ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมสำหรับการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่อง

น.ส.มนัสพร ตรีรุ่งโรจน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิทยาศาสตรมหาบัณฑิต
สาขาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

MIXED EFFECT MACHINE LEARNING MODEL FOR DISCRETE-TIME SURVIVAL ANALYSIS

.

MissManusaporn Treerungroj

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Statistics

Department of Statistics

FACULTY OF COMMERCE AND ACCOUNTANCY

Chulalongkorn University

Academic Year 2022

หัวข้อวิทยานิพนธ์	ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมสำหรับการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่อง
โดย	น.ส.มนัสพร ตรีรุ่งโรจน์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาหลัก	รองศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี
(ศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... อาจารย์ที่ปรึกษา
(รองศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... กรรมการภายนอก
(รองศาสตราจารย์ ดร.สันติ ถิรพัฒน์)

คำสำคัญ: การวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่อง,ข้อมูลที่ตรวจตัด,ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ,ตัวแบบอิทธิพลผสม

น.ส.มนัสพร ตรีรุ่งโรจน์ : ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมสำหรับการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่อง (MIXED EFFECT MACHINE LEARNING MODEL FOR DISCRETE-TIME SURVIVAL ANALYSIS) อ.ที่ปรึกษาหลัก: รองศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์

การวิเคราะห์การรอดชีพไม่ต่อเนื่องจะศึกษาบนข้อมูลตามยาวซึ่งชุดข้อมูลตามยาวมักถูกจัดเก็บเป็นตารางโดยข้อมูลแต่ละแถวแสดงถึงการจัดเก็บข้อมูลของบุคคลหนึ่ง ณ เวลาหนึ่งๆ ดังนั้น ข้อมูลจากบุคคลเดียวกันจึงประกอบไปด้วยข้อมูลหลายแถวซึ่งมีความสัมพันธ์กัน การใช้อัลกอริทึมการเรียนรู้ของเครื่องสำหรับการวิเคราะห์ชุดข้อมูลดังกล่าวมักมองข้ามความสัมพันธ์ของข้อมูลที่เกิดจากคนเดียวกัน แต่จะสมมติว่าข้อมูลแต่ละแถวเป็นอิสระต่อกัน งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการวิเคราะห์การรอดชีพไม่ต่อเนื่องโดยเปรียบเทียบผลลัพธ์จากการพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน โดยใช้ตัวแบบการสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม ที่พิจารณาเฉพาะอิทธิพลคงที่ และตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมที่พิจารณาทั้งอิทธิพลคงที่และอิทธิพลสุ่ม เพื่อพยากรณ์การเกิดเหตุการณ์บนข้อมูลการรอดชีพ 2 ชุด คือ ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ และข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานของประชากรไทย ซึ่งเป็นข้อมูลที่ขาดความสมดุลสูง ผลการศึกษาพบว่าสำหรับตัวแบบอิทธิพลคงที่ การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ประสิทธิภาพการพยากรณ์ที่ดีขึ้นเฉพาะเมื่อใช้ตัวแบบ CatBoost ในขณะที่ตัวแบบอิทธิพลผสมไม่ได้ให้ประสิทธิภาพการพยากรณ์ที่ดีขึ้นเสมอไปเมื่อเทียบกับตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่ โดยสรุป งานวิจัยนี้ได้แสดงให้เห็นว่าการพิจารณาความสัมพันธ์ของข้อมูลไม่ได้ส่งผลให้ประสิทธิภาพการพยากรณ์ดีขึ้นเสมอไป ทั้งบนตัวแบบอิทธิพลคงที่และตัวแบบอิทธิพลผสม ขึ้นอยู่ข้อจำกัดและปัจจัยต่างๆ เช่น ลักษณะข้อมูล ตัวแบบ การกำหนดตัวแปรอิทธิพลสุ่ม และวิธีการสกัดอิทธิพลคงที่จากตัวแบบ อย่างไรก็ตาม การใช้ตัวแบบอิทธิพลผสมร่วมกับการเรียนรู้ของเครื่องเป็นอีกหนึ่งวิธีการที่น่าลอง และสามารถทำให้ประสิทธิภาพการทำงานดีขึ้นจากการใช้เทคนิคการเรียนรู้ของเครื่องเพียงอย่างเดียว

สาขาวิชา สถิติ ลายมือชื่อนิสิต.....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก.....

ลายมือชื่อ อ.ที่ปรึกษาร่วม.....

6480472626 : MAJOR STATISTICS

KEYWORDS: Discrete-time Survival Analysis,Censored Data,Binary Classification Machine Learning Models,Mixed Effect Model

MissManusaporn Treerungroj : MIXED EFFECT MACHINE LEARNING MODEL FOR DISCRETE-TIME SURVIVAL ANALYSIS Advisor: Assoc. Prof.VITARA PUNGPAPONG, Ph.D.

The discrete-time survival analysis is a study of longitudinal data in which the data is typically organized as a table which each row represents a record of a person at a given time point. In other words, the data obtained from the same person consists of several rows in the table and they are dependent. Machine learning algorithms can be used to analyze those datasets. However, they typically ignore the dependency among records from the same person and assume independence among them instead. The purpose of this study is to compare prediction performance of methods with and without considering the relationships between data from the same individuals. Compared methods include fixed effect models, Random Forest, CatBoost, Artificial Neural Network, and a mixed effect machine learning model which considers both fixed and random effects. Here, we applied the aforementioned methods to predict event status from 2 datasets, Mayo Clinic primary biliary cholangitis dataset and diabetes screening dataset collected from Thai population. Our results show that, for the fixed effect model, considering the relationships between data from the same individuals resulted in improved prediction performance only when using CatBoost. While the mixed effect model does not result in improved prediction performance compared to the fixed effect model. In summary, this research shows that considering the relationships between data does not always lead to improved prediction performance, and depends on various limitations and factors such as data characteristics, model selection, random effect variables, and methods of fixed effect component extraction. However, using a mixed-effect model along with machine learning is worth trying and could improve predictive performance than using machine learning techniques alone.

Field of Study	Statistics	Student's Signature.....
Academic Year	2022	Advisor's Signature.....
		Co-advisor's Signature.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงด้วยดี ด้วยความกรุณาและความอนุเคราะห์เป็นอย่างดีจากคณาจารย์และผู้เกี่ยวข้องทุกท่าน โดยเฉพาะอย่างยิ่ง รองศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ผู้กรุณาเสียสละเวลาให้คำแนะนำปรึกษาและความช่วยเหลืออย่างสม่ำเสมอ ตลอดจนให้ความช่วยเหลือเพื่อแก้ไขปรับปรุงข้อบกพร่องต่างๆ อย่างเอาใจใส่มาโดยตลอด อีกทั้งยังช่วยส่งเสริมให้กำลังใจในการทำงานเป็นอย่างดีจนกระทั่งวิทยานิพนธ์ฉบับนี้เสร็จลุล่วง ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณท่าน ศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์ ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร. อัครินทร์ ไพบูลย์พานิช และรองศาสตราจารย์ ดร. สันติ ถิรพัฒน์ กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านได้กรุณาเสียสละเวลาและให้เกียรติเป็นกรรมการสอบครั้งนี้ ตลอดจนช่วยตรวจสอบและให้คำแนะนำเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์มากยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์ประจำวิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ได้ให้โอกาสทางการศึกษา ทำให้ผู้วิจัยสามารถนำความรู้ที่ได้รับมาประยุกต์ใช้ในการทำวิทยานิพนธ์ครั้งนี้ และขอกราบขอบพระคุณบุคลากรทุกท่านที่ได้อำนวยความสะดวกด้านการจัดการเอกสารและการประสานงานต่างๆ

ขอกราบขอบพระคุณสำนักงานหลักประกันสุขภาพแห่งชาติที่เอื้อเฟื้อข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานให้เป็นข้อมูลสำคัญในวิทยานิพนธ์ฉบับนี้

สุดท้ายนี้ ขอกราบขอบพระคุณบิดา มารดา พี่ชาย และนายทศพร บรรเจิดกิจ ผู้ให้ความห่วงใยและสนับสนุนเป็นกำลังใจให้ผู้วิจัยอย่างดีมาโดยตลอด ขอขอบคุณเพื่อนๆ ที่สนับสนุนช่วยเหลือ และให้กำลังใจกันและกัน โดยเฉพาะนางสาวศศิวิมล ศรีโรจน์ ผู้ให้การสนับสนุนคอมพิวเตอร์ระบบปฏิบัติการวินโดวส์ที่จำเป็นสำหรับการจัดการเอกสาร ทำให้ผู้วิจัยสำเร็จการศึกษาไปด้วยดี

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

การวิเคราะห์การรอดชีพ (Survival Analysis) คือวิธีการวิเคราะห์ข้อมูลและสร้างแบบจำลองทางสถิติบนข้อมูลระยะเวลาของการเกิดหรือไม่เกิดเหตุการณ์ ซึ่งถูกนำมาใช้อย่างหลากหลายในการวิเคราะห์ข้อมูลทางสุขภาพทางการแพทย์ เช่น ระยะเวลาที่ผู้ป่วยจะเป็นโรคเบาหวาน ระยะเวลาที่ผู้ป่วยจะเสียชีวิตด้วยโรคมะเร็ง ระยะเวลาที่อุปกรณ์ของเครื่องจักรจะหมดอายุการใช้งาน ซึ่งข้อมูลการรอดชีพจะเป็นลักษณะข้อมูลตามยาว (Longitudinal Data) ที่จะสังเกตและเก็บข้อมูลของตัวอย่างเรื่อยๆ เช่น บุคคลที่มีความเสี่ยงจะเป็นโรคเบาหวานจะมาตรวจคัดกรองการเป็นโรคเบาหวานอยู่เป็นประจำ ในการตรวจแต่ละครั้งจะเก็บข้อมูลสุขภาพ ได้แก่ อายุ น้ำหนัก ส่วนสูง ความดันโลหิต วิธีการตรวจคัดกรองโรคเบาหวาน ผลการตรวจระดับน้ำตาลในเลือด และข้อมูลประกอบอื่นๆ เช่น ผลการวินิจฉัยโดยแพทย์ และยาที่ได้รับในครั้งนั้น ซึ่งข้อมูลเหล่านี้จะสามารถนำไปเป็นตัวแปรร่วมในการวิเคราะห์และพยากรณ์การเกิดโรคเบาหวานได้

ในการวิเคราะห์การรอดชีพจะแบ่งเป็นการพยากรณ์การรอดชีพแบบเวลาต่อเนื่อง (Continuous-time Survival Prediction) คือการพยากรณ์ระยะเวลาจนกว่าจะเกิดเหตุการณ์ที่สนใจ ตัวอย่างอัลกอริทึมประเภทนี้ที่เป็นที่นิยม เช่น วิธีการ Kaplan-Meier (KM Method) และตัวแบบ Cox Proportional-hazards และอีกประเภทหนึ่งคือการพยากรณ์การรอดชีพแบบเวลาไม่ต่อเนื่อง (Discrete-time Survival Prediction) คือการแบ่งค่าเวลาการรอดชีพต่อเนื่องออกช่วงๆ และพยากรณ์ความเสี่ยงต่อการไม่รอดชีพ (Hazard) ในแต่ละช่วงเวลา ซึ่งจะเป็นความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจในช่วงเวลานั้น วิธีนี้ทำให้สามารถประยุกต์ใช้อัลกอริทึมการเรียนรู้ของเครื่องใดๆ ที่สามารถจำแนกประเภททวิมาพยากรณ์ได้ ตัวอย่างเช่น ต้นไม้การรอดชีพ (Survival Trees) วิธีการแบบเบย์ (Bayesian Methods) โครงข่ายประสาท (Neural Networks) เครื่องเวกเตอร์ค้ำยัน (Support Vector Machines) Gradient Boosted Machine และการเรียนรู้ของเครื่องขั้นสูงอื่นๆ

ในการเตรียมข้อมูลตามยาวสำหรับการวิเคราะห์การรอดชีพมักจะจัดเก็บข้อมูลในการตรวจแต่ละครั้งเป็นหนึ่งแถวในไฟล์ข้อมูล หรือก็คือข้อมูลจากผู้ป่วยคนเดียวกัน หากมาตรวจหลายครั้งก็จะมีข้อมูลหลายแถวสำหรับผู้ป่วยหนึ่งคน อย่างไรก็ตาม อัลกอริทึมการเรียนรู้ของเครื่องส่วนมากจะสมมติว่าข้อมูลแต่ละแถวนั้นมีการแจกแจงเหมือนกันและเป็นอิสระต่อกัน (Independent

Identically Distributed) แต่การสมมตินี้มักจะถูกละเมิดในการประยุกต์ใช้งานกับข้อมูลจริง ตัวอย่างเช่น ข้อมูลการสังเกตอาการของผู้ป่วย ซึ่งผู้ป่วยแต่ละคนจะมีการติดตามอาการและเก็บข้อมูลเรื่อยๆ ทำให้ข้อมูลแถวที่มาจากผู้ป่วยคนเดียวกันนั้นมีความสัมพันธ์กัน ดังนั้น การวิเคราะห์ข้อมูลของแต่ละบุคคลและพยากรณ์ว่าจะเกิดเหตุการณ์ที่สนใจหรือไม่ โดยเรียนรู้จากลักษณะประวัติอาการครั้งก่อนๆ ของบุคคลนั้นด้วยมีความสมเหตุสมผลในเชิงการนำเอาผลลัพธ์ไปประยุกต์ใช้จริง เช่น การพยากรณ์เหตุการณ์ที่อาจจะเกิดขึ้นของผู้ป่วยแบบเฉพาะบุคคล นำไปสู่การประมาณค่าใช้จ่ายและวางแผนล่วงหน้าได้

งานวิจัยนี้ต้องการศึกษาการวิเคราะห์การรอดชีพแบบเวลาไม่ต่อเนื่องโดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน และเปรียบเทียบผลลัพธ์กับตัวแบบที่ละเลยความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง ทั้งนี้ ในงานวิจัยนี้จะศึกษาผ่านข้อมูลการรอดชีพแบบไม่ต่อเนื่อง ได้แก่ ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ และข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานของประชากรไทยซึ่งเป็นข้อมูลจริง เพื่อศึกษาว่าผลลัพธ์ที่ได้สอดคล้องกันหรือไม่

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษาการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่อง (Discrete-time Survival Analysis) โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันในข้อมูลตามยาว

1.3 สมมติฐานการวิจัย

การใช้ตัวแบบอิทธิพลคงที่ (Fixed Effect Model) ด้วยตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ (Binary Classification Machine Learning Models) เมื่อเปรียบเทียบกับตัวแบบการเรียนรู้ของตัวแบบอิทธิพลผสม (Mixed Effect Model) ให้ผลลัพธ์การพยากรณ์ความเสี่ยงต่อการไม่รอดชีพที่แตกต่างกัน

1.4 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

งานวิจัยนี้จะทำให้ทราบถึงการพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลเดียวกันในข้อมูลตามยาวในการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนตัวแบบประเภทต่างๆ

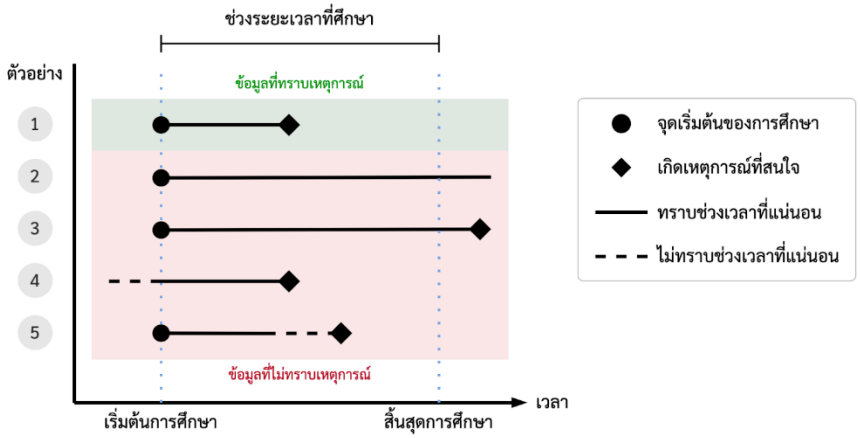
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดพื้นฐานเกี่ยวกับการวิเคราะห์การรอดชีพ (Introduction to Survival Analysis)

การวิเคราะห์การรอดชีพเป็นการวิเคราะห์เชิงสถิติบนข้อมูลที่ประกอบไปด้วยประเภทของเหตุการณ์ที่สนใจ (Event) และระยะเวลาจากจุดเริ่มต้นจนถึงจุดที่เกิดเหตุการณ์ที่สนใจ (Time to Event หรือ Survival Time) ซึ่งเป็นการศึกษาในช่วงระยะเวลาหนึ่ง เช่น การวิเคราะห์ข้อมูลของผู้ป่วยโรคเบาหวาน ซึ่งผู้ป่วยคนเดิมจะมาตรวจและเก็บข้อมูลอย่างสม่ำเสมอ ในที่นี้ เหตุการณ์ที่สนใจอาจเป็นการเสียชีวิตจากโรคเบาหวาน และระยะเวลาจากจุดเริ่มต้นจนถึงจุดเสียชีวิตจากโรคเบาหวาน ซึ่งบางกรณีจะไม่สามารถสังเกตเหตุการณ์ที่สนใจได้ เช่น สิ้นสุดช่วงระยะเวลาที่ศึกษาโดยที่ผู้ป่วยไม่เสียชีวิตจากโรคเบาหวาน ผู้ป่วยขาดการติดตามระหว่างระยะเวลาศึกษา ผู้ป่วยถอนตัวจากการติดตาม หรือผู้ป่วยเกิดเหตุการณ์อื่นทำให้ไม่สามารถเข้าร่วมการติดตามได้จนจบระยะเวลาศึกษา จึงทำให้ไม่สามารถสรุปได้ว่าท้ายสุดแล้วเกิดเหตุการณ์ที่สนใจหรือไม่ ข้อมูลลักษณะนี้เรียกว่าข้อมูลตรวจตัด (Censored Data)

ข้อมูลที่ไม่ทราบเวลาการเกิดเหตุการณ์จริงนี้แบ่งออกได้เป็น 3 ประเภทตามลักษณะการเกิด (Wang et al., 2019) ได้แก่ ข้อมูลที่ตรวจตัดด้านขวา (Right-censoring Data) คือข้อมูลที่เวลาการรอดชีพที่สังเกตได้น้อยกว่าหรือเท่ากับเวลาการรอดชีพจริง ข้อมูลตรวจตัดด้านซ้าย (Left-censoring Data) คือข้อมูลที่เวลาการรอดชีพที่สังเกตได้มากกว่าหรือเท่ากับเวลาการรอดชีพจริง และข้อมูลตรวจตัดแบบช่วง (Interval-censoring Data) คือข้อมูลที่ถูกตรวจตัดทั้งด้านขวาและด้านซ้าย จึงทราบเพียงว่าเกิดเหตุการณ์ที่สนใจในระหว่างระยะเวลาที่ศึกษา แต่ไม่ทราบช่วงเวลาเกิดที่แน่นอน

จากลักษณะของข้อมูล 3 ประเภทนี้ ข้อมูลที่ไม่ทราบเหตุการณ์ทางขวาเป็นข้อมูลที่พบเจอได้บ่อยที่สุด (Marubini & Valsecchi, 2004) ในการวิเคราะห์การรอดชีพ



