การจำแนกโรครากเน่าโคนเน่าของต้นทุเรียนด้วยอัลกอริทึมต้นไม้ตัดสินใจ
The Identification of Root and Stem Rot Disease of Durian Trees by Algorithm Decision Tree

ชัยศิริ สนิทพลกลาง[1*]
Chaisiri Sanitphonklang[1]
[1]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏจันทรเกษม
[1]Computer Science Program, Faculty of Science, Chandrakasem Rajabhat University
วันที่ส่งบทความ : 1 กรกฎาคม 2565 วันที่แก้ไขบทความ : 7 กันยายน 2565 วันที่ตอบรับบทความ : 22 พฤษภาคม 2566
Received: 1 July 2022, Revised: 7 September 2022, Accepted: 22 May 2023

บทคัดย่อ

ปัญหาสำคัญในการผลิตทุเรียนคือการระบาดของโรค โดยเฉพาะโรครากเน่าโคนเน่าของต้นทุเรียนเป็นโรคที่สำคัญที่สุด เนื่องจากโรคนี้สามารถทำให้ต้นทุเรียนที่โตเต็มที่และให้ผลผลิตยืนต้นตายได้ ปัญหาที่กล่าวมาข้างต้นนั้นจำเป็นเร่งด่วนในการหาทางช่วยเหลือ การวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาโมเดลการจำแนกโรครากเน่าโคนเน่าของต้นทุเรียนโดยใช้อัลกอริทึมต้นไม้ตัดสินใจภายใต้กระบวนการเรียนรู้ของเครื่อง จากข้อมูลเกษตรกร 50 ราย ในตำบลนาสัก อำเภอสวี จังหวัดชุมพร โดยมีข้อมูลตัวอย่างสำหรับเรียนรู้จำนวน 500 รายการ แบ่งออกเป็นสองกลุ่ม ได้แก่ รายการที่เป็นโรครากเน่าโคนเน่าจำนวน 150 รายการ และรายการปกติจำนวน 350 รายการ ซึ่งถูกเรียกว่า คลาส 1 และคลาส 0 ตามลำดับ สมดุลข้อมูลด้วยวิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE) และกระบวนการเรียนรู้ของเครื่องด้วยอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree: DT) ในการจำแนกข้อมูล พบว่า มีความถูกต้อง 99.33 % ซึ่งมีประสิทธิภาพสูงสุด เมื่อเปรียบเทียบความถูกต้องกับอัลกอริทึมแรนดอมฟอเรสต์ (Random Forest: RF) อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และอัลกอริทึมเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors: KNN) ที่มีความแม่นยำเป็น 95.33 % 93.66 % และ 91.33 % ตามลำดับ และค้นพบว่า ลักษณะอาการโรครากเน่าโคนเน่าที่มีประสิทธิภาพจำแนกว่าเป็นโรครากเน่าโคนเน่า 3 ลักษณะแรก คือ เนื้อไม้สีน้ำตาล แผลเน่าลุกลามไว และกลิ่นหืนในส่วนที่เน่า

คำสำคัญ : การเรียนรู้ของเครื่อง อัลกอริทึมต้นไม้ตัดสินใจ โรครากเน่าโคนเน่าของต้นทุเรียน SMOTE

[*]ที่อยู่ติดต่อ E-mail address: chaisiri.s@chandra.ac.th

Abstract

An essential problem in durian production is the outbreak of disease, especially root rot; stem rot of durian is the most critical disease. Due to this disease, mature and fruiting durian trees can die. The problems mentioned above must be urgently needed to find a way to help. This research aims to develop a model for the classification of root rot and stem rot disease of durian trees using a decision tree algorithm under the machine learning process. The data set originated from 50 agricultural people in Chumphon province, totaling 500 items. The datasets were divided into two groups: the first group, 150 items of root and stem rot disease of durian trees, referred to as class 1. The second group, the usual 350 items, referred to class 0. Before learning with decision tree algorithms, the data were balanced using Synthetic Minority Oversampling Technique (SMOTE). Decision Tree Algorithm (DT) used to classify the data was found to have an accuracy of 99.33 %, and accuracy comparisons were made with the Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor algorithm (KNN) at 95.33 %, 93.66 %, and 91.33 %, respectively. Discovered that the characteristics of root rot disease durian tree root rot that is effective for classification have three characteristics, namely brown wood, fast gangrene, and there is a rancid smell in the rotten part.

Keywords: Machine Learning, Decision Tree Algorithm, Root and Stem rot disease of durian, SMOTE

1. บทนำ

ทุเรียนได้ชื่อว่าเป็น “ราชาแห่งผลไม้” [1] และเป็นผลไม้ที่สำคัญทางเศรษฐกิจของประเทศ มีการปลูกในทุกภูมิภาคของประเทศไทย [2] โดยแหล่งปลูกที่สำคัญอยู่ที่ภาคตะวันออกและภาคใต้ นอกจากบริโภคภายในประเทศแล้วยังมีการส่งออกไปยังต่างประเทศ เช่น จีน ฮ่องกง สหรัฐอเมริกา เป็นต้น ซึ่งปัญหาของการผลิตทุเรียนที่สำคัญ คือการระบาดของโรคโดยเฉพาะโรครากเน่าโคนเน่าของต้นทุเรียนถือเป็นโรคที่สำคัญ เนื่องจากเป็นโรคที่ทำให้ต้นทุเรียนที่เจริญเติบโต และให้ผลผลิตแล้วยืนต้นตาย โดยมีลักษณะอาการส่วนบนของต้นทุเรียนที่เป็นโรครากเน่าโคนเน่ามีใบด้านสลด และไม่เป็นมัน สีใบเริ่มเหลืองและร่วง เปลือกที่บริเวณโคนมีสีน้ำตาล และอาจมีโคนเน่า [3] ต้นที่เริ่มเป็นโรคสามารถสังเกตได้ยาก จากปัญหาที่กล่าวมาผู้วิจัยเล็งเห็นความสำคัญของปัญหาดังกล่าวจึงได้คิดค้นหาวิธีช่วยป้องกัน และบรรเทาปัญหานี้ ด้วยการพัฒนาโมเดลสำหรับการเรียนรู้ของเครื่อง (Machine Learning) เพื่อจำแนกคุณลักษณะโรครากเน่าโคนเน่าของต้นทุเรียน ซึ่งสามารถพัฒนาในรูปแบบแอปพลิเคชันบนระบบอินเทอร์เน็ต [4] ในลำดับถัดไปได้ และจากการศึกษาที่ผ่านมาพบว่ามีงานวิจัยการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ (Decision Tree: DT) เกี่ยวกับโรคในมนุษย์ อาทิเช่น โรคเบาหวาน [5] ไทรอยด์ [6] สถานะทารกในครรภ์ [7] เป็นจำนวนมาก แต่ประเทศไทยมีอีกส่วนที่เป็นตัวขับเคลื่อนประเทศนั่นก็คือการเกษตร ซึ่งประสบกับปัญหา

