ดังสมการที่ (12) เพื่อหา

ค่าดัชนีจีนิของคุณลักษณะ  $x$  โดยจะพิจารณาเลือกจากค่าน้อยที่สุดเป็นคุณลักษณะที่จะนำมาใช้ในการแบ่งกลุ่ม ตัวชี้วัดนี้จะใช้สำหรับต้นไม้ตัดสินใจแบบ J48 และ CART

$$\text{Gini}(D) = 1 - \sum_{i=1}^m p_i^2 \quad (11)$$

$$\text{Gini}_x(D) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \quad (12)$$

งานวิจัยนี้เลือกใช้ต้นไม้ตัดสินใจต่อไปนี้

(1) ต้นไม้ตัดสินใจแบบ ID3 (Iterative Dichotomiser 3: ID3) จะใช้หลักการการใช้ทฤษฎีข่าวสาร โดยใช้ค่าเกนความรู้เป็นตัววัดว่าจะใช้คุณลักษณะใดในการจำแนกข้อมูล [18] งานวิจัยนี้ใช้ฟังก์ชัน `rpart with split = "information"` ของ `rpart` package ในโปรแกรม R

(2) ต้นไม้ตัดสินใจแบบ C4.5 (J48) จะใช้หลักการเดียวกับต้นไม้ตัดสินใจแบบ ID3 แต่มีการปรับปรุงให้ใช้ได้กับข้อมูลเชิงปริมาณ และข้อมูลที่มีค่าสูญหายได้ รวมทั้งสามารถทำการปรับแต่งต้นไม้ได้ (Pruning trees) โดยจะใช้ค่าอัตราส่วนเกนเป็นตัววัดว่าจะใช้คุณลักษณะใดในการจำแนกข้อมูล [19] งานวิจัยนี้ใช้ฟังก์ชัน `J48()` ของ `RWeka` package ในโปรแกรม R

(3) ต้นไม้ตัดสินใจแบบ CART (Classification and Regression Trees: CART) จะ

คล้ายคลึงกับต้นไม้ตัดสินใจแบบ C4.5 แต่สามารถใช้กับตัวแปรเชิงปริมาณได้และไม่ได้ใช้ rule sets และจะสร้างต้นไม้ตัดสินใจแบบสองทางเลือก (Binary trees) โดยจะใช้ดัชนีจีนิเป็นตัววัดว่าจะใช้คุณลักษณะใดในการจำแนกข้อมูล [20] งานวิจัยนี้ใช้ฟังก์ชัน `rpart with split = "gini"` ของ `rpart` package ในโปรแกรม R [21]

(4) ต้นไม้ตัดสินใจแบบอนุमानตามเงื่อนไข (Conditional inference Trees: CTree) เป็น Unbiased algorithm ถูกพัฒนาขึ้นเนื่องจากต้นไม้ตัดสินใจอื่น ๆ เช่น CART และ C4.5 มักจะมีอคติในการเลือกคุณลักษณะโดยมักจะเลือกคุณลักษณะที่มีจำนวนสมาชิกมากกว่า แต่ต้นไม้ตัดสินใจแบบ CTree จะใช้การทดสอบสมมติฐานทางสถิติและพิจารณาค่า p-value ที่มีค่าน้อยที่สุดเป็นตัวแบ่งข้อมูลแบบสองทางเลือก [22] งานวิจัยนี้ใช้ฟังก์ชัน `ctree` ของ `partykit` package ในโปรแกรม R [23]

(5) ต้นไม้ตัดสินใจแบบ C5.0 (C5.0) ถูกพัฒนาเพิ่มเติมจากต้นไม้ตัดสินใจแบบ C4.5 เพื่อเพิ่มความแม่นยำของการจำแนกด้วยการใช้วิธี Boosting และมีประสิทธิภาพดีกว่าในแง่ของความเร็วการใช้หน่วยความจำ และขนาดของต้นไม้ตัดสินใจ และสามารถใช้ได้กับข้อมูลที่มีค่าหลายค่า รวมทั้งรองรับชนิดข้อมูลใหม่ ๆ และข้อมูลขนาดใหญ่ [24] (ศึกษารายละเอียดเพิ่มเติมได้ที่ [www.rulequest.com](http://www.rulequest.com)) โดยจะใช้ค่าเกณฑ์ความรู้เป็นตัววัดว่าจะใช้คุณลักษณะใดในการจำแนกข้อมูล งานวิจัยนี้ใช้ฟังก์ชัน C5.0 ของ c50 package ในโปรแกรม R

### 3.3 Bootstrap Aggregating (Bagging)

เทคนิค Bagging เป็นเทคนิคหนึ่งในเทคนิคการรวมกลุ่มตัวจำแนกประเภท (Ensemble technique) เพื่อเพิ่มประสิทธิภาพการทำนาย โดยเริ่มจากสุ่มชุดข้อมูลฝึกฝนจำนวน  $c$  ชุดแบบสุ่มซ้ำ

