บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

แนวคิด ทฤษฎี เอกสารและงานวิจัย ที่เกี่ยวข้องกับการศึกษาในครั้งนี้ ผู้ศึกษาได้แบ่งออกเป็นหัวข้อต่างๆ ดังนี้

2.1 การตัดสินใจการให้สินเชื่อ

2.2 การทำเหมืองข้อมูล

2.3 เทคนิคการสุ่มตัวอย่างแบบมีระบบ

2.4 กฎการจำแนก

2.5 ต้นไม้ตัดสินใจ

2.6 ค่าอินฟอร์เมชันเกน

2.1 การตัดสินใจการให้สินเชื่อ

Yang Liu (2007) นำเสนอแนวคิดว่า การพิจารณาให้สินเชื่อโดยทั่วไปจะมี 2 วิธีคือ

1) Deductive Credit Scoring System เป็นการพิจารณาให้สินเชื่อโดยใช้ผู้เชี่ยวชาญ ที่ชำนาญในการพิจารณาสินเชื่อซึ่งจะพิจารณาโดยใช้หลักการทั่วไป โดยให้ค่าน้ำหนักในแต่ละคุณลักษณะของผู้กู้ ค่าน้ำหนักนั้นเกิดจากการประเมินโดยผู้เชี่ยวชาญและประสบการณ์ที่มีอยู่ คะแนนที่ได้จะถูกรวบรวมประมวลผลออกมาเพื่อใช้เป็นเกณฑ์ในการพิจารณาสมควรอนุมัติให้สินเชื่อหรือไม่

2) Empirical Credit Scoring System เป็นการพิจารณาให้สินเชื่อโดยใช้เกณฑ์การวัดเชิงปริมาณของประสิทธิภาพและคุณลักษณะของการปล่อยสินเชื่อที่ผ่านมา เพื่อจะพยากรณ์ประสิทธิภาพการปล่อยสินเชื่อด้วยคุณลักษณะที่คล้ายคลึงกัน โดยอาศัยฐานข้อมูลเดิมเป็นสิ่งสำคัญในการหาเกณฑ์การวัดและประเมินความเสี่ยงและพิจารณาว่าสมควรอนุมัติให้สินเชื่อหรือไม่

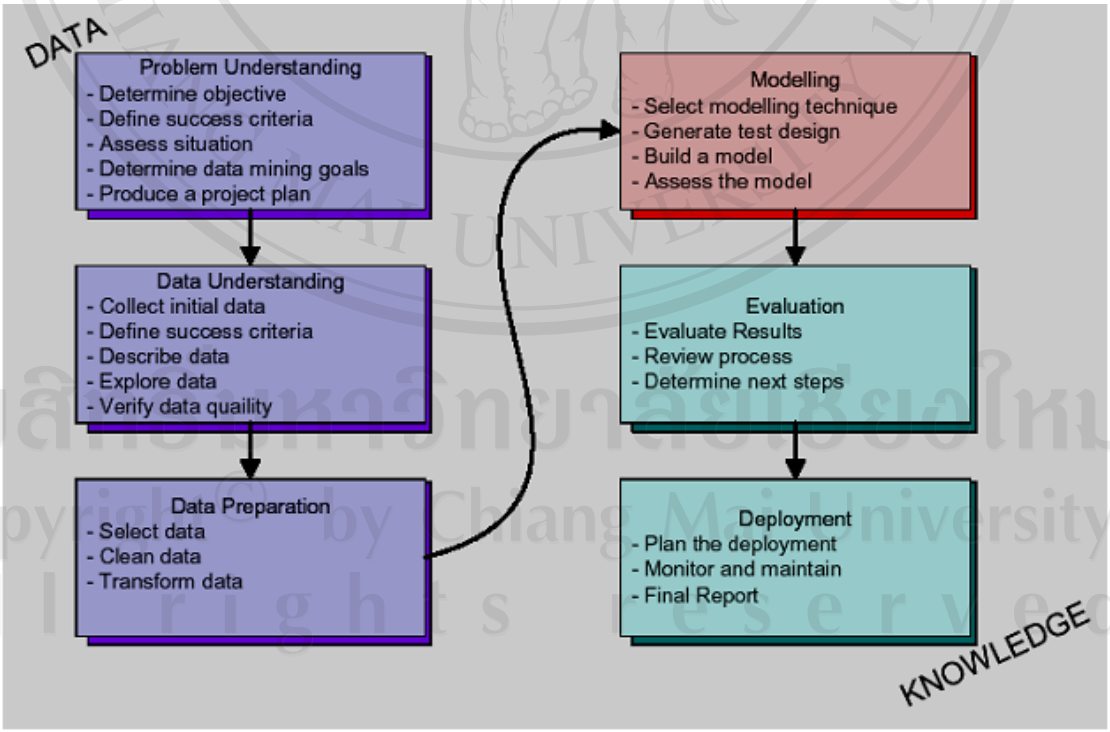
2.2 การทำเหมืองข้อมูล

การทำเหมืองข้อมูลคือกระบวนการค้นหาแนวโน้ม รูปแบบร่วมความสัมพันธ์ หรือความรู้ใหม่อื่นๆ จากข้อมูลจำนวนมาก การทำเหมืองข้อมูลเป็นเทคโนโลยีที่ช่วยให้การวิเคราะห์ข้อมูลทำได้โดยอัตโนมัติและมีประสิทธิภาพสูงขึ้นกว่าที่เคยเป็นมา จึงได้รับความสนใจนำไปใช้อย่างแพร่หลายในทุกวงการ โดยเฉพาะในกรณีที่ข้อมูลมีขนาดใหญ่มาก เช่นข้อมูลสำมะโน

ประชากร ข้อมูลที่ได้รับจากดาวเทียมสำรวจสภาพอากาศและพื้นผิวโลก ปัญหาของข้อมูลขนาดใหญ่เหล่านี้ คือ การวิเคราะห์ข้อมูลด้วยแรงงานของผู้เชี่ยวชาญเป็นสิ่งที่แทบจะเป็นไปไม่ได้ เพราะเป็นงานที่ใช้แรงงาน ทรัพยากรและเวลามาก แนวทางที่จะช่วยให้งานวิเคราะห์ข้อมูลทำได้รวดเร็วขึ้นและได้ผลการวิเคราะห์มาใช้ทันเวลา ก็คือทำให้กระบวนการวิเคราะห์ข้อมูลเป็นอัตโนมัติมากขึ้น ใช้แรงงานมนุษย์ให้น้อยที่สุด แนวคิดนี้จึงทำให้เกิดการทำเหมืองข้อมูล ซึ่งเป็นกระบวนการวิเคราะห์ข้อมูลโดยอัตโนมัติ โดยระบบคอมพิวเตอร์จะทำหน้าที่ค้นหาแนวโน้ม ลักษณะที่น่าสนใจต่างๆ ที่ปรากฏในข้อมูลส่วนใหญ่ และวิเคราะห์ความสัมพันธ์ระหว่างข้อมูล

การทำเหมืองข้อมูลนี้เปรียบเสมือนการทำเหมืองแร่ที่เราใช้เครื่องจักรคัดแยกแร่ที่เป็นที่ต้องการออกจากกองหิน กรวด ดินที่ปะปนมากับสายแร่ เพียงแต่การทำเหมืองข้อมูลนั้นสิ่งที่เราได้จากกองข้อมูลมหาศาล คือความรู้ที่ซ่อนอยู่ในกองข้อมูล ความรู้นี้จะช่วยให้เราเข้าใจลักษณะของข้อมูล และเข้าใจปัจจัยที่ทำให้เกิดลักษณะบางอย่างขึ้นในข้อมูลบางกลุ่ม ซึ่งจะช่วยให้เราสามารถทำนายแนวโน้มของข้อมูลใหม่ที่จะเกิดขึ้นในอนาคตได้ รวมถึงเข้าใจความสัมพันธ์ที่เชื่อมโยงข้อมูลแต่ละกลุ่มย่อยเข้าด้วยกัน

บุญเสริม กิจศิริกุล (2546) ได้กล่าวไว้ว่า โดยทั่วไปการทำเหมืองข้อมูลจะประกอบด้วย 6 ขั้นตอนหลัก ซึ่งแสดงได้ดังรูป 2.1



แหล่งที่มา : บุญเสริม กิจศิริกุล (2546)

รูป 2.1 ขั้นตอนการทำเหมืองข้อมูล

ขั้นตอนที่ 1 ทำความเข้าใจกับปัญหา (Problem Understanding) : ประกอบด้วยกระบวนการย่อย ดังนี้

- ตั้งเป้าหมายหลักว่าการทำเหมืองข้อมูลนี้เพื่อการแก้ปัญหาใด เช่น เพื่อช่วยในการตัดสินใจในการให้สินเชื่อ เพิ่มยอดขายสินค้า เป็นต้น
- ตั้งเกณฑ์วัดความสำเร็จในการทำเหมืองข้อมูล ซึ่งอาจเป็นได้ทั้งรูปธรรม เช่นเพิ่มยอดขายสินค้าขึ้นอีก 5% จากเดิม หรือนามธรรม เช่น สามารถค้นพบความรู้ใหม่ๆ จากฐานข้อมูลเดิม
- ตั้งเป้าหมายเชิงการทำเหมืองข้อมูล ซึ่งต่างจากเป้าหมายหลักในการแก้ไขปัญหา เช่น เป้าหมายหลักเพื่อช่วยในการตัดสินใจในการให้สินเชื่อ เป้าหมายเชิงการทำเหมืองข้อมูล คือ หาลักษณะของลูกค้าที่มีแนวโน้มที่จะเป็นหนี้เสีย เป็นต้น
- วางแผนการทำเหมืองข้อมูล การจัดเก็บข้อมูลอย่างไร ใช้ขั้นตอนวิธี (Algorithm) ใดในการทำเหมืองข้อมูล ซึ่งโดยทั่วไปขั้นตอนวิธีในการทำเหมืองข้อมูลสามารถแบ่งออกได้ 4 ประเภทคือ

<u>ประเภทที่ 1</u> ต้นแบบในการทำนาย (Predictive Modeling) เป็นต้นแบบที่ถูกสร้างขึ้นโดยอาศัยลักษณะคล้ายการเรียนรู้ของมนุษย์ ซึ่งต้นแบบนี้จะใช้ในการวิเคราะห์ข้อมูลที่มีอยู่เพื่อกำหนดคุณสมบัติที่สำคัญของข้อมูล ดังนั้นข้อมูลที่มีอยู่จะต้องเป็นข้อมูลที่สมบูรณ์ จึงจะทำให้ต้นแบบที่สร้างขึ้นนั้นสามารถให้คำทำนายที่ถูกต้องหรือใกล้เคียงได้ โดยเริ่มต้นจะต้องกำหนดคำตอบที่ถูกต้องพื้นฐานให้กับต้นแบบเสียก่อน จากนั้นจึงค่อยๆ สร้างการเชื่อมโยงหรือความสัมพันธ์ใหม่ๆ ซึ่งมีลักษณะคล้ายกับ If – Then

