การปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

IMPROVING ID3 ALGORITHM'S PROCESSING TIME BY CONSIDERING INFORMATION GAINS OF ATTRIBUTES

ภิญญรัตน์ ชื่นประเสริฐสุข

Pinyarat Chuenprasertsuk

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสมการปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2566

IMPROVING ID3 ALGORITHM'S PROCESSING TIME BY CONSIDERING INFORMATION GAINS OF ATTRIBUTES

Pinyarat Chuenprasertsuk

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTER ENGINEERING

SCHOOL OF ENGINEERING

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2023

KMITL-2023-EN-M-060-097

หัวข้อวิทยานิพนธ์ การปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

นักศึกษา ภิญญรัตน์ ชื่นประเสริฐสุข

รหัสประจำตัว 60601039

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชา วิศวกรรมคอมพิวเตอร์

พ.ศ. 2566

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ. ดร. เกียรติกูล เจียรนัยธนะกิจ

บทคัดย่อ

การศึกษานี้นำเสนอการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ เพื่อการปรับปรุงพัฒนาอัลกอริทึม ID3 แบบดั้งเดิมและงานวิจัยนี้จะเน้นไปที่การลดเวลาในการสร้างต้นไม้การตัดสินใจ โดยจะมีการอธิบายอัลกอริทึมรวมถึงอธิบายปัญหาของอัลกอริทึม รวมไปถึงการเตรียมสภาพแวดล้อมในการทำงาน พบว่าวิธีการปรับปรุงอัลกอริทึม ID3 แบบดั้งเดิมด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ ผลการทดลองยังระบุด้วยว่าวิธีการที่นำเสนอสามารถลดเวลาการทำงานลงได้มากกว่า 10 เปอร์เซ็นต์ ในขณะที่ความถูกต้องของการจัดหมวดหมู่แทบจะไม่ได้รับผลกระทบ สำหรับงานในอนาคต สามารถวางแผนที่จะปรับปรุงอัลกอริทึม ID3 เพื่อเพิ่มความถูกต้องการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ด้วยการนำอัตราส่วนของจำนวนอินสแตนซ์ย่อยที่มีค่าสูงสุดมาประกอบกับค่าอินฟอร์เมชันเกนเพื่อระบุแอตทริบิวต์ที่ดีที่สุด

Thesis	Improving ID3 algorithm's processing time by considering information gains of attributes.
Student	Pinyarat Chuenprasertsuk
Student ID.	60601039
Degree	Master of Engineering
Program	Computer Engineering
Year	2023
Thesis Advisor	Assoc.Prof.Dr. Kietikul Jearanaitanakij

ABSTRACT

This study presents the development of an ID3 algorithm by considering information gains of attributes to improve the traditional ID3 algorithm. This study will concentrate on reducing the time required to construct decision trees in order to enhance the ID3 algorithm. The algorithm and the problem of Algorithms will be explained, along with setting up the working environment. It was discovered that the traditional ID3 algorithm could be improved by considering information gains of attributes. Moreover, the experimental results indicate that the proposed method can reduce operation time by more than 10 percent with minimal impact on classification accuracy. For further work identifying the finest attribute can be made more precise by integrating the ratio of the highest sub-instance count to the forming gain value into the ID3 algorithm.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความกรุณาจากอาจารย์ที่ปรึกษา รศ. ดร. เกียรติกูล เจียรนัยธนะกิจ ที่ให้ความช่วยเหลือ ให้คำชี้แนะและช่วยแก้ปัญหาตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า

ขอขอบคุณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง คณะวิศวกรรมศาสตร์, และภาควิชาวิศวกรรมคอมพิวเตอร์ที่เอื้อเฟื้อสถานที่และอำนวยความสะดวกในการศึกษา ค้นหาข้อมูลและวิจัย ทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้ได้สำเร็จ

ขอขอบคุณบิดาและมารดาของข้าพเจ้าที่ได้ให้ความสนับสนุน, และกำลังใจในการศึกษาระดับมหาบัณฑิตตลอดมา

ขอขอบคุณรุ่นพี่วรวิทย์ แซ่ตั๊น, และนางสาวณิชา แก้วรอด ผู้แนะแนวทาง แนะนำ, และให้กำลังใจในการทำวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงไปได้ด้วยดี

สำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดามารดา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้และถ่ายทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

ภิญญูรัตน์ ชื่นประเสริฐสุข

สารบัญ

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์ข้อมูล (Data analysis) ถูกนำไปปรับใช้กับงานด้านต่าง ๆ, และบทบาทที่สำคัญที่สุดของการวิเคราะห์ข้อมูลคือการเรียนรู้ของเครื่อง (Machine Learning), และในปัจจุบันมีอัลกอริทึมของการเรียนรู้ของเครื่องอยู่มากมาย โดยแต่ละอัลกอริทึมจะมีความเหมาะสมในการใช้งานขึ้นอยู่กับชุดข้อมูล แผนผังการตัดสินใจ (Decision Tree) เป็นหนึ่งในอัลกอริทึมมากมายซึ่งเข้าใจง่ายและสะดวกสำหรับนักวิเคราะห์มือใหม่ บทความนี้จะเน้นไปที่การปรับปรุงเวลาการทำงานสำหรับสร้างแผนผังการตัดสินใจของอัลกอริทึม ID3

อัลกอริทึม ID3 ได้รับการปรับปรุงโดยนักวิจัยหลายคนด้วยวิธีการต่าง ๆ ที่ได้รับการวิจัย Chen Jim, และคณะ ได้ทำการปรับปรุงอัลกอริทึม ID3 ด้วยการปรับแต่งเกนความรู้ (Information Gain) กับแอตทริบิวต์ด้วยน้ำหนัก Kradesh, และ Jearanaitanakij นำเสนอวิธีการปรับปรุงอัลกอริทึมด้วยการรวมค่าของแอตทริบิวต์ที่มีความสำคัญเท่าเทียมกันซึ่งจะช่วยจัดการปัญหาเมื่ออัลกอริทึม ID3 เลือกโหนดปัจจุบันซึ่งจะเป็นแอตทริบิวต์ที่สำคัญที่สุด, และจะต้องมีแอตทริบิวต์ที่มีความสำคัญเท่าเทียมกันอย่างน้อยสองตัว ผลการทดลองของงานวิจัยได้แสดงให้เห็นว่าวิธีการนี้สามารถลดความลึกของแผนผังการตัดสินใจได้อย่างมีนัยสำคัญ Wang, Yu Lui, และ Lu Lui ได้พัฒนาอัลกอริทึม ID3 ด้วยการคำนวณความสอดคล้องของแอตทริบิวต์และการวิจัยจะทำการเลือกแอตทริบิวต์ที่มีความสอดคล้องสูงที่สุด จากผลการทดลองของงานวิจัยได้แสดงให้เห็นว่ามีการจัดประเภทที่ถูกต้อง แต่ยังคงมีข้อจำกัดเนื่องจากวิธีการนี้จะมีความเหมาะสมสำหรับปัญหาที่ไม่เกินสองคลาสเท่านั้น Kaewrod, และ Jearanaitanakij พบว่าอัลกอริทึม ID3 สร้างกฎการตัดสินใจที่เข้มงวดมากเกินไปและส่งผลให้อัลกอริทึม ID3 มีข้อจำกัด งานวิจัยนี้ได้ทำการปรับปรุงอัลกอริทึม ID3 โดยไม่สนใจอินสแตนซ์ย่อยและผลการทดลองก็สามารถสรุปได้ว่าค่าเฉลี่ยของจำนวนของกฎการตัดสินใจเพิ่มขึ้นมากกว่า 40 เปอร์เซ็นต์อย่างมีนัยสำคัญ

อัลกอริทึม ID3 แบบดั้งเดิมจะสร้างต้นไม้ตัดสินใจด้วยการหาแอตทริบิวต์ที่ดีที่สุดที่จะเป็นแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกน (Information Gain) สูงที่สุด แต่ถ้าหากว่ามีแอตทริบิวต์ที่ดีที่สุดมากกว่าหนึ่งตัวที่มีค่าอินฟอร์เมชันเกนเท่ากัน อัลกอริทึมจะไม่มีทางรู้เลยว่าแอตทริบิวต์ไหนที่จะเป็นแอตทริบิวต์ที่ดีที่สุด แต่วิธีการเลือกแอตทริบิวต์ที่ดีที่สุดของอัลกอริทึม ID3 แบบดั้งเดิมจะเลือกแอตทริบิวต์ตัวแรกสุดของชุดของแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดเพื่อจะนำมาสร้างต้นไม้ตัดสินใจ

เป้าหมายหลักของอัลกอริทึม ID3 แบบดั้งเดิมคือการเลือกแอตทริบิวต์ที่ดีที่สุดมากำหนดเป็นโหนด โดยค่าอินฟอร์เมชันเกนซึ่งจะแสดงถึงเอนโทรปีที่ใช้ในการกำหนดความแปรปรวนที่มีอยู่ในแอตทริบิวต์
จะเป็นการคำนวณในขั้นตอนแรกในอัลกอริทึม ID3 แบบดั้งเดิมตามสมการดังต่อไปนี้

$$IG(A) = -\sum_{x \in X} p(x) \log_2 p(x) - Remainder(A)$$

โดยที่ IG(A) หมายถึงค่าอินฟอร์เมชันเกนของแอตทริบิวต์ A, และ X หมายถึงกลุ่มของคลาสในข้อมูล, และสุดท้าย p(x) จะหมายถึงสัดส่วนของจำนวนองค์ประกอบในคลาส X ต่อจำนวนองค์ประกอบในข้อมูล

ส่วนที่เหลือ (Remainder) คือการคำนวณเอนโทรปีที่เหลือจากเอนโทรปีทั้งหมดหลังจากแต่ละแอตทริบิวต์ซึ่งจะเป็นส่วนหนึ่งของสมการการคำนวณค่าอินฟอร์เมชันเกน อัลกอริทึมจะทำการวนซ้ำตามจำนวนรายการที่ถูกจำแนกโดยใช้แต่ละส่วนของแอตทริบิวต์ที่ดีที่สุดเพื่อคำนวณหาจำนวนที่เหลืออยู่

$$Remainder(A) = \sum_{v \in V} \{-p(v). \sum_{x \epsilon X} p(x_v) \log_2 p(x_v)\}$$

โดยที่ Remainder(A) หมายถึงค่าของส่วนที่เหลือของ A, และ V หมายถึงกลุ่มของข้อมูลใน A, และ p(v) หมายถึงสัดส่วนของจำนวนองค์ประกอบของค่าใน A ต่อจำนวนองค์ประกอบของค่าทั้งหมดใน A, และ $p(x_v)$ หมายถึงสัดส่วนของจำนวนองค์ประกอบของคลาสของค่าใน A ต่อจำนวนองค์ประกอบของคลาสของค่าทั้งหมดใน A

อัลกอริทึม ID3 แบบดั้งเดิมจะเลือกแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุด ถ้าหากว่ามีแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดเพียงตัวเดียว อัลกอริทึมจะเลือกให้แอตทริบิวต์นั้นให้เป็นโหนดการตัดสินใจ แต่ถ้าหากมีแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงสุดหลายตัว อัลกอริทึมจะเลือกแอตทริบิวต์ตำแหน่งแรกจากชุดของแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงสุดมาเป็นโหนดการตัดสินใจ โดยที่แอตทริบิวต์สุดท้ายจะถูกเลือกให้เป็นโหนด

งานวิจัยนี้จะมุ่งเน้นไปที่การลดเวลาในการทำงานของอัลกอริทึม ID3 แบบดั้งเดิมด้วยการนำเสนอวิธีการในการปรับปรุงพัฒนาอัลกอริทึม เนื่องจากอัลกอริทึมไม่ทราบว่าแอตทริบิวต์ใดที่จะเป็นแอตทริบิวต์ที่ดีที่สุด เมื่ออัลกอริทึมได้รับชุดของแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมาซึ่งจะมีแอตทริบิวต์มากกว่า 1 แอตทริบิวต์ทำให้ไม่สามารถตัดสินใจเลือกแอตทริบิวต์ที่ดีที่สุดได้

1.2 วัตถุประสงค์ของการศึกษา

1.2.1 เพื่อศึกษา, ปรับปรุง, และพัฒนาอัลกอริทึม ID3

1.2.2 เพื่อศึกษาข้อจำกัดและช่องโหว่ของอัลกอริทึม ID3

1.2.3 เพื่อลดเวลาการทำงานของอัลกอริทึม ID3 โดยไม่ส่งผลกระทบกับค่าความถูกต้อง

1.2.4 เพื่อนำเสนอวิธีการพัฒนาอัลกอริทึม ID3

1.3 ขอบเขตของงานวิจัย

งานวิจัยนี้จะเน้นไปที่การลดเวลาในการสร้างต้นไม้การตัดสินใจ ในส่วนที่สองจะมีการอธิบายอัลกอริทึมและอธิบายถึงปัญหา ส่วนที่สามจะเป็นส่วนที่อธิบายการทำงานของการปรับปรุงพัฒนาอัลกอริทึม ID3 ขั้นตอนการเตรียมข้อมูล รวมไปถึงการเตรียมสภาพแวดล้อมในการทำงาน ส่วนผลการทดลองจะเปรียบเทียบอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ได้รับการปรับปรุงแล้ว, และส่วนสุดท้ายจะเป็นส่วนที่สรุปงานวิจัยและงานที่วางแผนว่าจะดำเนินการในอนาคต

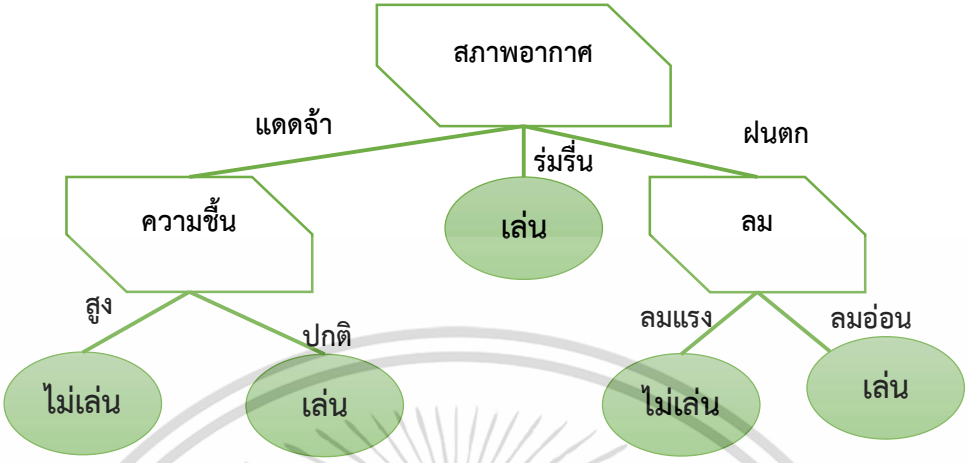
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การพัฒนาอัลกอริทึม ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ครั้งนี้ ผู้วิจัยได้ศึกษาแนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง เพื่อให้เกิดความรู้ความเข้าใจถึงรูปแบบและวิธีการได้อย่างถูกต้อง ผู้วิจัยได้เรียบเรียงเรื่องสรุปและนำเสนอเอกสารที่เกี่ยวข้องและจำเป็นต่อการศึกษา ดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ต้นไม้ตัดสินใจ (Decision Tree)

เทคนิคต้นไม้ตัดสินใจเป็นวิธีการรูปแบบหนึ่งของการจำแนกประเภท (Classification) ซึ่งเป็นวิธีการที่ทำการแบ่งประเภทหรือจำแนกหมวดหมู่ของข้อมูล ซึ่งการจำแนกประเภทเองก็เป็นเทคนิคหนึ่งของเหมืองข้อมูล โดยขั้นตอนและวิธีการของเหมืองข้อมูล (Data mining) จะมีวิธีการในการสร้างโครงร่างต้นไม้ตัดสินใจ ขึ้นมาเพื่อจำแนกหรือแบ่งประเภทของข้อมูลที่นำไปใช้ในการตัดสินใจในด้านต่าง ๆ เช่น ระบบวิเคราะห์พฤติกรรมการซื้อของใช้ในช่วงวัยต่าง ๆ, ระบบการทำงานแต่ละแผนกของบริษัท เป็นต้น ที่โดยส่วนมากแล้วประเภทของข้อมูลจะอยู่ในรูปแบบของเงื่อนไขในลักษณะการจำลอง ถ้าแอตทริบิวต์ที่หนึ่งสร้างเงื่อนไขหนึ่ง, และแอตทริบิวต์ที่สองสร้างอีกเงื่อนไขหนึ่งจะนำไปสู่ผลการตัดสินใจ เช่นตัวอย่างการตัดสินใจเล่นเทนนิส “ถ้า “สภาพอากาศ” เป็น “แดดจ้า” และ “ความชื้น” เป็น “สูง” เช่นนั้นแล้ว การตัดสินใจจะเลือก “ไม่เล่นเทนนิส” ” โดยการทำงานของต้นไม้ตัดสินใจจะทำการสร้างโครงร่างที่มีลักษณะคล้ายกับต้นไม้หัวกลับโดยจะเริ่มสร้างตั้งแต่โหนดแรกสุดจะแทนรากของต้นไม้ (Root node) แต่ละโหนดจะเป็นค่าของคุณลักษณะต่าง ๆ (attribute), และแต่ละกิ่งจะแสดงค่าผลในการทดสอบของคุณลักษณะนั้น ๆ, และสุดท้ายโหนดใบ (Leaf node) แสดงคลาสที่กำหนดหรือผลของการตัดสินใจ ซึ่งจะแสดงเป็นรูปภาพที่ 2.1 ตัวอย่างต้นไม้ตัดสินใจ (Decision Tree)



รูปที่ 2.1 ตัวอย่างต้นไม้ตัดสินใจ (Decision Tree) ของการตัดสินใจเล่นเทนนิสใน 14 วัน

ขั้นตอนวิธี ID3 (ID3 Algorithm)

ขั้นตอนวิธี ID3 เป็นขั้นตอนหรือวิธีการที่ใช้ในการสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยนำหลักการของทฤษฎีข่าวสารมาใช้ โดยค่าที่วัดได้จะนำมาประกอบการตัดสินใจเลือกจะใช้ตัวแปรใดในการแบ่งข้อมูล ซึ่งจะใช้วิธีการกำหนดโครงสร้างต้นไม้ตัดสินใจ โดยการสร้างโครงสร้างต้นไม้ตัดสินใจนั้น จะเริ่มด้วยการคัดเลือกคุณลักษณ์ (Attribute) ของชุดข้อมูลจากตัวชี้วัดหรือค่าอินฟอร์เมชันเกน (Gain) ที่มีค่าสูงที่สุดมากำหนดให้เป็นโหนดของต้นไม้ตัดสินใจ, และจะทำการวนหาจนครบทุกคุณลักษณะของชุดข้อมูลตัวอย่างเช่น การตัดสินใจเลือกเล่นเทนนิสใน 14 วัน ซึ่งจะมีคลาสเป้าหมายที่ต้องพิจารณ์อยู่ 2 กรณีคือ “เล่น” หรือ “ไม่เล่น” โดยจะแทนค่าคลาส “เล่น” เป็นค่า p, และแทนค่าคลาส “ไม่เล่น” เป็นค่า n ซึ่งจะใช้จำนวนบิตของข้อมูลในการแยกคลาส “p” หรือ “n” ตามสมการที่ (2.1)

$$I\,(p,n) = -\frac{p}{p+n} log_2 \left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right) log_2 \left(\frac{n}{p+n}\right) \quad (2.1)$$

ค่าคาดคะเนของข้อมูล (Entropy) คือค่าของคุณลักษณะหรือแอตทริบิวต์ A ซึ่งจะใช้ในการจำแนกข้อมูล โดยแอตทริบิวต์ A จะแบ่งข้อมูล D ออกเป็น D_1, D_2, D_3, ... D_n ที่ D_1 จะเป็นกรณีหนึ่งของคลาสตัวอย่าง จำนวน p1, และตัวอย่างจากคลาส n จำนวน n1 ดัง สมการที่ (2.2)

$$Entropy(I) = -\sum_{c \in C} p(X_c) \log_2 p(X_c) \quad (2.2)$$

เมื่อ

C หมายถึง เซทของคลาสคำตอบทั้งหมด

$P(X_c)$ หมายถึง อัตรส่วนของจำนวนกรณีตัวอย่างที่ตอบคลาส c ต่อจำนวนกรณีตัวอย่างทั้งหมดในเซท I