ตารางที่ 3 Confusion Matrix

| Predicted class | Actual class       |                    |
|-----------------|--------------------|--------------------|
|                 | Yes (+: เป็นโรค)   | No (-: ไม่เป็นโรค) |
| Yes (+)         | True Positive: TP  | False Positive: FP |
| No (-)          | False Negative: FN | True Negative: TN  |

การคำนวณค่าต่าง ๆ เพื่อวัดประสิทธิภาพของตัวแบบการทำนาย ดังต่อไปนี้

1. ค่าความแม่นยำ (Accuracy) เป็นการวัดประสิทธิภาพการพยากรณ์ของตัวแบบโดยรวมคำนวณโดย  $Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$

2. ค่าความไว (Sensitivity) หรือ True-Positive Rate เป็นการวัดประสิทธิภาพการพยากรณ์ถูกต้องในกลุ่มที่สนใจ คือ อัตราส่วนที่คนที่เป็นโรคจริง ๆ ทั้งหมดที่ตรวจได้ผลบวก คำนวณโดย  $Sensitivity = \frac{TP}{TP + FN}$

3. ค่าความจำเพาะ (Specificity) หรือ True-Negative Rate เป็นการวัดประสิทธิภาพการพยากรณ์อัตราส่วนถูกต้องในกลุ่มอื่น ๆ คือ อัตราส่วนที่คนที่ไม่เป็นโรคทั้งหมดที่ตรวจได้ผลลบ คำนวณโดย  $Specificity = \frac{TN}{FP + TN}$

(bootstraps replacement) แล้วนำมาสร้างตัวแบบด้วยเทคนิคเดียวกันทั้งหมด และจะได้ผลการทำนายของแต่ละชุดย่อย ( $\hat{y}_i$ ) ซึ่งอาจจะเป็นค่าความน่าจะเป็นหรือเป็นประเภทข้อมูล แล้วจึงนำมาพิจารณาเพื่อหาผลการทำนายร่วม ( $\hat{y}$ ) โดยการเฉลี่ย (averaging) หรือการหาค่าที่พบมากที่สุด (voting) ซึ่งการหาผลการทำนายร่วมกันของข้อมูลหลายชุดนั้นจะช่วยลดความแปรปรวนของการประมาณและเพิ่มความแม่นยำของตัวแบบได้ [25]

### 4. การเปรียบเทียบประสิทธิภาพของตัวแบบ

การวัดประสิทธิภาพของตัวแบบโดยใช้ตาราง Confusion Matrix ขนาด  $2 \times 2$  ดังตารางที่ 3 ซึ่งสรุปจำนวนตัวอย่างที่ตัวแบบจำแนกได้ถูกต้องและไม่ถูกต้อง

4. ค่าทำนายผลบวกหรือค่าพยากรณ์ผลบวก (Positive Predictive Value: PPV) คือ อัตราส่วนที่คนที่ถูกทำนายผลเป็นบวกที่เป็นโรคจริง ๆ คำนวณโดย  $PPV = \frac{TP}{TP + FP}$

5. ค่าทำนายผลลบหรือค่าพยากรณ์ผลลบ (Negative Predictive Value: NPV) คือ อัตราส่วนที่คนที่ถูกทำนายผลเป็นลบที่ไม่เป็นโรคจริง ๆ คำนวณโดย  $NPV = \frac{TN}{FN + TN}$

### ผลการศึกษา

1. ผลการวิเคราะห์ข้อมูลส่วนบุคคลของผู้สูงอายุที่เข้ารับบริการ

ผู้สูงอายุที่มาใช้บริการโรงพยาบาลสมเด็จพระปิ่นเกล้าในระหว่างวันที่ 1 มกราคม 2561 - 31 ธันวาคม 2561 จำนวน 28,928 คน ส่วนใหญ่เป็นเพศหญิง (57.36%) และมีอายุเฉลี่ย 71.88 ปี มีค่าดัชนีมวลกาย



มวลกายอยู่ในเกณฑ์ปกติ (56.93%) รองลงมา มีภาวะอ้วน (31.17%) ไม่ดื่มแอลกอฮอล์ (87.59%) ไม่สูบบุหรี่ (88.46%) และส่วนใหญ่ไม่เป็นโรคต่าง ๆ ที่ศึกษา ยกเว้นโรคความดันโลหิต

อัตราความชุกของการเกิดโรคหลอดเลือดสมองของผู้สูงอายุที่เข้ารับบริการโดยรวม เท่ากับ 5.50% (95% CI 5.24% -5.76%) และพบว่า ปัจจัยเพศ และโรคความดันโลหิตจะมีนัยสำคัญสำหรับผู้ที่เป็โรค

หลอดเลือดสมอง นั่นคือเพศชายมีโอกาสเป็นโรคหลอดเลือดสมองมากกว่าเพศหญิง ร้อยละ 2.90 (95% CI 2.72-3.10) และผู้ที่เป็นโรคความดันโลหิตมีโอกาสเป็นโรคหลอดเลือดสมองมากกว่าผู้ที่ไม่เป็นโรคความดันโลหิต ร้อยละ 3.48 (95% CI 3.28-3.70) ส่วนปัจจัยอื่น ๆ และโรคอื่น ๆ จะไม่มีนัยสำคัญต่อผู้ที่เป็นโรคหลอดเลือดสมอง

