การพัฒนาเวิร์กโฟลว์สำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุด

นายพงศ์ทวัส ฮั่นวัฒนวงศ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

A WORKFLOW DEVELOPMENT FOR THE OPTIMAL CLASSIFICATION TREE MODEL

Mr. Pongthawat Hanwathanawong

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	การพัฒนาเวิร์กโฟลว์สำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุด
โดย	นายพงศ์ทวัส ฮั่นวัฒนวงศ์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.นัท กุลวานิช)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... กรรมการ

(อาจารย์ ดร.สาวิตรี บุญพัชรนนท์)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.ดลชัย ละออนวล)

พงศ์ทวัส ฮั่นวัฒนวงศ์ : การพัฒนาเวิร์กโฟลว์สำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุด. (A WORKFLOW DEVELOPMENT FOR THE OPTIMAL CLASSIFICATION TREE MODEL) อ.ที่ปรึกษาหลัก : รศ. ดร.เสกสรร เกียรติสุไพบูลย์

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาเวิร์กโฟลว์สำหรับสร้างต้นไม้จำแนกประเภทที่ดีที่สุดด้วยตัวแบบเชิงเส้นจำนวนเต็มแบบผสม ทำการประเมินประสิทธิภาพของตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต และขยายตัวแบบให้รองรับชุดข้อมูลที่ตัวแปรต้นมีค่าสูญหายจำนวนมาก จากการพัฒนาเวิร์กโฟลว์พบว่าการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสมในงานวิจัยของ Lin และ Tang (2021) และกำหนดค่าพารามิเตอร์ความซับซ้อนตั้งต้นเป็นค่าบวกใกล้เคียงศูนย์ให้ผลลัพธ์เป็นที่น่าพอใจ จากการเปรียบเทียบประสิทธิภาพระหว่างตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดกับต้นไม้ตัดสินใจบนชุดข้อมูลเยอรมันเครดิต พบว่าต้นไม้จำแนกประเภทที่ดีที่สุดให้อัตราความถูกต้องสูงกว่าต้นไม้ตัดสินใจทั้งบนชุดข้อมูลสร้างตัวแบบและบนชุดข้อมูลทวนสอบ 0.4% ถึง 3.2% ข้อดีของการพัฒนาเวิร์กโฟลว์โดยใช้โปรแกรมหาคำตอบสำหรับปัญหาเชิงเส้นจำนวนเต็มแบบผสม คือความสามารถในการขยายตัวแบบให้รองรับเงื่อนไขเพิ่มเติมได้ ในงานวิจัยนี้จึงเสนอตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดที่ถูกขยายให้รองรับชุดข้อมูลที่มีตัวแปรต้นสูญหายจำนวนมาก และแสดงให้เห็นว่าตัวแบบที่ถูกขยายสามารถทำงานอย่างมีประสิทธิผลบนเวิร์กโฟลว์ที่พัฒนาขึ้น

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6380214026 : MAJOR STATISTICS

KEYWORD: Decision Tree, Mixed-Integer Optimization, Optimal Classification Tree

Pongthawat Hanwathanawong : A WORKFLOW DEVELOPMENT FOR THE OPTIMAL CLASSIFICATION TREE MODEL. Advisor: Assoc. Prof. SEKSAN KIATSUPAIBUL, Ph.D.

This research aims to develop a workflow for creating an optimal classification tree (OCT) model by using mixed-integer optimization (MIO), to evaluate the performance of the optimal classification tree model on German Credit dataset, and to extend the model to support datasets that contain explanatory variables with a lot of missing values. By developing the workflow, we found that creating an optimal classification tree by solving an MIO problem using Lin and Tang's (2021) formulation, with the complexity parameter as a positive value close to zero, provides satisfactory results. By comparing the performance between the OCT and CART on German credit dataset, we found that both in-sample and out-of-sample accuracy of the OCT is greater than CART by 0.4 – 3.2%. One advantage of creating the OCT model by using MIO is the ability to extend the model to support additional required conditions. In this research, we propose an extension to the OCT model that supports datasets containing explanatory variables with a lot of missing values and show that the extended model can work effectively on the workflow.

Field of Study: Statistics Student's Signature .....

Academic Year: 2022 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีเพราะความเมตตากรุณาและเอาใจใส่อย่างดียิ่งจาก รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์ ที่ให้ความกรุณารับเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ ช่วยให้คำปรึกษา ให้คำแนะนำ รวมทั้งให้องค์ความรู้และแนวทางในการศึกษาค้นคว้า ตลอดจนให้ความช่วยเหลือและอบรมสั่งสอนมาโดยตลอด ผู้วิจัยขอขอบพระคุณท่านอาจารย์เป็นอย่างสูงด้วยความเคารพอย่างยิ่ง

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.นัท กุลวานิช ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร.สาวิตรี บุญพัชรนนท์ และผู้ช่วยศาสตราจารย์ ดร.ดลชัย ลออนวล กรรมการสอบวิทยานิพนธ์ ที่ได้กรุณาสละเวลามาเป็นกรรมการในการสอบครั้งนี้ตลอดจนช่วยให้ความรู้ และคำแนะนำที่มีประโยชน์ยิ่งในการเขียนวิทยานิพนธ์ให้สมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ทุกท่านที่กรุณาถ่ายทอดวิชาความรู้ และให้คำแนะนำที่ดีเสมอมาจนสำเร็จการศึกษาในครั้งนี้

ขอขอบพระคุณ คุณยุทธพงศ์ ฮั่นวัฒนวงศ์ และคุณพรทิพย์ ไทยสม ผู้เป็นบิดาและมารดาของผู้วิจัย ซึ่งผู้ให้การอบรมเลี้ยงดู ให้โอกาสในการศึกษาที่ดี และเป็นผู้อยู่เบื้องหลังความสำเร็จของผู้วิจัยเสมอมา รวมถึงครอบครัวที่เป็นกำลังใจ และคอยให้การช่วยเหลือ ให้ผู้วิจัยสามารถจัดทำวิทยานิพนธ์เล่มนี้จนสำเร็จลุล่วง และขอขอบคุณกัลยาณมิตรทุกท่านที่คอยให้คำแนะนำ และเป็นกำลังใจตลอดมา

สุดท้ายนี้ผู้วิจัยหวังเป็นอย่างยิ่งว่าวิทยานิพนธ์เล่มนี้จะเป็นประโยชน์แก่ผู้ที่สนใจศึกษาค้นคว้าในเรื่องดังกล่าว คุณความดีใดที่เกิดขึ้นจากวิทยานิพนธ์เล่มนี้ผู้วิจัยขอมอบให้แก่บุคคลทุกท่านที่ได้กล่าวมาทั้งหมดนี้ ตลอดจนท่านผู้เขียนตำราที่ผู้วิจัยนำมาอ้างอิงและเรียบเรียงเป็นวิทยานิพนธ์เล่มนี้ หากวิทยานิพนธ์เล่มนี้มีข้อบกพร่องหรือผิดพลาดประการใด ผู้วิจัยขอน้อมรับไว้แต่เพียงผู้เดียวและขออภัยไว้ ณ โอกาสนี้

พงศ์ทวัส ฮั่นวัฒนวงศ์

สารบัญ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ต้นไม้ตัดสินใจแบบจำแนก (Classification Tree) เป็นหนึ่งในรูปแบบการเรียนรู้ของเครื่องที่เป็นที่นิยม และถูกนำมาใช้ในการสร้างตัวแบบ ปัจจุบันนั้นมีการพัฒนาขั้นตอนวิธี (Algorithm) ในการสร้างต้นไม้ตัดสินใจมากมาย โดยขั้นตอนวิธีแบบ CART (Breiman et al., 1984) เป็นหนึ่งในขั้นตอนวิธีพื้นฐานและเป็นที่นิยม มีข้อดีที่สำคัญคือความง่ายต่อการอธิบาย (Interpretability) ลักษณะขั้นตอนวิธีแบบ CART นั้นเป็นการค้นหาแบบละโมบ (Greedy Search) โดยจะทำการแบ่งชุดข้อมูลที่ถูกนำเข้ามาที่แต่ละโหนด (Node) เป็น 2 กลุ่ม จากบนลงล่าง (Top-Down) จนกว่าข้อมูลทั้งหมดถูกแบ่งไปยังโหนดใบหรือโหนดปลายทาง (leaf node) โดยการแบ่งกลุ่มข้อมูลแต่ละครั้ง จะใช้ค่าของตัวแปรต้น (Independent Variable) ที่ทำให้ค่าจีนี (Gini Impurity) ลดลงต่ำสุดเป็นเกณฑ์ ซึ่งค่าจีนีเป็นค่าที่บ่งบอกถึงความสามารถในการแบ่งแยกตัวแปรตาม (Dependent Variable)

ข้อบกพร่องหลักของขั้นตอนวิธีแบบ CART รวมถึงขั้นตอนวิธีแบบอื่นๆ ที่เป็นที่นิยมอย่าง C4.5 (Quinlan, 1993) และ ID3 (Quinlan, 1986) ซึ่งใช้การค้นหาแบบละโมบ (Greedy Search) คือ การแบ่งแต่ละครั้งจะไม่ได้คำนึงถึงผลกระทบที่อาจจะเกิดขึ้นต่อการแบ่งในครั้งถัดๆ ไป ซึ่งอาจจะทำให้ไม่สามารถจับลักษณะความสัมพันธ์ของชุดข้อมูลได้ดี และนำไปสู่ประสิทธิภาพของตัวแบบที่ไม่ดีเท่าที่ควรเมื่อนำใช้ในการแบ่งแยกข้อมูลในอนาคต อีกทั้งการแบ่งจากบนลงล่าง (Top-Down) ไม่สามารถที่จะใช้การจำแนกผิด (Misclassification) ซึ่งเป็นวัตถุประสงค์สุดท้ายของตัวแบบ มาเป็นเกณฑ์ในการจำแนกได้ เนื่องจากการแบ่งมีโอกาสที่จะหยุดก่อนถึงจุดที่เหมาะสม และการแบ่งจากบนลงล่างก็ไม่สามารถที่จะจัดการกับความซับซ้อน (Complexity) ของตัวแบบต้นไม้ได้อีกด้วย จึงจำเป็นต้องมีการตัดกิ่ง (Pruning) เพื่อลดความซับซ้อน ภายหลังจากการสร้างตัวแบบต้นไม้

จากการพยายามแก้ข้อบกพร่องดังกล่าว Bertsimas และ Dunn ได้นำเสนอตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุด (Bertsimas & Dunn, 2017) ซึ่งเป็นการมองการสร้างต้นไม้ตัดสินใจแบบจำแนก (Classification Tree) เป็นปัญหาเชิงเส้นจำนวนเต็มแบบผสม (Mixed-Integer Programing) โดยมีวัตถุประสงค์ (Objective) ของปัญหา คือ ค่าการจำแนกผิด (Misclassification) ที่ต่ำที่สุด จากการแก้ปัญหาดังกล่าวจะทำให้ตัวแบบต้นไม้ทั้งต้นถูกสร้างขึ้นในขั้นตอนเดียว การแบ่งแต่ละครั้งจะคำนึงถึงผลกระทบต่อการแบ่งครั้งอื่นๆ ทั้งหมด ทำให้ตัวแบบต้นไม้ตัดสินใจที่ได้เป็นตัวแบบที่เหมาะสมที่สุด (Optimal) สำหรับชุดข้อมูลสร้างตัวแบบ (Training Data)

ส่วนหนึ่งในงานวิจัยของ Lin และ Tang (2021) ได้นำตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดมาประยุกต์ใช้โดยมีวัตถุประสงค์เพื่อศึกษาและประเมินประสิทธิภาพของต้นไม้จำแนกประเภทที่พัฒนาโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสมเพิ่มเติม โดยตัวแบบที่ถูกนำไปประยุกต์ใช้ในงานวิจัยของ Lin และ Tang (2021) มีความแตกต่างจากในงานวิจัยของ Bertsimas และ Dunn (2017) เล็กน้อย

การพัฒนาตัวแบบการเรียนรู้ของเครื่องในหลายๆกรณีมักต้องการตัวแบบที่มีความง่ายต่อการอธิบาย เช่น การพัฒนาแบบจำลองคะแนนเครดิต แต่ตัวแบบที่มีความง่ายต่อการอธิบายในปัจจุบันนั้นมีเพียงไม่กี่ตัว ตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดมีลักษณะเช่นเดียวกับตัวแบบต้นไม้ตัดสินใจซึ่งเป็นตัวแบบที่เป็นที่นิยมและมีความง่ายต่อการอธิบายสูง

