

ประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรคหลอดเลือดในสมอง

The Efficiency of Data Mining Technique for the Prognosis of Cerebrovascular Disease

กิตติศักดิ์ ขำจิตร¹, ดารภา ใจคุ้มเก่า¹, วัชรพงษ์ ภูมิรัง¹, อาทิตยา สัตนาโค¹, อนูปงศ์ สุขประเสริฐ^{2,*}
Kittisak Kumjit¹, Darapa Jaikoomkao¹, Watcharaphong Phumirang¹,
Artitaya Sattanako¹, Anupong Sukprasert^{2,*}

¹ สาขาคอมพิวเตอร์ธุรกิจ คณะการบัญชีและการจัดการ มหาวิทยาลัยมหาสารคาม มหาสารคาม 44150 ประเทศไทย

² อาจารย์ประจำสาขาคอมพิวเตอร์ธุรกิจ คณะการบัญชีและการจัดการ มหาวิทยาลัยมหาสารคาม มหาสารคาม 44150 ประเทศไทย

¹ Department of Business Computer Maharakham Business School, Maharakham University, Maharakham 44150, Thailand

² Lecturer at Maharakham Business School, Maharakham University, Maharakham 44150, Thailand

* Corresponding Author: Anupong Sukprasert, anupong.s@acc.msu.ac.th

Received:

7 March 2022

Revised:

22 March 2022

Accepted:

27 April 2022

คำสำคัญ:

การเปรียบเทียบประสิทธิภาพ,
เทคนิคเหมืองข้อมูล, โรคหลอดเลือด
ในสมอง

Keywords:

Efficiency Comparison,
Data Mining Technique,
Cerebrovascular Disease

บทคัดย่อ: งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูล และหาเทคนิคที่เหมาะสมที่สุดในการสร้างตัวแบบสำหรับพยากรณ์การเกิดโรคหลอดเลือดในสมอง โดยข้อมูลที่ใช้ในงานวิจัยนี้ เป็นข้อมูลผู้ป่วยโรคหลอดเลือดในสมอง จำนวน 5,110 ราย ซึ่งเป็นข้อมูลที่ได้จากโรงพยาบาลคลินิกซานการโลส ประเทศสเปน ซึ่งได้ถูกรวบรวมไว้ในเว็บไซต์ www.kaggle.com จากนั้นผู้วิจัยได้นำข้อมูลมาทำการวิเคราะห์ตามกระบวนการมาตรฐานของการทำเหมืองข้อมูล (CRISP-DM) โดยใช้เทคนิคการทำเหมืองข้อมูลจำนวน 4 เทคนิค ได้แก่ เทคนิคนาอิวเบย์ (Naive Bayes) เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคต้นไม้แบบสุ่ม (Random Forest) เพื่อสร้างแบบจำลองสำหรับพยากรณ์การเกิดโรคหลอดเลือดในสมองที่เหมาะสมที่สุด โดยทำการเปรียบเทียบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยตัวชี้วัดทั้ง 3 ค่า ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าประสิทธิภาพโดยรวม (F-measure) และค่าความไว (Sensitivity) ผลการวิจัยพบว่า เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เป็นเทคนิคที่มีความเหมาะสมที่สุดในการสร้างแบบจำลองสำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมองในครั้งนี้ โดยให้ค่าความแม่นยำ เท่ากับ 95.11% ค่าวัดประสิทธิภาพโดยรวม เท่ากับ 97.49% และค่าความไว เท่ากับ 99.98%

Abstract: The purpose of this research is to compare the efficiency of data mining techniques to find the suitable model for prognosis of cerebrovascular disease. The 5110 patients with cerebrovascular disease was used in this research. The data was collected from Clínico San Carlos Hospital, Spain via www.kaggle.com. The data was analyzed in 4 different techniques of data mining consisting of (1) Naïve Bayes, (2) Deep Learning, (3) Decision Tree, and (4) Random Forest to simulate cerebrovascular disease prediction. The efficiency of data classification was compared in 3 different criteria which are accuracy, f-measure, and sensitivity to determine the most appropriate simulation for prognosis. The result suggested that the Deep Learning technique was the most appropriate technique to simulate prognosis of cerebrovascular disease with 95.11% of accuracy, 97.49% of f-measure, and 99.98% of sensitivity.

1. บทนำ

โรคหลอดเลือดในสมอง (Cerebrovascular Disease หรือ Stroke) เป็นโรคที่พบได้บ่อยที่สุดของโรคระบบประสาท ซึ่งสาเหตุเกิดจากการแตก ตีบ อุดตันของเส้นเลือดในสมอง ทำให้เนื้อสมองบริเวณนั้นเกิดการตายเนื่องจากการขาดเลือดไปเลี้ยง ทำให้การทำหน้าที่ของสมองส่วนนั้นลดลง และความผิดปกติของการไหลเวียนเลือดที่ไปเลี้ยงสมอง โดยที่หลอดเลือดอาจมีการอุดตัน ตีบ หรือแตก อาการมักเกิดขึ้นอย่างรวดเร็ว และอยู่นานเกิน 24 ชั่วโมง โรคหลอดเลือดสมองแบ่งตามลักษณะของพยาธิสรีรวิทยาออกเป็น 2 ประเภท คือ โรคหลอดเลือดในสมองที่เกิดจากการขาดเลือด (Ischemic Stroke) และโรคหลอดเลือดในสมองที่เกิดจากเลือดออก (Hemorrhagic Stroke) เมื่อหลอดเลือดที่ไปเลี้ยงสมองเกิดตีบ ตัน หรือแตกจะก่อให้เกิดการเปลี่ยนแปลงในเนื้อสมองเป็นผลจากการขาดเลือด ออกซิเจน และกลูโคส จะทำให้มีการเปลี่ยนแปลงของเมตะบอลิซึมรวมทั้งสูญเสียการทำงานของ Blood Brain Barrier พบว่าความรุนแรงของโรคจะขึ้นอยู่กับขนาด และบริเวณของสมองที่ขาดเลือด และระยะเวลาของการขาดเลือด ซึ่งสมองจะทนต่อภาวะการขาดเลือดได้ในช่วงเวลาหนึ่ง และจะเสียการทำงานอย่างถาวรเมื่อขาดเลือดนานกว่า 3-6 ชั่วโมง อาการของผู้ป่วยจึงขึ้นอยู่กับตำแหน่งของสมองด้านที่มีพยาธิสภาพ อาการที่

