การเปรียบเทียบประสิทธิภาพการจำแนกระดับฮีโมโกลบินของผู้บริจาคโลหิต ด้วยเทคนิคการเรียนรู้ของคอมพิวเตอร์

Efficiency Comparison of Machine Learning Classification Techniques for Hemoglobin Levels of Blood Donors

สาธิต เทศสมบูรณ์* และ เทวฤทธิ์ สะระชนะ

Sathit Tedsomboon* and Tewarit Sarachana

คณะสหเวชศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพ ประเทศไทย

Faculty of Allied Health Sciences, Chulalongkorn University, Bangkok, Thailand

**Corresponding author, E-mail: sathidball@gmail.com*

บทคัดย่อ

โลหิตและส่วนประกอบโลหิตที่เพียงพอมีความสำคัญกับผู้ป่วย จึงต้องรักษาผู้บริจาคโลหิตรายเดิมให้บริจาคโลหิตได้อย่างต่อเนื่อง ศูนย์บริการโลหิตแห่งชาติ สภากาชาดไทยมีผู้ถูกปฏิเสธให้บริจาคโลหิตประมาณร้อยละ 15-20 ส่วนใหญ่เกิดจากค่าฮีโมโกลบิน (hemoglobin; Hb) ไม่ผ่านเกณฑ์ หากสามารถพยากรณ์ผลตรวจ Hb ได้ล่วงหน้าจะช่วยลดผลกระทบปัญหาการถูกปฏิเสธ การศึกษานี้มีวัตถุประสงค์เปรียบเทียบประสิทธิภาพการจำแนกผลตรวจ Hb ในผู้บริจาคโลหิตระหว่างเทคนิคต้นไม้ตัดสินใจ (Decision tree) และโครงข่ายประสาทเทียม (Artificial neural networks) โดยชุดข้อมูลของผู้ที่บริจาคโลหิตจำนวน 40 ตัวแปร จากผู้มาบริจาคโลหิตตั้งแต่ 1 ต.ค. 2561 ถึง 31 ธ.ค. 2561 ณ ภาคบริการโลหิตแห่งชาติ 12 แห่งและสถานีกาชาดหัวหินเฉลิมพระเกียรติ จำนวน 1,908 ราย ทำการเตรียมข้อมูลให้เหมาะสมต่อการวิเคราะห์ โดยใช้เทคนิคต้นไม้ตัดสินใจ (Decision tree) และโครงข่ายประสาทเทียม (Artificial neural networks) ในวิเคราะห์ตัวแปรที่ผ่านการคัดเลือกแล้วเพื่อการจำแนกกลุ่มค่าฮีโมโกลบินของผู้บริจาคโลหิต พบว่าตัวแบบพยากรณ์ต้นไม้ตัดสินใจและตัวแบบพยากรณ์โครงข่ายประสาทเทียมให้ค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ค่าการพยากรณ์ผลบวก ค่าการพยากรณ์ผลลบและค่า AUC เท่ากับ 86.13/85.60, 23.44/37.50, 98.74/95.28, 78.95/61.54, 86.50/88.34 และ 0.844/0.865 ตามลำดับ จากการศึกษานี้อาจนำไปพัฒนาต่อเป็นระบบประเมินออนไลน์ก่อนเดินทางมาบริจาคโลหิต

คำสำคัญ: *ต้นแบบพยากรณ์, การเรียนรู้ของเครื่อง, ฮีโมโกลบิน, ธนาคารเลือด, เหมืองข้อมูล, สารสนเทศศาสตร์สุขภาพ*

Abstract

Adequate blood and blood components are important to patients. Therefore, the maintenance of the rate of repeated blood donors is a necessity. National Blood Centre have had rejected 15-20% of blood donors due to their low hemoglobin (Hb) level. If the hemoglobin level is predictable, the rejection will be less. This study aims to compare the efficiency between the decision tree and artificial neural networks for classification of Hb level in blood donors by using datasets regarding 40 variables of blood donor from 1 October to 31 December 2018 at 13 Regional Blood Centers, 1,908 cases. In addition, we were data cleaning and using Decision tree and Artificial neural networks classifier for analysis of selected variables to classify Hb level. We found Decision tree and Artificial neural networks show accuracy sensitivity specificity positive predictive value negative predictive value and AUC were 86.13/85.60, 23.44/37.50, 98.74/95.28, 78.95/61.54, 86.50/88.34 and 0.844/0.865, respectively. This study could lead to further development of an online assessment application.

Keywords: *predictive model, machine learning, hemoglobin, blood bank, data mining, health informatics*

1. บทนำ

เป็นที่ทราบดีว่าโลหิตและส่วนประกอบโลหิตเป็นสิ่งสำคัญในการรักษาผู้ป่วย โลหิตที่ปลอดภัยต้องเป็นโลหิตที่ได้รับจากผู้บริจาคโลหิตที่มีความเสี่ยงต่ำ (Organization, 2017) ซึ่งโลหิตที่มีความปลอดภัยสูงสุดคือโลหิตที่ได้รับจากผู้บริจาคโลหิตประจำไม่หวังสิ่งตอบแทน ดังนั้นหน่วยงานที่ทำหน้าที่ในการจัดหาโลหิตจำเป็นต้องรักษาผู้บริจาคโลหิตรายเก่าให้สามารถบริจาคโลหิตได้ต่อเนื่อง ทำให้ได้โลหิตที่มีคุณภาพมีปริมาณเพียงพออย่างสม่ำเสมอ

จากข้อมูลการจัดหาโลหิตของภาคบริการโลหิตแห่งชาติ สภากาชาดไทยพบว่ามีผู้ถูกปฏิเสธให้บริจาคโลหิตประมาณร้อยละ 15-20 โดยร้อยละ 50 ของผู้ที่ถูกปฏิเสธให้บริจาคโลหิตมีสาเหตุจากค่าฮีโมโกลบิน (hemoglobin; Hb) ต่ำกว่าเกณฑ์ ซึ่งมักพบในผู้บริจาคโลหิตประจำโดยเฉพาะผู้บริจาคโลหิตเพศหญิง ฮีโมโกลบินคือโมเลกุลของโปรตีนที่อยู่บนเม็ดเลือดแดง ทำหน้าที่ในการขนส่งออกซิเจนจากปอดไปยังเนื้อเยื่อต่างๆทั่วร่างกายและขนส่งคาร์บอนไดออกไซด์จากเนื้อเยื่อกลับเข้าสู่ปอด ศูนย์บริการโลหิตแห่งชาติ สภากาชาดไทยกำหนดให้ผู้หญิงต้องมีค่าไม่น้อยกว่า 12 mg/dl และผู้ชายต้องมีค่าไม่น้อยกว่า 13 mg/dl จากปัญหาดังกล่าวนอกจากจะทำให้การจัดหาโลหิตไม่เป็นไปตามเป้าหมายแล้ว ยังส่งผลให้จำนวนผู้บริจาคโลหิตรายเก่ามีจำนวนน้อยลงในทุกปีส่งผลกระทบต่อปริมาณโลหิต(วิชุดา กลิ่นหอม วรรณวิมล มีคงและสมรัก เพชรโฉมฉาย, 2015; วิไลภรณ์ วงษ์กิติโสภณ ศิริรักษ์ ศุภธีระธาดา และกัลยลักษณ์ คลับคล้าย, 2015)