<u>ประเภทที่ 2</u> การแบ่งกลุ่มข้อมูลเป็นการแบ่งข้อมูลออกเป็นกลุ่มๆ โดยไม่รู้ล่วงหน้าว่ามีจำนวนกลุ่มเป็นเท่าใด ทั้งนี้การจัดกลุ่มดังกล่าวได้จากการพิจารณาคุณสมบัติของข้อมูล ถ้ารายการใดของข้อมูลมีลักษณะคล้ายคลึงเป็นกลุ่มเดียวกันก็จะรวมเข้าด้วยกันเพื่อง่ายต่อการวิเคราะห์ เช่น การแบ่งลูกค้าออกตามอายุ เพศ หรือรายได้ เป็นต้น

<u>ประเภทที่ 3</u> การวิเคราะห์ความเชื่อมโยง (Link Analysis) เป็นการวิเคราะห์เพื่อหาความเชื่อมโยง หรือความสัมพันธ์ระหว่างข้อมูล เพื่อที่จะได้ทราบว่าข้อมูลแต่ละรายการมีความสัมพันธ์กันหรือไม่อย่างไร เช่น การเกิดอุบัติเหตุของรถยนต์มี

ความสัมพันธ์กับสีของรถยนต์ หรือเพศของผู้ขับรถยนต์หรือไม่ และถ้ามีเป็นแบบใด อย่างไร

ประเภทที่ 4 การตรวจหาค่าเบี่ยงเบน (Deviation Detection) เป็นการวิเคราะห์เพื่อค้นหาค่าที่แตกต่างไปจากค่ามาตรฐานหรือค่าที่กำหนดไว้ ว่ามากน้อยเพียงใด โดยทั่วไปจะใช้วิธีทางสถิติ ซึ่งสามารถนำไปใช้ในการตรวจสอบลายเซ็นที่อาจมีการปลอมแปลง รวมทั้งการหาจุดบกพร่องของชิ้นส่วนในการผลิตสำหรับโรงงานต่างๆ

ขั้นตอนที่ 2 ทำความเข้าใจข้อมูล (Data Understanding) : ประกอบด้วยกระบวนการย่อยดังนี้

- เก็บรวบรวมข้อมูล
- กำหนดคุณสมบัติของข้อมูลที่เก็บมาได้
- สำรวจดูข้อมูลอย่างคร่าวๆ เช่น ค่าสถิติต่างๆ ของข้อมูล
- ตรวจสอบข้อมูลขั้นต้น เช่น ความสมบูรณ์และความถูกต้องของข้อมูล

ขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation) : ประกอบด้วยกระบวนการย่อย ดังนี้

- คัดเลือกข้อมูลที่จะนำมาใช้
- ปรับเปลี่ยนรูปแบบของข้อมูล เช่น การนำตารางสองตารางในฐานข้อมูล มาเชื่อมต่อกัน
- ทำความสะอาดข้อมูล เป็นกระบวนการเตรียมข้อมูลให้เหมาะสมที่สุด เพื่อนำไปใช้ในขั้นตอนต่อไป มีได้หลายวิธี เช่น
- แก้ไขค่าว่างของข้อมูลโดยใส่ค่า 0
- ปรับเปลี่ยนข้อมูลให้เหมาะสมในการตัดสินใจเช่น จัดเปลี่ยนค่าข้อมูลให้เป็นกลุ่มเดียวกัน ยกตัวอย่าง “Coke” และ “Pepsi” เปลี่ยนค่าให้เป็น “น้ำอัดลม” เป็นต้น
- คัดเลือกเฉพาะข้อมูลที่สนใจและเกี่ยวข้องกับการแก้ปัญหาที่ต้องการ เช่น ไม่นำเอาข้อมูล ชื่อ-นามสกุล ลูกค้ามาเกี่ยวข้องกับการตัดสินใจการให้สินเชื่อ
- คอลัมน์ที่ไม่มีนัยสำคัญในการจำแนก เช่น สัญชาติ ซึ่งมีค่าซ้ำกันทั้งหมดคือ “ไทย” หรือหมายเลขบัตรประจำตัวประชาชน ที่มีค่าไม่ซ้ำกันเลย คอลัมน์เหล่านี้ควรตัดทิ้งไป ไม่ควรนำมาใช้ เป็นต้น

ขั้นตอนที่ 4 สร้างต้นแบบ (Modeling) : ประกอบด้วยกระบวนการย่อย ดังนี้

- เลือกขั้นตอนวิธีที่เหมาะสมในการทำเหมืองข้อมูล
- กำหนดรูปแบบและผลทดสอบ
- สร้างต้นแบบตามขั้นตอนวิธีที่เลือกไว้

- ทดสอบต้นแบบที่ได้ ว่ามีความถูกต้อง น่าเชื่อถือเพียงใด

ขั้นตอนที่ 5 การประเมิน (Evaluation) : เป็นขั้นตอนสำหรับการประเมินผลที่ได้จากต้นแบบว่าเหมาะสมตรงกับเป้าหมายหลักที่ได้วางไว้หรือไม่ โดยนำต้นแบบไปใช้กับสถานการณ์จริง หรือจำลองสถานการณ์ขึ้นมาให้เสมือนจริง กรณีผลที่ได้ไม่เหมาะสมจะทำให้เรากลับไปพิจารณาขั้นตอนก่อนหน้าที่ที่ได้ทำมา ถูกต้องหรือไม่ ต้องแก้ไขในขั้นตอนใด

ขั้นตอนที่ 6 การนำไปใช้ (Deployment) : เป็นขั้นตอนนำต้นแบบที่ได้สร้างไว้ไปใช้งานจริงกับสถานการณ์จริง

2.3 เทคนิคการสุ่มตัวอย่างแบบมีระบบ (Systematic Sampling)

สุพจน์ บุญแรง (2551) กล่าวว่า เทคนิคการสุ่มตัวอย่างแบบมีระบบ เป็นการสุ่มหาจำนวนประชากร เมื่อทราบจำนวนประชากร (N) และจำนวนตัวอย่าง (n) แล้วจะทำให้สามารถคำนวณช่วงการสุ่ม (sampling interval) คือ N/n = k และสัดส่วนของการสุ่ม (sampling fraction) = n/N = 1/k คือทุกๆ k หน่วยจะเลือกมา 1 หน่วย

ตัวอย่าง จำนวนประชากร (N) = 50 ต้องการตัวอย่างจำนวน (n) = 10 ช่วงของการสุ่ม = 50/10 = 5 หมายถึงต้องสุ่มในทุกๆ 5 หน่วย หรือ สุ่มทุก 1 ใน 5 หน่วย ซึ่งเรียกว่า sampling fraction = 10/50 = 1/5 ดังนั้นในช่วง 5 หน่วยแรกก็จะสุ่มอย่างธรรมดาว่าจะเริ่มต้นที่ใด เมื่อเลือกได้หน่วยที่ i เรียกว่า random start หน่วยถัดไปจะเลือกคือ i+k, i+2k, +……

จากตัวอย่างถ้าสุ่มได้หน่วยที่ 2 เป็น random start ก็จะได้หน่วยศึกษาดังนี้ 2, 2+5, 2+10,……, 2+45 นั่นคือหน่วยศึกษาที่ 2, 7, 12,….., 47

จักรกฤษณ์ สำราญใจ (2551) กล่าวว่า ในกรณีที่ N/n ไม่เป็นเลขจำนวนเต็มพอดี ให้เลือกเลขจำนวนเต็ม k ที่มีค่าใกล้เคียง N/n ให้มากที่สุด แล้วจัดเรียงลำดับหน่วยสมาชิกของประชากรเสมือนหนึ่งว่าสมาชิกเหล่านั้นเรียงกันเป็นวงกลม นั่นคือหมายเลข 1 จะไปต่อท้ายหมายเลข N จากนั้นให้สุ่มตัวเลขระหว่างเลข 1 ถึงเลข N ขึ้นมาหนึ่งตัว สมมติได้เลข p หน่วยตัวอย่างก็จะประกอบด้วยสมาชิกของประชากรลำดับที่ p, p+k, p+2k, p+3k,… ไปเรื่อยๆ โดยไม่ต้องสนใจว่าจะข้ามลำดับที่สุดท้ายไปอย่างไร จนได้จำนวนตัวอย่างครบ n หน่วยตามต้องการ เช่น N=70 , n=12 จะได้ N/n = 70/12 = 5.83 เลขจำนวนเต็มที่ใกล้เคียง 5.83 มากที่สุดคือ 6 จึงให้ k=6 แล้วสุ่มตัวเลขตั้งแต่ 1 ถึง 70 ขึ้นมาหนึ่งตัว สมมติได้เลข 42 จะได้หน่วยตัวอย่างคือสมาชิกประชากรลำดับที่ 42, 48, 54, 60, 66, 2, 8, 14, 20, 26, 32, 38 เป็นต้น

2.4 กฎการจำแนก (Classification Rules)

กฎการจำแนกเป็นขั้นตอนหนึ่งของการทำเหมืองข้อมูล นั่นคือขั้นตอนที่ 4 สร้างต้นแบบ เป็นศาสตร์ด้านการเรียนรู้ของเครื่อง (Machine Learning) ที่ได้รับความสนใจเป็นอย่างสูง เนื่องจากมนุษย์สามารถทำความเข้าใจได้ง่าย เทคนิคที่ได้รับความนิยมได้แก่ เทคนิคการสร้างต้นไม้ตัดสินใจโครงข่ายประสาทเทียม (Neural Network) เป็นต้น

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ, และ ธนาวินท์ รักธรรมานนท์ (2544) กล่าวว่า เป็นกระบวนการสร้างต้นแบบจัดการข้อมูล โดยนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ (Training Data) เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ ผลลัพธ์ที่ได้จากการเรียนรู้คือต้นแบบจำแนกข้อมูล (Classifier Model) และจะนำข้อมูลส่วนที่เหลือจากการสอนระบบมาเป็นข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบจะถูกนำมาเปรียบเทียบกับกับกลุ่มข้อมูลที่หามาได้จากต้นแบบเพื่อทดสอบความถูกต้องและปรับปรุงต้นแบบจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลเข้ามาใหม่เราจะนำมาผ่านต้นแบบ โดยต้นแบบจะสามารถทำนายกลุ่มข้อมูลของข้อมูลนี้ได้

กฎการจำแนกเป็นกระบวนการในการจัดแบ่งข้อมูล โดยการวิเคราะห์กลุ่มข้อมูล (Data object) ที่ยังไม่ได้จัดแบ่งประเภท วิเคราะห์เพื่อสร้างรูปแบบของข้อมูลออกมาเป็นชุดของข้อมูลซึ่งลักษณะของชุดข้อมูลนี้ถูกอธิบายได้โดยกลุ่มของคุณลักษณะ (Attribute) และกลุ่มของข้อมูลที่ใช้เรียนรู้ (Training Data Set) ผลลัพธ์ที่ได้ก็คือกฎการจำแนกและต้นแบบเพื่อการทำนาย (Prediction Model) เพื่อใช้ทำนายผลที่จะเกิดขึ้นสำหรับข้อมูลใหม่ โดยนำข้อมูลใหม่ที่ได้รับมาทำการเปรียบเทียบกับต้นแบบกฎการจำแนก วิเคราะห์และตัดสินใจเพื่อหาความเป็นไปได้ของข้อมูลนั้นๆ

การสร้างกฎการจำแนกจากฐานข้อมูล สามารถแบ่งเป็น 2 วิธีการใหญ่ๆ คือ

แบบที่หนึ่ง กฎการจำแนกแบบวัตถุประสงค์ (Objective Classification Rules) เป็นการสร้างกฎโดยอาศัยข้อมูลเดิมเป็นหลัก ดังนั้นถ้าข้อมูลเดิมมีความผิดพลาด อาจทำให้กฎที่ได้มีความคลาดเคลื่อนจากความเป็นจริง และในบางครั้งไม่ตรงกับความต้องการของผู้ใช้

