P-ISSN 0858-4435
E-ISSN 2651-1231

วารสารวิทยาศาสตร์และเทคโนโลยี Homepage : *http://li01.tci-thaijo.org/index.php/tstj*
ปีที่ 31 ฉบับที่ 4 (กรกฎาคม-สิงหาคม 2566) หน้า [1-17]

ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมสำหรับการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่อง: กรณีศึกษาการคัดกรองโรคเบาหวานในประชากรไทย

Mixed Effect Machine Learning Model for Discrete-Time Survival Analysis: A Case Study of Diabetes Screening Data of Thai Population

มนัสพร ตรีรุ่งโรจน์*, วิฐรา พึ่งพาพงศ์
สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10300
Manusaporn Treerungroj, Vitara Pungpapong
Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok 10330

Received 13 May 2023; Received in revised 28 June 2023; Accepted 5 July 2023

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการทำนายของการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่อง ระหว่างตัวแบบที่พิจารณาว่าข้อมูลตามยาวที่ถูกเก็บมาจากบุคคลคนเดียวกันนั้นมีความสัมพันธ์กันและไม่เป็นอิสระต่อกัน กับตัวแบบที่มองข้ามความสัมพันธ์นั้นและสมมติว่าข้อมูลที่เก็บจากบุคคลคนเดียวกันเป็นอิสระต่อกัน ทั้งนี้ ในงานวิจัยนี้ ผู้วิจัยพิจารณาการสุ่มป่าไม้กับตัวแบบ CatBoost ซึ่งพิจารณาเฉพาะอิทธิพลคงที่ และตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมที่พิจารณาทั้งอิทธิพลคงที่และอิทธิพลสุ่ม จากการวิเคราะห์ข้อมูลเพื่อพยากรณ์การเป็นโรคเบาหวานจากข้อมูลการคัดกรองโรคเบาหวานของกลุ่มตัวอย่างประชากรไทย ซึ่งเป็นข้อมูลที่ขาดความสมดุลสูง พบว่า มีเพียงตัวแบบ CatBoost ที่พิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยมองข้ามความสัมพันธ์ และมีเพียงการสุ่มป่าไม้ที่ใช้อิทธิพลผสมให้ประสิทธิภาพการพยากรณ์สูงกว่าการวิเคราะห์โดยใช้เพียงอิทธิพลคงที่ โดยสรุป งานวิจัยนี้แสดงให้เห็นว่าการพิจารณาความสัมพันธ์ของข้อมูลไม่ได้ส่งผลให้ประสิทธิภาพการพยากรณ์ดีขึ้นเสมอไป ทั้งบนตัวแบบอิทธิพลคงที่และตัวแบบอิทธิพลผสมขึ้นอยู่ข้อจำกัดและปัจจัยต่าง ๆ เช่น ลักษณะข้อมูล การเลือกตัวแบบ การกำหนดตัวแปรอิทธิพลสุ่ม และวิธีการสกัดอิทธิพลคงที่จากตัวแบบต้นไม้ ดังนั้น แม้ว่าตัวแบบการเรียนรู้ของเครื่องที่พิจารณาเฉพาะอิทธิพลคงที่นั้นมักจะถูกใช้ในการพยากรณ์ข้อมูลการรอดชีพแบบเวลาไม่ต่อเนื่อง การใช้ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมอาจเป็นอีกทางเลือกหนึ่งที่ทำให้การพยากรณ์มีความถูกต้องแม่นยำมากขึ้นได้

***ผู้รับผิดชอบบทความ:** m.treerungroj@gmail.com **doi:** .....

คำสำคัญ: การวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่อง; ข้อมูลที่ตรวจตัด; ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ; ตัวแบบอิทธิพลผสม

Abstract

The purpose of this study is to compare the prediction performance of discrete-time survival analysis methods, with and without considering the relationship between longitudinal data observed from the same individuals. In this research, we consider a Random Forest and CatBoost with fixed effects, as well as a mixed-effect machine learning model that considers both fixed and random effects. We applied these methods to predict diabetes status using a diabetes screening dataset collected from the Thai population. It was observed that the dataset is highly imbalanced. Our results show that, for the fixed effect model, considering the relationships among observations from the same individuals resulted in better prediction performance when using CatBoost. However, for the mixed effect model, only the fixed effect components extracted from the Random Forest model achieved higher prediction performance. In summary, this research demonstrates that considering the relationships between data does not always lead to the improvement of prediction performance depending on various limitations and factors such as data characteristics, model selection, random effect variables, and methods of fixed effect component extraction from tree-based models. Therefore, while fixed-effect models are commonly used in discrete-time survival analysis, using a mixed-effect model along with machine learning could be an alternative approach to improve predictive performance.

Keywords: Discrete-time survival analysis; Censored data; Binary classification machine learning models; Mixed effect model

1. บทนำ

การวิเคราะห์การรอดชีพ (Survival Analysis) คือวิธีการวิเคราะห์ทางสถิติเกี่ยวกับการสร้างตัวแบบการรอดชีพเพื่อประมาณค่าอัตราการรอดชีพเมื่อผ่านจุดเวลาที่กำหนด โดยข้อมูลที่ใช้ในการวิเคราะห์การรอดชีพจะมีลักษณะเป็นข้อมูลตามยาว (Longitudinal Data) ที่ข้อมูลแต่ละแถวแสดงถึงการจัดเก็บข้อมูลของตัวอย่างหนึ่ง ณ เวลาหนึ่ง ๆ ประกอบไปด้วยข้อมูลการเกิดเหตุการณ์ที่สนใจ ซึ่งมี 2 สถานะ คือ เกิดเหตุการณ์ และไม่เกิดเหตุการณ์ รวมถึงระยะเวลาการรอดชีพโดยวัดจากจุดเริ่มต้นที่สังเกตจนกระทั่งเกิดหรือไม่เกิดเหตุการณ์ [1]

การวิเคราะห์การรอดชีพถูกนำมาใช้ในหลายสาขาวิชารวมถึงสาขาทางการแพทย์เพื่อประมาณระยะเวลาการเกิดเหตุการณ์ที่สนใจ เช่น ในการพยากรณ์ระยะเวลาที่ผู้ป่วยจะเป็นโรคเบาหวาน บุคคลที่มีความเสี่ยงจะเป็นโรคเบาหวานจะมาตรวจคัดกรองการเป็นโรคเบาหวานอยู่เป็นประจำ ในการคัดกรองแต่ละครั้งจะมีการเก็บข้อมูลสุขภาพของผู้ป่วย ได้แก่ อายุ น้ำหนักส่วนสูง ความดันโลหิต ระดับน้ำตาลในเลือด และข้อมูลประกอบอื่น ๆ เช่น ผลการวินิจฉัยโดยแพทย์ และยาที่ได้รับในครั้งนั้น ซึ่งข้อมูลเหล่านี้จะสามารถนำไปเป็นตัวแปรร่วมในการวิเคราะห์และพยากรณ์การเกิดโรคเบาหวานได้

ในการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่อง (Discrete-time Survival Analysis) จะแบ่งระยะเวลาการรอดชีพออกเป็นช่วงที่มีระยะเวลาเท่า ๆ กัน เช่น ช่วงละ 1 ปี และกำกับสถานะของเหตุการณ์ ว่าในช่วงเวลาแต่ละช่วงได้เกิดเหตุการณ์ที่สนใจหรือไม่ กล่าวคือ ให้ค่าเป็น 1 เมื่อเกิดเหตุการณ์ที่สนใจในช่วงเวลาดังกล่าว และเป็น 0 เมื่อไม่เกิดเหตุการณ์ในช่วงเวลาดังกล่าว ตัวแปรที่สร้างขึ้นใหม่นี้จะถูกใช้เป็นตัวแปรตามในการวิเคราะห์การรอดชีพซึ่งสามารถประยุกต์ใช้อัลกอริทึมการเรียนรู้ของเครื่องใด ๆ ที่สามารถจำแนกประเภททวิในการพยากรณ์ได้ [2]