ปัจจุบันได้มีการนำข้อมูลต่างๆจำนวนมากมาวิเคราะห์เพื่อค้นหาความรู้ใหม่ที่เป็นประโยชน์ที่ซ้อนอยู่ ซึ่งเรียกว่าการทำเหมืองข้อมูล (data mining) และยังสามารถนำความรู้ที่ได้มาใช้พยากรณ์สิ่งต่างๆได้ล่วงหน้า (Kaur, Singh, & Josan, 2015) โดยอาศัยวิธีการเรียนรู้ของคอมพิวเตอร์ (machine learning; ML) เป็นวิธีการนำข้อมูลมาฝึกให้คอมพิวเตอร์ได้เรียนรู้วิธีหาคำตอบซึ่งเรียกว่าชุดข้อมูลฝึกหัด (training dataset) จากนั้นนำวิธีการหาคำตอบที่ได้มาทดสอบกับข้อมูลอีกชุดเพื่อประเมินประสิทธิภาพเรียกว่าชุดข้อมูลทดสอบ (testing dataset) เมื่อได้วิธีการหาคำตอบที่

มีประสิทธิภาพแล้วจึงนำไปใช้พยากรณ์สิ่งที่จะเกิดขึ้นจากข้อมูลชุดใหม่ (Gunčar et al., 2018; Kotu & Deshpande, 2015; R. Sathya, 2013) เทคนิคที่นิยมใช้มีหลากหลายวิธีขึ้นอยู่กับชนิดของข้อมูล เช่น การพยากรณ์การวินิจฉัยโรคไข้เลือดออกหรือการพยากรณ์โรค Metabolic syndrome ด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) การพยากรณ์การเปลี่ยนแปลงของตลาดหลักทรัพย์ด้วยโครงข่ายประสาทเทียม (artificial neural networks) และการศึกษาการจำแนก SNPs และอัตราการตายของลูกไก่ด้วยการจำแนกแบบเบย์อย่างง่าย (naïve bayesain classifier) เป็นต้น(Karimi-Alavijeh, Jalili, & Sadeghi, 2016; Long, Gianola, Rosa, Weigel, & Avendaño, 2009; Moghaddam, Moghaddam, & Esfandyari, 2016; Tanner et al., 2008) แต่ยังไม่มีการศึกษาใดนำเทคนิคเหล่านี้ไปใช้ในการพยากรณ์ผลตรวจฮีโมโกลบินในผู้บริจาคโลหิต ซึ่งค่าฮีโมโกลบินต่ำกว่าเกณฑ์คือสาเหตุหลักในการถูกปฏิเสธให้บริจาคโลหิต

เทคนิคต้นไม้ตัดสินใจ (decision tree) และเทคนิคโครงข่ายประสาทเทียม (artificial neural networks) เป็นเทคนิคที่อาศัยหลักการจำแนกที่แตกต่างกัน โดยเทคนิคต้นไม้ตัดสินใจ (decision tree) เป็นเทคนิคที่ให้ผลรวดเร็วโดยอาศัยความน่าจะเป็นในการจำแนกกลุ่มจากการคำนวณค่าเกนและสามารถนำผลไปแปลเป็นกฎได้ง่ายให้ความแม่ยำสูง ส่วนเทคนิคโครงข่ายประสาทเทียม (artificial neural networks) เป็นเทคนิคที่อาศัยการสุ่มถ่วงค่านำหนักตัวแปรจากนั้นนำมาคำนวณหาความสัมพันธ์ด้วยสมการเช่น Sigmoid Function ดังนั้นจึงศึกษาเปรียบเทียบการนำเทคนิค ML มาใช้พยากรณ์ค่าฮีโมโกลบินของผู้บริจาคโลหิตรายเก่าได้ล่วงหน้าซึ่งจะเป็นประโยชน์ทั้งต่อหน่วยงานที่จัดหาโลหิตและตัวผู้บริจาคโลหิตสามารถบริจาคโลหิตได้ยาวนานขึ้น ผู้บริจาคโลหิตจะไม่เสียเวลา ค่าใช้จ่ายในการเดินทางมาบริจาคโลหิตแต่ไม่สามารถบริจาคได้ ลดความผิดหวังของผู้บริจาคโลหิต

2. วัตถุประสงค์

เพื่อเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มค่าฮีโมโกลบินในผู้บริจาคโลหิตระหว่างเทคนิคต้นไม้ตัดสินใจ (decision tree) และโครงข่ายประสาทเทียม (artificial neural networks)

3. อุปกรณ์และวิธีการ / วิธีดำเนินการวิจัย

งานวิจัยนี้ได้รับการรับรองจากคณะกรรมการพิจารณาจริยธรรมการวิจัยในมนุษย์ ศูนย์บริการโลหิตแห่งชาติ สภากาชาดไทย Certificate Number NBC 17/2018 เลขที่โครงการ 17/2561

3.1. การเก็บข้อมูลผู้บริจาคโลหิต

เก็บข้อมูลผู้บริจาคโลหิตของภาคบริการโลหิตแห่งชาติ 12 แห่งและสถานีกาชาดหัวหินเฉลิมพระเกียรติด้วยแบบสอบถามตั้งแต่ 1 ตุลาคม ถึง 31 ธันวาคม 2561 จำนวน 2,061 ราย จำนวน 40 ตัวแปร ประกอบด้วย 1.เพศ (Gender) 2.อายุ (Age) 3.น้ำหนัก (Weight) 4.สถานะภาพ (Status) 5.จำนวนบุตร (Child) 6.ศาสนา (Religion) 7.อาชีพ (Occupation) 8.รายได้ (Income) 9.ประจำเดือน (Menstruation) 10.ที่อยู่ (Address) 11.ระดับการศึกษา (Education) 12.น้ำหนักเมื่อ 3 เดือนที่ผ่านมา (Old Weight 3 Month) 13.น้ำหนักที่เปลี่ยนแปลงใน 3 เดือน (Change Weight 3 Month) 14.ส่วนสูง (High) 15.BMI 16.หมู่โลหิต (ABO) 17.ความดันโลหิต Systolic (BP Sys) 18.ความดันโลหิต Diastolic (BP Dias) 19.อัตราการเต้นของหัวใจ (Pulse) 20.สถานที่บริจาคโลหิต (Donation Place) 21.ค่าฮีโมโกลบินครั้งปัจจุบัน (Now Hb) 22.ค่าฮีโมโกลบินครั้งที่ผ่านมา (Last Hb) 23.ประวัติเคยตรวจฮีโมโกลบินไม่ผ่านเกณฑ์ (Ever

Low Hb) 24.จำนวนครั้งที่บริจาคโลหิตแบบ WB (Donation) 25.ความถี่การบริจาค WB ในรอบปี (Donation Per Year) 26.ระยะห่างการบริจาคครั้งที่ผ่านมาถึงปัจจุบัน (Period Donation) 27.ประวัติการบริจาคเกล็ดเลือดหรือน้ำเลือดแบบ Single Donor (Single Donor) 28.จำนวนครั้งที่เคยบริจาคเกล็ดเลือดหรือน้ำเลือดแบบ Single Donor (Amount Single Donor) 29.ขนาดของถุงที่บริจาค (Bag Type) 30.ประวัติโรคประจำตัว (Disease) 31.การรับประทานอาหาร (Food Type) 32.การสูบบุรี่ (Smoke) 33.พฤติกรรมการผักผ่อน (Sleep Type) 34.ชั่วโมงการนอน (Sleep Hour) 35. การออกกำลังกาย (Exercise) 36.การดื่มแอลกอฮอล์ (Alcohol Take) 37.การรับประทานธาตุเหล็ก (Fe Take) 38.ช่องทางการได้รับข่าวสารการบริจาคโลหิต (Public to known) 39.สาเหตุการไม่ทานธาตุเหล็ก (Why not take) 40.น้ำหนักที่เปลี่ยนแปลงในรอบ 1 ปี (Change Weight a year)