หลากหลายรูปแบบและรอการแก้ไข ปัญหาบางปัญหายังมีข้อมูลปริมาณที่น้อยอาจไม่เพียงพอต่อการเรียนรู้ด้วยปัญญาประดิษฐ์ อันเป็นช่องทางที่สามารถเข้ามาช่วยแก้ปัญหาด้านการเกษตรได้จึงเป็นที่มาของการวิจัยในครั้งนี้

งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาโมเดลการจำแนกโรครากเน่าโคนเน่าของต้นทุเรียนด้วยอัลกอริทึมต้นไม้ตัดสินใจภายใต้สมมติฐานการเพิ่มข้อมูลกลุ่มน้อยมีผลต่อประสิทธิภาพของการจำแนกโรครากเน่าโคนเน่าของต้นทุเรียนด้วยอัลกอริทึมต้นไม้ตัดสินใจ ตามที่อัจฉรา สุมังเกษตร และคณะ [8] ได้พัฒนาระบบผู้เชี่ยวชาญที่สามารถจำแนกสมุนไพรด้วยเทคนิคต้นไม้ตัดสินใจที่มีความแม่นยำในการจำแนก 88.00 % และ Somvanshi และคณะ [9] ได้กล่าวไว้ว่า ต้นไม้ตัดสินใจทำงานอย่างมีประสิทธิภาพกับข้อมูลที่ไม่ต่อเนื่อง ซึ่งมีความสอดคล้องกับการวิจัยนี้ ในการวิจัยครั้งนี้ใช้ข้อมูลจากเกษตรกรในตำบลนาสัก อำเภอสวี จังหวัดชุมพร จำนวน 50 ราย มีจำนวนรายการคุณลักษณะโรครากเน่าโคนเน่าของต้นทุเรียนจำนวน 150 รายการ และรายการอาการปกติของต้นทุเรียนจำนวน 350 รายการ ถูกแบ่งออกเป็น 2 คลาส คือ คลาส 1 และคลาส 0 ตามลำดับ เมื่อพิจารณาข้อมูลทั้งสองคลาสเกิดความไม่สมดุลเนื่องจากคลาส 0 มีจำนวนมากกว่าหนึ่งเท่า จึงใช้เทคนิคการปรับเพิ่มข้อมูลแบบสุ่มโดยวิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE) สำหรับคลาส 1 ได้ข้อมูลใหม่เพิ่มขึ้นอีก 50% ทำการเพิ่มเฉพาะในข้อมูลทดสอบเท่านั้น นับว่าเป็นการเพิ่มประสิทธิภาพการเรียนรู้ของเครื่อง [10] จึงทำให้ข้อมูลตัวอย่างเพิ่มรวมเป็นข้อมูลทั้งสิ้น 575 รายการ แล้วนำข้อมูลชุดดังกล่าวสร้างโมเดลสำหรับจำแนกด้วยอัลกอริทึมต้นไม้ตัดสินใจที่มีความถูกต้อง 99.33 % เมื่อเทียบกับ อัลกอริทึมแรนดอมฟอเรสต์ (Random Forest: RF) อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) และอัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors: KNN) ที่มีความถูกต้องตามลำดับ ดังนี้ 95.33% 93.66% และ 91.33% และค้นพบว่า ลักษณะอาการโรครากเน่าโคนเน่าที่มีประสิทธิภาพจำแนกว่าเป็นโรครากเน่าโคนเน่า 3 ลักษณะ คือ เนื้อไม้สีน้ำตาล แผลเน่าลุกลามไว และกลิ่นหืนในส่วนที่เน่า ตามลำดับ

งานวิจัยนี้ถูกแบ่งออกเป็นส่วน ๆ ดังนี้ ส่วนแรก บทนำที่กล่าวถึงที่มาและปัญหางานวิจัย ส่วนที่สองวิธีการทดลอง ส่วนที่สามผลการทดลองและวิจารณ์ ส่วนสุดท้าย คือการสรุปผลการทดลอง

2. วิธีการทดลอง

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลสำหรับการเรียนรู้ของเครื่องเพื่อจำแนกโรครากเน่าโคนเน่าของต้นทุเรียน โดยมีขั้นตอนดำเนินการและทำการทดลองดังนี้

2.1 ชุดข้อมูล

การวิจัยนี้ใช้ข้อมูลคุณลักษณะอาการโรครากเน่าโคนเน่าของต้นทุเรียนจากเกษตรกรในตำบลนาสัก อำเภอสวี จังหวัดชุมพร จำนวน 50 ราย โดยมีจำนวนรายการคุณลักษณะโรครากเน่าโคนเน่าและต้นทุเรียนปกติทั้งสิ้นจำนวน 500 รายการ โดยบันทึกผ่านระบบออนไลน์ทางอินเทอร์เน็ต แบ่งเป็นข้อมูลของโรครากเน่าโคนเน่า จำนวน 150 รายการ และรายการข้อมูลต้นทุเรียนที่ไม่ได้เป็นโรครากเน่าโคนเน่าจำนวน 350 รายการ รวมเป็น 500 รายการ ดังแสดงในรูปที่ 1 โดยคุณลักษณะโรครากเน่าโคนเน่าของต้นทุเรียนประกอบไปด้วยหลายคุณลักษณะพร้อมทั้งนิยามความหมายให้แต่ละคุณลักษณะแทนด้วยอักษร

ภาษาอังกฤษ ดังแสดงในที่ตารางที่ 1 โดยกำหนดให้ข้อมูลถูกแบ่งออกเป็นสองคลาส คือคลาสโรครากเน่าโคนเน่าแทนด้วยคลาส 1 และคลาสปกติแทนด้วยคลาส 0

	A	B	C	...	L	class
1	1	0	1	0	1	1
2	1	1	0	1	1	1
3	1	1	1	1	1	1
				....		
499	0	0	1	0	0	0
500	0	0	0	0	0	0

รูปที่ 1. ชุดข้อมูลคุณลักษณะอาการโรครากเน่าโคนเน่าของต้นทุเรียนที่ถูกแปลงข้อมูล

ตารางที่ 1. สัญลักษณ์แทนคุณลักษณะอาการของโรครากเน่าโคนเน่าของต้นทุเรียน

สัญลักษณ์	ความหมาย	สัญลักษณ์	ความหมาย
A	ใบอ่อนเหี่ยวเหลือง	G	ถากเปลือกพบเปลือกเน่า
B	มีจุดแผลสีน้ำตาลอ่อนฉ่ำ	H	กลิ่นหืนในส่วนที่เน่า
C	แผลเน่าลุกลามไว	I	มอดเจาะ
D	เส้นใบสีน้ำตาล	J	เนื้อไม้สีน้ำตาล
E	จุดฉ่ำน้ำสีนำตาล	K	ใบสีน้ำตาลดำ
F	น้ำเยิ้มออกในช่วงเช้า	L	ยืนต้นตาย