ผู้วิจัยจึงมีความสนใจเป็นอย่างมากในการพัฒนาเวิร์กโฟลว์สำหรับการสร้างตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสมโดยศึกษาจากงานวิจัยของ Bertsimas และ Dunn (2017) และ Lin และ Tang (2021) และทดสอบประสิทธิผลของตัวแบบและเวิร์กโฟลว์บนชุดข้อมูลจริงและประเมินประสิทธิภาพ หนึ่งในข้อดีของการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสม คือความสามารถในการขยายตัวแบบให้รองรับเงื่อนไขเพิ่มเติม ผู้วิจัยจึงมีความสนใจในการขยายตัวแบบโดยเสนอวิธีการที่มีลักษณะที่เหมาะกับการแบ่งของต้นไม้ตัดสินใจในการรองรับชุดข้อมูลที่ตัวแปรต้นมีค่าสูญหายจำนวนมาก เพื่อเป็นแนวทางในการประยุกต์ใช้ ศึกษา และพัฒนาตัวแบบดังกล่าว

1.2 วัตถุประสงค์การวิจัย

เพื่อพัฒนาเวิร์กโฟลว์สำหรับสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสม (Mixed-Integer Programing) ทดสอบประสิทธิผลของตัวแบบและเวิร์กโฟลว์บนชุดข้อมูลจริงและประเมินประสิทธิภาพ และขยายตัวแบบให้รองรับชุดข้อมูลที่ตัวแปรต้นมีค่าสูญหายจำนวนมาก

1.3 ขอบเขตของการวิจัย

1.3.1 การพัฒนาเวิร์กโฟลว์สำหรับสร้างต้นไม้จำแนกประเภทที่ดีที่สุด

งานวิจัยในส่วนนี้จะทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลจำลองขนาดเล็ก โดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสม เพื่อทดสอบและเลือกใช้ตัวแบบและวิธีการสร้างที่ให้ประสิทธิผลที่ถูกต้อง และพัฒนาเวิร์กโฟลว์จากตัวแบบและวิธีการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดดังกล่าว

1.3.2 การประเมินประสิทธิภาพของต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต

ทดสอบประสิทธิผลของตัวแบบและเวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต (Hofmann, 1994) ซึ่งเป็นข้อมูลขนาด 1,000 ตัวอย่าง 1 ตัวแปรตามและ 20 ตัวแปรต้น และประเมินประสิทธิภาพเบื้องต้นโดยเปรียบเทียบความแม่นยำของตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดกับต้นไม้ตัดสินใจแบบ CART ที่ค่าพารามิเตอร์ความลึกสูงสุด 2 ถึง 4 บนชุดข้อมูลดังกล่าว

1.3.3 การขยายตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดให้รองรับชุดข้อมูลที่ตัวแปรต้นมีค่าสูญหายจำนวนมาก

ขยายตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดให้รองรับชุดข้อมูลที่มี ตัวแปรต้นที่มีค่าสูญหาย (Missing values) จำนวนมาก โดยเพิ่มเงื่อนไขแสดงขอบเขตที่ทำให้ตัวแปรที่มีข้อมูลสูญหายจำนวนมากถูกแยกข้อมูลส่วนที่

สูญหายและไม่สูญหายออกจากกันที่โหนดพ่อแม่ (Parent node) ก่อนที่ข้อมูลส่วนที่ไม่สูญหายของตัวแปรดังกล่าวถูกใช้ในการแบ่งที่โหนดลูก (Child node) เพื่อให้ตัวแปรที่มีข้อมูลสูญหายถูกใช้ในการสร้างตัวแบบโดยไม่ได้รับผลกระทบจากการแทนค่าข้อมูลส่วนที่สูญหาย ทดสอบประสิทธิผลของตัวแบบที่ถูกขยายดังกล่าวบนชุดข้อมูลเยอรมันเครดิตโดยเพิ่มตัวแปรจำลองที่สามารถใช้ในการจำแนกตัวแปรตาม (good credit risk/bad credit risk) ได้ดี แต่มีข้อมูลสูญหายจำนวนมาก

1.4 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการประยุกต์ใช้ศึกษา และพัฒนาการสร้างต้นไม้จำแนกประเภทที่ดีที่สุด (Optimal Classification Tree) โดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสม (Mixed-Integer Programing) และการขยายตัวแบบเพื่อให้รองรับเงื่อนไขเพิ่มเติม

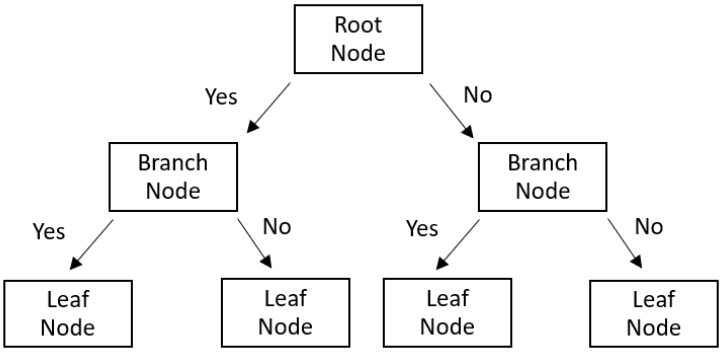
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การเรียนรู้แบบมีผู้สอน (Supervised learning)

การเรียนรู้แบบมีผู้สอน เป็นหนึ่งในรูปแบบของการเรียนรู้ของเครื่อง เพื่อให้เครื่องสามารถหาคำตอบได้ด้วยตัวเองโดนอาศัยข้อมูลสอน (Training Data) ซึ่งประกอบไปด้วยตัวแปรต้น (Independent Variable) และตัวแปรตาม (Dependent Variable) ในการเรียนรู้ผ่านขั้นตอนวิธีทางคณิตศาสตร์ (Algorithm) เพื่อให้ได้ตัวแบบ (Model) ที่สามารถนำไปใช้ในการทำนายตัวแปรตามจากการป้อนข้อมูลตัวแปรต้นลงในตัวแบบ หากตัวแปรตามเป็นข้อมูลเชิงคุณภาพ (Category) จะเรียกว่าเป็นการจำแนกประเภท (Classification) เช่น การจำแนกลูกค้าที่มีเครดิตดีและเครดิตเสีย หากตัวแปรตามเป็นข้อมูลเชิงปริมาณ (Numerical) จะเรียกว่าเป็นการถดถอย (Regression) เช่น การทำนายราคาบ้าน

2.2 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree) เป็นหนึ่งในรูปแบบการเรียนรู้ของเครื่อง (Machine Learning) แบบมีผู้สอน (Supervised Learning) ที่เป็นพื้นฐานและเป็นที่นิยมซึ่งมีข้อดีที่สำคัญคือมีความง่ายต่อการอธิบาย (Interpretable) สูง ต้นไม้ตัดสินใจจะมีลักษณะเหมือนต้นไม้กลับหัวโดยข้อมูลสอน (Training Data) ที่ถูกนำเข้ามาสร้างตัวแบบจะถูกแบ่งกลุ่มอย่างต่อเนื่อง (Recursive Partitioning) จากบนลงล่าง (Top-Down) โดยเริ่มที่โหนดราก (Root Node) ไปยังโหนดใบหรือโหนดปลายทาง (Leaf Node) เพื่อคำนวณหาตัวแปรและค่าของตัวแปรที่ใช้เป็นเกณฑ์ในการแบ่งที่แต่ละโหนดกิ่ง ซึ่งจะมีลักษณะโครงสร้าง ดังรูปที่ 2.1



รูปที่ 2.1 โครงสร้างของต้นไม้ตัดสินใจ

ในปัจจุบันขั้นตอนวิธีที่ใช้ในการสร้างตัวแบบต้นไม้ตัดสินใจ มีหลากหลายขั้นตอนวิธี CART (Breiman et al., 1984) เป็นหนึ่งในขั้นตอนวิธีที่เป็นที่นิยม ซึ่งมีลักษณะเป็นการค้นหาแบบละโมภ (Greedy Search) กล่าวคือการแบ่งกลุ่มของข้อมูลในแต่ละครั้งที่แต่ละโหนด สำหรับต้นไม้ตัดสินใจแบบจำแนก (Classification tree) จะใช้ค่าของตัวแปรต้น (Independent Variable) ที่ทำให้ค่าจีนี (Gini Impurity) หลังจากการแบ่งลดลงต่ำสุดเป็นเกณฑ์ ค่าจีนีเป็นค่าที่ใช้ในการบ่งบอกถึงความสามารถในการแบ่งแยกตัวแปรตาม (Dependent Variable) ค่าจีนีบนโหนด t ของต้นไม้ตัดสินใจที่สร้างบนชุดข้อมูลที่มีตัวแปรตามแบ่งเป็น n คลาส สามารถคำนวณได้ดังนี้

$$Gini\ Index\ (t) = 1 - \sum_{i=1}^{n} (p_i)^2$$

โดย p_i เป็นอัตราส่วนของตัวอย่างที่มีตัวแปรตามเป็นคลาส i

2.3 ปัญหาเชิงเส้นจำนวนเต็มแบบผสม (Mixed Integer Programming)

ปัญหาเชิงเส้น (Linear Programming) เป็นเทคนิคทางคณิตศาสตร์ที่ใช้ในการหาค่าที่เหมาะสมที่สุด มักถูกใช้ในการตัดสินใจในการใช้ทรัพยากรณ์ที่มีอย่างจำกัดให้เกิดประโยชน์สูงสุด เช่น การวางแผนการผลิตที่มีวัตถุดิบและเวลาอยู่อย่างจำกัดเพื่อให้ได้ต้นทุนในการผลิตต่ำที่สุด โดยลักษณะคำตอบจะอยู่ในรูปค่าของตัวแปรที่ต้องพิจารณา โดยการกำหนดรูปแบบของปัญหาเชิงเส้นจะประกอบไปด้วย

ฟังก์ชันวัตถุประสงค์ (Objective) เป็นวัตถุประสงค์ของตัวแบบซึ่งจะอยู่ในรูปแบบค่าที่สูงที่สุด (Maximize) หรือค่าต่ำที่สุด (Minimize) เช่น มีวัตถุประสงค์เพื่อหาต้นทุนที่ต่ำที่สุดในการผลิต

เงื่อนไขแสดงขอบเขต (Constraints) เป็นเงื่อนไขบังคับหรือข้อจำกัดของปัญหาในรูปแบบของอสมการ เช่น ปริมาณจำกัดของวัตถุดิบแต่ละชนิด

ตัวแปรตัดสินใจ (Decision Variable) เป็นตัวแปรที่ต้องพิจารณาเพื่อหาค่าที่เหมาะสมที่สุดตามวัตถุประสงค์ และข้อจำกัดที่กำหนด เช่น จำนวนของวัตถุดิบแต่ละชนิด

ซึ่งสามารถเขียนให้อยู่ในรูปแบบมาตรฐาน (Standard Form) ได้ดังนี้

วัตถุประสงค์

$$\text{ค่าสูงสุด} \quad z = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

ข้อจำกัด

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leq b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leq b_2$$
$$\vdots$$
$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leq b_m$$
$$x_1, x_2, \ldots x_n \geq 0$$

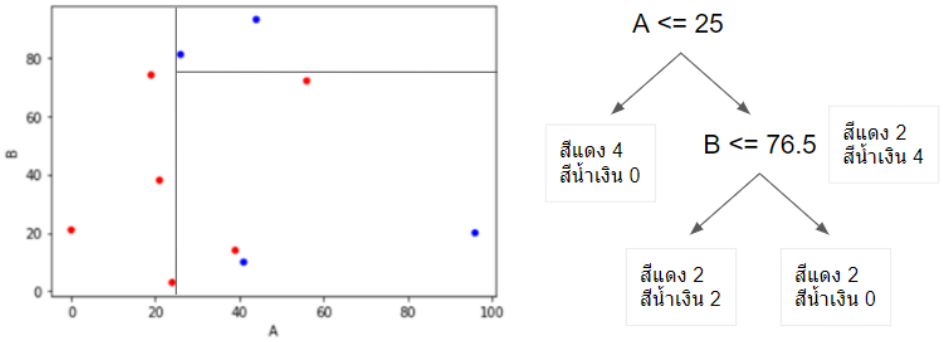
โดยที่ x_i เป็นตัวแปรตัดสินใจ (Decision Variable)

a_{ij}, b_i, c_i เป็นค่าคงที่

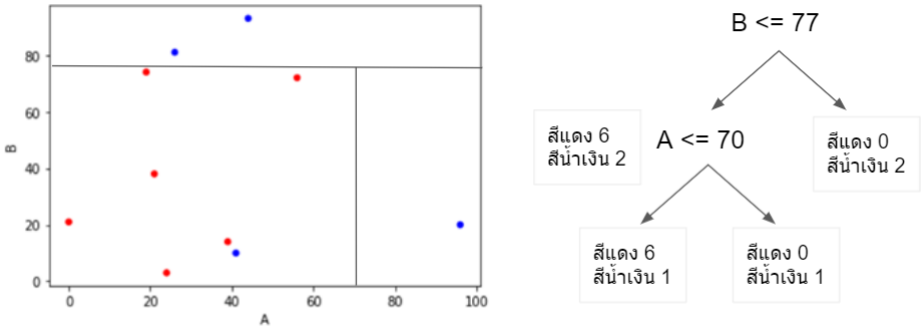
หากปัญหาเชิงเส้น (Linear Programming) มีตัวแปรตัดสินใจ (Decision Variable) บางส่วนเป็นจำนวนเต็มบวก ปัญหาเชิงเส้นดังกล่าวจะถูกเรียกว่าเป็น ปัญหาเชิงเส้นจำนวนเต็มแบบผสม (Mixed Integer Programming)