3.2 วิธีการ

3.2.1 การทำความสะอาดข้อมูลและการแปลงค่า (Data Cleaning and Preprocessing)

เก็บรวบรวมข้อมูลด้วยโปรแกรม Microsoft excel ตรวจสอบความถูกต้องครบถ้วนข้อมูลตัวแปรจำนวน 40 ตัวแปรของผู้บริจาคโลหิตแต่ละราย จากนั้นทำการแปลรหัสข้อมูลให้เหมาะสมต่อการนำเข้าตัวแบบพยากรณ์และกำหนดกลุ่มตัวแปรตาม (Class) ในการจำแนกด้วยตัวแปร ฮีโมโกลบินครั้งปัจจุบัน (Now Hb) โดยทำการแปลค่าเป็น 2 กลุ่มคือกลุ่มที่ผ่านเกณฑ์ (ผู้ชาย $\geq$ 13.0 mg/dl, ผู้หญิง $\geq$ 12.5 mg/dl) และกลุ่มที่ไม่ผ่านเกณฑ์ (ผู้ชาย < 13.0 mg/dl, ผู้หญิง < 12.5 mg/dl) และชดเชยค่าที่ว่างด้วยค่าเฉลี่ยของแต่ละตัวแปรด้วยโปรแกรม Rapidminer เวอร์ชั่น 9.2(Malik & Mishra, 2014; Wahyuni, Saputra S, & Iswan, 2017)

3.2.2 การคัดเลือกตัวแปรพยากรณ์

เป็นการเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Regression) เป็นวิธีที่มีความเหมาะสมในการพิจารณาคัดเลือกตัวแปรให้ได้ตัวแบบพยากรณ์ที่ประหยัดที่สุด จะทำการทดสอบตัวแปรพยากรณ์ที่เข้าสมการไปแล้วทุกครั้งที่มีการนำตัวแปรใหม่เข้าในสมการ หากพบว่าตัวแปรพยากรณ์ใดไม่ได้ส่งผลให้ค่า R^2 เพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติ(Chong & Jun, 2005; Derksen & Keselman, 1992; Haque, Rahman, Hagare, & Chowdhury, 2018; Zhang, 2016) จะขจัดตัวแปรนั้นออกจากสมการ การศึกษานี้คัดเลือกตัวแปรด้วยโปรแกรม IBM SPSS

3.2.3 การพัฒนาต้นแบบพยากรณ์ด้วยโปรแกรม Rapidminer เวอร์ชั่น 9.2

3.2.3.1 การพัฒนาตัวแบบพยากรณ์ ต้นไม้ตัดสินใจ (decision tree)

เป็นวิธีการจำแนกกลุ่มที่ง่ายด้วยการรวบรวมโหนดตัดสินใจ (decision node) เชื่อมต่อด้วยกิ่งก้านคล้ายต้นไม้ โหนดบนสุดเรียกว่าโหนดราก (root node) ส่วนโหนดที่ต่อเป็นกิ่งก้านเรียกว่าโหนดใบ (leaf node) แต่ละโหนดจะทำหน้าที่ทดสอบคุณลักษณะ (attribute) และตัดสินใจเลือกกลุ่มในแต่ละกิ่งก้านจะนำไปสู่โหนดตัดสินใจอีกโหนด โดยใช้ค่าเกน (gain) สูงที่สุดเป็นโหนดราก (root node) ค่าเกนพิจารณาจากความน่าจะเป็นของคุณสมบัติตัวแปรตามต่อคุณสมบัติของตัวแปรพยากรณ์(Gunčar et al., 2018; Kotu & Deshpande, 2015; Song & Lu, 2015; Vijay Kotu, 2015)

3.2.3.2 การพัฒนาตัวแบบพยากรณ์ โครงข่ายประสาทเทียม (artificial neural networks)

เทคนิคโครงข่ายประสาทเทียม (artificial neural networks) จุดเด่นคือสามารถที่จะเรียนรู้สิ่งต่างๆได้ มีรูปแบบการทำงานเรียนแบบสมองของมนุษย์(Walczak, 2005) เทคนิคโครงข่ายประสาทเทียมชนิด neural network ประกอบด้วย ชั้นรับข้อมูล (input Layer) ชั้นซ่อนเร้น (hidden Layer) และชั้นแสดงผล (output layer) กำหนดจำนวน

โหนดในชั้นรับข้อมูล (input node) ให้มีจำนวนเท่ากับจำนวนตัวแปรที่ผ่านการคัดเลือกแล้วว่ามีความสัมพันธ์ต่อประสิทธิภาพพยากรณ์ และกำหนดจำนวนโหนดในชั้นแสดงผล (output node) เท่ากับ 1 โหนด กำหนดจำนวนรอบในการเรียนรู้ (Epoch) เช่น 1,000 รอบ ค่าผิดพลาดที่ยอมรับได้เช่น 0.0001

3.2.4 การประเมินประสิทธิภาพตัวแบบพยากรณ์

ประเมินความถูกต้องด้วยการสุ่มแบ่งข้อมูลออกเป็นร้อยละ 80 สำหรับฝึกตัวแบบพยากรณ์และร้อยละ 20 ใช้ในการประเมินผลตัวแบบพยากรณ์และเปรียบเทียบประสิทธิภาพของตัวแบบพยากรณ์ด้วยค่าความถูกต้อง (accuracy) คือค่าร้อยละความถูกต้องในการจำแนกของตัวแบบทั้งในกลุ่มผู้ที่มีค่าฮีโมโกลบินผ่านและไม่ผ่านเกณฑ์ ค่าความไว (sensitivity) คือค่าร้อยละความถูกต้องในจำแนกของตัวแบบเฉพาะในกลุ่มผู้ที่มีค่าฮีโมโกลบินไม่ผ่านเกณฑ์ ค่าความจำเพาะ (specificity) คือค่าร้อยละความถูกต้องในการจำแนกของตัวแบบเฉพาะในกลุ่มที่มีค่าฮีโมโกลบินผ่านเกณฑ์ ค่าการทำนายผลบวก (positive predictive value;PPV) คือค่าร้อยละของความน่าจะเป็นว่าจะมีค่าฮีโมโกลบินไม่ผ่านเกณฑ์เมื่อถูกพยากรณ์จำแนกว่าไม่ผ่านเกณฑ์ ค่าการทำนายผลลบ (negative predictive value:NPV) คือค่าร้อยละของความน่าจะเป็นว่าจะมีค่าฮีโมโกลบินผ่านเกณฑ์เมื่อถูกพยากรณ์จำแนกว่าผ่านเกณฑ์ และค่า AUC ค่าคือพื้นที่ใต้เส้น Receiver Operating Characteristic curve (ROC curve) แสดงถึงความถูกต้องของการพยากรณ์ถ้ามีค่าใกล้ 1 แสดงว่าตัวแบบมีประสิทธิภาพ โดยทั้งหมดสามารถคำนวณได้จากตาราง confusion matrix