เกิดขึ้นทันทีทันใด ได้แก่ อาการอ่อนแรงของกล้ามเนื้อบริเวณหน้า แขน หรือขาอ่อนแรง และเป็นซีกใดซีกหนึ่ง เดินเซ หรือเสียการทรงตัว พูดจาสับสน พูดไม่ชัด การมองเห็นภาพไม่ชัดอาจเป็นข้างเดียวหรือสองข้าง หรือมีอาการปวดศีรษะรุนแรง โดยไม่ทราบสาเหตุ และกล่าวผู้ป่วยจะมีอาการคงที่จะใช้เวลาประมาณ 1-14 วัน อาจจะกลายเป็นอัมพาตทันที รู้สึกตัวแต่มีกล้ามเนื้อแขนขาข้างที่เป็นจะอ่อนปวกเปียก อาจไม่รู้สีกตัวร่วมด้วย และจากการศึกษาของกูลาริกพบว่า เมื่อสมองส่วนนั้นได้รับอันตราย มีการสูญเสียการทำงานที่ของสมอง อาการที่พบจะขึ้นอยู่กับขนาด และตำแหน่งของพยาธิสภาพที่เกิดขึ้น ถ้าตำแหน่งของโรคอยู่ซีกขวาที่เป็นสมองเดิน จะทำให้พูดไม่ได้ กลืนลำบาก สูญเสียการรับรู้ความรู้สึกซีกซ้าย อัมพาตของซีกซ้าย การรับรู้ลดลงโดยเฉพาะสิ่งใหม่ๆ ถ้าตำแหน่งของโรคอยู่ซีกซ้ายจะทำให้พูดไม่ชัด สูญเสียความรู้สึก อัมพาตของซีกขวา อารมณ์เปลี่ยนแปลง (สำนักงานพัฒนานโยบายสุขภาพระหว่างประเทศ, 2560) ซึ่งโรคหลอดเลือดในสมองถือเป็นปัญหาสาธารณสุขที่สำคัญปัญหาหนึ่งของประชากรโลก จากรายงานขององค์การอนามัยโลกพบว่า โรคหลอดเลือดในสมองเป็นสาเหตุการตาย 10 อันดับแรกทั่วโลก และเป็นสาเหตุการตายอันดับที่ 3 ของประเทศสหรัฐอเมริกา และจะมีจำนวน 5.7 ล้านคนในแต่ละปีซึ่งจะเพิ่มมากขึ้นเป็น 6.3 ล้านคนในปี ค.ศ. 2015 และ 7.8 ล้านคน

ในปี ค.ศ. 2030 สำหรับในประเทศไทยพบว่า โรคหลอดเลือดในสมองเป็นสาเหตุการตายอันดับที่ 4 โดยมีผู้เสียชีวิตด้วยโรคนี้ถึง 11,663 คนในปี พ.ศ. 2562 และคิดเป็นอัตราการตายถึง 18.9 รายต่อประชากร 100,000 คน และในปี พ.ศ. 2563 มีอัตราการเกิดโรคหลอดเลือดในสมองในประชากรไทยเท่ากับ 77.4 รายต่อประชากร 100,000 คน แม้ว่าโรคหลอดเลือดในสมองจะมีอัตราการเกิดโรคสูงในกลุ่มอายุ 45 ปีขึ้นไปแต่ปัจจุบันพบว่า โรคหลอดเลือดในสมองเกิดในผู้ที่มีอายุตั้งแต่ 30 ปีขึ้นไปซึ่งในปัจจุบันโรคหลอดเลือดในสมองเป็นสาเหตุการตายที่สำคัญ อันดับที่ 3 รองจากโรคหัวใจ และโรคมะเร็ง อุบัติการณ์จะสูงขึ้นตามอายุ โดยเฉพาะผู้ที่มีอายุเกิน 65 ปี และจากสถิติผู้ป่วยโรคหลอดเลือดสมองรายใหม่ พบในโรงพยาบาลนครนายกในปี พ.ศ. 2563 พบว่ามีจำนวน 337 รายต่อปี และโดยเฉลี่ย 28 รายต่อเดือน (รัตนพร สายตรี และคณะ, 2562)

จากปัญหาข้างต้น ผู้วิจัยจึงให้ความสำคัญในการนำทฤษฎีการทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นกระบวนการวิเคราะห์ข้อมูลเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้นๆ และในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท เช่น การพยากรณ์ผู้ป่วยเพื่อพยากรณ์โอกาสการเกิดโรคหลอดเลือดในสมอง โดยสร้างแบบจำลองสำหรับการพยากรณ์และเปรียบเทียบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูล สำหรับงานวิจัยนี้ผู้วิจัยได้ใช้เทคนิคการทำเหมืองข้อมูลทั้งหมด 4 เทคนิค ได้แก่ เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอ็ฟเบย์ (Naïve Bayes) และเทคนิคต้นไม้แบบสุ่ม (Random Forest) ในการสร้างแบบจำลองเพื่อพยากรณ์โอกาสการเกิดโรคหลอดเลือดในสมอง โดยการแบ่งข้อมูลด้วยสำหรับการสร้างแบบจำลอง (Training set) และการทดสอบประสิทธิภาพของแบบจำลอง (Testing set) ด้วยเทคนิค

10-fold cross validation และใช้เกณฑ์สำหรับการวัดประสิทธิภาพของการจำแนกประเภทข้อมูลทั้ง 3 ค่า ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าประสิทธิภาพโดยรวม (F-measure) และค่าความไว (Sensitivity) เพื่อหาตัวแบบที่มีความเหมาะสมสำหรับการคัดกรองผู้ป่วยที่มีโอกาสการเกิดโรคหลอดเลือดในสมองต่อไป

2. วัตถุประสงค์

2.1 เพื่อสร้างแบบจำลองสำหรับพยากรณ์การเกิดโรคหลอดเลือดในสมอง

2.2 เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองที่ใช้สำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมอง

3. ขอบเขตงานวิจัย

3.1 ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลผู้ป่วยโรคหลอดเลือดในสมองที่มารับบริการในปี พ.ศ. 2563 ของโรงพยาบาลคลินิกซานการโลส ประเทศสเปน เป็นเวลา 1 ปี จำนวน 5,110 ระเบียน 12 แอตทริบิวต์ ที่ได้ถูกรวบรวมไว้ในเว็บไซต์ <https://www.kaggle.com> และจัดเก็บอยู่ในไฟล์ชื่อ healthcare-dataset-stroke-data.csv (Rahman, 2564)