รูปที่ 2.1 ระยะเวลาจากจุดเริ่มต้นจนถึงจุดที่เกิดเหตุการณ์ที่สนใจ และรูปแบบของข้อมูลตรวจตัด

เพื่อให้เข้าใจลักษณะข้อมูลตรวจตัดที่ใช้ในการวิเคราะห์การรอดชีพมากขึ้น รูปที่ 2.1 แสดงการเก็บข้อมูลการเกิดเหตุการณ์ที่สนใจจาก 5 ตัวอย่างในช่วงระยะเวลาหนึ่ง สามารถอธิบายดังนี้

ตัวอย่างคนที่ 1 เกิดเหตุการณ์ที่สนใจในระหว่างช่วงระยะเวลาที่ศึกษา เป็นข้อมูลที่ทราบจุดเวลาในการเกิดเหตุการณ์ที่สนใจ ตัวอย่างคนที่ 2 และ 3 ไม่เกิดเหตุการณ์ที่สนใจในระหว่างช่วงระยะเวลาที่ศึกษา แต่อาจเกิดหรือไม่เกิดเหตุการณ์ที่สนใจหลังจากสิ้นสุดช่วงระยะเวลาที่ศึกษา เรียกข้อมูลลักษณะนี้ว่าข้อมูลที่ไม่ทราบเหตุการณ์ทางขวา ตัวอย่างคนที่ 4 มีจุดเริ่มต้นก่อนช่วงระยะเวลาที่ศึกษาและไม่ทราบช่วงเวลาที่แน่นอน เรียกข้อมูลลักษณะนี้ว่าข้อมูลที่ไม่ทราบเหตุการณ์ทางซ้าย และตัวอย่างคนที่ 5 ทราบเพียงว่าเกิดเหตุการณ์ที่สนใจในระหว่างระยะเวลาที่ศึกษา แต่ไม่ทราบช่วงเวลาเกิดที่แน่นอน เรียกข้อมูลลักษณะนี้ว่าข้อมูลที่ไม่ทราบเวลาในการเกิดเหตุการณ์ระหว่างศึกษา

ในการวิเคราะห์การรอดชีพ การไม่นำเอาข้อมูลตรวจตัดเหล่านี้ไปวิเคราะห์ด้วยอาจทำให้การวิเคราะห์และพยากรณ์เกิดความเอนเอียงและไม่มีประสิทธิภาพ (Kattan, 2003) จึงทำให้ตัวแบบที่ใช้พยากรณ์การรอดชีพแตกต่างจากตัวแบบที่ใช้พยากรณ์ปกติทั่วไป

ในการวิเคราะห์การรอดชีพ สามารถแสดงฟังก์ชันการรอดชีพได้ดังนี้

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

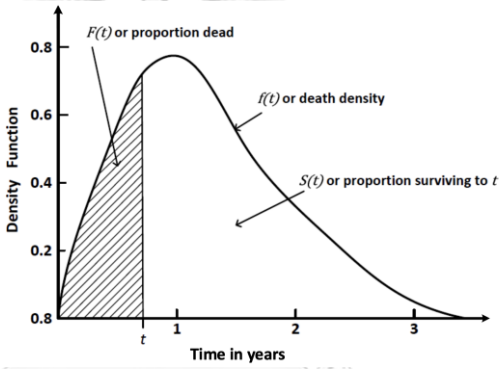
เมื่อ T คือเวลาการรอดชีพ และ $S(t)$ คือความน่าจะเป็นที่จะรอดชีพ ณ เวลา t โดย $S(t)$ จะมีค่า 1 เมื่อ $t = 0$ และ $S(t)$ จะลดลงแบบโมโนโทนเมื่อ t เพิ่มขึ้น ในขณะที่ฟังก์ชันการแจกแจงความน่าจะเป็นแบบสะสมสามารถแสดงได้ดังนี้

$$F(t) = 1 - S(t)$$

เมื่อ $F(t)$ คือความน่าจะเป็นที่เหตุการณ์ที่สนใจจะเกิดก่อนเวลา t และฟังก์ชันความหนาแน่นการเสียชีวิตสามารถแสดงได้ดังนี้

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

รูปที่ 2.2 สามารถแสดงความสัมพันธ์ระหว่าง $f(t)$, $F(t)$ และ $S(t)$



รูปที่ 2.2 ความสัมพันธ์ระหว่าง $f(t)$, $F(t)$ และ $S(t)$ (Wang et al., 2019)

นอกจากนี้ อีกฟังก์ชันหนึ่งที่มักถูกใช้ในการวิเคราะห์การรอดชีพคือฟังก์ชันพิบัติ (Hazard Function) สามารถแสดงได้ดังนี้

$$h(t) = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{S(t)} / \Delta t = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t)$$

เมื่อ $h(t)$ เป็นฟังก์ชันพิบัติ ซึ่งคือความเสี่ยงต่อการไม่รอดชีพที่แสดงอัตราที่จะเกิดเหตุการณ์ที่สนใจ ณ เวลา t เมื่อเหตุการณ์ที่สนใจนั้นยังไม่เคยเกิดขึ้นก่อนเวลา t โดย $h(t)$ ควรจะมีค่าไม่ติดลบเสมอ หรือ $h(t) \geq 0$ และสามารถแสดงฟังก์ชันพิบัติแบบสะสมได้ดังนี้

$$H(t) = \int_0^t h(s)ds = -\log S(t)$$

จากฟังก์ชันพิบัติแบบสะสม สามารถเขียนฟังก์ชันการรอดชีพให้อยู่ในรูปดังนี้ได้

$$S(t) = exp\big(-H(t)\big)$$

2.2 ตัวแบบการพยากรณ์การรอดชีพเวลาไม่ต่อเนื่อง (Discrete-time Survival Prediction Model)

ในปัจจุบันตัวแบบการพยากรณ์การรอดชีพถูกพัฒนามาอย่างหลากหลาย โดยเริ่มต้นจากแนวทางการวิเคราะห์การถดถอยแบบดั้งเดิมที่จะสนใจผลลัพธ์เป็นค่าเวลาการรอดชีพต่อเนื่อง คือ ระยะเวลาจากจุดเริ่มต้นจนถึงจุดที่เกิดเหตุการณ์ที่สนใจ ซึ่งวิธีการประมาณฟังก์ชันการรอดชีพแบบดั้งเดิมแบ่งออกเป็น 3 ประเภทตามวิธีการประมาณการกระจายตัวของข้อมูลเวลาการรอดชีพ ได้แก่ ข้อสมมติแบบไม่อิงพารามิเตอร์ (Non-parametric Assumption) ข้อสมมติแบบกึ่งการอิงพารามิเตอร์ (Semi-parametric Assumption) และข้อสมมติแบบอิงพารามิเตอร์ (Parametric Assumption)

ตัวแบบการพยากรณ์การรอดชีพแบบไม่อิงพารามิเตอร์จะไม่มีการตั้งข้อสมมติเกี่ยวกับ การกระจายตัวของข้อมูลเวลาการรอดชีพ และไม่สมมติว่าตัวแปรร่วมที่ศึกษามีความสัมพันธ์กับค่าเวลาการรอดชีพ ซึ่งหากไม่ทราบการกระจายตัวของข้อมูลเวลาการรอดชีพที่เหมาะสม วิธีการแบบไม่อิงพารามิเตอร์จะมีประสิทธิภาพมากกว่าวิธีการแบบอื่น แต่ข้อเสียของวิธีการนี้คือ ยากต่อการตีความผลลัพธ์ที่ได้ และอาจให้การประมาณที่ไม่แม่นยำ (Wang et al., 2019) ตัวอย่างตัวแบบประเภทนี้ เช่น วิธีการ Kaplan-Meier (KM Method) วิธีการ Nelson-Aalen (NA Method) และวิธีการ Life-table (LT Method)

ตัวแบบการพยากรณ์การรอดชีพแบบกึ่งการอิงพารามิเตอร์จะไม่จำเป็นต้องตั้งข้อสมมติเกี่ยวกับการกระจายตัวของข้อมูลเวลาการรอดชีพ โดยใช้วิธีประมาณการกระจายตัวของข้อมูล เวลาการรอดชีพด้วยการตั้งข้อสมมติแบบไม่อิงพารามิเตอร์ วิธีนี้แม้ว่าจะไม่จำเป็นต้องกำหนดค่าพิบัติพื้นฐาน (Baseline Hazard) หรือค่าการรอดชีพพื้นฐาน (Baseline Survival) แต่ยังคงต้องกำหนดค่าพารามิเตอร์การถดถอยสำหรับแต่ละตัวแปรร่วม จึงไม่เป็นวิธีการแบบอิงพารามิเตอร์ที่สมบูรณ์ วิธีการแบบกึ่งการอิงพารามิเตอร์นี้จะมีตัวประมาณที่คงเส้นคงวามากกว่าวิธีการแบบอิงพารามิเตอร์ และแม่นยำมากกว่าวิธีการแบบไม่อิงพารามิเตอร์ แต่ข้อเสียของวิธีการนี้คือจะไม่ทราบการกระจายตัวของผลลัพธ์ที่ได้ แต่การตีความผลลัพธ์ที่ได้ทำได้ไม่ง่าย ตัวอย่างตัวแบบประเภทนี้ เช่น Cox Proportional-hazards Regression

ตัวแบบการพยากรณ์การรอดชีพแบบอิงพารามิเตอร์จะตั้งข้อสมมติเกี่ยวกับการกระจายตัวของข้อมูลเวลาการรอดชีพจากความรู้เดิมในบริบททางคลินิกหรือวิทยาศาสตร์ ตัวแบบพยากรณ์นั้น

จะมีประสิทธิภาพและแม่นยำหากการกระจายตัวของข้อมูลเวลาการรอดชีพเป็นไปตามที่ตั้งสมมติฐานไว้ แต่ในขณะเดียวกัน หากตั้งสมมติฐานกระจายตัวของข้อมูลผิดพลาดก็อาจนำไปสู่การประมาณที่เอนเอียงได้ (Kleinbaum & Klein, 1996) ตัวอย่างตัวแบบประเภทนี้ เช่น การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) เวลาความล้มเหลวแบบเร่ง (Accelerated Failure Time)

แนวทางการวิเคราะห์การถดถอยแบบดั้งเดิมนั้นจะเน้นการอธิบายลักษณะการกระจายตัวของข้อมูลเหตุการณ์ที่สนใจและคุณสมบัติทางสถิติ การตีความผลลัพธ์ที่ได้ และการเข้าใจ อิทธิพลของตัวแปรร่วมแต่ละตัว ในขณะที่ปัจจุบันยังมีอัลกอริทึมการเรียนรู้ของเครื่องที่สามารถนำมาประยุกต์และพยากรณ์ความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจได้ ซึ่งจะเน้นไปที่ประสิทธิภาพการพยากรณ์และสามารถรองรับข้อมูลที่มีจำนวนมิติมากได้ดีกว่า ตัวอย่างอัลกอริทึมการเรียนรู้ของเครื่อง เช่น ต้นไม้การรอดชีพ วิธีการแบบเบย์ (Baysian Methods) โครงข่ายประสาท เครื่องเวกเตอร์ค้ำยัน และการเรียนรู้ของเครื่องขั้นสูงอื่นๆ

วิธีการที่กล่าวมาสามารถใช้ได้กับการวิเคราะห์ผลลัพธ์ที่เป็นค่าเวลาการรอดชีพต่อเนื่อง ในการวิเคราะห์ผลลัพธ์ที่เป็นค่าเวลาการรอดชีพไม่ต่อเนื่อง (Suresh et al., 2022) ข้อมูลจะยังคงเหมือนเดิม แต่มีการกำหนดฟังก์ชันความเสี่ยงต่อการไม่รอดชีพและการเชื่อมต่อระหว่างฟังก์ชันความเสี่ยงต่อการไม่รอดชีพกับฟังก์ชันการรอดชีพที่แตกต่างออกไป โดยจะแบ่งค่าเวลาการรอดชีพต่อเนื่องออกเป็น J ช่วงเวลา $(t_0, t_1], (t_1, t_2], \dots, (t_{j-1}, t_j]$ โดยที่ $t_0 = 0$ ในกรณีนี้ ความเสี่ยงต่อการไม่รอดชีพ (Hazard) ในแต่ละช่วงเวลาจะเป็นความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจในช่วงเวลานั้น โดยที่ยังคงรอดชีพ ณ เวลาเริ่มต้นของช่วงเวลานั้น

ดังนั้นในการวิเคราะห์ผลลัพธ์ที่เป็นค่าเวลาการรอดชีพไม่ต่อเนื่อง จึงมองว่าความเสี่ยงต่อการไม่รอดชีพนั้นเป็นความน่าจะเป็นแบบมีเงื่อนไข และมีค่าอยู่ระหว่าง 0 ถึง 1 ความเสี่ยงต่อการไม่รอดชีพจากตัวแปรร่วม X_i ในช่วงเวลา $A_j = (t_{j-1}, t_j]$ สามารถแสดงได้ดังนี้

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

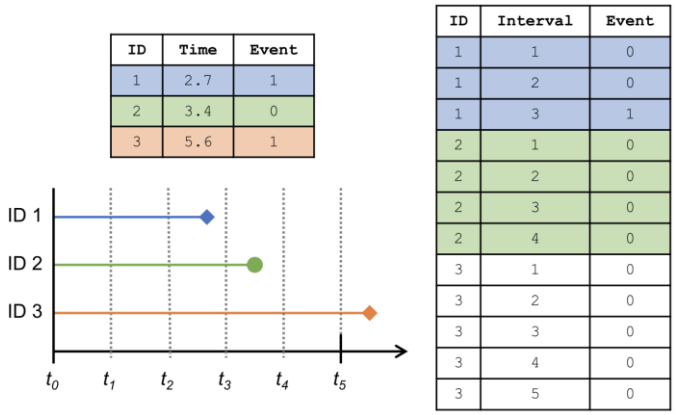
และฟังก์ชันความน่าจะเป็นแบบไม่ต่อเนื่อง (Discrete Probability Function) สามารถแสดงได้ดังนี้

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

โดยภาวะน่าจะเป็นของการรอดชีพ (Survival Likelihood) ของฟังก์ชันความน่าจะเป็นแบบไม่ต่อเนื่องจะสอดคล้องกับภาวะความน่าจะเป็นของตัวแบบแบบทวินามที่มีสมมติฐานว่าตัวบ่งชี้เหตุการณ์เป็นอิสระต่อกัน ภาวะน่าจะเป็นของการรอดชีพสามารถแสดงได้ดังนี้

$$L = \prod_{i=1}^{n} \prod_{j=1}^{ji} \lambda_{ij}\,(X_i)^{d_{ij}} \left(1 - \lambda_{ij}(X_i)\right)^{1-d_{ij}}$$

ดังนั้น การวิเคราะห์ผลลัพธ์ที่เป็นค่าเวลาการรอดชีพไม่ต่อเนื่องจึงสามารถประยุกต์ใช้ วิธีการจำแนกแบบทวิเพื่อคำนวณความน่าจะเป็นที่จะเกิดเหตุการณ์แบบทวิได้ แต่จำเป็นจะต้องจัดรูปข้อมูลเพื่อนำเข้าตัวแบบให้เป็นดังรูปที่ 2.3



ID	Time	Event
1	2.7	1
2	3.4	0
3	5.6	1

ID	Interval	Event
1	1	0
1	2	0
1	3	1
2	1	0
2	2	0
2	3	0
2	4	0
3	1	0
3	2	0
3	3	0
3	4	0
3	5	0

รูปที่ 2.3 ตัวอย่างการแปลงชุดข้อมูลการรอดชีพเวลาต่อเนื่อง เป็นชุดข้อมูลของคนตามช่วงเวลา (Suresh et al., 2022)

ลักษณะข้อมูลที่นำเข้าตัวแบบจะถูกแปลงจากข้อมูลการรอดชีพเวลาต่อเนื่องเป็นชุดข้อมูลของคนตามช่วงเวลา โดยแต่ละแถวข้อมูลจะประกอบไปด้วยตัวบ่งชี้ที่บอกถึงการเกิดเหตุการณ์ที่สนใจในช่วงระยะเวลานั้น (d_{ij} หรือ Event) ชุดของตัวแปรร่วม (X_i) และตัวแปรที่ใช้แบ่งกลุ่มว่าเป็นช่วงระยะเวลาใด (A_j หรือ Interval)

ด้วยการกำหนดโครงสร้างแบบทวิของฟังก์ชันภาวะน่าจะเป็นและข้อมูลเวลาการรอดชีพไม่ต่อเนื่องให้อยู่ลักษณะทั่วไปจึงสามารถเลือกใช้วิธีการที่หลากหลาย ไม่ว่าจะเป็นการวิเคราะห์การถดถอยแบบดั้งเดิมไปจนถึงแนวทางอัลกอริทึมการเรียนรู้ของเครื่องที่ซับซ้อนขึ้นเพื่อประมาณค่าพารามิเตอร์ได้

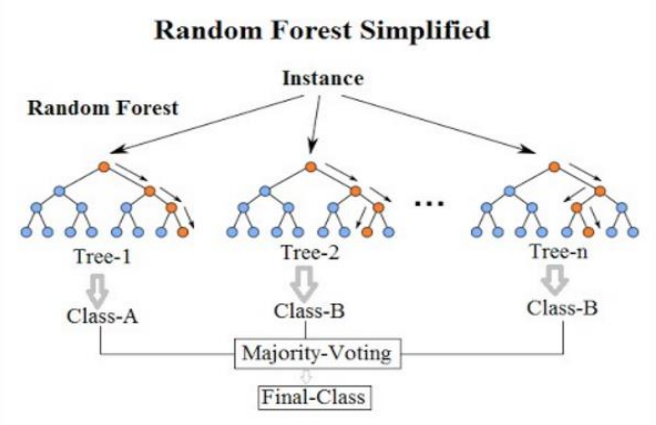
โดยสรุป ข้อดีของแนวทางการพยากรณ์การรอดชีพเวลาไม่ต่อเนื่องคือไม่มีการกำหนด ข้อสมมติการเกิดความพิบัติแบบสัดส่วนสำหรับการกระจายตัวของข้อมูลเวลาการรอดชีพ และผลลัพธ์ที่ได้จากตัวแบบสามารถตีความได้เข้าใจง่ายกว่า เนื่องจากฟังก์ชันพิบัติเองแสดงถึงความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจในช่วงเวลาหนึ่งโดยที่ยังคงรอดชีพ ณ เวลาเริ่มต้นของช่วงเวลานั้น

2.3 ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ (Binary Classification Machine Learning Models)

ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิเป็นวิธีการเรียนรู้โดยมีผู้สอนเพื่อจำแนกข้อมูล 2 กลุ่ม เช่น การจำแนกว่าอีเมลเป็นสแปมหรือไม่ การจำแนกว่าเป็นโรคหรือไม่ การจำแนกว่าเป็นธุรกรรมที่ทุจริตหรือไม่ เป็นต้น ปัจจุบันมีหลายตัวแบบการเรียนรู้ของเครื่องที่ถูกพัฒนาและต่อยอดเป็นตัวแบบที่ซับซ้อน สามารถเรียนรู้และพยากรณ์ได้ดียิ่งขึ้น เช่น