**ตารางที่ 4** จำนวน ร้อยละ ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของข้อมูลพื้นฐานผู้สูงอายุจำแนกตามผู้ป่วยที่เป็นและไม่เป็นโรคหลอดเลือดสมอง

| ข้อมูลพื้นฐาน         | โรคหลอดเลือดสมอง |        |           |       | รวม       |        |
|-----------------------|------------------|--------|-----------|-------|-----------|--------|
|                       | ไม่เป็น          |        | เป็น      |       |           |        |
| ค่าสถิติ              | $\bar{x}$        | $s$    | $\bar{x}$ | $s$   | $\bar{x}$ | $s$    |
| อายุ                  | 71.71            | 8.68   | 74.77     | 9.49  | 71.88     | 8.76   |
| ค่าสถิติ              | $n$              | %      | $n$       | %     | $n$       | %      |
| เพศ                   |                  |        |           |       |           |        |
| ชาย                   | 11496            | 39.74% | 840       | 2.90% | 12336     | 42.64% |
| หญิง                  | 15842            | 54.76% | 750       | 2.59% | 16592     | 57.36% |
| ดัชนีมวลกาย           |                  |        |           |       |           |        |
| น้อยกว่า 18.50        | 1131             | 3.91%  | 62        | 0.21% | 1193      | 4.12%  |
| ระหว่าง 18.50 - 22.99 | 15489            | 53.54% | 980       | 3.39% | 16469     | 56.93% |
| ระหว่าง 23 - 29.99    | 8541             | 29.53% | 477       | 1.65% | 9018      | 31.17% |
| มากกว่า 30            | 2177             | 7.53%  | 71        | 0.25% | 2248      | 7.77%  |
| การดื่มแอลกอฮอล์      |                  |        |           |       |           |        |
| ดื่ม                  | 3314             | 11.46% | 276       | 0.95% | 3590      | 12.41% |
| ไม่ดื่ม               | 24024            | 83.05% | 1314      | 4.54% | 25338     | 87.59% |
| การสูบบุหรี่          |                  |        |           |       |           |        |
| สูบ                   | 3029             | 10.47% | 305       | 1.05% | 3334      | 11.53% |
| ไม่สูบ                | 24304            | 84.02% | 1285      | 4.44% | 25589     | 88.46% |
| บุหรี่มือสอง          | 5                | 0.02%  | 0         | 0.00% | 5         | 0.02%  |
| โรคประจำตัว           |                  |        |           |       |           |        |
| ไม่เป็น               | 26110            | 90.26% | 1532      | 5.30% | 27642     | 95.55% |
| เป็น                  | 1228             | 4.25%  | 58        | 0.20% | 1286      | 4.45%  |
| ภาวะโลหิตจาง          |                  |        |           |       |           |        |
| ไม่เป็น               | 26647            | 92.11% | 1526      | 5.28% | 28173     | 97.39% |
| เป็น                  | 691              | 2.39%  | 64        | 0.22% | 755       | 2.61%  |