ค่าอินฟอร์เมชันเกน (Information Gain) จะใช้ในการระบุความสำคัญของแอตทริบิวต์แต่ละแอตทริบิวต์ดังสมการที่ (2.3)

$$Information\ Gain(I, A) = Entropy(I) - Remainder(I, A) \quad (2.3)$$

เมื่อ

Information Gain (I, A) หมายถึง ค่าอินฟอร์เมชันเกนความรู้ของ A ซึ่ง A หมายถึงแอตทริบิวต์ใด ๆ

I หมายถึง เซทของกรณีตัวอย่างปัจจุบัน

Entropy(I) หมายถึง ค่าเอนโทรปีของกรณีตัวอย่างก่อนที่จะถูกแบ่งไปตามแอตทริบิวต์ A

Remainder(I, A) หมายถึง ผลรวมของค่าเอนโทรปีของแต่ละค่าที่ไม่ซ้ำกันในแอตทริบิวต์ A หลังจากกรณีตัวอย่างถูกแบ่งไปตามค่าที่ไม่ซ้ำกันของแอตทริบิวต์ A

ค่าของส่วนที่ยังคงเหลือ (Remainder) จะถูกคำนวณได้จากสมการที่ (2.4)

$$Remainer(I, A) = \sum_{v \in V} p(v) Entropy(v) \quad (2.4)$$

เมื่อ

V หมายถึง เซทของ value หรือค่าที่ไม่ซ้ำกันของแอตทริบิวต์ A

p(v) หมายถึง อัตราส่วนของจำนวนกรณีตัวอย่างที่เป็นค่า v ในแอตทริบิวต์ A ต่อจำนวนกรณีตัวอย่างทั้งหมดในเซท A

Entropy(v) หมายถึง ค่าเอนโทรปีของกรณีตัวอย่างที่มีค่า v ในแอตทริบิวต์ A

ขั้นตอนการสร้างโมเดล Decision Tree ID3

การสร้างโมเดล Decision Tree ID3 จะทำด้วยวิธีการคัดเลือกจากความสัมพันธ์ของแอตทริบิวต์กับคลาส หาว่าแอตทริบิวต์ใดแอตทริบิวต์หนึ่งมีค่าความสัมพันธ์กับคลาสสูงที่สุดจะได้รับการคัดเลือกให้เป็นโหนดบนสุดของ tree (root node) หลังจากนั้นจึงจะทำการวนหาแอตทริบิวต์ถัดไปมากำหนดให้เป็นโหนดถัดไปเรื่อย ๆ ซึ่งตัวชี้วัดความสัมพันธ์ของแอตทริบิวต์กับคลาส ที่เรียกว่า Information Gain (IG) จะสามารถคำนวณได้จากสมการ (2.3), และ (2.4)

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคหนึ่งของ Classification ซึ่งเป็นวิธีการแบ่งประเภทหรือแยกหมวดหมู่ข้อมูล, และ Classification นั้นเป็นเทคนิคหนึ่งของเหมืองข้อมูล (Data Mining)

1) เหมืองข้อมูล (Data Mining) คือ รูปแบบหรือกระบวนการการวิเคราะห์ข้อมูล เพื่อแยกประเภท จำแนกรูปแบบและความสัมพันธ์ของข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่หรือคลังข้อมูล โดยมีเทคนิคต่าง ๆ หลายวิธี (กิตติ, 2550) เช่นการจะจำรูปแบบของข้อมูลเป็นต้น ซึ่งรูปแบบการทำเหมืองข้อมูลนั้นได้รวมความรู้จากหลายแขนงเข้าไว้ด้วยกันที่ประกอบด้วย ระบบการเรียนรู้ของเครื่องจักร (Machine Learning) รวมกับวิทยาศาสตร์สารสนเทศ (Information Science) สถิติ (Statistics), และระบบฐานข้อมูล โดยทั่วไปแล้วเทคนิคที่นำมาใช้ส่วนใหญ่มี 5 ประเภท (ศุภชัย, 2551)

1.1) เทคนิค Classification เป็นเทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่าง ๆ ที่ได้มีการกำหนดเป้าหมายไว้แล้ว เทคนิคประเภทนี้เหมาะกับข้อมูลที่มีลักษณะที่เป็นคำตอบตัวเลือก หรือกลุ่มข้อมูล ตัวอย่างเช่น คำตอบใช่หรือไม่ใช่, เป็นหรือไม่เป็น, คำตอบ ก, ข, ค, และ ง เป็นต้น เทคนิคนี้จะใช้การสร้างแบบจำลองเพื่อการพยากรณ์ค่าข้อมูล (Predictive Modeling) ในอนาคตจากการที่ได้จำแนกกลุ่มข้อมูลตัวอย่างไว้แล้ว ซึ่งในลักษณะดังกล่าวนี้เรียกว่า “Supervised Learning” เทคนิคการ Classification มี 2 รูปแบบ ได้แก่ Tree Induction, และ Neural Induction, และเป็นกระบวนการสร้างแบบจำลองเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด ตัวอย่างเช่น การแบ่งประเภทลูกค้าว่า เชื่อถือได้หรือไม่ การแบ่งประเภทการตัดสินใจตามเงื่อนไขอะไรบางอย่าง ซึ่งจะเป็นการสร้างแบบจำลองโดยการเรียนรู้จากข้อมูลที่ได้กำหนดกลุ่มไว้เรียบร้อยแล้ว

1.2) เทคนิค Association Rule Discovery เป็นเทคนิคที่ใช้กฎในการหาความสัมพันธ์ของข้อมูลภายในฐานข้อมูลขนาดใหญ่ โดยที่จะค้นหารูปแบบที่ซ้ำกัน ความสัมพันธ์ของข้อมูล หรือโครงสร้างเชิงสาเหตุ เพื่อนำมาทำการวิเคราะห์ข้อมูลและหาสิ่งที่ประกอบกันอยู่ภายในข้อมูลนั้น เช่น การวิเคราะห์ข้อมูลการซื้อขายในซูเปอร์มาร์เก็ต เพื่อทำการจัดทำการวางแผนเพื่อจัดการส่งเสริมการขาย (Promotion), และเตรียมการวางแผนตระเตรียมสินค้าบนชั้นวางสินค้า (Shelf) ที่สามารถเพิ่มปริมาณการซื้อเนื่องจากความสัมพันธ์กันของสินค้า เช่น การนำน้ำอัดลมมาเรียงเอาไว้ใกล้กับข้าวโพดคั่ว

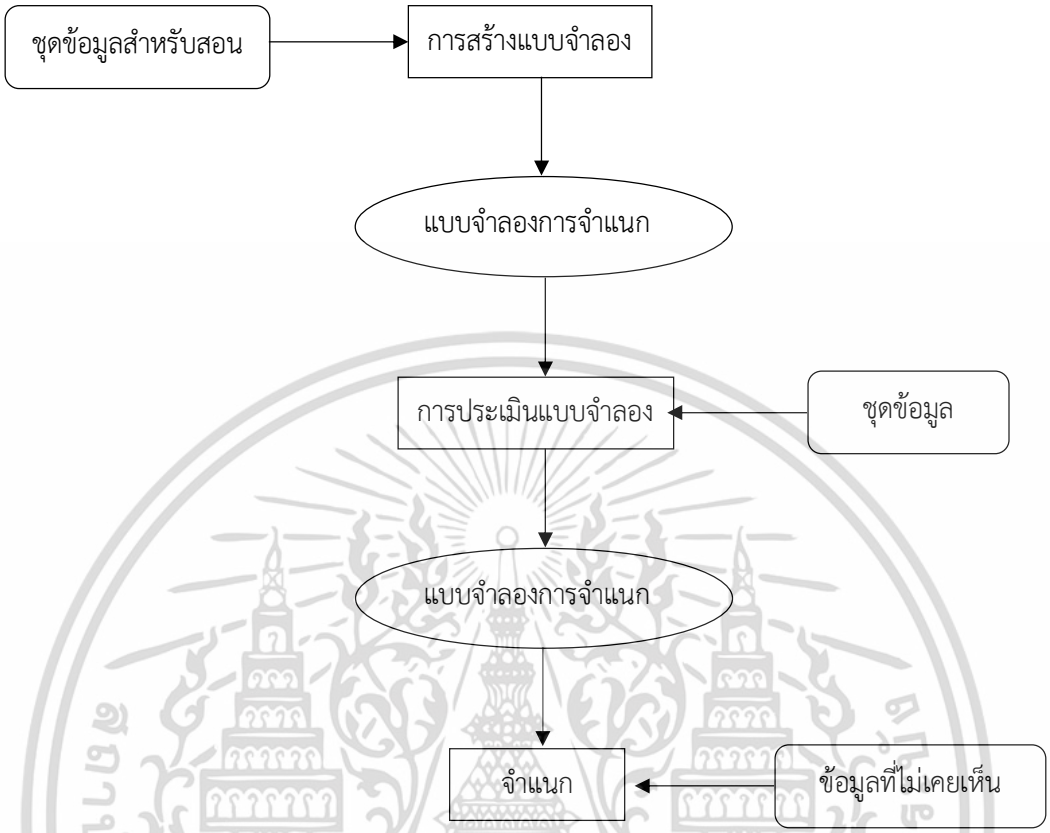
1.3) เทคนิค Clustering เป็นวิธีการลดขนาดของข้อมูลโดยที่จะทำการแบ่งข้อมูลออกเป็นประเภทหรือกลุ่มที่มีความสัมพันธ์กัน ทำให้สามารถค้นหาข้อมูลที่ตกหล่นหรือสูญหายจากการละเลยได้ เทคนิคนี้นิยมนำมาใช้เป็นขั้นตอนเบื้องต้นในการสร้างเหมืองข้อมูล, และเหมาะกับข้อมูลที่มีการผสมกันของข้อมูล, และยังไม่ได้ถูกจำแนกออกจากกันอย่างชัดเจนจึงนำวิธีการทำ Cluster เพื่อจำแนกข้อมูลออกเป็นกลุ่มต่าง ๆ โดยจำนวนของข้อมูลในแต่ละกลุ่มมักจะถูกแทนค่าด้วยตัวอักษร k หรือที่นิยมเรียกกันอีกชื่อว่า K-mean ในกลุ่มของผู้ที่ใช้

1.4) เทคนิค Deviation Detection เป็นวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน, ค่าทางสถิติ หรือค่าที่คาดคะเนไว้ มีความต่างกันมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพเพื่อให้ง่ายต่อการทำความเข้าใจ ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้คือ การตรวจสอบการลงชื่อหรือลายเซ็นปลอม หรือการตรวจสอบบัตรเครดิตปลอม เป็นต้น

1.5) เทคนิค Sequential Analysis เป็นเทคนิคในการวิเคราะห์ลำดับเพื่อค้นพบรูปแบบของการปรากฏของข้อมูล ซึ่งการปรากฏของข้อมูลในรายการที่ได้รับการจำแนกไปตามกลุ่มแล้ว ตัวอย่างเช่น ถ้าผู้ซื้อต้องการซื้อสินค้า A แล้วเขากลับมาซื้อสินค้า B ในภายหลัง เทคนิคนี้จะมีความแตกต่างจากเทคนิค Association Rule Discovery เนื่องจากเทคนิค Sequential Analysis จะให้ความสำคัญของลำดับการปรากฏของข้อมูลด้วย

วิธีการแบ่งประเภทหรือแยกหมวดหมู่ข้อมูล Classification คือ กระบวนการสร้างแบบจำลองเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด, เป็นการสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดไว้ล่วงหน้า, และสามารถพยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ แบบจำลองที่ได้อาจอยู่ในรูปแบบการตัดสินใจแบบต้นไม้ตัดสินใจ (Decision Tree) หรือแบบโครงข่ายประสาทเทียม (Neural Network)

ในการจัดหมวดหมู่จำเป็นต้องมีกลุ่มข้อมูลสำหรับการเรียนรู้ (Training Data) เพื่อให้ข้อมูลเรียนรู้, และนำข้อมูลเรียนรู้มาสร้างเป็นแบบจำลอง (Model Construction) หลังจากที่ได้แบบจำลองมาเป็นที่เรียบร้อย ก็จะทดสอบโดยกลุ่มข้อมูลสำหรับการทดสอบ (Testing Data) เพื่อประเมินความถูกต้องของโมเดล (Model Evaluation) อีกทั้งใช้ชุดข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen Data) เพื่อทำการกำหนด Class ให้กับข้อมูลใหม่ที่ได้มา หรือทำนายค่าออกมาตามที่ต้องการ เช่น การจัดหมวดหมู่ของผู้ยื่นขอเครดิต (Credits) เป็นระดับต่ำ, ระดับกลาง, และระดับสูง ของความเสี่ยงที่จะได้รับ, การอนุมัติบุคคลเข้ารับทำงานในลักษณะงานต่าง ๆ , การจัดกลุ่มโอกาสในการชนะเกมเอ็กซ์โอ หรือการจัดหมวดหมู่ของเว็บไซต์ที่มีโอกาสที่จะเป็นเว็บไซต์ที่ถูกปลอมแปลง เป็นต้น (พยุน, 2548)



รูปที่ 2.2 กระบวนการจำแนกข้อมูล

การสร้างต้นไม้ตัดสินใจ (Decision Tree) คือการนำข้อมูลมาทำการเรียนรู้, และนำข้อมูลนั้น ๆ สร้างเป็นต้นไม้ตามเงื่อนไขของข้อมูล โดยต้นไม้ตัดสินใจจะถูกสร้างจากด้านบนลงด้านล่าง (Top-Down) กล่าวคือเริ่มจากการสร้างรากของต้นไม้ (Root Node) ก่อนแล้วจึงแตกกิ่งไปจนถึงใบที่จะเป็นผลการตัดสินใจ ซึ่งคล้ายกับต้นไม้กลับหัว ซึ่งขั้นตอนในการสร้างต้นไม้ตัดสินใจจะทำได้ตามขั้นตอนด้านล่างดังนี้ (Han and Kamber, 2001)

1) ต้นไม้เริ่มต้นโดยกำหนดโหนดขึ้นมาหนึ่งโหนดเพื่อแสดงถึงชุดข้อมูลฝึก (Training Set)

2) ถ้าหากข้อมูลในโหนดเป็นกลุ่มข้อมูลเดียวกัน จะให้โหนดนั้น ๆ เป็นใบ โดยที่แต่ละใบมีชื่อจำแนกตามกลุ่มหัวข้อของข้อมูลนั้น ๆ

3) ถ้าหากในโหนดใดโหนดหนึ่งมีข้อมูลหลากหลายผสมรวมกันอยู่ ก็จำเป็นจะต้องคำนวณหาค่าอินฟอร์เมชันเกน (Gain) ของแต่ละแอตทริบิวต์เพื่อที่จะใช้ค่าอินฟอร์เมชันเกนมาประกอบการคัดเลือกแอตทริบิวต์ ที่มีความสามารถในการแบ่งกลุ่มของข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุด โดยแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือถูกเลือกให้เป็นแอตทริบิวต์ใช้ในการตัดสินใจ ซึ่งจะอยู่ในรูปแบบของโหนดต้นไม้

4) กิ่งของต้นไม้ จะสร้างขึ้นจากค่าความเป็นไปได้ต่าง ๆ ของแอตทริบิวต์แต่ละตัว, และจะสร้างกิ่งที่จำแนกออกตามข้อมูล เพื่อแสดงค่าแทนลงไปในแต่ละกิ่ง

5) ทำซ้ำต่อเพื่อเลือกเอาแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากำหนดเป็นโหนดสำหรับข้อมูลจะได้รับการจำแนกออกมาในรูปแบบของกิ่งเพื่อที่จะสร้างเส้นทางนำแอตทริบิวต์นั้น ๆ มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่แอตทริบิวต์ที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาซ้ำอีกสำหรับโหนดในความลึกถัดไป

6) ทำการวนซ้ำเพื่อจำแนกข้อมูลและต้นไม้ตัดสินใจจะขยายขนาดไปเรื่อย ๆ ตามกิ่งข้อมูล โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อพบว่ามีเงื่อนไขที่สามารถนำไปสู่การตัดสินใจ หรือเป้าหมายของการตัดสินใจนั้น ๆ

การคำนวณค่า Information Gain ต้นไม้ตัดสินใจ (Decision Tree) (ขจรศักดิ์ ศรีอ่อน, 2552) เป็นโครงสร้างในการจำแนกประเภทของข้อมูลที่จะใช้กฎจากการเรียนรู้ที่ได้จากชุดข้อมูลสำหรับเรียนรู้ ซึ่งข้อมูลชุดนี้จะถูกแบ่งออกจากชุดข้อมูลทั้งหมด โดยต้นไม้ตัดสินใจจะมีลักษณะเป็นโครงสร้างต้นไม้หัวกลับ โดยที่จะนำเอาคุณลักษณะ (Attribute) มากำหนดให้เป็นโหนดในการสร้างกฎของต้นไม้ตัดสินใจ การสร้างต้นไม้ตัดสินใจจำเป็นจะต้องให้ความสำคัญกับหลักเกณฑ์ในการพิจารณาเลือกคุณลักษณะ กล่าวคือการที่จะตัดสินใจเลือกแอตทริบิวต์ใดมากำหนดให้เป็นโหนดรากในแต่ละลำดับขั้นตอนของการสร้างต้นไม้และการสร้างต้นไม้ย่อย (Subtree) ของต้นไม้ตัดสินใจจำเป็นจะต้องประกอบไปด้วยค่า หรือเกณฑ์ที่จะใช้ในการพิจารณาเลือกแอตทริบิวต์มากำหนดเป็นโหนด ซึ่งก็คือการคำนวณหาค่ามาตรฐานเกน (Gain Criterion) ซึ่งจะเป็นค่าที่แสดงให้เห็นว่าแอตทริบิวต์นั้น ๆ สามารถใช้จำแนกกลุ่มของข้อมูลได้ดีมากหรือน้อยเพียงใด โดยทดสอบการเลือกแอตทริบิวต์ใดแอตทริบิวต์หนึ่งที่อาจจะเป็นไปได้จากชุดข้อมูลมากำหนดให้เป็นโหนดราก ซึ่งหากว่าแอตทริบิวต์ใดมีค่าอินฟอร์เมชันเกนสูงที่สุด จะหมายถึงแอตทริบิวต์นั้น ๆ สามารถแบ่งกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่าอินฟอร์เมชันเกน (Information Gain) เป็นการช่วยลดจำนวนการสร้างกฎของการทดสอบการจำแนกของข้อมูล อีกทั้งยังสามารถลดความซับซ้อนของต้นไม้ตัดสินใจ

$$I\ (S_1, S_2, \ldots, S_n) = \ -\sum_{i=1}^{n} \frac{s_1}{s}\ log_2 \frac{s_1}{s} \qquad (2.5)$$

เมื่อ

S เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s เรคคอร์ด

N เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น C_i

C_i แทนกลุ่มในลำดับ ที่ i โดย ที่ i มีค่าระหว่าง 1 ถึง n

s_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ S, และอยู่ในกลุ่ม C_i

sij แทนจำนวนข้อมูลที่เป็นสมาชิกของ S ในกลุ่ม C i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของแอตทริบิวต์ A

j ค่าระหว่าง 1 ถึง v

ค่าเอ็นโทรปีของแอตทริบิวต์ A ซึ่งมีค่าของแอตทริบิวต์เป็น (a_1, a_2, a_3, ..., a_v) หาได้ในสมการที่ 2.6

$$E\ (A) = \sum_{j=1}^{v} \frac{S_1+\cdots+S_{nj}}{S} I\ (S_{1j}, S_{2j}, \ldots, S_{nj}) \quad (2.6)$$