2.3.1 การสุ่มป่าไม้ (Random Forests)

การสุ่มป่าไม้เป็นหนึ่งในวิธีการเรียนรู้แบบกลุ่ม (Ensemble Learning Method) คือเทคนิคที่จะนำผลลัพธ์การพยากรณ์จากหลายตัวแบบมารวมกัน ทำให้ได้ผลลัพธ์ที่เสถียรและใช้ได้ทั่วไปมากขึ้น การสุ่มป่าไม้สามารถใช้วิเคราะห์ได้ทั้งการจำแนกประเภทและการพยากรณ์ค่าตัวเลข กลไกของอัลกอริทึมการสุ่มป่าไม้แสดงได้ดังรูปที่ 2.4 โดยจะสร้างต้นไม้ตัดสินใจ (Decision Trees) หลายๆ ต้น ที่เรียนรู้บนชุดข้อมูลที่เลือกโดยมีการคืนที่หลายๆ ชุดที่มีขนาดเท่ากัน ทำให้ได้ต้นไม้ตัดสินใจที่แตกต่างกันหลายต้น ในกรณีของการจำแนกประเภท ผลลัพธ์ที่ถูกเลือกโดยต้นไม้ตัดสินใจมากที่สุดจะเป็นผลลัพธ์สุดท้าย และในกรณีของการถดถอย ผลลัพธ์สุดท้ายจะเป็นค่าเฉลี่ยของผลลัพธ์จากต้นไม้ตัดสินใจทั้งหมด



รูปที่ 2.4 กลไกของอัลกอริทึมการสุ่มป่าไม้ (Azhari et al., 2019)

วิธีนี้ทำให้ตัวแบบการสุ่มป่าไม้มีความคงทนต่อค่าผิดปกติ ค่ารบกวน และไม่เกิดปัญหาเกินพอดี (Overfitting) (Breiman, 2001)

2.3.2 Categorical Boost (CatBoost)

CatBoost คือตัวแบบที่มีเค้าโครงของ Gradient Boosting พัฒนาต่อยอดมาจากต้นไม้ตัดสินใจ มีจุดเด่นในการจัดการกับข้อมูลที่จัดเป็นกลุ่ม นั่นคือ โดยทั่วไปในการทำ Gradient Boosting จะไม่สามารถนำข้อมูลที่จัดเป็นกลุ่มไปใช้ได้ตรงๆ เทคนิคที่นิยมใช้คือการแปลงข้อมูลที่จัดเป็นกลุ่มให้อยู่ในรูปแบบค่าทวิที่มีค่า 0 หรือ 1 (One-hot Encoding) ก่อนจะนำไปให้ตัวแบบเรียนรู้ได้

อีกวิธีหนึ่งคือการคำนวณค่าทางสถิติโดยใช้ข้อมูลผลเฉลย โดยสมมติให้มีชุดข้อมูล $D = \{(X_i, Y_i)\}_{i=1..n}$ โดยที่ $X_i = \left(X_{i,1}, \dots, X_{i,m}\right)$ คือเวกเตอร์ของ m คุณลักษณะที่มีทั้งข้อมูลตัวเลขและข้อมูลที่จัดเป็นกลุ่ม และ $Y_i \in R$ เป็นค่าผลเฉลย โดยจะทดแทนข้อมูลที่จัดเป็นกลุ่มด้วยค่าผลเฉลยเฉลี่ยของชุดข้อมูลในกลุ่มนั้นทั้งหมด แสดงได้ดังนี้

$$X_{i,k} = \frac{\sum_{j=1}^{n}\left[X_{j,k} = X_{i,k}\right] \cdot Y_j}{\sum_{j=1}^{n}\left[X_{j,k} = X_{i,k}\right]}$$

โดย $[\cdot]$ จะมีค่าเป็น 1 เมื่อ $X_{j,k} = X_{i,k}$ นอกจากนั้นจะมีค่าเป็น 0 แต่วิธีนี้อาจนำไปสู่การเกิดปัญหาเกินพอดีได้ เช่น กรณีที่กลุ่ม $X_{i,k}$ มีข้อมูลเพียง 1 ตัวอย่าง ค่าตัวเลขผลเฉลยเฉลี่ยก็จะมีค่าเท่ากับค่าผลเฉลยของข้อมูลตัวอย่างอันเดียวนั้น ปัญหานี้อาจแก้ได้โดยการแบ่งชุดข้อมูลออกเป็น 2 ส่วน ส่วนหนึ่งใช้สำหรับคำนวณค่าทางสถิติ และอีกส่วนหนึ่งใช้สำหรับให้ตัวแบบเรียนรู้ แม้จะสามารถลดการเกิดปัญหาเกินพอดีได้ แต่ปริมาณข้อมูลที่ให้ตัวแบบเรียนรู้ก็ลดลงเช่นเดียวกัน (Dorogush et al., 2018)

ตัวแบบ CatBoost จะใช้เทคนิคการเรียงสับเปลี่ยนแบบสุ่มบนชุดข้อมูล แล้วจึงคำนวณ ค่าผลเฉลยเฉลี่ยของตัวอย่างที่มีค่าผลเฉลยเดียวกัน สมมติให้ $\sigma = (\sigma_1, \dots, \sigma_n)$ คือผลลัพธ์การเรียงสับเปลี่ยนแบบสุ่ม จะแสดงค่าทดแทนข้อมูลที่จัดเป็นกลุ่มได้ดังนี้

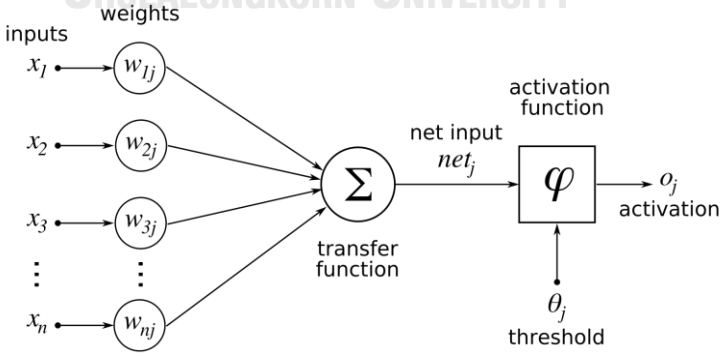
$$X_{\sigma_p,k} = \frac{\sum_{j=1}^{p-1}\left[X_{\sigma_j,k} = X_{\sigma_p,k}\right] Y_{\sigma_j} + a \cdot P}{\sum_{j=1}^{p-1}\left[X_{\sigma_j,k} = X_{\sigma_p,k}\right] + a}$$

โดยจะเพิ่มค่าก่อนหน้า P และพารามิเตอร์ $a > 0$ เป็นน้ำหนักของค่าก่อนหน้า การเพิ่มค่าก่อนหน้าจะช่วยลดการรบกวนจากกลุ่มข้อมูลที่มีจำนวนตัวอย่างน้อย (Cestnik, 1990) สำหรับการวิเคราะห์การจำแนกประเภท จะคำนวณค่าก่อนหน้าด้วยการหาค่าเฉลี่ยของค่าผลเฉลย และสำหรับการวิเคราะห์การถดถอย จะใช้ค่าความน่าจะเป็นเบื้องต้นที่จะเจอผลเฉลยเป็นบวกเป็นค่าก่อนหน้า (Micci-Barreca, 2001) ด้วยเทคนิคที่ตัวแบบ CatBoost ใช้ ทำให้ไม่เกิดปัญหาเกินพอดี (Dorogush et al., 2018)

ตัวแบบ CatBoost มีความยืดหยุ่นและสามารถนำไปประยุกต์ใช้ได้หลากหลายประเภทงาน เช่น การพยากรณ์แนวโน้มลูกค้าที่กำลังจะยกเลิกบริการ (Churn Prediction) การพยากรณ์สภาพอากาศ (Weather Prediction) ระบบแนะนำ (Recommendation Systems) รถยนต์ขับเคลื่อนด้วยตัวเอง (Self-driving Cars)

2.3.3 โครงข่ายประสาทเทียม (Artificial Neural Network)

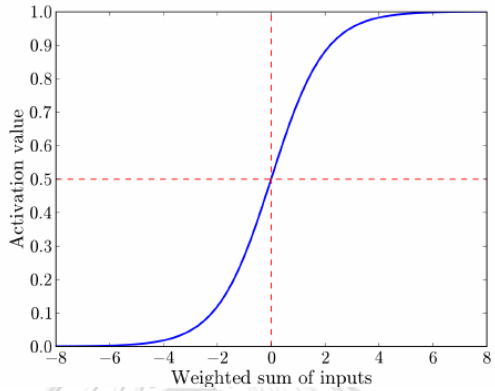
โครงข่ายประสาทเทียมคือระบบการคำนวณของคอมพิวเตอร์ที่จำลองการทำงานของโครงข่ายประสาทชีวภาพในสัตว์หรือมนุษย์ โครงสร้างของโครงข่ายประสาทเทียมแสดงได้ดังรูปที่ 2.5 โดยการประมวลผลต่างๆ จะเกิดในหน่วยประมวลผลย่อยที่เรียกว่า โหนด (Node) N_j จากรูปมีตั้งแต่โหนด 1 ถึง n ที่รับข้อมูลเข้า $X_1, X_2, \dots, X_n$ แต่ละโหนดประกอบไปด้วยเซ็ตของค่าน้ำหนักที่ปรับตัวได้ $W_{1j}, W_{2j}, \dots, W_{nj}$ เนื่องจากโครงข่ายประสาทเทียมจำลองลักษณะการทำงานมาจากเซลล์ส่งสัญญาณระหว่างแต่ละโหนด ดังนั้นแต่ละโหนดจึงเชื่อมต่อกันด้วยการเชื่อมโยงแบบมีน้ำหนัก และให้ค่าข้อมูลออก O_j ผ่านฟังก์ชันรวมการประมวลผล (Activation Function) ϕ



รูปที่ 2.5 โครงสร้างของโครงข่ายประสาทเทียมอย่างง่าย (Camuñas-Mesa et al., 2019)

ฟังก์ชันรวมการประมวลผลในโครงข่ายประสาทเทียมมีหลากหลายรูปแบบ ที่นิยมใช้จะเรียกว่าฟังก์ชันรวมการประมวลผลแบบริดจ์ (Ridge Activation Functions) ที่จะรวมข้อมูลเข้าแบบ

เส้นตรง เช่น ฟังก์ชัน Linear $\phi(v) = a + v'b$ ที่จะให้ค่าข้อมูลออกเป็นค่าเส้นตรง ฟังก์ชัน ReLu (Rectified Linear Unit Function) $\phi(v) = max(0, a + v'b)$ ที่จะให้ค่าข้อมูลออกเป็นค่าที่มากกว่า 0 และฟังก์ชัน Logistic $\phi(v) = \left(1 + exp(-1 - v'b)\right)^{-1}$ ที่จะให้ค่าข้อมูลออกเป็นค่าที่อยู่ระหว่าง 0 และ 1 ดังรูปที่ 2.6 ฟังก์ชัน Logistic นี้จึงสามารถนำไปประยุกต์ใช้ในการจำแนกประเภททวิได้



รูปที่ 2.6 ฟังก์ชันรวมการประมวลแบบ Logistic (Morello et al., 2014)

2.4 การเรียนรู้ของเครื่องอิทธิพลผสม (Mixed Effect Machine Learning)

ในวิธีศึกษาตามยาวหรือวิธีศึกษาระยะยาว (Longitudinal Study) บุคคลจะถูกสังเกตและเก็บข้อมูลหลายครั้งในหลายช่วงเวลา ข้อมูลเหล่านั้นจะใช้เป็นตัวแปรร่วมในการวิเคราะห์การรอดชีพ ซึ่งจะแบ่งเป็น 2 ประเภท ได้แก่ ตัวแปรร่วมคงที่ คือตัวแปรที่คงที่เสมอทุกครั้งที่เก็บข้อมูล เช่น เพศ เชื้อชาติ และตัวแปรร่วมผันแปรตามเวลา คือตัวแปรที่สังเกตได้ในแต่ละครั้งมีค่าเปลี่ยนไปเรื่อยๆ เช่น ผลการตรวจสุขภาพ

ตัวแบบอิทธิพลผสมเชิงเส้นนัยทั่วไป (Generalized Linear Mixed Models, GLMM) ถูกพัฒนาและขยายต่อเนื่องมาจากตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Models, GLM) สำหรับการวิเคราะห์ข้อมูลผลลัพธ์ที่ครอบคลุมทั้งแบบสัมพันธ์กัน เป็นอิสระต่อกัน ต่อเนื่อง และไม่ต่อเนื่อง ภายใต้การนำอิทธิพลแบบคงที่ (Fixed Effect) และอิทธิพลแบบสุ่ม (Random Effect) มาพิจารณาร่วมกัน จึงเรียกว่าอิทธิพลผสม (Mixed Effect) ซึ่งเป็นการวิเคราะห์ที่ให้ผลลัพธ์คลอบคลุมทั้งแบบค่าเฉลี่ยประชากร และแบบเฉพาะผลกระทบในระดับเฉพาะบุคคล (Sarakarn & Jumparway, 2020)

ในการวิเคราะห์ตัวแบบอิทธิพลผสมเชิงเส้นนัยทั่วไป ตัวแปรร่วมคงที่จะถูกใช้เป็นตัวแปรร่วมอิทธิพลคงที่และตัวแปรร่วมผันแปรตามเวลาจะถูกใช้เป็นตัวแปรร่วมอิทธิพลสุ่ม อย่างไรก็ตาม ตัวแปรอิทธิพลสุ่มควรจะเป็นข้อมูลที่จัดเป็นกลุ่มมากกว่า 5 กลุ่มขึ้นไป เนื่องจากการประมาณอิทธิพลสุ่มจะพยายามกำหนดค่าความแปรปรวนระหว่างแต่ละกลุ่ม จึงต้องการจำนวนกลุ่มที่เพียงพอต่อการประมาณอย่างแม่นยำและไม่ผิดพลาด หากมีจำนวนกลุ่มน้อยกว่า 5 กลุ่ม ควรพิจารณาเป็นตัวแปรอิทธิพลคงที่แทน (Bolker et al., 2009)

สำหรับช่วงเวลา t และบุคคล i จะมีตัวแปรร่วมอิทธิพลคงที่เป็นเวกเตอร์ x_{it} ขนาด p มิติ ตัวแปรร่วมอิทธิพลสุ่มเป็นเวกเตอร์ z_{it} ขนาด q มิติ และตัวแปรตอบสนอง y_{it} สามารถแสดงพารามิเตอร์ของประชากรอิทธิพลคงที่ได้ดังนี้

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

เมื่อเวกเตอร์ β คือพารามิเตอร์อิทธิพลคงที่ เวกเตอร์ β_i คือพารามิเตอร์อิทธิพลสุ่ม $\mu_{it} = E[y_{it}|b_i]$ คือความน่าจะเป็นที่จะสำเร็จ และ $g(\cdot)$ คือฟังก์ชันการเชื่อมโยงโลจิท (Logit Link Function)

ตัวแบบอิทธิพลผสมเชิงเส้นสามารถแสดงได้ดังนี้

$$y_{it}^* = \beta^T x_{it} + b_i^T z_{it} + \varepsilon_{it}^* \quad (2)$$

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

การเรียนรู้ของเครื่องอิทธิพลผสมนี้จะประมาณส่วนประกอบอิทธิพลคงที่ ($\beta^T x_{it}$) ด้วยอัลกอริทึมการเรียนรู้ของเครื่อง และประมาณอิทธิพลสุ่ม (b_i) ด้วยตัวแบบอิทธิพลผสมเชิงเส้นนัยทั่วไปจนกว่าจะตัวแบบจะลู่เข้า (Ngufor et al., 2019) ดังนั้นสามารถแสดงสมการ (1) และ (2) ได้ดังนี้

$$\eta_{it} = f(x_i) + b_i^T z_i \text{ และ } y_i^* = f(x_i) + b_i^T z_i + \varepsilon_i^*$$

เมื่อ $f(\cdot)$ เป็นฟังก์ชันที่ไม่ทราบ แต่จะประมาณได้ด้วยอัลกอริทึมการเรียนรู้ของเครื่อง เช่น การสุ่มป่าไม้ หรือ Gradient Boosted Machine เป็นต้น

บทที่ 3
ขอบเขตงานวิจัยและวิธีการดำเนินงานวิจัย

3.1 ขอบเขตงานวิจัย

1. ศึกษาภายใต้ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ (Primary Biliary Cholangitis: PBC) ประกอบไปด้วยผลการตรวจจากห้องปฏิบัติการ และสถานะของผู้ป่วยถูกเก็บจากผู้ป่วยท่อน้ำดีอักเสบปฐมภูมิจำนวน 312 คน ตั้งแต่ปี พ.ศ. 2517 ถึง พ.ศ. 2527 โดย Mayo Clinic สามารถเข้าถึงได้จาก R Package survival
2. ศึกษาภายใต้ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานของประชาชนอายุ 35 ปีขึ้นไปภายในประเทศไทยตั้งแต่ปี พ.ศ. 2557 ถึง พ.ศ. 2563 จากฐานข้อมูลสำนักงานหลักประกันสุขภาพแห่งชาติ (NHSO) จำนวน 1,235 คน

3.2 วิธีการดำเนินงานวิจัย

1. ค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง
2. ออกแบบโครงร่างและขอบเขตของการวิจัย
3. ศึกษาการวิเคราะห์ข้อมูล และเลือกตัวแบบหรือสถิติที่จะใช้วิเคราะห์
4. เตรียมข้อมูลเพื่อใช้สำหรับวิเคราะห์ตัวแบบที่สนใจศึกษา
5. วิเคราะห์ข้อมูลด้วยตัวแบบหรือสถิติที่เลือก
6. เปรียบเทียบผลการวิเคราะห์ และสรุปผล

3.3 แนวทางการวิเคราะห์ข้อมูลและสถิติที่ใช้ในการวิเคราะห์

งานวิจัยนี้ต้องการศึกษาการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบน 2 ชุดข้อมูล คือ ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ และข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน โดยเปรียบเทียบผลลัพธ์การพยากรณ์ความเสี่ยงต่อการไม่รอดชีพในแต่ละช่วงเวลาจากตัวแบบ ทั้งหมด 3 กลุ่ม ดังนี้

1. ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม โดยละเลยความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน และสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิงในการวิเคราะห์

2. ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันด้วยตัวแปรหุ่น (Dummy Variable) ในการวิเคราะห์
3. การเรียนรู้ของเครื่องอิทธิพลผสม โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันด้วยตัวแปรอิทธิพลสุ่มในการวิเคราะห์

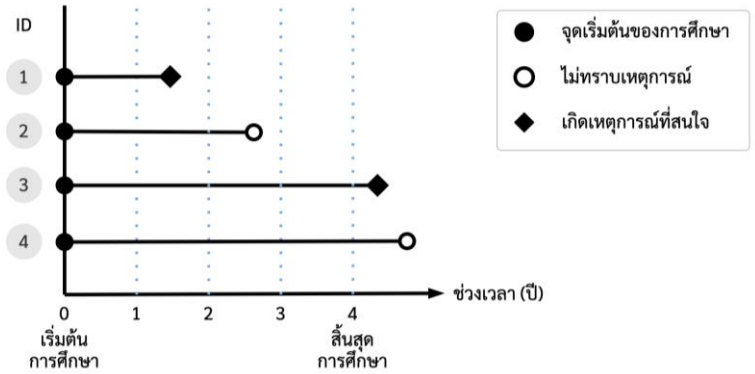
รูปที่ 3.1 แสดงขั้นตอนการดำเนินการวิจัย แบ่งเป็น 3 ขั้นตอน ดังนี้