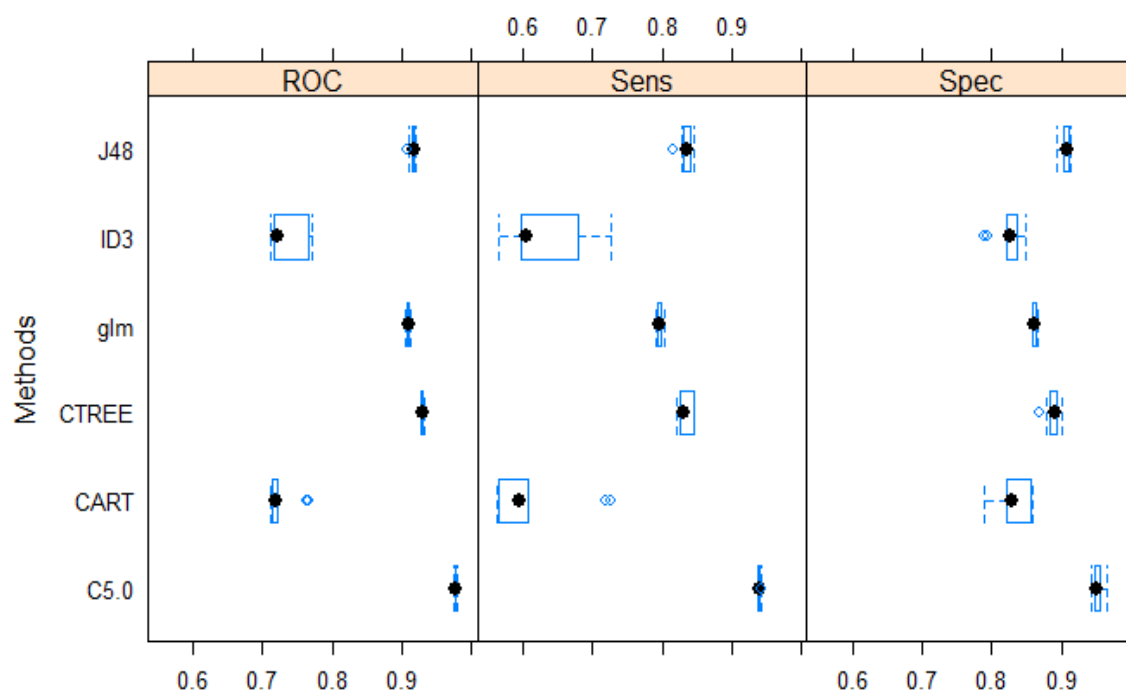
| ข้อมูลพื้นฐาน                                            | โรคหลอดเลือดสมอง |        |           |       | รวม       |        |
|----------------------------------------------------------|------------------|--------|-----------|-------|-----------|--------|
|                                                          | ไม่เป็น          |        | เป็น      |       |           |        |
| ค่าสถิติ                                                 | $\bar{x}$        | $s$    | $\bar{x}$ | $s$   | $\bar{x}$ | $s$    |
| ภาวะแข็งตัวของเลือดผิดปกติ และเลือดออก                   |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 27280            | 94.30% | 1581      | 5.47% | 28861     | 99.77% |
| เป็น                                                     | 58               | 0.20%  | 9         | 0.03% | 67        | 0.23%  |
| โรคเบาหวาน                                               |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 22265            | 76.97% | 1173      | 4.05% | 23438     | 81.02% |
| เป็น                                                     | 5073             | 17.54% | 417       | 1.44% | 5490      | 18.98% |
| กลุ่มโรคอ้วนและภาวะโภชนาการเกิน                          |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 27286            | 94.32% | 1589      | 5.49% | 28875     | 99.82% |
| เป็น                                                     | 52               | 0.18%  | 1         | 0.00% | 53        | 0.18%  |
| ภาวะเผาผลาญไขมันผิดปกติ                                  |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 27327            | 94.47% | 1589      | 5.49% | 28916     | 99.96% |
| เป็น                                                     | 11               | 0.04%  | 1         | 0.00% | 12        | 0.04%  |
| โรคลมชัก                                                 |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 27221            | 94.10% | 1515      | 5.24% | 28736     | 99.34% |
| เป็น                                                     | 117              | 0.40%  | 75        | 0.26% | 192       | 0.66%  |
| ภาวะหลอดเลือดสมองครั้งคราว                               |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 27190            | 93.99% | 1528      | 5.28% | 28718     | 99.27% |
| เป็น                                                     | 148              | 0.51%  | 62        | 0.21% | 210       | 0.73%  |
| กลุ่มโรคหัวใจ                                            |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 25102            | 86.77% | 1369      | 4.73% | 26471     | 91.51% |
| เป็น                                                     | 2236             | 7.73%  | 221       | 0.76% | 2457      | 8.49%  |
| โรคความดันโลหิต                                          |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 16749            | 57.90% | 582       | 2.01% | 17331     | 59.91% |
| เป็น                                                     | 10589            | 36.60% | 1008      | 3.48% | 11597     | 40.09% |
| โรคลิ่มเลือดอุดตันที่ปอด                                 |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 27309            | 94.40% | 1587      | 5.49% | 28896     | 99.89% |
| เป็น                                                     | 29               | 0.10%  | 3         | 0.01% | 32        | 0.11%  |
| กลุ่มโรคของหลอดเลือดแดง หลอดเลือดแดงย่อย และหลอดเลือดฝอย |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 27242            | 94.17% | 1580      | 5.46% | 28822     | 99.63% |
| เป็น                                                     | 96               | 0.33%  | 10        | 0.03% | 106       | 0.37%  |
| โรคไต                                                    |                  |        |           |       |           |        |
| ไม่เป็น                                                  | 25045            | 86.58% | 1425      | 4.93% | 26470     | 91.50% |
| เป็น                                                     | 2293             | 7.93%  | 165       | 0.57% | 2458      | 8.50%  |

| ข้อมูลพื้นฐาน                                  | โรคหลอดเลือดสมอง |        |           |       | รวม       |         |
|------------------------------------------------|------------------|--------|-----------|-------|-----------|---------|
|                                                | ไม่เป็น          |        | เป็น      |       |           |         |
| ค่าสถิติ                                       | $\bar{x}$        | $s$    | $\bar{x}$ | $s$   | $\bar{x}$ | $s$     |
| การบาดเจ็บที่ศีรษะ                             |                  |        |           |       |           |         |
| ไม่เป็น                                        | 26850            | 92.82% | 1543      | 5.33% | 28393     | 98.15%  |
| เป็น                                           | 488              | 1.69%  | 47        | 0.16% | 535       | 1.85%   |
| การมีอุปกรณ์ฝังและปลูกถ่ายของหัวใจและหลอดเลือด |                  |        |           |       |           |         |
| ไม่เป็น                                        | 27114            | 93.73% | 1574      | 5.44% | 28688     | 99.17%  |
| เป็น                                           | 224              | 0.77%  | 16        | 0.06% | 240       | 0.83%   |
| รวม                                            | 27338            | 94.50% | 1590      | 5.50% | 28928     | 100.00% |