อัลกอริทึม ID3 (ID3 Algorithm) (ศุภชัย ประคองศิลป์, 2551) เป็นอัลกอริทึมพื้นฐานที่นิยมนำมาปรับใช้งานในหลาย ๆ ด้าน หลักการของอัลกอริทึม ID 3 นั้นจะเป็นการสร้างการตัดสินใจแบบรูปแบบของต้นไม้หัวกลับซึ่งจะเลือกใช้ทฤษฎีข่าวสาร (Information Theory), และค่าที่คำนวณได้จะถูกนำมาประกอบการตัดสินใจว่าในการเลือกตัวแปรใดตัวแปรหนึ่งในการจำแนกประเภทของข้อมูล โดยที่จะแบ่งข้อมูลออกเป็นชุดตัวอย่าง (Sample) หมายถึงชุดของข้อมูลที่นำไปเรียนรู้ (Training Sample) ที่จะประกอบไปด้วยตัวแปรที่แสดงถึงการตัดสินใจ หรือเป้าหมาย (Target Attribute) ซึ่งหมายถึงตัวแปรที่สามารถนำค่าไปประกอบการทำนายผลการตัดสินใจในโครงสร้างต้นไม้ และแอตทริบิวต์ (Attributes) คือตัวแปรใด ๆ ที่สามารถนำมากำหนดให้เป็นโหนดในต้นไม้, และจะต้องไม่ใช่ตัวแปรที่แสดงถึงค่าการตัดสินใจ หรือเป้าหมาย (Target Attribute) (Tom, 1997) ซึ่งมีลักษณะของอัลกอริทึมดังรูปที่ 2.3

2.1.2 ตัวอย่างการสร้างต้นไม้ตัดสินใจ

1. ตัวอย่างชุดข้อมูลการตัดสินใจเล่นเทนนิส 14 วัน

ตารางที่ 2.1 ชุดข้อมูลการตัดสินใจเล่นเทนนิสใน 14 วัน

วัน	สภาพอากาศ	อุณหภูมิ	ความชื้น	ลม	การตัดสินใจเล่นเทนนิส
1	แดดจ้า	ร้อน	สูง	ลมอ่อน	ไม่เล่น
2	แดดจ้า	ร้อน	สูง	ลมแรง	ไม่เล่น
3	ร่มรื่น	ร้อน	สูง	ลมอ่อน	เล่น
4	ฝนตก	อุ่น	สูง	ลมอ่อน	เล่น
5	ฝนตก	เย็น	ปกติ	ลมอ่อน	เล่น
6	ฝนตก	เย็น	ปกติ	ลมแรง	ไม่เล่น
7	ร่มรื่น	เย็น	ปกติ	ลมแรง	ไม่เล่น
8	แดดจ้า	อุ่น	สูง	ลมอ่อน	เล่น
9	แดดจ้า	เย็น	ปกติ	ลมอ่อน	เล่น
10	ฝนตก	อุ่น	ปกติ	ลมอ่อน	เล่น
11	แดดจ้า	อุ่น	ปกติ	ลมแรง	เล่น
12	ร่มรื่น	อุ่น	สูง	ลมแรง	เล่น
13	ร่มรื่น	ร้อน	ปกติ	ลมอ่อน	เล่น
14	ฝนตก	อุ่น	สูง	ลมแรง	ไม่เล่น

จากตารางข้างต้น การตัดสินใจเล่นเทนนิสใน 14 วัน มีจำนวนอินสแตนซ์เท่ากับ 14 อินสแตนซ์, และจำนวนแอตทริบิวต์เท่ากับ 4 แอตทริบิวต์ ได้แก่ สภาพอากาศ, อุณหภูมิ, ความชื้น, และลม

การสร้างต้นไม้ตัดสินใจจำเป็นจะต้องคำนวณหาค่าอินฟอร์เมชันเกนมาใช้ประกอบการเลือกแอตทริบิวต์ที่ดีที่สุด โดยจะสามารถคำนวณหาค่าอินฟอร์เมชันเกนด้วยสมการที่ (2.3) ซึ่งจะเลือกแอตทริบิวต์สภาพอากาศ, และลมมาคำนวณหาค่าอินฟอร์เมชันเกน โดยสภาพอากาศจะประกอบไปด้วย แดดจ้า, ร่มรื่น, และฝนตก ส่วนลมจะประกอบไปด้วย ลมแรง, และลมอ่อน

ค่าเอนโทรปีของกรณีทั้งหมดเมื่อแทนค่าลงไปในสมการที่ 2.2 จะได้ออกมาเป็น

$$Entropy\ (I) = -\ [\ (9/14 \times \log_2 9/14) + (5/14 \times \log_2 5/14)\ ] = 0.94$$

แล้วจึงจะเริ่มคำนวณหาค่าอินฟอร์เมชันของสภาพอากาศซึ่งจะต้องคำนวณ

ค่าเอนโทรปีของกรณีที่สภาพอากาศเป็นแดดจ้าทั้งหมดเมื่อแทนค่าลงไปในสมการที่ 2.2 จะได้ออกมาเป็น

Entropy (สภาพอากาศ | แดดจ้า) = - (2/5 x $log_2$2/5) + (3/5 x $log_2$3/5) = 0.97

จากนั้นจึงคิดค่าเอนโทรปีของกรณีที่สภาพอากาศเป็นร่มรื่นทั้งหมดเมื่อแทนค่าลงไปในสมการที่ 2.2 จะได้ออกมาเป็น

Entropy (สภาพอากาศ | ร่มรื่น) = - (4/4 x $log_2$4/4) + (0/4 x $log_2$0/4) = 0

และจึงมาคำนวณหาค่าเอนโทรปีของกรณีที่สภาพอากาศเป็นฝนตกทั้งหมดเมื่อแทนค่าลงไปในสมการที่ 2.2 จะได้ออกมาเป็น

Entropy (สภาพอากาศ | ฝนตก) = - (3/5 x $log_2$3/5) + (2/5 x $log_2$2/5) = 0.97

เมื่อได้เอนโทรปีของกรณีย่อยในแอตทริบิวต์สภาพอากาศมาเป็นที่เรียบร้อยแล้ว จากนั้นให้นำค่าที่คำนวณมาได้นั้น มาแทนค่าลงในสมการที่ 2.4 จะได้ออกมาเป็น

Remainder (สภาพอากาศ) = (5/14 x 0.97) + (4/14 x 0) + (5/14 x 0.97)

= 0.69

ดังนั้นค่าอินฟอร์เมชันเกนของสภาพอากาศจะเท่ากับ

IG (I, สภาพอากาศ) = 0.94 – 0.69 = 0.25

เมื่อได้ค่าอินฟอร์เมชันเกนของสภาพอากาศมาเป็นที่เรียบร้อยแล้ว จึงจะมาคำนวณหาค่าอินฟอร์เมชันเกนของลม

จะเริ่มคำนวณหาค่าอินฟอร์มเมชันของ],ซึ่งจะต้องคำนวณ

ค่าเอนโทรปีของกรณีที่ลมเป็นลมแรงทั้งหมดเมื่อแทนค่าลงไปในสมการที่ 2.2 จะได้ออกมาเป็น

Entropy (ลม | ลมแรง) = - (3/6 x $log_2$3/6) + (3/6 x $log_2$3/6) = 1

จากนั้นจึงคิดค่าเอนโทรปีของกรณีที่ลมเป็นลมอ่อนทั้งหมดเมื่อแทนค่าลงไปในสมการที่ 2.2 จะได้ออกมาเป็น

Entropy (ลม | ลมอ่อน) = - (6/8 x $log_2$6/8) + (2/8 x $log_2$2/8) = 0.81

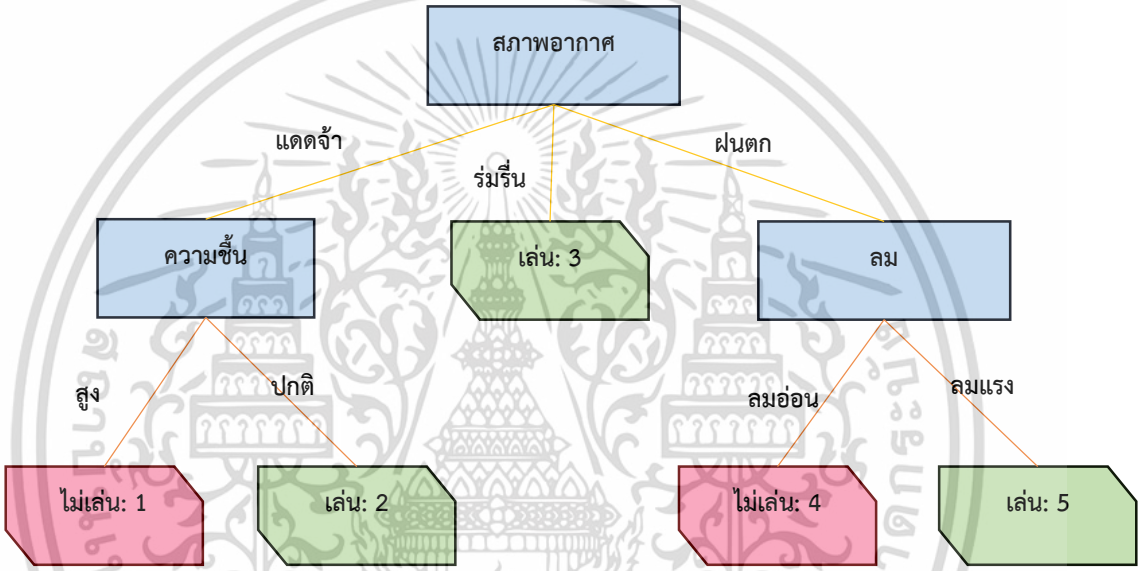
เมื่อได้เอนโทรปีของกรณีย่อยในแอตทริบิวต์ลมมาเป็นที่เรียบร้อยแล้ว จากนั้นให้นำค่าที่คำนวณมาได้นั้น มาแทนค่าลงในสมการที่ 2.4 จะได้ออกมาเป็น

Remainder(ลม) = = (6/14 x 1) + (8/14 x 0.81) = 0.89

ดังนั้นค่าอินฟอร์เมชันเกนของลมจะเท่ากับ

$$IG(I, \text{ลม}) = 0.94 - 0.89 = 0.05$$

นำค่าอินฟอร์เมชันเกนของแอตทริบิวต์สภาพอากาศและแอตทริบิวต์ลมมาเปรียบเทียบเพื่อหาว่าแอตทริบิวต์ตัวใดที่มีค่าอินฟอร์เมชันเกนมากกว่า ซึ่งจะพบว่าค่าอินฟอร์เมชันเกนของแอตทริบิวต์สภาพอากาศจะมากกว่าค่าอินฟอร์เมชันเกนของแอตทริบิวต์ลม จึงสามารถสร้างต้นไม้ตัดสินใจได้ดังภาพด้านล่าง



รูปที่ 2.3 ตัวอย่างต้นไม้ตัดสินใจของชุดข้อมูลการตัดสินใจเล่นเทนนิสใน 14 วัน

2. ทดสอบประสิทธิภาพของต้นไม้ตัดสินใจ

เมื่อนำชุดข้อมูลสำหรับสอน (Training Set) ไปสร้างเป็นต้นไม้ตัดสินใจเรียบร้อยแล้ว ให้นำชุดข้อมูลสำหรับการทดสอบ (Test Set) ซึ่งเป็นข้อมูลที่ถูกแยกออกมาและไม่ถูกนำไปสอนมาทดสอบกับต้นไม้ตัดสินใจ โดยนำค่าไปทดสอบกับต้นไม้ตัดสินใจ จะได้คลาสตัดสินใจเล่นเทนนิส (Play Tennis) มา จากนั้นจึงนำคลาสที่ได้จากการทดสอบมาเปรียบเทียบกับคลาสที่ถูกต้อง ขั้นตอนสุดท้ายจึงหา % ที่ทำนายได้ถูกต้องออกมา (Accuracy)

สมการคำนวณหาค่าความถูกต้อง (Accuracy)

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad (2.7)$$

เมื่อ

Number of correct predictions หมายถึง จำนวนของเหตุการณ์ที่คาดคะเนถูกต้อง

Total number of predictions หมายถึง จำนวนของเหตุการณ์ที่คาดคะเนไว้ทั้งหมด

2.1.3 ระบบสนับสนุนการตัดสินใจ (Decision Support System)

ระบบสนับสนุนการตัดสินใจ กล่าวคือ ระบบสารสนเทศที่จะจำลองการแก้ไขปัญหาและจะช่วยแบ่งข้อมูล, และข้อมูลที่มีความจำเป็นมาประกอบในการตัดสินใจ

2.1.3.1 ลักษณะของปัญหา

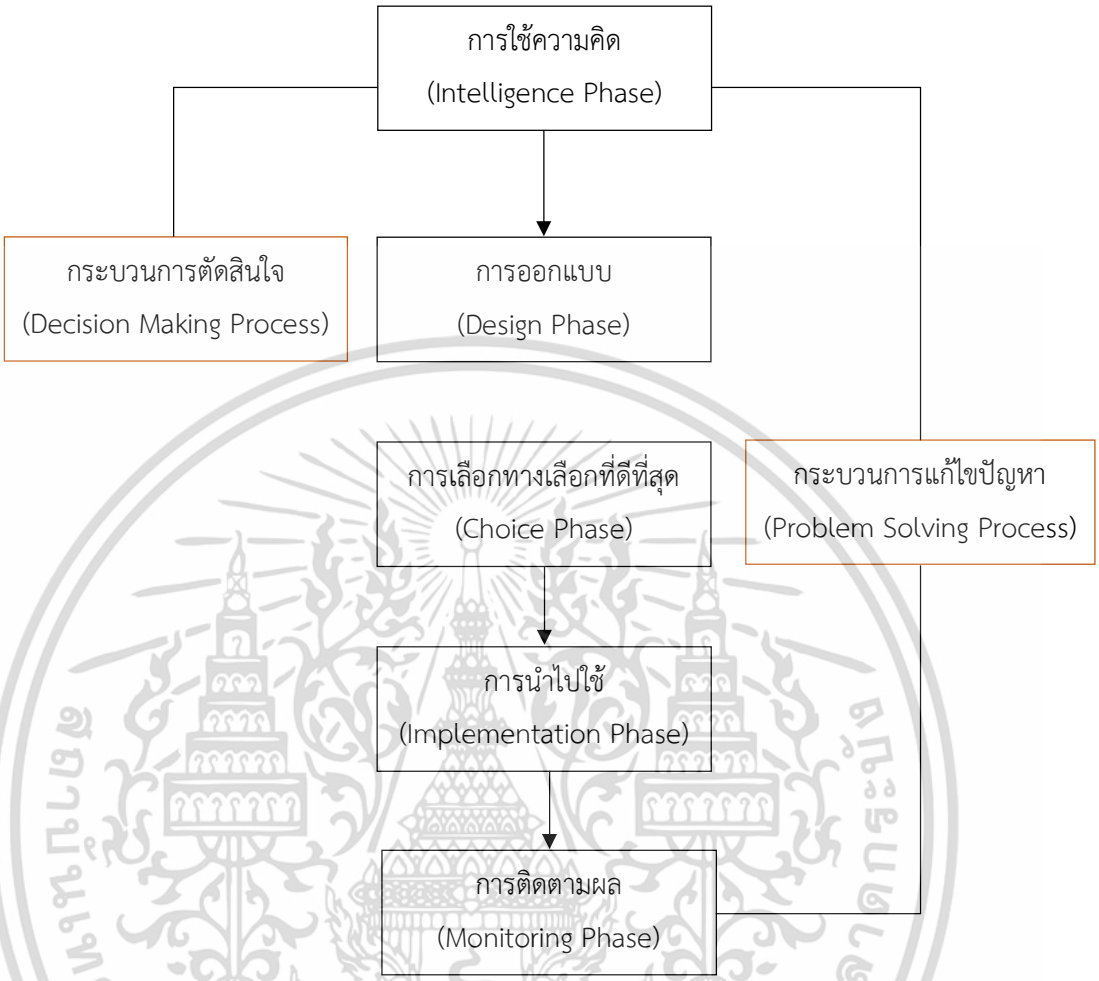
1) ปัญหาในลักษณะที่มีโครงสร้าง (Structured Problem) เป็นปัญหาที่มีแนวทางการแก้ไขปัญหาที่ชัดเจนแน่นอน หรือสามารถจำลองด้วยสมการทางคณิตศาสตร์ (แบบจำลองทางคณิตศาสตร์), และสามารถหาคำตอบที่ชัดเจนได้จากการแทนค่าในสมการ หรือปัญหาที่การตัดสินใจมีข้อมูลและสารสนเทศประกอบการตัดสินใจอย่างครบถ้วนและสามารถนำไปใช้แก้ปัญหาได้โดยการเขียนโปรแกรม (กิตติ, 2550)

2) ปัญหาที่ไม่มีโครงสร้าง (Unstructured Problem) เป็นปัญหาที่ไม่สามารถหาวิธีในการแก้ไขได้อย่างชัดเจน, และแน่นอน ไม่สามารถจำลองได้ด้วยสมการทางคณิตศาสตร์หรือปัญหาที่ผู้ตัดสินใจที่มีข้อมูล, และสารสนเทศไม่เพียงพอต่อการแก้ไขปัญหา จึงต้องอาศัยประสบการณ์ของผู้ตัดสินใจในการแก้ไขปัญหา (กิตติ, 2550)

3) ปัญหาแบบกึ่งโครงสร้าง (Semi Structured Problem) เป็นปัญหาแบบที่มีลักษณะเฉพาะ ส่วนมากจะไม่เกิดซ้ำและไม่มีกระบวนการดำเนินการมาตรฐาน หรือเป็นปัญหาที่มีวิธีในการแก้ไขเพียงบางส่วนเท่านั้น ส่วนที่เหลือจะต้องอาศัยประสบการณ์ หรือความชำนาญจากการเรียนรู้ในการตัดสินใจแก้ไขปัญหา ส่วนเทคโนโลยีสารสนเทศให้ได้แค่การสนับสนุนเท่านั้น (กิตติ, 2550)

2.1.3.2 การตัดสินใจและการแก้ปัญหา

การตัดสินใจจัดว่าเป็นหนึ่งวิธีของกระบวนการแก้ปัญหาของมนุษย์เมื่อพบว่ามีปัญหาเกิดขึ้นในเรื่องหนึ่งเรื่องใดแล้ว การแก้ปัญหาจะผ่านขั้นตอนการตัดสินใจเพื่อแก้ปัญหาต่อไป ส่วนกระบวนการตัดสินใจ (Decision Making Process) คือ การกำหนดขั้นตอนในการตัดสินใจแก้ปัญหาที่เกิดขึ้นภายในองค์กรอย่างมีหลักเกณฑ์ ด้วยการกำหนดขั้นตอนต่าง ๆ โดยผู้ที่กำหนดขั้นตอนคือ George Huber.ได้นำมารวมเข้ากับกระบวนการแก้ปัญหา จึงทำให้กระบวนการตัดสินใจมีทั้งหมด 5 ขั้นตอน ที่แสดงที่รูปที่ 2.4



รูปที่ 2.4 กระบวนการตัดสินใจและแก้ไขปัญหา

กระบวนการตัดสินใจ (Decision Making Process) จะประกอบด้วยการใช้ความคิด (Intelligence Phase) การออกแบบ (Design Phase) การเลือกทางเลือกที่ดีที่สุด (Choice Phase) โดยในส่วนของ การใช้ความคิดนั้นจะเป็นการระบุถึงปัญหาที่พบรวมทั้งจำแนกปัญหาออกมาเป็นส่วนย่อย ๆ, และคิดวิธีที่แก้ปัญหานั้นๆ โดยผลลัพธ์ในขั้นตอนนี้จะเรียกว่า “Decision Statement” ขั้นตอนต่อไปเป็นการออกแบบเพื่อวิเคราะห์ทางเลือกที่ใช้ในการตัดสินใจ โดยขั้นตอนนี้อาจมีการสร้างแบบจำลองการตัดสินใจมาแสดงให้เห็นถึงทางเลือกต่าง ๆ เช่นการสร้างการตัดสินใจออกมาในรูปแบบของต้นไม้ตัดสินใจ (Decision Tree) หรือตารางที่แสดงเงื่อนไขของการตัดสินใจ (Decision Table), และในขั้นตอนสุดท้ายของกระบวนการตัดสินใจ คือการนำทางเลือกที่ดีที่สุดในการนำวิธีการนั้น ๆ ไปแก้ปัญหา