รูปที่ 3.1 ขั้นตอนการดำเนินการวิจัย ประกอบด้วย ขั้นตอนเตรียมข้อมูล ขั้นตอนวิเคราะห์ข้อมูลและสร้างตัวแบบ และขั้นตอนเปรียบเทียบผลลัพธ์

3.3.1 ขั้นตอนการเตรียมข้อมูลการรอดชีพเวลาไม่ต่อเนื่อง

ข้อมูลการรอดชีพโดยปกติจะเป็นข้อมูลการรอดชีพเวลาต่อเนื่อง เพื่อให้เข้าใจลักษณะของข้อมูลและวิธีการเตรียมข้อมูล จะยกตัวอย่างข้อมูลดังแสดงในรูปที่ 3.2 ประกอบไปด้วย 4 บุคคล (ID) ที่มีจุดเริ่มต้นการศึกษา ณ ช่วงเวลาที่ 0 และมีการติดตามไปเรื่อยๆ ในช่วงเวลาแต่ละปี จนสิ้นสุดการศึกษาที่ช่วงเวลาที่ 4

บุคคลที่ 1 มีการติดตามเรื่อยๆ และเกิดเหตุการณ์ที่สนใจ ณ ช่วงเวลาที่ 1.5 บุคคลที่ 2 มีการติดตามเรื่อยๆ และไม่ทราบเหตุการณ์ ณ ช่วงเวลาที่ 2.6 บุคคลที่ 3 มีการติดตามเรื่อยๆ จนสิ้นสุดช่วงเวลาศึกษา และเกิดเหตุการณ์ที่สนใจ ณ ช่วงเวลาที่ 4.4 และบุคคลที่ 4 มีการติดตามเรื่อยๆ จนสิ้นสุดช่วงเวลาศึกษา และไม่ทราบเหตุการณ์ ณ ช่วงเวลาที่ 4.7



รูปที่ 3.2 ข้อมูลการรอดชีพของแต่ละบุคคลที่แสดงตามเส้นเวลา

ข้อมูลการรอดชีพข้างต้นสามารถแปลงเป็นตารางข้อมูลได้ดังรูปที่ 3.3 ประกอบไปด้วยคอลัมน์ ID คอลัมน์ Time คือระยะเวลาจากจุดเริ่มต้นจนถึงจุดที่เกิดเหตุการณ์ที่สนใจ และคอลัมน์ Event คือประเภทของเหตุการณ์ที่สนใจ

ID	Time	Event
1	1.5	1
2	2.6	0
3	4.4	1
4	4.7	0

รูปที่ 3.3 ข้อมูลการรอดชีพในรูปแบบตาราง

เพื่อวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องด้วยวิธีการจำแนกแบบทวิจะจัดรูปแบบข้อมูลการรอดชีพเวลาต่อเนื่องของแต่ละบุคคลให้เป็นข้อมูลการรอดชีพเวลาไม่ต่อเนื่องดังรูปที่ 3.4 โดยจะระบุว่าเป็นข้อมูลจากบุคคลเดียวกันด้วยคอลัมน์ ID หลังจากนั้นจะแบ่งค่าเวลาการรอดชีพต่อเนื่องออกเป็นช่วง ช่วงเวลาละ 1 ปี และเก็บเป็นค่าในคอลัมน์ Interval เริ่มต้นจากค่า 1, 2, ... ไปเรื่อยๆ จนกว่าจะสิ้นสุดช่วงเวลาของบุคคลนั้นหรือช่วงเวลาศึกษา และให้คอลัมน์ Event เป็นประเภทของเหตุการณ์ที่สนใจในช่วงเวลานั้น

ในการนำข้อมูลไปใช้ ข้อมูลแบบพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันจะนำไปใช้ทุกคอลัมน์ ดังรูปที่ 3.4 (ซ้าย) และข้อมูลแบบละเลยความสัมพันธ์ของข้อมูลระหว่างบุคคลเดียวกันจะไม่นำคอลัมน์ ID ไปใช้ ดังรูปที่ 3.4 (ขวา)

ID	Interval	Event
1	1	0
1	2	1
2	1	0
2	2	0
2	3	0
3	1	0
3	2	0
3	3	0
3	4	0
4	1	0
4	2	0
4	3	0
4	4	0

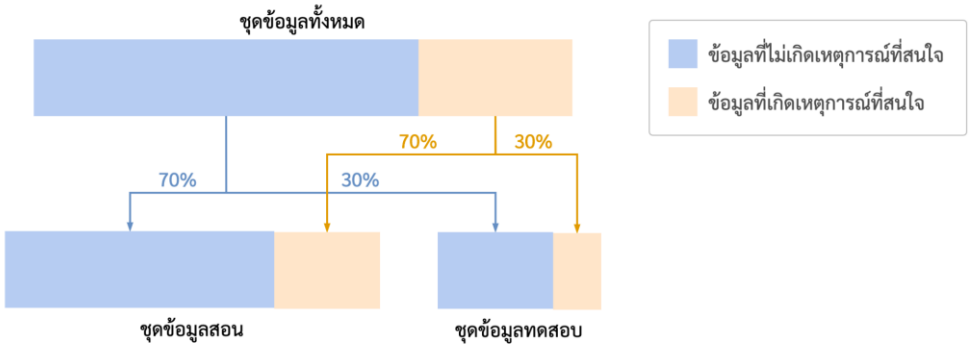
Interval	Event
1	0
2	1
1	0
2	0
3	0
1	0
2	0
3	0
4	0
1	0
2	0
3	0
4	0

รูปที่ 3.4 ตัวอย่างรูปแบบข้อมูลการรอดชีพเวลาไม่ต่อเนื่องเพื่อใช้ในการวิเคราะห์ด้วยการจำแนกแบบทวิ รูปซ้ายแสดงข้อมูลแบบพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน และรูปขวาแสดงข้อมูลแบบละเลยความสัมพันธ์ของข้อมูลระหว่างบุคคลเดียวกัน และสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง

สังเกตได้ว่าเมื่อไม่มีคอลัมน์ ID บ่งบอกว่าข้อมูลแถวใดสัมพันธ์กับแถวใดจะเปรียบเสมือนข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง ซึ่งเป็นสมมติฐานมาตรฐานของอัลกอริทึมการจำแนกส่วนมาก (Dundar et al., 2007) และเมื่อมีคอลัมน์ ID ข้อมูลนั้นจะสูญเสียคุณสมบัติการแจกแจงเหมือนกันและเป็นอิสระต่อกัน

เพื่อให้การวิเคราะห์ทุกตัวแบบเป็นไปในมาตรฐานเดียวกันและสามารถเปรียบเทียบผลลัพธ์กันได้ จะควบคุมชุดข้อมูลที่ใช้สอนและทดสอบตัวแบบ โดยแบ่งชุดข้อมูลทั้งหมดออกเป็น 2 ชุด คือ ชุดข้อมูลสอนสำหรับสอนตัวแบบ และชุดข้อมูลทดสอบสำหรับวัดประสิทธิภาพของตัวแบบ ด้วยวิธีการสุ่มตัวอย่างแบบแบ่งชั้นภูมิเพื่อคงอัตราส่วนประเภทของเหตุการณ์ที่สนใจให้เท่ากันในแต่ละชุดข้อมูล และจะกำหนดให้แต่ละ ID อยู่ในชุดข้อมูลสอนหรือชุดข้อมูลทดสอบเพียงอย่างใดอย่างหนึ่งด้วยวิธีการสุ่มแบบไม่ใส่คืน

วิธีการแบ่งชุดข้อมูลแสดงได้ดังรูปที่ 3.5 คือ จากชุดข้อมูลทั้งหมด จะถูกแบ่งเป็นชุดข้อมูลสอนในปริมาณ 70% ของข้อมูลทั้งหมด และชุดข้อมูลทดสอบในปริมาณ 30% ของข้อมูลทั้งหมด



รูปที่ 3.5 วิธีการแบ่งชุดข้อมูลเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบด้วยวิธีการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ

รายละเอียดของชุดข้อมูลและการเตรียมข้อมูลเป็นไปดังนี้

3.3.1.1 ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ

ข้อมูลท่อน้ำดีอักเสบปฐมภูมิเป็นข้อมูลที่เก็บโดย Mayo Clinic ตั้งแต่ปี พ.ศ. 2517 ถึง พ.ศ. 2527 โดยเก็บจากผู้ป่วยท่อน้ำดีอักเสบปฐมภูมิ จำนวน 312 คนที่เข้าร่วมการทดลองยา D-penicillamine ที่มีการควบคุมด้วยยาหลอกแบบสุ่ม ผู้ป่วยจะถูกติดตามข้อมูลผลการตรวจจากห้องปฏิบัติการ และสถานะของผู้ป่วยในการติดตามแต่ละครั้ง

ข้อมูลชุดนี้ประกอบไปด้วย 1,945 แถว แต่ละแถวคือข้อมูลผลการตรวจจากห้องปฏิบัติการแต่ละครั้งของแต่ละบุคคล โครงสร้างข้อมูลท่อน้ำดีอักเสบปฐมภูมิทั้ง 19 คอลัมน์แสดงดัง

ตารางที่ 3.1 โดยคอลัมน์ status เป็นสถานะของผู้ป่วยที่ต้องการพยากรณ์ด้วยตัวแบบ ทั้งนี้ในงานวิจัยนี้จะศึกษาเฉพาะสถานะของผู้ป่วยที่เป็นข้อมูลตรวจตัด (status=0) หรือเสียชีวิต (status=2) เท่านั้น ดังนั้นเมื่อคัดกรองข้อมูลแล้วจะเหลือข้อมูล 1,798 แถว ในที่นี้ ตัวแปรร่วมคงที่จะถูกนำไปใช้เป็นตัวแปรอิทธิพลคงที่ ตัวแปรร่วมผันแปรตามเวลาจะถูกนำไปใช้เป็นตัวแปรอิทธิพลสุ่ม ตัวแปร id ใช้สำหรับการแบ่งกลุ่มข้อมูล และจะไม่พิจารณาตัวแปร futime และ day ในการวิเคราะห์

ตารางที่ 3.1 โครงสร้างของข้อมูลท่อน้ำดีอักเสบปฐมภูมิ

ชื่อคอลัมน์	ประเภทข้อมูล	ประเภทตัวแปร	คำอธิบาย
id	Factor	-	หมายเลข ID ของผู้ป่วย
age	Numeric	คงที่	อายุ (ปี)
sex	Factor	คงที่	เพศ (m=ชาย, f=หญิง)
trt	Factor	คงที่	การรักษา (0=ได้รับยาหลอก, 1=ได้รับยา D-penicillmain)
futime	Integer	-	จำนวนวันตั้งแต่วันที่รับการรักษาจนถึงก่อนวันที่เสียชีวิต หรือปลูกถ่ายอวัยวะ หรือสิ้นสุดการศึกษา
status	Factor	-	สถานะของผู้ป่วย (0=ข้อมูลตรวจตัด, 1=ปลูกถ่ายอวัยวะ, 2=เสียชีวิต)
day	Integer	-	จำนวนวันตั้งแต่วันที่รับการรักษาจนถึงวันที่ได้รับการตรวจจากห้องปฏิบัติการแต่ละครั้ง
albumin	Numeric	ผันแปรตามเวลา	ระดับของ Albumin ในเลือด (mg/dl)
alk.phos	Integer	ผันแปรตามเวลา	ค่าเอนไซม์ Alkaline Phosphatase ที่พบในเลือด (U/liter)
ascites	Factor	ผันแปรตามเวลา	อาการท้องมาน (0=ไม่มีอาการ, 1=มีอาการ)
ast	Numeric	ผันแปรตามเวลา	ค่าเอนไซม์ Aspartate Aminotransferase หรือ SGOT (U/ml)
bili	Numeric	ผันแปรตามเวลา	ระดับของ Bilirubin ในเลือด (mg/dl)
chol	Integer	ผันแปรตามเวลา	ระดับของ Cholesterol ในเลือด (mg/dl)
edema	Integer	ผันแปรตามเวลา	อาการบวมน้ำ (0=ไม่มีอาการ, 0.5=มีอาการแต่ไม่ได้รักษาหรือรักษาแล้ว, 1=มีอาการบวมน้ำแม้จะมีการรักษาด้วยยาขับปัสสาวะ)
hepato	Factor	ผันแปรตามเวลา	อาการตับโต (0=ไม่มีอาการ, 1=มีอาการ)
platelet	Integer	ผันแปรตามเวลา	จำนวนของเกร็ดเลือดต่อ mL ในเลือด
protime	Numeric	ผันแปรตามเวลา	ระยะเวลาในการแข็งตัวของเลือด

spiders	Factor	ผันแปรตามเวลา	อาการรูปผิดปกติของหลอดเลือดใน ผิวหนัง (0=ไม่มีอาการ, 1=มีอาการ)
stage	Integer	ผันแปรตามเวลา	ระยะของโรคจากการผ่าตัดชิ้นเนื้อเพื่อ วินิจฉัย

ตัวอย่างของข้อมูลท่อน้ำดีอักเสบปฐมภูมิของ 3 บุคคลแสดงดังรูปที่ 3.6 จะสังเกตได้ว่า จำนวนแถวข้อมูลของแต่ละบุคคลไม่เท่ากัน และมีข้อมูลบางส่วนสูญหายไป (NA)

```
> tail(pbcseq, 14)
      id futime status trt      age sex  day ascites hepato spiders edema bili chol albumin alk.phos ast platelet protime stage
1785 310   1608      0   1 62.33265   f    0       0      0       0   0.5  1.7  434    3.35     1713 171      234    10.2     2
1786 310   1608      0   1 62.33265   f  260       0      0       0   0.5  1.7  516    3.24     1661  82      249     9.9     3
1787 310   1608      0   1 62.33265   f  623       0      1       0   0.5  2.2  386    2.94     1808  81      268    10.9     3
1788 310   1608      0   1 62.33265   f  988       0      1       0   0.5  1.4  367    2.71     1584  72      260    11.3     3
1789 310   1608      0   1 62.33265   f 1353       0      1       0   0.5  1.8  364    3.19     1350  65      272    11.3     3
1790 311   1508      0   1 37.99863   f    0       0      0       0   0.0  2.0  247    3.16     1050 117      335    10.5     2
1791 311   1508      0   1 37.99863   f  187       0   <NA>    <NA>   0.0  1.5   NA    3.41     2562 123      382    10.5     2
1792 311   1508      0   1 37.99863   f  397       0      0       0   0.0  1.9  424    3.57     2516 166      408    10.6     3
1793 311   1508      0   1 37.99863   f 1098       0      0       0   0.0  0.6  391    3.40     2322 191      337    11.4     3
1794 312   1457      0   0 33.15264   f    0       0      0       1   0.0  6.4  576    3.79     2115 136      200    10.8     2
1795 312   1457      0   0 33.15264   f  206       0      0       0   0.0  5.5   NA    3.20     1678 124      189    10.9     2
1796 312   1457      0   0 33.15264   f  390       0      0       0   0.0  7.4  312    3.56     1767 166      148    11.7     2
1797 312   1457      0   0 33.15264   f  775       0      0       1   0.5 16.3  688    3.34     2460 173      138    13.0     2
1798 312   1457      0   0 33.15264   f 1075       0      0       1   0.5 23.4  741    3.42     3012 200      128    13.4     3
```

รูปที่ 3.6 ตัวอย่างของข้อมูลท่อน้ำดีอักเสบปฐมภูมิ

เมื่อตรวจสอบข้อมูลพบว่ามี 6 คอลัมน์ที่มีข้อมูลสูญหายไป ได้แก่ ascites, hepato, spiders, chol, alk.phos และ platelet แสดงจำนวนข้อมูลสูญหายดังตารางที่ 3.2 โดยจะประมาณค่าข้อมูลสูญหายเหล่านี้ด้วยวิธี K-nearest Neighbors Imputation

วิธี K-nearest Neighbors Imputation ที่ k จะใช้หลักการประมาณค่าข้อมูลสูญหายด้วยข้อมูล k แถวที่มีความใกล้เคียงกันมากที่สุด โดยวัดความใกล้เคียงด้วยระยะทางโกเวอร์ (Gower Distance) ซึ่งมีวิธีคำนวณที่แตกต่างกันตามประเภทของข้อมูล ระยะทางโกเวอร์ระหว่าง X_1 และ X_2 สามารถแสดงได้ดังนี้

$$D_{Gower}(X_1, X_2) = 1 - \left(\frac{1}{k} \sum_{j=1}^{k} S_j(X_1, X_1) \right)$$

เมื่อ $S_j(X_1, X_2)$ เป็นฟังก์ชันความเหมือนระหว่าง X_1 และ X_2 ในกรณีที่เป็นข้อมูลตัวเลข $S_j(X_1, X_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j}$ โดย R_j คือพิสัยของข้อมูล และในกรณีที่เป็นข้อมูลที่จัดเป็นกลุ่ม $S_j(X_1, X_2) = 1$ เมื่อเป็นกลุ่มเดียวกัน และ $S_j(X_1, X_2) = 0$ เมื่อต่างกลุ่มกัน ระยะห่างโกเวอร์จะมีค่าระหว่าง 0 ถึง 1 โดยค่า 0 หมายถึงข้อมูลเหมือนกันทั้งหมด และค่า 1 หมายถึงข้อมูลแตกต่างกัน

มากที่สุด รูปที่ 3.7 แสดงตัวอย่างผลลัพธ์ของข้อมูลที่ประมาณค่าด้วยวิธี K-nearest Neighbors Imputation ที่ k=5 แล้ว

ตารางที่ 3.2 จำนวนข้อมูลสูญหายของข้อมูลท่อน้ำดีอักเสบปฐมภูมิ

ชื่อคอลัมน์	ประเภทข้อมูล	จำนวนข้อมูลสูญหาย (แถว)	จำนวนข้อมูลสูญหาย (ร้อยละ)
ascites	Factor	56	3.1146
hepato	Factor	57	3.1702
spiders	Factor	54	3.0033
chol	Integer	766	42.6029
alk.phos	Integer	54	3.0033
platelet	Integer	68	3.7819

```
> tail(pbcseq.imputed, 14)
      id futime status trt      age sex  day ascites hepato spiders edema bili chol albumin alk.phos ast platelet protime stage
1785 310   1608      0   1 62.33265   f    0       0      0       0   0.5  1.7  434    3.35     1713 171      234    10.2     2
1786 310   1608      0   1 62.33265   f  260       0      0       0   0.5  1.7  516    3.24     1661  82      249     9.9     3
1787 310   1608      0   1 62.33265   f  623       0      1       0   0.5  2.2  386    2.94     1808  81      268    10.9     3
1788 310   1608      0   1 62.33265   f  988       0      1       0   0.5  1.4  367    2.71     1584  72      260    11.3     3
1789 310   1608      0   1 62.33265   f 1353       0      1       0   0.5  1.8  364    3.19     1350  65      272    11.3     3
1790 311   1508      0   1 37.99863   f    0       0      0       0   0.0  2.0  247    3.16     1050 117      335    10.5     2
1791 311   1508      0   1 37.99863   f  187       0      0       0   0.0  1.5  335    3.41     2562 123      382    10.5     2
1792 311   1508      0   1 37.99863   f  397       0      0       0   0.0  1.9  424    3.57     2516 166      408    10.6     3
1793 311   1508      0   1 37.99863   f 1098       0      0       0   0.0  0.6  391    3.40     2322 191      337    11.4     3
1794 312   1457      0   0 33.15264   f    0       0      0       1   0.0  6.4  576    3.79     2115 136      200    10.8     2
1795 312   1457      0   0 33.15264   f  206       0      0       0   0.0  5.5  281    3.20     1678 124      189    10.9     2
1796 312   1457      0   0 33.15264   f  390       0      0       0   0.0  7.4  312    3.56     1767 166      148    11.7     2
1797 312   1457      0   0 33.15264   f  775       0      0       1   0.5 16.3  688    3.34     2460 173      138    13.0     2
1798 312   1457      0   0 33.15264   f 1075       0      0       1   0.5 23.4  741    3.42     3012 200      128    13.4     3
```