## 2. การเปรียบเทียบประสิทธิภาพของตัวแบบ

การตรวจสอบข้อสมมติของตัวแบบ glm พบว่า ตัวแปรส่วนใหญ่มีระดับความสัมพันธ์ต่ำมากถึงค่อนข้างสูง แต่ยังไม่เกิน 0.80 ตามเกณฑ์ของ Stevens [17] จึงยังไม่ก่อให้เกิดปัญหาความสัมพันธ์เชิงเส้นแบบพหุ และเนื่องจากจำนวนผู้ป่วยที่เป็นและไม่เป็นโรคหลอดเลือดสมองมีขนาดแตกต่างกันมาก (1,590/27,338) ผู้วิจัยจึงใช้เทคนิค SMOTE เพื่อเพิ่ม

ข้อมูลกลุ่มน้อยจนได้จำนวนข้อมูลของทั้งสองกลุ่มใกล้เคียงกัน (27,984 และ 27,030) เป็นจำนวนรวมทั้งสิ้น 55,014 คน จากนั้นแบ่งข้อมูลเป็น 2 ชุด ชุดแรกคือ ชุดข้อมูลฝึกฝนร้อยละ 80 (44,011 คน) และใช้เทคนิค Bagging โดยสุ่มชุดข้อมูลจากชุดข้อมูลฝึกฝนเป็นจำนวนทั้งหมด 10 ชุด เพื่อใช้สร้างตัวแบบด้วยเทคนิคจำแนกเดียวกัน และชุดที่สองคือ ชุดข้อมูลทดสอบร้อยละ 20 (11,003 คน) เพื่อใช้ทดสอบตัวแบบ



รูปที่ 1 ประสิทธิภาพการจำแนกผู้ป่วยโรคหลอดเลือดสมองสำหรับชุดข้อมูลฝึกฝนจำแนกตามตัวแบบ

ตัวแบบที่นำมาศึกษาในการจำแนกการเกิดโรคหลอดเลือดสมอง คือ ตัวแบบ glm และเทคนิคต้นไม้ตัดสินใจแบบ ID3, CART, J48 (C4.0), CTREE และ C5.0 ร่วมกับเทคนิค Bagging โดยพบว่าประสิทธิภาพของการจำแนกผู้ป่วยโรคหลอดเลือดสมองด้วยตัวแบบดังกล่าวข้างต้นสามารถจำแนกผู้ป่วยโรคหลอดเลือดสมองได้แม่นยำค่อนข้างสูง และเมื่อพิจารณาในแต่ละวิธีพบว่าเทคนิคต้นไม้ตัดสินใจแบบ C5.0 มีค่า ROC ( $\bar{x} = 97.75\%$ ,  $s = 1.01\%$ ) ค่าความไว ( $\bar{x} = 99.93\%$ ,  $s = 1.65\%$ ) และค่าความจำเพาะ ( $\bar{x} = 95.23\%$ ,  $s = 6.14\%$ ) สูงกว่า

ตัวแบบอื่น ๆ และมีค่าความแปรปรวนของค่า ROC และค่าความไวจากการทำ Bagging จำนวน 10 ชุด น้อยกว่าตัวแบบอื่น ๆ (รูปที่ 1)

ในชุดข้อมูลทดสอบให้ผลสรุปการเปรียบเทียบเหมือนกับชุดข้อมูลฝึกฝน คือ เทคนิคต้นไม้ตัดสินใจแบบ C5.0 มีประสิทธิภาพในการจำแนกสูงกว่าตัวแบบอื่น ๆ โดยมีค่าความแม่นยำในการจำแนกร้อยละ 95.31 ค่าความไว ร้อยละ 94.48 ค่าความจำเพาะ ร้อยละ 96.12 ค่าทำนายผลบวก ร้อยละ 95.93 และค่าทำนายผลลบ ร้อยละ 94.73 (ตารางที่ 5)

ตารางที่ 5 ประสิทธิภาพการจำแนกผู้ป่วยโรคหลอดเลือดสมองสำหรับชุดข้อมูลทดสอบจำแนกตามตัวแบบ

| ตัวแบบ | Accuracy | Sensitivity | Specificity | PPV   | NPV   |
|--------|----------|-------------|-------------|-------|-------|
| ID3    | 72.88    | 61.75       | 83.66       | 78.55 | 69.30 |
| CART   | 72.88    | 61.75       | 83.66       | 78.55 | 69.30 |
| glm    | 83.62    | 79.78       | 87.35       | 85.94 | 81.68 |
| J48    | 86.36    | 83.90       | 88.74       | 87.84 | 85.05 |
| CTREE  | 87.24    | 84.34       | 90.05       | 89.15 | 85.58 |
| C5.0   | 95.31    | 94.48       | 96.12       | 95.93 | 94.73 |