แบบที่สอง กฎการจำแนกแบบอัตวิสัย (Subjective Classification Rules) เป็นการแก้ปัญหาของวิธีการแบบที่หนึ่ง คือเป็นการสร้างกฎโดยมีผู้เชี่ยวชาญเข้ามามีส่วนร่วมในกระบวนการสร้างกฎ ในการสร้างกฎขั้นแรกจะให้ผู้เชี่ยวชาญเป็นผู้กำหนดรูปแบบที่คิดว่าเป็นจริงและสนใจ กฎที่ได้รับจะเป็นเงื่อนไขในการกลั่นกรองข้อมูล และจำแนกกฎต่อไป ผลลัพธ์ที่ได้จะตรงกับความต้องการของผู้ใช้ และใกล้เคียงความจริงมากกว่า แต่ในบางครั้งหากใช้ความรู้ของ

ผู้เชี่ยวชาญอาจทำให้กฎที่ได้รับไม่ครอบคลุมข้อมูลที่นำมาใช้เรียนรู้ เนื่องจากมีความลำเอียง (Bias) ตามความเชื่อของผู้เชี่ยวชาญในการสร้างกฎ

2.5 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเป็นเทคนิคหนึ่งในกฎการจำแนก โดยมีลักษณะเหมือนโครงสร้างต้นไม้ โดยที่แต่ละโหนดแสดงคุณลักษณะที่ใช้ทดสอบข้อมูล แต่ละกิ่งแสดงผลในการทดสอบ และลีฟโหนด (Leaf Node) แสดงกลุ่มหรือคลาสที่กำหนดไว้ ขั้นตอนวิธีพื้นฐานของการสร้างต้นไม้ตัดสินใจคือ การค้นหาของที่ดีที่สุดเชิงละโมบ (Greedy Approach) โดยจะสร้างต้นไม้จากบนลงล่างแบบวนซ้ำ (Recursive) รูปแบบของต้นไม้จะประกอบด้วยโหนดแรกสุดที่เรียกว่า โหนดราก (Root Node) ซึ่งเป็นคุณลักษณะที่แบ่งกลุ่มได้ดีที่สุด กลุ่มข้อมูลจะถูกแบ่งเป็นกลุ่มย่อยไปเรื่อยๆ จนกระทั่งการแบ่งดังกล่าวไม่มีนัยสำคัญทางสถิติหรือไม่สามารถแบ่งย่อยต่อไปได้ กระบวนวิธีที่ใช้แยกกลุ่มข้อมูลมีได้หลายกระบวนวิธีด้วยกัน เช่น

- Automatic Interaction Detection (AID) ใช้ตัวสถิติทดสอบ t ในการจำแนก
- AID แบบไคสแควร์ ใช้ตัวสถิติทดสอบไคสแควร์ในการจำแนก
- Classification and Regression Tree (CART) ใช้ดัชนีความผันแปร (Index of Diversity) ในการจำแนก
- ขั้นตอนวิธี ID3 (Quinlan, J.R. 1986) ใช้ ค่าวัดที่อิง Entropy (Entropy-based measure) ซึ่งรู้จักกันในนามค่าอินฟอร์เมชันเกน (Information Gain) ในการเลือกตัวแปรที่ใช้เป็นตัวจำแนก

ต้นไม้ตัดสินใจสามารถนำไปสร้างกฎการจำแนกโดยเขียนอยู่ในรูป If-Then ซึ่งสามารถนำไปใช้งานต่อในลำดับถัดไป

2.6 ค่าอินฟอร์เมชันเกน

ค่าอินฟอร์เมชันเกนหรือเรียกอีกชื่อหนึ่งว่า ID3 เป็นกระบวนวิธีหนึ่งที่ใช้แยกกลุ่มข้อมูล พวงทิพย์ แท่นแสง(2550) อธิบายว่า ค่าอินฟอร์เมชันเกนเป็นขั้นตอนวิธีในการสร้างต้นไม้ตัดสินใจ ที่ใช้หลักการของทฤษฎีข่าวสาร ค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการแบ่งข้อมูล โดยวิธีกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นการเลือกข้อมูลตามลำดับของตัวชี้วัดหรือค่าเกน (Gain) สูงที่สุดเป็นข้อมูลเริ่มต้น และข้อมูลถัดไปที่มีค่าลดหลั่นกันตามลำดับ โดยมีหลักการคำนวณดังนี้

กำหนดให้ P และ N : เป็นชุดข้อมูลที่ใช้แบ่งตัวอย่างข้อมูล

p : จำนวนตัวอย่าง ที่อยู่ในชุดข้อมูล P

n : จำนวนตัวอย่าง ที่อยู่ในชุดข้อมูล N

A : คุณลักษณะ

ค่าอินฟอร์เมชันเกนของคุณลักษณะ A เป็นดังสมการ

$$Gain(A) = I(p,n) - E(A)$$

โดยที่

$I(p,n)$ คือค่าสารสนเทศของกลุ่มข้อมูล เป็นค่าคาดคะเนที่กลุ่มข้อมูลตัวอย่างต้องใช้จำนวนบิต ในการแยกชุดข้อมูล P และ N นิยามโดย

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

$E(A)$ คือค่าคาดคะเนของข้อมูล(Entropy) ที่แยกโดยการใช้คุณลักษณะ A โดยที่ A เป็นคุณลักษณะที่แบ่ง S ออกเป็น $\{S_1, S_2, S_3, \ldots, S_v\}$ โดยที่ Si มีข้อมูลตัวอย่างจากชุดข้อมูล P อยู่จำนวน Pi และตัวอย่างจากชุดข้อมูล N อยู่จำนวน Ni ดังสมการ

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Simon Colton(2004) อธิบายว่า ในกรณีที่ชุดข้อมูลที่ใช้แบ่งตัวอย่างข้อมูลมีมากกว่า 2 ค่านั้น ค่าสารสนเทศของกลุ่มข้อมูลและค่าคาดคะเนของข้อมูล สามารถคำนวณได้จากสูตรดังนี้

$$I(S) = \sum_{i=1}^{n} -\frac{p_i}{S}\log_2\left(\frac{p_i}{S}\right)$$

โดยที่ S : จำนวนรายการข้อมูลทั้งหมดในกลุ่มข้อมูลนั้นๆ

P : จำนวนรายการที่จำแนกตามชุดข้อมูลนั้นๆ

n : จำนวนชุดข้อมูลที่ใช้แบ่งตัวอย่างข้อมูล

$$E(A) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$$

โดยที่ A : คุณลักษณะที่ต้องการหาค่าคาดคะเน

v : ค่าของคุณลักษณะ A

เพื่อความเข้าใจขอยกตัวอย่างประกอบ โดยตัวอย่างข้อมูลการซื้อเครื่องปรับอากาศของลูกค้าแต่ละราย ดังตาราง 2.1

ตาราง 2.1 ตารางข้อมูลตัวอย่างการซื้อเครื่องปรับอากาศของลูกค้า

อายุ (Age)	รายได้ (Income)	การมีงานทำ (Working)	ความน่าเชื่อถือ (Credit)	ซื้อ/ไม่ซื้อ (Buy)
≤ 30	สูง	ไม่มี	ปานกลาง	ไม่ซื้อ
≤ 30	สูง	ไม่มี	สูง	ไม่ซื้อ
31 – 39	สูง	ไม่มี	ปานกลาง	ซื้อ
≥ 40	ปานกลาง	ไม่มี	ปานกลาง	ซื้อ
≥ 40	ต่ำ	มี	ปานกลาง	ซื้อ
≥ 40	ต่ำ	มี	สูง	ไม่ซื้อ
31 – 39	ต่ำ	มี	สูง	ซื้อ
≤ 30	ปานกลาง	ไม่มี	ปานกลาง	ไม่ซื้อ
≤ 30	ต่ำ	มี	ปานกลาง	ซื้อ
≥ 40	ปานกลาง	มี	ปานกลาง	ซื้อ
≤ 30	ปานกลาง	มี	สูง	ซื้อ
31 – 39	ปานกลาง	ไม่มี	สูง	ซื้อ
31 – 39	สูง	มี	ปานกลาง	ซื้อ
≥ 40	ปานกลาง	ไม่มี	สูง	ไม่ซื้อ

<u>คุณลักษณะ</u> ได้แก่ อายุ, รายได้, การมีงานทำ, ความน่าเชื่อถือ โดยที่

อายุ ประกอบด้วยค่าของข้อมูล ดังนี้ ≤ 30 , $31 - 39$, ≥ 40

รายได้ ประกอบด้วยค่าของข้อมูล ดังนี้

สูง หมายถึงรายได้ตั้งแต่ 3 หมื่นบาทต่อเดือนขึ้นไป

กลาง หมายถึงรายได้ที่อยู่ในช่วง 1 ถึง 3 หมื่นบาทต่อเดือน

ต่ำ หมายถึงรายได้ที่ต่ำกว่า 1 หมื่นบาทต่อเดือน

การมีงานทำ ประกอบด้วยค่าของข้อมูล ดังนี้ มี, ไม่มี

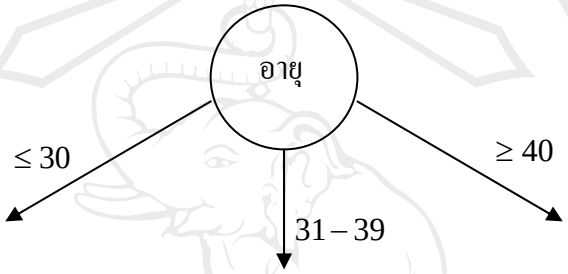
ความน่าเชื่อถือ ประกอบด้วยค่าของข้อมูล ดังนี้ ปานกลาง, สูง

<u>ชุดข้อมูล</u> คือ การซื้อเครื่องปรับอากาศ ประกอบด้วยค่าของข้อมูล ดังนี้ ซื้อ, ไม่ซื้อ

ในการสร้างต้นไม้ตัดสินใจ จะต้องหาค่าอินฟอร์เมชันเกนเพื่อใช้แบ่งคุณลักษณะทั้ง 4 คือ อายุ รายได้ การมีงานทำ และความน่าเชื่อถือ แล้วนำมาพิจารณาคำตอบว่าการซื้อเครื่องปรับอากาศตามชุดข้อมูลว่า ซื้อ หรือ ไม่ซื้อ

กระบวนการสร้างต้นไม้ตัดสินใจโดยใช้ค่าอินฟอร์เมชันเกน สามารถสรุปเป็นขั้นตอนคร่าวๆ ได้ดังนี้