รูปที่ 3.7 ตัวอย่างของข้อมูลท่อน้ำดีอักเสบปฐมภูมิที่ประมาณค่าข้อมูลสูญหายด้วยวิธี K-nearest Neighbors Imputation

หลังจากนั้นจะแบ่งชุดข้อมูลนี้ออกเป็น 2 ชุด คือ ชุดข้อมูลสอน และชุดข้อมูลทดสอบ ด้วยอัตราส่วน 70 : 30 โดยจะสรุปสถานะเหตุการณ์สุดท้ายของผู้ป่วยแต่ละคน และกำหนดให้ผู้ป่วยแต่ละคนอยู่ในชุดข้อมูลสอนหรือชุดข้อมูลทดสอบเพียงอย่างใดอย่างหนึ่งด้วยวิธีการสุ่มแบบไม่ใส่คืนและแบ่งชั้นภูมิ เพื่อคงอัตราส่วนประเภทของเหตุการณ์ที่สนใจให้เท่าๆ กันในแต่ละชุดข้อมูล

จากการแบ่งชุดข้อมูลด้วยวิธีดังกล่าว มีผู้ป่วยจำนวน 198 คนอยู่ในกลุ่มข้อมูลสอน และ 85 คนอยู่ในกลุ่มข้อมูลทดสอบ โดยมีอัตราส่วนประเภทของเหตุการณ์ ข้อมูลตรวจตัด : เสียชีวิต ของผู้ป่วยทั้งสองกลุ่มเป็น 49.5 : 50.5

ในการเตรียมชุดข้อมูลสอน จะคัดกรองข้อมูลทั้งหมดด้วย ID ของผู้ป่วยกลุ่มข้อมูลสอน ได้เป็นชุดข้อมูลสอนจำนวน 1,260 แถว มีอัตราส่วนประเภทของเหตุการณ์ ข้อมูลตรวจตัด : เสียชีวิต เป็น 61.27 : 38.73 และเตรียมชุดข้อมูลทดสอบโดยคัดกรองข้อมูลทั้งหมดด้วย ID ของผู้ป่วยกลุ่มข้อมูลทดสอบ ได้เป็นชุดข้อมูลทดสอบจำนวน 538 แถว มีอัตราส่วนประเภทของเหตุการณ์ ข้อมูลตรวจตัด : เสียชีวิต เป็น 59 : 41 ซึ่งจะใช้ชุดข้อมูลสอนและชุดข้อมูลทดสอบนี้สำหรับวิเคราะห์และสร้างตัวแบบต่อไป

3.3.1.2 ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน

ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานเป็นข้อมูลที่เก็บโดยสำนักงานหลักประกันสุขภาพแห่งชาติ (NHSO) ตั้งแต่ปี พ.ศ. 2557 ถึง พ.ศ. 2563 โดยเก็บข้อมูลจากประชาชนในประเทศไทยที่มีอายุ 35 ปีขึ้นไป ณ วันที่เริ่มเข้ารับการคัดกรองโรคเบาหวานจำนวนทั้งหมด 1,181 คน

ข้อมูลชุดนี้ประกอบไปด้วย 6,848 แถว แต่ละแถวคือข้อมูลการตรวจคัดกรองโรคเบาหวานรายปีของแต่ละบุคคล โดยเริ่มต้นปีแรกด้วยสถานะไม่เป็นโรคเบาหวาน และได้รับการตรวจคัดกรองเรื่อยๆ ทุกปีจนถึงปีที่สิ้นสุดระยะเวลาที่เก็บข้อมูล โครงสร้างข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานทั้ง 13 คอลัมน์แสดงดังตารางที่ 3.3 ในที่นี้ คอลัมน์ next_event เป็นสถานะของบุคคลนั้นในปีถัดไปที่ต้องการพยากรณ์ด้วยตัวแบบ ตัวแปรร่วมคงที่จะถูกนำไปใช้เป็นตัวแปรอิทธิพลคงที่ ตัวแปรร่วมผันแปรตามเวลาจะถูกนำไปใช้เป็นตัวแปรอิทธิพลสุ่ม ตัวแปร PID_EN ใช้สำหรับการแบ่งกลุ่มข้อมูล และจะไม่พิจารณาตัวแปร year ในการวิเคราะห์

ตารางที่ 3.3 โครงสร้างของข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน

ชื่อคอลัมน์	ประเภทข้อมูล	ประเภทตัวแปร	คำอธิบาย
PID_EN	Character	-	หมายเลข ID ที่ถูกเข้ารหัสไว้
gender	Factor	คงที่	เพศ (1=ชาย, 2=หญิง)
age	Integer	คงที่	อายุ (ปี)
year	Integer	-	ปี ค.ศ. ที่เก็บข้อมูล
BMI	Numeric	ผันแปรตามเวลา	ค่าดัชนีมวลกาย
WAIST_CM	Numeric	ผันแปรตามเวลา	ค่าความยาวรอบเอว (เซนติเมตร)
SBP	Numeric	ผันแปรตามเวลา	ค่าความดันโลหิตสูงสุดขณะหัวใจห้องล่าง

			บีบตัว (mm)
DBP	Numeric	ผันแปรตามเวลา	ค่าความดันโลหิตต่ำสุดขณะหัวใจห้องล่างคลายตัว (Hg)
SMOKE	Factor	คงที่	ประวัติการสูบบุหรี่ (1=ไม่สูบ, 2=สูบนานๆ ครั้ง, 3=สูบเป็นครั้งคราว, 4=สูบเป็นประจำ)
DMFAMILY	Factor	คงที่	ประวัติเบาหวานในญาติสายตรง (0=ไม่มี, 1=มี)
HTFAMILY	Factor	คงที่	ประวัติความดันโลหิตสูงในญาติสายตรง (0=ไม่มี, 1=มี)
HT	Factor	คงที่	ภาวะความดันโลหิตสูง (0=ไม่มีภาวะ, 1=มีภาวะ)
next_event	Factor	-	สถานะในปีถัดไป (0=ไม่เป็นเบาหวาน, 1=เป็นเบาหวาน)

ตัวอย่างของข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานของ 3 บุคคลแสดงดังรูปที่ 3.8 โดยมีการจัดการค่าผิดปกติ ดังนี้ จากความรู้เฉพาะทางการแพทย์ ตัวแปร SBP ควรจะมีค่าอยู่ในช่วง 75 ถึง 250 และตัวแปร DBP ควรจะมีค่าอยู่ในช่วง 30 ถึง 150 และจากค่ารั้วบนและรั้วล่างของแผนภาพกล่อง ตัวแปร BMI ควรมีค่าอยู่ในชวง 14.5 ถึง 31.2 และตัวแปร WAIST_CM ควรจะมีค่าอยู่ในช่วง 60 ถึง 100 หากค่าสังเกตอยู่นอกเหนือจากช่วงเหมาะสมก็จะแทนที่ด้วยค่าสุดขอบของช่วง นอกจากนี้ จะสังเกตได้ว่าจำนวนแถวข้อมูลของแต่ละบุคคลไม่เท่ากัน และมีข้อมูลบางส่วนสูญหายไป (NA)

```
> tail(ncd, 15)
                      PID_EN gender age year      BMI WAIST_CM SBP DBP SMOKE DMFAMILY HTFAMILY HT next_event
5307 zf5wIkfMR5qK5+YWEqjLqQ==      1  45 2014 23.23346       78 113  62     1        0        0 NA          0
5308 zf5wIkfMR5qK5+YWEqjLqQ==      1  46 2015       NA       NA  NA  NA    NA       NA       NA NA          0
5309 zf5wIkfMR5qK5+YWEqjLqQ==      1  48 2017 23.23346       80 113  74     1        0       NA NA          1
5310 zf5wIkfMR5qfIOrHdhmfGw==      1  72 2014 20.02884       81 130  86     1        0        0 NA          0
5311 zf5wIkfMR5qfIOrHdhmfGw==      1  73 2015 20.02884       81 118  65    NA        0       NA NA          0
5312 zf5wIkfMR5qfIOrHdhmfGw==      1  74 2016 20.02884       81 110  60    NA        0       NA NA          0
5313 zf5wIkfMR5qfIOrHdhmfGw==      1  75 2017 20.02884       81 120  70     1        0       NA NA          0
5314 zf5wIkfMR5qfIOrHdhmfGw==      1  76 2018 20.42942       74 120  72     1        0       NA NA          0
5315 zf5wIkfMR5qfIOrHdhmfGw==      1  77 2019       NA       NA  NA  NA    NA       NA       NA  1          0
5316 zf5wIkfMR5rk32TukZEFJQ==      1  68 2014 28.90625       98 112  74     1        0        0 NA          0
5317 zf5wIkfMR5rk32TukZEFJQ==      1  69 2015 28.90625       98 132  87    NA        0       NA NA          0
5318 zf5wIkfMR5rk32TukZEFJQ==      1  70 2016 28.51562      100 127  87    NA        0       NA NA          0
5319 zf5wIkfMR5rk32TukZEFJQ==      1  71 2017 28.90625       98 118  74     1        0       NA NA          0
5320 zf5wIkfMR5rk32TukZEFJQ==      1  72 2018 26.56250      101 136  82     1        0       NA NA          0
5321 zf5wIkfMR5rk32TukZEFJQ==      1  73 2019 26.56250       88 117  68     1        0       NA NA          0
```

รูปที่ 3.8 ตัวอย่างของข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน

เมื่อตรวจสอบข้อมูลพบว่ามี 8 คอลัมน์ที่มีข้อมูลสูญหายไป ได้แก่ BMI, WAIST_CM, SBP, DBP, SMOKE, DMFAMILY, HTFAMILY และ HT แสดงจำนวนข้อมูลสูญหายดังตารางที่ 3.4 โดยจะประมาณค่าข้อมูลสูญหายสำหรับคอลัมน์ BMI, WAIST_CM, SBP และ DBP ด้วยค่ากลางของแต่ละบุคคล และประมาณค่าข้อมูลสูญหายสำหรับคอลัมน์ SMOKE, DMFAMILY, HTFAMILY และ HT ด้วยค่าอื่นที่พบในข้อมูลของบุคคลนั้น โดยตั้งสมมติฐานว่าพฤติกรรมการสูบบุหรี่ ประวัติทางสุขภาพของญาติ และภาวะความดันโลหิตสูงของบุคคลไม่เปลี่ยนแปลง รูปที่ 3.9 แสดงตัวอย่างผลลัพธ์ของข้อมูลที่ประมาณค่าแล้ว

ตารางที่ 3.4 จำนวนข้อมูลสูญหายของข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน

ชื่อคอลัมน์	ประเภทข้อมูล	จำนวนข้อมูลสูญหาย (แถว)	จำนวนข้อมูลสูญหาย (ร้อยละ)
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	Factor	3,603	52.6139
DMFAMILY	Factor	3,220	47.0210
HTFAMILY	Factor	6,142	89.6904
HT	Factor	2,466	36.0105

```
> tail(ncd, 15)
                      PID_EN gender age year      BMI WAIST_CM SBP  DBP SMOKE DMFAMILY HTFAMILY HT next_event
5307 zf5wIkfMR5qK5+YWEqjLqQ==      1  45 2014 23.23346     78.0 113 62.0     1        0        0  1          0
5308 zf5wIkfMR5qK5+YWEqjLqQ==      1  46 2015 23.23346     74.5 113 74.5     1        0        0  1          0
5309 zf5wIkfMR5qK5+YWEqjLqQ==      1  48 2017 23.23346     80.0 113 74.0     1        0        0  1          1
5310 zf5wIkfMR5qfIOrHdhmfGw==      1  72 2014 20.02884     81.0 130 86.0     1        0        0  1          0
5311 zf5wIkfMR5qfIOrHdhmfGw==      1  73 2015 20.02884     81.0 118 65.0     1        0        0  1          0
5312 zf5wIkfMR5qfIOrHdhmfGw==      1  74 2016 20.02884     81.0 110 60.0     1        0        0  1          0
5313 zf5wIkfMR5qfIOrHdhmfGw==      1  75 2017 20.02884     81.0 120 70.0     1        0        0  1          0
5314 zf5wIkfMR5qfIOrHdhmfGw==      1  76 2018 20.42942     74.0 120 72.0     1        0        0  1          0
5315 zf5wIkfMR5qfIOrHdhmfGw==      1  77 2019 20.02884     81.0 119 71.0     1        0        0  1          0
5316 zf5wIkfMR5rk32TukZEFJQ==      1  68 2014 28.90625     98.0 112 74.0     1        0        0  1          0
5317 zf5wIkfMR5rk32TukZEFJQ==      1  69 2015 28.90625     98.0 132 87.0     1        0        0  1          0
5318 zf5wIkfMR5rk32TukZEFJQ==      1  70 2016 28.51562    100.0 127 87.0     1        0        0  1          0
5319 zf5wIkfMR5rk32TukZEFJQ==      1  71 2017 28.90625     98.0 118 74.0     1        0        0  1          0
5320 zf5wIkfMR5rk32TukZEFJQ==      1  72 2018 26.56250    101.0 136 82.0     1        0        0  1          0
5321 zf5wIkfMR5rk32TukZEFJQ==      1  73 2019 26.56250     88.0 117 68.0     1        0        0  1          0
```

รูปที่ 3.9 ตัวอย่างของข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน
ที่ประมาณค่าข้อมูลสูญหายแล้ว

ข้อมูลที่ต้องการใช้ในการวิจัยนี้คือข้อมูลการตรวจคัดกรองโรคเบาหวานรายปีของแต่ละบุคคล โดยเริ่มต้นปีแรกด้วยสถานะไม่เป็นโรคเบาหวาน และได้รับการตรวจคัดกรองเรื่อยๆ ทุกปีจนถึงปีที่ตรวจคัดกรองแล้วพบว่าเป็นโรคเบาหวาน หรือสิ้นสุดระยะเวลาที่เก็บข้อมูลเท่านั้น เมื่อกรองข้อมูลแล้วจึงเหลือ 5,027 แถว จากประชากรจำนวน 1,175 คน

หลังจากนั้นจะแบ่งชุดข้อมูลนี้ออกเป็น 2 ชุด คือ ชุดข้อมูลสอน และชุดข้อมูลทดสอบ ด้วยอัตราส่วน 70 : 30 โดยจะสรุปสถานะเหตุการณ์สุดท้ายของแต่ละบุคคล และกำหนดให้แต่ละบุคคลอยู่ในชุดข้อมูลสอนหรือชุดข้อมูลทดสอบเพียงอย่างใดอย่างหนึ่งด้วยวิธีการสุ่มแบบไม่ใส่คืนและแบ่งชั้นภูมิ เพื่อคงอัตราส่วนประเภทของเหตุการณ์ที่สนใจให้เท่าๆ กันในแต่ละชุดข้อมูล

จากการแบ่งชุดข้อมูลด้วยวิธีดังกล่าว มีบุคคลจำนวน 865 คนอยู่ในกลุ่มข้อมูลสอน และ 370 คนอยู่ในกลุ่มข้อมูลทดสอบ โดยมีอัตราส่วนประเภทของเหตุการณ์ ไม่เป็นโรคเบาหวาน : เป็นโรคเบาหวาน ของทั้งสองกลุ่มเป็นประมาณ 83.37 : 16.63

ในการเตรียมชุดข้อมูลสอน จะคัดกรองข้อมูลทั้งหมดด้วย ID ของบุคคลกลุ่มข้อมูลสอน ได้เป็นชุดข้อมูลสอนจำนวน 3,525 แถว มีอัตราส่วนประเภทของเหตุการณ์ ไม่เป็นโรคเบาหวาน : เป็นโรคเบาหวาน เป็น 96.05 : 3.95

ในการเตรียมชุดข้อมูลทดสอบ จะคัดกรองข้อมูลทั้งหมดด้วย ID ของบุคคลกลุ่มข้อมูลทดสอบ ได้เป็นชุดข้อมูลทดสอบจำนวน 1,502 แถว มีอัตราส่วนประเภทของเหตุการณ์ ไม่เป็นโรคเบาหวาน : เป็นโรคเบาหวาน เป็น 96.29 : 3.71 ซึ่งจะใช้ชุดข้อมูลสอนและชุดข้อมูลทดสอบนี้สำหรับวิเคราะห์และสร้างตัวแบบต่อไป