2.2 วิธีสุ่มเพิ่มข้อมูลแบบ SMOTE

เทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มแบบ SMOTE จัดการความไม่สมดุลของข้อมูล (Imbalanced data) ที่มีปริมาณแตกต่างกัน ข้อมูลในการทดลองครั้งมีข้อมูลคลาส 1 เพียง 150 รายการ ส่วนคลาส 0 มีจำนวน 350 รายการ แตกต่างกันหนึ่งเท่า ซึ่งมีผลต่อประสิทธิภาพในการจำแนกข้อมูล วิธี SMOTE เพิ่มจำนวนข้อมูลในคลาสที่มีปริมาณน้อย โดยการสุ่มค่าข้อมูลขึ้นมา 1 รายการจากกลุ่มคลาส 1 หลังจากนั้นพิจารณาค่าข้อมูลเพื่อนบ้านใกล้สุด k ค่า แล้วคำนวณหาระยะห่างระหว่างค่าที่สุ่มกับข้อมูลเพื่อนบ้านใกล้สุดแต่ละค่า เพื่อหาค่าระยะห่างที่น้อยที่สุดระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงตัวที่มีระยะห่างน้อยสุด [11] นำเสนอโดย Chawla และคณะ ทำการเปรียบเทียบวิธีการสุ่มตัวอย่างเพิ่มข้อมูลเดิมอย่างสุ่ม (Random Over-sampling: ROS) และSMOTE ใช้กับการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจสำหรับชุดข้อมูลผลการตรวจเต้านม ซึ่งเป็นข้อมูลที่ไม่สมดุล ผลการทดลองพบว่าความถูกต้องในการทำนายของข้อมูลกลุ่มน้อยโดยวิธี ROS น้อยกว่าวิธี SMOTE ปัจจุบัน SMOTE ได้รับความนิยม และถูกปรับปรุงให้มีประสิทธิภาพขึ้นโดยการใช้วิธีระยะทาง Minkowski ที่ถ่วงน้ำหนักสำหรับนิยามเพื่อนบ้านใกล้เคียง วิธีนี้นิยามเพื่อนบ้านใกล้เคียงที่ดีที่สุดได้เพราะมีการจัดลำดับความสำคัญของคุณลักษณะที่เกี่ยวข้องกับการจำแนก นำเสนอโดย Maldonado และคณะ [12]

สมการที่ (1) ใช้เพื่อสร้างตัวอย่างใหม่

$$x' = x + rand(0,1) * |x - x_k| \quad (1)$$

เมื่อ

x'	แทน	ข้อมูลที่สังเคราะห์ขึ้นมาใหม่
x	แทน	ข้อมูลเดิมที่สุ่มมา
rand (0, 1)	แทน	ตัวเลขสุ่มระหว่าง 0 ถึง 1
x_k	แทน	ข้อมูลที่เป็นเพื่อนบ้านของ x

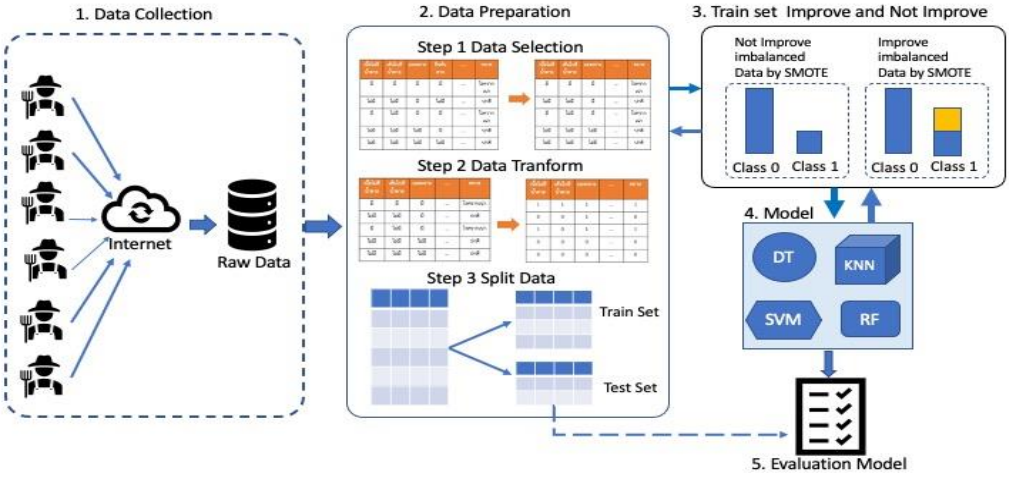
ตัวอย่างการเพิ่มข้อมูลกลุ่มน้อยด้วยวิธี SMOTE แสดงดังตารางที่ 2 พิจารณาข้อมูล (6,4) และให้ (4,3) เป็นเพื่อนบ้านใกล้สุด ข้อมูล (6,4) คือข้อมูลที่สุ่มมาจากกลุ่มน้อยเพื่อนำมาใช้หาเพื่อนบ้านใกล้สุด ข้อมูล (4,3) คือหนึ่งในเพื่อนบ้านใกล้สุด

ตารางที่ 2. ตัวอย่างสร้างชุดข้อมูลใหม่ด้วยวิธี SMOTE

ขั้นตอนวิธีดำเนินการสุ่มเพิ่มกลุ่มตัวอย่างแบบวิธี SMOTE
Start
$\boldsymbol{DS1_1 = 6,\ DS1_2 = 4}$
$\boldsymbol{DS1_1 - DS1_2 = -2}$
$\boldsymbol{DS2_1 = 4, DS2_2 = 3}$
$\boldsymbol{DS2_1 - DS2_2 = -1}$
The new samples will be generated as
$(\boldsymbol{DS1', DS2' = (6,4) + rand(0-1) * (-2,-1)})$
and(0-1)generates a random number between 0 and 1.
End

2.3 ขั้นตอนการดำเนินการ

ขั้นตอนวิธีดำเนินการตามหลักวงจรชีวิตการเรียนรู้ของเครื่อง และเพื่อให้การอธิบายวิธีการดำเนินการวิจัยให้เข้าใจง่าย ผู้วิจัยได้ออกแบบวิธีดำเนินการวิจัยดังแสดงในรูปที่ 2 โดยมีรายละเอียดของขั้นตอนดังต่อไปนี้



รูปที่ 2. ขั้นตอนวิธีดำเนินการ

2.3.1 การเก็บและรวบรวมข้อมูล (Data collection)