Rules: .... Rule 110: (516/55, lift 1.8)  
 Rule 1: (724, lift 2.0)  
 age > 69  
 age <= 69.99133  
 -> class Pos [0.999]  
 Rule 2: (839, lift 2.0)  
 age > 70  
 age <= 70.99721  
 -> class Pos [0.999]  
 Rule 3: (846, lift 2.0)  
 age > 71.00771  
 age <= 71.99906  
 -> class Pos [0.999]  
 Rule 111: (2, lift 1.5)  
 cigar = 3  
 -> class Neg [0.750]  
 Rule 112: (41057/18771, lift 1.1)  
 x8 = Neg  
 -> class Neg [0.543]  
 Default class: Neg

รูปที่ 2 ตัวอย่างกฎในการจำแนกผู้ป่วยโรคหลอดเลือดสมองของต้นไม้ตัดสินใจแบบ C5.0

3. ปัจจัยที่ส่งผลต่อการจำแนกผู้ป่วยโรคหลอดเลือดสมอง

จากการเปรียบเทียบประสิทธิภาพของตัวแบบ 6 ตัวแบบ พบว่า ตัวแบบการจำแนกที่มีประสิทธิภาพที่สุด คือ ต้นไม้ตัดสินใจแบบ C5.0 จึงศึกษาปัจจัยที่ส่งผลต่อการจำแนกผู้ป่วยโรคหลอดเลือดสมองด้วยต้นไม้ตัดสินใจแบบ C5.0 ซึ่งใช้ Rule based สร้างกฎในการจำแนกรวมทั้งสิ้น 112 กฎ ดัง

รูปที่ 2 และมีค่าผิดพลาดในการจำแนกร้อยละ 5.4 พบว่า ปัจจัยที่ส่งผลต่อการจำแนกผู้ป่วยโรคหลอดเลือดสมองโดยเรียงตามลำดับความสำคัญมีดังนี้

"x8-ภาวะหลอดเลือดสมองครึ่งคราว" "age-อายุ" "x2-ภาวะโลหิตจาง" "x7-โรคลมชัก" "cigar-การสูบบุหรี่" "x3-ภาวะแข็งตัวของเลือดผิดปกติและเลือดออก" "x14-การบาดเจ็บที่ศีรษะ" "x9-กลุ่มโรคหัวใจ" "x1-โรคมะเร็ง" "drink-การดื่มแอลกอฮอล์"

"x13-โรคไต" "x15-การมีอุปกรณ์ฝังและปลูกถ่ายของหัวใจและหลอดเลือด" "sex-เพศ" "x10-โรคความดันโลหิต" "x4-โรคเบาหวาน" "BMI-ค่าดัชนีมวลกาย" "x12-กลุ่มโรคของหลอดเลือดแดง หลอดเลือดแดงย่อย และหลอดเลือดฝอย" และ "x11-โรคลิ้นเลือดอุดตันที่ปอด"

ส่วนปัจจัยที่ไม่ส่งผลต่อการจำแนกผู้ป่วยโรคหลอดเลือดสมอง คือ "x5-กลุ่มโรคอ้วนและภาวะโภชนาการเกิน" และ "x6-ภาวะเผาผลาญไขมันผิดปกติ"

## วิจารณ์

การเปรียบเทียบตัวแบบ glm และต้นไม้ตัดสินใจแบบ ID3, CART, J48 (C4.0), CTREE และ C5.0 ร่วมกับเทคนิค Bagging เพื่อจำแนกการเกิดโรคหลอดเลือดสมองของผู้สูงอายุที่มีอายุตั้งแต่ 60 ปีขึ้นไป ที่เข้ารับบริการที่โรงพยาบาลสมเด็จพระปิ่นเกล้าในระหว่างวันที่ 1 มกราคม 2561 - 31 ธันวาคม 2561 จำนวน 28,928 คน พบว่าประสิทธิภาพการจำแนกผู้ป่วยโรคหลอดเลือดสมองด้วยเทคนิคต้นไม้ตัดสินใจแบบ C5.0 สูงที่สุด โดยมีค่าความแม่นยำสูงถึง 95.31% ซึ่งสอดคล้องกับงานวิจัยของ Maung และ Aye ซึ่งพบว่าต้นไม้ตัดสินใจแบบ C5.0 มีความแม่นยำมากกว่า ต้นไม้ตัดสินใจแบบ CART [24] และต้นไม้ตัดสินใจแบบ C4.5 [26] ซึ่งเป็นเพราะตัวแบบนี้ถูกพัฒนาขึ้นมาเพื่อให้มีการทำงานที่ดีขึ้นไม่ว่าจะเป็นการใช้หน่วยความจำที่เล็กกว่าและการทำงานที่เร็วขึ้น [26]