4. ผลการวิจัย

พบข้อมูลจำนวน 1,908 รายจากจำนวน 2,061 รายที่มีข้อมูลตัวแปรต่างๆถูกต้องและตัวแปรพยากรณ์จำนวน 40 ตัวแปรพบว่ามี 2 ตัวแปรที่มีข้อมูลน้อยกว่าร้อยละ 70 คือตัวแปรพฤติกรรมการพักผ่อนและน้ำหนักที่เปลี่ยนแปลงในรอบ 1 ปีคงเหลือตัวแปร 38 ตัวแปรและทำการเติมค่าว่างด้วยค่าเฉลี่ยของแต่ละตัวแปร นอกจากนี้ยังพบจำนวน 820 รายที่ไม่มีค่าฮีโมโกลบินครั้งที่ผ่านมาเนื่องจากเป็นการตรวจด้วยสารคอปเปอร์ซัลเฟต แต่ให้ประวัติในแบบสอบถามว่าไม่เคยถูกปฏิเสธสาเหตุจากค่าฮีโมโกลบินไม่ผ่านเกินจึงปรับค่าฮีโมโกลบินครั้งที่ผ่านมาให้มีค่าเท่ากับครั้งปัจจุบัน

เมื่อนำข้อมูล 38 ตัวแปรของจำนวนผู้บริจาคโลหิต 1,908 พบ 24 ตัวแปรที่มีความแตกต่างกันระหว่างกลุ่มอย่างมีนัยสำคัญทางสถิติที่ p-value น้อยกว่า 0.05 รายระหว่างกลุ่มที่มีผลการตรวจฮีโมโกลบินผ่านและไม่ผ่านเกณฑ์ จากนั้นคัดเลือกตัวแปรด้วยเทคนิค Stepwise Regression พบว่าคงเหลือเพียง 9 ตัวแปรที่ให้ค่าการพยากรณ์ที่ดีประกอบด้วย ประวัติเคยมีค่าฮีโมโกลบินต่ำกว่าเกณฑ์ (EverLowHb) ค่าฮีโมโกบิลครั้งที่ผ่านมา (LastHb) ความถี่ในการบริจาคในรอบปี (DonationPerYear) ระยะเวลานอนพักผ่อน (Sleep Hour) การมีประจำเดือน (Menstruation) ระยะห่างการบริจาคครั้งที่ผ่านมาถึงปัจจุบัน (Period Donation) ดัชนีมวลกาย (BMI) และหมู่โลหิต (ABO) (ตารางที่ 1)

ตารางที่ 1 Selected factors by stepwise regression

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	EverLowHb	.813	1	1	1906.000	439.755	1	1906.000	0.000

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
2	LastHb	.774	2	1	1906.000	278.876	2	1905.000	0.000
3	DonationPerYear	.767	3	1	1906.000	192.859	3	1904.000	0.000
4	Sleep hour	.761	4	1	1906.000	149.216	4	1903.000	0.000
5	Menstruation	.756	5	1	1906.000	122.511	5	1902.000	0.000
6	PeriodDonation	.754	6	1	1906.000	103.107	6	1901.000	0.000
7	DonationPlace	.753	7	1	1906.000	89.208	7	1900.000	0.000
8	BMI	.751	8	1	1906.000	78.734	8	1899.000	0.000
9	ABO	.749	9	1	1906.000	70.537	9	1898.000	0.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.[a,b,c,d]

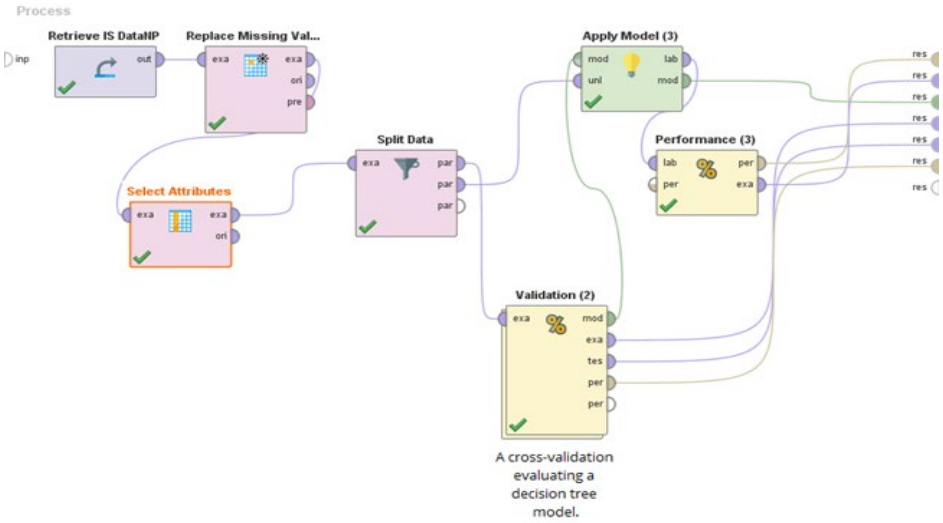
a. Maximum number of steps is 76.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

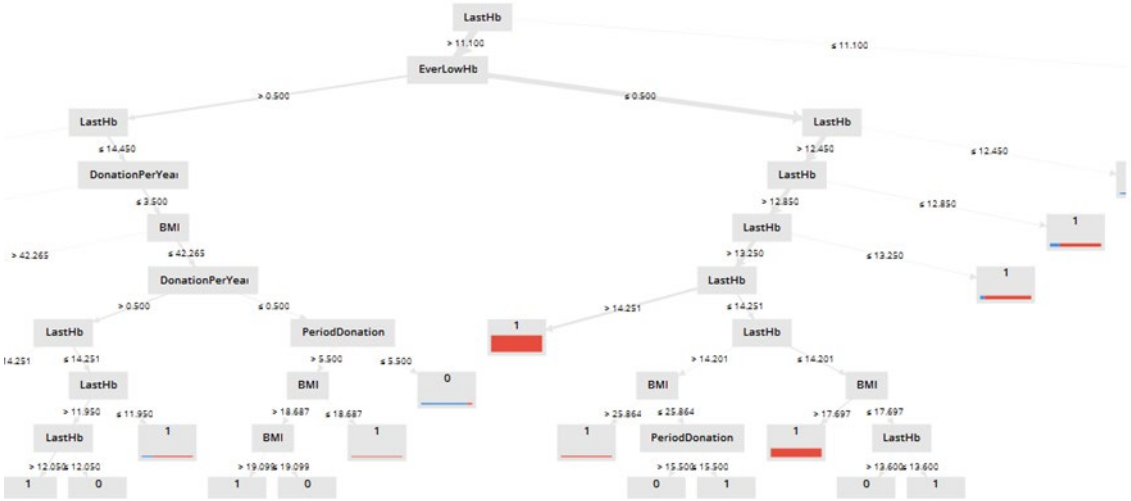
d. F level, tolerance, or VIN insufficient for further computation.