กระบวนการแก้ไขปัญหา (Problem Solving Process) จะเป็นการดำเนินงานต่อเนื่องมาจากการทำกระบวนการตัดสินใจซึ่งจะเสริมการนำไปใช้ (Implementation Phase) เข้าไปรวมไปถึงการ

ติดตามผล (Monitoring Phase) โดยนำกระบวนการเหล่านี้ไปใช้กับข้อมูลจริงเพื่อแก้ปัญหาจริงโดยอาจจะประสบความสำเร็จ หรือไม่ประสบความสำเร็จก็ได้ โดยขั้นตอนที่จะบอกสิ่งนี้ได้คือขั้นตอนการติดตามผล ที่จะสามารถสรุปผลการดำเนินการว่าประสบความสำเร็จหรือไม่

2.2 งานวิจัยที่เกี่ยวข้อง

ณัฐกิจ เจนการ (2563) การพัฒนาแบบจำลองในการตรวจจับข้อความภาษาไทยที่เป็นการกลั่นแกล้งทางไซเบอร์ โดยใช้วิธีซัพพอร์ตเวกเตอร์แมชชีน โดยงานวิจัยฉบับนี้ได้ใช้เทคนิควิธีการทำเหมืองข้อมูล (Data Mining), และการเรียนรู้ของเครื่อง (Machine Learning) มาประยุกต์เพื่อพัฒนาแบบจำลองในการจำแนกข้อความที่แสดงถึงการกลั่นแกล้ง การดูถูก คำหยาบคาย รวมไปถึงการพูดถึงผู้อื่นในด้านที่ไม่ดี ที่สามารถเป็นการกลั่นแกล้งทางไซเบอร์ได้ ซึ่งงานวิจัยชิ้นนี้จะเลือกพัฒนากับข้อความที่เป็นภาษาไทย เพื่อที่จะนำไปพัฒนาต่อยอดเพื่อป้องกันการกลั่นแกล้งจากข้อความที่แสดงถึงการกลั่นแกล้ง การดูถูก คำหยาบคาย รวมไปถึงการพูดถึงผู้อื่นในด้านที่ไม่ดีทางไซเบอร์บนสื่อโซเชียลออนไลน์ได้, และเพื่อลดใช้ข้อความในการกลั่นแกล้งกันทางไซเบอร์อีกด้วย ซึ่งในงานวิจัยชิ้นนี้ได้พัฒนาแบบจำลองที่ใช้อัลกอริทึม Support Vector Machine , K-Nearest Neighbour, และอัลกอริทึม Naïve Bayes ซึ่งเป็นอัลกอริทึมที่เป็นที่รู้จัก ซึ่งผลการทดสอบประสิทธิภาพของทั้งสามอัลกอริทึมจะถูกนำมาเปรียบเทียบกัน โดยการนำ Confusion Matrix มาวัดผลการเปรียบเทียบ ซึ่งจากผลของการเปรียบเทียบประสิทธิภาพของทั้งสามอัลกอริทึมสามารถสรุปได้ว่าอัลกอริทึม Support Vector Machine มีประสิทธิภาพในสูงที่สุดเมื่อเปรียบเทียบกับอัลกอริทึม K-Nearest Neighbour, และอัลกอริทึม Naïve Bayes ซึ่งมีค่าความถูกต้องอยู่ที่ร้อยละ 83.91

อนันต์ ปินะเต (2563) การวิเคราะห์สารสนเทศเพื่อการพัฒนาระบบสนับสนุนการวางแผนการคัดเลือกบุคคลเข้าศึกษาในระบบ TCAS มหาวิทยาลัยมหาสารคาม เนื่องมาจากการบุคคลเข้าศึกษาในสถาบันอุดมศึกษามีการออกนโยบายระบบใหม่ หรือที่เรียกกันว่า TCAS ซึ่งเป็นนโยบายที่มีการปฏิรูปการศึกษาอยู่ 3 ประการ คือ นักเรียนควรจะได้รับสิทธิ์ในการศึกษาไปจนจบมัธยมศึกษาปีที่ 6 เมื่อนักเรียกได้สำเร็จการศึกษาในระดับมัธยมแล้วนั้น นักเรียนจะได้รับสิทธิ์ในการเลือกที่จะตอบรับในสาขาวิชาที่ตนเองต้องการเพียงคนละ 1 สิทธิ์เพื่อสร้างความเสมอภาคกันในหมู่นักเรียน, และสถาบันอุดมศึกษาในเครือข่ายที่ประชุมอธิการบดีแห่งประเทศไทย (ทปอ.) ทุกแห่งจะต้องเข้าร่วมระบบเคลียริงเฮ้าส์ (Clearing House) ด้วยเช่นกันเพื่อเป็นบริหาร 1 สิทธิ์ของนักเรียนที่มีสิทธิ์เข้าศึกษาต่ออย่างเท่าเทียม, และเนื่องมาจากนโยบายการคัดเลือกบุคคลเข้าศึกษาของมหาวิทยาลัยมหาสารคาม ซึ่งเป็นมหาวิทยาลัยของภาครัฐที่อยู่ในเครือข่ายที่ประชุมอธิการบดีแห่งประเทศไทย (ทปอ.) ซึ่งจำเป็นจะต้องมีการดำเนินการตามนโยบายในการคัดเลือกนิสิตที่จะเข้ามาศึกษาระดับ

ปริญญาตรีระบบ TCAS, และจากประเด็นปัญหาการคัดเลือกของผู้ที่มีสิทธิ์เข้าศึกษาต่อในระบบใหม่ พบว่าจำนวนผู้มีสิทธิ์เข้าศึกษาที่ได้ทำการยืนยันสิทธิ์ (Clearing house) นั้นมีจำนวนที่น้อยกว่าแผนสำหรับการรับนักเรียนเข้าศึกษาต่อที่วางแผนร่วมกันเอาไว้ก่อนหน้า เมื่อเปรียบเทียบจำนวนนักเรียนที่มีความสนใจเข้าศึกษาต่อ, และจำนวนของนักเรียนที่สมัครเข้ามาศึกษาต่อที่มีจำนวนมาก ผู้ทำวิจัยจึงเลือกนำเทคนิคเหมืองข้อมูล (Data mining) มาช่วยจัดการแก้ไขปัญหาที่เกิดขึ้นคือจำนวนผู้ยืนยันสิทธิ์น้อยกว่าที่คาดการณ์เอาไว้ โดยเทคนิคเหมืองข้อมูล (Data mining) เป็นเทคนิคในการวิเคราะห์ข้อมูลนักเรียนที่เป็นผู้มีสิทธิ์ในการยืนยันสิทธิ์เข้าศึกษาต่อที่ถูกจัดเก็บไว้ในรูปแบบของข้อมูลสารสนเทศ เพื่อนำรูปแบบที่ได้มาพัฒนาระบบในการวางแผนการคัดเลือกนิสิตใหม่ในระบบ TCAS โดยมีข้อมูลจากเหมืองข้อมูล (Data mining) มาช่วยสนับสนุนการวางแผนนี้เพื่อลดปัญหาจำนวนการยืนยันสิทธิ์ที่น้อยกว่าที่คาดการณ์ไว้ โดยเหมืองข้อมูลจะช่วยคาดการณ์ได้ว่าจะมีจำนวนการยืนยันสิทธิ์มากน้อยเพียงใด ผลการวิจัยพบว่า วิธีสร้างต้นไม้ตัดสินใจ, และการค้นหากฎความสัมพันธ์จากชุดข้อมูลที่นำมาทดลองจากแต่ละสาขาวิชา โดยจะเลือกมาทดลองสามสาขาวิชา ซึ่งจะประกอบไปด้วยสาขาวิชามนุษยศาสตร์และสังคมศาสตร์ที่ได้รับการทดสอบแล้วพบว่ามีค่าความถูกต้องอยู่ที่ 82.85 เปอร์เซ็นต์, และสามารถสร้างเป็นกฎความสัมพันธ์จากชุดข้อมูลได้ทั้งหมด 89 กฎ, กลุ่มสาขาวิชาวิทยาศาสตร์สุขภาพได้รับการทดสอบแล้วพบว่ามีค่าความถูกต้องอยู่ที่ 80.88 เปอร์เซ็นต์, และสามารถสร้างเป็นกฎความสัมพันธ์จากข้อมูลได้ทั้งหมด 85 กฎ, และสาขาวิชาสุดท้ายคือกลุ่มสาขาวิชาวิทยาศาสตร์เทคโนโลยีได้รับการทดสอบแล้วพบว่ามีค่าความถูกต้องอยู่ที่ 78.85 เปอร์เซ็นต์, และสามารถสร้างเป็นกฎความสัมพันธ์ของข้อมูลได้ทั้งหมด 85 กฎ จึงสรุปได้ว่าจากผลการทดลอง จะพบว่าสามารถนำผลการทดลองมาใช้ในการพัฒนาเป็นระบบสนับสนุนการวางแผนการคัดเลือกนิสิตใหม่ในระบบ TCAS ได้

สัณห์ชัย หยีวิยม (2564) การพัฒนาระบบตรวจสอบความถูกต้องของข้อมูลเพื่อให้บริการผ่านเว็บไซต์ สำหรับการสนับสนุนการพัฒนาพื้นที่เป้าหมายชายแดนภาคเหนือ 4 จังหวัด งานวิจัยชิ้นนี้ได้ทำการพัฒนาระบบตรวจสอบความถูกต้องของข้อมูลเพื่อให้บริการผ่านเว็บไซต์ สำหรับการสนับสนุนการพัฒนาพื้นที่เป้าหมายชายแดนภาคเหนือ 4 จังหวัด โดยการพัฒนาระบบครั้งที่ 1 ผู้ทำการวิจัยได้พบประเด็นเรื่องการคัดกรองข้อมูลต่าง ๆ จากในฐานข้อมูลที่เป็นปัญหา ทำให้เกิดความผิดพลาดของการนำข้อมูลมาประมวลผล, และเนื่องด้วยความไม่สมบูรณ์ของข้อมูลที่นักพัฒนาชุมชนนำไปใช้ในการสร้างแผนพัฒนาพื้นที่ทำให้เกิดเป็นปัญหา การวิจัยนี้จึงมุ่งเน้นไปที่เพื่อพัฒนาระบบสำหรับการตรวจสอบ, และ คัดกรองการกรอกข้อมูลแบบสอบถามผ่านเว็บไซต์ ซึ่งการพัฒนาระบบสำหรับตรวจสอบความถูกต้องนี้จะให้ความสำคัญอยู่ 2 ประเด็น ได้แก่ 1. การควบคุมความถูกต้องของข้อมูล, และ 2. การรองรับการตรวจสอบข้อมูลแบบสอบถามด้วยมนุษย์ ซึ่งจะส่งผลให้ข้อมูลที่จะได้รับมีความถูกต้องและมีความสมบูรณ์มากยิ่งขึ้น รวมถึงการถูกตรวจสอบด้วยมนุษย์จะส่งผลให้ข้อมูลมีความถูกต้องก่อนที่จะถูกนำเข้าสู่ฐานข้อมูล ผลการศึกษาและวิจัยของเว็บไซต์การ

พัฒนาระบบตรวจสอบความถูกต้องของข้อมูลเพื่อให้บริการผ่านเว็บไซต์ สำหรับการสนับสนุนการพัฒนาพื้นที่เป้าหมายชายแดนภาคเหนือ 4 จังหวัด พบว่า เว็บไซต์สามารถตรวจสอบความถูกต้องของข้อมูลจากแบบสอบถามซึ่งจะได้รับการตรวจสอบจากผู้ตรวจสอบของภาครัฐ รวมไปถึงความสามารถของเว็บไซต์ที่สามารถตรวจสอบความถูกต้องของความสัมพันธ์ตามธรรมชาติของข้อมูลแบบสอบถาม โดยมี 14 ความสัมพันธ์เป็นส่วนประกอบ โดยที่แต่ละแบบสอบถามจะมีแอตทริบิวต์มากกว่า 700 แอตทริบิวต์ที่นำมาใช้วิเคราะห์ถึงความถูกต้องของข้อมูล นอกจากนี้เมื่อได้เปรียบเทียบกับการศึกษาด้านอื่น ๆ เช่น ด้านการรองรับการบันทึกข้อมูลที่มีจำนวนมากกว่า 10,000 แบบสอบถาม, ด้านการเก็บข้อมูลที่ถูกออกแบบให้เป็นไปตามแบบสอบถาม, และด้านการตรวจสอบความสัมพันธ์ของข้อมูลในการบันทึกข้อมูล

กนิษฐา อินธิชิต, และคณะ (2566) ได้พัฒนาแอปพลิเคชันช่วยตัดสินใจในการเลือกเรียนสาขาวิชาคอมพิวเตอร์ ในมหาวิทยาลัยราชภัฏศรีสะเกษบนระบบปฏิบัติการแอนดรอยด์ โดยใช้เทคนิคต้นไม้ตัดสินใจ จากวารสารวิชาการการจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม โดยงานวิจัยชิ้นนี้จะมุ่งเน้นไปที่การนำต้นไม้ตัดสินใจมาช่วยตัดสินใจในการเลือกเรียนสาขาวิชาคอมพิวเตอร์ในมหาวิทยาลัยราชภัฏศรีสะเกษโดยแอปพลิเคชันบนระบบปฏิบัติการแอนดรอยด์, และให้ผู้เชี่ยวชาญตรวจสอบความถูกต้อง พร้อมประเมินความเหมาะสมของแอปพลิเคชันช่วยตัดสินใจในการเลือกเรียนสาขาวิชาคอมพิวเตอร์ในมหาวิทยาลัยราชภัฏศรีสะเกษบนระบบปฏิบัติการแอนดรอยด์ รวมถึงการประเมินความพึงพอใจของการใช้งานแอปพลิเคชันช่วยตัดสินใจในการเลือกเรียนสาขาวิชาคอมพิวเตอร์ในมหาวิทยาลัยราชภัฏศรีสะเกษบนระบบปฏิบัติการแอนดรอยด์ จากผู้ใช้งาน ผลการวิจัยพบว่าแอปพลิเคชันที่ได้นำต้นไม้ตัดสินใจมาพัฒนาเพื่อวิเคราะห์ข้อมูลที่ผู้ใช้ผ่านการทำแบบทดสอบแล้วเก็บมาเป็นคะแนน, และนำคะแนนที่ได้มาเปรียบเทียบหาความเหมาะสมของผู้ใช้กับสาขาวิชาและแสดงผลการเปรียบเทียบที่ได้จากการวิเคราะห์ข้อมูลของแอปพลิเคชันออกมาในรูปแบบของข้อความ โดยจะเรียงลำดับความเหมาะสมของผู้ใช้กับสาขาวิชาตามคะแนนที่ได้สูงสุด 3 อันดับแรก นอกจากนั้นผลการตรวจสอบและประเมินความเหมาะสมจากผู้เชี่ยวชาญพบว่าแอปพลิเคชันมีความเหมาะสมในการนำมาช่วยตัดสินใจในการเลือกเรียนสาขาวิชาคอมพิวเตอร์ ในมหาวิทยาลัยราชภัฏศรีสะเกษอยู่ในระดับสูง ซึ่งค่าเฉลี่ย 4.03, และมีส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 0.49 รวมถึงการประเมินความพึงพอใจ จากผู้ใช้งานโดยทั่วไปจากการทำแบบสอบถาม สามารถนำมาสรุปได้ว่า ผู้ใช้งานมีแอปพลิเคชันช่วยตัดสินใจในการเลือกเรียนสาขาวิชาคอมพิวเตอร์ในมหาวิทยาลัยราชภัฏศรีสะเกษบนระบบปฏิบัติการแอนดรอยด์มีความพึงพอใจอยู่ในระดับสูง โดยมีค่าเฉลี่ยเท่ากับ 4.53, และส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 0.69

บทที่ 3
วิธีดำเนินการวิจัย

การดำเนินการพัฒนาอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้นได้นำแนวทางการปฏิบัติผ่านกระบวนการการวิเคราะห์และออกแบบระบบ (System Analysis and Design) รวมถึงการออกแบบระบบฐานข้อมูล (Database System) เพื่อให้การบริการจัดการระบบทำงานได้อย่างมีประสิทธิภาพ, รวดเร็ว, เหมาะสม, และมีความถูกต้อง โดยจะต้องสามารถรายงาน, และสรุปผลตามของรูปแบบต่าง ๆ ได้โดยมีรายละเอียดการดำเนินการต่าง ๆ ดังนี้

3.1 การเตรียมข้อมูล

ก่อนที่จะเริ่มสร้างต้นไม้การตัดสินใจของอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ได้รับการพัฒนาจำเป็นจะต้องทำตามขั้นตอนต่อไปนี้

1) อ่านข้อมูลจากไฟล์ดาต้าเซต
2) กำหนดชุดของหัวตารางให้เป็นคอลัมน์
3) กำหนดให้คลาสเป้าหมายคือคอลัมน์สุดท้ายของชุดหัวข้อ
4) กำหนดคอลัมน์ให้เป็นชุดของคลาส
5) ใช้ K-fold cross-validation ที่ K = 10 เพื่อหลีกเลี่ยงปัญหาความหนาแน่นมากเกินไป

3.2 วิธีดำเนินการ

ผู้จัดทำได้ทำการศึกษา วิเคราะห์, และศึกษาถึงการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ เพื่อหาแนวทางในการพัฒนาระบบงานที่เหมาะสม โดยมีรายละเอียดและวิธีการดำเนินการดังนี้

1. กำหนดปัญหา (Problem Definition)
    1.1 รับรู้สภาพปัญหาของอัลกอริทึม ID3 แบบดั้งเดิม
    1.2 ศึกษาและสรุปสาเหตุของปัญหา
    1.3 รวบรวมทฤษฎีหรืองานวิจัยที่เกี่ยวข้อง
    1.4 สรุปข้อกำหนดต่าง ๆ ให้มีความชัดเจน, ถูกต้อง, และเป็นที่ยอมรับ
2. วิเคราะห์ (Analysis)
    2.1 วิเคราะห์การดำเนินการของอัลกอริทึม ID3 แบบดั้งเดิม
    2.2 กำหนดข้อจำกัดและช่องโหว่ของอัลกอริทึม ID3 แบบดั้งเดิม

2.3 สร้างแบบการทดลองทั้ง ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ได้รับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

3. ออกแบบ (Design)

3.1 ออกแบบการรายงานผล (Output Design)

3.2 ออกแบบข้อมูลนำเข้า, และรูปแบบการรับข้อมูล (Input Design)

