

การจำแนกรูปแบบพันธุกรรมของสายพันธุ์ไวรัส SARS-CoV-2
(Identifying genomic patterns across SARS-CoV-2 lineages)

นายยงยุทธ อุ่นยาว
รหัส 620510522

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
มหาวิทยาลัยเชียงใหม่
ภาคการศึกษาที่ 1 ปีการศึกษา 2565

การจำแนกรูปแบบพันธุกรรมของสายพันธุ์ไวรัส SARS-CoV-2
(Identifying genomic patterns across SARS-CoV-2 lineages)

ยงยุทธ อุ๋นยาว

ได้รับการพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต
สาขาคณิตศาสตร์

คณะกรรมการควบคุมการค้นคว้าอิสระ:

..... นัทธ์รัตน์ ยิงทวิสิทธิกุล ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นัทธ์รัตน์ ยิงทวิสิทธิกุล)

.....  กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.เบน วงศ์สายใจ)

วันที่ 7 เดือน พฤศจิกายน พ.ศ. 2565

การจำแนกรูปแบบพันธุกรรมของสายพันธุ์ไวรัส SARS-CoV-2
(Identifying genomic patterns across SARS-CoV-2 lineages)

นายยงยุทธ อุ่นยาว
รหัส 620510522

งานค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

กิตติกรรมประกาศ

งานค้นคว้าอิสระฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ในกระบวนวิชา 206499 ซึ่งผู้ค้นคว้าได้ศึกษา เรื่อง การจำแนกรูปแบบพันธุกรรมของสายพันธุ์ไวรัส SARS-CoV-2

ผู้ค้นคว้าขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.หทัยรัตน์ ยิ่งทวีสิทธิกุล อาจารย์ที่ปรึกษางานค้นคว้าอิสระที่คอยช่วยเหลือให้คำอธิบายในส่วนที่ผู้ค้นคว้าไม่เข้าใจ ให้ความรู้ ให้คำแนะนำ และให้คำปรึกษาตลอดจนตรวจสอบความถูกต้องพิจารณาแก้ไข จนทำให้การศึกษาค้นคว้าอิสระนี้เสร็จสมบูรณ์ด้วยดี

ขอขอบคุณอาจารย์ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ทุกท่าน ที่ได้อบรม สั่งสอน และให้ความรู้ทั้งในและนอกตำราเรียนให้แก่ผู้ค้นคว้า อันเป็นประโยชน์และสามารถนำมาประยุกต์ใช้ในการศึกษาค้นคว้าอิสระครั้งนี้ ขอขอบคุณนายภานุพงษ์ เกียงคำ และเพื่อน ๆ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ และผู้เกี่ยวข้องที่ช่วยเหลือสนับสนุนให้คำปรึกษาและช่วยเหลือขณะที่ผู้ค้นคว้ากำลังศึกษาค้นคว้าอิสระชิ้นนี้

สุดท้ายนี้ ขอขอบคุณบิดา มารดา และครอบครัว รวมทั้งขอขอบคุณไปยังทุกท่านที่มีส่วนเกี่ยวข้องกับงานค้นคว้าอิสระ ฉบับนี้ ซึ่งผู้ค้นคว้าไม่ได้เอ่ยนาม มา ณ ที่นี้ด้วย

หากผลงานการค้นคว้าอิสระฉบับนี้ได้รับคำชื่นชมหรือเกิดคุณประโยชน์ประการใด ผู้ค้นคว้าขอมอบ คำชื่นชม และคุณความดีนั้นแก่ ครู อาจารย์ทุกท่านที่ได้อบรมสั่งสอน ในทางตรงกันข้ามหากผลงานค้นคว้าอิสระนี้ เกิดข้อผิดพลาดประการใด ผู้ค้นคว้าขออภัยไว้ ณ ที่นี้ และขอน้อมรับความผิดพลาดไว้เพียงผู้เดียว

ยังยุทธ อุ่นยาว

หัวข้อ (ภาษาไทย)	การจำแนกรูปแบบพันธุกรรมของสายพันธุ์ไวรัส SARS-CoV-2
(ภาษาอังกฤษ)	Identifying genomic patterns across SARS-CoV-2 lineages
ชื่อผู้ทำการค้นคว้าอิสระ	นายยงยุทธ อุ่นยาว รหัสนักศึกษา 620510522
ชื่ออาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.หทัยรัตน์ ยิ่งทวีสิทธิกุล
ชื่อกรรมการสอบ	ผู้ช่วยศาสตราจารย์ ดร.เบน วงศ์สายใจ

บทคัดย่อ

ในงานค้นคว้าอิสระนี้ ผู้ศึกษาได้ศึกษาการสร้างโมเดลต้นไม้ตัดสินใจและ single linkage agglomerative hierarchical algorithms เพื่อจำลองการจำแนกพันธุกรรมของเชื้อไวรัส SARS-CoV-2 และเพื่อจัดกลุ่มผู้ป่วยโดยอ้างอิงจากข้อมูลสายพันธุกรรม

Abstract

In this independent study, we study decision trees and single linkage agglomerative hierarchical algorithms for identifying genomic patterns across SARS-CoV-2 lineages and determining patient clusters based on genomic information.