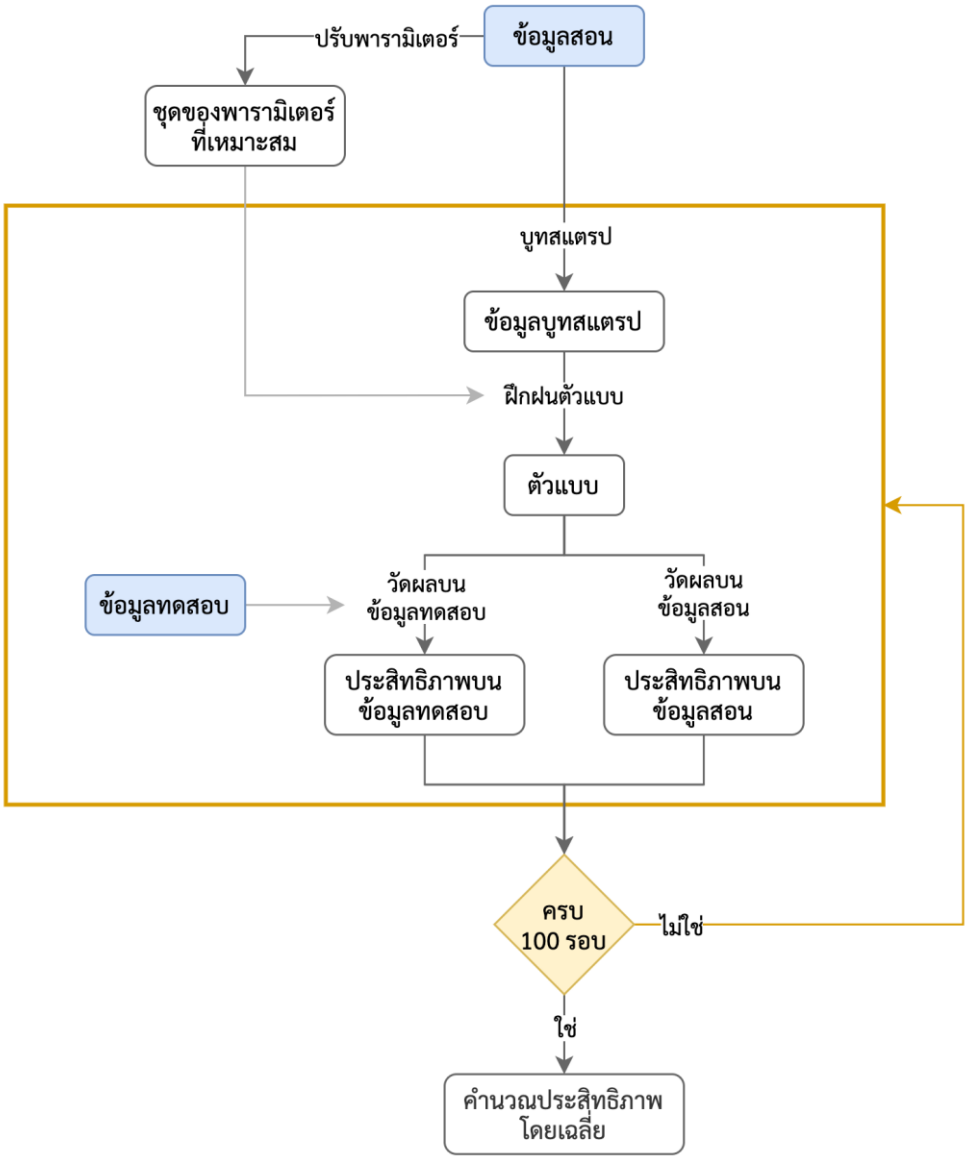
อย่างไรก็ตาม เนื่องจากได้มีการกำหนดให้แต่ละ ID อยู่ในชุดข้อมูลสอนหรือชุดข้อมูลทดสอบเพียงอย่างใดอย่างหนึ่งเท่านั้น ดังนั้น ตัวแบบที่ใช้ตัวแปร ID ในการวิเคราะห์จะไม่สามารถพยากรณ์ข้อมูลสอนที่มี ID อยู่นอกเหนือจากในข้อมูลสอนได้ จึงได้ใช้อัลกอริทึม K-nearest Neighbor กำหนดบุคคลในข้อมูลสอนที่ใกล้เคียงกับบุคคลในข้อมูลทดสอบมากที่สุด โดยใช้เพียงข้อมูลแถวแรกของแต่ละบุคคลในการคำนวณความเหมือน ส่วนข้อมูลแถวที่เหลือจะถูกกำหนดเป็นบุคคลเดียวกันกับแถวแรก ด้วยวิธีนี้ตัวแบบจะสามารถพยากรณ์ข้อมูลสอนได้ นอกจากนี้ ในการนำไปใช้งานจริงยังสามารถประยุกต์ใช้วิธีนี้เพื่อพยากรณ์บุคคลใหม่นอกเหนือจากที่มีข้อมูลสอนได้

3.3.2 ขั้นตอนการวิเคราะห์ข้อมูลและสร้างตัวแบบ

ขั้นตอนการวิเคราะห์ข้อมูลและสร้างตัวแบบ จะแบ่งเป็น 2 กลุ่ม คือ

1. กลุ่มที่วิเคราะห์โดยละเลยความสัมพันธ์ของข้อมูลระหว่างบุคคลเดียวกัน และสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิงในการวิเคราะห์ โดยใช้ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม
2. กลุ่มที่วิเคราะห์โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันในการวิเคราะห์ โดยใช้ตัวแบบการเรียนรู้ของเครื่องในการจำแนกประเภททวิ ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost, โครงข่ายประสาทเทียม และการเรียนรู้ของเครื่องอิทธิพลผสมโดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันด้วยตัวแปรอิทธิพลสุ่ม

กระบวนการวิเคราะห์สำหรับตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่ ได้แก่ การสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม แสดงดังรูปที่ 3.10 โดยเริ่มต้นจากการฝึกฝนตัวแบบด้วยข้อมูลสอนทั้งหมดเพื่อหาชุดของพารามิเตอร์ที่เหมาะสมก่อน หลังจากนั้นจะเข้าสู่ลูปการวิเคราะห์ตัวแบบ ในแต่ละรอบจะทำบูทสแตรปบนข้อมูลสอน คือสุ่มแถวข้อมูลแบบซ้ำกันได้ในกลุ่มของตัวเองและฝึกฝนตัวแบบด้วยข้อมูลบูทสแตรปนั้น หลังจากนั้นจะวัดประสิทธิภาพของตัวแบบที่ได้บนทั้งข้อมูลสอนและข้อมูลทดสอบ จึงจะถือว่าจบหนึ่งลูปการวิเคราะห์ เมื่อวิเคราะห์ครบทั้งหมด 100 รอบ จึงจะสรุปตัววัดผลด้วยวิธีการหาค่าเฉลี่ยจากตัววัดผลทั้งหมด



รูปที่ 3.10 กระบวนการวิเคราะห์สำหรับตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่

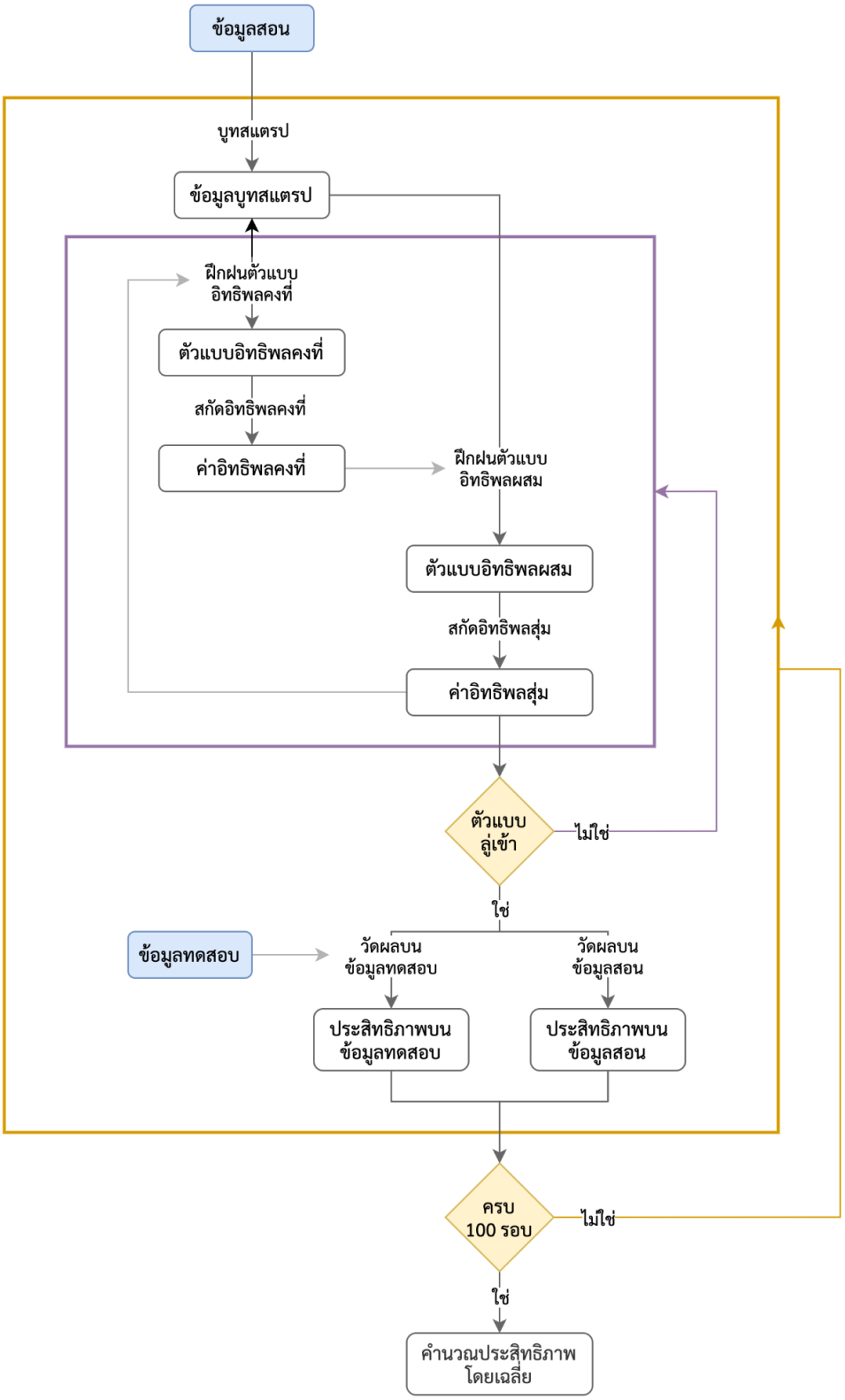
กระบวนการวิเคราะห์สำหรับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมแสดงดังรูปที่ 3.11 ซึ่งจะวิเคราะห์ทั้งหมด 100 รอบ และแต่ละรอบทำบูทสแตรปบนข้อมูลสอนเช่นกัน ในการวิเคราะห์ตัวแบบอิทธิพลผสม จะเริ่มต้นจากการฝึกฝนตัวแบบอิทธิพลคงที่บนทุกตัวแปรและตัวแปรค่าอิทธิพลสุ่ม โดยใช้ชุดของพารามิเตอร์ที่เหมาะสมชุดเดียวกันกับที่หาได้ หลังจากนั้นจะใช้แพ็กเกจ inTrees สำหรับตัวแบบการสุ่มต้นไม้ และแพ็กเกจ treeshap สำหรับตัวแบบ CatBoost ในการสกัดเส้นทางทำนายจากตัวแบบอิทธิพลคงที่

เส้นทางทำนายที่สกัดมาได้จะถูกนำไปใช้เป็นตัวแปรอิทธิพลคงที่ในการวิเคราะห์ตัวแบบอิทธิพลผสมในรูปแบบดังนี้

$$Y \sim Treecondition + (rand.var1 + rand.var2 + \cdots | group)$$

เมื่อ Y คือสถานะเหตุการณ์ที่ต้องการพยากรณ์ Treecondition คือตัวแปรกลุ่มเส้นทางทำนายจากตัวแบบอิทธิพลคงที่ rand.var คือตัวแปรอิทธิพลผสม และ group คือตัวแปรที่กำหนดกลุ่ม หลังจากนั้นจะสกัดค่าอิทธิพลสุ่มจากตัวแบบอิทธิพลผสมด้วยแพ็กเกจ lme4 ค่าอิทธิพลสุ่มที่ได้นี้จะถูกใช้เป็นตัวแปรอิทธิพลสุ่มในการฝึกฝนตัวแบบอิทธิพลคงที่ ลูปการฝึกฝนตัวแบบอิทธิพลผสมนี้จะสิ้นสุดเมื่อตัวแบบอิทธิพลผสมลู่เข้า หรือครบจำนวนรอบสูงสุดที่ตั้งไว้

ในแต่ละลูปการวิเคราะห์จะวัดประสิทธิภาพของตัวแบบอิทธิพลผสมบนทั้งข้อมูลสอนและข้อมูลทดสอบ และเมื่อวิเคราะห์ครบ 100 รอบจึงจะสรุปตัววัดผลด้วยวิธีการหาค่าเฉลี่ยจากตัววัดผลทั้งหมด



รูปที่ 3.11 กระบวนการวิเคราะห์สำหรับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสม

3.3.3 ขั้นตอนการเปรียบเทียบผลลัพธ์

จากขั้นตอนการวิเคราะห์และสร้างตัวแบบ จะให้ผลลัพธ์การพยากรณ์ความเสี่ยงต่อการไม่รอดชีพแบบทวิซึ่งมีผลลัพธ์ที่เป็นไปได้เพียง 2 ค่า คือ เกิดเหตุการณ์ และไม่เกิดเหตุการณ์ จึงสามารถใช้ตัววัดผลสำหรับวิธีการจำแนกแบบทวิทั่วไปได้

ตัววัดผลที่จะใช้ในการเปรียบเทียบประสิทธิภาพของตัวแบบ ได้แก่ ความแม่นยำ (Accuracy) ความเที่ยงตรง (Precision หรือ Positive Predictive Value: PPV) ความไว (Sensitivity หรือ Recall หรือ True Positive Rate: TPR) ความจำเพาะ (Specificity หรือ Selectibity หรือ True Negative Rate: TNR) 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)

3.3.3.1 ความแม่นยำ ความเที่ยงตรง ความไว ความจำเพาะ และคะแนน F1

ผลลัพธ์การพยากรณ์ความเสี่ยงต่อการไม่รอดชีพแบบทวิสามารถแสดงในรูปแบบของ Confusion Matrix ได้ดังรูปที่ 3.12

		สถานะเหตุการณ์ที่พยากรณ์	
		Positive (ไม่รอดชีพ)	**Negative (รอดชีพ)**
สถานะเหตุการณ์จริง	**Positive (ไม่รอดชีพ)**	True Positive (TP)	True Negative (TN)
	Negative (รอดชีพ)	False Positive (FP)	False Negative (FN)

รูปที่ 3.12 ผลลัพธ์การพยากรณ์ความเสี่ยงต่อการไม่รอดชีพแบบทวิในรูปแบบของ Confusion Matrix

จาก Confusion Matrix สามารถเรียกผลลัพธ์ของการพยากรณ์ได้ดังนี้ True Positive (TP) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่าไม่รอดชีพ และพยากรณ์ได้ถูกต้อง False Positive (FP) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่าไม่รอดชีพ แต่พยากรณ์ผิดพลาด True Negative (TN) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่ารอดชีพ และพยากรณ์ได้ถูกต้อง และ False Negative (FN) คือ ผลลัพธ์เมื่อตัวแบบพยากรณ์ว่ารอดชีพ แต่พยากรณ์ผิดพลาด

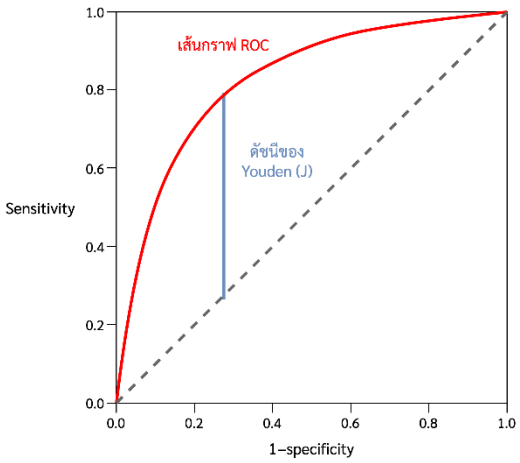
ผลลัพธ์ของการพยากรณ์สามารถนำไปคำนวณตัววัดผลได้ดังนี้

1. ความแม่นยำ $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ คือ อัตราส่วนของการพยากรณ์ที่ถูกต้องเทียบกับการพยากรณ์ทั้งหมด
2. ความเที่ยงตรง $Precision = \frac{TP}{TP+FP}$ 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 ค่าเฉลี่ยฮาร์โมนิก ระหว่างตัววัดความเที่ยงตรงและความไว

เนื่องจากตัววัดผลเหล่านี้มีค่าขึ้นกับจุดตัด จึงเปรียบเทียบโดยใช้วิธีเลือกจุดตัด 2 แบบ แบบแรกคือจุดตัดจากดัชนีของ Youden (Youden's index หรือ Youden’s J Statistic: J) สามารถคำนวณค่าดัชนีของ Youden ได้ดังนี้

$$J = sensitivity + specificity - 1 = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1$$

จุดตัดจากดัชนีของ Youden เป็นจุดตัดที่ให้ระยะห่างระหว่างจุด (sensitivity, 1-specificity) บนเส้นกราฟ ROC กับเส้นทแยงสูงที่สุด แสดงดังรูปที่ 3.13 วิธีนี้เป็นวิธีให้สมดุลระหว่างการระบุเหตุการณ์รอดชีพและไม่รอดชีพอย่างถูกต้องสูงที่สุด



รูปที่ 3.13 ตัวอย่างจุดตัดจากดัชนีของ Youden บนเส้นกราฟ ROC เมื่อเส้นประคือเส้นทแยงมุมและเส้นตั้งตรง J คือจุดที่ให้ค่าดัชนีของ Youden สูงที่สุด

อีกแบบหนึ่งจะใช้จุดตัดจากอัตราส่วนของจำนวนแถวข้อมูลสอนที่เกิดเหตุการณ์เทียบกับจำนวนแถวข้อมูลสอนทั้งหมด คือ ค่า 0.41 สำหรับข้อมูลท่อน้ำดีอักเสบปฐมภูมิ และค่า 0.04 สำหรับข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน

3.3.3.2 R-squared จากคะแนน Brier

คะแนน Brier เป็นค่าความคลาดเคลื่อนในการพยากรณ์เมื่อเทียบกับสถานะเหตุการณ์จริง สามารถแสดงได้ดังนี้

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

เมื่อ N คือจำนวนข้อมูลสังเกตทั้งหมด $D(t)$ คือสถานะเหตุการณ์จริงภายในเวลา t และ $p(t)$ คือความน่าจะเป็นที่พยากรณ์ว่าจะเกิดเหตุการณ์ภายในเวลา t ทั้งนี้ คะแนน Brier จะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยคะแนน Brier ที่มีค่าน้อยจะแสดงถึงประสิทธิภาพการพยากรณ์ของตัวแบบที่ดี

คะแนน Brier สามารถนำไปคำนวณเป็นตัววัด R-squared ที่แสดงถึงค่าความผันแปรของเศษเหลือที่สามารถอธิบายได้ดังนี้

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

เมื่อ $BS_0(t)$ คือคะแนน Brier ของตัวแบบตั้งต้นที่ให้ความน่าจะเป็นที่พยากรณ์ว่าจะเกิดเหตุการณ์เหมือนกันในทุกข้อมูลสังเกต หรือก็คือ $BS_0(t)$ จะเป็นคะแนน Brier สูงสุดที่เป็นไปได้ของตัวแบบและชุดข้อมูลสังเกตนี้ ดังนั้น R-squared ที่มีค่ามากจะแสดงถึงประสิทธิภาพการพยากรณ์ของตัวแบบที่ดี

3.3.3.3 พื้นที่ใต้กราฟ ROC และพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลา

กราฟ ROC เป็นตัววัดผลการพยากรณ์ที่แสดงในรูปแบบของกราฟความสัมพันธ์ระหว่างความไวและความจำเพาะบนช่วงของเกณฑ์การพยากรณ์ พื้นที่ใต้กราฟ ROC สามารถบ่งบอกได้ว่าตัวแบบมีความสามารถในการแยกแยะความแตกต่างระหว่างกลุ่มได้ดีเพียงใด จะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยที่พื้นที่ใต้กราฟ ROC ที่มีค่ามากแสดงถึงความสามารถในการแยกแยะความแตกต่างระหว่างกลุ่มได้ดี

การวัดผลด้วยพื้นที่ใต้กราฟ ROC แบบปกติ จะทำให้ทราบประสิทธิภาพการพยากรณ์ของตัวแบบโดยรวม อย่างไรก็ตาม ในการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่อง สถานะเหตุการณ์ของแต่ละบุคคลจะเปลี่ยนไปตามช่วงเวลาหรือครั้งที่ติดตาม ดังนั้นจึงจะวัดพื้นที่ใต้กราฟ ROC แบ่งตามครั้งที่ติดตามด้วย โดยข้อมูลท่อน้ำดีอักเสบปฐมภูมิมีทั้งหมด 10 ครั้งติดตาม และข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานมีทั้งหมด 6 ครั้งติดตาม ตัววัดผลนี้จะทำให้ทราบถึงแนวโน้มประสิทธิภาพการพยากรณ์ของตัวแบบเมื่อจำนวนแถวข้อมูลในกลุ่มเพิ่มขึ้น หรือจำนวนครั้งที่ติดตามที่เพิ่มขึ้น

บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษาการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่อง โดยเปรียบเทียบผลลัพธ์การพยากรณ์ความเสี่ยงต่อการไม่รอดชีพในแต่ละช่วงเวลา ระหว่างการวิเคราะห์โดยละเลยความสัมพันธ์ของข้อมูลระหว่างบุคคลเดียวกัน และสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิงในการวิเคราะห์ ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม ซึ่งเป็นตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่ และการวิเคราะห์โดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันในข้อมูลตามยาว ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost, โครงข่ายประสาทเทียม ซึ่งเป็นตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่ และการเรียนรู้ของเครื่องอิทธิพลผสมโดยพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันด้วยตัวแปรอิทธิพลสุ่ม

ในการเปรียบเทียบประสิทธิภาพของตัวแบบจะใช้ตัววัดผล ได้แก่ ความแม่นยำ, ความเที่ยงตรง, ความไว, ความจำเพาะ, คะแนน F1, R-squared จากคะแนน Brier พื้นที่ใต้กราฟ ROC และพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลา โดยตัววัดผลที่ค่าขึ้นกับจุดตัดจะใช้จุดตัดจากดัชนีของ Youden และจุดตัดจากอัตราส่วนของจำนวนแถวข้อมูลสอนที่เกิดเหตุการณ์เทียบกับจำนวนแถวข้อมูลสอนทั้งหมด

ผลการวิจัยจะแบ่งเป็น 2 ส่วนตามชุดข้อมูลที่ศึกษา ได้แก่ ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ และข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน

4.1 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลท่อน้ำดีอักเสบปฐมภูมิ

ผลการวิเคราะห์จะแสดงแยกตามตัวแบบที่ศึกษา ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม

4.1.1 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลท่อน้ำดีอักเสบปฐมภูมิด้วยตัวแบบการสุ่มป่าไม้

จากผลการวิเคราะห์แสดงดังตารางที่ 4.1 พบว่าสำหรับตัวแบบการสุ่มป่าไม้ที่พิจารณาเฉพาะอิทธิพลคงที่ การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ผลลัพธ์ที่แย่ลง โดยทุกตัววัดผลให้ผลลัพธ์ไปในทางเดียวกัน

สำหรับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากตัวแบบการสุ่มป่าไม้ ได้ศึกษาโดยใช้ตัวแปร bili, ast, albumin และ protime เป็นตัวแปรอิทธิพลสุ่ม อย่างไรก็ตาม พบว่าตัวแบบที่พิจารณาอิทธิพลผสมให้ผลลัพธ์ที่แย่กว่าตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่ โดยทุกตัววัดผลให้ผลลัพธ์ไปในทางเดียวกัน

ในแง่ของจุดตัด เนื่องจากทั้งสองจุดตัดให้ผลลัพธ์ที่ใกล้เคียงกัน จึงยังไม่สามารถสรุปได้ว่าจุดตัดแบบใดเหมาะสมกว่าสำหรับตัวแบบการสุ่มป่าไม้

ตารางที่ 4.1 ผลการวิเคราะห์บนข้อมูลท่อน้ำดีอักเสบปฐมภูมิด้วยตัวแบบการสุ่มป่าไม้

อิทธิพล	ความสัมพันธ์	จุดตัด	ความแม่นยำ	ความเที่ยงตรง	ความไว	ความจำเพาะ	คะแนน F1	R-squared	AUC
คงที่	✗	Y	0.7678*	0.6785*	0.7866	0.7556*	0.7272	0.8422*	0.8422*
		0.41	0.7674	0.6728	0.7982	0.7473	0.7300*		
	✓	Y	0.7341	0.6313	0.8007*	0.6907	0.7040	0.8252	0.8252
		0.41	0.7406	0.6392	0.7863	0.7109	0.7050		
ผสม	✓	Y	0.5539	0.4341	0.4352	0.6311	0.4344	0.5083	0.5083
		0.41	0.5525	0.4291	0.4094	0.6456	0.4189		

หมายเหตุ: “Y” แทนจุดตัดจากดัชนีของ Youden และ “*” แสดงตัววัดผลที่ดีที่สุด

4.1.2 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลท่อน้ำดีอักเสบปฐมภูมิด้วยตัวแบบ CatBoost

จากผลการวิเคราะห์แสดงดังตารางที่ 4.2 พบว่าสำหรับตัวแบบ CatBoost ที่พิจารณาเฉพาะอิทธิพลคงที่ การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ผลลัพธ์ที่ดีขึ้น โดยทุกตัววัดผลให้ผลลัพธ์ไปในทางเดียวกัน

สำหรับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากตัวแบบ CatBoost ได้ศึกษาโดยใช้ตัวแปร bili, ast, albumin และ protime เป็นตัวแปรอิทธิพลสุ่ม อย่างไรก็ตาม พบว่าตัวแบบที่พิจารณาอิทธิพลผสมให้ผลลัพธ์ที่แย่กว่าตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่ โดยทุกตัววัดผลให้ผลลัพธ์ไปในทางเดียวกัน

ในแง่ของจุดตัด พบว่าการใช้จุดตัดจากดัชนีของ Youden ให้ผลลัพธ์ที่ดีกว่า โดยจุดตัดอยู่ในช่วง 0.3-0.5

ตารางที่ 4.2 ผลการวิเคราะห์บนข้อมูลท่อน้ำดีอักเสบปฐมภูมิด้วยตัวแบบ CatBoost

อิทธิพล	ความสัมพันธ์	จุดตัด	ความแม่นยำ	ความเที่ยงตรง	ความไว	ความจำเพาะ	คะแนน F1	R-squared	AUC
คงที่	✗	Y	0.7707	0.6865	0.7775	0.7662	0.7274	0.7414	0.8476
		0.41	0.7677	0.6703	0.809	0.7408	0.7330		
	✓	Y	0.8181*	0.7631*	0.7815	0.8420*	0.7720	0.7903*	0.8992*
		0.41	0.8179	0.7319	0.8496*	0.7974	0.7862*		
ผสม	✓	Y	0.5466	0.4251	0.4283	0.6236	0.4267	0.2632	0.5339
		0.41	0.5467	0.4298	0.4195	0.6294	0.4188		

หมายเหตุ: “Y” แทนจุดตัดจากดัชนีของ Youden และ “*” แสดงตัววัดผลที่ดีที่สุด

4.1.3 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลท่อน้ำดีอักเสบปฐมภูมิด้วยตัวแบบโครงข่ายประสาทเทียม

จากผลการวิเคราะห์แสดงดังตารางที่ 4.3 พบว่าสำหรับตัวแบบโครงข่ายประสาทเทียมที่พิจารณาเฉพาะอิทธิพลคงที่ ทั้งตัวแบบที่พิจารณาและไม่พิจารณาความสัมพันธ์ของข้อมูลให้ผลลัพธ์ที่แข่งขันกันได้ ขึ้นอยู่กับว่าเปรียบเทียบด้วยตัววัดผลใด อย่างไรก็ตาม หากพิจารณาด้วยตัววัดพื้นที่ใต้กราฟ ROC ซึ่งวัดประสิทธิภาพโดยรวมในหลายจุดตัด พบว่าการพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ผลลัพธ์ที่ดีขึ้น

ในแง่ของจุดตัด พบว่าการใช้จุดตัดจากอัตราส่วนของจำนวนแถวข้อมูลสอนที่เกิดเหตุการณ์เทียบกับจำนวนแถวข้อมูลสอนทั้งหมดให้ผลลัพธ์ที่ดีกว่า

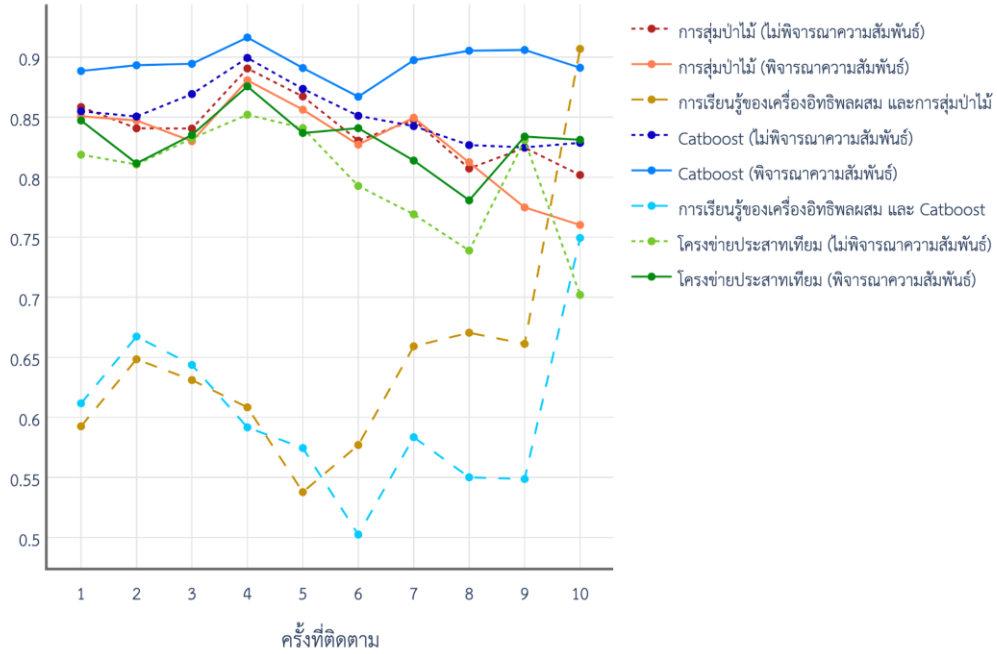
อย่างไรก็ตาม ผู้วิจัยพบว่ามีข้อจำกัดในการวิเคราะห์ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากตัวแบบโครงข่ายประสาทเทียม โดยจะอภิปรายในส่วนถัดไป

ตารางที่ 4.3 ผลการวิเคราะห์บนข้อมูลท่อน้ำดีอักเสบปฐมภูมิด้วยตัวแบบโครงข่ายประสาทเทียม

อิทธิพล	ความสัมพันธ์	จุดตัด	ความแม่นยำ	ความเที่ยงตรง	ความไว	ความจำเพาะ	คะแนน F1	R-squared	AUC
คงที่	X	Y	0.7291	0.6296	0.7620	0.7077	0.6892	0.7141*	0.8058
		0.41	0.7355*	0.6416*	0.7460	0.7287*	0.6898*		
	✓	Y	0.7089	0.5981	0.8117*	0.6421	0.6874	0.7103	0.8277*

หมายเหตุ: “Y” แทนจุดตัดจากดัชนีของ Youden และ “*” แสดงตัววัดผลที่ดีที่สุด

นอกจากนี้ หากพิจารณาตัววัดพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลาสำหรับทุกตัวแบบซึ่งแสดงดังรูปที่ 4.1 พบว่า สำหรับตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่ (เส้นจุดและเส้นทึบ) ประสิทธิภาพการพยากรณ์มีแนวโน้มค่อยๆ ลดลงเรื่อยๆ เมื่อครั้งที่ติดตามเพิ่มขึ้น ในขณะที่สำหรับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสม (เส้นประ) ประสิทธิภาพการพยากรณ์มีแนวโน้มลดลงในช่วงครั้งติดตามที่ 1 ถึง 5 และ 6 หลังจากนั้นจึงเปลี่ยนเป็นแนวโน้มเพิ่มขึ้นเรื่อยๆ เมื่อครั้งที่ติดตามเพิ่มขึ้น



รูปที่ 4.1 กราฟเส้นแสดงผลการวิเคราะห์พื้นที่ใต้กราฟ ROC
ที่ขึ้นกับเวลาบนข้อมูลท่อน้ำดีอักเสบปฐมภูมิ โดยแกนนอนคือครั้งที่ติดตาม
และแกนตั้งคือพื้นที่ใต้กราฟ ROC

4.2 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน

ผลการวิเคราะห์จะแสดงแยกตามตัวแบบที่ศึกษา ประกอบไปด้วยตัวแบบการสุ่มป่าไม้, CatBoost และโครงข่ายประสาทเทียม

4.2.1 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานตัวแบบการสุ่มป่าไม้

จากผลการวิเคราะห์แสดงดังตารางที่ 4.4 พบว่าสำหรับตัวแบบการสุ่มป่าไม้ที่พิจารณาเฉพาะอิทธิพลคงที่ การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ผลลัพธ์ที่แย่ลง เมื่อพิจารณาด้วยตัววัดพื้นที่ใต้กราฟ ROC

สำหรับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากตัวแบบการสุ่มป่าไม้ได้ศึกษาโดยใช้ตัวแปร BMI, SBP และ DBP เป็นตัวแปรอิทธิพลสุ่ม โดยพบว่าตัวแบบที่พิจารณาอิทธิพลผสมให้ผลลัพธ์ที่ดีกว่าตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่ เมื่อพิจารณาด้วยตัววัดพื้นที่ใต้กราฟ ROC

ในแง่ของจุดตัด เนื่องจากทั้งสองจุดตัดให้ผลลัพธ์ที่ใกล้เคียงกัน จึงยังไม่สามารถสรุปได้ว่าจุดตัดแบบใดเหมาะสมกว่าสำหรับตัวแบบการสุ่มป่าไม้

ตารางที่ 4.4 ผลการวิเคราะห์บนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานด้วยตัวแบบการสุ่มป่าไม้

อิทธิพล	ความสัมพันธ์	จุดตัด	ความแม่นยำ	ความเที่ยงตรง	ความไว	ความจำเพาะ	คะแนน F1	R-squared	AUC
คงที่	✗	Y	0.8294	0.0508	0.1842*	0.8562	0.0779*	0.9593*	0.5564
		0.04	0.9236	0.0617*	0.0622	0.9594	0.0598		
	✓	Y	0.8816	0.0279	0.0578	0.9158	0.0371	0.9584	0.4897
		0.04	0.9337*	0.0251	0.0158	0.9719*	0.0274		
ผสม	✓	Y	0.9052	0.0478	0.0698	0.9400	0.0540	0.9481	0.5608*
		0.04	0.9098	0.0468	0.0635	0.9450	0.0532		

หมายเหตุ: “Y” แทนจุดตัดจากดัชนีของ Youden และ “*” แสดงตัววัดผลที่ดีที่สุด

4.2.2 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานตัวแบบ CatBoost

จากผลการวิเคราะห์แสดงดังตารางที่ 4.5 พบว่าสำหรับตัวแบบ CatBoost ที่พิจารณาเฉพาะอิทธิพลคงที่ การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ผลลัพธ์ที่ดีขึ้น โดยทุกตัววัดผลให้ผลลัพธ์ไปในทางเดียวกัน

สำหรับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากตัวแบบ CatBoost ได้ศึกษาโดยใช้ตัวแปร SBP และ DBP เป็นตัวแปรอิทธิพลสุ่ม อย่างไรก็ตาม พบว่าตัวแบบที่พิจารณาอิทธิพลผสมให้ผลลัพธ์ที่แย่กว่าตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่ เมื่อพิจารณาด้วยตัววัดพื้นที่ใต้กราฟ ROC

ในแง่ของจุดตัด เนื่องจากทั้งสองจุดตัดให้ผลลัพธ์ที่ใกล้เคียงกัน จึงยังไม่สามารถสรุปได้ว่าจุดตัดแบบใดเหมาะสมกว่าสำหรับตัวแบบ CatBoost

ตารางที่ 4.5 ผลการวิเคราะห์บนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานด้วยตัวแบบ CatBoost

อิทธิพล	ความสัมพันธ์	จุดตัด	ความแม่นยำ	ความเที่ยงตรง	ความไว	ความจำเพาะ	คะแนน F1	R-squared	AUC
คงที่	✗	Y	0.7816	0.0552	0.2742	0.8027	0.0909	0.9599	0.5824
		0.04	0.7309	0.0526	0.3378	0.7473	0.0907		
	✓	Y	0.6300	0.0548	0.5067*	0.6351	0.0986*	0.9601*	0.6051*
		0.04	0.8025	0.0615*	0.2663	0.8248	0.0957		
ผสม	✓	Y	0.8956*	0.0449	0.0772	0.9297*	0.0507	0.9365	0.5299
		0.04	0.8483	0.0476	0.1385	0.8779	0.0568		

หมายเหตุ: “Y” แทนจุดตัดจากดัชนีของ Youden และ “*” แสดงตัววัดผลที่ดีที่สุด

4.2.3 ผลการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานตัวแบบโครงข่ายประสาทเทียม

จากผลการวิเคราะห์แสดงดังตารางที่ 4.6 พบว่าสำหรับตัวแบบโครงข่ายประสาทเทียมที่พิจารณาเฉพาะอิทธิพลคงที่ การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันให้ผลลัพธ์ที่แย่ลง เมื่อพิจารณาด้วยตัววัดพื้นที่ใต้กราฟ ROC

ในแง่ของจุดตัด พบว่าการใช้จุดตัดจากอัตราส่วนของจำนวนแถวข้อมูลสอนที่เกิดเหตุการณ์เทียบกับจำนวนแถวข้อมูลสอนทั้งหมดให้ผลลัพธ์ที่ดีกว่า

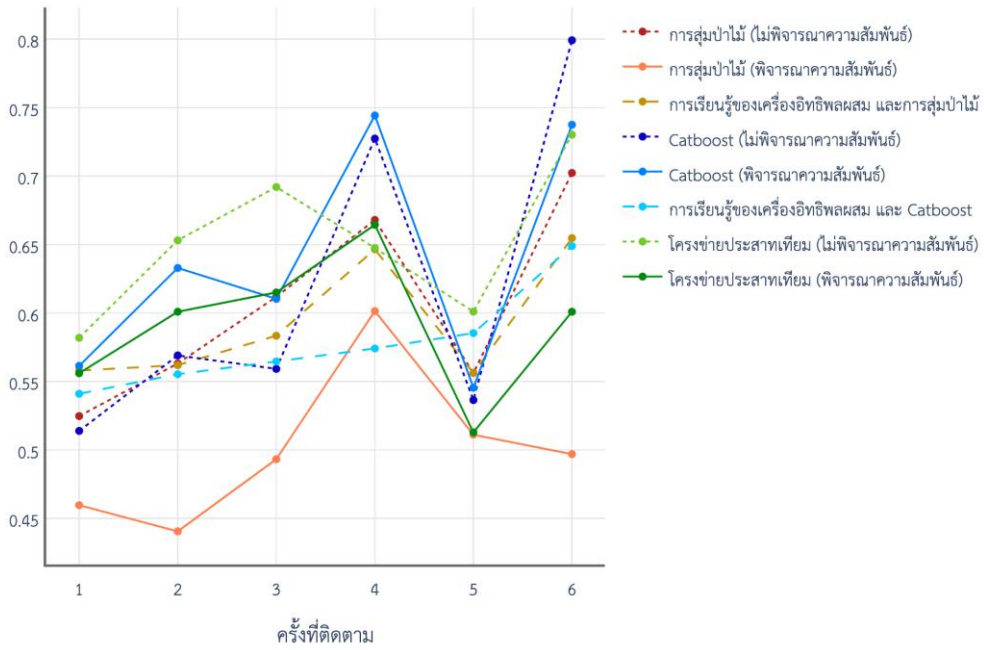
อย่างไรก็ตาม ผู้วิจัยพบว่ามีข้อจำกัดในการวิเคราะห์ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมโดยใช้อิทธิพลคงที่จากตัวแบบโครงข่ายประสาทเทียม โดยจะอภิปรายในส่วนถัดไป