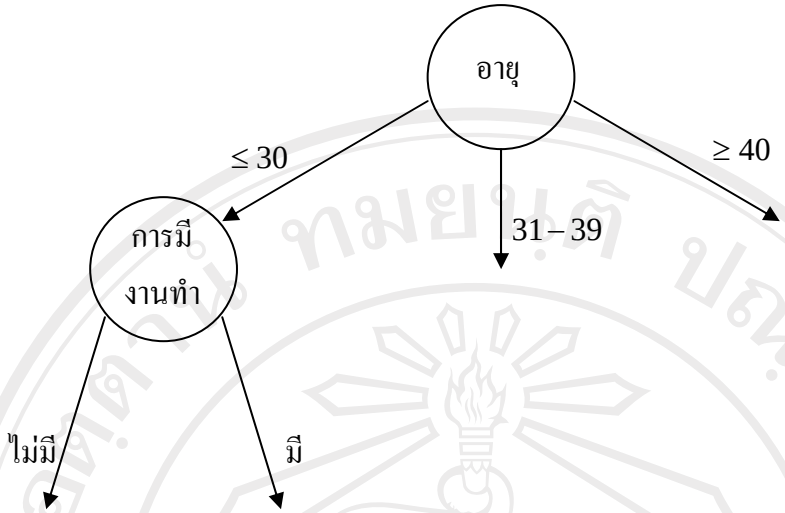
ขั้นตอนที่ 1 หาโหนดรากโดยพิจารณาจากคุณลักษณะที่มีค่าอินฟอร์เมชันเกนสูงที่สุดจากข้อมูลตัวอย่างดังกล่าวข้างต้น สามารถหาได้ว่า คุณลักษณะอายุมีค่าอินฟอร์เมชันเกนสูงที่สุดและถูก พิจารณาให้เป็นโหนดรากใช้สำหรับการแบ่งข้อมูลออกได้เป็น 3 ทางเลือกดังรูป 2.2



รูป 2.2 รูปคุณลักษณะอายุ ถูกเลือกให้เป็นโหนดราก

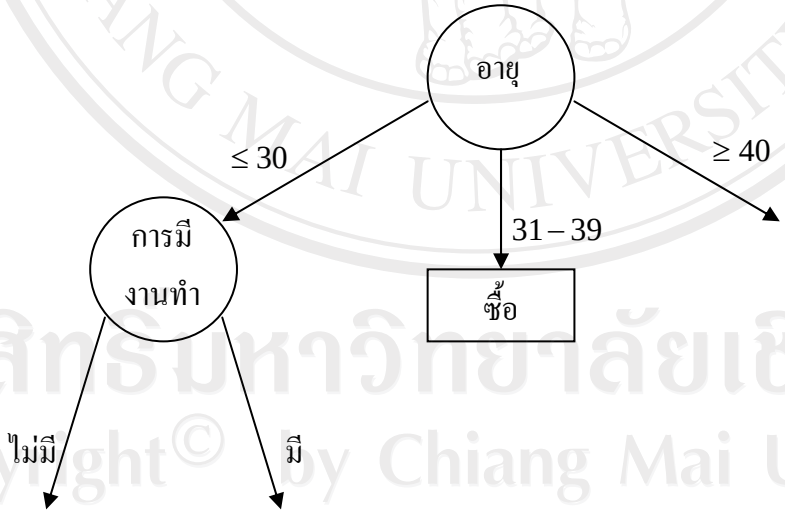
ขั้นตอนที่ 2 หาคุณลักษณะที่จะใช้แบ่งข้อมูลในอันดับถัดไปจากโหนดรากโดยพิจารณาจากการแบ่งอายุที่มีทางเลือก 3 ทางเลือก หรือ เรียกอีกอย่างว่า 3 กิ่งดังกล่าว ให้เราพิจารณาทีละกิ่งว่าโหนดถัดไปของแต่ละกิ่งควรจะใช้คุณลักษณะใดในการแบ่งข้อมูลในลำดับถัดไป

ขั้นตอนที่ 2.1 พิจารณาจากกิ่งที่ 1 ค่าสารสนเทศของกลุ่มข้อมูล(อายุ ≤ 30) มีค่ามากกว่าศูนย์ ฉะนั้นจะต้องหาคุณลักษณะที่จะใช้แบ่งข้อมูลในลำดับถัดไป จากข้อมูลตัวอย่างสามารถหาได้ว่า คุณลักษณะการมีงานทำมีค่าอินฟอร์เมชันเกนสูงที่สุด และถูกพิจารณาให้เป็นโหนดถัดไปจากโหนดราก และใช้สำหรับการแบ่งข้อมูลออกได้เป็น 2 ทางเลือกดังรูป 2.3



รูป 2.3 รูปคุณลักษณะการมีงานทำ ถูกเลือกให้เป็นโหนดถัดมาจากกิ่งที่ 1 ของอายุ

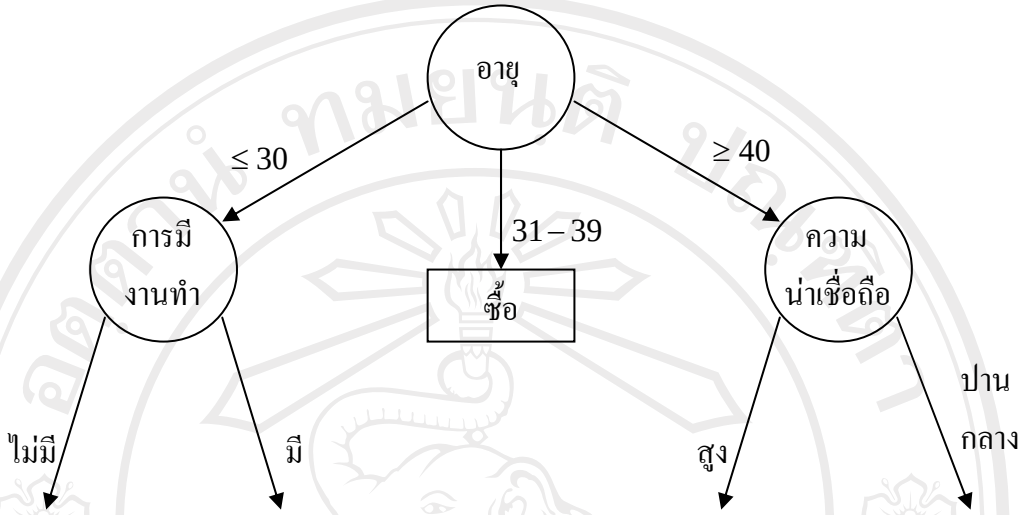
ขั้นตอนที่ 2.2 พิจารณาจากกิ่งที่ 2 ค่าสารสนเทศของกลุ่มข้อมูล ($31 \leq$ อายุ ≤ 39) มีค่าเป็นศูนย์ หมายถึงทุกรายการอยู่ในชุดข้อมูลเดียวกันทั้งหมด เมื่อพิจารณาจากข้อมูลตัวอย่างจะเห็นได้ว่าเป็นรายการซื้อทั้งหมด ดังนั้นไม่ต้องมีการแบ่งข้อมูลอีกต่อไปสามารถสรุปได้ว่า ถ้าลูกค้าที่มีอายุระหว่าง 31-39 ปี แล้วจะซื้อเครื่องปรับอากาศ แสดงได้ดังรูป 2.4



รูป 2.4 รูปคุณลักษณะอายุระหว่าง 31-39 ปี (กิ่งที่ 2 ของอายุ)ไม่ต้องหาโหนดถัดไป

ขั้นตอนที่ 2.3 พิจารณาจากกิ่งที่ 3 ค่าสารสนเทศของกลุ่มข้อมูล(อายุ ≥ 40) มีค่ามากกว่าศูนย์ ฉะนั้นจะต้องหาคุณลักษณะที่จะใช้แบ่งข้อมูลในลำดับถัดไป จากข้อมูลตัวอย่างสามารถหา

ได้ว่า คุณลักษณะความน่าเชื่อถือมีค่าอินฟอร์เมชันเกนสูงที่สุด และถูกพิจารณาให้เป็นโหนดถัดไปจากโหนดราก และใช้สำหรับการแบ่งข้อมูลออกได้เป็น 2 ทางเลือกดังรูป 2.5



รูป 2.5 รูปคุณลักษณะความน่าเชื่อถือ ถูกเลือกให้เป็นโหนดถัดมาจากกิ่งที่ 3 ของอายุ

จากขั้นตอนดังกล่าวข้างต้น สามารถแสดงรายละเอียดการคำนวณและการสร้างต้นไม้ตัดสินใจจากข้อมูลตัวอย่างการซื้อเครื่องปรับอากาศดังกล่าวข้างต้น ซึ่งมีทั้งหมด(All) 14 รายการ สามารถหาค่าสารสนเทศของกลุ่มข้อมูลได้ดังนี้

ค่าสารสนเทศของกลุ่มข้อมูล = I(p,n) = I(9,5) โดยที่ p = 9 (ซื้อ), n = 5 (ไม่ซื้อ)

$$I(All) = I(9,5) = -\frac{9}{9+5}\log_2\left(\frac{9}{9+5}\right) - \frac{5}{9+5}\log_2\left(\frac{5}{9+5}\right) = 0.9403$$

ขั้นตอนที่ 1. หาคุณลักษณะที่จะใช้แบ่งข้อมูลอันดับแรก หรือหาโหนดราก ตามลำดับดังนี้

1.1) หาค่าอินฟอร์เมชันเกนของอายุ โดยพิจารณาข้อมูลทั้ง 14 รายการ แล้วจำแนกตามอายุที่มี 3 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.2

ตาราง 2.2 ตารางแจกแจงข้อมูลจำแนกตามอายุของลูกค้า

อายุ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
≤ 30	5	2	3
31 – 39	4	4	0
≥ 40	5	3	2

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะอายุ

$$E(Age) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.6935$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะอายุ แสดงดังสมการ

$$Gain(Age) = I(All) - E(Age) = 0.9403 - 0.6935 = 0.2468$$

1.2) ค่าอินฟอร์เมชันเกนของรายได้ โดยพิจารณาข้อมูลทั้ง 14 รายการ แล้วจำแนกตามรายได้ที่มี 3 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.3

ตาราง 2.3 ตารางแจกแจงข้อมูลจำแนกตามรายได้ของลูกค้า

รายได้	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
สูง	4	2	2
ปานกลาง	6	4	2
ต่ำ	4	3	1

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะรายได้

$$E(Income) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2) + \frac{4}{14} I(3,1) = 0.9111$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะรายได้ แสดงดังสมการ

$$Gain(Income) = I(All) - E(Income) = 0.9403 - 0.9111 = 0.0292$$

1.3) ค่าอินฟอร์เมชันเกนของการมีงานทำ โดยพิจารณาข้อมูลทั้ง 14 รายการ แล้วจำแนกตามการมีงานทำที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.4

ตาราง 2.4 ตารางแจกแจงข้อมูลจำแนกตามการมีงานทำของลูกค้า

การมีงานทำ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
มี	7	6	1
ไม่มี	7	3	4

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะการมีงานทำ

$$E(Working) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{7}{14} I(6,1) + \frac{7}{14} I(3,4) = 0.7885$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะการมีงานทำ แสดงดังสมการ

$$Gain(Working) = I(All) - E(Working) = 0.9403 - 0.7885 = 0.1518$$

1.4) ค่าอินฟอร์เมชันเกนของความน่าเชื่อถือ โดยพิจารณาข้อมูลทั้ง 14 รายการ แล้วจำแนกตามความน่าเชื่อถือที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.5

ตาราง 2.5 ตารางแจกแจงข้อมูลจำแนกตามความน่าเชื่อถือของลูกค้า

ความน่าเชื่อถือ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
ปานกลาง	8	6	2
สูง	6	3	3