ในการเตรียมข้อมูลตามยาวสำหรับการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องมักจะจัดเก็บข้อมูลในการตรวจแต่ละครั้งเป็นหนึ่งแถวข้อมูล หากมาตรวจหลายครั้งก็จะมีข้อมูลหลายแถวสำหรับผู้ป่วยหนึ่งคน อย่างไรก็ตาม อัลกอริทึมการเรียนรู้ของเครื่องส่วนมากจะสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกัน แต่ข้อสมมตินี้ไม่เป็นไปตามความเป็นจริงที่ว่าข้อมูลหลายแถวที่เกิดจากบุคคลคนเดียวกันมีความสัมพันธ์กัน [3] ดังนั้นการวิเคราะห์ข้อมูลของแต่ละบุคคลและพยากรณ์ว่าจะเกิดเหตุการณ์ที่สนใจหรือไม่ โดยเรียนรู้จากประวัติอาการครั้งก่อน ๆ ของบุคคลนั้นด้วยจึงน่าจะมีความสมเหตุสมผลในเชิงการประยุกต์ใช้จริงมากกว่า

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการพยากรณ์ของการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องระหว่างกรณีที่พิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน กับกรณีที่มองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง โดยตัวแบบที่พิจารณาประกอบไปด้วยตัวแบบการเรียนรู้ของเครื่องที่ใช้เพียงอิทธิพลคงที่ รวมถึงตัวแบบการเรียนรู้ของเครื่องที่ใช้อิทธิพลผสมซึ่งพิจารณาทั้งอิทธิพลคงที่และอิทธิพลสุ่ม ทั้งนี้ ในงานวิจัยนี้จะศึกษาผ่านข้อมูลการคัดกรองโรคเบาหวานของกลุ่มตัวอย่างประชากรไทยอายุ 35 ปีขึ้นไปภายในประเทศไทย ตั้งแต่ปี พ.ศ. 2557 ถึง พ.ศ. 2563 จากฐานข้อมูลสำนักงานหลักประกันสุขภาพแห่งชาติ (NHSO) จำนวน 1,175 คน

2. วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการพยากรณ์ของการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องเพื่อพยากรณ์การเป็นโรคเบาหวานบนข้อมูลการคัดกรองโรคเบาหวานของกลุ่มตัวอย่างประชากรไทย โดยเปรียบเทียบระหว่างตัวแบบต่อไปนี้

1) ตัวแบบการเรียนรู้ของเครื่องอิทธิพลคงที่ประกอบด้วยการสุ่มป่าไม้กับตัวแบบ CatBoost โดยมองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง

2) ตัวแบบการเรียนรู้ของเครื่องอิทธิพลคงที่ประกอบด้วยการสุ่มป่าไม้กับตัวแบบ CatBoost โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน

3) ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันด้วยตัวแปรอิทธิพลสุ่ม ซึ่งใช้อิทธิพลคงที่จากตัวแบบอิทธิพลคงที่ ประกอบด้วยการสุ่มป่าไม้กับตัวแบบ CatBoost

3. การทบทวนวรรณกรรม

3.1 แนวคิดพื้นฐานเกี่ยวกับการวิเคราะห์การรอดชีพ

การวิเคราะห์การรอดชีพเป็นวิธีการทางสถิติเกี่ยวกับการสร้างตัวแบบการรอดชีพที่วิเคราะห์บนข้อมูลที่ประกอบไปด้วยสถานะของเหตุการณ์ที่สนใจ (Event) และระยะเวลาจากจุดเริ่มต้นจนถึงจุดที่เกิดเหตุการณ์ที่สนใจ (Time to Event) ซึ่งเป็นการศึกษาในช่วงระยะเวลาหนึ่ง เช่น การวิเคราะห์ข้อมูลของผู้ป่วยโรคเบาหวาน ผู้ป่วยคนเดิมจะตรวจและเก็บข้อมูลอย่างสม่ำเสมอ ในที่นี้เหตุการณ์ที่สนใจอาจเป็นการเสียชีวิตจากโรคเบาหวาน และระยะเวลาจากจุดเริ่มต้นจนถึงจุดเสียชีวิตจากโรคเบาหวาน ซึ่งบางกรณีจะไม่สามารถสังเกตเหตุการณ์ที่สนใจได้ เช่น สิ้นสุดช่วงระยะเวลาที่ศึกษาโดยที่ผู้ป่วย

ไม่เสียชีวิตจากโรคเบาหวาน ขาดการติดตามระหว่างศึกษา จึงทำให้ไม่สามารถสรุปได้ว่าท้ายสุดแล้วเกิดเหตุการณ์ที่สนใจหรือไม่ เรียกข้อมูลลักษณะนี้ว่าข้อมูลตรวจตัดด้านขวา (Right-censored Data) ซึ่งการไม่นำเอาข้อมูลตรวจตัดเหล่านี้ไปวิเคราะห์ด้วยอาจทำให้การวิเคราะห์และพยากรณ์เกิดความเอนเอียงและไม่มีประสิทธิภาพ [4] จึงทำให้ตัวแบบที่ใช้พยากรณ์การรอดชีพแตกต่างจากตัวแบบที่ใช้พยากรณ์ปกติทั่วไป สามารถแสดงฟังก์ชันการรอดชีพได้ดังนี้

$$S(t) = P(T \geq t) \quad (1)$$

เมื่อ T คือตัวแปรสุ่มเวลา และ $S(t)$ คือความน่าจะเป็นที่ตัวอย่างรอดชีพหลังจุดเวลา t เมื่อ $t \geq 0$ โดย $S(t)$ จะมีค่า 1 เมื่อ $t=0$ และ $S(t)$ จะมีค่าเพิ่มขึ้นหรือคงที่เท่านั้น โดยไม่มีการลดลงเมื่อ t เพิ่มขึ้น ในขณะที่ฟังก์ชันการแจกแจงความน่าจะเป็นแบบสะสมสามารถแสดงได้ดังนี้

$$F(t) = 1 - S(t) \quad (2)$$

เมื่อ $F(t)$ คือความน่าจะเป็นสะสมที่เวลาที่เหตุการณ์ที่สนใจจะเกิดมีค่าน้อยกว่าจุดเวลา t และฟังก์ชันความหนาแน่นการเสียชีวิตแสดงได้ดังนี้

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) \quad (3)$$

3.2 ตัวแบบการพยากรณ์การรอดชีพแบบเวลาไม่ต่อเนื่อง (Discrete-time Survival Prediction Model)

ในการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องจะแบ่งค่าเวลาการรอดชีพแบบเวลาต่อเนื่องออกเป็น J ช่วงเวลา $(t_0, t_1], (t_1, t_2], I, (t_{J-1}, t_J]$ โดยที่ $t_0=0$ [2] กรณีที่สมมติว่าตัวแปรร่วมของแต่ละบุคคลไม่ผันแปรตามเวลา (Time-invarying Covariates) อัตราพิบัติ (Hazard Rate) ในช่วงเวลาที่ j หรือ $A_j = (t_{j-1}, t_j]$ เมื่อพิจารณาเวกเตอร์ตัวแปรร่วมของบุคคลที่ i หรือ X_i สามารถแสดงได้ดังนี้

$$\lambda_{ij}(X_i) = P(T_i \in A_j | T_i > t_{j-1}, X_i)$$
$$= P(t_{j-1} < T_i \leq t_j | T_i > t_{j-1}, X_i) \quad (4)$$