ตารางที่ 4.6 ผลการวิเคราะห์บนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานด้วยตัวแบบโครงข่ายประสาทเทียม

อิทธิพล	ความสัมพันธ์	จุดตัด	ความแม่นยำ	ความเที่ยงตรง	ความไว	ความจำเพาะ	คะแนน F1	R-squared	AUC
คงที่	✗	Y	0.6835	0.0613	0.4652*	0.6926	0.1063	0.9601*	0.6209*
		0.04	0.7461	0.0631*	0.3752	0.7616	0.1067*		
	✓	Y	0.7514	0.0475	0.2812	0.7709	0.0803	0.9594	0.5887
		0.04	0.7957*	0.0479	0.2197	0.8196*	0.0770		

หมายเหตุ: “Y” แทนจุดตัดจากดัชนีของ Youden และ “*” แสดงตัววัดผลที่ดีที่สุด

นอกจากนี้ หากพิจารณาตัววัดพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลาสำหรับทุกตัวแบบซึ่งแสดงดังรูปที่ 4.2 พบว่า ทุกตัวแบบมีแนวโน้มไปในทิศทางเดียวกัน คือประสิทธิภาพการพยากรณ์มีแนวโน้มเพิ่มขึ้นเมื่อครั้งที่ติดตามเพิ่มขึ้น โดยสังเกตว่า ณ ครั้งติดตามที่ 5 ทุกตัวแบบจะมีประสิทธิภาพการพยากรณ์ต่ำ ซึ่งเป็นผลมาจากจำนวนแถวข้อมูลทดสอบที่มีน้อยลงเรื่อยๆ เมื่อจำนวนครั้งที่ติดตามเพิ่มขึ้น ดังนั้น หากเกิดการพยากรณ์ผิดพลาดจะส่งผลต่อตัววัดผลสูง อาจกล่าวได้ว่า ตัววัดพื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลามีความน่าเชื่อถือน้อยลงในครั้งที่ติดตามท้ายๆ



รูปที่ 4.2 กราฟเส้นแสดงผลการวิเคราะห์พื้นที่ใต้กราฟ ROC ที่ขึ้นกับเวลาบนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน โดยแกนนอนคือครั้งที่ติดตาม และแกนตั้งคือพื้นที่ใต้กราฟ ROC

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้ศึกษาการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องบนตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่ และตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมที่พิจารณาทั้งอิทธิพลแบบคงที่และอิทธิพลแบบสุ่ม โดยเปรียบเทียบทั้งแบบพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและแบบละเลยความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันและสมมติว่าข้อมูลแต่ละแถวเป็นอิสระกันอย่างสิ้นเชิง วิเคราะห์บนข้อมูล 2 ชุด คือ ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ และข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานของประชากรไทยซึ่งเป็นข้อมูลจริง

5.1 สรุปผลการวิจัย

จากการศึกษาพบว่าข้อมูลทั้ง 2 ชุดให้ผลลัพธ์การพยากรณ์ที่แตกต่างกัน ตารางที่ 5.1 สรุปผลลัพธ์การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันบนตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่ พบว่า มีเพียงตัวแบบ CatBoost ที่การพิจารณาความสัมพันธ์ให้ผลลัพธ์ที่ดีขึ้นบนทั้ง 2 ชุดข้อมูล ในขณะที่ตัวแบบการสุ่มป่าไม้ให้ผลลัพธ์ที่แย่ลงบนทั้ง 2 ชุดข้อมูล และตัวแบบโครงข่ายประสาทเทียมให้ผลลัพธ์ที่ดีขึ้นเฉพาะบนข้อมูลท่อน้ำดีอักเสบปฐมภูมิ

ตารางที่ 5.1 สรุปผลลัพธ์การพิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกันบนตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่

ตัวแบบ	ข้อมูล	
	ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ	ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน
การสุ่มป่าไม้	-	-
CatBoost	การพิจารณาความสัมพันธ์ให้ผลลัพธ์ดีขึ้น	การพิจารณาความสัมพันธ์ให้ผลลัพธ์ดีขึ้น
โครงข่ายประสาทเทียม	การพิจารณาความสัมพันธ์ให้ผลลัพธ์ดีขึ้น	-

สำหรับผลลัพธ์การพิจารณาอิทธิพลผสมเมื่อเทียบกับตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่แสดงดังตารางที่ 5.2 มีเพียงตัวแบบอิทธิพลผสมโดยใช้อิทธิพลคงที่จากตัวแบบการสุ่มป่าไม้บน

ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานที่ให้ประสิทธิภาพการพยากรณ์ที่ดีขึ้นเมื่อเทียบกับตัวแบบการสุ่มป่าไม้ที่พิจารณาเฉพาะอิทธิพลคงที่

ตารางที่ 5.2 สรุปผลลัพธ์การพิจารณาอิทธิพลผสมเมื่อเทียบกับตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่

ตัวแบบ	**ข้อมูล**	
	ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ	**ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน**
การสุ่มป่าไม้	-	การพิจารณาอิทธิพลผสมให้ผลลัพธ์ดีขึ้น
CatBoost	-	-

โดยสรุป การพิจารณาความสัมพันธ์ระหว่างบุคคลคนเดียวกันบนตัวแบบที่พิจารณาเฉพาะอิทธิพลแบบคงที่ไม่ได้ให้ประสิทธิภาพการพยากรณ์ที่ดีขึ้นเสมอไป และการพิจารณาอิทธิพลผสมเมื่อเทียบกับตัวแบบอิทธิพลคงที่ไม่ได้ให้ประสิทธิภาพการพยากรณ์ที่ดีขึ้นเสมอไป ขึ้นอยู่กับตัวแบบที่เลือกใช้และลักษณะของข้อมูลด้วย

ในแง่ของการเลือกจุดตัด พบว่าแต่ละตัวแบบอาจจะเหมาะสมกับวิธีการเลือกจุดตัดที่แตกต่างกัน ในที่นี้พิจารณาว่าจุดตัดที่เหมาะสมคือจุดตัดที่ให้ตัววัดความแม่นยำ ความเที่ยงตรง และคะแนน F1 ที่สูงกว่า ตารางที่ 5.3 แสดงผลสรุปวิธีการเลือกจุดตัดที่เหมาะสมในแต่ละตัวแบบ สำหรับตัวแบบการสุ่มป่าไม้ไม่สามารถสรุปวิธีการเลือกจุดตัดที่เหมาะสมได้ เนื่องจากทั้งสองวิธีให้ผลลัพธ์ค่อนข้างใกล้เคียงกัน สำหรับตัวแบบ CatBoost จุดตัดจากดัชนีของ Youden ให้ผลลัพธ์ที่ดีกว่าบนข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน แต่ไม่สามารถสรุปวิธีการเลือกจุดตัดที่เหมาะสมได้บนข้อมูลท่อน้ำดีอักเสบปฐมภูมิ และสำหรับตัวแบบโครงข่ายประสาทเทียม จุดตัดจากอัตราส่วนของจำนวนแถวข้อมูลสอนที่เกิดเหตุการณ์เทียบกับจำนวนแถวข้อมูลสอนทั้งหมดให้ผลลัพธ์ที่ดีกว่าบนทั้ง 2 ชุดข้อมูล

ตารางที่ 5.3 ผลสรุปวิธีการเลือกจุดตัดที่เหมาะสม

ตัวแบบ	ข้อมูล	
	ข้อมูลท่อน้ำดีอักเสบปฐมภูมิ	**ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวาน**
การสุ่มป่าไม้	ไม่สามารถสรุปได้	ไม่สามารถสรุปได้
CatBoost	ไม่สามารถสรุปได้	จุดตัดจากดัชนีของ Youden
โครงข่ายประสาทเทียม	จุดตัดจากอัตราส่วนของจำนวนแถวข้อมูลสอนที่เกิดเหตุการณ์เทียบกับจำนวนแถวข้อมูลสอนทั้งหมด	จุดตัดจากอัตราส่วนของจำนวนแถวข้อมูลสอนที่เกิดเหตุการณ์เทียบกับจำนวนแถวข้อมูลสอนทั้งหมด

5.2 อภิปรายผลการวิจัย

การวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องที่มีลักษณะเป็นข้อมูลตามยาวด้วยตัวแบบที่ผสมผสานระหว่างอิทธิพลคงที่และอิทธิพลสุ่มนั้นมีความน่าสนใจ และมีความสมเหตุสมผลในเชิงการนำเอาผลลัพธ์ไปประยุกต์ใช้จริง โดยเฉพาะกับการวิเคราะห์บนข้อมูลทางการแพทย์หรือสุขภาพ อย่างไรก็ตาม จากการศึกษาบนชุดข้อมูลสุขภาพทั้ง 2 ชุดในงานวิจัยนี้ พบว่าผลลัพธ์ที่ได้แตกต่างกันในทุกตัวแบบ ทำให้ไม่สามารถสรุปผลอย่างทั่วไปได้ชัดเจน ทั้งนี้ ผู้วิจัยคาดว่าขึ้นอยู่กับหลากหลายปัจจัย เช่น ข้อมูล ตัวแบบที่เลือกใช้ การปรับพารามิเตอร์ของตัวแบบ การกำหนดตัวแปรอิทธิพลสุ่ม และวิธีการสกัดสิ่งสำคัญจากตัวแบบอิทธิพลคงที่ซึ่งเป็นปัจจัยสำคัญที่ส่งผลต่อการวิเคราะห์ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสม

งานวิจัยนี้ได้ศึกษาวิธีการสกัดอิทธิพลคงที่จากตัวแบบการสุ่มป่าไม้ที่มีพื้นฐานเป็นต้นไม้ตัดสินใจ (Ngufor et al., 2019) โดยเริ่มจากการสกัดกฎการตัดสินใจจากต้นไม้ทุกต้น เลือกชุดของกฎการตัดสินใจที่ให้ Empty Loss น้อยที่สุดเพื่อนำมาสร้างเป็นต้นไม้อย่างง่ายเพียงต้นเดียว แล้วจึงสกัดเส้นทางทำนายจากต้นไม้อย่างง่ายนั้นว่าพยากรณ์ไปที่โหนดสุดท้ายโหนดใด เพื่อนำเส้นทางทำนายนั้นไปใช้เป็นตัวแปรกลุ่ม

ผู้วิจัยได้นำวิธีการสกัดเส้นทางทำนายลักษณะเดียวกันนี้ไปประยุกต์ใช้กับตัวแบบ CatBoost ซึ่งมีพื้นฐานเป็นต้นไม้ตัดสินใจเหมือนกัน แต่เนื่องจากวิธีการแปลงข้อมูลกลุ่มของ CatBoost ทำให้ได้กฎการตัดสินใจที่แตกต่างกันในต้นไม้ทุกต้น ขั้นตอนการรวมให้อยู่ในรูปแบบต้นไม้อย่างง่ายจึงทำได้

ยาก อย่างไรก็ตาม ผู้วิจัยได้ตัดสินใจเลือกใช้เส้นทางทำนายจากต้นไม้ต้นสุดท้าย เนื่องจากในหลักการเรียนรู้ของ CatBoost ต้นไม้ต้นถัดไปจะเรียนรู้ข้อผิดพลาดของต้นไม้ก่อนหน้าแล้วปรับปรุงให้ดีขึ้น จึงตั้งสมมติฐานว่าต้นไม้ต้นสุดท้ายเกิดการเรียนรู้ที่ดีที่สุดจากต้นไม้ทั้งหมด

อย่างไรก็ตาม วิธีนี้การสกัดเส้นทางทำนายนั้นมีข้อจำกัดสำหรับตัวแบบโครงข่ายประสาทเทียม เนื่องจากเป็นตัวแบบที่มีความซับซ้อนสูง และการสกัดเส้นทางทำนายด้วยวิธีเดียวกันนั้นเป็นไปได้ยาก จึงได้ละไว้ในงานวิจัยครั้งนี้

จากผลการวิจัย จะเห็นว่าตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมยังให้ผลลัพธ์ที่ไม่ค่อยดี ผู้วิจัยคาดว่าลักษณะข้อมูลก็เป็นปัจจัยหนึ่งที่สำคัญ เช่น ข้อมูลการคัดกรองและผลการคัดกรองโรคเบาหวานที่เป็นข้อมูลจริงที่ถูกเก็บภายในประเทศไทยและป้อนข้อมูลโดยผู้รับผิดชอบ จึงอาจมีความผิดพลาดในการใส่ข้อมูล ข้อมูลสูญหาย ข้อมูลผิดปกติ รวมถึงเป็นข้อมูลที่ขาดความสมดุลสูง สังเกตได้จากตัววัดผลที่ค่าขึ้นกับจุดตัดมีความเอียงไปทางค่าใดค่าหนึ่งสูงมาก

นอกจากนี้ ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมยังมีข้อจำกัดในเชิงข้อมูลที่มีปริมาณน้อย เนื่องจากเป็นการศึกษาอิทธิพลสุ่มของแต่ละกลุ่ม ซึ่งแถวข้อมูลในกลุ่มเดียวกันมีความสัมพันธ์กัน ดังนั้นหากมีจำนวนกลุ่มเยอะ แต่จำนวนแถวข้อมูลในกลุ่มมีน้อยเกินไป ก็จะไม่สามารถกำหนดตัวแปรอิทธิพลสุ่มหลายตัวได้ ในที่นี้ ผู้วิจัยได้พยายามกำหนดตัวแปรสุ่มจำนวนเยอะที่สุดเท่าที่จะเป็นไปได้ตามแต่ละข้อมูลและตัวแบบ

ในแง่ของต้นทุนและเวลาในการคำนวณ พบว่าตัวแบบที่พิจารณาความสัมพันธ์ของข้อมูลระหว่างบุคคลคนเดียวกัน โดยเฉพาะตัวแบบการสุ่มป่าไม้ ใช้เวลาในการเรียนรู้นานขึ้น เนื่องจากมีจำนวนตัวแปรเยอะขึ้น เช่นเดียวกับตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมใช้เวลาในการเรียนรู้นาน เพราะต้องสอนจนกว่าตัวแบบจะลู่เข้า

โดยสรุป งานวิจัยนี้พบว่าการวิเคราะห์ด้วยตัวแบบที่พิจารณาอิทธิพลผสมไม่ได้ให้ประสิทธิภาพการพยากรณ์ที่ดีกว่าตัวแบบที่พิจารณาเฉพาะอิทธิพลคงที่เสมอไป และเมื่อจำนวนแถวข้อมูลในกลุ่มเพิ่มขึ้น หรือจำนวนครั้งที่ติดตามที่เพิ่มขึ้น ตัวแบบการเรียนรู้ของเครื่องอิทธิพลผสมไม่ได้ให้การประสิทธิภาพการพยากรณ์ที่มีแนวโน้มคงที่หรือเพิ่มขึ้นเสมอไป ซึ่งแตกต่างจากผลลัพธ์ของงานวิจัยที่เกี่ยวข้องที่พบว่า การพิจารณาอิทธิพลผสมให้ประสิทธิภาพการพยากรณ์ที่ดีกว่า และมีแนวโน้มคงที่หรือเพิ่มขึ้นเมื่อจำนวนครั้งที่ติดตามที่เพิ่มขึ้น (Ngufor et al., 2019) โดยผู้วิจัยคาดว่าด้วยปัจจัยต่างๆ ดังที่กล่าวข้างต้น จึงทำให้ได้ผลลัพธ์ที่แตกต่างกัน

5.3 ข้อเสนอแนะ

เนื่องจากการศึกษาครั้งนี้มีปัจจัยหลากหลายที่ส่งผลต่อการวิเคราะห์การรอดชีพเวลาไม่ต่อเนื่องดังที่ได้กล่าวไปในส่วนการอภิปรายผลการวิจัย ทางผู้วิจัยได้พยายามควบคุมปัจจัยให้คงที่ในทุกตัวแบบเพื่อความเท่าเทียมและเปรียบเทียบกันได้ อย่างไรก็ตาม ผู้ที่สนใจสามารถศึกษาเพิ่มเติมได้ในเรื่องของวิธีการสกัดสิ่งสำคัญจากตัวแบบอิทธิพลคงที่ โดยเฉพาะตัวแบบ CatBoost และโครงข่ายประสาทเทียม วิธีการเลือกว่าจะกำหนดตัวแปรใดเป็นตัวแปรอิทธิพลสุ่ม รวมถึงอาจศึกษาบนตัวแบบอื่นๆ และข้อมูลการรอดชีพอื่นๆ เพื่อให้เห็นผลลัพธ์โดยทั่วไปได้

รายการอ้างอิง

Azhari, M., Alaoui, A., Achraoui, Z., Ettaki, B., & Zerouaoui, J. (2019). Adaptation of the random forest method: solving the problem of pulsar search. Proceedings of the 4th International Conference on Smart City Applications,

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3), 127-135.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Camuñas-Mesa, L. A., Linares-Barranco, B., & Serrano-Gotarredona, T. (2019). Neuromorphic spiking neural networks and their memristor-CMOS hardware implementations. *Materials*, 12(17), 2745.

Cestnik, B. (1990). Estimating probabilities: A crucial task in machine learning. Proc. 9th European Conf. on Artificial Intelligence,

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:*1810.11363.

Dundar, M., Krishnapuram, B., Bi, J., & Rao, R. B. (2007). Learning classifiers when the training data is not IID. IJCAI,

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.

Kleinbaum, D. G., & Klein, M. (1996). *Survival analysis a self-learning text*. Springer.

Marubini, E., & Valsecchi, M. G. (2004). *Analysing survival data from clinical trials and observational studies* (Vol. 15). John Wiley & Sons.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3(1), 27-32.

Morello, V., Barr, E., Bailes, M., Flynn, C., Keane, E., & van Straten, W. (2014). SPINN: a straightforward machine learning solution to the pulsar candidate selection problem. *Monthly Notices of the Royal Astronomical Society*, 443(2), 1651-1662.

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., & McCoy, R. G. (2019). Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. *Journal of biomedical informatics*, 89, 56-67.

Sarakarn, P., & Jumparway, D. (2020). Coverage and flexibility : issues should be considered for analyzing by generalized linear model in health science research. *Journal of Health Science and Community Public Health*, 3(2), 144-158. https://he01.tci-thaijo.org/index.php/jhscph/article/view/244276

Suresh, K., Severn, C., & Ghosh, D. (2022). Survival prediction models: an introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22(1), 207.

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6), 1-36.

ประวัติผู้เขียนวิทยานิพนธ์

ชื่อ-นามสกุล นางสาวมนัสพร ตรีรุ่งโรจน์

วัน เดือน ปี เกิด 1 สิงหาคม 2539

สถานที่เกิด

วุฒิการศึกษา

ที่อยู่ปัจจุบัน

ผลงานวิจัย

รางวัลหรือทุนการศึกษาที่ได้รับ