2.4 ต้นไม้จำแนกประเภทที่ดีที่สุด (Optimal Classification Tree)

ขั้นตอนวิธีที่ใช้ในการพัฒนาต้นไม้ตัดสินใจ ที่นิยมอย่าง CART (Breiman et al., 1984), C4.5 (Quinlan, 1993) และ ID3 (Quinlan, 1986) มีข้อบกพร่องหลักร่วมกันอยู่อย่างนึง คือ การแบ่งจากบนลงล่าง (top-down) โดยใช้การค้นหาแบบละโมบ (Greedy Search) กล่าวคือการแบ่งกลุ่มของข้อมูลในแต่ละครั้งที่แต่ละโหนด จะใช้ค่าของตัวแปรต้น (Independent Variable) ที่ทำให้ค่าจีนี (Gini Impurity) ลดลงต่ำสุดเป็นเกณฑ์ ทำให้การแบ่งแต่ละครั้ง จะไม่คำนึงถึงผลกระทบที่อาจจะเกิดขึ้นในการแบ่งในครั้งถัดๆ ไป อีกทั้งการแบ่งจากบนลงล่างยังไม่สามารถใช้การจำแนกผิด (misclassification) ซึ่งเป็นวัตถุประสงค์สุดท้ายของตัวแบบ มาเป็นเกณฑ์ในการจำแนกได้ และยังไม่สามารถจะจัดการกับความซับซ้อน (Complexity) ของตัวแบบต้นไม้ได้ จึงจำเป็นต้องมีขั้นตอนการตัดกิ่ง (Pruning) ภายหลังจากการสร้างตัวแบบเพิ่มเติม เพื่อลดความซับซ้อน



รูปที่ 2.2 ตัวอย่างต้นไม้ตัดสินใจที่พัฒนาจากขั้นตอนวิธี CART



รูป 2.3 ตัวอย่างต้นไม้ตัดสินใจที่ใช้การสุ่มเกณฑ์ในการแบ่ง

ตัวอย่างข้างต้นเป็นการพัฒนาต้นไม้ตัดสินใจ ที่ใช้ในการจำแนกจุดสีน้ำเงินและจุดสีแดง โดยตัวอย่างในรูป 2.2 เป็นต้นไม้ตัดสินใจที่พัฒนาโดยใช้ขั้นตอน CART ซึ่งค่าจีนีหลังจากการแบ่งครั้งแรกเท่ากับ 0.267 หลังจากการแบ่งครั้งที่สองเท่ากับ 0.2 ตัวอย่างใน รูป 2.3 เป็นต้นไม้ตัดสินใจที่มาจากการสุ่มเกณฑ์ในการแบ่ง มีค่าจีนีหลังจากการแบ่งครั้งแรกเท่ากับ 0.3 หลังจากการแบ่งครั้งที่สองเท่ากับ 0.172 จะเห็นได้ว่าตัวแบบต้นไม้ตัดสินใจที่พัฒนาโดย CART ซึ่งใช้การค้นหาแบบละโมบ (greedy) จะมีค่าจีนีจากการแบ่งครั้งแรก ต่ำกว่า ต้นไม้ตัดสินใจที่มาจากการสุ่ม แต่หลังจากการแบ่งครั้งที่สองแล้ว ค่าจีนีของต้นไม้ตัดสินใจที่มาจากการสุ่ม มีค่าต่ำกว่าต้นไม้ตัดสินใจที่พัฒนาโดย CART จากตัวอย่างดังกล่าวแสดงให้เห็นถึงข้อบกพร่องของการใช้การค้นหาแบบละโมบ ซึ่งการแบ่งแต่ละครั้งจะใช้ค่าของตัวแปรต้น ที่ทำให้ค่าจีนีลดลงต่ำสุดเป็นเกณฑ์ โดยไม่คำนึงถึงผลลัพธ์จากการแบ่งในครั้งถัดไป

ต้นไม้จำแนกประเภทที่ดีที่สุด หรือ Optimal Classification Tree (OCT) เป็นขั้นตอนวิธีที่พัฒนาโดย Bertsimas และ Dunn (2017) มีวัตถุประสงค์เพื่อแก้ข้อบกพร่องของขั้นตอนวิธีดังกล่าว โดยมองการสร้างต้นไม้ตัดสินใจแบบจำแนก (Classification Tree) เป็นปัญหาเชิงเส้นจำนวนเต็มแบบผสม (Mixed-Integer Programing) มีวัตถุประสงค์ (Objective) ของปัญหา คือ ค่าการจำแนกผิด (Misclassification) ที่ต่ำที่สุด จากการแก้ปัญหาเชิงเส้นจำนวนเต็มแบบผสมดังกล่าวจะทำให้ตัวแบบต้นไม้ทั้งต้นถูกสร้างขึ้นในขั้นตอนเดียว การแบ่งแต่ละครั้งจะ

คำนึงถึงผลกระทบต่อการแบ่งครั้งอื่นๆ ทั้งหมด ทำให้ตัวแบบต้นไม้ตัดสินใจที่ได้เป็นตัวแบบที่เหมาะสมที่สุด (Optimal) สำหรับชุดข้อมูลสอน

โดยในงานวิจัยของ Bertsimas และ Dunn (2017) ได้เสนอตัวแบบเชิงเส้นจำนวนเต็มแบบผสมสำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดไว้ ดังนี้

$$\min \ \frac{1}{\hat{L}} \sum_{t \in \tau_L} L_t + \alpha \sum_{t \in \tau_B} d_t \tag{1}$$

$$s.t. \ \ L_t \geq N_t - \ N_{kt} - n(1 - c_{kt}), \ \ k = 1, \ldots, K, \ \ \forall t \ \in \ \tau_L, \tag{2}$$

$$L_t \leq N_t - \ N_{kt} + nc_{kt}, \ \ k = 1, \ldots, K, \ \ \forall t \ \in \ \tau_L, \tag{3}$$

$$L_t \geq 0, \ \ \forall t \ \in \ \tau_L, \tag{4}$$

$$N_{kt} = \frac{1}{2} \sum_{i=1}^{n} (1 + Y_{ik}) z_{it}, \ \ k = 1, \ldots, K, \ \ \forall t \ \in \ \tau_L, \tag{5}$$

$$N_t = \sum_{i=1}^{n} z_{it}, \ \ \forall t \ \in \ \tau_L, \tag{6}$$

$$\sum_{k=1}^{K} c_{kt} = l_t, \ \ \forall t \ \in \ \tau_L, \tag{7}$$

$$a_m^T x_i \geq b_m - (1 - z_{it}), \ \ i = 1, \ldots, n, \forall t \ \in \ \tau_B, \ \ \forall m \ \in \ A_R(t), \tag{8}$$

$$a_m^T (x_i + \epsilon) \leq b_m + (1 + \epsilon_{max})(1 - z_{it}), \ \ i = 1, \ldots, n, \ \ \forall t \ \in \ \tau_B, \ \ \forall m \ \in \ A_L(t), \tag{9}$$

$$\sum_{t \in \tau_L} z_{it} = 1, \ \ i = 1, \ldots, n, \tag{10}$$

$$z_{it} \leq l_t, \ \ \forall t \ \in \ \tau_L, \tag{11}$$

$$\sum_{i=1}^{n} z_{it} \geq N_{min} l_t, \ \ \forall t \ \in \ \tau_L, \tag{12}$$

$$\sum_{i=1}^{n} a_{jt} = d_t, \ \ \forall t \ \in \ \tau_B, \tag{13}$$

$$0 \leq b_t \leq d_t, \ \ \forall t \ \in \ \tau_B, \tag{14}$$

$$d_t \leq d_{p(t)}, \ \ \forall t \ \in \ \tau_B \backslash \{1\}, \tag{15}$$

$$z_{it}, l_t \in \{0, 1\}, \ \ i = 1, \ldots, n, \ \ \forall t \ \in \ \tau_L \tag{16}$$

$$a_{jt}, d_t \in \{0, 1\}, \ \ j = 1, \ldots, p, \ \ \forall t \ \in \ \tau_B \tag{17}$$

โดยมีตัวแปรตัดสินใจ พารามิเตอร์ และค่าคงที่สำหรับตัวแบบ ดังตารางที่ 2.1 2.2 และ 2.3 ตามลำดับ

ตารางที่ 2.1 ตัวแปรตัดสินใจสำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดและคำอธิบาย

ตัวแปรตัดสินใจ	คำอธิบาย
L_t	จำนวนการทำนายผิดที่โหนดใบ t
N_{kt}	จำนวนตัวอย่างที่เป็นคลาส k ที่โหนดใบ t
c_{kt}	คลาส k เป็นค่าในการทำนายที่โหนดใบ t หรือไม่
a_{jt}	ตัวแปรตำแหน่งที่ j ถูกแบ่งที่โหนดใบ t หรือไม่
z_{it}	ตัวอย่างที่ i ตกที่โหนดใบ t หรือไม่
l_t	มีตัวอย่างตกที่โหนดใบ t หรือไม่
d_t	โหนดใบ t มีการแบ่งหรือไม่
b_t	ค่าของตัวแปรที่ใช้เป็นเกณฑ์ในการแบ่งที่โหนดใบ t

ตารางที่ 2.2 พารามิเตอร์สำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดและคำอธิบาย

พารามิเตอร์	คำอธิบาย
N_{min}	จำนวนตัวอย่างที่โหนดใบต่ำสุด
$Max\ depth\ (D)$	ความลึกสูงสุดของต้นไม้ตัดสินใจ
$alpha\ (\alpha)$	ค่าความซับซ้อนของต้นไม้ตัดสินใจ ค่า α ที่ให้สูงขึ้นส่งผลให้จำนวนครั้งในการแบ่งลดลง

ตาราง 2.3 ค่าคงที่สำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดและคำอธิบาย

ค่าคงที่	คำอธิบาย
n	จำนวนตัวอย่างของชุดข้อมูลสอน
p	จำนวนคอลัมน์ (column) ของชุดข้อมูลสอน
K	จำนวนคลาสของตัวแปรตาม
τ_B	ตำแหน่งของแต่ละโหนดใบ
τ_L	ตำแหน่งของแต่ละโหนดกิ่ง
$\hat{L}$	ค่าทำนายผิดพื้นฐานเท่ากับค่าทำนายผิดหากทำนายด้วยคลาสที่มีจำนวนเยอะที่สุด
Y_{ik}	เมทริกที่ใช้ระบุว่าจำนวนตัวอย่างที่มีตัวแปรตามเป็นคลาส k ที่โหนดใบ
ε_j	ค่าส่วนต่างบวกที่น้อยที่สุดของแต่ละตัวอย่างในตัวแปร j

คำอธิบายของวัตถุประสงค์ (Objective) และเงื่อนไขแสดงขอบเขต (Constraints) สมการที่ (1) – (17) มีกล่าวโดยละเอียดในงานวิจัยของ Bertsimas และ Dunn (2017) ซึ่งสามารถอธิบายโดยสังเขปได้ดังนี้

ตัวแบบมีวัตถุประสงค์ (1) เพื่อหาค่าการทำนายผิดรวมที่โหนดใบ โดยมีค่าพารามิเตอร์ความซับซ้อนของต้นไม้ตัดสินใจในการควบคุมจำนวนครั้งในการแบ่ง และมีเงื่อนไขแสดงขอบเขต (2) – (4) เพื่อระบุค่าทำนายผิดที่แต่ละโหนดใบ (5) – (6) เพื่อระบุจำนวนตัวอย่างทั้งหมดและจำนวนตัวอย่างของแต่ละคลาสที่แต่ละโหนดใบ (7) เพื่อระบุค่าการทำนายที่แต่ละโหนดใบ (8) – (9) เพื่อระบุตัวแปรและค่าของตัวแปรที่ใช้ในการแบ่งตัวอย่างที่แต่ละโหนดกิ่ง (10) – (12) เพื่อให้มั่นใจว่าแต่ละตัวอย่างตกลงที่โหนดใบโหนดใดโหนดหนึ่ง (13) – (14) เพื่อให้มั่นใจว่าตัวแปรและค่าของตัวแปรที่ใช้ในการแบ่งถูกระบุเฉพาะที่โหนดกิ่งที่มีการแบ่ง (15) เพื่อให้มั่นใจว่าโหนดลูก (Child node) จะถูกแบ่งได้ก็ต่อเมื่อโหนดพ่อแม่ (Parent node) มีการแบ่งเท่านั้น (16) – (17) ตัวแปรตัดสินใจ z_{it}, l_t, a_{jt} และ d_t เป็นตัวแปรประเภทไบนารี (Binary Variable)

2.5 ข้อมูลสูญหาย (Missing Data)