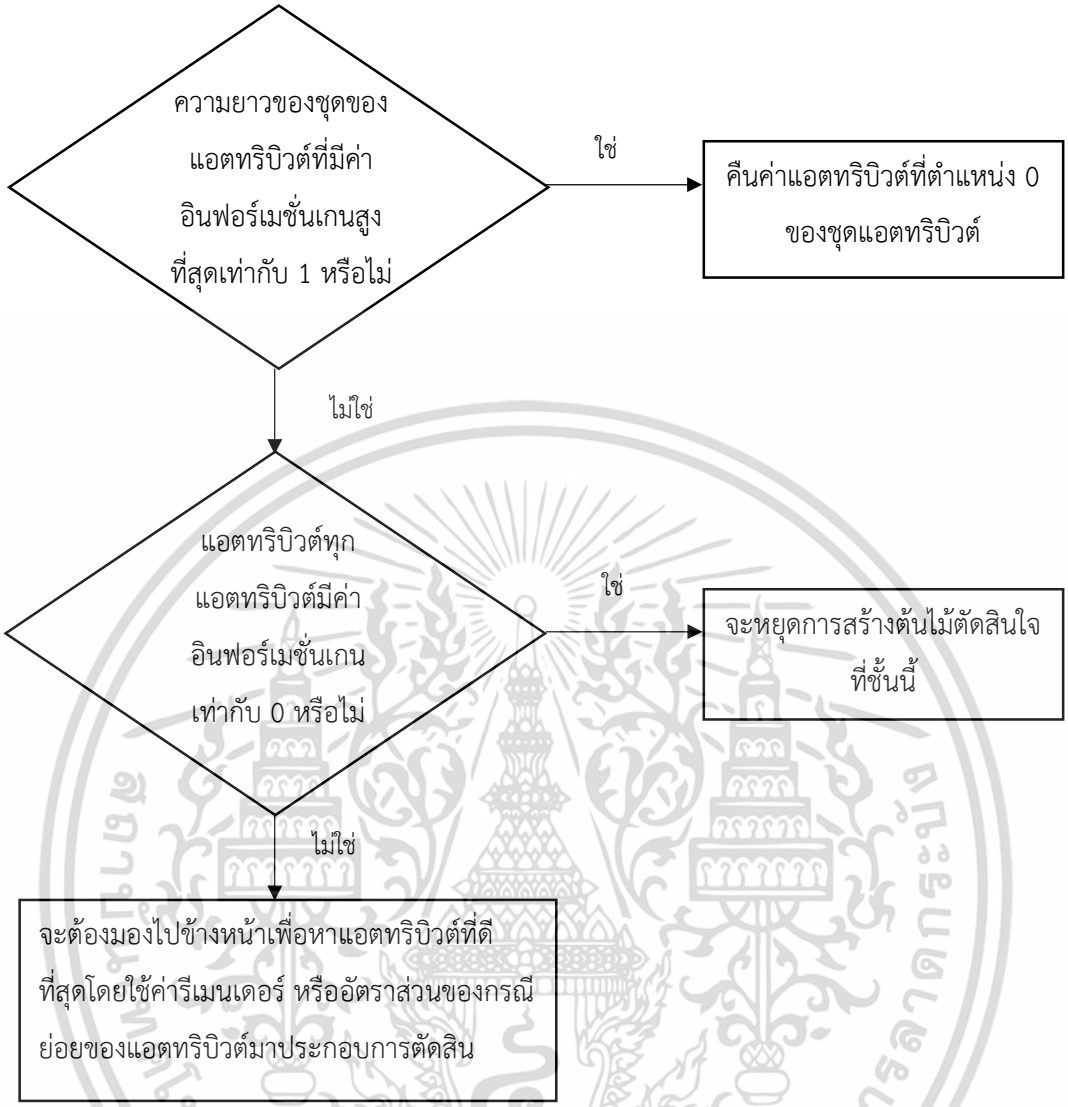
3.3 ออกแบบผังระบบ (System Flowchart)

3.4 ออกแบบฐานข้อมูล (Database Design)

3.5 สร้างพจนานุกรมข้อมูล (Data Dictionary)

3.3 วิธีการปรับปรุงอัลกอริทึม ID3

การปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์จำเป็นจะต้องเพิ่มขั้นตอนบางขึ้นตอนเพื่อที่จะได้รับค่าที่มีความสำคัญเพิ่มขึ้น ในหัวข้อนี้เราจะทำการเพิ่มขั้นตอนใหม่สำหรับการเลือกแอตทริบิวต์ หรือฟีเจอร์ที่ดีที่สุด โดยสามขั้นตอนจะถูกอธิบายเพิ่มเติมด้านล่าง



รูปที่ 3.1 เงื่อนไขแรกใช้สำหรับหาแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนที่สูงที่สุดเพียงหนึ่งเดียว

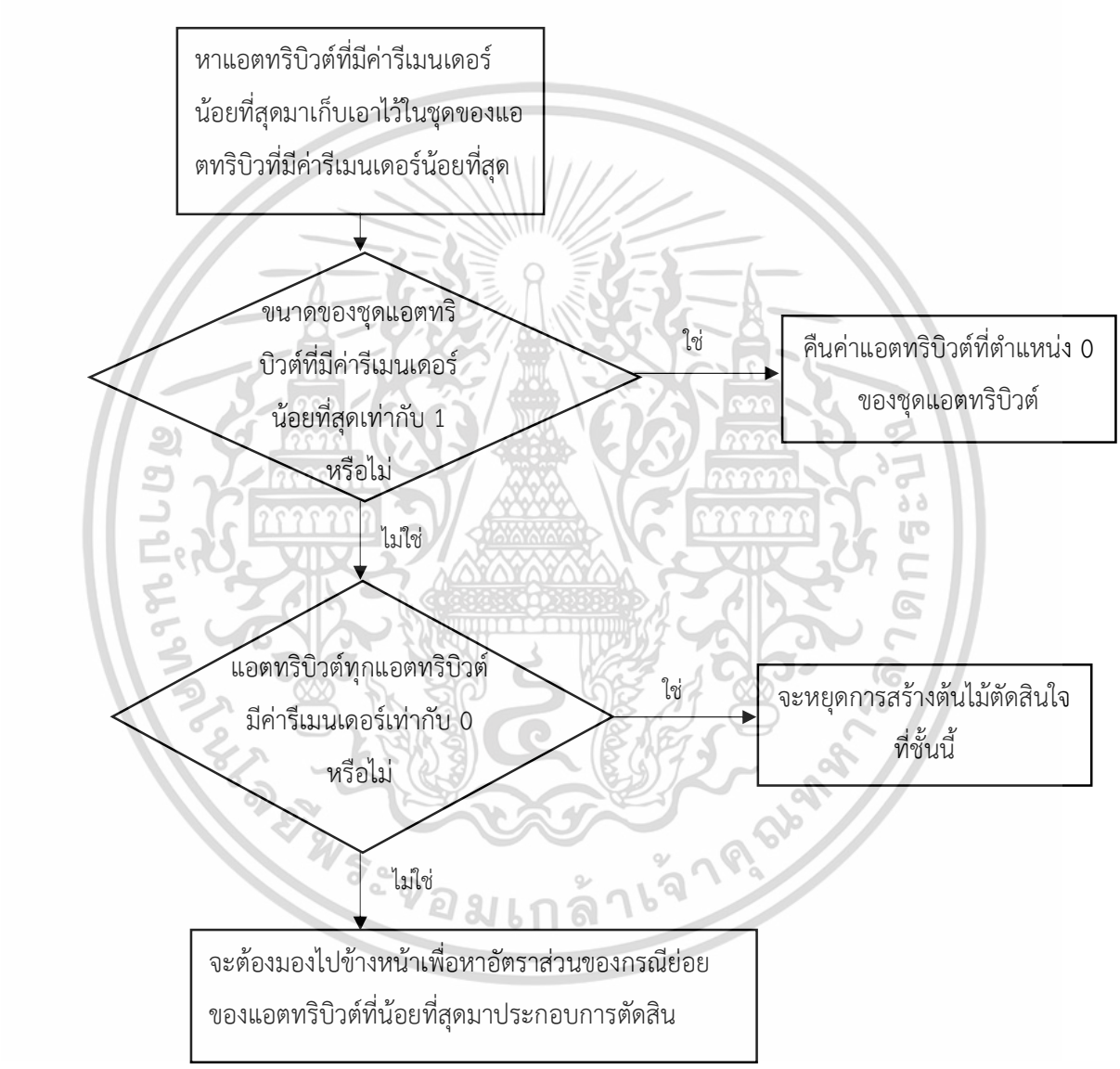
ขั้นตอนแรกจะเน้นไปที่การหาแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดเพียงหนึ่งแอตทริบิวต์ โดยจะตรวจสอบจากขนาดของชุดของแอตทริบิวต์ ที่ได้อธิบายเงื่อนไขการตรวจสอบไว้ในรูปที่ 3.1 ที่ได้แสดงถึงขั้นตอนที่หนึ่งที่เริ่มจากการเพิ่มเงื่อนไขเพื่อตรวจสอบขนาดของชุดของแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุด

ถ้าขนาดของชุดแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดเท่ากับ 1 นั้นจะหมายถึงว่าชุดข้อมูลจะประกอบไปด้วยแอตทริบิวต์ที่ดีที่สุดเพียงตัวเดียวเท่านั้น ดังนั้นเงื่อนไขก็จะทำการรีเทิร์นแอตทริบิวต์ที่ดัชนีตัวแรกเป็นแอตทริบิวต์ที่ดีที่สุด

อย่างไรก็ตามถ้าหากขนาดของชุดแอตทริบิวต์มากกว่า 1 จะเข้าเงื่อนไขที่ 2 เพื่อตรวจสอบว่าแอตทริบิวต์ทุกตัวในชุดแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมีค่าเท่ากับ 0 หรือไม่ หากว่า

แอตทริบิวต์ทุกตัวมีค่าเป็น 0 จะไม่สร้างต้นไม้ต่อในชั้นถัด ๆ ไป จะวนกลับขึ้นไปที่โหนดก่อนหน้าเพื่อหาเส้นทางอื่น ๆ

แต่หากขนาดของชุดแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1, และแอตทริบิวต์ทุกตัวในชุดแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดไม่เท่ากับ 0 จะมองต่อไปข้างหน้าด้วยค่ารีเมนเดอร์ที่น้อยที่สุดที่จะแสดงไว้ในรูปที่ 3.2

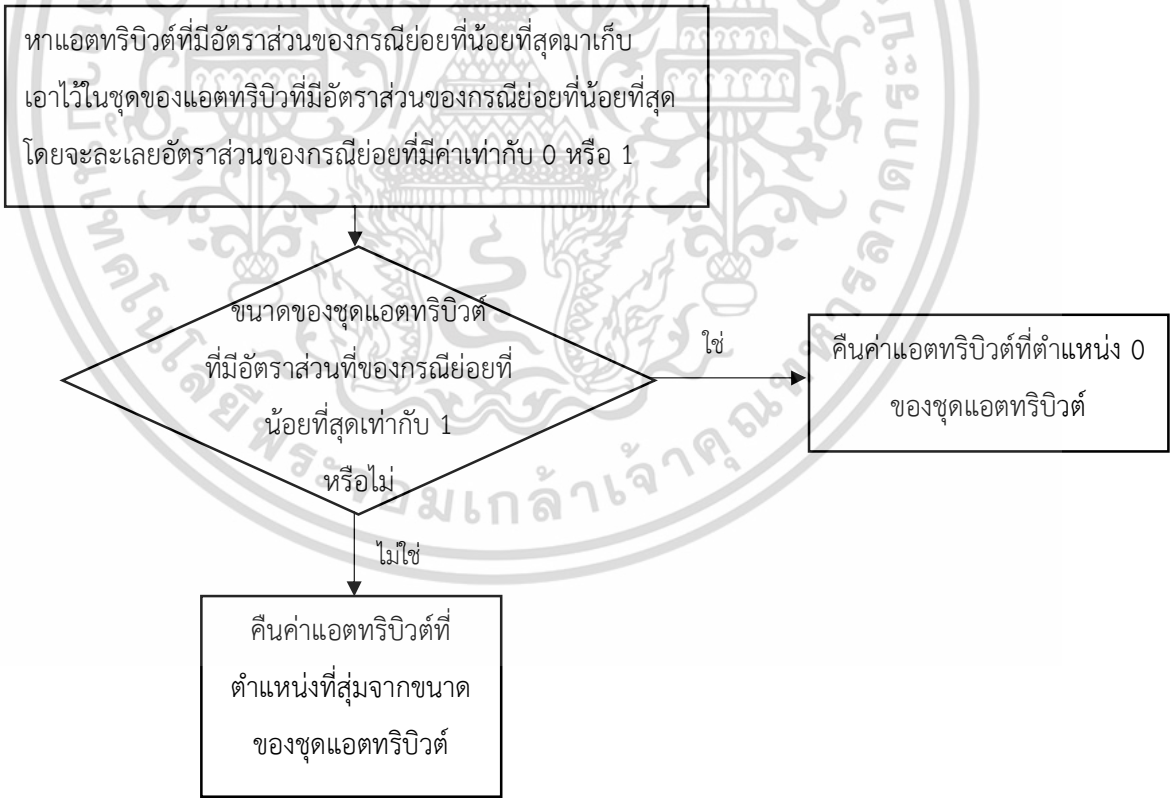


รูปที่ 3.2 เงื่อนไขที่สองใช้สำหรับหาแอตทริบิวต์ที่มีค่ารีเมนเดอร์ที่น้อยที่สุดเพียงหนึ่งเดียว

ขั้นตอนที่สองจะรับเอาชุดแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมาตรวจสอบว่ามีแอตทริบิวต์ตัวใดบ้างที่มีค่ารีเมนเดอร์ที่น้อยที่สุด โดยจะนำแอตทริบิวต์นั้น ๆ มาสร้างเป็นชุดของแอตทริบิวต์ที่มีค่ารีเมนเดอร์ที่น้อยที่สุด จากนั้นจึงตรวจสอบด้วยเงื่อนไข ถ้าขนาดของชุดข้อมูลเท่ากับ 1 นั้นจะหมายถึงว่าชุดข้อมูลจะประกอบไปด้วยแอตทริบิวต์ที่ดีที่สุดเพียงตัวเดียวเท่านั้น ดังนั้นเงื่อนไขก็จะทำการส่งแอตทริบิวต์ที่ดัชนีตัวแรกกลับไปเป็นแอตทริบิวต์ที่ดีที่สุด

อย่างไรก็ตามถ้าหากขนาดของชุดแอตทริบิวต์มากกว่า 1 จะเข้าเงื่อนไขที่ 2 เพื่อตรวจสอบว่าแอตทริบิวต์ทุกตัวในชุดแอตทริบิวต์ที่มีค่ารีเมนเดอร์ที่น้อยที่สุดมีค่าเท่ากับ 0 หรือไม่ หากว่าแอตทริบิวต์ทุกตัวมีค่ารีเมนเดอร์เป็น 0 จะไม่สร้างต้นไม้ต่อในชั้นถัด ๆ ไป จะวนกลับขึ้นไปที่โหนดก่อนหน้าเพื่อหาเส้นทางอื่น ๆ

แต่หากขนาดของชุดแอตทริบิวต์ที่มีค่ารีเมนเดอร์ที่น้อยที่สุดมากกว่า 1, และแอตทริบิวต์ทุกตัวในชุดแอตทริบิวต์ที่มีค่ารีเมนเดอร์ที่น้อยที่สุดไม่เท่ากับ 0 จะมองต่อไปข้างหน้าด้วยอัตราส่วนของกรณีย่อยของแอตทริบิวต์ที่น้อยที่สุดที่จะแสดงไว้ในรูปที่ 3.3



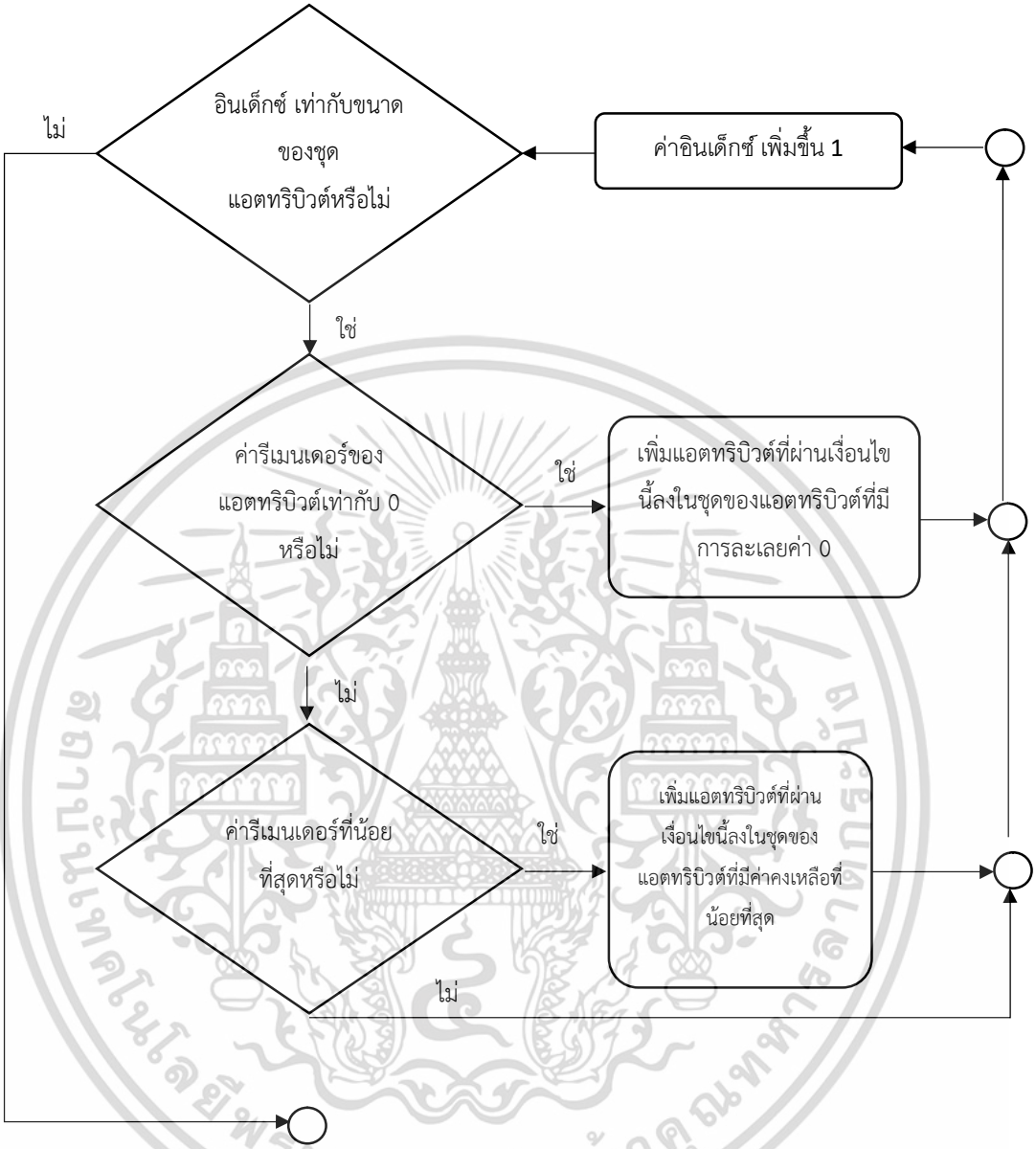
รูปที่ 3.3 เงื่อนไขสุดท้ายใช้สำหรับหาแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยน้อยที่สุดเพียงหนึ่งเดียว

ขั้นตอนสุดท้ายจะรับเอาชุดแอตทริบิวต์ที่มีค่ารีเมนเดอร์ที่น้อยที่สุดมาตรวจสอบว่ามีแอตทริบิวต์ตัวใดบ้างที่มีอัตราส่วนของกรณีย่อยที่น้อยที่สุดโดยจะละเลยอัตราส่วนของกรณีย่อยที่มีค่าเท่ากับ 0 หรือ 1 โดยจะนำแอตทริบิวต์นั้น ๆ มาสร้างเป็นชุดของแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยที่น้อยที่สุด จากนั้นจึงตรวจสอบด้วยเงื่อนไข ถ้าขนาดของชุดข้อมูลเท่ากับ 1 นั้นจะหมายถึงว่าชุดข้อมูลจะประกอบไปด้วยแอตทริบิวต์ที่ดีที่สุดเพียงตัวเดียวเท่านั้น ดังนั้นเงื่อนไขก็จะทำการส่งแอตทริบิวต์ที่ดัชนีตัวแรกกลับไปเป็นแอตทริบิวต์ที่ดีที่สุด