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะความน่าเชื่อถือ

$$E(Credit) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3) = 0.8922$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะความน่าเชื่อถือ แสดงดังสมการ

$$Gain(Credit) = I(All) - E(Credit) = 0.9403 - 0.8922 = 0.0481$$

เมื่อหา ค่าอินฟอร์เมชันเกนของคุณลักษณะครบทั้ง 4 ตัวแล้วสรุปผลได้ดังนี้

$Gain(Age) = 0.2468$ $Gain(Income) = 0.0292$

$Gain(Working) = 0.1518$ $Gain(Credit) = 0.0481$

เมื่อเปรียบเทียบแล้วจะเห็นว่าค่าอินฟอร์เมชันเกนของอายุ มีค่ามากที่สุด ซึ่งเราจะเลือกคุณลักษณะอายุมาเป็นทางเลือกแรกในการแบ่งข้อมูล คุณลักษณะอายุจะมีคุณสมบัติเป็นโหนดราก

ขั้นตอนที่ 2. หาคุณลักษณะที่จะใช้แบ่งข้อมูลในอันดับถัดไปจากโหนดรากโดยพิจารณาจากการแบ่งอายุที่มีทางเลือก 3 ทางนั้น แสดงว่าสามารถแบ่งข้อมูลออกไปได้ 3 กิ่ง ให้เราพิจารณาทีละกิ่งว่าโหนดถัดไปของแต่ละกิ่งควรจะใช้คุณลักษณะใดในการแบ่งข้อมูลต่อ

2.1) จากกิ่งแรก อายุ ≤ 30 ปี เมื่อพิจารณาแล้วมีทั้งหมด 5 รายการ ซื้อ 2 รายการ ไม่ซื้อ 3 รายการ ค่าสารสนเทศของกลุ่มข้อมูล = I(p,n) = I(2,3) โดยที่ p = 2 (ซื้อ), n = 3 (ไม่ซื้อ)

$$I(Age \leq 30) = I(2,3) = -\frac{2}{2+3}\log_2\left(\frac{2}{2+3}\right) - \frac{3}{2+3}\log_2\left(\frac{3}{2+3}\right) = 0.9709$$

2.1.1) หาค่าอินฟอร์เมชันเกนของอายุ โดยพิจารณาข้อมูลทั้ง 5 รายการออกมาได้ข้อมูลดังตาราง 2.6

ตาราง 2.6 ตารางแจกแจงข้อมูลจำแนกตามอายุของลูกค้า เฉพาะอายุ ≤ 30 ปี

อายุ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
≤ 30	5	2	3

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะอายุ

$$E(Age \leq 30 \rightarrow Age) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{5}{5} I(2,3) = 0.9709$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะอายุ แสดงดังสมการ

$$Gain(Age \leq 30 \rightarrow Age) = I(Age \leq 30) - E(Age \leq 30 \rightarrow Age)$$

$$= 0.9709 - 0.9709 = 0$$

2.1.2) หาค่าอินฟอร์เมชันเกนของรายได้ โดยพิจารณาข้อมูลทั้ง 5 รายการ แล้วจำแนกตามรายได้ที่มี 3 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.7

ตาราง 2.7 ตารางแจกแจงข้อมูลจำแนกตามรายได้ของลูกค้า เฉพาะอายุ ≤ 30 ปี

รายได้	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
สูง	2	0	2
ปานกลาง	2	1	1
ต่ำ	1	1	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะรายได้

$$E(Age \leq 30 \rightarrow Income) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$= \frac{2}{5} I(0,2) + \frac{2}{5} I(1,1) + \frac{1}{5} I(1,0) = 0.4000$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะรายได้ แสดงดังสมการ

$$Gain(Age \leq 30 \rightarrow Income) = I(Age \leq 30) - E(Age \leq 30 \rightarrow Income)$$

$$= 0.9709 - 0.4000 = 0.5709$$

2.1.3) หาค่าอินฟอร์เมชันเกนของการมีงานทำ โดยพิจารณาข้อมูลทั้ง 5 รายการ แล้วจำแนกตามการมีงานทำที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.8

ตาราง 2.8 ตารางแจกแจงข้อมูลจำแนกตามการมีงานทำของลูกค้า เฉพาะอายุ ≤ 30 ปี

การมีงานทำ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
มี	2	2	0
ไม่มี	3	0	3

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะการมีงานทำ

$$E(Age \leq 30 \rightarrow Working) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i) = \frac{2}{5} I(2,0) + \frac{3}{5} I(0,3) = 0$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะการมีงานทำ แสดงดังสมการ

$$Gain(Age \leq 30 \rightarrow Working) = I(Age \leq 30) - E(Age \leq 30 \rightarrow Working)$$

$$= 0.9709 - 0 = 0.9709$$

2.1.4) หาค่าอินฟอร์เมชันเกนของความน่าเชื่อถือ โดยพิจารณาข้อมูลทั้ง 5 รายการ แล้วจำแนกตามการมีงานทำที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.9

ตาราง 2.9 ตารางแจกแจงข้อมูลจำแนกตามความน่าเชื่อถือของลูกค้า เฉพาะอายุ ≤ 30 ปี

ความน่าเชื่อถือ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
ปานกลาง	3	1	2
สูง	2	1	1

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะความน่าเชื่อถือ

$$E(Age \le 30 \rightarrow Credit) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{3}{5} I(1,2) + \frac{2}{5} I(1,1) = 0.9510$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะความน่าเชื่อถือ แสดงดังสมการ

$$Gain(Age \le 30 \rightarrow Credit) = I(Age \le 30) - E(Age \le 30 \rightarrow Credit)$$
$$= 0.9709 - 0.9510 = 0.0199$$

จากกิ่งแรก อายุ ≤ 30 ปี เมื่อหา ค่าอินฟอร์เมชันเกนของคุณลักษณะครบทั้ง 4 ตัวแล้วสรุปผลได้ดังนี้

$Gain(Age \le 30 \rightarrow Age) = 0$

$Gain(Age \le 30 \rightarrow Income) = 0.5709$

$Gain(Age \le 30 \rightarrow Working) = 0.9709$

$Gain(Age \le 30 \rightarrow Credit) = 0.0199$

เมื่อเปรียบเทียบแล้วจะเห็นว่าค่าอินฟอร์เมชันเกน $Age \le 30 \rightarrow Working$ มีค่ามากที่สุด ซึ่งเราจะใช้ การมีงานทำ เป็นตัวแบ่งข้อมูลในลำดับถัดไป

2.2) จากกิ่งที่สอง อายุในช่วง 31-39 ปี เมื่อพิจารณาแล้วมีทั้งหมด 4 รายการ ซื้อทั้งหมด 4 รายการ ไม่มีรายการใด ไม่ซื้อ รายการค่าสารสนเทศของกลุ่มข้อมูล = I(p,n) = I(4,0) โดยที่ p = 4 (ซื้อ), n = 0 (ไม่ซื้อ)

$$I(31 \le Age \le 39) = I(4,0) = -\frac{4}{4+0}\log_2\left(\frac{4}{4+0}\right) - \frac{0}{4+0}\log_2\left(\frac{0}{4+0}\right) = 0$$

และพิจารณาข้อมูลทั้ง 4 รายการออกมาได้ข้อมูลดังตาราง 2.10

ตาราง 2.10 ตารางแจกแจงข้อมูลจำแนกตามอายุของลูกค้า เฉพาะอายุช่วง 31-39 ปี

อายุ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
31 – 39	4	4	0

จะเห็นว่ามีแต่รายการซื้อทั้งหมด นั่นคือทุกรายการอยู่ในชุดข้อมูลเดียวกันทั้งหมด จึงไม่ต้องมีการแบ่งข้อมูลอีกต่อไป สามารถให้คำตอบได้ทันทีว่า ถ้าลูกค้าที่มีอายุระหว่าง 31-39 ปี แล้วจะซื้อเครื่องปรับอากาศ

2.3) จากกิ่งที่สาม อายุ ≥ 40 ปี เมื่อพิจารณาแล้วมีทั้งหมด 5 รายการ ซื้อ 3 รายการ ไม่ซื้อ 2 รายการค่าสารสนเทศของกลุ่มข้อมูล = I(p,n) = I(3,2) โดยที่ p = 3 (ซื้อ), n = 2 (ไม่ซื้อ)

$$I(Age \le 30) = I(3,2) = -\frac{3}{3+2}\log_2\left(\frac{3}{3+2}\right) - \frac{2}{3+2}\log_2\left(\frac{2}{3+2}\right) = 0.9709$$

2.3.1) หาค่าอินฟอร์เมชันเกนของอายุ โดยพิจารณาข้อมูลทั้ง 5 รายการออกมาได้ข้อมูลดังตาราง 2.11

ตาราง 2.11 ตารางแจกแจงข้อมูลจำแนกตามอายุของลูกค้า เฉพาะอายุ ≥ 40 ปี

อายุ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
≥ 40	5	3	2

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ อายุ

$$E(Age \geq 40 \rightarrow Age) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i) = \frac{5}{5} I(3,2) = 0.9709$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะอายุ แสดงดังสมการ

$$Gain(Age \geq 40 \rightarrow Age) = I(Age \geq 40) - E(Age \geq 40 \rightarrow Age)$$

$$= 0.9709 - 0.9709 = 0$$

2.3.2) หาค่าอินฟอร์เมชันเกนของรายได้ โดยพิจารณาข้อมูลทั้ง 5 รายการ แล้วจำแนกตามรายได้ที่มี 3 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.12

ตาราง 2.12 ตารางแจกแจงข้อมูลจำแนกตามรายได้ของลูกค้า เฉพาะอายุ ≥ 40 ปี

รายได้	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
สูง	0	0	0
ปานกลาง	3	2	1
ต่ำ	2	1	1

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ รายได้

$$E(Age \geq 40 \rightarrow Income) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$= \frac{0}{5} I(0,0) + \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1) = 0.9510$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะรายได้ แสดงดังสมการ

$$Gain(Age \geq 40 \rightarrow Income) = I(Age \geq 40) - E(Age \geq 40 \rightarrow Income)$$

$$= 0.9709 - 0.9510 = 0.0199$$

2.3.3) หาค่าอินฟอร์เมชันเกนของ การมีงานทำ โดยพิจารณาข้อมูลทั้ง 5 รายการ แล้วจำแนกตามการมีงานทำที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.13

ตาราง 2.13 ตารางแจกแจงข้อมูลจำแนกตามการมีงานทำของลูกค้า เฉพาะอายุ ≥ 40 ปี

การมีงานทำ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
มี	3	2	1
ไม่มี	2	1	1

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะการมีงานทำ

$$E(Age \geq 40 \rightarrow Working) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1) = 0.9510$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะการมีงานทำ แสดงดังสมการ

$$Gain(Age \geq 40 \rightarrow Working) = I(Age \geq 40) - E(Age \geq 40 \rightarrow Working)$$

$$= 0.9709 - 0 = 0.9510 = 0.0199$$

2.3.4) หาค่าอินฟอร์เมชันเกนของความน่าเชื่อถือ โดยพิจารณาข้อมูลทั้ง 5 รายการ แล้วจำแนกตามการมีงานทำที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.14