ข้อมูลสูญหาย เป็นหนึ่งในปัญหาที่พบได้บ่อยในการสร้างตัวแบบการเรียนรู้ของเครื่อง ข้อมูลสูญหายเกิดได้จากหลายปัจจัย เช่น การที่ผู้ทำแบบสอบถามหรือผู้ที่กรอกข้อมูลในระบบให้ข้อมูลไม่ครบถ้วน หรืออาจจะเกิดจากความผิดพลาดของระบบฐานข้อมูล ข้อมูลสูญหายโดยทั่วไปมีอยู่ 3 รูปแบบ ได้แก่

ข้อมูลสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely At Random : MCAR) ข้อมูลสูญหายเกิดขึ้นอย่างสุ่มโดยไม่ขึ้นกับกลุ่มใดกลุ่มหนึ่งหรือตัวแปรอื่นๆ อาจจะเกิดได้จากความผิดพลาดในการนำข้อมูลเข้าระบบ

ข้อมูลสูญหายแบบสุ่ม (Missing At Random : MAR) ข้อมูลสูญหายเกิดขึ้นอยางสุ่มแต่เกิดภายในบางกลุ่มของข้อมูลหรือขึ้นอยู่กับตัวแปรอื่นๆ เช่น ข้อมูลน้ำหนักบางส่วนสูญหายในกลุ่มเพศหญิงเกิดจากกลุ่มผู้ให้ข้อมูลเพศหญิงมักจะไม่ให้ข้อมูลน้ำหนัก

ข้อมูลสูญหายแบบไม่สุ่ม (Not Missing At Random : NMAR) ข้อมูลสูญหายเกิดขึ้นอยู่กับข้อมูลที่สูญหายเอง เช่น ผู้ที่มีน้ำหนักตัวเยอะมักจะไม่ให้ข้อมูลน้ำหนัก

การจัดการกับข้อมูลส่วนที่สูญหายเพื่อให้ชุดข้อมูลนั้นสามารถนำไปใช้ในการสร้างตัวแบบการเรียนรู้ของเครื่องได้นั้นมีหลายวิธี เช่น การเก็บข้อมูลเพิ่มเติม การนำตัวแปรที่มีข้อมูลสูญหายจำนวนมากออก หรือการแทนค่าข้อมูลส่วนที่สูญหาย ขึ้นอยู่กับลักษณะการหายของข้อมูลและความจำเป็นในการใช้ข้อมูลดังกล่าว

บทที่ 3
ขอบเขตและวิธีการดำเนินงานวิจัย

3.1 ขอบเขตงานวิจัย

3.3.1 การพัฒนาเวิร์กโฟลว์สำหรับสร้างต้นไม้จำแนกประเภทที่ดีที่สุด

งานวิจัยในส่วนนี้จะทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลจำลองขนาดเล็ก โดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสม เพื่อทดสอบและเลือกใช้ตัวแบบและวิธีการสร้างที่ให้ประสิทธิผลที่ถูกต้อง และพัฒนาเวิร์กโฟลว์จากตัวแบบและวิธีการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดดังกล่าว

3.3.2 การประเมินประสิทธิภาพของต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต

ทดสอบประสิทธิผลของตัวแบบและเวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตซึ่งเป็นข้อมูลขนาด 1,000 ตัวอย่าง 1 ตัวแปรตามและ 20 ตัวแปรต้น และประเมินประสิทธิภาพเบื้องต้นโดยเปรียบเทียบความแม่นยำของตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดกับต้นไม้ตัดสินใจแบบ CART ที่ค่าพารามิเตอร์ความลึกสูงสุด 2 ถึง 4 บนชุดข้อมูลดังกล่าว

3.3.3 การขยายตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดให้รองรับชุดข้อมูลที่ตัวแปรต้นมีค่าสูญหายจำนวนมาก

ขยายตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดให้รองรับชุดข้อมูลที่มี ตัวแปรต้นที่มีค่าสูญหาย (Missing values) จำนวนมาก โดยเพิ่มเงื่อนไขแสดงขอบเขตที่ทำให้ตัวแปรที่มีข้อมูลสูญหายจำนวนมากถูกแยกข้อมูลส่วนที่สูญหายและไม่สูญหายออกจากกันที่โหนดพ่อแม่ (Parent node) ก่อนที่ข้อมูลส่วนที่ไม่สูญหายของตัวแปรดังกล่าวถูกใช้ในการแบ่งที่โหนดลูก (Child node) เพื่อให้ตัวแปรที่มีข้อมูลสูญหายถูกใช้ในการสร้างตัวแบบโดยไม่ได้รับผลกระทบจากการแทนค่าข้อมูลส่วนที่สูญหาย ทดสอบประสิทธิผลของตัวแบบที่ถูกขยายดังกล่าวบนชุดข้อมูลเยอรมันเครดิตโดยเพิ่มตัวแปรจำลองที่สามารถใช้ในการจำแนกตัวแปรตาม (good credit risk/bad credit risk) ได้ดี แต่มีข้อมูลสูญหายจำนวนมาก

3.2 วิธีการดำเนินการวิจัย

3.2.1 การพัฒนาเวิร์กโฟลว์สำหรับสร้างต้นไม้จำแนกประเภทที่ดีที่สุด

1. ศึกษาทฤษฎี และการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสมจากเอกสารและงานวิจัยที่เกี่ยวข้อง

2. ทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสมบนชุดข้อมูลขนาด 6 ตัวอย่าง 2 ตัวแปรต้น และ 1 ตัวแปรตาม

3. แปลงผลลัพธ์ของตัวแปรตัดสินใจที่ได้จากการแก้ปัญหาตัวแบบเชิงเส้นจำนวนเต็มแบบผสมให้อยู่ในรูปแบบของต้นไม้ตัดสินใจ

4. ทดสอบประสิทธิผลของตัวแบบและวิธีที่ใช้ในการสร้างต้นไม้จำแนกประเภทที่ดีที่สุด โดยตรวจสอบว่าต้นไม้ตัดสินใจที่ได้มีโครงสร้างที่ถูกต้องหรือไม่

5. พัฒนาเวิคโฟลว์จากตัวแบบและวิธีการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดที่ให้ประสิทธิผลที่ถูกต้อง

3.2.2 การประเมินประสิทธิภาพของต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต

1. จัดเตรียมชุดข้อมูลเยอรมันเครดิตให้อยู่ในรูปแบบที่พร้อมสำหรับการสร้างตัวแบบ โดยการทำให้เป็นปกติน้อยที่สุด-มากที่สุด (Min-Max Normalization) และสร้างตัวแปรหุ่น (Dummy Variable Encoding) สำหรับตัวแปรที่เป็นข้อมูลเชิงคุณภาพ

2. สุ่มแบ่งชุดข้อมูลเป็น 2 ส่วน ข้อมูลสร้างตัวแบบ 75% และชุดข้อมูลทวนสอบ 25%

3. แบ่งชุดข้อมูลสร้างตัวแบบออกเป็น 3 ส่วน สร้างต้นไม้จำแนกประเภทที่ดีที่สุดและต้นไม้จำแนกประเภทแบบ CART ใช้เทคนิคทดสอบแบบไขว้ (K-fold cross-validate) โดยใช้ 2 ส่วนสำหรับการสร้างตัวแบบ 1 ส่วนสำหรับการทดสอบ เพื่อหาค่าพารามิเตอร์ความซับซ้อนหรือจำนวนครั้งในการแบ่ง (C) ซึ่งมีค่าตั้งแต่ 1 ถึง $2^D - 1$ ที่ให้ความแม่นยำเฉลี่ยสูงสุดสำหรับแต่ละค่าพารามิเตอร์ความลึกสูงสุด (D) 2 ถึง 4 กำหนดค่าพารามิเตอร์จำนวนตัวอย่างที่โหนดใบต่ำสุด (N_{min}) เป็น 5% ของจำนวนตัวอย่างของชุดข้อมูลสร้างตัวแบบ

4. สร้างต้นไม้จำแนกประเภทที่ดีที่สุด และต้นไม้จำแนกประเภทแบบ CART บนชุดข้อมูลสร้างตัวแบบ 75% โดยกำหนดค่าพารามิเตอร์ความลึกสูงสุด 2 ถึง 4 กำหนดค่าพารามิเตอร์จำนวนครั้งในการแบ่งสูงสุดที่ได้จากการทำการทดสอบแบบไขว้ (K-fold cross-validate) และกำหนดค่าพารามิเตอร์จำนวนตัวอย่างที่โหนดใบต่ำสุดเป็น 5% ของจำนวนตัวอย่างของชุดข้อมูลสร้างตัวแบบ

5. ประเมินประสิทธิภาพโดยวัดความแม่นยำของตัวแบบบนชุดข้อมูลทวนสอบ 25% ที่แต่ละค่าพารามิเตอร์ความลึกสูงสุด 2 ถึง 4 ของต้นไม้จำแนกประเภทที่ดีที่สุดและต้นไม้จำแนกประเภทแบบ CART

6. เพื่อลดผลการทบจากการสุ่มแบ่งข้อมูลที่มีผลต่อการประเมินประสิทธิภาพ ดำเนินขั้นตอน 2 ถึง 5 ซ้ำ 3 ครั้ง โดยสุ่มแบ่งชุดข้อมูลใหม่และหาค่าความแม่นยำเฉลี่ยจากการทดลองซ้ำทั้ง 3 ครั้ง

7. เปรียบเทียบค่าความแม่นยำเฉลี่ยของต้นไม้จำแนกประเภทที่ดีที่สุดกับต้นไม้จำแนกประเภทแบบ CART ที่แต่ละค่าพารามิเตอร์ความลึกสูงสุด 2 ถึง 4 และสรุปผล

3.2.3 การขยายตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดให้รองรับชุดข้อมูลที่มีตัวแปรต้นสูญหายจำนวนมาก

1. กำหนดวิธีการที่ทำให้ตัวแบบต้นไม้จำแนกประเภทดีที่สุดสามารถรองรับตัวแปรต้นที่มีข้อมูลสูญหาย (Missing Value) จำนวนมาก

2. เพิ่มเงื่อนไขแสดงขอบเขตสำหรับวิธีการดังกล่าวลงบนตัวแบบเชิงเส้นจำนวนเต็มแบบผสมสำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุด

3. เพิ่มตัวแปรจำลองที่สามารถใช้ในการจำแนกตัวแปรตามของชุดข้อมูลเยอรมันเครดิต (good credit risk/bad credit risk) ได้ดีโดยกำหนดให้ตัวแปรดังกล่าวมีข้อมูลสูญหายถึง 50% และตัวแปรสำหรับการแบ่งแยกข้อมูลส่วนที่สูญหายและไม่สูญหายของตัวแปรดังกล่าวลงบนชุดข้อมูลเยอรมันเครดิต

4. จัดเตรียมชุดข้อมูลให้อยู่ในรูปแบบที่พร้อมสำหรับการสร้างตัวแบบ โดยการทำให้เป็นปกติน้อยที่สุด-มากที่สุด (Min-Max Normalization) และสร้างตัวแปรหุ่น (Dummy Variable Encoding) สำหรับตัวแปรที่เป็นข้อมูลเชิงคุณภาพ สำหรับตัวแปรที่มีข้อมูลสูญหายให้แทนที่ข้อมูลส่วนที่สูญหายด้วยค่าคงที่ค่าใดค่าหนึ่ง

4. สร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตที่ถูกเพิ่มตัวแปรจำลอง โดยใช้ตัวแบบที่ถูกขยาย

5. ทดสอบประสิทธิผลของตัวแบบที่ถูกขยายจากความถูกต้องของโครงสร้างของต้นไม้ตัดสินใจที่ได้และสรุปผล

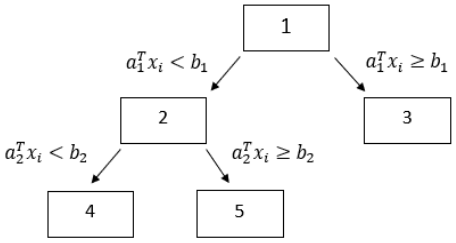
บทที่ 4
ผลงานวิจัย

4.1 เวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุด

ในงานวิจัยส่วนนี้ผู้วิจัยได้ทำการทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลขนาดเล็กซึ่งจะทำให้เห็นโครงสร้างของต้นไม้ตัดสินใจได้อย่างชัดเจน เพื่อที่จะทดสอบและเลือกใช้ตัวแบบและวิธีการที่ให้ประสิทธิผลที่ถูกต้อง

ต้นไม้จำแนกประเภทที่ดีที่สุด เป็นการมองการสร้างต้นไม้ตัดสินใจแบบจำแนกเป็นตัวแบบเชิงเส้นจำนวนเต็มแบบผสม (Mixed-Integer Optimization) ซึ่งจากการแก้ปัญหาตัวแบบดังกล่าวจะให้ผลลัพธ์ในรูปแบบของค่าของตัวแปรตัดสินใจ (Decision Variable) โดยตัวแปรตัดสินใจของตัวแบบเชิงเส้นจำนวนเต็มแบบผสมที่บอกลักษณะการแบ่งของต้นไม้จำแนกประเภทที่ดีที่สุด ได้แก่ d_t a_{jt} และ b_t ซึ่งมีคำอธิบายดังตารางที่ 2.1