3.2 ปัจจัยในการวิเคราะห์ผู้ป่วยหลอดเลือดในสมองในประเทศสเปน คือ โรคความผิดปกติทางร่างกายและพฤติกรรม มีจำนวน 8 แอตทริบิวต์ ได้แก่ รหัสผู้ป่วย (ID) เพศ (Gender) อายุ (Age) ความดันโลหิต (Hypertension) โรคหัวใจ (Heart_disease) ค่าดัชนีมวลกาย (BMI) สถานะการสูบบุหรี่ของผู้ป่วย (Smoking_status) และผลสรุปหลอดเลือดสมอง (Stroke)

3.3 เทคนิคการทำเหมืองข้อมูลที่ใช้วิเคราะห์คือ เทคนิคการจำแนกประเภทข้อมูล 4 เทคนิค ได้แก่ เทคนิคนาอ็ฟเบย์ (Naïve Bayes) เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เทคนิคต้นไม้ตัดสินใจ

(Decision Tree) และเทคนิคต้นไม้แบบสุ่ม (Random Forest)

3.4 โปรแกรมที่ใช้ในงานวิจัย คือ โปรแกรม RapidMiner Studio สำหรับการเตรียมข้อมูล การวิเคราะห์ข้อมูล และการทดสอบประสิทธิภาพของแบบจำลองสำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมอง (อนุพงศ์ สุขประเสริฐ, 2563)

4. แนวคิดและทฤษฎีที่เกี่ยวข้อง

การทำเหมืองข้อมูล (Data Mining) คือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหา รูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์รวมทั้งในด้านเศรษฐกิจและสังคม การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย มาสู่การจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล (ชนิดาภา บุญประสม และจรัญ แสนราช, 2561) ประเภทข้อมูลที่ใช้ทำเหมืองข้อมูลเทคนิคในการทำเหมืองข้อมูลมีดังนี้

4.1 กฎความสัมพันธ์ (Association Rule) เป็นการแสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุที่เกิดขึ้นพร้อมกัน ตัวอย่างของการประยุกต์ใช้กฎเชื่อมโยง เช่น การวิเคราะห์ข้อมูลการขายสินค้าโดยเก็บข้อมูลจากระบบ ณ จุดขาย (Point of Sales: POS) หรือร้านค้าออนไลน์ แล้วพิจารณาสินค้าที่ผู้ซื้อมักจะซื้อพร้อมกัน เช่น ถ้าพบว่าคนที่ซื้อเทปวีดีโอ มักจะซื้อเทปกาต้มน้ำ ร้านค้าก็อาจจะจัดร้านให้สินค้าสองอย่าง

อยู่ใกล้กันเพื่อเพิ่มยอดขาย (อนุพงศ์ สุขประเสริฐ, 2563)

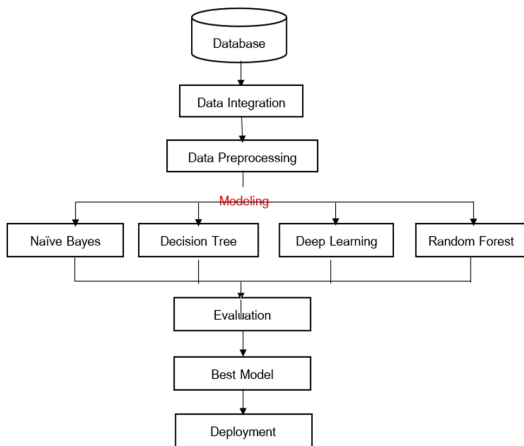
4.2 การจำแนกประเภทข้อมูล (Data Classification) เป็นการหากฎเพื่อระบุประเภทของวัตถุจากคุณสมบัติของวัตถุ เช่น หากความสัมพันธ์ระหว่างผลการตรวจร่างกายต่างๆ กับการเกิดโรค โดยใช้ข้อมูลผู้ป่วยและการวินิจฉัยของแพทย์ที่เก็บไว้ เพื่อนำมาช่วยวินิจฉัยโรคของผู้ป่วย หรือการวิจัยทางการแพทย์ ในทางธุรกิจจะใช้เพื่อดูคุณสมบัติของลูกค้าที่จะก่อหนี้ดีหรือหนี้เสีย เพื่อประกอบการพิจารณาการอนุมัติเงินกู้

4.3 การแบ่งกลุ่มข้อมูล (Data Clustering) เป็นการแบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่ม เช่นการแบ่งกลุ่มผู้ป่วยที่เป็นโรคเดียวกันตามลักษณะอาการ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์หาสาเหตุของโรค โดยพิจารณาจากผู้ป่วยที่มีอาการคล้ายคลึงกัน

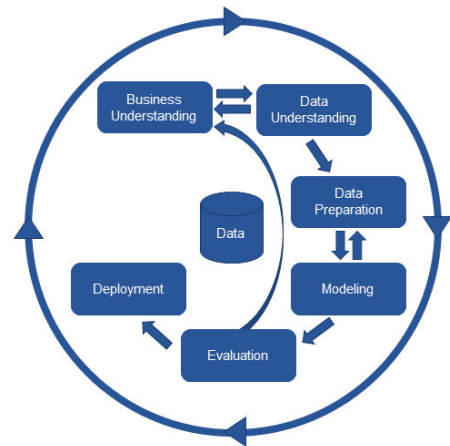
4.4 การสร้างมโนภาพ (Visualization) เป็นการสร้างภาพคอมพิวเตอร์กราฟิกที่สามารถนำเสนอข้อมูลได้อย่างครบถ้วนแทนการใช้ข้อความนำเสนอข้อมูลที่ อาจพบข้อมูลที่ซ่อนเร้นเมื่อดูข้อมูลชุดนั้นด้วยจินตทัศน์ (อนุพงศ์ สุขประเสริฐ, 2563)

5. วิธีดำเนินงานวิจัย

งานวิจัยนี้มีกรอบแนวคิดสำหรับการสร้างตัวแบบในการพยากรณ์การเกิดโรคหลอดเลือดในสมอง โดยใช้เทคนิคการทำเหมืองข้อมูล เพื่อให้ทราบลักษณะการเกิดโรคหลอดเลือดในสมอง และเพื่อหาแนวทางการรักษาผู้ป่วยโรคหลอดเลือดในสมองโดยแพทย์ผู้เชี่ยวชาญรักษาโดยตรง โดยมีขั้นตอนการดำเนินงาน ตามกระบวนการมาตรฐานในการทำเหมืองข้อมูล ดังภาพประกอบ 1