และฟังก์ชันความน่าจะเป็นแบบไม่ต่อเนื่อง (Discrete Probability Function) สามารถแสดงได้ดังนี้

$$f_{ij} = P(T_i \in A_j | X_i)$$
$$= S(t_{j-1}|X_i) - S(t_j|X_i) \quad (5)$$

โดยภาวะน่าจะเป็นของการรอดชีพ (Survival Likelihood) ของฟังก์ชันความน่าจะเป็นแบบไม่ต่อเนื่องจะสอดคล้องกับภาวะความน่าจะเป็นของตัวแบบแบบทวินามที่มีสมมติฐานว่าตัวบ่งชี้เหตุการณ์เป็นอิสระต่อกัน แสดงได้ดังนี้

$$L = \prod_{i=1}^{n} \prod_{j=1}^{j_i} \lambda_{ij}(X_i)^{y_{ij}} (1 - \lambda_{ij}(X_i))^{1-y_{ij}} \quad (6)$$

เมื่อ n คือจำนวนบุคคลในชุดข้อมูล j_i คือจำนวนช่วงเวลาของบุคคลที่ i ในชุดข้อมูล และ y_{ij} คือสถานะของเหตุการณ์ที่สนใจของบุคคลที่ i ในช่วงเวลาที่ j โดยที่ $y_{ij}=1$ เมื่อเกิดเหตุการณ์ที่สนใจกับบุคคลที่ i ในช่วงเวลาที่ j และ $y_{ij}=0$ ในกรณีอื่น ๆ

ในการวิเคราะห์เวลาการรอดชีพแบบเวลาไม่ต่อเนื่องจึงสามารถประยุกต์ใช้วิธีการจำแนกประเภททวิเพื่อคำนวณความน่าจะเป็นที่จะเกิดเหตุการณ์ได้ โดยให้ตัวแปรตาม คือ การเกิดเหตุการณ์ที่สนใจในช่วงระยะเวลานั้น (d_{ij} หรือ Event) นั่นเอง

3.3 ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ (Binary Classification Machine Learning Models)

ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิเป็นวิธีการเรียนรู้โดยมีผู้สอนเพื่อจำแนกข้อมูล 2 กลุ่ม ในงานวิจัยนี้จะศึกษาโดยใช้การสุ่มป่าไม้ (Random Forests) และตัวแบบ CatBoost (Categorical Boost)

3.3.1. การสุ่มป่าไม้

การสุ่มป่าไม้เป็นหนึ่งในวิธีการเรียนรู้แบบกลุ่ม (Ensemble Learning Method) ที่จะนำผลลัพธ์การพยากรณ์จากหลาย ตัวแบบมารวมกันทำให้ได้ผลลัพธ์ที่มีความแกร่ง (Robustness) จึงเป็นที่นิยมอย่างแพร่หลายการสุ่มป่าไม้สามารถใช้วิเคราะห์ได้ทั้งการจำแนกประเภทและการพยากรณ์ค่าตัวเลข กลไกของการสุ่มป่าไม้จะสร้างต้นไม้ตัดสินใจ (Decision Trees) หลาย ๆ ต้นที่เรียนรู้บนชุดข้อมูลที่สุ่มแบบคืนที่หลาย ๆ ชุดที่มีขนาดตัวอย่างเท่ากัน ทำให้ได้ต้นไม้ตัดสินใจที่แตกต่างกันหลายต้น ในกรณีของการจำแนกประเภท ผลลัพธ์ที่ถูกทำนายโดยต้นไม้ตัดสินใจมากที่สุดจะเป็นผลลัพธ์สุดท้าย และในกรณีของการถดถอย ผลลัพธ์สุดท้ายจะเป็นค่าเฉลี่ยของผลลัพธ์จากต้นไม้ตัดสินใจทั้งหมด การสุ่มป่าไม้มีความคงทนต่อค่าผิดปกติค่ารบกวน และลดปัญหาเกินพอดี (Overfitting) [5]

3.3.2. ตัวแบบ CatBoost

ตัวแบบ CatBoost คือตัวแบบที่มีเค้าโครงจากวิธี Gradient Boosting ซึ่งพัฒนาต่อยอดมาจากต้นไม้ตัดสินใจ มีจุดเด่นในการจัดการกับตัวแปรจำแนกประเภท (Categorical Variable) ด้วยเทคนิคการเรียงสับเปลี่ยนแบบสุ่มบนชุดข้อมูล จากนั้นจึงคำนวณค่าทางสถิติโดยใช้ข้อมูลผลเฉลยเพื่อทดแทนข้อมูลจำแนกประเภทนั้น และใช้เทคนิคการลดการรบกวนจากกลุ่มข้อมูลที่มีจำนวนตัวอย่างน้อย [6, 7] ทำให้สามารถลดปัญหาเกินพอดีได้ [8]

3.4 ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสม (Mixed Effect Machine Learning Model)

เนื่องจากบุคคลหนึ่ง ๆ จะถูกสังเกตและเก็บข้อมูลหลายครั้งในหลายช่วงเวลา ข้อมูลเหล่านั้นจะใช้เป็นตัวแปรร่วมในการวิเคราะห์การรอดชีพซึ่งแบ่งเป็น 2 ประเภท ได้แก่ 1) ตัวแปรร่วมไม่ผันแปรตามเวลา (Time-invariant Covariates) คือ ตัวแปรที่คงที่เสมอทุกครั้งที่เก็บข้อมูล เช่น เพศ เชื้อชาติ และ 2) ตัวแปรร่วมผันแปรตามเวลา (Time-varying Covariates) คือ ตัวแปรที่มีค่าเปลี่ยนไปในแต่ละครั้งที่สังเกตหรือเก็บข้อมูล เช่น ในการตรวจสุขภาพแต่ละครั้ง ระดับไขมันหรือน้ำตาลแต่ละครั้งจะมีค่าเปลี่ยนแปลงไป ไม่คงที่ เป็นต้น

ตัวแบบอิทธิพลผสมเชิงเส้นนัยทั่วไป (Generalized Linear Mixed Models, GLMM) ถูกพัฒนามาจากตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Models, GLM) สำหรับการวิเคราะห์ข้อมูลผลลัพธ์ที่ครอบคลุมทั้งแบบสัมพันธ์กัน เป็นอิสระต่อกัน ต่อเนื่องและไม่ต่อเนื่อง ภายใต้การนำอิทธิพลแบบคงที่ (Fixed Effect) และอิทธิพลแบบสุ่ม (Random Effect) มาพิจารณาร่วมกัน จึงเรียกว่าอิทธิพลผสม (Mixed Effect) ซึ่งจะได้ผลลัพธ์ครอบคลุมทั้งแบบค่าเฉลี่ยประชากร และผลกระทบในระดับเฉพาะบุคคล [9]

ในการวิเคราะห์ตัวแบบอิทธิพลผสมเชิงเส้นนัยทั่วไป จะใช้ตัวแปรร่วมคงที่เป็นตัวแปรร่วมอิทธิพลคงที่ และใช้ตัวแปรร่วม ผันแปรตามเวลาเป็นตัวแปรร่วมอิทธิพลสุ่ม อย่างไรก็ตาม ตัวแปรอิทธิพลสุ่มควรจะเป็นข้อมูลที่จัดเป็นกลุ่มมากกว่า 5 กลุ่มขึ้นไป เนื่องจากการประมาณอิทธิพลสุ่มจะพยายามกำหนดค่าความแปรปรวนระหว่างแต่ละกลุ่ม จึงต้องการจำนวนกลุ่มที่เพียงพอต่อการประมาณอย่างแม่นยำและไม่ผิดพลาด หากมีจำนวนกลุ่มน้อยกว่า 5 กลุ่ม ควรพิจารณาเป็นตัวแปรอิทธิพลคงที่แทน [10]