ผลการพัฒนาตัวแบบพยากรณ์ด้วยโปรแกรม Rapidminer เวอร์ชั่น 9.2 มีด้วยกัน 7 ขั้นตอน ประกอบด้วย 1. การนำข้อมูลเข้า 2.การตรวจสอบค่าว่างและการเติมค่า 3.การเลือกตัวแปร 4.การแบ่งข้อมูลออกเป็นร้อยละ 80 จำนวน 1,526 รายสำหรับ Training Dataset และร้อยละ 20 จำนวน 382 รายสำหรับ Testing Dataset 5.การ Validation ตัวแบบพยากรณ์ต้นไม้ตัดสินใจ (decision tree) และโครงข่ายประสาทเทียม (artificial neural networks) 6.การสร้างตัวแบบพยากรณ์และ7.การทดสอบประสิทธิภาพตัวแบบพยากรณ์ (รูปที่ 1)



รูปที่ 1 Rapidminer process to validation and performance models

โดยตัวแบบพยากรณ์ต้นไม้ตัดสินใจ (decision tree) เมื่อทำการฝึกหัดสามารถลดตัวแปรลงเหลือเพียง 5 ตัวแปรได้เป็นต้นไม้ตัดสินใจ 21 โหนด โดยมีค่าฮีโมโกลบินที่ครั้งที่ผ่านมา (LastHb) เป็นโหนดรากและ 20 โหนดกิ่งประกอบด้วยค่าฮีโมโกลบินที่ครั้งที่ผ่านมา (LastHb) 10 โหนด ประวัติเคยตรวจฮีโมโกลบินไม่ผ่านเกณฑ์ (EverLowHb) 1 โหนด ความถี่การบริจาคในรอบปี (DonationPerYear) 2 โหนด ดัชนีมวลกาย (BMI) 5 โหนด และระยะห่างการบริจาคครั้งที่ผ่านมาถึงปัจจุบัน (Period Donation) 2 โหนด (รูปที่ 2)



รูปที่ 2 Outcome of decision tree model as classified passed and not passed Hb level

จากรูปที่ 2 สามารถอธิบายเป็นกฎการตัดสินใจพยากรณ์ผลว่ามีค่าฮีโมโกลบินไม่ผ่านเกณฑ์ได้ 7 ข้อดังนี้ กฎข้อ 1 ถ้าค่าอีโมโกลบิลครั้งที่ผ่านมา≤11.10 กฎข้อที่ 2 ถ้าค่าอีโมโกลบิลครั้งที่ผ่านมา≤12.05 และไม่เคยมีประวัติตรวจไม่ผ่านเกณฑ์และบริจาคโลหิต≤3.5 ครั้งต่อปีและมีค่า BMI ≤42.265 กฎข้อที่ 3 ถ้าค่าอีโมโกลบิลครั้งที่ผ่านมา ≤14.450 และไม่เคยมีประวัติตรวจไม่ผ่านเกณฑ์และบริจาคโลหิต≤3.5 ครั้งต่อปีและมีค่า BMI ≤42.265 และระยะห่างการบริจาคครั้งที่ผ่านมาถึงปัจจุบัน ≤5.5 เดือน กฎข้อที่ 4 ถ้าค่าอีโมโกลบิลครั้งที่ผ่านมา≤14.450 และไม่เคยมีประวัติตรวจไม่ผ่านเกณฑ์และบริจาคโลหิต≤3.5 ครั้งต่อปีและมีค่า BMI≤19.099 และระยะห่างการบริจาคครั้งที่ผ่านมาถึงปัจจุบัน > 5.5 เดือน กฎข้อที่ 5 เคยมีประวัติตรวจไม่ผ่านเกณฑ์และค่าอีโมโกลบิลครั้งที่ผ่านมา≤12.450 กฎข้อที่ 6 เคยมีประวัติตรวจไม่ผ่านเกณฑ์และค่าอีโมโกลบิลครั้งที่ผ่านมา< 14.201 แต่ไม่เกิน 14.251 และมีค่า BMI≤25.864 และระยะห่างการบริจาคครั้งที่ผ่านมาถึงปัจจุบัน >15.5 เดือนและกฎข้อที่ 7 เคยมีประวัติตรวจไม่ผ่านเกณฑ์และค่าอีโมโกลบิลครั้งที่ผ่านมา< 13.60 แต่ไม่เกิน 14.201 และมีค่า BMI≤17.697

ทำการ Validation ตัวแบบพยากรณ์ด้วยข้อมูล Training dataset พบว่าให้ค่าความถูกต้อง (accuracy) เท่ากับร้อยละ 85.71 ค่าความไว (sensitivity) ร้อยละ 23.53 ค่าความจำเพาะ (specificity) ร้อยละ 98.19 ค่าการทำนายผลบวก (positive predictive value) ร้อยละ 72.29 และค่าการทำนายผลลบ (negative predictive value) ร้อยละ 86.49 (ตารางที่ 2) และประเมินประสิทธิภาพตัวแบบพยากรณ์ด้วยข้อมูล Testing dataset พบว่าให้ค่าความถูกต้อง (accuracy) เท่ากับร้อยละ 86.13 ค่าความไว (sensitivity) ร้อยละ 23.44 ค่าความจำเพาะ (specificity) ร้อยละ 98.74 ค่าการทำนายผลบวก

(positive predictive value) ร้อยละ 78.95 และค่าการทำนายผลลบ (negative predictive value) ร้อยละ 86.50 (ตารางที่ 3) โดยมีค่า AUC เท่ากับ 0.844 (รูปที่ 3)

ตารางที่ 2 Confusion matrix of training dataset using decision tree model

	Hb ไม่ผ่านเกณฑ์	**Hb ผ่านเกณฑ์**	**Class precision****
พยากรณ์ Hb ไม่ผ่านเกณฑ์	60	23	72.29
พยากรณ์ Hb ผ่านเกณฑ์	195	1248	86.49
Class recall*	23.53	98.19	

*Class recall ประกอบด้วยค่า Sensitivity=23.53 และ Specificity=98.19

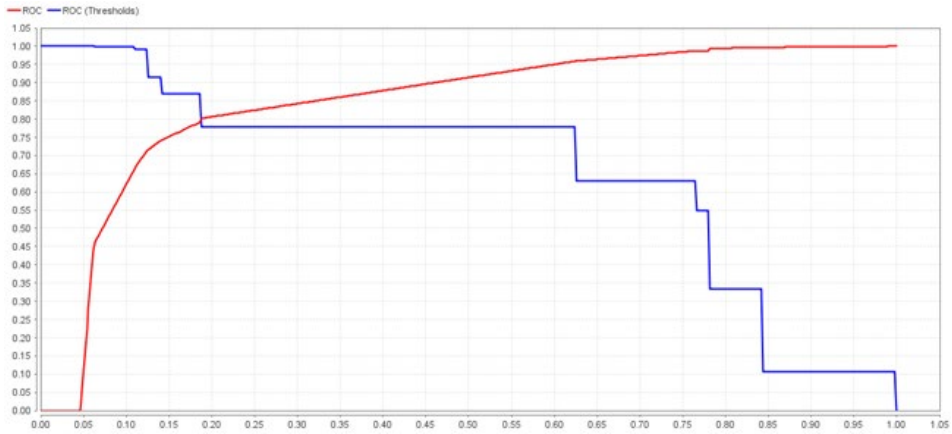
**Class precision ประกอบด้วยค่า PPV=72.29 และค่า NPV=86.45

ตารางที่ 3 Confusion matrix of testing dataset using decision tree model

	Hb ไม่ผ่านเกณฑ์	**Hb ผ่านเกณฑ์**	**Class precision****
พยากรณ์ Hb ไม่ผ่านเกณฑ์	15	4	78.95
พยากรณ์ Hb ผ่านเกณฑ์	49	314	86.50
Class recall*	23.44	98.74	