ภาพประกอบ 1 ขั้นตอนการดำเนินงาน



ภาพประกอบ 2 กระบวนการมาตรฐาน
การทำเหมืองข้อมูล

6. กระบวนการมาตรฐานในการทำเหมืองข้อมูล (Cross Standard Process for Data Mining)

วิธีการดำเนินงานวิจัยสำหรับการแบ่งกลุ่มผู้ป่วยโรคหลอดเลือดในสมองโดยใช้เทคนิคการทำเหมืองข้อมูล โดยอ้างอิงตามกระบวนการมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) ทั้ง 6 ขั้นตอน

6.1 การทำความเข้าใจเกี่ยวกับธุรกิจ (Business Understanding) ผู้วิจัยได้ศึกษาข้อมูลเกี่ยวกับการเกิดโรคหลอดเลือดในสมอง พบปัญหาที่เกิดขึ้นจริงเกี่ยวกับเกิดโรคหลอดเลือดในสมอง คือ ผู้ที่ป่วยโรคหลอดเลือดในสมอง ไม่ทราบว่าลักษณะของอาการที่จะเกิดการเกิดโรคหลอดเลือดในสมองมีลักษณะอาการใดบ้างที่มีความเสี่ยงที่จะเกิดโรคหลอดเลือดในสมอง สำหรับงานวิจัยนี้จึงจัดทำขึ้นเพื่อสร้างแบบจำลองสำหรับการพยากรณ์โอกาสการเกิดโรคหลอดเลือดในสมอง เพื่อหาแนวทางป้องกันให้กับผู้ที่มีโอกาสเสี่ยงเป็นโรคหลอดเลือดในสมอง เพื่อลดการสูญเสียที่จะเกิดขึ้นในอนาคต

6.2 การทำความเข้าใจเกี่ยวกับข้อมูล (Data Understanding) งานวิจัยนี้ใช้ชุดข้อมูลจริงเกี่ยวกับผู้ป่วยโรคหลอดเลือดในสมอง ที่เข้ารับการรักษา

ณ โรงพยาบาลคลินิกซานการ์โลส ประเทศสเปน ประจำปี พ.ศ. 2563 จำนวน 5,110 แถว และมีจำนวน 12 แอตทริบิวต์ ได้แก่ รหัสผู้ป่วย เพศ อายุ ความดันโลหิตสูง การเป็นโรคหัวใจ สถานะแต่งงาน ประเภทงานของผู้ป่วย ประเภทที่อยู่อาศัยของผู้ป่วย ระดับน้ำตาลในเลือด ค่าดัชนีที่ใช้วัดความสมดุลของน้ำหนักตัว สถานะการสูบบุหรี่ของผู้ป่วย สรุปผลโรคหลอดเลือดสมอง และได้ถูกรวบรวมไว้ในเว็บไซต์ www.kaggle.com ซึ่งจัดเก็บอยู่ในรูปแบบไฟล์ CSV (Rahman, 2020)

6.3 การเตรียมข้อมูล (Data Preparation) เป็นการเตรียมข้อมูลก่อนนำไปวิเคราะห์ในงานวิจัย ได้แบ่งขั้นตอนการเตรียมข้อมูลออกเป็น 4 ขั้นตอน

6.3.1 การคัดเลือกข้อมูล (Data Selection) ผู้วิจัยได้ตัดแอตทริบิวต์ที่ไม่สามารถอธิบายค่าข้อมูลอื่นๆ ได้ ซึ่งได้แก่ แอตทริบิวต์สถานะแต่งงาน (ever_married) และแอตทริบิวต์ ประเภทที่อยู่อาศัยของผู้ป่วย (Residence) ซึ่งจะเหลือแอตทริบิวต์ที่ใช้ในการวิเคราะห์ทั้งหมดมีจำนวน 10 แอตทริบิวต์ จำนวน 5,110 ระเบียบ

6.3.2 การกลั่นกรองข้อมูล (Data Cleaning) หลังจากสำรวจข้อมูล (Explore Data) พบว่ามีข้อมูลที่ไม่ทราบค่าเกิดขึ้นในแอตทริบิวต์ค่า

ตาราง 1 แสดงแอตทริบิวต์ (Attribute) ที่ใช้ในการวิเคราะห์

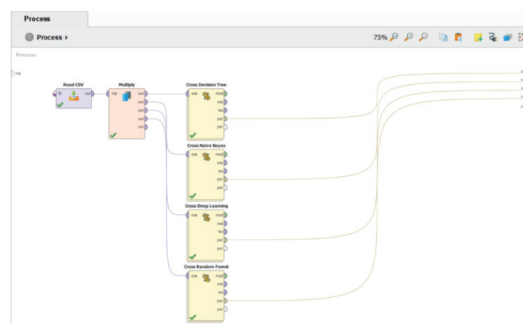
No.	Name	Data Type	Description
1	Id (ID)	Integer	รหัสผู้ป่วย
2	Gender	binominal	เพศ
3	Age	Integer	อายุ
4	Hypertension	binominal	ความดันโลหิตสูง
5	heart_disease	binominal	เป็นโรคหัวใจ
6	work_type	polynominal	ประเภทงานของผู้ป่วย
7	avg_glucose_level	real	ระดับน้ำตาลในเลือด
8	bmi	real	ค่าดัชนีที่ใช้ชี้วัดความสมดุลของน้ำหนักตัว
9	smoking_status	polynominal	สถานะการสูบบุหรี่ของผู้ป่วย
10	stroke (Label)	binominal	สรุปผลโรคหลอดเลือดสมอง

ดัชนีที่ใช้ชี้วัดความสมดุลของน้ำหนักตัว (bmi) ที่มีค่าเป็น “N/A” และในแอตทริบิวต์สถานะการสูบบุหรี่ของผู้ป่วย (smoking_status) ที่มีค่าเป็น “Unknown” ผู้วิจัยได้กำหนดให้ค่า “N/A” และ “Unknown” นี้เป็นค่าสูญหาย และไม่นำค่าสูญหายมาใช้ในการวิเคราะห์ในครั้งนี้

6.3.3 กำหนดหน้าที่ให้กับแอตทริบิวต์รหัสผู้ป่วย (ID) ให้มีหน้าที่เป็น “ID” เพื่อใช้ระบุกับข้อมูลที่มีความเป็นเอกลักษณ์ หรือไม่ซ้ำกับข้อมูลในแถวอื่นๆ ดังนั้น “รหัสผู้ป่วย” จึงไม่ได้ใช้สำหรับการวิเคราะห์ในครั้งนี้ และกำหนดหน้าที่ให้กับแอตทริบิวต์สรุปผลโรคหลอดเลือดสมอง (stroke) ให้มีหน้าที่เป็น “Label” เพื่อใช้ระบุกับข้อมูลที่เป็นคลาสคำตอบ