สำหรับช่วงเวลา t และบุคคล i จะมีตัวแปรร่วมอิทธิพลคงที่เป็นเวกเตอร์ X_{it} ขนาด p มิติ ตัวแปรร่วมอิทธิพลสุ่ม เป็นเวกเตอร์ z_{it} ขนาด q มิติ และตัวแปรตาม y_{it} สามารถแสดงพารามิเตอร์ของประชากรอิทธิพลคงที่ได้ดังนี้

$$\eta_{it} = g(\mu_{it}) = log(\frac{\mu_{it}}{1-\mu_{it}})$$
$$= \beta^T x_{it} + b_i^T z_{it} \quad (7)$$

เมื่อเวกเตอร์ β คือ พารามิเตอร์อิทธิพลคงที่ เวกเตอร์ b_i คือพารามิเตอร์อิทธิพลสุ่ม $\mu_{it} = E[y_{it}|b_i]$ คือ ค่าคาดหวังของตัวแปรสุ่มทวินาม y_{it} เมื่อกำหนดให้ b_i เป็นค่าคงที่ ดังนั้น μ_{it} ก็คือความน่าจะเป็นที่

จะเกิดเหตุการณ์ของบุคคล i ในช่วงเวลา t และ $g(\cdot)$ คือ ฟังก์ชันการเชื่อมโยงโลจิท (Logit Link Function)

การประมาณค่า β สามารถทำได้โดยใช้วิธี Penalized Quasi-likelihood (PQL) [11] ซึ่งประมาณตัวแปรตามโดยใช้ผลรวมระหว่างค่าเฉลี่ย μ_{it} และค่าคลาดเคลื่อน ε_{it} จากนั้นจึงใช้การกระจายเทย์เลอร์ (Taylor Expansion) รอบ ๆ ค่าประมาณ $(\hat{\beta}, \hat{b}_i)$ ดังนี้

$$y_{it} = \mu_{it} + \varepsilon_{it} \quad (8)$$

$$= h(\eta_{it}) + \varepsilon_{it} \quad (9)$$

$$\approx h(\hat{\eta}_{it}) + h'(\hat{\eta}_{it})(\eta_{it} - \hat{\eta}_{it}) + \varepsilon_{it} \quad (10)$$

โดยที่ฟังก์ชัน $h(\cdot)$ คือฟังก์ชันผกผันของฟังก์ชัน $g(\cdot)$ ซึ่งหากจัดสมการที่ (10) จะสามารถแสดงสมการของตัวแบบอิทธิพลผสมเชิงเส้นได้ดังนี้

$$y_{it}^* = \beta^T x_{it} + b_i^T z_{it} + \varepsilon_{it}^* \quad (11)$$

เมื่อ $y_{ii}^* = (y_{it} - \hat{\mu}_{it})g'(\hat{\mu}_{it}) + g(\hat{\mu}_{it})$

และ $\varepsilon_{it}^* = g'(\hat{\mu}_{it})\varepsilon_{it}$ ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมนี้ จะประมาณส่วนประกอบอิทธิพลคงที่ $(\beta^T x_{it})$ ด้วยอัลกอริทึมการเรียนรู้ของเครื่อง และประมาณอิทธิพลสุ่ม (b_i) ด้วยตัวแบบอิทธิพลผสมเชิงเส้นนัยทั่วไปจนกว่าตัวแบบจะลู่เข้า [12]

4. วิธีการดำเนินการวิจัย

ขั้นตอนการดำเนินการวิจัยแบ่งเป็น 3 ขั้นตอน

4.1 ขั้นตอนการเตรียมข้อมูลการรอดชีพแบบเวลาไม่ต่อเนื่อง

ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานนี้เป็นข้อมูลตัวอย่างได้ที่รับการอนุญาตเพื่อนำมาใช้ในการวิจัยจากสำนักงานหลักประกันสุขภาพแห่งชาติ ประกอบไปด้วย 6,848 แถว จากกลุ่มตัวอย่าง 1,175 คน ซึ่งถูกสุ่มมาจากผู้ที่เคยเข้ารับการตรวจคัดกรองโรคเบาหวานในประเทศไทยทั้งหมด โดยแต่ละแถวคือข้อมูลการตรวจคัดกรองรายปีของแต่ละบุคคล ข้อมูลที่จัดเก็บในปีแรกของผู้เข้ารับการตรวจคัดกรองแต่ละคนจะเริ่มต้นจากสถานะไม่เป็นโรคเบาหวาน และได้เข้ารับการตรวจคัดกรองทุกปีจนกระทั่งสิ้นสุดระยะเวลาเก็บข้อมูล โครงสร้างข้อมูลแสดงดัง Table 1 โดยคอลัมน์ next_event เป็นสถานะของบุคคลในปีถัดไปที่ต้องการพยากรณ์หรือคือตัวแปรตามในงานวิจัยนี้

Table 1 Schema of Diabetes screening dataset

Variable names	Data types	Description
PID_EN	Text	Encoded ID number
gender	Nominal	Gender (1=male, 2=female)
age	Integer	Age (years)
year	Integer	Year of data collection (CE)
BMI	Numeric	Body mass index
WAIST_CM	Numeric	Waist circumference (centimeters)
SBP	Numeric	Systolic blood pressure during contraction of the heart (mm)

Table 1 Schema of Diabetes screening dataset (Cont.)

Variable names	Data types	Description
DBP	Numeric	Diastolic blood pressure during relaxation of the heart (mm)
SMOKE	Nominal	Smoking history (1=never smoked, 2=smoked in the past, 3=occasional smoker, 4=regular smoker)
DMFAMILY	Nominal	Family history of diabetes (0=no, 1=yes)
HTFAMILY	Nominal	Family history of high blood pressure (0=no, 1=yes)
HT	Nominal	High blood pressure status (0=no, 1=yes)
next_event	Nominal	Status in the following year (0=not diagnosed with diabetes, 1=diagnosed with diabetes)

จากข้อมูลพบว่าจำนวนแถวข้อมูลของแต่ละบุคคลไม่เท่ากัน และมี 8 คอลัมน์ที่มีข้อมูลสูญหาย ดัง Table 2 โดยจะประมาณค่าข้อมูลสูญหายสำหรับคอลัมน์ BMI, WAIST_CM, SBP และ DBP ด้วยค่ากลางของแต่ละบุคคล และประมาณค่าข้อมูลสูญหายสำหรับคอลัมน์ SMOKE, DMFAMILY, HTFAMILY และ HT ด้วยค่าอื่นที่พบในข้อมูลของบุคคลนั้น โดยตั้งสมมติฐานว่าพฤติกรรมการสูบบุหรี่ ประวัติทางสุขภาพของญาติ และภาวะความดันโลหิตสูงของบุคคลไม่เปลี่ยนแปลง

Table 2 Number of missing data in diabetes screening dataset

Variable names	Data types	Number of missing data (rows)	Number of missing data (%)
BMI	Numeric	2,506	36.5946
WAIST_CM	Numeric	2,506	36.5946
SBP	Numeric	2,504	36.5654
DBP	Numeric	2,504	36.5654
SMOKE	Nominal	3,603	52.6139
DMFAMILY	Nominal	3,220	47.0210
HTFAMILY	Nominal	6,142	89.6904
HT	Nominal	2,466	36.0105