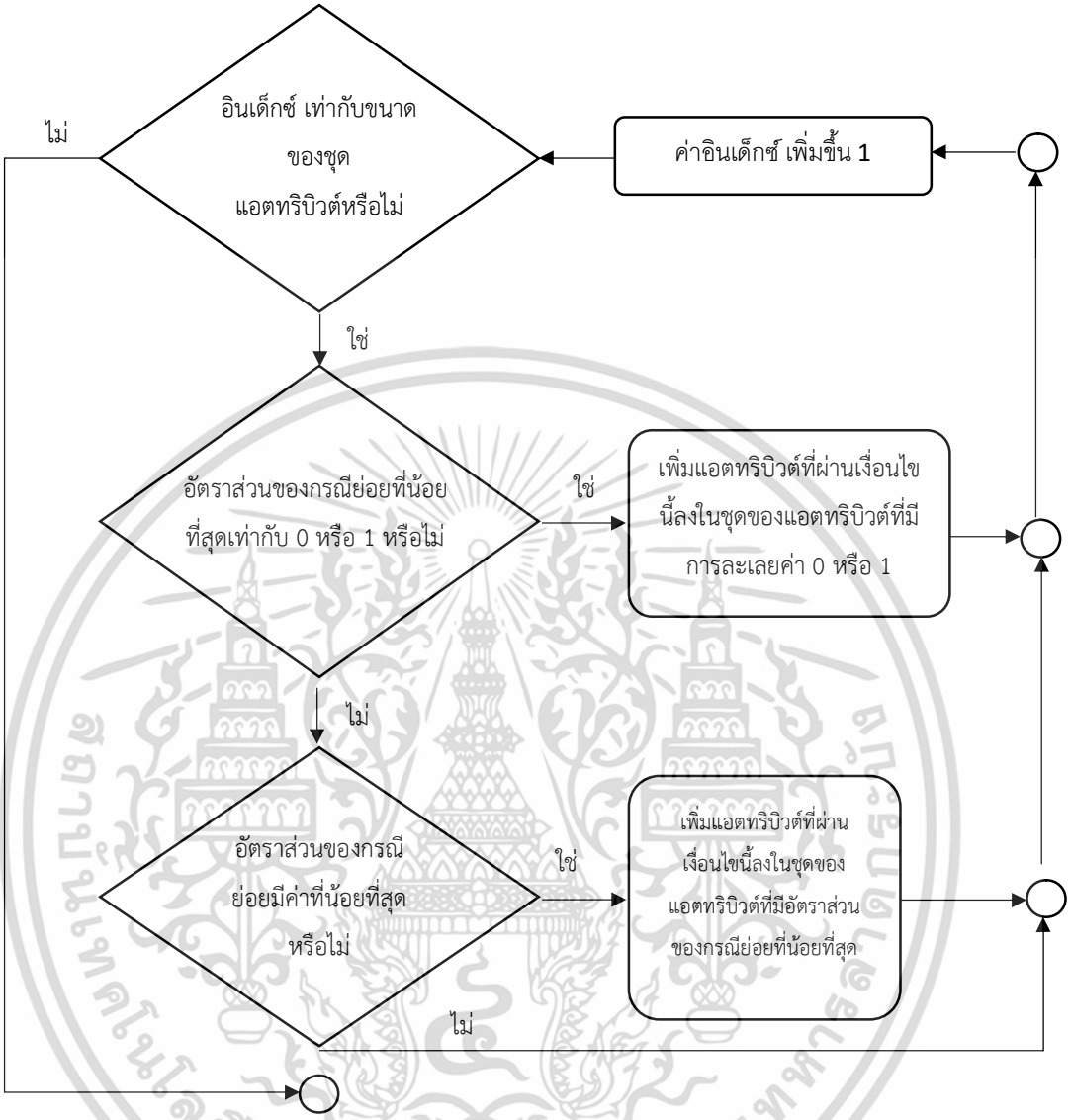
อย่างไรก็ตามถ้าหากขนาดของชุดแอตทริบิวต์มากกว่า 1 นั้นจะทำการสุ่มเลือกแอตทริบิวต์ 1 ตัวจากชุดของแอตทริบิวต์มากำหนดให้เป็นโหนด ตามที่ได้แสดงไว้ในรูปที่ 3.3

ไม่เพียงแค่สามขั้นตอนที่ถูกเพิ่มเติมเข้ามาในฟังก์ชันที่ทำการหาแอตทริบิวต์ที่ดีที่สุด ยังได้เพิ่มเงื่อนไขที่จะแสดงอยู่ในรูปที่ 3.4 แสดงให้เห็นถึงวิธีการหาแอตทริบิวต์ที่มีค่ารีเมนเดอร์น้อยที่สุดมาเก็บเอาไว้ในชุดของแอตทริบิวต์ที่มีค่ารีเมนเดอร์น้อยที่สุด แล้วจึงละเว้นกรณีแอตทริบิวต์ที่มีค่ารีเมนเดอร์เท่ากับ 0 โดยการวนตรวจสอบทุกแอตทริบิวต์ในชุดของแอตทริบิวต์ที่มีค่ารีเมนเดอร์สูงที่สุด เพื่อลดจำนวนของแอตทริบิวต์ลง

จากรูปที่ 3.5 แสดงให้เห็นถึงวิธีการหาแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยน้อยที่สุดมาเก็บเอาไว้ในชุดของแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยน้อยที่สุด แล้วจึงละเว้นกรณีแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยเท่ากับ 0 หรือ 1 โดยการวนตรวจสอบทุกแอตทริบิวต์ในชุดของแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยน้อยที่สุด เพื่อหาแอตทริบิวต์ที่ดีที่สุด



รูปที่ 3.4 เงื่อนไขการวนซ้ำสำหรับการละเว้นกรณีที่ค่ารีเมนเดอร์น้อยที่สุด เท่ากับ 0 ในฟังก์ชัน



รูปที่ 3.5 เงื่อนไขการวนซ้ำสำหรับการละเว้นกรณีที่อัตราส่วนของกรณีย่อยที่น้อยที่สุดเท่ากับ 0 หรือ 1 ในฟังก์ชัน

3.4 หลักการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

3.4.1 เงื่อนไขของการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

1. ชุดข้อมูลที่นำมาสร้างต้นไม้ตัดสินใจนั้น มีจำนวนแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 ตัวหรือไม่

2. ถ้ามีจำนวนแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 ตัว หากเป็นอัลกอริทึม ID3 แบบดั้งเดิมจะทำการเลือกแอตทริบิวต์ตัวแรกของชุดของแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากำหนดให้เป็นโหนด แต่อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้น จะยังไม่ทำการเลือกแอตทริบิวต์จากจำนวนแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1, และจะละเลยกรณีที่ค่าอินฟอร์เมชันเกนของทุกแอตทริบิวต์มีค่าไม่เท่ากับ 0 โดยจะนำค่ารีเมนเดอร์ (Remainder), และอัตราส่วนของกรณีย่อยของแอตทริบิวต์นั้นมาประกอบกับค่าอินฟอร์เมชัน โดยที่จะไม่ทำการหาแอตทริบิวต์ที่ดีที่สุดต่อ หากว่าค่าอินฟอร์เมชันเกนสูงที่สุดเท่ากับ 0

3. เมื่อจำนวนแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 แอตทริบิวต์และค่าอินฟอร์เมชันเกนของทุกแอตทริบิวต์ไม่เท่ากับ 0 แล้ว อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้น จะนำค่ารีเมนเดอร์ (Remainder) ของแอตทริบิวต์ทุกตัวมาเปรียบเทียบหาว่า แอตทริบิวต์ตัวใดมีค่ารีเมนเดอร์น้อยที่สุดหรือไม่ หากว่ามีเพียงหนึ่งแอตทริบิวต์จากจำนวนแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 มีค่ารีเมนเดอร์ (Remainder) น้อยที่สุด ก็จะกำหนดให้แอตทริบิวต์ตัวนั้น ๆ เป็นแอตทริบิวต์ที่มีความสำคัญ อย่างไรก็ตามหากว่าค่ารีเมนเดอร์ (Remainder) น้อยที่สุดยังมีมากกว่า 1 แอตทริบิวต์ จะทำการหาแอตทริบิวต์ที่ดีที่สุดต่อด้วยอัตราส่วนของกรณีย่อยของแอตทริบิวต์ โดยจะละเว้นในกรณีที่ค่ารีเมนเดอร์ (Remainder) ที่น้อยที่สุดเท่ากับ 0

4. แต่ถ้ายังไม่สามารถหาแอตทริบิวต์ที่มีค่ารีเมนเดอร์ (Remainder) น้อยที่สุดได้ก็จำเป็นจะต้องเปรียบเทียบอัตราส่วนของกรณีย่อยของแอตทริบิวต์ โดยจะหาแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยของแอตทริบิวต์ที่น้อยที่สุด โดยที่จะละเลยแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยของแอตทริบิวต์เท่ากับ 0 หรือ 1

5. เมื่อยังไม่สามารถหาแอตทริบิวต์ที่มีอัตราส่วนของกรณีย่อยของแอตทริบิวต์ที่น้อยที่สุดเพียง 1 ตัวได้ จะทำการสุ่มเลือกแอตทริบิวต์จากจำนวนแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 แอตทริบิวต์มากำหนดให้เป็นโหนด

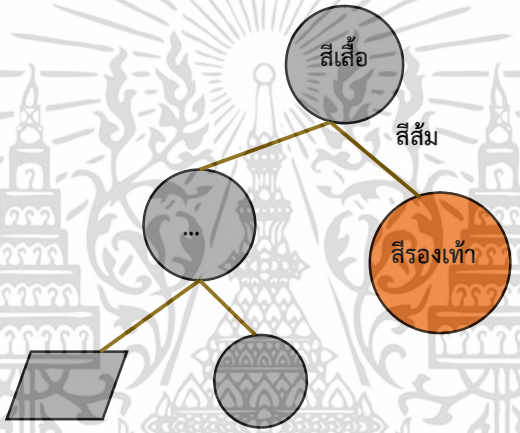
3.4.2 ตัวอย่างข้อมูลกรณีจำนวนแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 แอตทริบิวต์

ตัวอย่างข้อมูลการตัดสินใจเลือกกินขนมเพียง 1 อย่างของเด็กผู้หญิงทั้งหมด 50 คน โดยมีแอตทริบิวต์ทั้งหมด 8 แอตทริบิวคือ “สีเสื้อ”, “การเลือกใส่กระโปรงหรือกางเกง”, “สีรองเท้า”, “เลือกถือกระเป๋าหรือตุ๊กตา”, “มาซื้อกับคุณพ่อหรือคุณแม่”, “ตำแหน่งของชั้นวางขนม”, “สีของถุงขนม”,

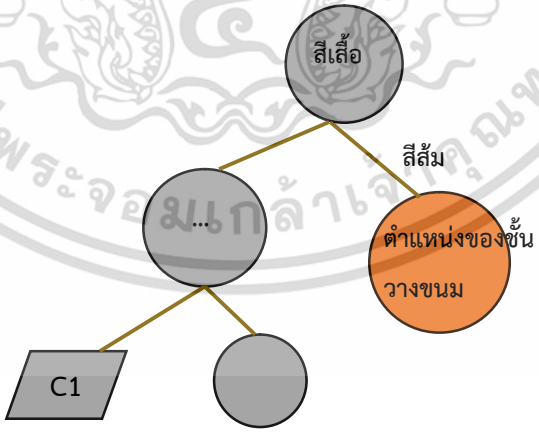
และ “ขนมที่มีของแถมและไม่มีของแถม” โดยจะมีคลาสการตัดสินใจอยู่ที่ ซื้อ, ไม่ซื้อ, และไม่สามารถเลือกขนมเพียง 1 ชิ้นได้

เริ่มต้นแอตทริบิวต์ “มาซื้อกับคุณพ่อหรือคุณแม่” เป็นแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุด จึงกำหนดให้เป็นโหนดราก ทำให้จะเหลือแอตทริบิวต์เพียง 7 ตัวที่สามารถนำไปกำหนดให้เป็นโหนดได้

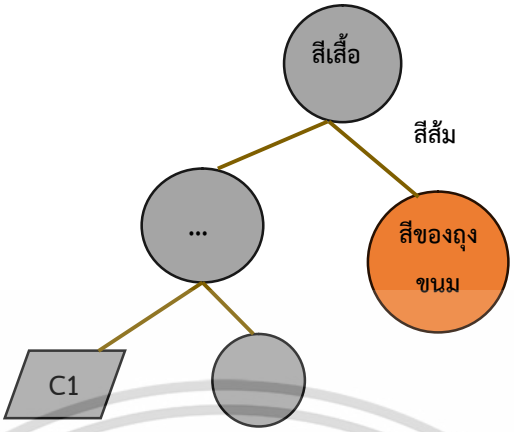
เมื่อผ่านการสร้างต้นไปจนพบว่ามีแอตทริบิวที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 แอตทริบิวต์ คือ “สีรองเท้า”, “ตำแหน่งของชั้นวางขนม”, และ “สีของถุงขนม” เมื่อ “สีเสื้อ” มีค่าเป็น “สีส้ม” ตามที่ได้แสดงอยู่ในรูป 3.6, 3.7, และ 3.8 ตามลำดับ



รูปที่ 3.6 ต้นไม้การตัดสินใจที่กำหนดให้ “สีรองเท้า” เป็นโหนด เมื่อ “สีเสื้อ” มีค่าเป็น “สีส้ม”



รูปที่ 3.7 ต้นไม้การตัดสินใจที่กำหนดให้ “ตำแหน่งของชั้นวางขนม” เป็นโหนด เมื่อ “สีเสื้อ” มีค่าเป็น “สีส้ม”



รูปที่ 3.8 ต้นไม้การตัดสินใจที่กำหนดให้ “สีของถุงขนม” เป็นโหนด เมื่อ “สีเสื้อ” มีค่าเป็น “สีส้ม”

เมื่อแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงที่สุดมากกว่า 1 แอตทริบิวต์ คือ “สีรองเท้า”, “ตำแหน่งของชั้นวางขนม”, และ “สีของถุงขนม” อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ จะมองไปข้างหน้าเพื่อหาค่ารีเมนเดอร์ที่น้อยที่สุดมาประกอบกับค่าอินฟอร์เมชันเกนในการตัดสินใจเลือกแอตทริบิวต์ตัวใดตัวหนึ่งขึ้นมา

อย่างไรก็ตามแอตทริบิวต์ “สีรองเท้า”, “ตำแหน่งของชั้นวางขนม”, และ “สีของถุงขนม” ก็มีค่ารีเมนเดอร์ที่น้อยที่สุดเท่ากันเช่นเดิม จึงจำเป็นจะต้องมองไปข้างหน้าในระดับของอัตราส่วนของกรณีย่อยของแต่ละแอตทริบิวต์

แอตทริบิวต์ “สีรองเท้า” มี 35 อินสแตนซ์ เมื่อ “สีเสื้อ” มีค่าเป็น “สีส้ม” ที่สามารถแบ่งเป็นกรณีย่อยได้ดังนี้

1) “สีแดง” มี 16 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 8 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 8 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 0 กรณีย่อย

   $Entropy$ (สีรองเท้า| “สีแดง”) $= -[(8/16 \times \log_2 8/16) + (8/16 \times \log_2 8/16) + (0/16 \times \log_2 0/16)] = 1$

   P (สีรองเท้า| “สีแดง”) $= 16/35 = 0.45$

2) “สีขาว” มี 9 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 3 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 3 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 3 กรณีย่อย

*Entropy (*สีรองเท้า| “สีขาว”*) = - [(3/9 x $log_2$3/9) + (3/9 x $log_2$3/9) + (3/9 x $log_2$3/9)] = 1.58*

*P (*สีรองเท้า| “สีขาว”*) = 9/35 = 0.26*

3) “สีฟ้า” มี 10 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 3 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 3 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 4 กรณีย่อย

*Entropy (*สีรองเท้า| “สีฟ้า”*) = - [(3/10 x $log_2$3/10) + (3/10 x $log_2$3/10) + (4/10 x $log_2$4/10)] = 1.57*

*P (*สีรองเท้า| “สีฟ้า”*) = 10/35 = 0.29*

นำข้อมูลของกรณีย่อยมาคำนวณเป็นเอนโทรปีของกรณีส่วนย่อยของสีรองเท้ากรณี “สีแดง”, “สีขาว”, “สีฟ้า” ซึ่งได้แสดงไว้ในตารางที่ 3.1 ค่าเอนโทรปีของกรณีย่อยที่มีค่าเป็นคลาส “ซื้อ”, “ไม่ซื้อ”, และ “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ของแอตทริบิวต์ “สีรองเท้า” มาเปรียบเทียบ โดยจะละเลยกรณีย่อยที่มีผลรวมของค่าเอนโทรปีเท่ากับ 0 หรือ 1 ซึ่งกรณีย่อย “สีแดง” จะถูกละเว้นเนื่องจากมีผลรวมของค่าเอนโทรปีเท่ากับ 0 หรือ 1, และกรณีย่อยที่มีผลรวมของค่าเอนโทรปีน้อยที่สุดคือ “สีฟ้า” ที่มีผลรวมของค่าเอนโทรปีเท่ากับ 1.57

ตารางที่ 3.1 อัตราส่วนของแต่ละกรณีย่อยของแอตทริบิวต์ “สีรองเท้า”

กรณีย่อย / คลาส	**“สีแดง”**	**“สีขาว”**	**“สีฟ้า”**
“ซื้อ”	0.50	0.53	0.53
“ไม่ซื้อ”	0.50	0.53	0.52
“ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้”	0.00	0.53	0.52
ผลรวม	1.00	1.59	1.57

แอตทริบิวต์ถัดมาคือแอตทริบิวต์ “ตำแหน่งของชั้นวางขนม” ซึ่งจะนับจากชั้นสูงที่สุดสุดไล่ต่ำลงมาตามลำดับ โดยอินสแตนซ์ทั้งหมด 35 อินสแตนซ์ที่สามารถแบ่งเป็นกรณีย่อยคือ “ชั้นที่ 1” มี

7 อินสแตนซ์, “ชั้นที่ 2” มี 7 อินสแตนซ์, “ชั้นที่ 3” มี 14 อินสแตนซ์, และ “ชั้นที่ 4” มี 7 อินสแตนซ์ เมื่อ “สีเสื้อ” มีค่าเป็น “สีส้ม” โดยที่แต่ละกรณีย่อยจะมีจำนวนของคลาสที่แตกต่างกัน

1) “ชั้นที่ 1” มี 7 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 3 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 1 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 3 กรณีย่อย

   Entropy (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 1”***) = - [(3/7 x $log_2$3/7) + (1/7 x $log_2$1/7) + (3/7 x $log_2$3/7)] = 1.45***

   P (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 1”***) = 7/35 = 0.20***

2) “ชั้นที่ 2” มี 7 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 3 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 3 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 1 กรณีย่อย

   Entropy (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 2”***) = - [(3/7 x $log_2$3/7) + (3/7 x $log_2$3/7) + (1/7 x $log_2$1/7)] = 1.45***

   P (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 2”***) = 7/35 = 0.20***

3) “ชั้นที่ 3” มี 7 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 6 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 1 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 0 กรณีย่อย

   Entropy (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 3”***) = - [(6/7 x $log_2$6/7) + (1/7 x $log_2$1/7) + (0/7 x $log_2$0/7)] = 0.59***

   P (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 3”***) = 7/35 = 0.20***

4) “ชั้นที่ 4” มี 14 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 6 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 3 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 5 กรณีย่อย

   Entropy (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 4”***) = - [(6/14 x $log_2$6/14) + (3/14 x $log_2$3/14) + (5/14 x $log_2$5/14)] = 1.53***

   P (ตำแหน่งของชั้นวางขนม| “ชั้นที่ 4”***) = 14/35 = 0.40***

นำข้อมูลของกรณีย่อยมาคำนวณเป็นเอนโทรปีของกรณีส่วนย่อยของตำแหน่งของชั้นวางขนม กรณี “ชั้นที่ 1”, “ชั้นที่ 2”, “ชั้นที่ 3”, และ “ชั้นที่ 4” ซึ่งได้แสดงไว้ในตารางที่ 3.2 ค่าเอนโทรปีของกรณีย่อยที่มีค่าเป็นคลาส “ซื้อ”, “ไม่ซื้อ”, และ “ไม่สามารถเลือกขนม

เพียง 1 ชิ้นได้” ของแอตทริบิวต์ “ตำแหน่งของชั้นวางขนม” มาเปรียบเทียบ โดยจะละเลยกรณีย่อยที่มีผลรวมของค่าเอนโทรปีเท่ากับ 0 หรือ 1 ซึ่งไม่มีกรณีย่อยถูกละเว้นเนื่องจากไม่มีกรณีใดผลรวมของค่าเอนโทรปีเท่ากับ 0 หรือ 1, และกรณีย่อยที่มีผลรวมของค่าเอนโทรปีน้อยที่สุดคือ “ชั้นที่ 3” ที่มีผลรวมของค่าเอนโทรปีเท่ากับ 0.59

ตารางที่ 3.2 ตารางค่าเอนโทรปีของแต่ละกรณีย่อยของแอตทริบิวต์ “ตำแหน่งของชั้นวางขนม”