6.3.4 การแปลงรูปข้อมูล (Data Transformation) งานวิจัยนี้บันทึกข้อมูลจัดเก็บในโปรแกรม Microsoft Excel อยู่ในรูปไฟล์นามสกุล .CSV ดังแสดงในตาราง 1 เพื่อนำไปใช้วิเคราะห์ข้อมูลต่อไป

6.4 การสร้างแบบจำลอง (Modeling) ขั้นตอนนี้จะทำการแบ่งข้อมูลสำหรับการสร้างแบบจำลองและทดสอบประสิทธิภาพของโมเดลโดยใช้เทคนิค 10-fold cross validation ด้วยโปรแกรม RapidMiner Studio ในการสร้างแบบจำลองและประเมินประสิทธิภาพของแบบจำลองในแต่ละแบบจำลอง ดังแสดงในภาพประกอบ 3



ภาพประกอบ 3 แสดงขั้นตอนการสร้างโมเดลการวัดประสิทธิภาพโมเดล

สำหรับเทคนิคการจำแนกประเภทข้อมูลที่
ใช้ในงานวิจัย ประกอบด้วย 4 เทคนิค ได้แก่ เทคนิค
นาอ็ฟเบย์ (Naïve Bayes) เทคนิคการเรียนรู้เชิงลึก
(Deep Learning) เทคนิคต้นไม้ตัดสินใจ (Decision
Tree) และเทคนิคต้นไม้แบบสุ่ม (Random Forest)
ซึ่งมีรายละเอียด ดังนี้

6.4.1 เทคนิคนาอ็ฟเบย์ (Naïve Bayes)
เป็นเทคนิคที่ใช้ทฤษฎีความน่าจะเป็นตามกฎของเบย์
(Bayes' Theorem) (รุ่งโรจน์ บุญมา และ นิเวศ
จิระวิจิตชัย, 2562) เพื่อหาสมมติฐานใดน่าจะถูกต้อง
ที่สุด โดยใช้ความรู้ก่อนหน้า (Prior Knowledge)
ได้แก่ ความน่าจะเป็นก่อนหน้าสำหรับสมมติฐาน
หนึ่งๆ ร่วมกับข้อมูล เช่น ความน่าจะเป็นที่สังเกตได้
สำหรับสมมติฐานหนึ่งๆ เพื่อหาสมมติฐานที่ดีที่สุด
การเรียนรู้แบบเบย์อาศัยหลักการของการคำนวณ
ความน่าจะเป็นของแต่ละสมมติฐาน ในที่นี้คือคลาส
เป้าหมายหรือผลลัพธ์การทำนายโดยการเรียนรู้แบบ
เบย์เป็นการเรียนรู้เพิ่มเติม เนื่องจากตัวอย่างใหม่ที่ได้
มาถูกนำมาปรับเปลี่ยนการแจกแจงซึ่งมีผลต่อ
การเพิ่มหรือลดความน่าจะเป็นทำให้มีการเรียนรู้
ที่เปลี่ยนไป วิธีการนี้ตัวแบบจะถูกปรับเปลี่ยนไปตาม
ตัวอย่างใหม่ที่ได้โดยผนวกกับความรู้เดิมที่มี ซึ่งการ
ทำนายค่าคลาสเป้าหมายของตัวอย่างใช้ความน่าจะเป็น
มากที่สุดของทุกสมมติฐานจากทฤษฎีของเบย์
เราสามารถคำนวณความน่าจะเป็นของสมมติฐาน
ต่างๆ โดยใช้สมการ ดังสมการ (1)

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (1)$$

เมื่อ

$P(h)$ คือ ความน่าจะเป็นก่อนหน้าของ
สมมติฐาน h

$P(D)$ คือ ความน่าจะเป็นก่อนหน้าของชุด
ข้อมูลตัวอย่าง D

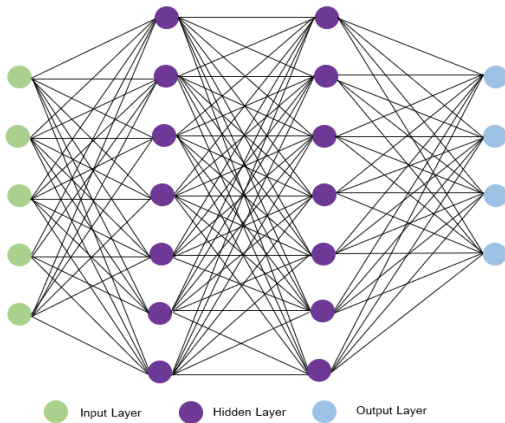
$P(h|D)$ คือ ความน่าจะเป็นแบบมีเงื่อนไข
ของสมมติฐาน h ภายใต้ข้อมูล D

$P(D|h)$ คือ ความน่าจะเป็นแบบมีเงื่อนไข
ของชุดข้อมูล D ภายใต้สมมติฐาน h

ความน่าจะเป็นที่ h เกิดก่อน และ D เกิด
ตามมาความ น่าจะเป็นที่ D และ h เกิดร่วมกัน โดย
ใช้สมการ เนอ็ฟเบย์ ดังสมการ (2)

$$P(h|D) = \frac{P(h \cap D)}{P(D)} \quad (2)$$

6.4.2 เทคนิคการเรียนรู้เชิงลึก (Deep
Learning) เป็นเทคนิคในกลุ่มโครงข่ายประสาท
เทียม (Artificial Neural Network: ANN) ที่มี
โครงสร้างขนาดใหญ่ประกอบด้วยนิรอนและชั้นซ่อน
จำนวนมากเป็นอัลกอริทึมที่ถูกสร้างขึ้นมาเพื่อการ
เรียนรู้ของเครื่องจักรแต่ละระดับ Hidden Layer ของ
การเรียนรู้เชิงลึกมีมากกว่า ANN ซึ่ง แต่ละเลเยอร์
จะเปรียบเสมือนประกอบด้วยเซลล์ประสาท (Neural)
จำนวนมากที่มีหน้าที่ในการประมวลผลโดยเลเยอร์
แรกสุดจะทำหน้าที่ในการรับข้อมูล (Input Layer)
และส่งข้อมูลที่ประมวลผลเสร็จแล้วไปยังเลเยอร์
สุดท้าย (Output Layer) การส่งข้อมูลแบบนี้มีข้อดีคือ
แต่ละเลเยอร์อาจทำให้มีค่าถ่วงน้ำหนัก (Weight) ค่า
ความเอนเอียงของข้อมูล (Bias) และวิธีการประมวล
ผลทางคณิตศาสตร์ (Activation Function) เป็นอิสระ
ต่อกันถ้าป้อนข้อมูลเข้าไปให้กับโมเดลมากเท่าไร
แต่ละเลเยอร์ก็จะสามารถสกัดคุณลักษณะที่มีความ
ซับซ้อนมากขึ้นทำให้ระบบสามารถตัดสินใจได้ใกล้เคียง
กับมนุษย์มากยิ่งขึ้น (สมศักดิ์ ศรีสวการย์ และ
สมัย ศรีสว, 2563) ดังภาพประกอบ 4