เมื่อคัดกรองเฉพาะแถวข้อมูลของบุคคลที่มาตรวจคัดกรองทุกปีตั้งแต่ปีแรกที่เข้ารับการตรวจคัดกรองจนถึงปีที่ตรวจคัดกรองแล้วพบว่าเป็นโรคเบาหวาน หรือสิ้นสุดระยะเวลาที่เก็บข้อมูลเท่านั้นจะเหลือข้อมูลจำนวน 5,027 แถว จากกลุ่มตัวอย่างจำนวน 1,175 คน หลังจากนั้นจึงแบ่งชุดข้อมูลออกเป็นข้อมูลสอน 70% และข้อมูลทดสอบ 30% โดยจะสรุปสถานะเหตุการณ์สุดท้ายของแต่ละบุคคล ซึ่งมีสถานะที่เป็นไปได้ 2 สถานะ คือ เป็นโรคเบาหวาน และไม่เป็นโรคเบาหวาน แล้วทำการสุ่มตัวอย่างแบบแบ่งชั้นภูมิจากข้อมูลสรุปสถานะเหตุการณ์สุดท้ายของแต่ละบุคคลนั้น โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มตามสถานะ แล้วจึงสุ่มตัวอย่างจากแต่ละกลุ่มสถานะตามสัดส่วนที่ได้วางไว้ เพื่อคงสัดส่วนสถานะของเหตุการณ์ที่สนใจให้เท่ากัน และจะกำหนดให้แต่ละ ID อยู่ในชุดข้อมูลใดชุดข้อมูลหนึ่งด้วยวิธีการสุ่มแบบไม่ใส่คืน จะได้ข้อมูลสอนจำนวน 3,525 แถวจาก 865 คน และข้อมูลทดสอบจำนวน 1,502 แถวจาก 370 คน มีสัดส่วนสถานะของเหตุการณ์ ไม่เป็นโรคเบาหวาน : เป็นโรคเบาหวาน ประมาณ 96 : 4

4.2 ขั้นตอนการวิเคราะห์ข้อมูลและสร้างตัวแบบ

การวิเคราะห์ข้อมูลและสร้างตัวแบบจะเริ่มต้นจากการฝึกฝนตัวแบบด้วยข้อมูลสอนทั้งหมดเพื่อหาชุดของพารามิเตอร์ที่เหมาะสม หลังจากนั้นจะเข้าสู่ขั้นตอนการวนซ้ำเพื่อวิเคราะห์ตัวแบบ ในแต่ละรอบจะทำบูตสแทร็ปบนข้อมูลสอน โดยแบ่งข้อมูลสอนออกเป็นกลุ่มตามบุคคล ในแต่ละกลุ่มจะทำการสุ่มแถวข้อมูลแบบซ้ำกันได้ตามจำนวนแถวข้อมูลทั้งหมดในกลุ่ม แล้วจึงรวมแถวข้อมูลที่สุ่มได้จากทุกกลุ่มเข้าด้วยกัน เรียกว่า ข้อมูลบูตสแทร็ป หลังจากนั้นจึงฝึกฝนตัวแบบด้วยข้อมูลบูตสแทร็ปนั้น และวัดประสิทธิภาพของตัวแบบที่ได้บนข้อมูลทดสอบ จึงจะถือว่าจบหนึ่งรอบการวิเคราะห์ เมื่อวิเคราะห์ครบทั้งหมด 100 รอบ จะสรุปตัววัดผลด้วยวิธีการหาค่าเฉลี่ยจากตัววัดผลทั้งหมด แบ่งการวิเคราะห์เป็น 2 กรณี คือ

1) กรณีที่วิเคราะห์โดยมองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน และสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง โดยใช้การสุ่มป่าไม้กับตัวแบบ CatBoost ซึ่งพิจารณาเฉพาะอิทธิพลคงที่

2) กรณีที่วิเคราะห์โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันในการวิเคราะห์ โดยใช้การสุ่มป่าไม้กับตัวแบบ CatBoost ซึ่งพิจารณาเฉพาะอิทธิพลคงที่และตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันด้วยตัวแปรอิทธิพลสุ่ม

สำหรับการวิเคราะห์ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมที่จะใช้อิทธิพลคงที่จากการสุ่มป่าไม้กับตัวแบบ CatBoost และใช้อิทธิพลสุ่มจากตัวแบบอิทธิพลผสม [12] มีขั้นตอนการวิเคราะห์ดังนี้

1) ฝึกฝนตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่บนทุกตัวแปร และตัวแปรอิทธิพลสุ่ม

2) สกัดเส้นทางทำนายจากตัวแบบอิทธิพลคงที่ เพื่อใช้เป็นตัวแปรกลุ่มในการวิเคราะห์ตัวแบบอิทธิพลผสม

3) ฝึกฝนตัวแบบอิทธิพลผสม

4) สกัดค่าอิทธิพลสุ่มจากตัวแบบอิทธิพลผสม เพื่อใช้เป็นตัวแปรอิทธิพลสุ่มในการฝึกฝนตัวแบบอิทธิพลคงที่

5) ทำขั้นตอนที่ 1-4จนกระทั่งตัวแบบอิทธิพลผสมลู่เข้าหรือครบจำนวนรอบสูงสุดที่ตั้งไว้

4.3 ขั้นตอนการเปรียบเทียบผลลัพธ์

การเปรียบเทียบผลลัพธ์จะใช้ตัววัดผลสำหรับวิธีการจำแนกประเภททวิ ได้แก่ ความแม่น (Accuracy) ความเที่ยง (Precision) ความไว (Sensitivity หรือ Recall) ความจำเพาะ (Specificity) คะแนน F1 (F1-score) R-squared จากคะแนน Brier (R-squared Measure of Brier Score) พื้นที่ใต้กราฟ ROC (Area Under the ROC Curve: AUC) และพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลา (Time-dependent Area Under the ROC Curve: Time-dependent AUC_T)

4.3.1. ความแม่น ความเที่ยง ความไว ความจำเพาะ และคะแนน F1

ผลลัพธ์การพยากรณ์ความเสี่ยงต่อการไม่รอดชีพ

แบบทวิสามารถแสดงอยู่ในรูปตารางสับสน (Confusion Matrix) ได้ดัง Table 3 ทั้งนี้ตารางสับสนจะแสดงให้เห็นถึงความถี่ของผลลัพธ์ที่ทำนายถูกต้องหรือผิดพลาด [13] ดังนี้

1) True Positive (TP) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่าไม่รอดชีพและพยากรณ์ได้ถูกต้อง

2) False Positive (FP) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่าไม่รอดชีพแต่พยากรณ์ผิดพลาด

3) True Negative (TN) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่ารอดชีพและพยากรณ์ได้ถูกต้อง

4) False Negative (FN) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่ารอดชีพแต่พยากรณ์ผิดพลาด

Table 3 Binary confusion matrix

		Predicted condition	
		Positive (Not survive)	**Negative (Survive)**
Actual condition	**Positive (Not survive)**	True Positive (TP)	False Negative (FN)
	Negative (Survive)	False Positive (FP)	True Negative (TN)

จากความถี่ของผลลัพธ์ของการพยากรณ์ข้างต้น สามารถนำไปคำนวณตัววัดผลเพื่อวัดประสิทธิภาพการทำนายได้ ดังนี้

1) ความแม่น $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ คือ อัตราส่วนของการพยากรณ์ที่ถูกต้องเทียบกับการพยากรณ์ทั้งหมด

2) ความเที่ยง $Precision = \frac{TP}{TP+FP}$ คือ อัตราส่วนของการพยากรณ์ว่าไม่รอดชีพที่ถูกต้องเทียบกับการพยากรณ์ว่าไม่รอดชีพทั้งหมด

3) ความไว $Sensitivity = \frac{TP}{TP+FN}$ คือ อัตราส่วนของการพยากรณ์ว่าไม่รอดชีพที่ถูกต้องเทียบกับสถานะเหตุการณ์จริงที่ไม่รอดชีพทั้งหมด

4) ความจำเพาะ $Specificity = \frac{TN}{TN+FP}$ คือ อัตราส่วนของการพยากรณ์ว่ารอดชีพที่ถูกต้องเทียบกับสถานะเหตุการณ์จริงที่รอดชีพทั้งหมด