กรณีย่อย / คลาส	**“ชั้นที่ 1”**	**“ชั้นที่ 2”**	**“ชั้นที่ 3”**	**“ชั้นที่ 4”**
“ซื้อ”	0.53	0.52	0.40	0.43
“ไม่ซื้อ”	0.40	0.40	0.19	0.14
“ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้”	0.52	0.52	0.00	0.43
ผลรวม	1.44	1.44	0.59	1.53

แอตทริบิวต์ถัดมาคือแอตทริบิวต์ “สีของถุงขนม” มี 35 อินสแตนซ์ที่สามารถแบ่งเป็นกรณีย่อยคือ “สีชมพู” มี 13 อินสแตนซ์, “สีเงิน” มี 10 อินสแตนซ์, และ “สีเหลือง” มี 12 อินสแตนซ์ เมื่อ “สีเสื้อ” มีค่าเป็น “สีส้ม” โดยที่แต่ละกรณีย่อยจะมีจำนวนของคลาสที่แตกต่างกัน

1) “สีชมพู” มี 13 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 8 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 3 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 2 กรณีย่อย

   Entropy (สีของถุงขนม | “สีชมพู”***) = - [(8/13 x $log_2$8/13) + (3/13 x $log_2$3/13) + (2/13 x $log_2$2/13)] = 1.07***

   P (สีของถุงขนม “สีชมพู”***) = 13/35 = 0.37***

2) “สีเงิน” มี 10 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 4 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 2 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 4 กรณีย่อย

   Entropy (สีของถุงขนม | “สีเงิน”***) = - [(4/10 x $log_2$4/10) + (2/10 x $log_2$2/10) + (4/10 x $log_2$4/10)] = 1.52***

*P (*สีของถุงขนม | “สีเงิน”*) = 10/35 = 0.29*

3) “สีเหลือง” มี 10 อินสแตนซ์ซึ่งจะมีกรณีที่เป็นคลาส “ซื้อ” ทั้งหมด 6 กรณีย่อย, กรณีที่เป็นคลาส “ไม่ซื้อ” ทั้งหมด 2 กรณีย่อย, และ กรณีที่เป็นคลาส “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ทั้งหมด 4 กรณีย่อย

*Entropy (*สีของถุงขนม | “สีเหลือง”*) = - [(6/12 x* log_2*6/12) + (2/12 x* log_2*2/12) + (4/12 x* log_2*4/12)] = 1.52*

*P (*สีของถุงขนม | “สีเหลือง”*) = 12/35 = 0.34*

นำอัตราส่วนของกรณีย่อยที่มีค่าเป็นคลาส “ซื้อ”, “ไม่ซื้อ”, และ “ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้” ของแอตทริบิวต์ “สีของถุงขนม” มาเปรียบเทียบ โดยจะละเลยกรณีย่อยที่มีย่อยที่มีผลรวมของค่าเอนโทรปีเป็น 0 หรือ 1 ซึ่งไม่มีกรณีย่อยถูกละเว้นเนื่องจากไม่มีกรณีใดผลรวมของค่าเอนโทรปีเท่ากับ 0 หรือ 1, และกรณีย่อยที่มีผลรวมของค่าเอนโทรปีน้อยที่สุดคือ “สีชมพู” ที่คลาส “ไม่ซื้อ” ซึ่งมีผลรวมของค่าเอนโทรปีเท่ากับ 1.07 ที่ได้แสดงไว้ในตารางที่ 3.3

ตารางที่ 3.3 ตารางค่าเอนโทรปีของแต่ละกรณีย่อยของแอตทริบิวต์ “สีของถุงขนม”

กรณีย่อย / คลาส	**“สีชมพู”**	**“สีเงิน”**	**“สีเหลือง”**
“ซื้อ”	0.35	0.53	0.48
“ไม่ซื้อ”	0.34	0.46	0.52
“ไม่สามารถเลือกขนมเพียง 1 ชิ้นได้”	0.40	0.53	0.39
ผลรวม	1.07	1.52	1.39

จากตัวอย่างของแอตทริบิวต์ “สีรองเท้า”, “ตำแหน่งของชั้นวางขนม”, และ “สีของถุงขนม” จะพบว่าทั้ง 3 แอตทริบิวต์มีค่าอินฟอร์เมชันเกนและค่ารีเมนเดอร์ที่เท่ากัน ทำให้ต้องมองต่อไปข้างหน้าในระดับของกรณีย่อย ซึ่งอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้น จะนำผลรวมของค่าเอนโทรปีของกรณีย่อยที่น้อยที่สุดของแอตทริบิวต์มาเข้าเงื่อนไขละเลยกรณีที่ผลรวมของค่าเอนโทรปีเท่ากับ 0 หรือ 1, และเปรียบเทียบกันโดยที่จะเลือกแอตทริบิวต์ที่มีค่าอัตราส่วนของกรณีย่อยที่น้อยที่สุดไปกำหนดให้เป็นโหนด ซึ่งแอตทริบิว “สีรองเท้า” มีกรณี

ย่อย “สีฟ้า” ที่มีผลรวมของค่าเอนโทรปีที่น้อยที่สุดเท่ากับ 1.57, แอตทริบิวต์ “ตำแหน่งของชั้นวางขนม” มีกรณีย่อย “ชั้นที่ 3” ที่มีผลรวมของค่าเอนโทรปีที่น้อยที่สุดเท่ากับ 0.59, และแอตทริบิวต์ “สีของถุงขนม” มีกรณีย่อย “สีชมพู” ที่มีผลรวมของค่าเอนโทรปีที่น้อยที่สุดเท่ากับ 1.07 ดั้งนั้นเมื่อนำผลรวมของค่าเอนโทรปีของกรณีย่อยของแอตทริบิวต์ทั้ง 3 แอตทริบิวต์มาเปรียบเทียบกันจะพบว่าแอตทริบิวต์ “ตำแหน่งของชั้นวางขนม” ที่กรณีย่อย “ชั้นที่ 3” มีผลรวมของค่าเอนโทรปีที่น้อยที่สุด จึงถูกเลือกให้เป็นโหนดถัดไปของแอตทริบิวต์ “สีเสื้อ”

บทที่ 4
การทดลองและผลการทดลอง

การพัฒนาอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ ผู้วิจัยได้นำเสนอการทดสอบการเปรียบเทียบความถูกต้องระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับอัลกอริทึม ID3 ที่ปรับปรุงแล้ว ที่ผู้วิจัยได้ทำการพัฒนา ซึ่งมีวิธีการทดสอบประสิทธิภาพของอัลกอริทึม ขั้นตอนการเตรียมข้อมูล รวมไปถึงการเตรียมสภาพแวดล้อมในการทำงาน ผลการทดลองการเปรียบเทียบอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ได้รับการปรับปรุงแล้ว โดยมีรายละเอียดแต่ละแบบดังต่อไปนี้

4.1 การเตรียมการทดลอง

4.1.1 ขั้นตอนการเตรียมการทดลอง

1. เตรียมชุดข้อมูลสำหรับทดลอง
2. สร้างสภาพแวดล้อมสำหรับทำการทดลอง – เนื่องจากการทดลองนี้มีเป้าหมายเพื่อลดความเร็วในการสร้างต้นไม้ตัดสินใจ จึงจำเป็นจะต้องกำหนดคุณลักษณะของอุปกรณ์ เพราะประสิทธิภาพของอุปกรณ์ที่ใช้ในการทำทดลองอาจจะส่งผลต่อเวลาในการสร้างต้นไม้ตัดสินใจ

   Device spec:

   CPU: Octa-core

   Memory: 8 gigabytes

   SSD: 50 gigabytes

   OS: Ubuntu

3. นำอัลกอริทึม ID3 ที่ได้รับการพัฒนามาทดลอง
4. เก็บผลการทดลอง

4.1.2 คัดเลือกชุดข้อมูลสำหรับทำการทดลอง

การคัดเลือกชุดข้อมูล จะแบ่งชุดข้อมูลออกเป็น 3 กลุ่ม เพื่อให้มีความหลากหลายของรูปแบบของข้อมูล, จำนวนอินสแตนซ์, จำนวนแอตทริบิวต์, ชนิดของอินสแตนซ์, และขนาดของข้อมูลที่แตกต่างกันโดยชุดข้อมูลทั้ง 6 ชุดข้อมูลตามตารางที่ 4.1, 4.2, และ 4.3 เป็นชุดข้อมูลที่ไม่เล็กจนเกินไปรวมถึงเป็นชุดข้อมูลที่มีกรณีของการมีแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงสุดมากกว่า 1 แอตทริบิวต์

1. ชุดข้อมูลขนาดใหญ่ (มีจำนวนอินสแตนซ์มากกว่า 5,000 อินสแตนซ์ขึ้นไป)

ตารางที่ 4.1 ชุดข้อมูลขนาดใหญ่ที่มีจำนวนอินสแตนซ์มากกว่า 5,000 อินสแตน์ขึ้นไป

ชุดข้อมูล	จำนวนอินสแตนซ์	จำนวนแอตทริบิวต์
Insurance	9,822	85
Nursery	12,960	8

2. ชุดข้อมูลขนาดกลาง (มีจำนวนอินสแตนซ์ตั้งแต่ 500 อินสแตนซ์ขึ้นไป แต่ไม่เกิน 5,000 อินสแตนซ์)

ตารางที่ 4.2 ชุดข้อมูลขนาดกลางที่มีจำนวนอินสแตนซ์ตั้งแต่ 500 อินสแตนซ์ขึ้นไป แต่ไม่เกิน 5,000 อินสแตนซ์

ชุดข้อมูล	จำนวนอินสแตนซ์	จำนวนแอตทริบิวต์
Diabetes Data	520	16
Tic Tac Toe	958	9

3. ชุดข้อมูลขนาดเล็ก (มีจำนวนอินสแตนซ์น้อยกว่า 500 อินสแตนซ์)

ตารางที่ 4.3 ชุดข้อมูลขนาดเล็กที่มีจำนวนอินสแตนซ์น้อยกว่า 500 อินสแตนซ์ขึ้นไป

ชุดข้อมูล	จำนวนอินสแตนซ์	จำนวนแอตทริบิวต์
Soybean	307	35
Divorce	170	54

4.1.2 เงื่อนไขในการทำการทดลอง

1. ชุดข้อมูลทั้งหมดถูกแบ่งเป็น 2 ส่วนได้แก่ Training set, และ Test set
2. ใช้ the K-fold cross-validation ที่ K = 10 เพื่อหลีกเลี่ยงปัญหา Overfitting
3. ทดลอง 200 ครั้ง โดยใช้ Training set, และ Test set เหมือนกันทั้ง ID3 แบบดั้งเดิม, และ อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกน

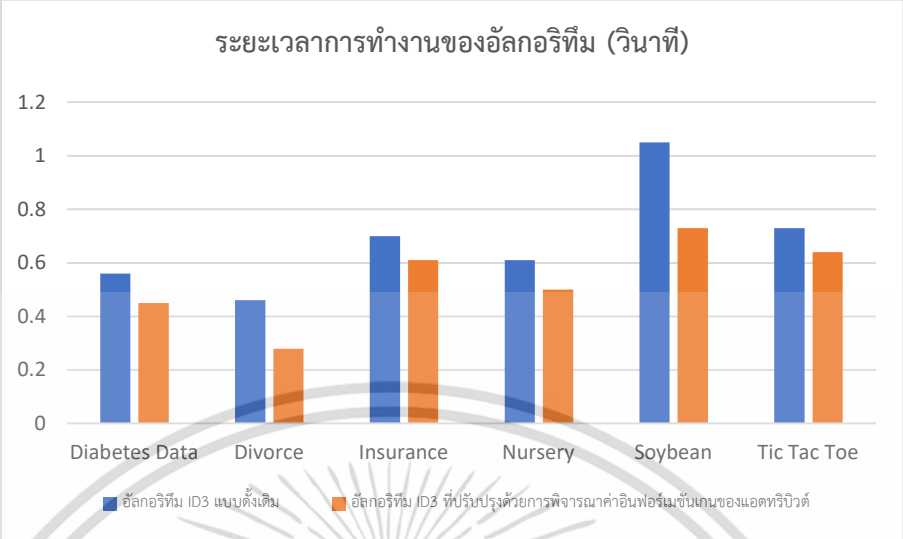
4.2 ผลการทดลอง

4.2.1. การทดลองเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิม

การทดลองทั้งอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงแล้ว การเปรียบเทียบเวลาทำงานระหว่างสองอัลกอริทึมแสดงไว้ในตารางที่ 4.4 เห็นได้ชัดว่าอัลกอริทึม ID3 ที่ปรับปรุงแล้วสามารถลดเวลาการทำงานลงได้มากกว่า 10 เปอร์เซ็นต์ ทำซ้ำการทดสอบ 200 ครั้งและรายงานค่าเฉลี่ยของเวลาทำงานและความถูกต้อง อัลกอริทึมที่ได้รับการปรับปรุงจะทำงานเร็วกว่าอัลกอริทึม ID3 แบบเดิม

ตารางที่ 4.4 การวัดเวลาทำงานโดยเปรียบเทียบระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

ชุดข้อมูล	การวัดผล		
	เวลาการทำงาน (วินาที)		
	อัลกอริทึม ID3 แบบดั้งเดิม	อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์	เปอร์เซ็นต์การปรับปรุง
Diabetes Data	0.56	0.45	19.64
Divorce	0.46	0.28	39.13
Insurance	0.70	0.61	12.86
Nursery	0.61	0.50	18.03
Soybean	1.05	0.73	30.48
Tic Tac Toe	0.73	0.64	12.33

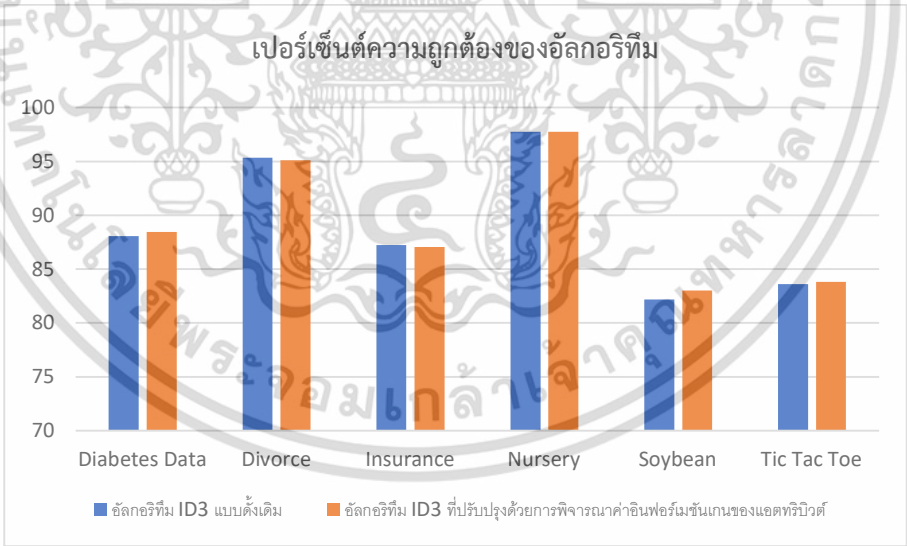


รูปที่ 4.1 กราฟเปรียบเทียบระยะเวลาการทำงานระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับ อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

จากตารางที่ 4.4, และรูปที่ 4.1 แสดงการเปรียบเทียบเวลาทำงานระหว่างอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ซึ่งจะสังเกตได้ว่าเวลาการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้นน้อยกว่าเวลาการทำงานของอัลกอริทึม ID3 แบบเดิมอย่างสังเกตุได้สำหรับชุดข้อมูลโดยที่เวลาการทำงานของอัลกอริทึมลดลงไปมากกว่า 10 เปอร์เซ็นต์

ตารางที่ 4.5 การวัดความถูกต้องโดยเปรียบเทียบระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

ชุดข้อมูล	การวัดผล		
	เปอร์เซ็นต์ความถูกต้อง		
	อัลกอริทึม ID3 แบบดั้งเดิม	อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์	เปอร์เซ็นต์การปรับปรุง
Diabetes Data	88.07	88.45	0.43
Divorce	95.32	95.10	0.41
Insurance	87.22	87.05	-0.20
Nursery	97.73	97.71	-0.03
Soybean	82.17	83.01	1.02
Tic Tac Toe	83.59	83.83	0.24



รูปที่ 4.2 กราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้องระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

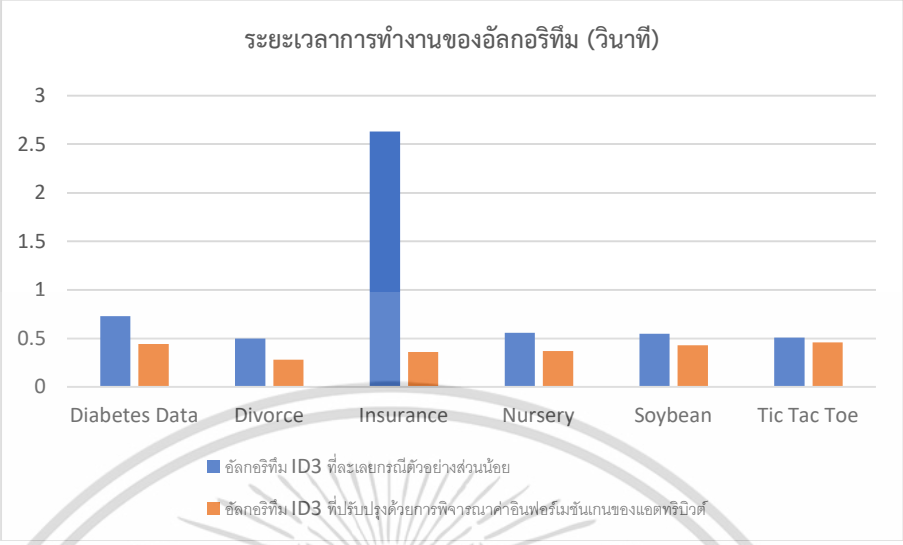
จากตารางที่ 4.5, และรูปที่ 4.2 แสดงการเปรียบเทียบความถูกต้องระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับอัลกอริทึม ID3 ที่ปรับปรุงแล้ว ความถูกต้องในการปรับปรุงอัลกอริทึม ID3 ไม่ลดลงหรือเพิ่มขึ้นเกิน 1 เปอร์เซ็นต์ เมื่อนำมาเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิม ดังนั้นอัลกอริทึม ID3 ที่ปรับปรุงแล้วไม่ส่งผลต่อความถูกต้อง

4.2.2. การทดลองเปรียบเทียบกับงานวิจัยการปรับปรุงอัลกอริทึม ID3 โดยการละเลยกรณีตัวอย่างส่วนน้อย (Improving ID3 Algorithm by Ignoring Minor Instances)

การทดลองทั้งอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อยและอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ การเปรียบเทียบเวลาทำงานระหว่างสองอัลกอริทึมแสดงไว้ในตารางที่ 4.6 เห็นได้ชัดว่าอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ แล้วสามารถลดเวลาการทำงานลงได้มากกว่า 15 เปอร์เซ็นต์ ทำซ้ำการทดสอบ 200 ครั้งและรายงานค่าเฉลี่ยของเวลาทำงานและความถูกต้อง อัลกอริทึมที่ได้รับการปรับปรุงจะทำงานเร็วกว่าอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อย

ตารางที่ 4.6 การวัดเวลาทำงานโดยเปรียบเทียบระหว่างอัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อยกับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

ชุดข้อมูล	การวัดผล เวลาการทำงาน (วินาที) อัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อย	อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์	เปอร์เซ็นต์การปรับปรุง
Diabetes Data	0.73	0.44	39.05
Divorce	0.50	0.28	43.01
Insurance	2.63	0.36	86.19
Nursery	0.56	0.37	34.18
Soybean	0.55	0.43	21.98
Tic Tac Toe	0.51	0.46	15.26