ตาราง 2.14 ตารางแจกแจงข้อมูลจำแนกตามความน่าเชื่อถือของลูกค้า เฉพาะอายุ ≥ 40 ปี

ความน่าเชื่อถือ	จำนวนรวม	จำนวนผู้ซื้อ	จำนวนผู้ไม่ซื้อ
ปานกลาง	3	3	0
สูง	2	0	2

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ ความน่าเชื่อถือ

$$E(Age \geq 40 \rightarrow Credit) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2) = 0$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะความน่าเชื่อถือ แสดงดังสมการ

$$Gain(Age \geq 40 \rightarrow Credit) = I(Age \geq 40) - E(Age \geq \rightarrow Credit)$$

$$= 0.9709 - 0 = 0.9709$$

จากกิ่งที่สาม อายุ ≥ 40 ปี เมื่อหา ค่าอินฟอร์เมชันเกนของคุณลักษณะครบทั้ง 4 ตัวแล้ว สรุปผลได้ดังนี้

$$Gain(Age \geq 40 \rightarrow Age) = 0$$

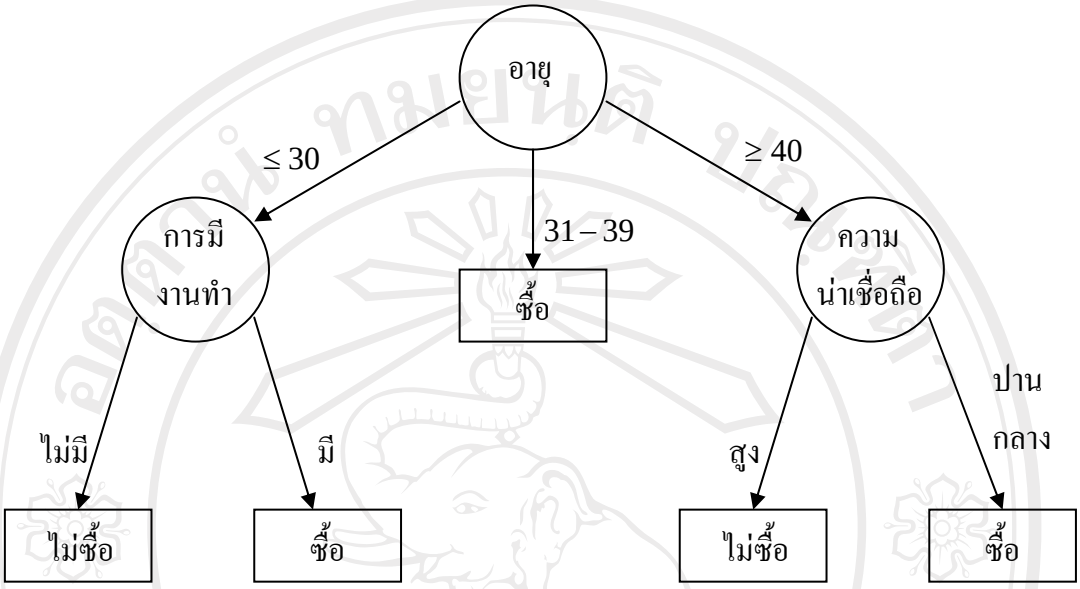
$$Gain(Age \geq 40 \rightarrow Income) = 0.0199$$

$$Gain(Age \geq 40 \rightarrow Working) = 0.0199$$

$$Gain(Age \geq 40 \rightarrow Credit) = 0.9709$$

เมื่อเปรียบเทียบแล้วจะเห็นว่าค่าอินฟอร์เมชันเกน $Age \geq 40 \rightarrow Credit$ มีค่ามากที่สุด ซึ่งเราจะใช้ ความน่าเชื่อถือ เป็นตัวแบ่งข้อมูลในลำดับถัดไป

ท้ายที่สุดสามารถแสดงรูปของต้นไม้ตัดสินใจกรณีตัวอย่างข้อมูลการซื้อเครื่องปรับอากาศของลูกค้าแต่ละราย ได้ดังรูป 2.6



รูป 2.6 รูปต้นไม้ตัดสินใจกรณีตัวอย่างข้อมูลการซื้อเครื่องปรับอากาศของลูกค้า

จากต้นไม้ตัดสินใจสามารถนำมาเขียนในรูป If-Then ได้ดังนี้

IF Age="≤ 30" and Working="ไม่มี" THEN Buy="ไม่ซื้อ"

IF Age="≤ 30" and Working="มี" THEN Buy="ซื้อ"

IF Age="31 – 39" THEN Buy="ซื้อ"

IF Age="≥ 40" and Credit="สูง" THEN Buy="ไม่ซื้อ"

IF Age="≥ 40" and Credit="ปานกลาง" THEN Buy="ซื้อ"

เพื่อเพิ่มความเข้าใจกรณีการแบ่งชุดข้อมูลมากกว่า 2 ค่า จะขอยกตัวอย่างประกอบ โดยตัวอย่างข้อมูลการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์ ดังตาราง 2.15

ตาราง 2.15 ตารางข้อมูลตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์

Weekend	Weather	Parent	Money	Activity
1	Sunny	Yes	Rich	Cinema
2	Sunny	No	Rich	Tennis
3	Windy	Yes	Rich	Cinema
4	Rainy	Yes	Poor	Cinema
5	Rainy	No	Rich	Stay in

ตาราง 2.15 ตารางข้อมูลตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์(ต่อ)

Weekend	Weather	Parent	Money	Activity
6	Rainy	Yes	Poor	Cinema
7	Windy	No	Poor	Cinema
8	Windy	No	Rich	Shopping
9	Windy	Yes	Rich	Cinema
10	Sunny	No	Rich	Tennis

<u>คุณลักษณะ</u> ได้แก่ Weather, Parent และ Money โดยที่

Weather ประกอบด้วยค่าของข้อมูล ดังนี้ Sunny, Windy และ Rainy

Parents ประกอบด้วยค่าของข้อมูล ดังนี้ Yes และ No

Money ประกอบด้วยค่าของข้อมูล ดังนี้ Rich และ Poor

<u>ชุดข้อมูล</u> คือ Activity ประกอบด้วยค่าของข้อมูล ดังนี้ Cinema, Tennis, Stay in และ Shopping

ในการสร้างต้นไม้ตัดสินใจ จะต้องหาค่าอินฟอร์เมชันเกนเพื่อแบ่งคุณลักษณะทั้ง 3 คือ Weather, Parent และ Money แล้วนำมาพิจารณาคำตอบ Activity ตามชุดข้อมูลว่า Cinema, Tennis, Stay in หรือ Shopping

กระบวนการสร้างต้นไม้ตัดสินใจโดยใช้ค่าอินฟอร์เมชันเกน สามารถแสดงรายละเอียดการคำนวณและการสร้างต้นไม้ตัดสินใจจากข้อมูลตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์ดังกล่าวข้างต้น ซึ่งมีทั้งหมด 10 รายการ สามารถหาค่าสารสนเทศของกลุ่มข้อมูลได้ดังนี้

ค่าสารสนเทศของกลุ่มข้อมูล = I(Cinema,Tennis,Stay in,Shopping) = I(6,2,1,1)

โดยที่ Cinema = 6 รายการ

Tennis = 2 รายการ

Stay in = 1 รายการ

Shopping = 1 รายการ

$$I(S) = I(6,2,1,1) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{1}{10}\log_2\frac{1}{10} - \frac{1}{10}\log_2\frac{1}{10}$$

$$I(6,2,1,1) = 1.571$$

ขั้นตอนที่ 1. หาคุณลักษณะที่จะใช้แบ่งข้อมูลอันดับแรก หรือหาโหนดราก ตามลำดับดังนี้

1.1) หาค่าอินฟอร์เมชันเกนของ Weather โดยพิจารณาข้อมูลทั้ง 10 รายการ แล้วจำแนกตาม Weather ที่มี 3 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.16

ตาราง 2.16 ตารางแจกแจงข้อมูลจำแนกตาม Weather

Weather	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Sunny	3	1	2	0	0
Windy	4	3	0	0	1
Rainy	3	2	0	1	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Weather

$$E(Weather) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{3}{10} I(1,2,0,0) + \frac{4}{10} I(3,0,0,1) + \frac{3}{10} I(2,0,1,0)$$

$$I(1,2,0,0) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} - \frac{0}{3}\log_2\frac{0}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0.918$$

$$I(3,0,0,1) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{0}{4}\log_2\frac{0}{4} - \frac{0}{4}\log_2\frac{0}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$$

$$I(2,0,1,0) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{0}{3}\log_2\frac{0}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0.918$$

$$E(Weather) = \frac{3}{10}0.918 + \frac{4}{10}0.811 + \frac{3}{10}0.918 = 0.875$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Weather แสดงดังสมการ

$$Gain(Weather) = I(All) - E(Weather) = 1.571 - 0.875 = 0.696$$

1.2) ค่าอินฟอร์เมชันเกนของ Parents โดยพิจารณาข้อมูลทั้ง 10 รายการ แล้วจำแนกตาม Parents ที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.17

ตาราง 2.17 ตารางแจกแจงข้อมูลจำแนกตาม Parent

Parents	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Yes	5	5	0	0	0
No	5	1	2	1	1

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Parents

$$E(Parents) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{5}{10} I(5,0,0,0) + \frac{5}{10} I(1,2,1,1)$$

$$I(5,0,0,0) = -\frac{5}{5}\log_2 \frac{5}{5} - \frac{0}{0}\log_2 \frac{0}{0} - \frac{0}{0}\log_2 \frac{0}{0} - \frac{0}{0}\log_2 \frac{0}{0} = 0.000$$

$$I(1,2,1,1) = -\frac{1}{5}\log_2 \frac{1}{5} - \frac{2}{5}\log_2 \frac{2}{5} - \frac{1}{5}\log_2 \frac{1}{5} - \frac{1}{5}\log_2 \frac{1}{5} = 1.922$$

$$E(Parents) = \frac{5}{10} 0.000 + \frac{5}{10} 1.922 = 0.961$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Parents แสดงดังสมการ

$$Gain(Parents) = I(All) - E(Parents) = 1.571 - 0.961 = 0.610$$

1.3) ค่าอินฟอร์เมชันเกนของ Money โดยพิจารณาข้อมูลทั้ง 10 รายการ แล้วจำแนกตาม Money ที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.18

ตาราง 2.18 ตารางแจกแจงข้อมูลจำแนกตาม Money

Money	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Rich	7	3	2	1	1
Poor	3	3	0	0	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Money

$$E(Money) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{7}{10} I(3,2,1,1) + \frac{3}{10} I(3,0,0,0)$$

$$I(3,2,1,1) = -\frac{3}{7}\log_2 \frac{3}{7} - \frac{2}{7}\log_2 \frac{2}{7} - \frac{1}{7}\log_2 \frac{1}{7} - \frac{1}{7}\log_2 \frac{1}{7} = 1.842$$

$$I(3,0,0,0) = -\frac{3}{3}\log_2 \frac{3}{3} - \frac{0}{3}\log_2 \frac{0}{3} - \frac{0}{3}\log_2 \frac{0}{3} - \frac{0}{3}\log_2 \frac{0}{3} = 0.000$$

$$E(Money) = \frac{7}{10} 1.842 + \frac{3}{10} 0.000 = 1.290$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Money แสดงดังสมการ

$$Gain(Money) = I(All) - E(Money) = 1.571 - 1.290 = 0.281$$

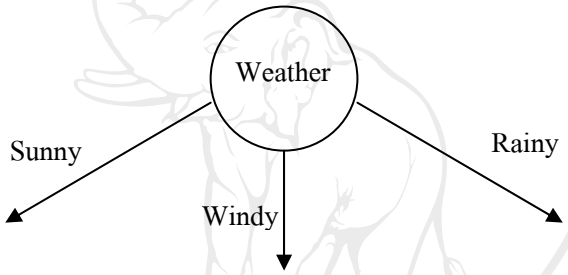
เมื่อหา ค่าอินฟอร์เมชันเกนของคุณลักษณะครบทั้ง 3 ตัวแล้วสรุปผลได้ดังนี้

$$Gain(Weather) = 0.696$$

$$Gain(Parents) = 0.610$$

$$Gain(Money) = 0.281$$

เมื่อเปรียบเทียบแล้วจะเห็นว่าค่าอินฟอร์เมชันเกนของ Weather มีค่ามากที่สุด ซึ่งเราจะเลือกคุณลักษณะ Weather มาเป็นทางเลือกแรกในการแบ่งข้อมูล และนั่นหมายถึงคุณลักษณะ Weather จะมีคุณสมบัติเป็นโหนดราก ซึ่งสามารถแสดงได้ดังรูป 2.7