*Class recall ประกอบด้วยค่า Sensitivity=23.44 และ Specificity=98.74

**Class precision ประกอบด้วยค่า PPV=78.95 และค่า NPV=86.50



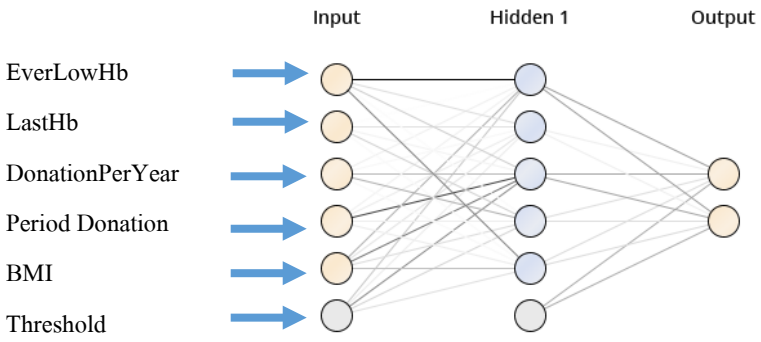
รูปที่ 3 Area under the curve (AUC) of testing dataset using decision tree model

ตัวแบบพยากรณ์โครงข่ายประสาทเทียม (artificial neural networks) มี 3 layer เมื่อผ่านการฝึกหัดคงเหลือตัวแปรทั้งหมด 5 ตัวแปรใน Input layer และ Hidden layer ประกอบด้วย 6 โหนดคือค่าฮีโมโกลบิลครั้งที่ผ่านมา (LastHb) ประวัติเคยตรวจฮีโมโกลบินไม่ผ่านเกณฑ์ (EverLowHb) ความถี่การบริจาคในรอบปี (DonationPerYear) ระยะห่างการบริจาคครั้งที่ผ่านมาถึงปัจจุบัน (Period Donation) ดัชนีมวลกาย (BMI) และ Threshold (Bias) โหนด Output layer ประกอบด้วย 2 โหนดคือฮีโมโกลบินผ่านเกณฑ์และไม่ผ่านเกณฑ์ (รูปที่ 4) โดยแต่ละโหนดในชั้น Hidden layer ให้ค่าน้ำหนักของตัวแปรจากชั้น Input layer ดังนี้ โหนดที่ 1 ค่า LastHb เท่ากับ 20.395 ค่า EverLowHb เท่ากับ -6.379 ค่า DonationPerYear เท่ากับ -1.168 ค่า Period Donation เท่ากับ -1.099 ค่า BMI เท่ากับ -0.191 และค่า Bias เท่ากับ -4.461

โหนดที่ 2 ค่า LastHb เท่ากับ 2.661 ค่า EverLowHb เท่ากับ -2.192 ค่า DonationPerYear เท่ากับ 0.523 ค่า Period Donation เท่ากับ 1.503 ค่า BMI เท่ากับ 2.438 และค่า Bias เท่ากับ -1.504 โหนดที่ 3 ค่า LastHb เท่ากับ 3.910 ค่า EverLowHb เท่ากับ 10.972 ค่า DonationPerYear เท่ากับ 5.607 ค่า Period Donation เท่ากับ 15.295 ค่า BMI เท่ากับ 0.929 และค่า Bias เท่ากับ 9.376 โหนดที่ 4 ค่า LastHb เท่ากับ 0.346 ค่า EverLowHb เท่ากับ 2.834 ค่า DonationPerYear เท่ากับ 6.270 ค่า Period Donation เท่ากับ -1.330 ค่า BMI เท่ากับ 3.559 และค่า Bias เท่ากับ -3.327 โหนดที่ 5 ค่า LastHb เท่ากับ 10.065 ค่า EverLowHb เท่ากับ -4.762 ค่า DonationPerYear เท่ากับ -0.511 ค่า Period Donation เท่ากับ 1.464 ค่า BMI เท่ากับ 0.438 และค่า Bias เท่ากับ -2.511 และแต่ละโหนดในชั้น Out layer ให้ค่าน้ำหนักของตัวแปรจากชั้น Hidden layer ดังนี้ กลุ่มไม่ผ่านเกณฑ์ โหนดที่ 1 2 3 4 5 และ 6 เท่ากับ -7.794 -1.290 -7.213 -2.615 -2.904 และ 6.670 ตามลำดับ และกลุ่มผ่านเกณฑ์ โหนดที่ 1 2 3 4 5 และ 6 เท่ากับ 7.791 1.258 7.213 2.616 2.915 และ -6.669 ตามลำดับ

ทำการ Validation ตัวแบบพยากรณ์ด้วยข้อมูล Training dataset พบว่าให้ค่าความถูกต้อง (accuracy) เท่ากับร้อยละ 85.78 ค่าความไว (sensitivity) ร้อยละ 34.51 ค่าความจำเพาะ (specificity) ร้อยละ 96.07 ค่าการทำนายผลบวก (positive predictive value) ร้อยละ 63.77 และค่าการทำนายผลลบ (negative predictive value) ร้อยละ 87.97 (ตารางที่ 4)

ประเมินประสิทธิภาพตัวแบบด้วยข้อมูล Testing dataset พบว่าให้ค่าความถูกต้อง (accuracy) เท่ากับร้อยละ 85.60 ค่าความไว (sensitivity) ร้อยละ 37.50 ค่าความจำเพาะ (specificity) ร้อยละ 95.28 ค่าการทำนายผลบวก (positive predictive value) ร้อยละ 61.54 และค่าการทำนายผลลบ (negative predictive value) ร้อยละ 88.34 (ตารางที่ 5) โดยมีค่า AUC เท่ากับ 0.865 (รูปที่ 5)



รูปที่ 4 Outcome of artificial neural networks model as classified passed and not passed Hb level

ตารางที่ 4 Confusion matrix of training dataset using artificial neural networks model

	Hb ไม่ผ่านเกณฑ์	**Hb ผ่านเกณฑ์**	**Class precision****
พยากรณ์ Hb ไม่ผ่านเกณฑ์	88	50	63.77
พยากรณ์ Hb ผ่านเกณฑ์	167	1221	87.97
Class recall*	34.51	96.07	

*Class recall ประกอบด้วยค่า Sensitivity=34.51 และ Specificity=96.07

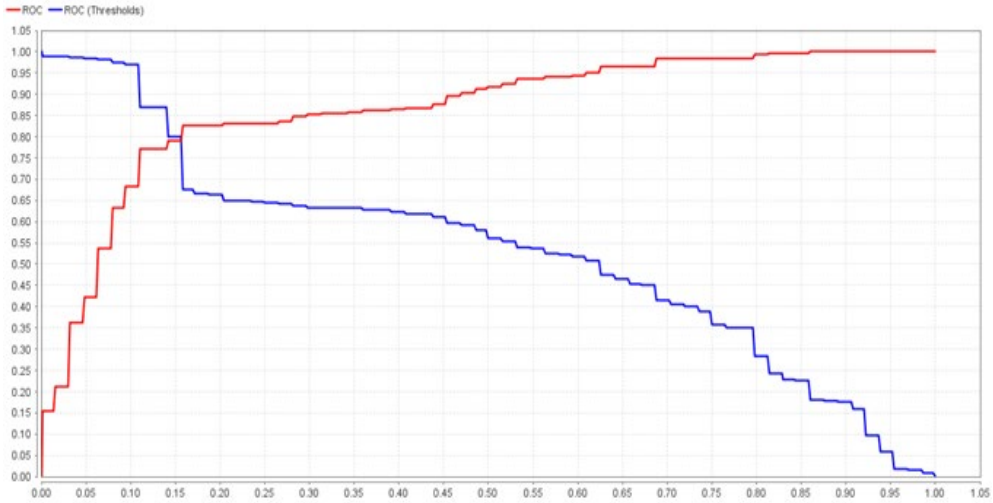
**Class precision ประกอบด้วยค่า PPV=63.77 และค่า NPV=57.97