5) คะแนน F1 $F1\ score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$ คือ ค่าเฉลี่ยฮาร์โมนิกระหว่างตัววัดความเที่ยงและความไว

4.3.2. R-squared จากคะแนน Brier

คะแนน Brier เป็นค่าความคลาดเคลื่อนในการพยากรณ์เมื่อเทียบกับสถานะเหตุการณ์จริง สามารถแสดงได้ดังนี้

$$BS = E[(D(t) - p(t))^2]$$
$$= \frac{1}{N}\sum_{t=1}^{N}(D(t) - p(t))^2 \quad (12)$$

เมื่อ N คือจำนวนข้อมูลสังเกตทั้งหมด $D(t)$ คือสถานะเหตุการณ์จริงที่เกิดภายในเวลา t โดยมีค่าเป็น 1 เมื่อเกิดเหตุการณ์ที่สนใจ และเป็น 0 เมื่อไม่เกิดเหตุการณ์ และ $p(t)$ คือความน่าจะเป็นที่พยากรณ์ว่าจะเกิดเหตุการณ์ภายในเวลา t คะแนน Brier จะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยค่าน้อยจะแสดงถึงประสิทธิภาพการพยากรณ์ที่ดี

R-squared คือ ค่าสัมประสิทธิ์การตัดสินใจพหุคูณที่สามารถบ่งชี้ถึงความสอดคล้องของตัวแบบกับข้อมูลที่ใช้สร้างตัวแบบ โดยเป็นสัดส่วนของความแปรปรวนที่อธิบายโดยตัวแบบต่อความแปรปรวนทั้งหมดของข้อมูล [13] ซึ่งสามารถคำนวณ R-squared จากคะแนน Brier ได้ดังนี้

$$R^2 = 1 - \frac{BS(t)}{BS_0(t)} \quad (13)$$

เมื่อ $BS_0(t)$ คือคะแนน Brier ของตัวแบบตั้งต้นที่ให้ความน่าจะเป็นที่พยากรณ์ว่าจะเกิดเหตุการณ์เหมือนกัน ในทุกข้อมูลสังเกต หรือก็คือ $BS_0(t)$ จะเป็นคะแนน Brier สูงสุดที่เป็นไปได้ของตัวแบบและชุดข้อมูลสังเกตนี้ [2] ดังนั้น R-squared ที่มีค่ามากจะแสดงถึงตัวแบบที่สามารถอธิบายความผันแปรของค่าตัวแปรตามได้มาก

4.3.3. พื้นที่ใต้กราฟ ROC และพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลา

กราฟ ROC เป็นตัววัดผลการพยากรณ์ที่แสดงในรูปแบบของกราฟความสัมพันธ์ระหว่างความไวและความจำเพาะบนช่วงของเกณฑ์การพยากรณ์ พื้นที่ใต้กราฟ ROC สามารถบ่งบอกได้ว่าตัวแบบมีความสามารถในการแยกแยะความแตกต่างระหว่างกลุ่มได้ดีเพียงใด มีค่าอยู่ระหว่าง 0 ถึง 1 โดยค่ามากแสดงถึงความสามารถในการแยกแยะความแตกต่างระหว่างกลุ่มได้ดี การวัดผลด้วยพื้นที่ใต้กราฟ ROC แบบปกติจะทำให้ทราบประสิทธิภาพการพยากรณ์โดยรวม [13] อย่างไรก็ตาม ในการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่อง สถานะเหตุการณ์ของแต่ละบุคคลเปลี่ยนไปตามช่วงเวลา ดังนั้นจะวัดพื้นที่ใต้กราฟ ROC แบ่งตามครั้งที่ติดตามด้วยเพื่อทราบถึงแนวโน้มประสิทธิภาพการพยากรณ์เมื่อจำนวนแถวข้อมูลในกลุ่มหรือจำนวนครั้งที่ติดตามเพิ่มขึ้น

5. ผลการวิจัย

ในการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบ ตัววัดผลที่ค่าขึ้นกับจุดตัดจะใช้จุดตัดจากดัชนีของ Youden (Youden's Index) ผลการวิจัยจะแบ่งเป็น 2 ส่วนตามตัวแบบที่ศึกษา ได้แก่ การสุ่มป่าไม้ และตัวแบบ CatBoost

5.1 ผลการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องด้วยการสุ่มป่าไม้

ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ของการสุ่มป่าไม้บนข้อมูลทดสอบแสดงดังTable 4 เนื่องจากข้อมูลนี้ขาดความสมดุลสูง จึงสังเกตได้ว่าตัววัดผลที่ค่าขึ้นกับจุดตัดบนข้อมูลนั้นมีความเอนเอียงไปทางค่าใดค่าหนึ่งสูง โดยพบว่า สำหรับการสุ่มป่าไม้ซึ่งพิจารณาเฉพาะอิทธิพลคงที่ การวิเคราะห์โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยมองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง เฉพาะเมื่อพิจารณาจากความแม่นเท่านั้น

ในอีกแง่หนึ่ง พบว่าประสิทธิภาพการพยากรณ์ของการสุ่มป่าไม้โดยมองข้ามความสัมพันธ์ของข้อมูลนั้นกับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากการสุ่มป่าไม้สามารถเทียบเคียงกันได้ อย่างไรก็ตาม เนื่องจากคะแนน F1 เป็นค่าเฉลี่ยฮาร์โมนิกระหว่างความเที่ยงกับความไว จึงอาจสามารถเปรียบเทียบโดยใช้คะแนน F1 แทนความเที่ยงกับความไวได้ ซึ่งสรุปได้ว่า การวิเคราะห์โดยพิจารณาอิทธิพลผสมให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยพิจารณาเพียงอิทธิพลคงที่ เมื่อพิจารณาจากความแม่นความจำเพาะ และพื้นที่ใต้กราฟ ROC

นอกเหนือจากนี้ ในเชิงทางการแพทย์จะให้ความสำคัญกับอัตราการค้นพบผิดพลาด (False Discovery Rate = 1 - Precision) และอัตราการไม่ตรวจพบผลลัพธ์ที่จริง (False Negative Rate = 1 - Sensitivity) จากการศึกษาพบว่า ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมสามารถลดอัตราการค้นพบผิดพลาด และอัตราการไม่ตรวจพบผลลัพธ์ที่จริงได้ เมื่อเปรียบเทียบกับตัวแบบอิทธิพลคงที่โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน

Table 4 Performance evaluation of the Random Forest model on the test data

Models	Considering relationships	Accuracy	Precision	Sensitivity	Specificity	F1 score	R-squared	AUC
Random Forest	X	0.8294	0.0508*	0.1842*	0.8562	0.0779*	0.9593*	0.5564
	✓	0.8816	0.0279	0.0578	0.9158	0.0371	0.9584	0.4897
Mixed Effect Model with Random Forest	✓	0.9052*	0.0478	0.0698	0.9400*	0.0540	0.9481	0.5608*

"*" indicates the best metric.

Note. "Considering relationships" means the consideration of relationships between data from the same individuals, where the symbol " X " indicates that the model is ignoring the relationships, and the symbol " ✓ " indicates that the model is considering the relationships.