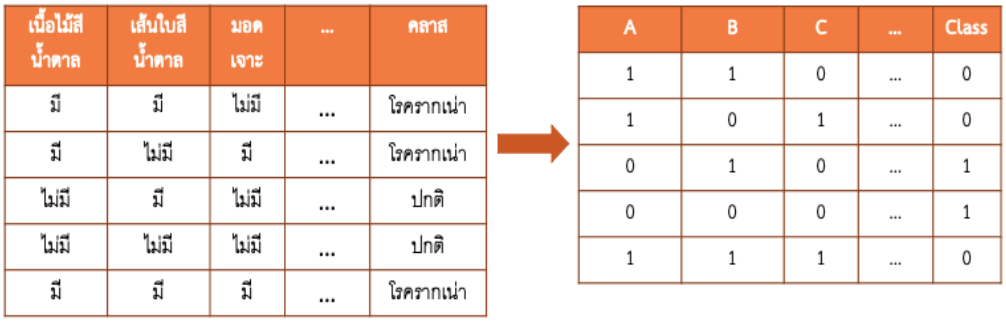
ผู้วิจัยสร้างแบบเก็บข้อมูลในรูปแบบ Google Form เพื่อให้เกษตรกรทำการบันทึกอาการโรครากเน่าโคนเน่าของต้นทุเรียน ถ้าปรากฏคุณลักษณะอาการต่าง ๆ ให้บันทึกเป็นค่า “มี” ถ้าไม่ปรากฏให้บันทึกเป็นค่า “ไม่มี” โดยเก็บข้อมูลจากต้นทุเรียนที่เป็นโรครากเน่าโคนเน่า และต้นทุเรียนปกติ

2.3.2 การเตรียมข้อมูล (Data preparation)

ขั้นตอนนี้เป็นการเตรียมข้อมูลให้สะอาดและมีความสมบูรณ์เพื่อนำไปใช้กับโมเดลสำหรับจำแนก [13] ซึ่งส่งผลต่อความแม่นยำ [14] มีขั้นตอนดังนี้

ขั้นตอนแรก เริ่มต้นด้วยการวิเคราะห์เพื่อพิจารณาเลือกคุณลักษณะของอาการโรครากเน่าโคนเน่าของต้นทุเรียนที่เป็นอาการสังเกตได้ว่าจะเป็นโรครากเน่าโคนเน่า พบว่ามีคุณลักษณะยืนต้นตายไม่ใช่อาการที่ใช้สังเกตการเกิดโรคแต่กลับเป็นผลที่เกิดจากโรค ดังนั้นคุณลักษณะอาการยืนต้นตายจึงไม่ถูกนำไปใช้ในครั้งนี้ทำให้คุณลักษณะอาการที่ถูกนำไปใช้ประกอบไปด้วย ใบอ่อนเหี่ยวเหลือง มีจุดแผลสีน้ำตาลอ่อนฉ่ำใบสีน้ำตาลดำ แผลเน่าลุกลามไว จุดฉ่ำน้ำสีนำตาล น้ำเยิ้มออกในช่วงเช้า ถากเปลือกพบเปลือกเน่า กลิ่นหืนในส่วนที่เน่า มอดเจาะ เนื้อไม้สีน้ำตาล และเส้นใบสีน้ำตาล

ขั้นตอนที่สอง การแปลงข้อมูล เนื่องจากข้อมูลปฐมภูมิไม่เหมาะต่อการเรียนรู้ของโมเดลจึงทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมดังแสดงในรูปที่ 3



รูปที่ 3. การแปลงข้อมูลปฐมภูมิ

ขั้นตอนที่สาม ทำการแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนแรกไว้สำหรับสอนเครื่อง และส่วนที่สองไว้สำหรับทดสอบความถูกต้องของการจำแนก การจำแนกข้อมูลจะเกิดความแม่นยำถ้าข้อมูลที่นำมาใช้สอนมีโดเมนที่ชัดเจน และมีความพอดีกับโมเดลที่ใช้เรียนรู้จึงทำให้เกิดมีการแยกข้อมูล [15]-[16] ในการวิจัยครั้งนี้ได้ใช้หลักการขั้นตอนวิธี QUEST (Quick, Unbiased, Efficient, Statistical Tree) สำหรับการจำแนกโครงสร้างต้นไม้ [17] กฎการแยกในขั้นตอนวิธีนี้มีสมมติฐานว่าตัวแปรเป้าหมายเป็นตัวแปรต่อเนื่อง และสามารถหลีกเลี่ยงอคติที่อาจเกิดขึ้น ซึ่งเหมาะสำหรับตัวแปรหลายหมวดหมู่ [18] ผลการแบ่งข้อมูลได้เป็น 2 ส่วน โดยส่วนแรกคิดเป็น 80% แบ่งไว้สำหรับใช้สอนเครื่อง และส่วนที่สองคิดเป็น 20% จากข้อมูลทั้งหมด แบ่งไว้สำหรับการทดสอบความแม่นยำของเครื่องเรียนรู้

ขั้นตอนที่สี่ การเพิ่มจำนวนข้อมูลให้เกิดความสมดุลระหว่างสองคลาสที่ใช้เรียนรู้ของเครื่อง สำหรับงานวิจัยนี้คลาส 1 หรือคลาสที่เป็นโรครากเน่าโคนเน่ามีจำนวนข้อมูลที่น้อยกว่าคลาส 0 หรือคลาสปกติถึงหนึ่งเท่าตัว จึงทำการเพิ่มจำนวนข้อมูลตัวอย่างในกลุ่มน้อยด้วยวิธี SMOTE หลังจากดำเนินการได้ข้อมูลตัวอย่างใหม่ในคลาส 0 เพิ่มขึ้น 75 รายการ คิดเป็น 50% โดยในการวิจัยนี้ทำการเพิ่มสัดส่วนสุ่มเพิ่มตัวอย่างเป็น 10%, 20%, 30% และ 50% ตามลำดับ ผลปรากฏว่าการเพิ่มสัดส่วนสุ่มกลุ่มตัวอย่างในข้อมูลส่วนน้อยที่เพิ่มขึ้นทำให้การเรียนรู้ของเครื่องกับคลาสที่มีข้อมูลน้อยมีประสิทธิภาพในการจำแนกกลุ่มข้อมูลน้อยได้ดีขึ้น [19] ดังนั้นจึงเลือกเพิ่มจำนวนข้อมูลกลุ่มน้อย 50% ดังแสดงในตารางที่ 3

ตารางที่ 3. รายการชุดข้อมูลก่อนและหลังการเพิ่มด้วยวิธี SMOTE

คลาส	ข้อมูลตั้งต้น	ข้อมูลที่ผ่าน SMOTE	รวม
0	150	75	225
1	350	0	350
รวมทั้งสิ้น			575

2.3.3 สร้างโมเดล (Model)

การทดลองครั้งนี้ใช้อัลกอริทึมต้นไม้ตัดสินใจสำหรับจำแนกข้อมูล โดยเปรียบเทียบประสิทธิภาพการจำแนกกับอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน อัลกอริทึมแรนดอมฟอเรสต์ และอัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด ทั้ง 4 อัลกอริทึมจัดอยู่ในวิธีการจำแนกข้อมูล โดยขั้นตอนการสร้างต้นไม้ตัดสินใจ [20] มีดังนี้