ตารางที่ 5 Confusion matrix of testing dataset using artificial neural networks model

	Hb ไม่ผ่านเกณฑ์	Hb ผ่านเกณฑ์	Class precision**
พยากรณ์ Hb ไม่ผ่านเกณฑ์	24	15	61.54
พยากรณ์ Hb ผ่านเกณฑ์	40	303	88.34
Class recall*	37.50	95.28	

*Class recall ประกอบด้วยค่า Sensitivity=37.50 และ Specificity=95.28

**Class precision ประกอบด้วยค่า PPV=61.54 และค่า NPV=88.34



รูปที่ 5 Area under the curve (AUC) of testing dataset using artificial neural networks model

5. การอภิปรายผล

ในพัฒนาตัวแบบพยากรณ์ได้คัดเลือกตัวแปรด้วย Stepwise regression คงเหลือ 9 ตัวแปรจาก 24 ตัวแปรแต่จากการฝึกหัดพบว่า 4 ตัวแปรสามารถนำออกได้อีกโดยยังคงให้ผลการฝึกหัดที่ดี ทำให้ตัวแบบพยากรณ์มีความประหยัด ประกอบด้วยตัวแปรระยะเวลาพักผ่อน (Sleep Hour) การมีประจำเดือน (Menstruation) สถานที่บริจาค (Donation Place) และหมู่โลหิต (ABO) พบว่าตัวแบบพยากรณ์ต้นไม้ตัดสินใจ (decision tree) และตัวแบบพยากรณ์โครงข่ายประสาทเทียม (artificial neural networks) ให้ค่า Accuracy Sensitivity Specificity PPV NPV และ AUC เท่ากับ 86.13/85.60, 23.44/37.50, 98.74/95.28, 78.95/61.54, 86.50/88.34 และ 0.844/0.865 ตามลำดับ พบว่าตัวแบบทั้งสองมีค่าความไว (sensitivity) ต่ำและค่าความจำเพาะ (specificity) สูงเช่นเดียวกันหมายความว่าผู้ที่มีค่าฮีโมโกลบินต่ำกว่าเกณฑ์จะถูกพยากรณ์จำแนกว่าไม่ผ่านเกณฑ์ได้ในอัตราที่ต่ำแต่ผู้ที่มีค่าฮีโมโกลบินผ่านเกณฑ์มีโอกาสจะถูกพยากรณ์ว่าผ่านเกณฑ์ในอัตราที่สูง ทั้งนี้อาจเนื่องมาจากระดับค่าฮีโมโกลบินในกลุ่มที่มีระดับค่าใกล้เคียงเกณฑ์อาจมีการเปลี่ยนแปลงหรือคลาดเคลื่อนจากปัจจัยอื่นๆในแต่ละรอบของการบริจาคโลหิต ซึ่งไม่ได้เก็บข้อมูลตัวแปรเหล่านั้นในการการศึกษานี้เช่น ชนิดของอาหารที่รับประทาน ปริมาณธาตุเหล็กสะสมของแต่ละคน ความถูกต้อง ความแม่นยำของวิธีการตรวจฮีโมโกลบินในแต่ละครั้ง เป็นต้น ซึ่งในความคาดหวังของเครื่องมือทดสอบที่พัฒนาขึ้นควรจะมีค่าความไว (sensitivity) และค่าความจำเพาะ (specificity) ที่สูง นั้นแสดงว่าประสิทธิภาพของตัวแบบทั้งสองในการ

จำแนกผู้ที่มีค่าฮีโมโกลบินไม่ผ่านเกณฑ์ของการศึกษานี้ยังไม่ดีพอแม้ว่าค่าความถูกต้อง (accuracy) ค่าการพยากรณ์ผลลบ (negative predictive value) และค่า AUC จะอยู่ในช่วงที่ยอมรับได้ก็ตาม เนื่องจากข้อมูลส่วนใหญ่เป็นผู้บริจาคโลหิตที่มีผลตรวจฮีโมโกลบินผ่านเกณฑ์จำนวน 1,589 รายจากจำนวน 1,908 ราย ทำให้ผลกระทบของความผิดพลาดในการจำแนกกลุ่มที่มีผลไม่ผ่านเกณฑ์ต่อตัวชี้วัดทั้ง 2 มีน้อยมากต่างจากค่าความไว (sensitivity) แต่การศึกษานี้มุ่งเน้นประสิทธิภาพผลพยากรณ์การจำแนกกลุ่มผู้ที่มีค่าฮีโมโกลบินไม่ผ่านเกณฑ์ พบว่าตัวแบบพยากรณ์ต้นไม้ตัดสินใจ (decision tree) ได้ค่าการพยากรณ์ผลบวก (positive predictive value) มากกว่าตัวแบบพยากรณ์โครงข่ายประสาทเทียม (artificial neural networks) ที่ร้อยละ 78.95 และร้อยละ 61.54 ตามลำดับ จึงอาจสรุปได้ว่าตัวแบบพยากรณ์ต้นไม้ตัดสินใจ (decision tree) มีความเหมาะสมในการนำไปใช้พยากรณ์ผลตรวจฮีโมโกลบินมากกว่าตัวแบบพยากรณ์โครงข่ายประสาทเทียม (artificial neural networks)

เมื่อเปรียบเทียบการพยากรณ์ค่าฮีโมโกลบินของการศึกษานี้กับการศึกษาอื่นเกี่ยวข้องกับการพยากรณ์ค่าฮีโมโกลบินในผู้บริจาคโลหิต ได้แก่ Kazem Nasserinejad และคณะ (ค.ศ. 2013) ซึ่งศึกษาการพยากรณ์ด้วยเทคนิค Multiple linear regression model และ Transition model พบกว่าค่า AUC ในเพศชายอยู่ในช่วงที่ดีคือ 0.83 และ 0.81 ในขนาดที่ค่า AUC ในเพศหญิงอยู่ในช่วงที่ยอมรับได้คือ 0.73 และ 0.72 ตามลำดับ(Nasserinejad et al., 2013) ซึ่งใกล้เคียงกับการศึกษานี้ และการศึกษาของ Jesse Fokkinga (ค.ศ. 2018) ที่ใช้เทคนิคต้นไม้ตัดสินใจ (decision tree) ชนิด Random forest ได้ค่า AUC อยู่ในช่วงที่ยอมรับได้คือ 0.717 ในเพศหญิงและ 0.690 ในเพศชายซึ่งน้อยกว่าการศึกษานี้ (Fokkinga & Paap, 2019)