ต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลสร้างตัวแบบขนาด n ตัวอย่าง p ตัวแปรต้น และ 1 ตัวแปรตามซึ่งแบ่งเป็น K คลาส เขียนแทนโดย $x_i = [x_i^1, x_i^2, \dots, x_i^p],\ i = 1, \dots, n$ และ $y_i \in \{1, \dots, K\}$ สามารถเขียนเป็นต้นไม้ตัดสินใจได้ ดังรูปที่ 4.1



รูปที่ 4.1 ต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดที่มีการแบ่งที่โหนด 1 และโหนด 2

เพื่อทดสอบประสิทธิผลของตัวแบบเชิงเส้นจำนวนเต็มแบบผสมที่ใช้ในการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดที่เสนอโดย Bertsimas และ Dunn (2017) ผู้วิจัยจึงทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลจำลองขนาด 6 ตัวอย่าง 2 ตัวแปรต้นและตัวแปรตามซึ่งแบ่งเป็น 2 คลาส เขียนแทนโดย $x_i = [x_i^1, x_i^2],\ i = 1, \dots, 6$ และ $y_i \in \{1,2\}$ โดยในงานวิจัยนี้ผู้วิจัยใช้ gurobi 9.5.1 (Gurobi Optimization Inc, 2021a) ในการแก้ปัญหาตัวแบบเชิงเส้นจำนวนเต็มแบบผสม ซึ่งเป็นไลบรารี่ (Library) ชื่อ gurobipy บนภาษาไพธอน (Python) โดยใช้คอมพิวเตอร์ที่มีตัวประมวลผล 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz แรม 16 GB จากการแก้ปัญหาตัวแบบพบว่าผลลัพธ์ของตัวแปรตัดสินใจ (Decision Variable) ที่ได้ไม่อยู่ในรูปแบบที่สามารถนำไปสร้างเป็นต้นไม้ตัดสินใจได้ โดยตัวแปรตัดสินใจ d_t, a_{jt}, b_t ซึ่งเป็นตัวแปรที่บอกลักษณะการแบ่งของต้นไม้ตัดสินใจ มีค่าเป็น 0 ทั้งหมด กล่าวคือไม่มีการแบ่งเกิดขึ้นบนต้นไม้ตัดสินใจ อีกทั้งยังไม่สอดคล้องกับผลลัพธ์

ของตัวแปรตัดสินใจที่เหลือ ผู้วิจัยจึงได้ทดลองสร้างโดยใช้ตัวแบบเชิงเส้นจำนวนเต็มแบบผสมในงานวิจัยของ Lin และ Tang (2021) ซึ่งมีวัตถุประสงค์ (Objective) และเงื่อนไขแสดงขอบเขต (Constraints) ดังนี้

$$\min \ \frac{1}{\hat{L}} \sum_{t \in \tau_L} L_t + \alpha \sum_{t \in \tau_B} d_t \tag{1}$$

$$s.t. \ \ L_t \geq N_t - N_{kt} - n(1 - c_{kt}), \ \ k = 1, \ldots, K, \ \ \forall t \in \tau_L, \tag{2}$$

$$L_t \leq N_t - N_{kt} + n c_{kt}, \ \ k = 1, \ldots, K, \ \ \forall t \in \tau_L, \tag{3}$$

$$L_t \geq 0, \ \ \forall t \in \tau_L, \tag{4}$$

$$N_{kt} = \frac{1}{2} \sum_{i=1}^{n} (1 + Y_{ik}) z_{it}, \ \ k = 1, \ldots, K, \ \ \forall t \in \tau_L, \tag{5}$$

$$N_t = \sum_{i=1}^{n} z_{it}, \ \ \forall t \in \tau_L, \tag{6}$$

$$\sum_{k=1}^{K} c_{kt} = l_t, \ \ \forall t \in \tau_L, \tag{7}$$

$$a_m^T x_i \geq b_m - (1 - z_{it}), \ \ i = 1, \ldots, n, \forall t \in \tau_B, \ \ \forall m \in A_R(t), \tag{8}$$

$$a_m^T (x_i + \epsilon) \leq b_m + (1 + \epsilon_{max})(d_t - z_{it}), \ \ i = 1, \ldots, n, \ \ \forall t \in \tau_B, \ \ \forall m \in A_L(t), \tag{9}$$

$$\sum_{t \in \tau_L} z_{it} = 1, \ \ i = 1, \ldots, n, \tag{10}$$

$$z_{it} \leq l_t, \ \ \forall t \in \tau_L, \tag{11}$$

$$\sum_{n=1}^{n} z_{it} \geq N_{min} l_t, \ \ \forall t \in \tau_L, \tag{12}$$

$$\sum_{i=1}^{n} a_{jt} = d_t, \ \ \forall t \in \tau_B, \tag{13}$$

$$0 \leq b_t \leq d_t, \ \ \forall t \in \tau_B, \tag{14}$$

$$d_t \leq d_{p(t)}, \ \ \forall t \in \tau_B \backslash \{1\}, \tag{15}$$

$$z_{it}, l_t \in \{0, 1\}, \ \ i = 1, \ldots, n, \ \ \forall t \in \tau_L \tag{16}$$

$$a_{jt}, d_t \in \{0, 1\}, \ \ j = 1, \ldots, p, \ \ \forall t \in \tau_B \tag{17}$$

ซึ่งแตกต่างจากตัวแบบที่เสนอโดย Bertsimas และ Dunn (2017) ที่เงื่อนไขแสดงขอบเขต (8) จากตัวแบบข้างต้น ซึ่งเป็นเงื่อนไขที่ใช้ในการกำหนดลักษณะการแบ่งของต้นไม้ตัดสินใจ

จาก

$$a_m^T(x_i + \epsilon) \leq b_m + (1 + \epsilon_{max})(1 - z_{it}),\ \ i = 1, \ldots, n,\ \ \forall t \in \tau_B,\ \ \forall m \in A_L(t),$$

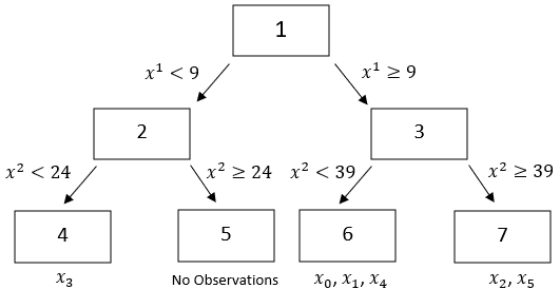
เป็น

$$a_m^T(x_i + \epsilon) \leq b_m + (1 + \epsilon_{max})(d_t - z_{it}),\ \ i = 1, \ldots, n,\ \ \forall t \in \tau_B,\ \ \forall m \in A_L(t),$$

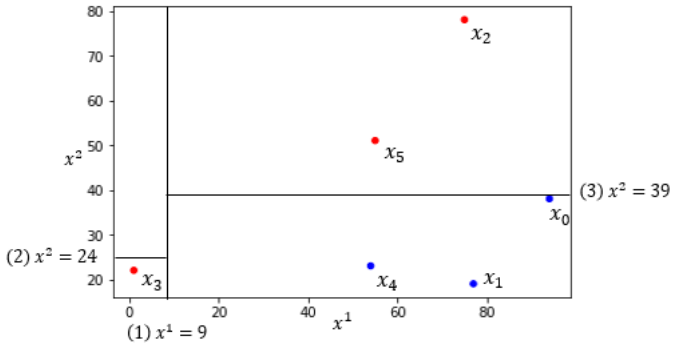
จากการแก้ปัญหาตัวแบบดังกล่าวให้ผลลัพธ์ของตัวแปรตัดสินใจที่บอกโครงสร้างของต้นไม้ตัดสินใจ d_t, a_{jt}, b_t ดังนี้

Decision Variable	Value	Decision Variable	Value
a_{11}	1	d_1	1
a_{12}	0	d_2	1
a_{13}	0	d_3	1
a_{21}	0	b_1	0.085
a_{22}	1	b_2	0.085
a_{23}	1	b_3	0.339

ผลลัพธ์ของตัวแปรตัดสินใจ d_t สำหรับทุกโหนดกิ่ง t ให้ค่าเป็น 1 อธิบายได้ว่ามีการแบ่งเกิดขึ้นที่ทุกโหนดกิ่ง สามารถนำไปสร้างเป็นต้นไม้ตัดสินใจได้ ดังรูปที่ 4.2 และ 4.3



รูปที่ 4.2 ต้นไม้ตัดสินใจจากการใช้ตัวแบบของ Lin และ Tang (2021)



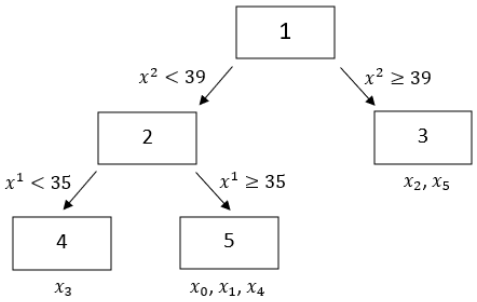
รูปที่ 4.3 แสดงข้อมูลเป็นส่วนจากต้นไม้ติดสินใจรูปที่ 4.2

ซึ่งพบว่า มีการแบ่งที่โหนด 2 ดังรูปที่ 4.2 และ 4.3 ซึ่งเป็นโหนดที่ไม่ควรจะมีการแบ่งเกิดขึ้นเนื่องจากมีจำนวนตัวอย่างเท่ากับ 1 ต้นไม้ติดสินใจดังกล่าวจึงมีลักษณะที่ไม่ถูกต้อง ผู้วิจัยจึงได้ทำการกำหนดค่าพารามิเตอร์ความซับซ้อน (α) ตั้งต้นจากเดิมเป็น 0 ให้เป็นค่าบวกใกล้เคียง 0 ซึ่งเป็นค่าที่ไม่ส่งผลต่อความซับซ้อนของต้นไม้ตัดสินใจ เพื่อเป็นการกำหนดให้โหนดกิ่งไม่มีการแบ่งหากไม่ทำให้ค่าการจำแนกผิดลดลงตามวัตถุประสงค์ของตัวแบบ

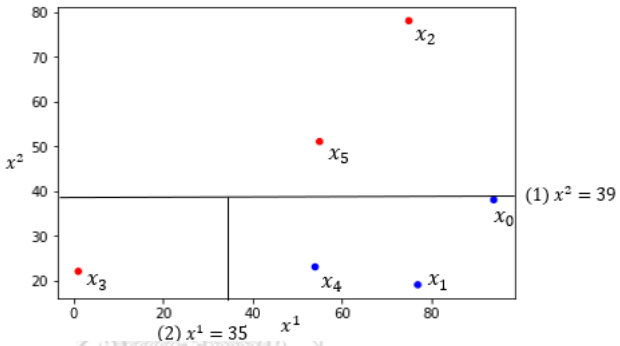
จากการแก้ปัญหาตัวแบบดังกล่าวให้ผลลัพธ์ของตัวแปรตัดสินใจที่บอกโครงสร้างของต้นไม้ตัดสินใจ d_t, a_{jt}, b_t ดังนี้

Decision Variable	Value	Decision Variable	Value
a_{11}	0	d_1	1
a_{12}	1	d_2	1
a_{13}	0	d_3	0
a_{21}	1	b_1	0.339
a_{22}	0	b_2	0.368
a_{23}	0	b_3	0

ซึ่งผลลัพธ์ที่ได้เป็นที่น่าพอใจ สามารถนำไปสร้างเป็นต้นไม้ตัดสินใจที่มีลักษณะโครงสร้างถูกต้อง ดังรูปที่ 4.4 และ 4.5



รูปที่ 4.4 ต้นไม้ตัดสินใจจากการใช้ตัวแบบของ Lin และ Tang (2021) โดยกำหนดค่าพารามิเตอร์ความซับซ้อนเป็นค่าบวกใกล้เคียง 0



รูปที่ 4.5 แสดงข้อมูลเป็นส่วนจากต้นไม้ติดสินใจรูปที่ 2.4

ในงานวิจัยของ Bertsimas และ Dunn (2017) ได้เสนออีกหนึ่งวิธีในการกำหนดค่าพารามิเตอร์ความซับซ้อน (α) สำหรับสร้างต้นไม้จำแนกประเภทที่ดีที่สุด สามารถทำได้โดยเพิ่มเงื่อนไขแสดงขอบเขตเพื่อกำหนดจำนวนครั้งในการแบ่งสูงสุดของโหนดกิ่ง (τ_B) ดังสมการ (18)

$$\sum_{t \in \tau_B} d_t \leq C, \tag{18}$$

ผู้วิจัยจึงใช้ค่าพารามิเตอร์จำนวนครั้งในการแบ่งสูงสุด (C) ที่เงื่อนไขแสดงขอบเขต (18) เป็นการกำหนดค่าพารามิเตอร์ความซับซ้อน และกำหนดให้ค่าพารามิเตอร์ความซับซ้อน (α) ที่วัตถุประสงค์ของตัวแบบ (1) เป็นค่าคงที่บวกใกล้เคียง 0

สรุปตัวแบบสำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดที่ให้ประสิทธิผลเป็นที่น่าพอใจซึ่งจะใช้สำหรับการพัฒนาเวิร์กโฟลว์สำหรับสร้างต้นไม้จำแนกประเภทที่ดีที่สุดในงานวิจัยฉบับนี้ ได้ดังนี้

$$\min \frac{1}{\hat{L}} \sum_{t \in \tau_L} L_t + \alpha \sum_{t \in \tau_B} d_t$$

$s.t.$ $\sum_{t \in \tau_B} d_t \leq C,$

$L_t \geq N_t - N_{kt} - n(1 - c_{kt}),\ k = 1, \ldots, K,\ \forall t \in \tau_L,$

$L_t \leq N_t - N_{kt} + nc_{kt},\ k = 1, \ldots, K,\ \forall t \in \tau_L,$

$L_t \geq 0,\ \forall t \in \tau_L,$

$N_{kt} = \frac{1}{2} \sum_{i=1}^{n} (1 + Y_{ik}) z_{it},\ k = 1, \ldots, K,\ \forall t \in \tau_L,$

$N_t = \sum_{i=1}^{n} z_{it},\ \forall t \in \tau_L,$

$\sum_{k=1}^{K} c_{kt} = l_t,\ \forall t \in \tau_L,$

$a_m^T x_i \geq b_m - (1 - z_{it}),\ i = 1, \ldots, n, \forall t \in \tau_B,\ \forall m \in A_R(t),$

$a_m^T (x_i + \epsilon) \leq b_m + (1 + \epsilon_{max})(d_t - z_{it}),\ i = 1, \ldots, n,\ \forall t \in \tau_B,\ \forall m \in A_L(t),$

$\sum_{t \in \tau_L} z_{it} = 1,\ i = 1, \ldots, n,$

$z_{it} \leq l_t,\ \forall t \in \tau_L,$

$\sum_{n=1}^{n} z_{it} \geq N_{min} l_t,\ \forall t \in \tau_L,$

$\sum_{i=1}^{n} a_{jt} = d_t,\ \forall t \in \tau_B,$

$0 \leq b_t \leq d_t,\ \forall t \in \tau_B,$

$d_t \leq d_{p(t)},\ \forall t \in \tau_B \backslash \{1\},$

$z_{it}, l_t \in \{0, 1\},\ i = 1, \ldots, n,\ \forall t \in \tau_L$

$a_{jt}, d_t \in \{0, 1\},\ j = 1, \ldots, p,\ \forall t \in \tau_B$

โดยมีความลึกสูงสุด (D) จำนวนครั้งในการแบ่งสูงสุด (C) และจำนวนตัวอย่างที่โหนดใบต่ำสุด (N_{min}) เป็นค่าพารามิเตอร์สำหรับต้นไม้จำแนกประเภทที่ดีที่สุด และกำหนดให้ค่าพารามิเตอร์ความซับซ้อน (α) เป็นค่าคงที่บวกใกล้เคียง 0 ดังตารางที่ 4.1 และ 4.2

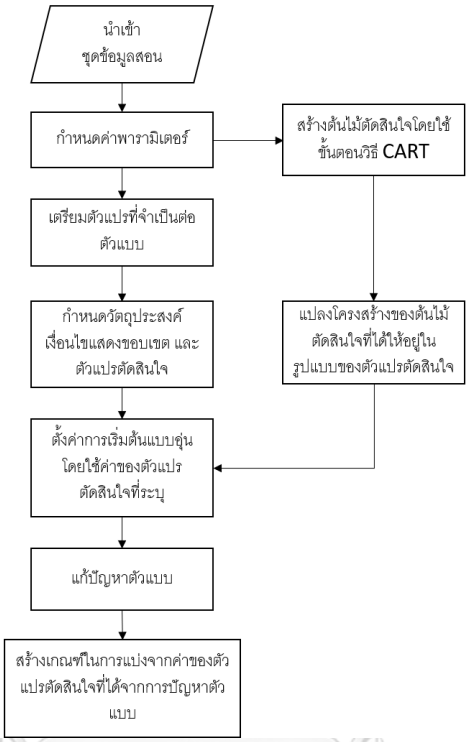
ตารางที่ 4.1 พารามิเตอร์สำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดและคำอธิบาย

พารามิเตอร์	คำอธิบาย
N_{min}	จำนวนตัวอย่างที่โหนดใบต่ำสุด
$Max\ depth\ (D)$	ความลึกสูงสุดของต้นไม้ตัดสินใจ
$Number\ of\ split\ (C)$	จำนวนครั้งในการแบ่งของต้นไม้ตัดสินใจ

ตาราง 4.2 ค่าคงที่สำหรับตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดและคำอธิบาย

ค่าคงที่	คำอธิบาย
n	จำนวนตัวอย่างของชุดข้อมูลสอน
p	จำนวนคอลัมน์ (column) ของชุดข้อมูลสอน
K	จำนวนคลาสของตัวแปรตาม
τ_B	ตำแหน่งของแต่ละโหนดใบ
τ_L	ตำแหน่งของแต่ละโหนดกิ่ง
$\hat{L}$	ค่าทำนายผิดพื้นฐานเท่ากับค่าทำนายผิดหากทำนายด้วยคลาสที่มีจำนวนเยอะที่สุด
Y_{ik}	เมทริกที่ใช้ระบุว่าจำนวนตัวอย่างที่มีตัวแปรตามเป็นคลาส k ที่โหนดใบ
ε_j	ค่าส่วนต่างบวกที่น้อยที่สุดของแต่ละตัวอย่างในตัวแปร j
α	ค่าพารามิเตอร์ความซับซ้อนเท่ากับค่าคงที่บวกใกล้เคียง 0

จากการศึกษาขั้นตอนการพัฒนาในงานวิจัยของ Bertsimas และ Dunn (2017) และการทดลองเพื่อให้ได้ตัวแบบและวิธีการที่ให้ประสิทธิผลที่ถูกต้อง ผู้วิจัยได้ทำการพัฒนาเวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบข้างต้นซึ่งสามารถเขียนเป็นผังงานได้ ดังรูปที่ 4.6



รูปที่ 4.6 ผังงานเวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุด

โดยมีรายละเอียดของเวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดดังต่อไปนี้

นำเข้าชุดข้อมูลสอน (x_i, y_i) ในรูปแบบที่ทำให้เป็นปกติ (Normalization) โดยใช้การทำให้เป็นปกติน้อยที่สุด-มากที่สุด (Min-Max Normalization)

กำหนดค่าพารามิเตอร์ความลึกสูงสุด (D) จำนวนครั้งในการแบ่งสูงสุด (C) และจำนวนตัวอย่างที่โหนดใบต่ำสุด (N_{min}) ดังตาราง 4.1 โดยค่าตั้งต้นของจำนวนครั้งในการแบ่งสูงสุดมีค่าเท่ากับ $2^D - 1$ ค่าตั้งต้นของจำนวนตัวอย่างที่โหนดใบต่ำสุดเท่ากับ 1

สร้างต้นไม้ตัดสินใจโดยใช้ขั้นตอนวิธี Classification and Regression (CART) จากชุดข้อมูลสอน และค่าพารามิเตอร์ที่กำหนดในขั้นตอนก่อนหน้า

เตรียมค่าของตัวแปรที่จำเป็นต่อตัวแบบเชิงเส้นจำนวนเต็มแบบผสม สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดดังตาราง 4.2 โดยจำนวนคลาสของตัวแปรตาม (k) จำนวนตัวแปรต้น (p) และขนาดของตัวอย่าง (n) เป็นไปตามชุดข้อมูลสอน จำนวนโหนดทั้งหมด (T) มีค่าเท่ากับ $2^{D+1} - 1$ โดย D เท่ากับความลึกสูงสุดที่กำหนด ตำแหน่ง

ของโหนดกิ่ง (τ_B) เท่ากับ $\{1, \dots, \lfloor T/2 \rfloor\}$ และโหนดใบ (τ_L) เท่ากับ $\{\lfloor T/2 \rfloor + 1, \dots, T\}$ ค่าเอปซีลอน (ε_j) หรือค่าส่วนต่างบวกที่น้อยที่สุดของแต่ละตัวอย่างในตัวแปร j มีค่าเท่ากับ $\min\{x_j^{(i+1)} - x_j^{(i)} | x_j^{(i+1)} - x_j^{(i)}, i = 1, \dots, n-1\}$ และค่าเอปซีลอนสูงสุด (ε_{max}) มีค่าเท่ากับ $\max_j\{\varepsilon_j\}$ เมทริกซ์ Y (Y_{ik}) เป็นเมทริกที่ใช้ระบุว่าจำนวนตัวอย่างที่มีตัวแปรตามเป็นคลาส k ที่โหนดใบโดย Y_{ik} เท่ากับ +1 เมื่อ $y_i = k$ และ เท่ากับ -1 เมื่อ $y_i \neq k$ ค่าการทำนายผิดพื้นฐาน ($\hat{L}$) เท่ากับค่าการทำนายผิดหากทำนายชุดข้อมูลสอนด้วยค่าของตัวแปรตามที่มีจำนวนตัวอย่างมากที่สุด

เตรียมวัตถุประสงค์ (Objective) เงื่อนไขแสดงขอบเขต (Constraints) ของตัวแบบ และตัวแปรตัดสินใจ (Decision Variable) ที่ใช้ในตัวแบบตามตารางที่ 2.1

นำต้นไม้ตัดสินใจที่ได้จากการสร้างโดยใช้ขั้นตอนวิธี CART ในขั้นตอนก่อนหน้ามาแปลงให้อยู่ในรูปแบบของค่าของตัวแปรตัดสินใจ และกำหนดค่าการเริ่มต้นแบบอุ่น (Warm Start) ของตัวแบบโดยใช้ค่าของตัวแปรตัดสินใจดังกล่าว

กำหนดเวลาจำกัด (Time limit) ในการแก้ปัญหาเชิงเส้นจำนวนเต็มแบบผสม แก้ปัญหาตัวแบบและแปลงค่าของตัวแปรตัดสินใจที่ได้จากการแก้ปัญหาให้อยู่ในรูปแบบของต้นไม้ตัดสินใจโดยสร้างเกณฑ์ในการแบ่งที่แต่ละโหนดกิ่งและระบุค่าทำนายตัวแปรตามที่แต่ละโหนดใบ

4.2 ต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลแบบจำลองคะแนนเครดิต

ในส่วนนี้ผู้วิจัยได้ทำการสร้างต้นไม้จำแนกประเภทที่ดีที่สุด บนชุดข้อมูลเยอรมันเครดิตจาก *UCI* Machine Learning *Repository* (Hofmann, 1994) ขนาด 1,000 ตัวอย่าง โดยมีตัวแปรตาม 1 ตัวแปร ประกอบไปด้วย 2 คลาส คือ good credit risk และ bad credit risk และมีตัวแปรต้นจำนวน 20 ตัวแปร เขียนแทนโดย $x_i = [x_i^1, x_i^2, \dots, x_i^{20}]$ และ $y_i \in \{0,1\}$ โดย $i = 1, \dots, 1000$ ตัวแปรต้นมีคำอธิบายดังตาราง 4.1

ตาราง 4.1 ตัวแปรต้นและคำอธิบายของชุดข้อมูลเยอรมันเครดิต

ชื่อตัวแปร	คำอธิบาย
Status of existing checking account	สถานะของบัญชี
duration in month	ระยะเวลาการกู้ยืม
credit history	ประวัติเครดิต
purpose	วัตถุประสงค์การกู้ยืม
credit amount	วงเงิน
savings account/bonds	สถานะบัญชีออมทรัพย์/พันธบัตร
present employment since	ระยะเวลาการเป็นพนักงานในที่ทำงานปัจจุบัน
installment rate in percentage of disposable income	อัตราการผ่อนชำระคิดเป็นเปอร์เซ็นต์ต่อรายได้
Personal status and sex	สถานะและเพศ
Other debtors/guarantors	ลูกหนี้/ผู้ค้ำประกัน

Present residence since	ระยะเวลาการอยู่อาศัยในที่อยู่ปัจจุบัน
Property	ทรัพย์สิน
Age in years	อายุ
Other installment plans	แผนการผ่อนชำระ
Housing	สถานะที่อยู่
Number of existing credits at this bank	จำนวนเครดิตที่เหลือในธนาคารนี้
Job	ประเภทอาชีพ
Number of people being liable to provide maintenance for	จำนวนผู้ที่อยู่ในความรับผิดชอบเลี้ยงดู
Telephone	มีข้อมูลโทรศัพท์หรือไม่
foreign worker	เป็นแรงงานต่างชาติหรือไม่

เพื่อทดสอบผลลัพธ์จากวิธีการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลจริงโดยใช้ตัวแบบและเวิร์กโฟลว์จากงานวิจัยในส่วนที่ 2 และเปรียบเทียบประสิทธิภาพกับต้นไม้จำแนกประเภทที่สร้างโดยใช้ขั้นตอนวิธี Classification and Regression (CART) บนชุดข้อมูลดังกล่าว