1. เลือกแอตทริบิวต์ที่เหมาะสมเป็นโหนดรากจากแอตทริบิวต์ทั้งหมด ซึ่งแอตทริบิวต์ที่นำมาเป็นโหนดรากต้องมีความสัมพันธ์กับคลาสมากที่สุดโดยให้ดูจากค่า อินฟอร์เมชันเกน (Information gain) ที่มีค่ามากที่สุดของแต่ละแอตทริบิวต์ ดังแสดงในสมการที่ (2) แต่ก่อนจะทราบค่าอินฟอร์เมชันเกนต้องทำการนิยามค่า เอนโทรปี (Entropy) ตามสมการที่ (3) ก่อน

$$Gain(S,A) = E(S) - \sum_{v=value(A)} \frac{|S_v|}{|S|} E(S_v) \quad (2)$$

เมื่อ

S = ตัวอย่างที่ประกอบด้วยชุดข้อของตัวแปรต้นและตัวแปรตามหลาย ๆ กรณี

E = เอนโทรปีของตัวอย่างสามารถคำนวณโดยใช้สมการที่ (3)

A = ตัวแปรต้นที่พิจารณา

value (A) = เซตของค่าของ A ที่เป็นไปได้

S_v = ตัวอย่างที่ A มีค่า v ทั้งหมด

$$E(S) = -\sum_{j=1}^{n} ps(j) log_2\, ps(j) \quad (3)$$

เมื่อ

S = ตัวอย่างที่ประกอบด้วยชุดข้อของตัวแปรต้นและตัวแปรตามหลาย ๆ กรณี

ps(j) = อัตราส่วนของกรณีใน S ที่ตัวแปรตามหรือผลลัพธ์มีค่า j

2. พิจารณาค่าแอตทริบิวต์ที่เลือกเป็นโหนดราก เพื่อหาความบริสุทธิ์ของค่าในแอตทริบิวต์ และถ้าแอตทริบิวต์ใดมีความบริสุทธิ์แล้วไม่ต้องแตกกิ่งต้นไม้ ซึ่งโหนดนั้นเรียกว่าโหนดใบ โดยค่าบริสุทธิ์ คือค่าที่มีคำเฉลยเดียวไม่มีคำเฉลยอื่นปะปน

3. สำหรับแอตทริบิวต์ที่ไม่บริสุทธิ์ให้ทำการแยกกิ่งต้นไม้พร้อมสร้างโหนดลำดับถัดไป โดยการวนกลับไปทำตามขั้นที่หนึ่งจนกว่าค่าแต่ละแอตทริบิวต์แตกกิ่งเป็นโหนดใบและมีค่าบริสุทธิ์

การวิจัยนี้ทำกระบวนการเรียนรู้ของเครื่องด้วยอัลกอริทึมต้นไม้ตัดสินใจ โดยแบ่งการเรียนรู้ออกเป็น 3 แบบ คือ

แบบที่ 1 ใช้ข้อมูลที่ไม่มีการเพิ่มกลุ่มตัวอย่างด้วยวิธี SMOTE สำหรับการเรียนรู้ของเครื่อง

แบบที่ 2 ใช้ข้อมูลที่มีการเพิ่มกลุ่มตัวอย่างกลุ่มน้อยด้วย SMOTE สำหรับการเรียนรู้ของเครื่อง

แบบที่ 3 ใช้ข้อมูลที่มีการเพิ่มกลุ่มตัวอย่างกลุ่มน้อยด้วย SMOTE สำหรับการเรียนรู้ของเครื่องด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน อัลกอริทึมแรนดอมฟอเรสต์ และอัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด เพื่อ

มาเปรียบเทียบประสิทธิภาพกับการเรียนรู้ของเครื่องด้วยอัลกอริทึมต้นไม้ตัดสินใจ โดยผลการทดลองถูกนำเสนอในหัวข้อถัดไป

2.3.4 การประเมินประสิทธิภาพ (Evaluation model)

การประเมินประสิทธิภาพของโมเดลสำหรับการวิจัยใช้เครื่องมือที่ชื่อว่าตาราง Confusion matrix ประเมินผลลัพธ์ของการทำนายสำหรับกระบวนการเรียนรู้ของเครื่อง ซึ่งมีแนวคิดจากการวัดว่าสิ่งที่โมเดลทำนายกับสิ่งที่เกิดขึ้นจริงมีสัดส่วนเป็นอย่างไร [21] ตารางที่ 4 คือตาราง Confusion matrix มีการประยุกต์ใช้วัดประสิทธิภาพของโมเดลกันอย่างแพร่หลาย อาทิเช่น การจำแนกผู้สมัครที่ลังเลในการตัดสินใจศึกษาต่อเพื่อสร้างกลยุทธ์สนับสนุนการตัดสินใจให้เข้าศึกษาต่อ กรณีศึกษามหาวิทยาลัยราชภัฏจันทรเกษม [22] ที่มีความแม่นยำในการจำแนกถึง 95.85 % โดยใช้เครื่องมือ Confusion matrix ในการประเมินความแม่นยำ

ตารางที่ 4. Confusion matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

โดย Confusion matrix คำนวณประสิทธิภาพการทำนายของโมเดลในรูปแบบค่าต่าง ๆ ดังนี้

1. ค่าความถูกต้อง (Accuracy) คือ ความถูกต้องที่ทำนายได้ตรงกับสิ่งที่เกิดขึ้น ด้วยสมการที่ (4)

$$Accuracy = \frac{TPs+TNs}{TPs+TNs+FPs+FNs} \quad (4)$$

2. ค่าความแม่นยำ (Precision) คือ ความแม่นยำที่ทำนายว่าใช่แล้วถูกต้องมากแค่ไหน ในการทำนายว่าใช่ทั้งหมด สามารถคำนวณได้จากสมการที่ (5)

$$Precision = \frac{TPs}{TPs+FPs} \quad (5)$$

3. ค่าความระลึก (Recall) คือ ค่าความไวต่อการทำนายว่าใช่แล้วถูกต้องแค่ไหน ในผลของจริงที่ใช่ทั้งหมดสามารถคำนวณได้จากสมการที่ (6)

$$Recall = \frac{TPs}{TPs+FNs} \quad (6)$$

4. ค่าเฉลี่ยฮาร์โมนิกของความแม่นยำ (F1 score) คือ ค่าเฉลี่ยระหว่าง Precision และ Recall แบบค่าเฉลี่ย (Harmonic mean) ของการสร้าง F1 เป็นการวัดความสามารถของโมเดล ดังสมการที่ (7)

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)} \quad (7)$$

การวิจัยนี้ได้นำตาราง Confusion matrix มาใช้วัดประสิทธิภาพผลการทำนายของโมเดลต้นไม้ตัดสินใจทั้งแบบก่อนและหลังที่มีการปรับเพิ่มกลุ่มตัวอย่างด้วย SMOTE และยังใช้กับโมเดลการเรียนรู้ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน อัลกอริทึมแรนดอมฟอเรสต์ และอัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด เพื่อนำประสิทธิภาพมาเปรียบเทียบกับอัลกอริทึมต้นไม้ตัดสินใจ ซึ่งผลการประเมินถูกนำเสนอในหัวข้อถัดไป

3. ผลการทดลองและวิจารณ์

3.1 ผลการทดลอง

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อสร้างโมเดลจำแนกโรคราก เน่าโคนเน่าของต้นทุเรียนภายใต้กระบวนการเรียนรู้ของเครื่อง มีผลดำเนินการ ดังนี้