ภาพประกอบ 4 แสดงโครงข่ายประสาทเทียม
แบบเชิงลึก

6.4.3 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคในกลุ่มการจำแนกประเภทข้อมูล ซึ่งจะจำลองลักษณะกึ่งก้านของต้นไม้จะประกอบด้วยโหนดแทนคุณลักษณะ และโหนดล่างสุดแทนหมวดหมู่การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าที่ใช้จะมาจากการคำนวณจากค่า Information Gain การสร้างต้นไม้ตัดสินใจ C4.5 ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ถ้าให้ชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1, m_2, \dots, m_n\}$ และให้ความน่าจะเป็นที่จะเกิดค่า m_i มีค่าเท่ากับ $P(m_i)$ จะได้ค่าเกนสารสนเทศ (Information Gain) ของ M เขียนแทนด้วย $I(M)$ (เดช ธรรมศิริ และพยุ่ง มีสัจ, 2556) คำนวณได้ดังสมการ (3)

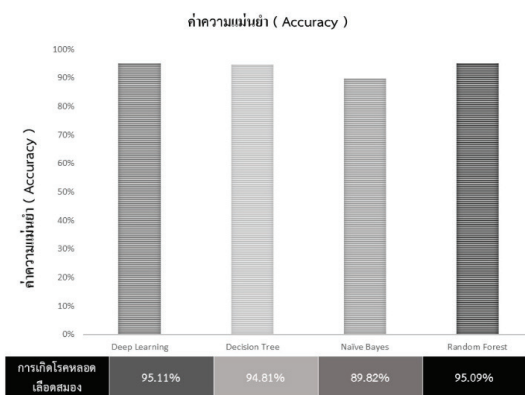
$$I(M) = \sum_{i=1}^n -P(m_i) \log_2 P(m_i) \quad (3)$$

คำนวณค่ามาตรฐานอัตราส่วนเกน (Gain ratio) ได้จาก $\text{Gain Ratio} = \text{Gain} - \text{Split Information}$ ท้ายสุดจึงเลือกค่า Gain ratio สูงสุดเป็นคุณลักษณะเริ่มต้น และเลือกคุณสมบัติถัดไปตามค่า Gain ratio

น้อยลงตามลำดับ (วัชรวัชรณ จิตต์สกุล และ สุนันทา สดสี, 2560)

6.4.4 เทคนิคต้นไม้แบบสุ่ม (Random Forest) เป็นอัลกอริทึมประเภทหนึ่งของอัลกอริทึมต้นไม้ตัดสินใจที่มีลักษณะแบบ Unpruned หรือ Regression Trees ซึ่งถูกสร้างจากการนำข้อมูลไปสุ่มเลือกตัวอย่างข้อมูล หลักการของ Random Forest คือ สร้าง model จาก Decision Tree หลายๆ model (ตั้งแต่ 10 model ถึง มากกว่า 1000 model) โดยแต่ละ model จะได้รับ data set ไม่เหมือนกัน ซึ่งเป็น subset ของ data set ทั้งหมด ตอนทำ prediction ก็ให้แต่ละ Decision Tree ทำ prediction แต่ละตัว และคำนวณผล prediction ด้วยการ vote output ที่ถูกเลือกโดย Decision Tree มากที่สุด หรือ หาค่าเฉลี่ยจาก output ของแต่ละ Decision Tree (กรณี regression) Decision Tree แต่ละ model ใน Random Forest ถือว่าเป็น weak learner เป็น model ที่ไม่เก่งเท่าไร แต่เมื่อนำเอาแต่ละ Decision Tree มาทำการพยากรณ์ร่วมกัน ก็จะได้ model ที่มีความเก่ง และแม่นยำมากกว่า Decision Tree ที่ทำการพยากรณ์แบบเดี่ยว (ปิยวรรณ นิลธนน และคณะ, 2563)

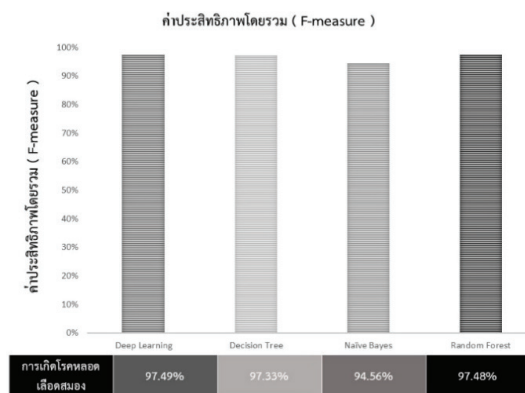
6.5 การประเมินผล (Evaluation) สำหรับการประเมินผลเพื่อทดสอบประสิทธิภาพของการสร้างแบบจำลองการพยากรณ์การเกิดโรคหลอดเลือดในสมอง ได้ทำการแบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธี Cross Validation คือ ส่วนสำหรับสร้างตัวแบบ (training set) และส่วนสำหรับการทดสอบ (testing set) โดยมีการแบ่งข้อมูลออกเป็น 10 ส่วน (10-fold cross validation) วัดค่าประสิทธิภาพของการจำแนกประเภทข้อมูล ด้วย ค่าความแม่นยำ (Accuracy) ค่าประสิทธิภาพโดยรวม (F-measure) และค่าความไว (Sensitivity) โดยมีผลการทดสอบประสิทธิภาพแสดงจากภาพประกอบ 5-7 ดังนี้