เพื่อกำหนดเวลาจำกัด (Time limit) ในการแก้ปัญหาตัวแบบเชิงเส้นจำนวนเต็มแบบผสมสำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต ผู้วิจัยได้สร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตขนาด 1,000 ตัวอย่าง โดยกำหนดพารามิเตอร์ความลึกสูงสุดเท่ากับ 4 จำนวนครั้งในการแบ่งเท่ากับ 15 และจำนวนตัวอย่างที่โหนดใบต่ำสุดเท่ากับ 5% ของจำนวนตัวอย่างข้อมูลที่ใช้สร้างตัวแบบ จากการกำหนดเวลาจำกัดไว้ที่ 300 ถึง 10,800 วินาที ให้ความแม่นยำบนชุดข้อมูลสอนได้ผลดังตาราง 4.2

ตาราง 4.2 ค่าความถูกต้องบนชุดข้อมูลสอนของต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตที่การกำหนดเวลาจำกัด (Time limit) ในการสร้างตัวแบบ 300 ถึง 10,800 วินาที

Time limit (Seconds)	In-sample accuracies (%)
300	71.8
600	72.6
1,800	73
3,600	73.8
7,200	75.8
9,000	75.8
10,800	75.8

การให้เวลาจำกัด (Time limit) ในการแก้ปัญหาตัวแบบที่มากขึ้นทำให้ผลลัพธ์ของต้นไม้ตัดสินใจดีขึ้นซึ่งหมายถึงค่าการทำนายผิดที่น้อยลงบนชุดข้อมูลสอนตามวัตถุประสงค์ (Objective) ของตัวแบบ จากตารางที่ 4.2 การกำหนดเวลาจำกัดที่มากกว่า 7,200 วินาที หรือ 2 ชม. ให้ผลลัพธ์ของต้นไม้ตัดสินใจไม่ต่างจากเดิม ในงานวิจัยนี้ผู้วิจัยจึงกำหนดเวลาจำกัดไว้ที่ 2 ชม. สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต

เพื่อเปรียบเทียบประสิทธิภาพกับต้นไม้จำแนกประเภทที่สร้างโดยใช้ขั้นตอนวิธี Classification and Regression (CART) บนชุดข้อมูลดังกล่าว ทำการกำหนดการทดลองโดยสุ่มแบ่งชุดข้อมูลเป็น 2 ส่วน ข้อมูลสร้างตัวแบบ 75% และข้อมูลชุดทวนสอบ 25% และแบ่งชุดข้อมูลสร้างตัวแบบออกเป็น 3 ส่วน สำหรับทำการทดสอบแบบไขว้ (K-fold cross-validate) เพื่อหาค่าพารามิเตอร์ความซับซ้อน หรือจำนวนครั้งในการแบ่ง (C) ที่ให้ผลลัพธ์ที่ดีที่สุดสำหรับแต่ละค่าพารามิเตอร์ความลึกสูงสุด (D) 1 ถึง 4 และเลือกใช้ค่าพารามิเตอร์ที่ได้ในการสร้างตัวแบบ สำหรับค่าพารามิเตอร์จำนวนตัวอย่างที่โหนดใบต่ำสุด ($N_{\min}$) จะกำหนดไว้ที่ 5% ของจำนวนตัวอย่างของข้อมูลที่ใช้สร้างตัวแบบ กำหนดเวลาจำกัด (Time limit) ในการแก้ปัญหาตัวแบบสำหรับต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดไว้ที่ 2 ชม. และกำหนดค่าพารามิเตอร์ในลักษณะเดียวกันสำหรับการสร้างต้นไม้ตัดสินใจโดยใช้ขั้นตอนวิธี CART ทำการทดลองซ้ำ 3 ครั้งโดยสุ่มแบ่งชุดข้อมูลใหม่เพื่อลดผลกระทบต่อการประเมินประสิทธิภาพจากการแบ่งชุดข้อมูล และประเมินค่าความแม่นยำเฉลี่ยจากการทดลองทั้ง 3 ครั้ง

ตารางที่ 4.3 ค่าความถูกต้องเฉลี่ยบนชุดข้อมูลสอนและชุดข้อมูลทวนสอบของต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดและต้นไม้ตัดสินใจที่สร้างโดยใช้ขั้นตอนวิธี CART ที่ค่าความลึกสูงสุด 1 ถึง 4

Max. depth	Mean in-sample accuries		Mean out-of-sample accuracies	
	CART (%)	OCT (%)	CART (%)	OCT (%)
1	69.8	71	70.7	71.1
2	71	73.5	70.5	71.1
3	72.8	73.9	70.7	73.9
4	74.2	74.6	71.2	72.4

จากการทดลองพบว่าตัวแบบและเวิร์กโฟลว์จากงานวิจัยในส่วนที่ 2 สามารถใช้ในการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดได้อย่างถูกต้องบนชุดข้อมูลจริง และจากตารางที่ 4.3 พบว่าค่าความถูกต้องจากการสร้างต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตมีค่าสูงกว่าต้นไม้ตัดสินใจที่สร้างโดยใช้ขั้นตอนวิธี CART บนทุกค่าพารามิเตอร์ความลึกสูงสุด 0.4 ถึง 3.2%

4.3 การขยายตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดให้รองรับชุดข้อมูลที่ตัวแปรต้นมีค่าสูญหายจำนวนมาก

ข้อดีของการพัฒนาเวิร์กโฟลว์โดยใช้โปรแกรมหาคำตอบสำหรับปัญหาเชิงเส้นจำนวนเต็มแบบผสมคือความสามารถในการขยายตัวแบบให้รองรับเงื่อนไขเพิ่มเติมได้ งานวิจัยในส่วนนี้ผู้วิจัยจึงทดลองขยายตัวแบบให้รองรับชุดข้อมูลที่มีตัวแปรต้นสูญหาย (Missing Value) จำนวนมาก เพื่อเป็นหนึ่งทางเลือกในการนำตัวแปรดังกล่าวมาเป็นส่วนหนึ่งในการสร้างตัวแบบโดยไม่ได้รับผลกระทบจากการแทนค่าข้อมูลส่วนที่สูญหาย

โดยการเพิ่มตัวแปรที่ใช้ในการแบ่งแยกข้อมูลส่วนที่สูญหายและไม่สูญหาย และเพิ่มเงื่อนไขแสดงขอบเขต (18) ลงในตัวแบบในงานวิจัยส่วนที่ 2 เพื่อใช้ตัวแปรดังกล่าวในการแบ่งแยกข้อมูลส่วนที่สุญหายและไม่สูญหายออกจากกันที่โหนดพ่อแม่ (Parent node) ก่อนที่จะใช้ข้อมูลส่วนที่ไม่สูญหายในการแบ่งแยกที่โหนดลูก (Child node)

$$\sum_{t=2p(t)}^{2p(t)+1} a_{jt} = a_{f(j)p(t)} \tag{19}$$

$p(t) \in$ ตำแหน่งของโหนดพ่อแม่ของโหนด t, $\forall t \in \tau_B \backslash \{1\}$

$j \in$ ตำแหน่งของตัวแปรที่มีข้อมูลสูญหาย

$f(j) \in$ ตำแหน่งของตัวแปรที่ใช้ในการแบ่งแยกข้อมูลส่วนที่สูญหายและไม่สูญหายของตัวแปร j

โดย a_{jt} เป็นตัวแปรตัดสินใจที่กำหนดว่าตัวแปรต้นตำแหน่งที่ j ถูกใช้ในการแบ่งที่โหนด t หรือไม่

เงื่อนไขแสดงของเขต (19) หากตัวแปรที่มีข้อมูลสูญหาย j ถูกใช้ในการแบ่งที่โหนด t ตัวแปรที่ใช้ในการจำแนกข้อมูลส่วนที่สูญหายและไม่สูญหายของตัวแปร j จะถูกใช้ในการแบ่งที่โหนดพ่อแม่ของโหนด t

เพื่อทดสอบประสิทธิผลของตัวแบบที่ถูกขยาย ผู้วิจัยจึงทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบที่ถูกขยายบนชุดข้อมูลเยอรมันเครดิตจากงานวิจัยในส่วนที่ 2 โดยจำลองตัวแปรที่มีความสามารถในการจำแนกตัวแปรตาม (good credit risk/bad credit risk) ได้ดี แต่มีข้อมูลสูญหายถึง 50% และเพิ่มตัวแปรในการแบ่งแยกข้อมูลส่วนที่สูญหายและไม่สูญหายของตัวแปรดังกล่าว กำหนดชื่อตัวแปรและคำอธิบายดังตาราง 4.4

ตาราง 4.4 ชื่อและคำอธิบายของตัวแปรจำลองและตัวแปรที่ใช้ในการแบ่งแยกข้อมูลสวนที่สูญหายและไม่สูญหายของตัวแปรจำลอง

ชื่อตัวแปร	คำอธิบาย
A mock variable	ตัวแปรที่จำแนกตัวแปรตาม (good credit/bad credit) ได้ดี แต่มีข้อมูลสูญหายจำนวนมาก
Is value of a mock variable not missing?	ค่าของตัวแปร A mock variable เป็นข้อมูลไม่สูญหายใช่หรือไม่

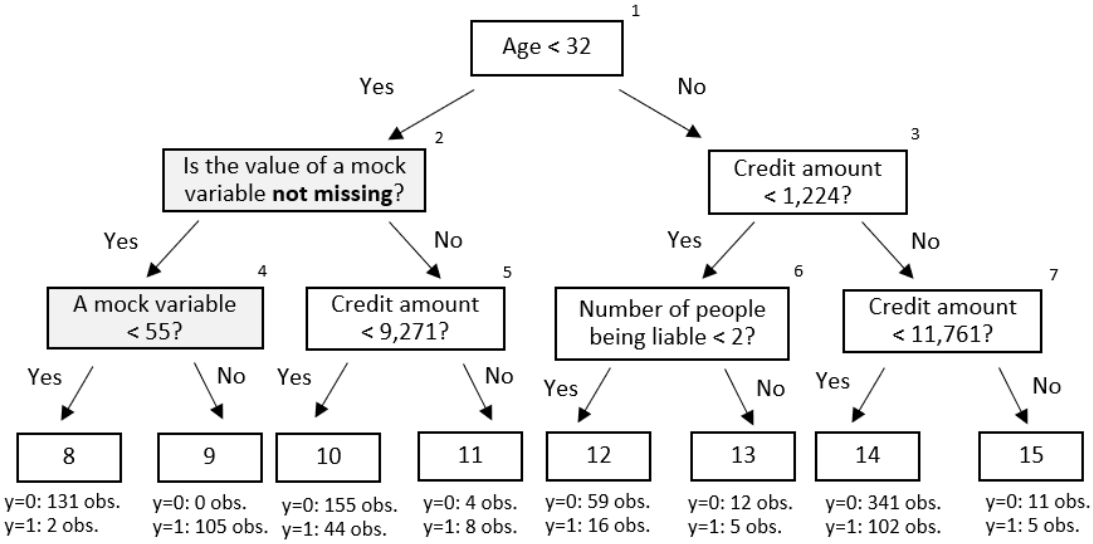
เพิ่มตัวแปรดังกล่าวเป็นตัวแปรต้นลำดับที่ 21 และ 22 ลงบนชุดข้อมูลเยอรมันเครดิต ชุดข้อมูลดังกล่าว เขียนแทนโดย $x_i = [x_i^1, x_i^2, \dots, x_i^{22}]$ และ $y_i \in \{0,1\}$ โดย $i = 1, \dots, 1000$ จากการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้ตัวแบบที่ถูกขยายบนชุดข้อมูลดังกล่าวโดยกำหนดค่าพารามิเตอร์ความลึกสูงสุดเป็น 3 มีตำแหน่งของโหนดกิ่ง $\tau_B \in \{1, \dots, 7\}$ จากเงื่อนไขแสดงขอบเขต (19) โดยมี $t \in \{2, \dots, 7\}$, $p(t) \in \{1,2,3\}$, $j \in \{21\}$ และ $f(j) \in \{22\}$ สามารถเขียนเป็นสมการได้ดังนี้

$$a_{21,2} + a_{21,3} = a_{22,1}$$

$$a_{21,4} + a_{21,5} = a_{22,2}$$

$$a_{21,6} + a_{21,7} = a_{22,3}$$

จากการแก้ปัญหาตัวแบบ ค่าของตัวแปรตัดสินใจ $a_{22,2}$ และ $a_{21,4}$ มีค่าเป็น 1 อธิบายได้ว่า ตัวแปรต้นตำแหน่งที่ 22 (Is value of a mock variable missing?) ถูกใช้ในการแบ่งที่โหนด 2 และ ตัวแปรต้นตำแหน่งที่ 21 (A mock variable) ถูกใช้ในการแบ่งที่โหนด 4 ได้ผลดังรูปที่ 4.7



รูปที่ 4.7 ต้นไม้ตัดสินใจจำแนกประเภทที่ถูกขยายให้รองรับตัวแปรต้นที่มีข้อมูลสูญหายจำนวนมาก