รูป 2.7 รูปคุณลักษณะ Weather ถูกเลือกให้เป็นโหนดราก

ขั้นตอนที่ 2. หาคุณลักษณะที่จะใช้แบ่งข้อมูลในอันดับถัดไปจากโหนดรากโดยพิจารณาจากการแบ่ง Weather ที่มีทางเลือก 3 ทางนั้น แสดงว่าสามารถแบ่งข้อมูลออกไปได้ 3 กิ่ง ให้เราพิจารณาทีละกิ่งว่าโหนดถัดไปของแต่ละกิ่งควรจะใช้คุณลักษณะใดในการแบ่งข้อมูลต่อ

2.1) จากกิ่งแรก Weather มีค่าเป็น Sunny เมื่อพิจารณาแล้วมีทั้งหมด 3 รายการ ดังตาราง 2.19

ตาราง 2.19 ตารางข้อมูลตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์ เฉพาะ Weather มีค่าเป็น Sunny

Weekend	Weather	Parent	Money	Activity
1	Sunny	Yes	Rich	Cinema
2	Sunny	No	Rich	Tennis
10	Sunny	No	Rich	Tennis

ค่าสารสนเทศของกลุ่มข้อมูล = I(Cinema,Tennis,Stay in,Shopping) = I(1,2,0,0)

โดยที่ Cinema = 1 รายการ

Tennis = 2 รายการ

Stay in = 0 รายการ

Shopping = 0 รายการ

$$I(S_{Weather=Sunny}) = I(1,2,0,0) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} - \frac{0}{3}\log_2\frac{0}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$I(1,2,0,0) = 0.918$$

2.1.1) หาค่าอินฟอร์เมชันเกนของ Parents โดยพิจารณาข้อมูลทั้ง 3 รายการออกมาได้ข้อมูลดังตาราง 2.20

ตาราง 2.20 ตารางแจกแจงข้อมูลจำแนกตาม Parents เฉพาะ Weather มีค่าเป็น Sunny

Parents	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Yes	1	1	0	0	0
No	2	0	2	0	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Parents

$$E(S_{Weather=Sunny} \rightarrow Parents) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{1}{3} I(1,0,0,0) + \frac{2}{3} I(0,2,0,0)$$

$$I(1,0,0,0) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0.000$$

$$I(0,2,0,0) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0.000$$

$$E(S_{Weather=Sunny} \rightarrow Parents) = \frac{1}{3}0.000 + \frac{2}{3}0.000 = 0.000$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Parents แสดงดังสมการ

$$Gain(S_{Weather=Sunny} \rightarrow Parents) = I(S_{Weather=Sunny}) - E(S_{Weather=Sunny} \rightarrow Parents)$$

$$= 0.918 - 0.000 = 0.918$$

2.1.2) หาค่าอินฟอร์เมชันเกนของ Money โดยพิจารณาข้อมูลทั้ง 3 รายการ แล้วจำแนกตาม Money ที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.21

ตาราง 2.21 ตารางแจกแจงข้อมูลจำแนกตาม Money เฉพาะ Weather มีค่าเป็น Sunny

Money	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Rich	3	1	2	0	0
Poor	0	0	0	0	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Money

$$E(S_{Weather=Sunny} \rightarrow Money) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{3}{3} I(1,2,0,0) + \frac{0}{3} I(0,0,0,0)$$

$$I(1,2,0,0) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} - \frac{0}{3}\log_2\frac{0}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0.918$$

$$I(0,0,0,0) = -\frac{0}{0}\log_2\frac{0}{0} - \frac{0}{0}\log_2\frac{0}{0} - \frac{0}{0}\log_2\frac{0}{0} - \frac{0}{0}\log_2\frac{0}{0} = 0.000$$

$$E(S_{Weather=Sunny} \rightarrow Money) = \frac{3}{3}0.918 + \frac{0}{3}0.000 = 0.918$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Money แสดงดังสมการ

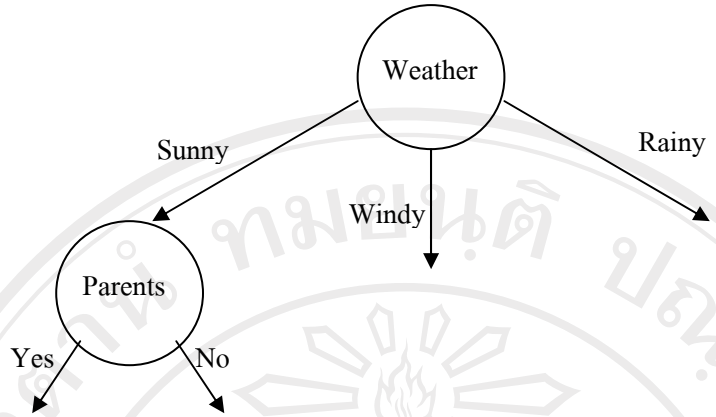
$$Gain(S_{Weather=Sunny} \rightarrow Money) = I(S_{Weather=Sunny}) - E(S_{Weather=Sunny} \rightarrow Money)$$
$$= 0.918 - 0.918 = 0.000$$

จากกิ่งแรก Weather มีค่าเป็น Sunny เมื่อหา ค่าอินฟอร์เมชันเกนของคุณลักษณะครบทั้ง 2 ตัวแล้วสรุปผลได้ดังนี้

$$Gain(S_{Weather=Sunny} \rightarrow Parents) = 0.918$$

$$Gain(S_{Weather=Sunny} \rightarrow Money) = 0.000$$

เมื่อเปรียบเทียบแล้วจะเห็นว่าค่าอินฟอร์เมชันเกน $Gain(S_{Weather=Sunny} \rightarrow Parents)$ มีค่ามากที่สุด ซึ่งเราจะใช้ Parents เป็นตัวแบ่งข้อมูลในลำดับถัดไป ซึ่งสามารถแสดงได้ดังรูป 2.8



รูป 2.8 รูปคุณลักษณะ Parents ถูกเลือกให้เป็นโหนดถัดมาจากกิ่งที่ 1 ของ Weather

2.2) จากกิ่งที่สอง Weather มีค่าเป็น Windy เมื่อพิจารณาแล้วมีทั้งหมด 4 รายการ ดังตาราง 2.22

ตาราง 2.22 ตารางข้อมูลตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์ เฉพาะ Weather มีค่าเป็น Windy

Weekend	Weather	Parent	Money	Activity
3	Windy	Yes	Rich	Cinema
7	Windy	No	Poor	Cinema
8	Windy	No	Rich	Shopping
9	Windy	Yes	Rich	Cinema

ค่าสารสนเทศของกลุ่มข้อมูล = I(Cinema,Tennis,Stay in,Shopping) = I(3,0,0,1)

โดยที่ Cinema = 3 รายการ

Tennis = 0 รายการ

Stay in = 0 รายการ

Shopping = 1 รายการ

$$I(S_{Weather=Windy}) = I(3,0,0,1) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{0}{4}\log_2\frac{0}{4} - \frac{0}{4}\log_2\frac{0}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$I(3,0,0,1) = 0.811$$

2.2.1) หาค่าอินฟอร์เมชันเกนของ Parents โดยพิจารณาข้อมูลทั้ง 4 รายการออกมาได้ข้อมูลดังตาราง 2.23

ตาราง 2.23 ตารางแจกแจงข้อมูลจำแนกตาม Parents เฉพาะ Weather มีค่าเป็น Windy

Parents	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Yes	2	2	0	0	0
No	2	1	0	0	1

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Parents

$$E(S_{Weather=Windy} \rightarrow Parents) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{2}{4} I(2,0,0,0) + \frac{2}{4} I(1,0,0,1)$$

$$I(2,0,0,0) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0.000$$

$$I(1,0,0,1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1.000$$

$$E(S_{Weather=Windy} \rightarrow Parents) = \frac{2}{4}0.000 + \frac{2}{4}1.000 = 0.500$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Parents แสดงดังสมการ

$$Gain(S_{Weather=Windy} \rightarrow Parents) = I(S_{Weather=Windy}) - E(S_{Weather=Windy} \rightarrow Parents)$$
$$= 0.811 - 0.500 = 0.311$$

2.2.2) หาค่าอินฟอร์เมชันเกนของ Money โดยพิจารณาข้อมูลทั้ง 4 รายการ แล้วจำแนกตาม Money ที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.24

ตาราง 2.24 ตารางแจกแจงข้อมูลจำแนกตาม Money เฉพาะ Weather มีค่าเป็น Windy

Money	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Rich	3	2	0	0	1
Poor	1	1	0	0	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Money

$$E(S_{Weather=Windy} \rightarrow Money) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{3}{4} I(2,0,0,1) + \frac{1}{4} I(1,0,0,0)$$

$$I(2,1,0,0) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{0}{3}\log_2\frac{0}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0.918$$

$$I(1,0,0,0) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0.000$$

$$E(S_{Weather=Windy} \rightarrow Money) = \frac{3}{4}0.918 + \frac{1}{4}0 = 0.689$$

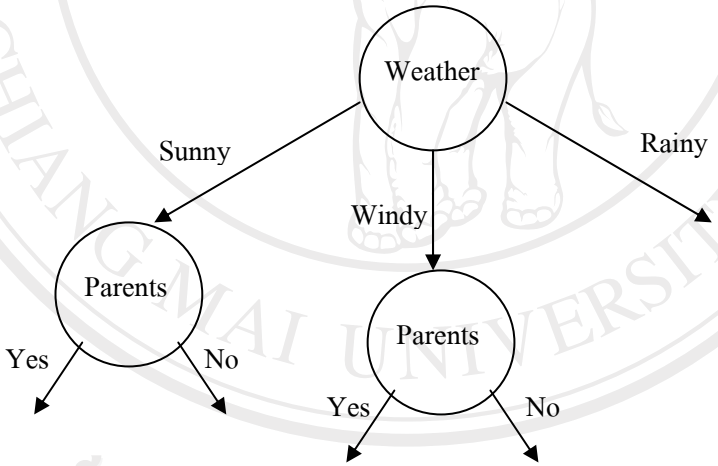
ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Money แสดงดังสมการ

$$Gain(S_{Weather=Windy} \rightarrow Money) = I(S_{Weather=Windy}) - E(S_{Weather=Windy} \rightarrow Money)$$
$$= 0.811 - 0.689 = 0.122$$