ต้นไม้ตัดสินใจแบบ C5.0 ได้สร้างกฎการจำแนกทั้งหมด 112 กฎ และพบว่าปัจจัยที่ส่งผลต่อการจำแนกผู้ป่วยโรคหลอดเลือดสมองเรียงตามลำดับความสำคัญ คือ ภาวะหลอดเลือดสมองครั้งคราว อายุ ภาวะโลหิตจาง โรคลมชัก การสูบบุหรี่ ภาวะแข็งตัวของเลือดผิดปกติและเลือดออก การบาดเจ็บที่ศีรษะ กลุ่มโรคหัวใจ โรคกระดูก การดื่มแอลกอฮอล์ โรคไต การมีอุปกรณ์ฝังและปลูกถ่ายของหัวใจและหลอดเลือด เพศ โรคความดันโลหิต โรคเบาหวาน ดัชนีมวลกาย กลุ่มโรคของหลอดเลือดแดง หลอดเลือดแดงย่อย และหลอดเลือดฝอย และโรคลิ้นเลือดอุดตันที่ปอด ซึ่งสอดคล้องกับผลการศึกษามากมายฉบับ [27], [28], [29], [30] เช่น ปัจจัยการเกิดภาวะหลอดเลือดสมอง

ครั้งคราวนั้นเป็นสัญญาณเตือนที่สำคัญของการเป็นโรคหลอดเลือดสมอง จึงเป็นปัจจัยที่สำคัญในการจำแนกผู้ป่วยโรคหลอดเลือดสมองได้ดี รองลงมาคือปัจจัยอายุ เนื่องจากเมื่ออายุมากขึ้น หลอดเลือดก็จะเสื่อมตามไปด้วย โดยผิวชั้นในของหลอดเลือดจะหนาและแข็งขึ้นจากการที่มีไขมันและหินปูนมาเกาะทำให้รูที่เลือดไหลผ่านจะแคบลงเรื่อย ๆ จึงทำให้ความเสี่ยงการเป็นโรคหลอดเลือดสมองเพิ่มขึ้น ซึ่งสอดคล้องกับรายงานที่พบว่าหลังอายุ 55 ปี โอกาสการเป็นโรคนี้จะเพิ่มขึ้นเป็นสองเท่าในแต่ละช่วง 10 ปี และพบว่ามากกว่า 75% ของผู้ป่วยโรคหลอดเลือดสมองมีอายุมากกว่า 65 ปี [29] ส่วนปัจจัยภาวะโลหิตจางจะทำให้อาการของโรคหลอดเลือดสมองแย่ลงเนื่องจากอาจจะเกิดขาดธาตุเหล็ก สำหรับปัจจัยการสูบบุหรี่นั้นเพิ่มความเสี่ยงต่อการเป็นโรคหลอดเลือดสมอง เนื่องจากในบุหรี่มีสารนิโคตินและคาร์บอนมอนอกไซด์ซึ่งมีส่วนทำให้ปริมาณออกซิเจนลดลง และเป็นตัวทำลายผนังหลอดเลือดทำให้หลอดเลือดแข็งตัว ในขณะที่ปัจจัยภาวะแข็งตัวของเลือดผิดปกติและเลือดออกนั้นจะเพิ่มความเสี่ยงต่อการเกิดภาวะหลอดเลือดสมองตีบเนื่องจากจะเกิดลิ่มเลือดไหลไปอุดตันหลอดเลือดในสมองได้หรือปอดได้ เป็นต้น

ส่วนปัจจัยที่ตัวแบบไม่ได้นำไปใช้จำแนกผู้ป่วยโรคหลอดเลือดสมอง คือ กลุ่มโรคอ้วนและภาวะโภชนาการเกิน และภาวะเผาผลาญไขมันผิดปกติ อาจจะเนื่องมาจากจำนวนผู้ป่วยสูงอายุที่เข้ารับการรักษาที่โรงพยาบาลสมเด็จพระปิ่นเกล้ามีภาวะทั้งสองโรคค่อนข้างน้อยมากเมื่อเทียบกับปัจจัยอื่น ๆ โดยเฉพาะในกลุ่มผู้ป่วยที่เป็นโรคหลอดเลือดสมอง

## กิตติกรรมประกาศ

ขอขอบคุณโรงพยาบาลสมเด็จพระปิ่นเกล้า กรมแพทย์ทหารเรือ สำหรับความอนุเคราะห์ข้อมูลเวชระเบียนในการศึกษาครั้งนี้