จากรูปที่ 4.7 ตัวแบบที่ถูกขยายและเวิร์กโฟลวที่พัฒนาขึ้นในงานวิจัยส่วนที่ 2 สามารถทำงานได้อย่างมีประสิทธิผลบนชุดข้อมูลเยอรมันเครดิตที่ถูกเพิ่มตัวแปรจำลองที่มีข้อมูลสูญหายจำนวนมาก โดยตัวแปรที่มีข้อมูลสูญหายจำนวนมากถูกแบ่งแยกข้อมูลส่วนที่สูญหายและไม่สูญหายออกจากกันที่โหนด 2 และข้อมูลส่วนที่ไม่สูญหายของตัวแปรดังกล่าวถูกใช้ในการแบ่งที่โหนด 4

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

5.1.1 เวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุด

งานวิจัยในส่วนนี้ผู้วิจัยได้ทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลขนาดเล็กโดยใช้ตัวแบบในงานวิจัยของ Bertsimas และ Dunn (2017) และ Lin และ Tang (2021) การกำหนดค่าพารามิเตอร์ความซับซ้อนและค่าคงที่ เพื่อให้ได้ตัวแบบและวิธีการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดที่ให้ประสิทธิผลที่ถูกต้อง และเสนอเวิร์กโฟลว์สำหรับการสร้างต้นไม้จำแนกประเภทที่ดีที่สุดจากตัวแบบและวิธีการดังกล่าว

จากการทดลองพบว่าการใช้ตัวแบบจากงานวิจัยของ Lin และ Tang (2021) โดยกำหนดค่าพารามิเตอร์ความซับซ้อน (α) เป็นค่าบวกใกล้เคียง 0 และใช้การกำหนดจำนวนครั้งในการแบ่ง (C) เป็นค่าพารามิเตอร์ความซับซ้อนเป็นรูปแบบที่ให้ประสิทธิผลที่ถูกต้อง

5.1.2 ต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิต

เพื่อทดสอบประสิทธิผลเพิ่มเติมและประเมินประสิทธิภาพเบื้องต้น งานวิจัยในส่วนนี้ผู้วิจัยได้สร้างต้นไม้จำแนกประเภทที่ดีที่สุดโดยใช้เวิร์กโฟลว์ที่พัฒนาขึ้นในงานวิจัยส่วนที่ 1 บนชุดข้อมูลเยอรมันเครดิตขนาด 1,000 ตัวอย่าง 20 ตัวแปรต้น โดยจากการทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุด ที่ความลึกสูงสุด 4 โดยกำหนดเวลาจำกัดในการแก้ปัญหาตัวแบบเชิงซ้อนจำนวนเต็มแบบผสม ไว้ที่ 300 ถึง 10,800 วินาที ได้ผลความแม่นยำบนชุดข้อมูลสอนดังตาราง 5.1 และเปรียบเทียบประสิทธิภาพกับต้นไม้ตัดสินใจแบบ CART ที่ความลึกสูงสุด 1-4 ได้ผลลัพธ์ดังตารางที่ 5.2

ตาราง 5.1 ค่าความแม่นยำบนชุดข้อมูลสอนของต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตที่การกำหนดเวลาจำกัด (Time limit) ในการสร้างตัวแบบ 300 ถึง 10,800 วินาที

Time limit (Seconds)	In-sample accuracies (%)
300	71.8
600	72.6
1,800	73
3,600	73.8
7,200	75.8
9,000	75.8
10,800	75.8

ตารางที่ 5.2 ค่าความแม่นยำเฉลี่ยบนชุดข้อมูลสอนและชุดข้อมูลทวนสอบของต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดและต้นไม้ตัดสินใจที่สร้างโดยใช้ขั้นตอนวิธี CART ที่ค่าความลึกสูงสุด 1 ถึง 4

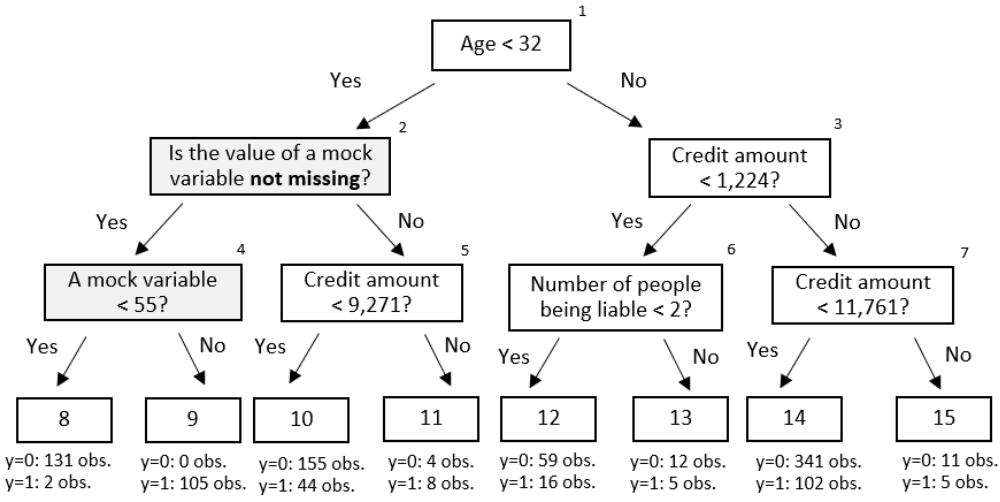
Max. depth	Mean in-sample accuries		Mean out-of-sample accuracies	
	CART (%)	OCT (%)	CART (%)	OCT (%)
1	69.8	71	70.7	71.1
2	71	73.5	70.5	71.1
3	72.8	73.9	70.7	73.9
4	74.2	74.6	71.2	72.4

จากผลดังตารางที่ 5.1 พบว่าที่การกำหนดเวลาจำกัดในการแก้ปัญหาตัวแบบที่มากกว่า 7,200 วินาที หรือ 2 ชม. บนชุดข้อมูลดังกล่าวให้ผลลัพธ์ของต้นไม้ตัดสินใจไม่ต่างจากเดิม ในการสร้างต้นไม้สร้างต้นไม้จำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตเพื่อเปรียบเทียบประสิทธิภาพ ผู้วิจัยจึงใช้เวลาจำกัดในการแก้ปัญหาไว้ที่ 2 ชม.

จากตารางที่ 5.2 ค่าความแม่นยำจากการสร้างต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดบนชุดข้อมูลเยอรมันเครดิตมีค่าสูงกว่าต้นไม้ตัดสินใจที่สร้างโดยใช้ขั้นตอนวิธี CART บนทุกค่าพารามิเตอร์ความลึกสูงสุด 0.4 ถึง 3.2%

5.1.3 การขยายตัวแบบต้นไม้จำแนกประเภทที่ดีที่สุดให้รองรับชุดข้อมูลที่ตัวแปรต้นมีค่าสูญหายจำนวนมาก

งานวิจัยในส่วนนี้ผู้วิจัยจึงทดลองขยายตัวแบบให้รองรับชุดข้อมูลที่มีตัวแปรต้นสูญหาย (Missing Value) จำนวนมาก เพื่อเป็นหนึ่งทางเลือกในการสร้างตัวแบบบนชุดข้อมูลที่มีลักษณะดังกล่าว โดยจากการทดลองสร้างต้นไม้จำแนกประเภทที่ดีที่สุดจากตัวแบบที่ถูกขยาย บนชุดข้อมูลเยอรมันเครดิตที่ถูกเพิ่มตัวแปรจำลองที่มีข้อมูลสูญหายจำนวนมากได้ผลดังรูป 5.1



รูปที่ 5.1 ต้นไม้ตัดสินใจจำแนกประเภทที่ถูกขยายให้รองรับตัวแปรต้นที่มีข้อมูลสูญหายจำนวนมาก

ตัวแบบที่ถูกขยายสามารถทำงานได้อย่างมีประสิทธิผลบนชุดข้อมูลดังกล่าว โดยจากรูปที่ 5.1 ตัวแปรที่มีข้อมูลสูญหายจำนวนมากถูกแบ่งแยกข้อมูลส่วนที่สูญหายและไม่สูญหายออกจากกันที่โหนด 2 และข้อมูลส่วนที่ไม่สูญหายของตัวแปรดังกล่าวถูกใช้ในการแบ่งที่โหนด 4

5.2 สรุปและอภิปลายผล

งานวิจัยนี้ได้เสนอเวิร์กโฟลว์ของการสร้างต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดที่ให้ผลลัพธ์ที่สามารถนำมาสร้างเป็นต้นไม้ตัดสินใจได้อย่างถูกต้อง และเปรียบเทียบประสิทธิภาพกับต้นไม้ตัดสินใจแบบ CART บนชุดข้อมูลเยอรมันเครดิต พบว่าต้นไม้จำแนกประเภทที่ดีที่สุดให้อัตราความถูกต้องสูงกว่าต้นไม้ตัดสินใจทั้งบนชุดข้อมูลสร้างตัวแบบและบนชุดข้อมูลทวนสอบที่ 0.4% ถึง 3.2% และจากการขยายตัวแบบของต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดให้รองรับตัวแปรต้นที่มีข้อมูลสูญหายจำนวนมาก ตัวแบบที่ถูกขยายสามารถทำงานอย่างมีประสิทธิผลบนเวิร์กโฟลว์ที่พัฒนาขึ้น

แม้ว่าประสิทธิภาพของต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดจากการทดลองในงานวิจัยนี้ และงานวิจัยที่ผ่านมาทั้งของ Bertsimas และ Dunn (2017) และ Lin และ Tang (2021) พบว่าต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดให้อัตราความถูกต้องสูงกว่าต้นไม้ตัดสินใจแบบ CART แต่ต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดยังมีข้อจำกัดในเรื่องของเวลาที่ใช้ในแก้ปัญหาเชิงเส้นจำนวนเต็มแบบผสมซึ่งเป็นขั้นตอนหลักในการสร้างต้นไม้ตัดสินใจดังกล่าว ข้อมูลที่ใช้ในงานวิจัยชุดนี้มีขนาด 1,000 ตัวอย่าง 20 ตัวแปรต้น พบว่าเวลาจำกัดที่เหมาะสมในการสร้างต้นไม้ตัดสินใจจำแนกประเภทที่ดีที่สุดบนข้อมูลดังกล่าวอยู่ที่ 2 ชม. ในขณะที่ต้นไม้ตัดสินใจแบบ CART ใช้เวลาในการสร้างน้อยกว่า 1 วินาที แต่จากความสามารถในการคำนวณของตัวแก้ปัญหาตัวแบบเชิงเส้นจำนวนเต็มแบบผสมที่พัฒนาขึ้นอย่างต่อเนื่อง (Gurobi Optimization Inc, 2021b) อีกทั้งประสิทธิภาพของคอมพิวเตอร์ที่เพิ่มมากขึ้น จึงอาจจะทำให้ข้อจำกัดดังกล่าวลดน้อยลงในอนาคต และจากข้อดีด้านความยืดหยุ่นของต้นไม้จำแนกประเภทที่ดีที่สุดที่สามารถปรับแต่งและขยายตัวแบบให้รองรับเงื่อนไขเพิ่มเติม รวมถึงความง่ายต่อการอธิบายตัวแบบเช่นเดียวกับต้นไม้ตัดสินใจแบบ CART ผู้วิจัยจึงมองว่าต้นไม้จำแนกประเภทที่ดีที่สุดเป็นอีกหนึ่งตัวเลือกที่น่าสนใจสำหรับการพัฒนาตัวแบบการเรียนรู้ของเครื่องที่ต้องการความง่ายต่อการอธิบาย

บรรณานุกรม

1. Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine learning*, *106*(7), 1039-1082.
2. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis. https://books.google.co.th/books?id=JwQx-WOmSyQC
3. Gurobi Optimization Inc. (2021a). *Gurobi 9.5 Optimizer Reference Manual* https://www.gurobi.com
4. Gurobi Optimization Inc. (2021b). *Gurobi 9.5 performance benchmarks*. https://www.gurobi.com/wp-content/uploads/Performance-Gurobi-9_5.pdf
5. Hofmann, H. (1994). *Statlog (German Credit Data)*.
6. Lin, B., & Tang, B. (2021). *Optimal Decision Tree - MIP Formulations, Solution Methods, and Stability*. https://github.com/LucasBoTang/Optimal_Classification_Trees/blob/main/Report.pdf
7. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.
8. Quinlan, J. R. (1993). C 4.5: Programs for machine learning. *The Morgan Kaufmann Series in Machine Learning*.

ประวัติผู้เขียน

ชื่อ-สกุล	พงศ์ทวัส ฮั่นวัฒนวงศ์
วัน เดือน ปี เกิด	6 มีนาคม 2539
สถานที่เกิด	จังหวัดนครศรีธรรมราช
วุฒิการศึกษา	วิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
ที่อยู่ปัจจุบัน	9/268 ซอยพหลโยธิน21 แขวงลาดยาว เขตจตุจักร กรุงเทพมหานคร 10900