แบบที่ 1 ใช้ข้อมูลที่ไม่มีการเพิ่มกลุ่มตัวอย่างด้วยวิธี SMOTE สำหรับการเรียนรู้ของเครื่องด้วยอัลกอริทึมต้นไม้ตัดสินใจ โดยทำการทดลองทั้งหมด 3 รอบ ซึ่งมีประสิทธิภาพการจำแนกข้อมูลก่อนเพิ่มกลุ่มข้อมูลตัวอย่างกลุ่มน้อย (คลาส 1) จากการวัดประสิทธิภาพด้วยเครื่องมือ Confusion matrix คิดเป็นค่าเฉลี่ยดังนี้ ค่าความถูกต้อง 97.33 % ค่าความแม่นยำ 97.66 % ค่าความระลึก 99.00 % และค่าเฉลี่ยฮาร์โมนิกของความแม่นยำ 98.33 % ดังแสดงในตารางที่ 5

ตารางที่ 5. ประสิทธิภาพการจำแนกข้อมูลที่ไม่มีการเพิ่มกลุ่มตัวอย่างด้วยวิธี SMOTE ด้วยอัลกอริทึมต้นไม้ตัดสินใจ

มิติการประเมินประสิทธิภาพ (%)	รอบที่ 1	รอบที่ 2	รอบที่ 3	ค่าเฉลี่ย
ค่าความถูกต้อง	97.00	98.00	97.00	**97.33**
ค่าความแม่นยำ	97.00	98.00	98.00	**97.66**
ค่าความระลึก	99.00	99.00	99.00	**99.00**
ค่าเฉลี่ยฮาร์โมนิกของความแม่นยำ	98.00	99.00	98.00	**98.33**

แบบที่ 2 ใช้ข้อมูลที่มีการเพิ่มกลุ่มตัวอย่างกลุ่มน้อยด้วยวิธี SMOTE สำหรับการเรียนรู้ของเครื่องโดยเพิ่มในส่วนข้อมูลสำหรับกระบวนการเรียนรู้ของเครื่อง แล้วทำการทดลองทั้งหมด 3 รอบ ซึ่งมีประสิทธิภาพการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจหลังเพิ่มกลุ่มข้อมูลตัวอย่างกลุ่มน้อย (คลาส 1) จากการวัดประสิทธิภาพด้วยเครื่องมือ Confusion matrix คิดเป็นค่าเฉลี่ยดังนี้ ค่าความถูกต้อง 99.33 % ค่าความแม่นยำ 98.33 % ค่าความระลึก 100.00% และค่าเฉลี่ยฮาร์โมนิกของความแม่นยำ 98.66 % ดังแสดงในตารางที่ 6

ตารางที่ 6. ประสิทธิภาพการจำแนกข้อมูลที่มีการเพิ่มกลุ่มตัวอย่างด้วยวิธี SMOTE ด้วยอัลกอริทึมต้นไม้ตัดสินใจ

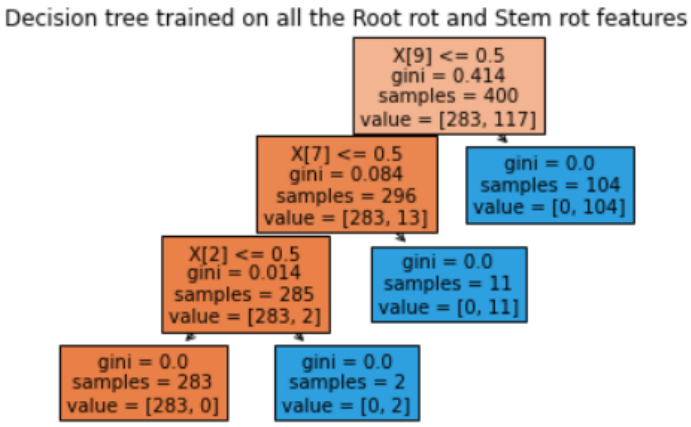
มิติการประเมินประสิทธิภาพ (%)	รอบที่ 1	รอบที่ 2	รอบที่ 3	ค่าเฉลี่ย
ค่าความถูกต้อง	99.00	99.00	100.00	**99.33**
ค่าความแม่นยำ	98.00	98.00	99.00	**98.33**
ค่าความระลึก	100.00	100.00	100.00	**100.00**
ค่าเฉลี่ยฮาร์โมนิกของความแม่นยำ	98.00	98.00	99.00	**98.66**

แบบที่ 3 ใช้ข้อมูลที่มีการเพิ่มกลุ่มตัวอย่างกลุ่มน้อยด้วย SMOTE สำหรับสอนเครื่องเรียนรู้ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน อัลกอริทึมแรนดอมฟอเรสต์ และอัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด โดยทุกอัลกอริทึมดำเนินการประมวลผลจำนวน 3 รอบ โดยใช้ข้อมูลชุดเดียวกันกับการทดลองแบบที่ 2 เพื่อมาเปรียบเทียบประสิทธิภาพการจำแนกหลังจากเรียนรู้ด้วยอัลกอริทึมต้นไม้ตัดสินใจ จากการวัดประสิทธิภาพด้วยเครื่องมือ Confusion matrix ได้ผลค่าความถูกต้องเฉลี่ยดังนี้ การจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจมีค่าความถูกต้องเฉลี่ยเท่ากับ 99.33 % อัลกอริทึมแรนดอมฟอเรสต์ มีค่าความถูกต้องเฉลี่ยเท่ากับ 95.33 % อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ Kernel Function แบบ Linear และกำหนดค่า Hyperparameter C เท่ากับ 200 (ถ้าค่า Hyperparameter C มีค่ามากแสดงว่ายอมให้มีขอบเขตเส้นที่กว้างขึ้น) มีค่าความถูกต้องเฉลี่ยเท่ากับ 93.66 % และอัลกอริทึมเพื่อนบ้านใกล้ที่สุด โดยกำหนดค่า k (จำนวนเพื่อนบ้านใกล้ที่สุด) เท่ากับ 3 มีค่าความถูกต้องเฉลี่ยเท่ากับ 91.33 % ดังแสดงในตารางที่ 7

ตารางที่ 7. เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลสำหรับทั้ง 4 อัลกอริทึม

มิติการประเมิน (%)	Method/Algorithms			
	DT	RF	SVM	KNN
ค่าความถูกต้อง	99.33	95.33	93.66	91.33
ค่าความแม่นยำ	98.33	95.33	92.33	93.66
ค่าความระลึก	100.00	99.33	96.33	94.66
ค่าเฉลี่ยฮาร์โมนิกของความแม่นยำ	98.66	97.66	94.33	94.33