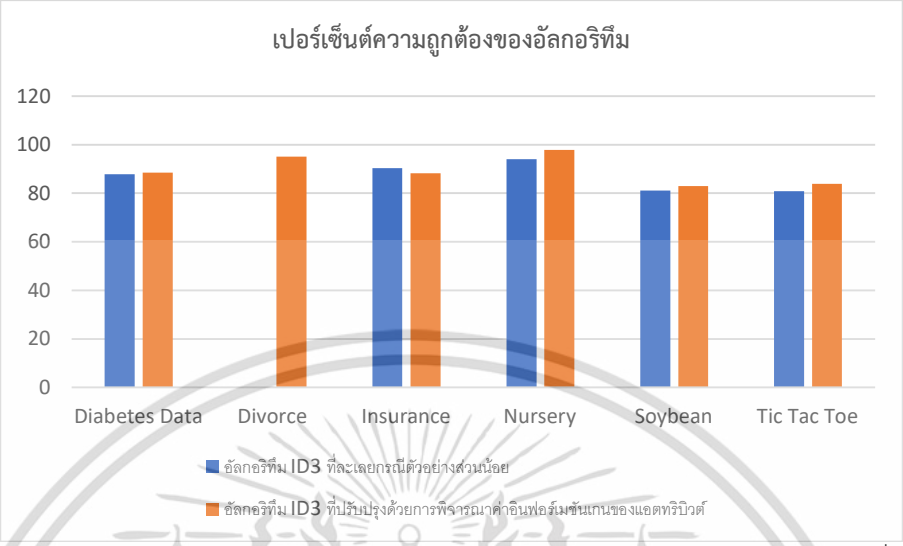


รูปที่ 4.3 กราฟเปรียบเทียบระยะเวลาการทำงานระหว่างอัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อย กับ อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

จากตารางที่ 4.6, และรูปที่ 4.3 แสดงการเปรียบเทียบเวลาทำงานระหว่างอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อยและอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ นั้นน้อยกว่าเวลาการทำงานของอัลกอริทึม ID3 ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อยสำหรับชุดข้อมูล เวลาการทำงานของอัลกอริทึมลดลงไปมากกว่า 15 เปอร์เซ็นต์

ตารางที่ 4.7 การวัดความถูกต้องโดยเปรียบเทียบระหว่างอัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อยกับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

ชุดข้อมูล	การวัดผล		
	เปอร์เซ็นต์ความถูกต้อง		
	อัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อย	อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์	เปอร์เซ็นต์การปรับปรุง
Diabetes Data	87.87	88.45	0.66
Divorce	(ไม่สามารถหาค่าได้)	95.19	(ไม่สามารถหาค่าได้)
Insurance	90.27	88.19	-2.30
Nursery	94.07	97.85	3.03
Soybean	81.19	83.01	2.24
Tic Tac Toe	80.92	83.83	3.60



รูปที่ 4.4 กราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของอัลกอริทึมระหว่างอัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อยกับ อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

จากตารางที่ 4.7, และรูปที่ 4.4 แสดงเปอร์เซ็นต์ความถูกต้องระหว่างอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อยและอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ ความถูกต้องในการปรับปรุงอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ไม่ลดลงหรือเพิ่มขึ้นเกิน 3 เปอร์เซ็นต์ เมื่อนำมาเปรียบเทียบกับอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อย ดังนั้นอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ไม่ส่งผลต่อความถูกต้อง

4.3 สรุปผลการทดลอง, และข้อจำกัด

4.3.1 สรุปผลการทดลอง, และข้อจำกัด

จากการทำการทดลองเปรียบเทียบระยะเวลาการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์กับอัลกอริทึม ID3 แบบดั้งเดิม, และอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อย ตามตารางที่ 4.4, และตารางที่ 4.6 นั้น พบว่าเวลาการทำงานของอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้นลดลงจากระยะเวลาการทำงานของอัลกอริทึม ID3 แบบดั้งเดิม, และอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อย

สาเหตุเนื่องมาจากอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้นจะไม่พยายามที่จะหาแอตทริบิวต์ที่ดีที่สุดต่อหากพบว่าค่าอินฟอร์เมชันเกนมีค่าเท่ากับ 0

ถึงแม้ว่าจะยังคงเหลือแอตทริบิวต์มากกว่า 1 ตัวก็ตาม จึงเป็นการลดระยะเวลาการทำงานของอัลกอริทึมในส่วนของแอตทริบิวต์ที่ยังเหลืออยู่ลงไป ซึ่งหากว่าเป็นอัลกอริทึม ID3 แบบดั้งเดิม, และอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อยนั้น ต่อให้แอตทริบิวต์เหล่านั้นจะมีค่าเท่ากับ 0 แล้วก็ยังคงค้นหาแอตทริบิวต์ที่ดีที่สุดมากำหนดให้เป็นโหนดจนกว่าครบทุกแอตทริบิวต์ของชุดข้อมูล

ข้อจำกัดของอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

1) หากชุดข้อมูลที่เป็นข้อมูลเชิงปริมาณที่มีความต่อเนื่อง (Continuous Value) ปะปนมาในชุดข้อมูล จะส่งผลกระทบต่อค่าความถูกต้อง
2) หากชุดข้อมูลมีขนาดเล็ก, และไม่มีกรณีที่มีแอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงสุดมากกว่า 1 แอตทริบิวต์ อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ ก็จะไม่สามารถลดระยะเวลาในการทำงานได้

ตัวอย่างเช่นชุดข้อมูล Diagnosis ที่มีจำนวนแอตทริบิวต์ 7 แอตทริบิวต์, และมีจำนวนของอินสแตนซ์ทั้งหมด 106 อินสแตนซ์ นอกจากนั้นยังมีกรณีย่อยของแอตทริบิวต์ที่มีค่าเชิงปริมาณที่มีความต่อเนื่อง (Continuous Value) ที่แสดงไว้ในตารางที่ 4.8, และเมื่อนำระยะเวลาการทำงานของอัลกอริทึมและค่าความถูกต้องมาเปรียบเทียบกับอัลกอริทึม ID3 แบบดั้งเดิมจะพบว่า อัลกอริทึม ID3 ที่ได้รับการปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ใช้ระยะเวลาการทำงานมากกว่าาอัลกอริทึม ID3 แบบดั้งเดิมดังที่แสดงไว้ในตารางที่ 4.9, และได้ความถูกต้องน้อยกว่าอัลกอริทึม ID3 แบบดั้งเดิมดังที่แสดงไว้ในตารางที่ 4.10

ตารางที่ 4.8 ตัวอย่างชุดข้อมูล Diagnosis

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Class
35.5	no	yes	no	no	no	no	no
35.9	no	no	yes	yes	yes	yes	no
35.9	no	yes	no	no	no	no	no
36	no	no	yes	yes	yes	yes	no
36	no	yes	no	no	no	no	no
36	no	yes	no	no	no	no	no
36.2	no	no	yes	yes	yes	yes	no
36.2	no	yes	no	no	no	no	no
36.3	no	no	yes	yes	yes	yes	no
36.6	no	no	yes	yes	yes	yes	no
36.6	no	no	yes	yes	yes	yes	no
36.6	no	yes	no	no	no	no	no
36.6	no	yes	no	no	no	no	no
36.7	no	no	yes	yes	yes	yes	no
36.7	no	yes	no	no	no	no	no
36.7	no	yes	no	no	no	no	no
36.8	no	no	yes	yes	yes	yes	no
36.8	no	no	yes	yes	yes	yes	no
36.9	no	no	yes	yes	yes	yes	no
36.9	no	yes	no	no	no	no	no
37	no	no	yes	yes	no	yes	no
37	no	no	yes	yes	no	yes	no
37	no	yes	no	no	no	no	no
37	no	no	yes	yes	yes	yes	no

ตารางที่ 4.9 การวัดเวลาทำงานโดยเปรียบเทียบระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

ชุดข้อมูล	การวัดผล		
	เวลาการทำงาน (วินาที)		
	อัลกอริทึม ID3 แบบดั้งเดิม	อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์	เปอร์เซ็นต์การปรับปรุง
Diagnosis	0.03	0.04	-33.05

ตารางที่ 4.10 การวัดความถูกต้องโดยเปรียบเทียบระหว่างอัลกอริทึม ID3 แบบดั้งเดิมกับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์

ชุดข้อมูล	การวัดผล		
	เปอร์เซ็นต์ความถูกต้อง		
	อัลกอริทึม ID3 แบบดั้งเดิม	อัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์	เปอร์เซ็นต์การปรับปรุง
Diagnosis	70.05	62.68	-10.52

ในทางเดียวกันนั้นจากผลการทดลองในตารางที่ 4.7 เมื่อเปรียบเทียบความถูกต้องระหว่างอัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อยกับการปรับปรุงเวลาการประมวลผลของอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ จะพบว่ามี 1 ชุดข้อมูลคือ Divorce ที่เมื่อนำมาใช้กับอัลกอริทึม ID3 ที่ได้รับการพัฒนาให้ละเลยกรณีตัวอย่างส่วนน้อยแล้วพบว่าไม่สามารถหาค่าได้นั้น พบว่ามีข้อสังเกตดังนี้

- ชุดข้อมูลจะแทนที่กรณีย่อยที่หายไปด้วยสัญลักษณ์ “?”

- ชุดข้อมูลมีชนิดของข้อมูลเป็นทศนิยม ผสมอยู่กับข้อมูลชนิดข้อความ ซึ่งคาดว่าชนิดของข้อมูลที่ต่างกันผสมปนอยู่ในแอตทริบิวต์เดียวกัน จะทำให้เกิดเหตุการที่ไม่สามารถหาค่าของชุดข้อมูลนี้ได้

บทที่ 5
สรุปผลการวิจัย, และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อการพัฒนาอัลกอริทึม ID3 ด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ เพื่อการปรับปรุงพัฒนาอัลกอริทึม ID3, และงานวิจัยนี้จะเน้นไปที่การลดเวลาในการสร้างต้นไม้การตัดสินใจ มีการอธิบายอัลกอริทึมและอธิบายถึงปัญหา รวมไปถึงการเตรียมสภาพแวดล้อมในการทำงานสามรถสรุปผลได้ดังนี้

5.1 สรุปผลการวิจัย

ผลการทดลองทั้งอัลกอริทึม ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ การเปรียบเทียบเวลาทำงานระหว่างสองอัลกอริทึมแสดงไว้ในตารางที่ 4.4, และผลการทดลองทั้งอัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อยและอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ การเปรียบเทียบเวลาทำงานระหว่างสองอัลกอริทึมแสดงไว้ในตารางที่ 4.6 จะสังเกตได้ว่าอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์สามารถลดเวลาการทำงานลงได้มากกว่า 10 เปอร์เซ็นต์ ทำซ้ำการทดสอบ 200 ครั้งและรายงานค่าเฉลี่ยของเวลาทำงานและความถูกต้อง อัลกอริทึมที่ได้รับการปรับปรุงจะทำงานเร็วกว่าอัลกอริทึม ID3 แบบเดิมและ อัลกอริทึม ID3 ที่ละเลยกรณีตัวอย่างส่วนน้อย

วิธีการปรับปรุงอัลกอริทึม ID3 แบบดั้งเดิมด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ ผลการทดลองยังระบุด้วยว่าวิธีการที่นำเสนอสามารถลดเวลาการทำงานลงได้มากกว่า 10 เปอร์เซ็นต์ ในขณะที่ความถูกต้องของการจัดหมวดหมู่แทบจะไม่ได้รับผลกระทบ สำหรับงานในอนาคตสามารถวางแผนที่จะปรับปรุงอัลกอริทึม ID3 โดยการใช้อัตราส่วนของกรณีย่อยที่มีค่าสูงที่สุดมาเป็นเงื่อนไขในการหาแอตทริบิวต์ที่ดีที่สุด

ข้อจำกัดของอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์นั้นจำเป็นจะต้องใช้กับชุดข้อมูลที่มีขนาดใหญ่, และมีความหลากหลายของคลาสซึ่งจะทำให้เกิดกรณีที่แอตทริบิวต์ที่มีค่าอินฟอร์เมชันเกนสูงสุดมากกว่า 1 แอตทริบิวต์, และจะต้องไม่เป็นชุดข้อมูลที่เป็นข้อมูลเชิงปริมาณที่มีความต่อเนื่อง (Continuous Value) โดยข้อจำกัดนี้อาจจะส่งผลกระทบถึงระยะเวลาในการทำงานของอัลกอริทึม, และความถูกต้องของอัลกอริทึม

5.2 ข้อเสนอแนะ

ทดลองทำการทดลองทั้ง ID3 แบบดั้งเดิมและอัลกอริทึม ID3 ที่ปรับปรุงด้วยการพิจารณาค่าอินฟอร์เมชันเกนของแอตทริบิวต์ สามารถเพิ่มชุดข้อมูลเรียนรู้ให้มีความหลากหลาย สามารถเพิ่มจำนวนรอบของการทำซ้ำ รวมถึงสามารถปรับเปลี่ยนค่า K ของ K-fold cross-validation ให้เป็นค่าอื่น ๆ นอกเหนือจากค่า K ที่เท่ากับ 10 เพื่อศึกษาเพิ่มเติมว่าจะสามารถใช้วิธีการอื่น ๆ ที่สามารถเพิ่มความถูกต้องของอัลกอริทึม, และการตัดสินใจที่ดีมีประสิทธิภาพ โดยที่ไม่ส่งผลกระทบต่อเวลาการทำงานของอัลกอริทึม หรืออาจจะลดเวลาการทำงานของอัลกอริทึมลง

บรรณานุกรม

กนิษฐา อินธิชิต, และคณะ. (2566). การพัฒนาแอปพลิเคชันช่วยตัดสินใจในการเลือกเรียนสาขาวิชาคอมพิวเตอร์ ในมหาวิทยาลัยราชภัฏศรีสะเกษบนระบบปฏิบัติการแอนดรอยด์ โดยใช้เทคนิคต้นไม้ตัดสินใจ. วารสารวิชาการการจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม, 9(2), 97-107.

ขจรศักดิ์ ศรีอ่อน. (2552). การทำนายสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้อง โดยใช้เทคนิคการทำเหมืองข้อมูลในระบบจำหน่ายของการไฟฟ้าส่วนภูมิภาค เขต 1 ภาคกลาง. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมไฟฟ้า) สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์.

ณัฏฐพันธ์ เขจรนันทน์, และไพบูลย์ เกียรติโกมล. (2542). ระบบสารสนเทศเพื่อการจัดการ. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย.

ณัฐกิจ เจนการ. (2563). การพัฒนาแบบจำลองในการตรวจจับข้อความภาษาไทยที่เป็นการกลั่นแกล้งทางไซเบอร์ โดยใช้วิธีซัพพอร์ตเวกเตอร์แมชชีน.วารสารวิทยาศาสตร์และเทคโนโลยี, และนวัตกรรม, 1(1), 24-34.

นฤพนต์ ว่องประชานุกูล. (2548). วิธีที่เหมาะสมสำหรับการตัดกิ่งต้นไม้ตัดสินใจของการทำเหมืองข้อมูลทางด้านวิทยาศาสตร์. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

พยูน พาณิชย์กุล. (2548). การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Decision Tree. โครงการพัฒนาระบบงานปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ แขนงวิทยาการสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง.

ศุภชัย ประคองศิลป์. (2551). การออกแบบและพัฒนาระบบสนับสนุนการตัดสินใจในการอนุมัติลูกบ้านเข้าโครงการโดยใช้เทคนิคต้นไม้ตัดสินใจ กรณีศึกษา มูลนิธิที่อยู่อาศัยเพื่อมนุษยชาติ. ปัญหาพิเศษปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศคณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

สัณห์ชัย หยีวิยม. (2564). การพัฒนาระบบตรวจสอบความถูกต้องของข้อมูลเพื่อให้บริการผ่านเว็บไซต์สำหรับการสนับสนุนการพัฒนาพื้นที่เป้าหมายชายแดนภาคเหนือ 4 จังหวัด. วารสารวิชาการเพื่อการพัฒนานวัตกรรมเชิงพื้นที่, 2 (3), 17-27.

อนันต์ ปินะเต. (2563). การวิเคราะห์สารสนเทศเพื่อการพัฒนาระบบสนับสนุนการวางแผนการคัดเลือกบุคคลเข้าศึกษาในระบบ TCAS มหาวิทยาลัยมหาสารคาม. Journal of Science and Technology Mahasarakham University, 39(1), 78-89.

Chen Jin, Luo De-lin, Mu Fen Xiang, “An Improved ID3 Decision Tree Algorithm,” in 2009 4th International Conference on Computer Science & Education, 2009, pp.127-130

S. Kraidesh, K. Jearanaitanakij, “Improving ID3 Algorithm by Combining Values from Equally Important Attributes,” 2017 21st International Computer Science and Engineering Conference (ICSEC), Bangkok, Thailand, pp.102-105

Z. Wang, Yu Lium Lu Liu, “A new way to choose splitting attribute in ID3 algorithm,” in 2017 IEEE 2nd Information Technology, Network, Electronic and Automation Control Conference (ITNEC), Chengdu, China, pp.659-636

N. Kaewrod, K. Jearanaitanakij, “Improving ID3 Algorithm by Ignoring Minor Instance”, in 2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, pp. 1-5

M. Kahn. UCI Machine Learning Repository[online]. Available: http://archive ics.uci.edu/ml/datasets.html

Guillermo Arria-Devoe. “Building a ID3 Decision Tree Classifier with Python” [online] Available: https://guillermoarriadevoe.com/blog/building-a-id3 -decision-tree-classifier-with-python/

Yaser Sakkaf. “Decision Trees: ID3 Algorithm Explained” [online] Available: https://bit.ly/3wPWKh5

Bernardo Garcia del Rio. “ID3 Decision Tree Classifier from scratch in Python” [online] Available: https://towardsdatascience.com/id3-decision-tree-classifier-from-scratch-in-python-b38ef145fd

Brain Ambielli. “Information Entropy and Information Gain” [online] Available: https://bambielli.com/til/2017-10-22-information-gain/

ประวัติผู้เขียน

ชื่อ-นามสกุล นางสาวภิญญรัตน์ ชื่นประเสริฐสุข
วัน เดือน ปีเกิด 20 ตุลาคม 2537 ที่ชลบุรี
ที่อยู่ 98/15 หมู่ที่ 1 ตำบลเสม็ด อำเภอเมืองชลบุรี
จังหวัดชลบุรี 20000 โทร 098-246-8326

ประวัติการศึกษา

2560 วิศวกรรมศาสตรบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ความชำนาญเฉพาะด้าน

1) การพัฒนาโปรแกรมด้วยภาษา Python
2) การวิเคราะห์ข้อมูล (Data Analysis)

ประสบการณ์การทำงานและผลงานวิจัย

มิถุนายน พ.ศ. 2559 - กรกฎาคม พ.ศ. 2559 IT Support
(ฝึกงานระยะเวลา 2 เดือน) ที่บริษัท Greenline Synergy
บริษัท กรีนไลน์ ซินเนอร์ยี่ จำกัด เป็นบริษัทในเครือของ BDMS (กรุงเทพดุสิตเวชการ)
ซึ่งก่อตั้งขึ้นในปี พ.ศ. 2551 เพื่อทำหน้าที่เป็นศูนย์กลางด้านเทคโนโลยีสารสนเทศ
เพื่อให้ BDMS มีสภาพแวดล้อมด้านไอทีที่ปลอดภัย ยืดหยุ่น แข็งแกร่ง, และคุ้มค่าผ่าน
มาตรฐานบริการระดับมืออาชีพด้านไอทีและโซลูชันนวัตกรรมด้านไอทีด้านการดูแล
สุขภาพ เพื่อให้เกิดการเปลี่ยนแปลงทางดิจิทัลและความสามารถในการทำงานร่วมกัน
สำหรับการส่งมอบการดูแลที่เชื่อมต่อกัน

ผลงานที่ได้รับการตีพิมพ์

1. P. Chuenprasertsuk, P. Lertpunyavuttikul, and S. Glomglome (2017) "Usage-based Insurance Using IoT Platform", 2017 21st International Computer Science and Engineering Conference (ICSEC), 2017, pp. 286-289, 978-1-5386-0787-9/17/$31.00 ©2017 IEEE.

2. P. Chuenprasertsuk, and K. Jearanaitanakij (2022) "Improving the ID3 Algorithm By Filtering Out Attributes With Values Of 0 or 1", 2022 6th International Conference on Information Technology (InCIT), 2022, pp. 173-176, DOI: 10.1109/InCIT56086.2022.10067614