เมื่อพิจารณากราฟเส้นแสดงประสิทธิภาพการพยากรณ์ด้วยพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลาดัง Figure 1 พบว่าทุกตัวแบบมีแนวโน้มของประสิทธิภาพกาพยากรณ์ขึ้นลงไปในทิศทางเดียวกันเมื่อเวลาที่ติดตามเพิ่มขึ้น ซึ่งสอดคล้องกับพื้นที่ใต้กราฟ ROC ใน Table 4 กล่าวคือ การสุ่มป่าไม้โดยมองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง กับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากการสุ่มป่าไม้สามารถเทียบเคียงกันได้

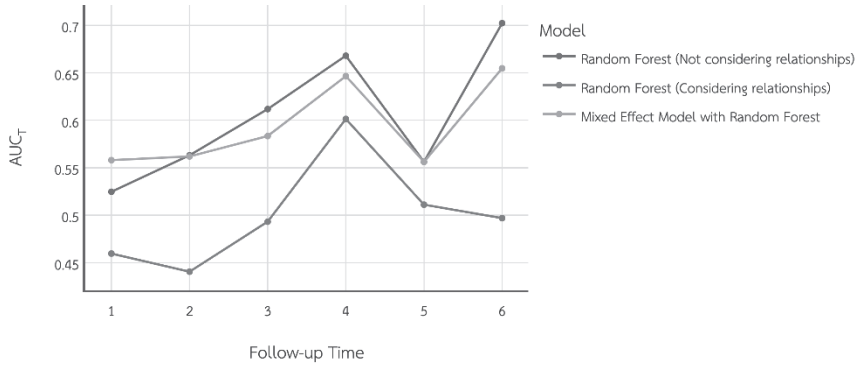

Figure 1 Time-dependent AUC of the Random Forest model on the test data. The horizontal axis is the follow-up time, and the vertical axis is the AUC_T

5.2 ผลการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องด้วยตัวแบบ CatBoost

ผลการเปรียบเทียบประสิทธิภาพของตัวแบบ CatBoost บนข้อมูลทดสอบแสดงดัง Table 5 โดยที่ตัววัดผลที่ค่าขึ้นกับจุดตัดบนข้อมูลมีความเอนเอียงไปทางค่าใดค่าหนึ่งสูง จากการศึกษาพบว่า สำหรับตัวแบบ CatBoost ซึ่งพิจารณาเฉพาะอิทธิพลคงที่การวิเคราะห์โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยมองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง เมื่อพิจารณาจากคะแนน F1, R-squared และพื้นที่ใต้กราฟ ROC ในขณะที่การวิเคราะห์โดยพิจารณาอิทธิพลผสมให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยพิจารณาเพียงอิทธิพลคงที่ เฉพาะเมื่อพิจารณาจากความแม่น และความจำเพาะเท่านั้น

Table 5 Performance evaluation of the CatBoost model on the test data

Models	Considering relationships	Accuracy	Precision	Sensitivity	Specificity	F1 score	R-squared	AUC
CatBoost	X	0.7816	0.0552*	0.2742	0.8027	0.0909	0.9599	0.5824
	✓	0.6300	0.0548	0.5067*	0.6351	0.0986*	0.9601*	0.6051*
Mixed Effect Model with CatBoost	✓	0.8956*	0.0449	0.0772	0.9297*	0.0507	0.9365	0.5299

"*" indicates the best detection.

Note. "Considering relationships" means the consideration of relationships between data from the same individuals, where the symbol " X " indicates that the model is ignoring the relationships, and the symbol " ✓ " indicates that the model is considering the relationships.

เมื่อพิจารณากราฟเส้นแสดงพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลาดัง Figure 2 พบว่าทั้งสองตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่มีประสิทธิภาพการพยากรณ์ที่เพิ่มขึ้นสลับกับลดลง กล่าวคือ ประสิทธิภาพการพยากรณ์มีแนวโน้มเพิ่มขึ้นในช่วงเวลาติดตามที่ 1 ถึง 4 ก่อนที่ประสิทธิภาพการพยากรณ์จะลดลง ณ เวลาติดตามที่ 5 และเพิ่มขึ้นอีกครั้ง ณ เวลาติดตามที่ 6 ในขณะที่ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมมีแนวโน้มของประสิทธิภาพการพยากรณ์ที่เพิ่มขึ้นเมื่อเวลาที่ติดตามเพิ่มขึ้น

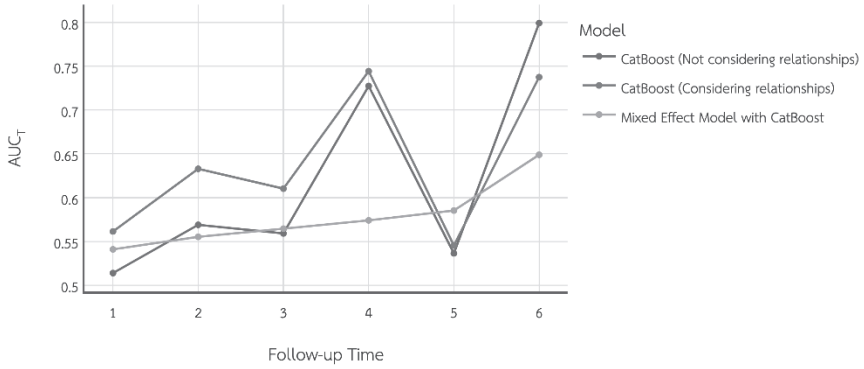

Figure 2 Time-dependent AUC of the CatBoost model on the test data. The horizontal axis is the follow-up time, and the vertical axis is the AUC_T

6. สรุป

6.1 สรุปผลการวิจัย

จากการศึกษาพบว่าทั้ง 2 ตัวแบบให้ผลลัพธ์การพยากรณ์ที่แตกต่างกัน กล่าวคือ

1) การวิเคราะห์โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยมองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิงเฉพาะตัวแบบ CatBoost แต่ให้ประสิทธิภาพการพยากรณ์ที่แย่ลงสำหรับการสุ่มป่าไม้ เมื่อพิจารณาจากคะแนน F1, R-squared และพื้นที่ใต้กราฟ ROC

2) การวิเคราะห์โดยพิจารณาอิทธิพลผสมให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยพิจารณาเฉพาะอิทธิพลคงที่เฉพาะการสุ่มป่าไม้ เมื่อพิจารณาจากความแม่น ความจำเพาะ และพื้นที่ใต้กราฟ ROC แต่ให้ประสิทธิภาพการพยากรณ์ที่แย่ลงสำหรับตัวแบบ CatBoost เมื่อพิจารณาจากคะแนน F1, R-squared และพื้นที่ใต้กราฟ ROC

จากผลการศึกษาดังกล่าว สามารถสรุปผลการวิจัยได้ดัง Table 6 คือ มีเพียงตัวแบบ CatBoost ที่การวิเคราะห์โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยมองข้ามความสัมพันธ์นั้นและสมมติว่าข้อมูลที่เก็บจากบุคคลคนเดียวกันเป็นอิสระต่อกัน และมีเพียงการสุ่มป่าไม้ที่การวิเคราะห์โดยใช้อิทธิพลผสมให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าการวิเคราะห์โดยใช้เพียงอิทธิพลคงที่

Table 6 Conclusion of considering the relationships compared to ignoring the relationships

Models	Types of effect	Result
Random Forest	Fixed	-
Mixed Effect Model with Random Forest	Mixed	Better
CatBoost	Fixed	Better
Mixed Effect Model with CatBoost	Mixed	-

ดังนั้นจึงสรุปได้ว่าการพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวทั้งการใช้ตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่ และตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมไม่ได้ให้ผลลัพธ์การพยากรณ์ที่ดีขึ้นเสมอไป ขึ้นอยู่กับตัววัดผลที่เลือกพิจารณา ทั้งนี้ปัจจัยต่าง ๆ ที่เกี่ยวข้องจะกล่าวถึงในส่วนถัดไป