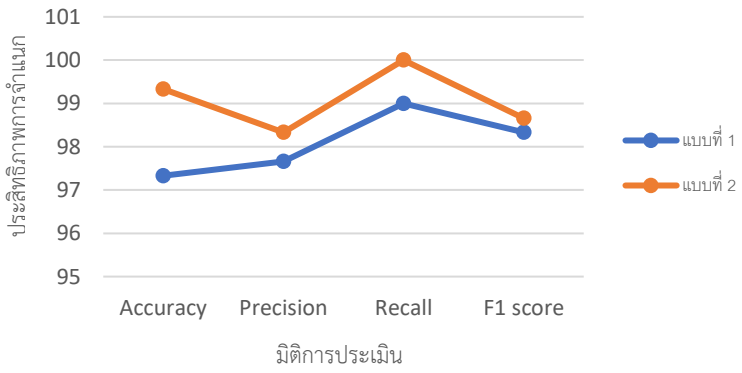
การเรียนรู้ของเครื่องจากข้อมูลโรครากเน่าโคนเน่าของต้นทุเรียนจำนวน 575 รายการ หลังจากทำการเพิ่มกลุ่มข้อมูลตัวอย่างของกลุ่มข้อมูลส่วนน้อยเพิ่ม โดยแบ่งข้อมูลสำหรับการเรียนรู้ 80% และอีก 20% ใช้สำหรับทดสอบนั้น ผลลัพธ์ที่เกิดขึ้นคือ กฎของต้นไม้ตัดสินใจสามารถจำแนกโรครากเน่าโคนเน่าของต้นทุเรียนได้ดังรูปที่ 4 และพบว่ามีลักษณะอาการโรครากเน่าโคนเน่าที่มีประสิทธิภาพอีกทั้งยังสามารถจำแนกว่าเป็นโรครากเน่าโคนเน่าได้ 3 ลักษณะ คือ เนื้อไม้สีน้ำตาล แผลเน่าลุกลามไว และกลิ่นหืนในส่วนที่เน่า



รูปที่ 4. กฎของต้นไม้ตัดสินใจสำหรับจำแนกโรครากเน่าโคนเน่าของต้นทุเรียน

3.2 วิจารณ์ผลการทดลอง

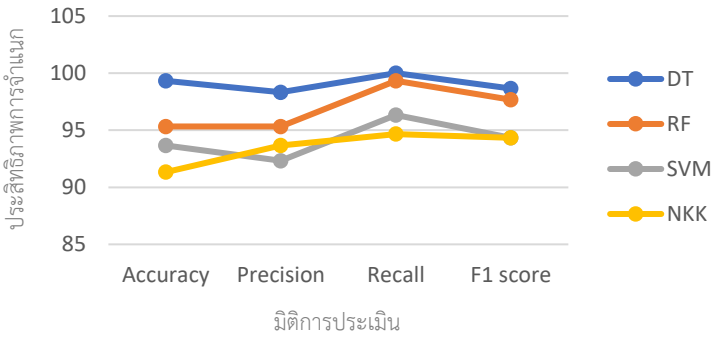
จากผลการทดลองเมื่อเปรียบเทียบระหว่างกระบวนการเรียนรู้ของเครื่อง แบบที่ 1 ใช้ข้อมูลที่ไม่มีการเพิ่มกลุ่มตัวอย่างด้วยวิธี SMOTE สำหรับกระบวนการเรียรู้ของเครื่อง และแบบที่ 2 ใช้ข้อมูลที่มีการเพิ่มกลุ่มตัวอย่างกลุ่มน้อยด้วยวิธี SMOTE สำหรับกระบวนการเรียนรู้ของเครื่อง พบว่าการเพิ่มกลุ่มตัวอย่างด้วยวิธี SMOTE มีประสิทธิภาพการจำแนกดีกว่า อันเกิดมาจากมีปริมาณข้อมูลเรียนรู้ที่เพิ่มขึ้นดังแสดงในรูปที่ 5



รูปที่ 5. ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลแบบที่ 1 และแบบที่ 2

เมื่อนำข้อมูลที่มีการเพิ่มกลุ่มตัวอย่างกลุ่มน้อยด้วยวิธี SMOTE ทำการเรียนรู้และทดสอบการจำแนกด้วย อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน อัลกอริทึมแรนดอมฟอเรสต์ และอัลกอริทึมเพื่อนบ้านใกล้ที่สุด พบว่าประสิทธิภาพการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจมีประสิทธิภาพทุกมิติดีกว่าทั้ง 3

อัลกอริทึม ดังกราฟที่แสดงในรูปที่ 6 ซึ่งเกิดจากอัลกอริทึมต้นไม้ตัดสินใจเหมาะกับการจำแนกข้อมูลแบบไม่ต่อเนื่อง โดยข้อมูลในการทดลองครั้งนี้มีลักษณะข้อมูลแบบไม่ต่อเนื่อง



รูปที่ 6. ผลการเปรียบเทียบประสิทธิภาพการจำแนกของ 4 อัลกอริทึม

4. สรุปผลการทดลอง

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลการจำแนกโรครากเน่าโคนเน่าของต้นทุเรียนด้วยอัลกอริทึมต้นไม้ตัดสินใจภายใต้กระบวนการเรียนรู้ของเครื่อง ใช้ข้อมูลในการเรียนรู้จำนวน 575 รายการ ด้วยการสุ่มเพิ่มกลุ่มตัวอย่างข้อมูลกลุ่มน้อยด้วยวิธี SMOTE จากเกษตรกร 50 ราย ในตำบลนาสัก อำเภอสวี จังหวัดชุมพร ผลการทดลองพบว่าประสิทธิภาพการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจก่อนเพิ่มข้อมูลตัวอย่างข้อมูลกลุ่มน้อยด้วยวิธี SMOTE เท่ากับ 97.33 % และหลังการเพิ่มข้อมูลตัวอย่างข้อมูลกลุ่มน้อยด้วยวิธี SMOTE มีประสิทธิภาพจำแนก เท่ากับ 99.33 % จะเห็นว่าการปรับเพิ่มความสมดุลข้อมูลกลุ่มน้อยส่งผลให้การจำแนกด้วยอัลกอริทึมต้นไม้ตัดสินใจมีความแม่นยำเพิ่มขึ้น และเมื่อนำไปเปรียบเทียบการจำแนกด้วยอัลกอริทึมแรนดอมฟอเรสต์ อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน อัลกอริทึมเพื่อนบ้านใกล้ที่สุด พบว่าประสิทธิภาพการจำแนกของอัลกอริทึมต้นไม้ตัดสินใจยังมีประสิทธิภาพสูงสุด โดยสามารถจำแนกลักษณะอาการโรครากเน่าโคนเน่าที่มีประสิทธิภาพได้เป็น 3 ลักษณะแรก คือ เนื้อไม้สีน้ำตาล แผลเน่าลุกลามไว และกลิ่นหืนในส่วนที่เน่า

สำหรับการทดลองในอนาคตควรมีการพิจารณาเปรียบเทียบระหว่างการเพิ่มจำนวนกลุ่มตัวอย่างสำหรับชุดข้อมูลกลุ่มน้อย หรือการลดปริมาณกลุ่มตัวอย่างข้อมูลกลุ่มมากให้ใกล้เคียงกับข้อมูลกลุ่มน้อย ว่าวิธีการใดที่สามารถทำให้การจำแนกข้อมูลดีขึ้น