## เอกสารอ้างอิง

- [1] International Health Policy Program (IHPP). The study of national burden of diseases and injuries among the Thai population in 2014. Nonthaburi: Graphico Systems; 2017.
- [2] Bureau of Non Communicable Diseases. Number and rate of patients in 2016 - 2018 (hypertension, diabetes, coronary heart disease, stroke, COPD); 2019. [Cites 16 June 2021]. Assessed from <http://www.thaincd.com/2016/mission/documents-detail.php?id=13684&tid=32&gid=1-020>.
- [3] Strategy and Planning Division, Office of the Permanent Secretary Ministry of Public Health. Public Health Statistics A.D.2019. Nonthaburi: Strategy and Planning Division; 2020.
- [4] Zhuo Y, Wu J, Qu Y, Yu H, Huang X, Zee B, et al. Clinical risk factors associated with recurrence of ischemic stroke within two years: A cohort study. *Medicine*. 2020;99(26):e20830.
- [5] Muntham D, Ingsrisawang L. An Application of Decision Tree Algorithms for Diagnosis of the Respiratory System: A Case Study of Pranakorn Sri Ayudthaya Hospital. *Journal of Health Systems Research*. 2010;4(1):73-81.
- [6] Hongboonmee N, Trepanichkul P. Comparison of Data Classification Efficiency to Analyze Risk Factors that Affect the Occurrence of Hyperthyroid using Data Mining Techniques. *Journal of Information Science and Technology*. 2019; 9(1):41-51.
- [7] Thanathamath P, Sirisathitkul Y. Improved Classification Techniques for Imbalanced Data Sets of Elderly's Knee Osteoarthritis. *Thai Science and Technology Journal*. 2019;27(6):1164-1178.
- [8] Boonchuay K, Sinapiromsaran K, Lursinsap C. Minority split and gain ratio for a class imbalance. *Proceeding of Eighth International Conference on Fuzzy Systems and Knowledge Discovery*; 2011 July 26-28; Shanghai, China.
- [9] Chawla NV, Bowyer KW, Hall LO, Kegelmayer WP. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligent Research*. 2002;16:321-357.
- [10] Paranya P. Improving Decision Tree Technique in Imbalanced Data Sets Using SMOTE for Internet Addiction Disorder Data. *Information Technology Journal*. 2016;12(1):54-63.
- [11] Gosain A, Sardana S. Handling class imbalance problem using oversampling techniques: A review. *Proceeding 2017 International Conference on Advances in Computing, Communications and Informatics*. 2017 September 13-16; Udupi, India. IEEE Xplore; 2017.
- [12] Fernández A, García S, Herrera F, Chawla NV. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15 - year Anniversary. *Journal Of Artificial Intelligence Research*. 2018;61:863-905.
- [13] RStudio Team. RStudio: Integrated Development for R. MA: RStudio, PBC; 2020. <http://www.rstudio.com>.
- [14] Torgo L. *Data Mining using R: learning with case studies*, CRC Press; 2010.
- [15] Gareth J, Daniela W, Trevor H, Robert T. *An introduction to statistical learning: with applications in R*, Springer; 2013.
- [16] Kaiyawan Y. Principle and Using Logistic Regression Analysis for Research. *Rajamangala University of Technology Srivijaya Research Journal*. 2012;4(1):1-12.

- [17] Stevens J. Applied multivariate statistics for the social science. New Jersey: Lawrence Erlbaum Associate, Inc; 1996.
- [18] Quinlan R. Introduction of decision trees. Machine Learning. 1986;1(1): 81-106.
- [19] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.
- [20] Breiman L, Friedman JH, Olshen R, Stone CJ. Classification and Regression Trees. California: Wadsworth International Group; 1984.
- [21] Therneau T, Atkinson B, Ripley B. The rpart package. (Version 4.1-13) [Software]; 2018. <https://cran.r-project.org/package=rpart>.
- [22] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics. 2006;15(3):651-674. doi: 10.1180/106186006X133933.
- [23] Hothorn T, Seibold H, Zeileis A. Partykit: A toolkit for recursive partytioning. (Version 1.2-2) [Software]; 2018. <https://CRAN.R-project.org/package=partykit>.
- [24] Maung ETW, Aye ZM. Comparison of Data Mining Classification Algorithms, C5.0 and CART for Car Evaluation and Credit Card Information Datasets. National Journal of Parallel and Soft Computing. 2019;1(1):75-80.
- [25] Breiman L. Bagging Predictors. Machine Learning. 1996;24(2):123-140.
- [26] Upadhayay A, Shukla S, Kumar S. Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set, International Journal of Computer Science & Communication Networks. 2013;3(1):64-68.
- [27] Nilnate N. Risk Factors and Prevention of Stroke in Hypertensive Patients. Journal of The Royal Thai Army Nurses. 2019;20(2):51-57.
- [28] Reddy HP, Jaganath A, Nagaraj N, Visweswara RYJ. A study of age as a risk factor in ischemic stroke of elderly. International Journal of Research in Medical Sciences. 2019;7(5):1553-1557.
- [29] Yousufuddin M, Young N. Aging and ischemic stroke. AGING. 2019;11(9):2542-2544.
- [30] Zaorsky NG, Zhang Y, Tchelebi LT, Mackley HB, Chinchilli VM, Zacharia BE. Stroke among cancer patients. Nature communications. 2019;10(1):5172. <https://doi.org/10.1038/s41467-019-13120-6>.