6.2 อภิปรายผลการวิจัย

การวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องด้วยตัวแบบที่ผสมผสานระหว่างอิทธิพลคงที่และอิทธิพลสุ่มนั้นมีความน่าสนใจ และมีความสมเหตุสมผลในเชิงการนำไปประยุกต์ใช้จริง โดยเฉพาะกับการวิเคราะห์ข้อมูลทางการแพทย์หรือสุขภาพ อย่างไรก็ตาม จากการศึกษาบนชุดข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานนี้ พบว่ามีหลายปัจจัยที่อาจส่งผลต่อประสิทธิภาพการพยากรณ์ของตัวแบบ เพื่อให้การเปรียบเทียบมีความยุติธรรม ดังนั้น ผู้วิจัยจึงได้ควบคุมปัจจัยเหล่านั้นให้เหมือนกันในทุกตัวแบบ ได้แก่

1) ชุดข้อมูล โดยใช้ข้อมูลสอน ข้อมูลทดสอบ และข้อมูลบูตสแทร็ปในแต่ละรอบการวิเคราะห์เป็นข้อมูลชุดเดียวกัน

2) วิธีการปรับพารามิเตอร์ของตัวแบบ โดยการฝึกฝนตัวแบบด้วยข้อมูลสอนทั้งหมด แล้วจึงเลือกชุดของพารามิเตอร์ที่ให้พื้นที่ใต้กราฟ ROC สูงสุดเป็นชุดของพารามิเตอร์ที่เหมาะสมที่สุด

3) การกำหนดตัวแปรอิทธิพลสุ่ม โดยกำหนดตัวแปรอิทธิพลสุ่มเป็นตัวแปรเดียวกัน

4) วิธีการสกัดสิ่งสำคัญจากตัวแบบอิทธิพลคงที่ ซึ่งเป็นปัจจัยสำคัญที่ส่งผลต่อตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสม โดยผู้วิจัยได้ศึกษาวิธีการสกัดอิทธิพลคงที่จากการสุ่มป่าไม้ที่มีพื้นฐานเป็นต้นไม้ตัดสินใจ ซึ่งจะรวมการสุ่มต้นไม้ให้อยู่ในรูปแบบต้นไม้อย่างง่าย แล้วจึงสกัดเส้นทางทำนายจากต้นไม้อย่างง่ายนั้น [12] และนำวิธีการเดียวกันนี้ไปประยุกต์ใช้กับตัวแบบ CatBoost ซึ่งมีพื้นฐานเป็นต้นไม้ตัดสินใจเหมือนกัน

อย่างไรก็ตาม ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมนั้นมีข้อจำกัดสำหรับข้อมูลขนาดเล็ก เนื่องจากเป็นการศึกษาอิทธิพลสุ่มของแต่ละบุคคล ซึ่งแถวข้อมูลที่ถูกเก็บมาจากบุคคลคนเดียวกันนั้นมีความสัมพันธ์กันและไม่เป็นอิสระต่อกัน ดังนั้นหากมีจำนวนบุคคลมาก แต่จำนวนแถวข้อมูลของแต่ละบุคคลมีน้อยเกินไป ก็จะไม่สามารถกำหนดตัวแปรอิทธิพลสุ่มหลายตัวแปรได้ อีกทั้ง ในแง่ของต้นทุนและเวลาในการคำนวณ ตัวแบบที่พิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน โดยเฉพาะการสุ่มป่าไม้ จะใช้เวลาในการเรียนรู้นานขึ้นด้วย

นอกจากนี้ งานวิจัยนี้พบว่าการวิเคราะห์ด้วยตัวแบบที่พิจารณาอิทธิพลผสมไม่ได้ให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่เสมอไป แม้ว่าจะมีข้อมูลเวลาที่ติดตามเพิ่มขึ้น ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมก็ไม่ได้ให้ประสิทธิภาพการพยากรณ์ที่มีแนวโน้มคงที่หรือเพิ่มขึ้นเสมอไปเช่นกัน ผลลัพธ์ที่ได้จากงานนี้จึงแตกต่างจาก

ผลลัพธ์ของงานวิจัยของ Che Ngufor และคณะที่พบว่าตัวแบบที่พิจารณาอิทธิพลผสมให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าตัวแบบที่พิจารณาเพียงอิทธิพลคงที่ และประสิทธิภาพการพยากรณ์มีแนวโน้มคงที่หรือเพิ่มขึ้นเมื่อจำนวนครั้งที่ติดตามที่เพิ่มขึ้น [12] โดยผู้วิจัยคาดว่า ลักษณะของข้อมูลที่แตกต่างกันเป็นปัจจัยสำคัญหนึ่งที่ทำให้ได้ผลลัพธ์แตกต่างกัน สำหรับข้อมูลการคัดกรองโรคเบาหวานเป็นข้อมูลจริงที่เกิดขึ้นในประเทศไทย โดยการวิเคราะห์ข้อมูลชุดนี้มีความท้าทายหลายประการ ไม่ว่าจะเป็น ข้อมูลสูญหาย ข้อมูลผิดปกติ รวมถึงเป็นข้อมูลที่ขาดความสมดุลสูง

อนึ่ง ในปัจจุบันการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องมักจะถูกวิเคราะห์โดยมองข้ามความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง หรือพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันด้วยตัวแบบอิทธิพลคงที่เท่านั้น อย่างไรก็ตาม งานวิจัยนี้พบว่า แม้ว่าการเลือกวิเคราะห์ด้วยตัวแบบอิทธิพลผสมโดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันเพื่อให้เหมาะสมกับลักษณะของข้อมูลตามยาวจะไม่ได้ให้ประสิทธิภาพการพยากรณ์ที่ดีขึ้นเสมอไป แต่การใช้ตัวแบบอิทธิพลผสมร่วมกับเทคนิคการเรียนรู้ของเครื่องก็เป็นอีกทางเลือกหนึ่งที่ไม่ควรมองข้ามในการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องด้วย เพราะอาจจะสามารถเพิ่มประสิทธิภาพการพยากรณ์ในบางตัววัดผลได้ ทั้งนี้ ขึ้นกับวัตถุประสงค์หรือตัววัดผลที่ให้ความสำคัญในการวิเคราะห์นั้น ๆ ด้วย

7. References

[1] Wang, P., Li, Y. and Reddy, C.K., 2019, Machine learning for survival analysis: A survey, ACM Computing Surveys (CSUR), 51(6): 136-.

[2] Suresh, K., Severn, C. and Ghosh, D., 2022, Survival prediction models: anintroduction to discrete-time modeling, BMC Medical Research Methodology, 22(1): 207.

[3] Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM, 55(10), pp.78-87.

[4] Kattan, M.W., 2003, Comparison of Cox regression with other methods for determining prediction models and nomograms, The Journal of urology, 170(6): S6-S10.

[5] Breiman, L., 2001, Random forests, Machine learning, 45: 532-.

[6] Cestnik, B., 1990, Estimating Probabilities: A Crucial Task in Machine Learning, ECAI: 147149-.

[7] Micci-Barreca, D., 2001, A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems, ACM SIGKDD Explorations Newsletter, 3(1): 2732-.

[8] Dorogush, A.V., Ershov, V. and Gulin, A., 2018, CatBoost: gradient boosting with categorical features support, arXiv preprint arXiv:1810.11363.

[9] Sarakarn, P. and Jumparway, D., 2020, Coverage and flexibility: issues should be considered for analyzing by generalized linear model in health science research, J Health Sci Comm Publ Health, 3(2): 144158-. (in Thai)

[10] Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. and White, J.S.S., 2009, Generalized linear mixed models: a practical guide for ecology and evolution, Trends in ecology & evolution, 24(3): 127135-.

[11] Breslow, N.E. and Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American statistical Association, 88(421), pp.925-.

[12] Ngufor, C., Van Houten, H., Caffo, B.S., Shah, N.D. and McCoy, R.G., 2019, Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c, Journal of biomedical informatics, 89: 5667-.

[13] Google for developers, Machine Learning Glossary, Available Source: https://developers.google.com machine-learning/glossary, February 21, 2023.