กิตติกรรมประกาศ

ขอขอบพระคุณมหาวิทยาลัยราชภัฏจันทรเกษมที่อำนวยความสะดวกให้ใช้สถานที่ ห้องปฏิบัติการ เครื่องอำนวยความสะดวกต่าง ๆ ชาวเกษตรกร ตำบลนาสัก อำเภอสวี จังหวัดชุมพร ที่สนับสนุนชุดข้อมูลในการทดลองจนทำให้การวิจัยครั้งนี้เสร็จสมบูรณ์ตามเป้าประสงค์ของนักวิจัย และขอบพระคุณ พ่อ แม่ ญาติ มิตรทุกท่านที่คอยให้กำลังใจตลอดยะระเวลาดำเนินการวิจัยครั้งนี้

เอกสารอ้างอิง (References)

[1] วันดี กฤษณพันธ์. 2541. สมุนไพรน่ารู้. พิมพ์ครั้งที่ 3, สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร. [Wandee Gritsanapan. 1998. Interesting herbs to know. 3td ed, Chulalongkorn University Publisher, Bangkok. (in Thai)]

[2] สำนักงานเศรษฐกิจการเกษตร. 2564. สารสนเทศเศรษฐกิจการเกษตรรายสินค้าปี 2564. สำนักงานเศรษฐกิจการเกษตรกระทรวงเกษตรและสหกรณ์ เอกสารสถิติการเกษตรเลขที่ 402 แหล่งข้อมูล : https://www.oae.go.th/assets/portals/1/files/jounal/2565/commodity 2564.pdf. สืบค้นเมื่อวันที่ 20 มิถุนายน 2565.

[3] อมรรัตน์ ภู่ไพบูลย์. 2557. โรคราก่เน่า โคนเน่า และโรคผลเน่าของทุเรียน. กรมวิชาการเกษตร, แหล่งข้อมูล : https://www.doa.go.th/share/. ค้นเมื่อวันที่ 25 มิถุนายน 2565.

[4] Bawack, R. E., Wamba, S. F., & Carillo, K. D. A. 2021. A framework for understanding aritificial intelligence research: insights from practice. *Journal of Enterprise Information Management*, 34(2), 645-678.

[5] นพรัตน์ นนท์ศิริ. 2022. การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิควิธีแบบร่วมกันตัดสินใจและวิธีเลือกคุณลักษณะเด่นไปข้างหน้า. วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏอุดรธานี, 10(2), 107-122. [Nonsiri N., Chaichitwanidchakol P. and Somkantha K. 2022. Data Classification for Diabetes Risk Diagnosis using Majority Voting Ensemble Method and Forward Feature Selection Method. *Udon Thani Rajabhat University Journal of Sciences and Technology*, 10(2), 107-122. (In Thai)]

[6] ณัฐวดี หงษ์บุญมี และ ประภาศิริ ตรีพาณิชย์กุล. 2019. การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยความเสี่ยงที่ส่งผลต่อการ เกิดโรคไฮเปอร์ไทรอยด์ด้วยเทคนิคเหมืองข้อมูล. *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, 9(1), 41-51. [Hongboonmee, N., & Trepanichkul, P. 2019. Comparison of Data Classification Efficiency to Analyze Risk Factors that Affect the Occurrence of Hyperthyroid using Data Mining Techniques. *Journal of Information Science and Technology*, 9(1), 41-51. (In Thai)]

[7] นฤมล ชูเมือง. 2021. การจำแนกสถานะทารกในครรภ์โดยการใช้เครื่องมือติดตามอัตราการเต้นของหัวใจและการหดรัดตัวของมดลูกด้วยเทคนิค ต้นไม้ตัดสินใจ. *วารสารวิทยาศาสตร์เทคโนโลยีและนวัตกรรม*, 2(4), 36-46. [Chumuang N. 2021. Fetal Status Classification with Heart Rate Tracker and CTG by using Decision Tree. *Journal of Science Technology and Innovation (STIJ).* 2(4), 36-46. (In Thai)]

[8] อัจฉรา สุมังเกษตร และณัฐวุฒิ ศรีวิบูลย์. 2021. ระบบผู้เชี่ยวชาญการใช้สมุนไพรรักษาโรคบนฐานความรู้ภูมิปัญญาสุขภาพด้วยเทคนิคการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ. *วารสารศรีปทุมปริทัศน์,* 13(1), 115-127. [Achara Sumungkaset and Nattavut Sriwiboon. 2021. The Expert System for Herbs Usage Based on Wisdom-Knowledge in Terms of Health Using

Tree Classification Decision Techniques. Sripatum Review of Science and Technology, 13(1), 115-127. (In Thai)]

[9] Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. 2016. A review of machine learning techniques using decision tree and support vector machine. In 2016 international conference on computing communication control and automation (ICCUBEA) , IEEE Pune Section, Pune, India. 1-7.

[10] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61, 863-905.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

[12] Maldonado, S., Vairetti, C., Fernandez, A., & Herrera, F. 2022. FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. *Pattern Recognition*, 124, 108511.

[13] Pyle, D. 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, United States of America.

[14] Brownlee, J (Jason Brownlee). 2022. Data Preparation for machine learning. Available at: http://14.99.188.242:8080/jspui/bitstream/123456789/14557/1/Data_Preparation_for_Machine_Learning_Data_Cleaning_C_Feature_Selection_C_and_Data_by_Jason_Brownlee.pdf.

[15] Wang, H., Hsu, H., Diaz, M., & Calmon, F. P. 2021. To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67(10), 6733-6757.

[16] Rácz, A., Bajusz, D., & Héberger, K. 2021. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. *Molecules*, 26(4), 1111.

[17] Loh, W. Y., & Shih, Y. S. 1997. Split selection methods for classification trees. *Statistica sinica,* 7(4), 815 - 840.

[18] Lin, C.L. and Fan, C.L. 2019. Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering*, 18(6), 539-553.

[19] Kesornsit, W., Lorchirachoonkul, V., & Jitthavech, J. 2018. Imbalanced data problem solving in classification of diabetes patients. *Khon Kaen University Research Journal*, 18(3), 11-21.

[20] Rokach, L., & Maimon, O. 2005. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), 476-487.

[21] Stehman, Stephen V. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62 (1), 77–89.

[22] ชัยศิริ สนิทพลกลาง และ สัตวัฏฏ์ คำศิริ. 2021. การจำแนกผู้สมัครที่ลังเลในการ ตัดสินใจศึกษาต่อเพื่อสร้างกลยุทธ์สนับสนุนการตัดสินใจให้เข้าศึกษาต่อ กรณี ศึกษามหาวิทยาลัยราชภัฏจันทรเกษม. วารสารวิทยาศาสตร์ลาดกระบัง, 30(2), 58-73. [Sanitphonklang, C., & Kumsiri, S. 2021. Identification of Applicants Who are Hesitant to Decide to Study in Order to Create Strategies to Support Decision-making for Admission Case Study of Chandrakasem Rajabhat University. *Journal of Science Ladkrabang*, 30(2), 58-73. (In Thai)]