ในการศึกษาครั้งนี้ยังพบว่าประวัติเคยตรวจฮีโมโกลบินไม่ผ่านเกณฑ์ (EverLowHb) และข้อมูลผลการตรวจฮีโมโกลบินในครั้งที่ผ่านมา (LastHb) เป็นตัวแปรที่มีความสำคัญโดยพิจารณาจากค่า F ทางสถิติที่มีค่า 439.733 และ 210.201 แต่พบว่าผู้บริจาคโลหิตจำนวนมากให้ข้อมูลไม่ครบถ้วนถูกต้องหรือระบุเพียงว่าผ่านหรือไม่ผ่าน เนื่องมาจากในอดีตการตรวจคัดกรองฮีโมโกลบินใช้วิธีวัดความถ่วงจำเพาะด้วยสารละลายคอปเปอร์ซัลเฟต ทำให้ต้องใช้ค่าปัจจุบันแทนซึ่งอาจส่งผลให้ตัวแบบพยากรณ์มีความผิดพลาดในการพยากรณ์และในการศึกษานี้ไม่ได้ศึกษาแยกตัวแบบพยากรณ์ระหว่างเพศชายและหญิงที่มีค่าเกณฑ์และปัจจัยต่างๆที่แตกต่างกันก็อาจส่งผลต่อความคลาดเคลื่อนด้วยเช่นกันเมื่อนำไปใช้งานจริง ในอนาคตควรทำการศึกษาเพิ่มเติมโดยใช้กลุ่มประชากรที่มากขึ้นและเก็บข้อมูลเฉพาะผู้ที่มีข้อมูลค่าผลตรวจฮีโมโกลบินเท่านั้น อีกทั้งยังอาจศึกษาโดยใช้เทคนิคอื่นๆ ทั้งการคัดเลือกตัวแปรและตัวแบบพยากรณ์ ซึ่งอาจทำให้ค่าความถูกต้องแม่นยำของการพยากรณ์ที่ดียิ่งขึ้น

6. บทสรุป

ตัวแบบพยากรณ์ต้นไม้ตัดสินใจ (decision tree) เหมาะสมในการนำไปใช้พยากรณ์มากกว่าตัวแบบโครงข่ายประสาทเทียม (artificial neural networks) เนื่องจากให้ค่าความถูกต้อ (accuracy) ค่าความจำเพาะ (specificity) และค่าพยากรณ์ผลบวก (positive predictive value) สูงกว่าหมายคามว่าถ้าพยากรณ์จำแนกว่าไม่ผ่านเกณฑ์มีโอกาสที่จะเป็นจริงมากกว่าและในกลุ่มผู้ที่ผ่านเกณฑ์จะพยากรณ์ว่าผ่านได้มากกว่าด้วยเช่นกัน ถึงแม้ว่าค่าความไว (sensitivity) ของตัวแบบโครงข่ายประสาทเทียม (artificial neural networks) จะมากกว่าก็ตาม เพราะถึงแม้ว่าจะพยากรณ์จำนวนผู้ไม่ผ่านเกณฑ์ได้มากกว่าแต่โอกาสจะเป็นผู้ที่มีค่าฮีโมโกลบินต่ำจริงเพียงร้อยละ 61.54 เท่านั้น

7. กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณหัวหน้าภาคบริการโลหิตแห่งชาติ 12 แห่งและหัวหน้างานบริการโลหิตสถานีกาชาดหัวหิน เฉลิมพระเกียรติ ตลอดจนเจ้าหน้าที่ทุกท่านที่ได้ให้ความอนุเคราะห์ช่วยเหลือเก็บข้อมูลผู้บริจาคโลหิต

8. เอกสารอ้างอิง

วิชุดา กลิ่นหอม วรรณวิมล มีคง และ สมรัก เพชรโฉมฉาย. (2015). การศึกษาผู้บริจาคโลหิตที่ไม่ผ่านการคัดกรองสุขภาพก่อนบริจาคโลหิตเพื่อมุ่งเน้นการเพิ่มจำ นวนผู้บริจาคโลหิตของภาคบริการโลหิตแห่งชาติที่ 4 จังหวัดราชบุรี. *Thai J Hematol Transf Med,* 25(3), 1.

วิไลภรณ์ วงษ์กิติโสภณ ศิริรักษ์ ศุภธีระธาดา และ กัลยลักษณ์ คลับคล้าย. (2015). การศึกษาอัตราผู้บริจาคโลหิตที่ไม่ผ่านเกณฑ์การคัดเลือกผู้บริจาคโลหิต : ศูนย์บริการโลหิตแห่งชาติ สภากาชาดไทย. *Thai J Hematol Transf Med,* 25(3), 1.

Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems,* 78(1), 103-112. doi:https://doi.org/10.1016/j.chemolab.2004.12.011

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology,* 45(2), 265-282. doi:10.1111/j.2044-8317.1992.tb00992.x

Fokkinga, J. J., & Paap, R. (2019). *Modelling hemoglobin levels of blood donors.* Retrieved from http://hdl.handle.net/2105/45039

Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., & Notar, M. (2018). An application of machine learning to haematological diagnosis. *Scientific Reports,* 8(1), 411. doi:10.1038/s41598-017-18564-8

Haque, M. M., Rahman, A., Hagare, D., & Chowdhury, K. R. (2018). A Comparative Assessment of Variable Selection Methods in Urban Water Demand Forecasting. *Water,* 10(4). doi:10.3390/w10040419

Karimi-Alavijeh, F., Jalili, S., & Sadeghi, M. (2016). Predicting metabolic syndrome using decision tree and support vector machine methods. *ARYA atherosclerosis,* 12(3), 146-152.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science,* 57, 500-508. doi:https://doi.org/10.1016/j.procs.2015.07.372

Kotu, V., & Deshpande, B. (2015). Chapter 1 - Introduction. In V. Kotu & B. Deshpande (Eds.), *Predictive Analytics and Data Mining* (pp. 1-16). Boston: Morgan Kaufmann.

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. A., & Avendaño, S. (2009). Comparison of classification methods for detecting associations between SNPs and chick mortality. *Genetics, selection, evolution : GSE,* 41(1), 18-18. doi:10.1186/1297-9686-41-18

Malik, H., & Mishra, S. (2014, 11-13 Dec. 2014). *Feature selection using RapidMiner and classification through probabilistic neural network for fault diagnostics of power transformer.* Paper presented at the 2014 Annual IEEE India Conference (INDICON).

Moghaddam, A. H., Moghaddam, M. H., & Esfandyari, M. (2016). Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science,* 21(41), 89-93. doi:https://doi.org/10.1016/j.jefas.2016.07.002

Nasserinejad, K., de Kort, W., Baart, M., T Komárek, A., van Rosmalen, J., & Lesaffre, E. (2013). *Predicting hemoglobin levels in whole blood donors using transition models and mixed effects models* (Vol. 13).

Organization, W. H. (2017). The 2016 global status report on blood safety and availability, 166. Retrieved from http://www.who.int/iris/handle/10665/254987

R. Sathya, A. A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence,* 2(2), 5.

Song, Y.-y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry,* 27(2), 130-135. doi:10.11919/j.issn.1002-0829.215044

Tanner, L., Schreiber, M., Low, J. G. H., Ong, A., Tolfvenstam, T., Lai, Y. L., . . . Ooi, E. E. (2008). Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness. *PLOS Neglected Tropical Diseases,* 2(3), e196. doi:10.1371/journal.pntd.0000196

Vijay Kotu, B. D. (2015). *Predictive Analytics and Data Mining* S. Elliot (Ed.) *Concepts and Practice with RapidMiner* (pp. 425).

Wahyuni, S., Saputra S, K., & Iswan, M. (2017). *THE IMPLEMENTATION OF DECISION TREE ALGORITHM C*4.5 *USING RAPIDMINER IN ANALYZING DROPOUT STUDENTS.*

Walczak, S. (2005). Artificial neural network medical decision support tool: predicting transfusion requirements of ER patients. *IEEE Transactions on Information Technology in Biomedicine,* 9(3), 468-474. doi:10.1109/TITB.2005.847510

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine,* 4(7), 136. doi:10.21037/atm.2016.03.35