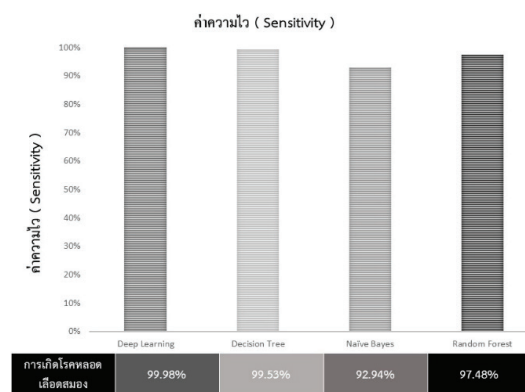
ภาพประกอบ 5 แสดงการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลด้วยค่าความแม่นยำ (Accuracy)

จากภาพประกอบ 5 แสดงการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลด้วยค่าความแม่นยำ (Accuracy) โดยใช้เทคนิคนาอิวเบย์ (Naive Bayes) เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคต้นไม้แบบสุ่ม (Random Forest) ในการพยากรณ์การเกิดโรคผลปรากฏว่าในข้อมูลการเกิดโรคหลอดเลือดในสมอง พบว่า เทคนิคการเรียนรู้เชิงลึกสามารถสร้างแบบจำลองได้ค่าความแม่นยำ ในการพยากรณ์สูงสุดเท่ากับ 95.11% เทคนิคต้นไม้ตัดสินใจให้ค่าความแม่นยำ เท่ากับ 94.81% เทคนิค เทคนิคต้นไม้แบบสุ่ม ให้ค่าความแม่นยำ เท่ากับ 95.09% และคือเทคนิคนาอิวเบย์ ให้ค่าความแม่นยำ เท่ากับ 89.82% ตามลำดับ

จากภาพประกอบ 6 แสดงการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลด้วยค่าประสิทธิภาพโดยรวม (F-measure) โดยใช้เทคนิคนาอิวเบย์ (Naive Bayes) เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และ เทคนิคต้นไม้แบบสุ่ม (Random Forest) ในการพยากรณ์การเกิดโรคผลปรากฏว่าในข้อมูลการเกิดโรคหลอดเลือดในสมอง พบว่า เทคนิคการเรียนรู้เชิงลึก สามารถสร้างแบบจำลองได้ค่าประสิทธิภาพโดยรวม ในการพยากรณ์สูงสุดเท่ากับ 97.49%



ภาพประกอบ 6 แสดงการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลด้วยค่าประสิทธิภาพโดยรวม (F-measure)



ภาพประกอบ 7 แสดงการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลด้วยค่าความไว (Sensitivity)

เทคนิคต้นไม้ตัดสินใจ ให้ค่าประสิทธิภาพโดยรวม เท่ากับ 97.33% เทคนิคต้นไม้แบบสุ่ม ให้ค่าประสิทธิภาพโดยรวม เท่ากับ 97.48% และน้อยที่สุดคือเทคนิคนาอิวเบย์ ให้ค่าประสิทธิภาพโดยรวม เท่ากับ 94.56%

ภาพประกอบ 7 แสดงการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลด้วยค่าความไว (Sensitivity) โดยใช้เทคนิคนาอิวเบย์ (Naive Bayes) เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคต้นไม้แบบสุ่ม (Random Forest) ในการพยากรณ์การ

เกิดโรคผลปรากฏว่าในข้อมูลการเกิดโรคหลอดเลือดในสมอง พบว่า เทคนิคการเรียนรู้เชิงลึก สามารถสร้างแบบจำลองที่ค่าความไว ในการพยากรณ์สูงที่สุดเท่ากับ 99.98% เทคนิคต้นไม้ตัดสินใจ ให้ค่าความไวเท่ากับ 99.53% เทคนิคต้นไม้แบบสุ่ม ให้ค่าความไวเท่ากับ 97.48% และน้อยที่สุดคือเทคนิคนาอ์ฟเบย์ ให้ค่าความไว เท่ากับ 97.48%

6.5 การนำไปใช้งาน (Deployment)

เมื่อทำการวิเคราะห์ตามกระบวนการมาตรฐานในการทำเหมืองข้อมูลทั้ง 5 ขั้นตอนแล้ว พบว่าเทคนิคการจำแนกประเภทข้อมูลด้วยเทคนิคการเรียนรู้เชิงลึก (Deep Learning) เป็นเทคนิคที่มีความเหมาะสมที่สุดในการสร้างแบบจำลองสำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมอง เนื่องจากการทดสอบประสิทธิภาพของแบบจำลอง ค่าความแม่นยำ (Accuracy) ของเทคนิค 3 เทคนิคมีความใกล้เคียงกัน ได้แก่ เทคนิคการเรียนรู้เชิงลึก ให้ค่าความแม่นยำเท่ากับ 95.17% เทคนิคต้นไม้แบบสุ่ม (Random Forest) ให้ค่าความแม่นยำเท่ากับ 95.09% และเทคนิคต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความแม่นยำเท่ากับ 94.81% ดังนั้น ผู้วิจัยจึงเลือกเทคนิคการเรียนรู้เชิงลึก เนื่องจากเมื่อเปรียบเทียบกับสองเทคนิคแล้วค่าประสิทธิภาพไม่ต่างกันมาก เนื่องจากเป็นเทคนิคที่ไม่มีความซับซ้อนในการวิเคราะห์

และอ่านผลลัพธ์ได้ง่าย จึงเหมาะสำหรับการนำไปใช้สร้างตัวแบบในการพยากรณ์ การพยากรณ์การเกิดโรคหลอดเลือดสมองต่อไป

7. สรุปผล

จากการเปรียบเทียบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูลสำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมอง ครั้งนี้ ใช้ชุดข้อมูลจริงเกี่ยวกับผู้ป่วยโรคหลอดเลือดในสมองของประเทศสเปน จากเว็บไซต์ www.kaggle.com ศึกษาประสิทธิภาพของมาวิเคราะห์ตามกระบวนการมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) โดยใช้เทคนิคการจำแนกประเภทข้อมูล 4 เทคนิค ได้แก่ เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เทคนิคนาอ์ฟเบย์ (Naïve Bayes) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคต้นไม้แบบสุ่ม (Random Forest) นำมาสร้างแบบจำลองสำหรับพยากรณ์การเกิดโรคหลอดเลือดในสมองและทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลทั้ง 4 เทคนิคนี้ด้วยเกณฑ์การวัดประสิทธิภาพทั้ง 3 ค่า ได้แก่ ค่าความแม่นยำ (Accuracy) ค่า ประสิทธิภาพโดยรวม (F-measure) และค่าความไว (Sensitivity) ซึ่งผลของการวิเคราะห์ประสิทธิภาพของตัวแบบจำลองสำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมอง ดังแสดงในตาราง 2