จากกิ่งที่สอง Weather มีค่าเป็น Windy เมื่อหา ค่าอินฟอร์เมชันเกนของคุณลักษณะครบทั้ง 2 ตัวแล้วสรุปผลได้ดังนี้

$$Gain(S_{Weather=Windy} \rightarrow Parents) = 0.311$$
$$Gain(S_{Weather=Windy} \rightarrow Money) = 0.122$$

เมื่อเปรียบเทียบแล้วจะเห็นว่าค่าอินฟอร์เมชันเกน $Gain(S_{Weather=Windy} \rightarrow Parents)$ มีค่ามากที่สุด ซึ่งเราจะใช้ Parents เป็นตัวแบ่งข้อมูลในลำดับถัดไป ซึ่งสามารถแสดงได้ดังรูป 2.9



รูป 2.9 รูปคุณลักษณะ Parents ถูกเลือกให้เป็นโหนดถัดมาจากกิ่งที่ 2 ของ Weather

2.3) จากกิ่งที่สาม Weather มีค่าเป็น Rainy เมื่อพิจารณาแล้วมีทั้งหมด 3 รายการ ดังตาราง 2.25

ตาราง 2.25 ตารางข้อมูลตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์ เฉพาะ Weather มีค่าเป็น Rainy

Weekend	Weather	Parent	Money	Activity
4	Rainy	Yes	Poor	Cinema
5	Rainy	No	Rich	Stay in
6	Rainy	Yes	Poor	Cinema

ค่าสารสนเทศของกลุ่มข้อมูล = I(Cinema,Tennis,Stay in,Shopping) = I(2,0,1,0)

โดยที่ Cinema = 2 รายการ

Tennis = 0 รายการ

Stay in = 1 รายการ

Shopping = 0 รายการ

$$I(S_{Weather=Rainy}) = I(2,0,1,0) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{0}{3}\log_2\frac{0}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$I(2,0,1,0) = 0.918$$

2.3.1) หาค่าอินฟอร์เมชันเกนของ Parents โดยพิจารณาข้อมูลทั้ง 3 รายการออกมาได้ข้อมูลดังตาราง 2.26

ตาราง 2.26 ตารางแจกแจงข้อมูลจำแนกตาม Parents เฉพาะ Weather มีค่าเป็น Rainy

Parents	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Yes	2	2	0	0	0
No	1	0	0	1	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Parents

$$E(S_{Weather=Rainy} \rightarrow Parents) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{2}{3} I(2,0,0,0) + \frac{1}{3} I(0,0,1,0)$$

$$I(2,0,0,0) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0.000$$

$$I(0,0,1,0) = -\frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0.000$$

$$E(S_{Weather=Rainy} \rightarrow Parents) = \frac{2}{3}0.000 + \frac{1}{3}0.000 = 0.000$$

ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Parents แสดงดังสมการ

$$Gain(S_{Weather=Rainy} \rightarrow Parents) = I(S_{Weather=Rainy}) - E(S_{Weather=Rainy} \rightarrow Parents)$$

$$= 0.918 - 0.000 = 0.918$$

2.3.2) หาค่าอินฟอร์เมชันเกนของ Money โดยพิจารณาข้อมูลทั้ง 3 รายการ แล้วจำแนกตาม Money ที่มี 2 กลุ่ม ออกมาได้ข้อมูลดังตาราง 2.27

ตาราง 2.27 ตารางแจกแจงข้อมูลจำแนกตาม Money เฉพาะ Weather มีค่าเป็น Rainy

Money	จำนวนรวม	Cinema	Tennis	Stay in	Shopping
Rich	1	0	0	1	0
Poor	2	2	0	0	0

หาค่าคาดคะเนของข้อมูลที่แยกโดยการใช้คุณลักษณะ Money

$$E(S_{Weather=Rainy} \rightarrow Money) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v) = \frac{1}{3} I(0,0,1,0) + \frac{2}{3} I(2,0,0,0)$$

$$I(0,0,1,0) = -\frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0.000$$

$$I(2,0,0,0) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0.000$$

$$E(S_{Weather=Rainy} \rightarrow Money) = \frac{1}{3}0.000 + \frac{2}{3}0 = 0.000$$

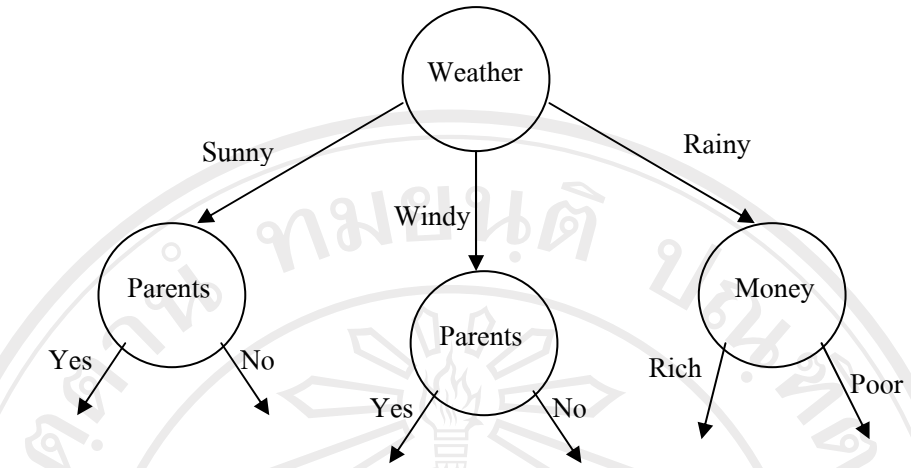
ดังนั้น ค่าอินฟอร์เมชันเกนของคุณลักษณะ Money แสดงดังสมการ

$$Gain(S_{Weather=Rainy} \rightarrow Money) = I(S_{Weather=Rainy}) - E(S_{Weather=Rainy} \rightarrow Money)$$
$$= 0.918 - 0.000 = 0.918$$

จากกิ่งที่สาม Weather มีค่าเป็น Rainy เมื่อหา ค่าอินฟอร์เมชันเกนของคุณลักษณะครบทั้ง 2 ตัวแล้วสรุปผลได้ดังนี้

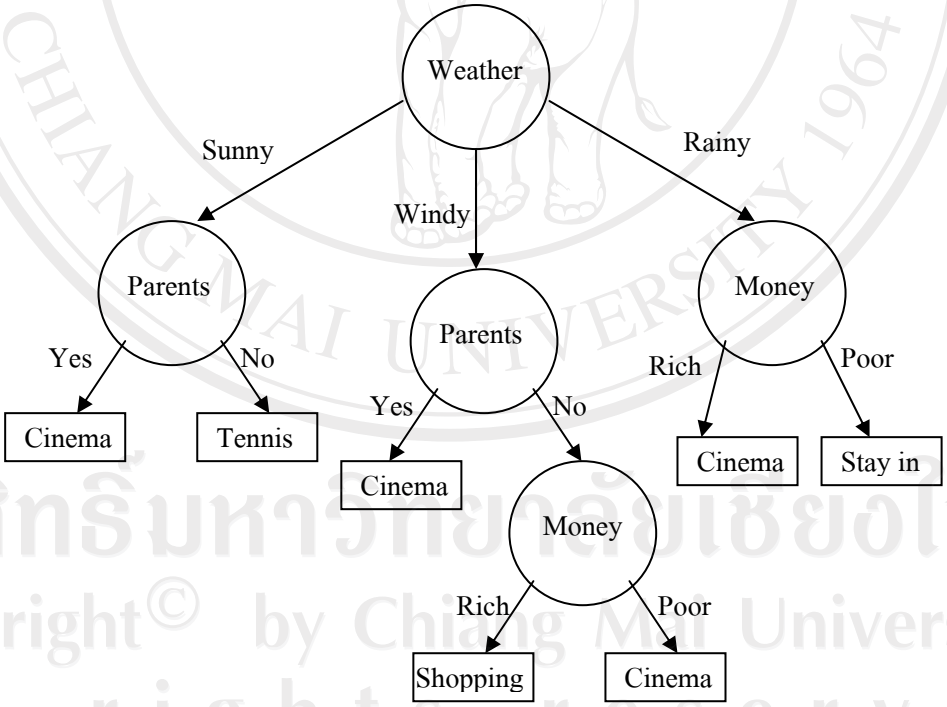
$$Gain(S_{Weather=Rainy} \rightarrow Parents) = 0.918$$
$$Gain(S_{Weather=Rainy} \rightarrow Money) = 0.918$$

เมื่อเปรียบเทียบแล้วจะเห็นว่าค่าอินฟอร์เมชันเกน $Gain(S_{Weather=Windy} \rightarrow Parents)$ มีค่าเท่ากัน ซึ่งเราสามารถเลือกใช้ Parents หรือ Money ตัวใดตัวหนึ่งก็ได้ ให้เป็นตัวแบ่งข้อมูลในลำดับถัดไป จากตัวอย่างนี้ได้เลือก Money เป็นตัวแบ่งข้อมูลในลำดับถัดไป ซึ่งสามารถแสดงได้ดังรูป 2.10



รูป 2.10 รูปคุณลักษณะ Money ถูกเลือกให้เป็นโหนดถัดมาจากกิ่งที่ 3 ของ Weather

ท้ายที่สุดสามารถแสดงรูปของต้นไม้ตัดสินใจกรณีตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์ ได้ดังรูป 2.11



รูป 2.11 รูปต้นไม้ตัดสินใจกรณีตัวอย่างการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์

จากรูป 2.11 รูปต้นไม้ตัดสินใจกรณีตัวอย่างข้อมูลการทำกิจกรรมต่างๆ ในวันหยุดสุดสัปดาห์ สามารถนำมาเขียนในรูป If-Then ได้ดังนี้

IF Weather=”Sunny” and Parents=”Yes” THEN Decision=”Cinema”

IF Weather=”Sunny” and Parents=”No” THEN Decision=”Tennis”

IF Weather=”Windy” and Parents=”Yes” THEN Decision=”Cinema”

IF Weather=”Windy” and Parents=”No” and Money=”Rich” THEN

Decision=”Shopping”

IF Weather=”Windy” and Parents=”No” and Money=”Poor” THEN Decision=”Cinema”

IF Weather=”Rainy” and Money=”Rich” THEN Decision=”Cinema”

IF Weather=”Rainy” and Money=”Poor” THEN Decision=”Stay in”

เอกสารและงานวิจัย ที่เกี่ยวข้องกับการศึกษาในครั้งนี้พบว่า ปรีชา ยามันสะบีดีน, บุญเสริม กิจศิริกุล, ปิยะวัฒน์ จิระพงษ์สุวรรณ และ ประสงค์ ประณีตพลกรัง (2551) ได้ใช้เทคนิคต้นไม้ตัดสินใจ เพื่อสร้างต้นแบบสำหรับจำแนกคุณลักษณะของนักศึกษาเพื่อใช้ทำนายสถานภาพของนักศึกษา เพื่อช่วยในการวางแผนการศึกษาให้กับนักศึกษาให้เหมาะสมกับตัวนักศึกษา และช่วยในการตัดสินใจการดำเนินงานของมหาวิทยาลัย ด้านการแนะแนวการศึกษา การจัดการสอนเสริม และกิจกรรมการเรียนให้กับนักศึกษา