ตาราง 2 การเปรียบเทียบค่าทดสอบประสิทธิภาพของตัวแบบจำลองสำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมอง

เทคนิค การจำแนกประเภทข้อมูล	ค่าทดสอบประสิทธิภาพการจำแนกประเภทข้อมูล		
	Accuracy	F-measure	Sensitivity
Naïve Bayes	89.82%	94.56%	92.94%
Decision Tree	94.81%	97.33%	99.53%
Random Forest	95.09%	97.48%	99.96%
Deep Learning*	95.17%	97.52%	99.99%

* เทคนิคที่มีความเหมาะสมสำหรับการสร้างแบบจำลองการพยากรณ์การเกิดโรคหลอดเลือดในสมอง

จากตาราง 2 พบว่า เทคนิคการเรียนรู้เชิงลึก (Deep Learning) เป็นเทคนิคที่มีความเหมาะสมที่สุดในการสร้างแบบจำลองสำหรับการพยากรณ์การเกิดโรคหลอดเลือดในสมอง โดยให้ความแม่นยำสูงสุด เท่ากับ 95.17% ค่าประสิทธิภาพโดยรวม เท่ากับ 97.52% และค่าความไว เท่ากับ 99.99% จึงสรุปได้ว่าเทคนิคการเรียนรู้เชิงลึก มีความเหมาะสมในการนำสร้างแบบจำลองเพื่อพยากรณ์โอกาสการเกิดโรคหลอดเลือดในสมอง เพราะสามารถจัดการกับข้อมูลที่มีหลายตัวแปรได้เป็นอย่างดี เป็นเทคนิคที่ไม่มีความซับซ้อนในการวิเคราะห์ อ่านผลลัพธ์ได้ง่าย และมีประสิทธิภาพน่าเชื่อถือมากที่สุด

8. ข้อเสนอแนะ

1. ชุดข้อมูลที่ใช้ในงานวิจัยครั้งนี้เป็นข้อมูลผู้ป่วยที่ได้มาจากประเทศสเปน ซึ่งถ้าหากนำมาใช้สำหรับการพยากรณ์ความเสี่ยงโรคหลอดเลือดในสมองกับผู้ป่วยในประเทศไทยก็อาจจะไม่เหมาะสม ดังนั้นจึงต้องศึกษาข้อมูลลักษณะประชากรของผู้ป่วยในประเทศไทย

2. ข้อมูลที่นำมาวิเคราะห์ในครั้งนี้ยังมีน้อยจำนวนเกินไป และค่า accuracy มีค่าที่สูงมาก จึงมีข้อจำกัดเกี่ยวกับ แอตทริบิวต์ ที่นำมาใช้ในการพยากรณ์ ซึ่งผลสรุปของงานวิจัยไม่สามารถไปใช้กับข้อมูลที่มี แอตทริบิวต์ อื่นๆ นอกเหนือจากนี้ได้

3. การเตรียมข้อมูลมีความสำคัญมากในการทำเหมืองข้อมูล ผู้วิจัยควรศึกษาข้อมูลให้ครบถ้วนและทำความเข้าใจในข้อมูล ความสำคัญของแอตทริบิวต์ต่างๆ ที่มีความสำคัญในการนำมาวิเคราะห์ข้อมูลในการทำเหมืองข้อมูล การผสมผสานข้อมูลเข้าด้วยกัน และจัดทำข้อมูลให้เป็นกลุ่มเดียวกัน

9. เอกสารอ้างอิง

ชนิตาภา บุญประสม และจรัญ แสนราช. (2561). การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล. *วารสารวิชาการครุศาสตร์อุตสาหกรรม พระจอมเกล้าพระนครเหนือ*, 9(1). 142-151.

เดช ธรรมศิริ และพยุ่ง มีสัจ. (2556). การจำแนกข้อมูลด้วยวิธีแบบร่วมกันตัดสินใจจากพื้นฐานของเทคนิคต้นไม้ตัดสินใจเทคนิคโครงข่ายประสาทเทียม และเทคนิคซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการเลือกตัวแทนที่เหมาะสมด้วยขั้นตอนวิธีเชิงพันธุกรรม. *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, 21(2), 293-303.

ปิยวรรณ นิลถนนอม, ธนพร มาลัย และสายชล สินสมบุญทอง. (2563). การเปรียบเทียบประสิทธิภาพการทำนายผลการแปลงข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล. *Thai Journal of Science and Technology*, 10(1), 14-25.

รัตนพร สายตรี, ปุณณพัฒน์ ไชยมะล และสมเกียรติยศ วรเดช. (2562). ความสามารถในการประกอบกิจวัตรประจำวันของผู้ป่วยโรคหลอดเลือดสมอง. *วารสารวิชาการสาธารณสุขชุมชน*, 5(2), 1-13.

รุ่งโรจน์ บุญมา และนิเวศ จิระวิจิตชัย. (2562). การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล. *วารสารวิชาการชาชนนุศาสตร์*, 3(2), 11-19.

วัชรวิวรรณ จิตต์สกุล และสุนันทา สดสี. (2560). การวิเคราะห์การจำแนกข้อความด้วยการเปรียบเทียบความเสถียรของอัลกอริทึม. *ศรีปทุมปริทัศน์ ฉบับวิทยาศาสตร์และเทคโนโลยี*, 9, 19-31.

สมศักดิ์ ศรีสุวรรณ และสมัย ศรีสว. (2563). การวิเคราะห์เหมืองความคิดเห็นโดยใช้เทคนิคการสกัดคำ. *วารสารวิชาการการประยุกต์ใช้เทคโนโลยีสารสนเทศ*, ปีที่ 6 (2), 95-104.

สำนักงานพัฒนานโยบายสุขภาพระหว่างประเทศ. (2560). รายงานภาระโรคและการบาดเจ็บของประชากรไทย พ.ศ. 2557. *Burden of Disease Thailand*. สืบค้น 5 ตุลาคม 2564, สืบค้นจาก <http://bodthai.net/download/รายงานภาระโรคและการบาดเจ็บ>

อนุพงศ์ สุขประเสริฐ. (2563). *คู่มือการทำเหมืองข้อมูลด้วยโปรแกรม RapidMiner Studio*. พิมพ์ครั้งที่ 3. สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะการบัญชีและการจัดการมหาวิทยาลัยมหาสารคาม:มหาสารคาม.

Rahman, M. (2020). Stroke Prediction Dataset. *kaggle*. Retrieved 25 September 2021, Retrieved from <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/version/1>