สารบัญ

กิตติกรรมประกาศ	i
บทคัดย่อ	ii
สารบัญ	iii
1 บทนำ (Introduction)	1
1.1 ที่มาและความสำคัญ	1
1.2 วัตถุประสงค์	2
1.3 ประโยชน์ที่คาดว่าจะได้รับ	2
1.4 ขอบเขตและวิธีการดำเนินการ	2
2 ความรู้พื้นฐาน (Preliminaries)	3
2.1 ต้นไม้ตัดสินใจ (Decision Tree)	3
2.1.1 ส่วนประกอบของผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจ	3
2.1.2 เครื่องมือที่ใช้ในการสร้างต้นไม้ตัดสินใจ	4
2.1.3 ขั้นตอนวิธีทำต้นไม้ตัดสินใจ	5
2.2 Agglomerative Hierarchical Cluster Analysis	126
3 ผลการศึกษา (Results)	132
4 สรุปผลการศึกษา (Conclusion)	139
บรรณานุกรม	140
ภาคผนวก	142

บทที่ 1

บทนำ (Introduction)

1.1 ที่มาและความสำคัญ

การระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 (Coronavirus disease 2019 หรือ COVID-19) เริ่มขึ้นตั้งแต่วันที่ 30 ธันวาคม พ.ศ. 2562 ที่มีรายงานกลุ่มผู้ป่วยด้วยโรคปอดอักเสบไม่ทราบสาเหตุเกิดขึ้นที่เมืองอู่ฮั่น มณฑลหูเป่ย์ สาธารณรัฐประชาชนจีน ต่อมาในวันที่ 7 มกราคม พ.ศ. 2563 ศูนย์ควบคุมและป้องกันโรคแห่งสาธารณรัฐประชาชนจีน (Chinese Center for Disease Control and Prevention; China CDC) ได้ออกมาประกาศอย่างเป็นทางการว่าสาเหตุของโรคปอดอักเสบดังกล่าวนี้ เกิดจากเชื้อไวรัสโคโรนาชนิดใหม่ ที่ไม่ใช่ไวรัสซาร์ส-โควี (Severe Acute Respiratory Syndrome-Coronavirus; SARS-CoV) ซึ่งเป็นไวรัสที่ก่อให้เกิดโรคทางเดินหายใจเฉียบพลันรุนแรงหรือโรคซาร์ส และไม่ใช่ไวรัสเมอร์ส-โควี (Middle East Respiratory Syndrome-Coronavirus; MERS-CoV) ซึ่งทำให้เกิดโรคทางเดินหายใจตะวันออกกลางหรือโรคเมอร์ส จึงตั้งชื่อไวรัสชนิดนี้ในตอนแรกว่า 2019-Novel Corona Virus (2019-nCoV) และในวันที่ 11 กุมภาพันธ์ พ.ศ. 2563 ทาง The International Committee on Taxonomy of Viruses (ICTV) ได้ออกมาประกาศชื่อใหม่ของไวรัสว่าไวรัสซาร์ส-โควี-2 (Severe Acute Respiratory Syndrome-Coronavirus-2; SARS-CoV-2) และตั้งชื่อโรคที่เกิดจากการติดเชื้อไวรัสดังกล่าวว่า โรคติดเชื้อไวรัสโคโรนา 2019 หรือโรคโควิด-19 (Coronavirus disease 2019; COVID-19) การระบาดของโรคโควิด-19 ได้กระจายไปอย่างรวดเร็วจากเริ่มแรกมีการระบาดแค่ในสาธารณรัฐประชาชนจีนเพียงประเทศเดียว ต่อมาได้กระจายไปยังประเทศและภูมิภาคอื่นๆ ทั่วโลก โดยในวันที่ 11 มีนาคม พ.ศ. 2563 องค์การอนามัยโลกได้ประกาศให้โรคโควิด-19 เป็นโรคระบาดใหญ่ทั่วโลก (pandemic) ไวรัสโคโรนาจัดเป็นไวรัสที่มีขนาดใหญ่ที่สุดในกลุ่มไวรัสที่มีสารพันธุกรรมอาร์เอ็นเอ (RNA virus) จากการศึกษาลำดับพันธุกรรมและลำดับนิวคลีโอไทด์ทั้งหมดในจีโนม (whole genome sequencing) ของเชื้อไวรัสซาร์ส-โควี-2 พบว่ามีลำดับนิวคลีโอไทด์ทั้งหมดประมาณ 29,903 คู่เบส ไวรัสซาร์สที่พบก่อนหน้านี้โดยเมื่อวิเคราะห์ความสัมพันธ์เชิงวิวัฒนาการ (phylogenetic tree) พบว่าเชื้อไวรัสซาร์ส-โควี-2 มีความใกล้เคียงกับเชื้อไวรัสซาร์สของค้างคาว (bat SARS-like coronavirus) ที่พบในประเทศจีน ประมาณร้อยละ 96 และไวรัสซาร์สของมนุษย์ (SARS-CoV) ประมาณร้อยละ 80

เนื่องจากไวรัสซาร์ส-โควี-2 เป็นไวรัสชนิดอาร์เอ็นเอ (RNA virus) จึงมีโอกาสกลายพันธุ์ (mutation) ได้สูงมาก เริ่มแรกได้มีการระบาดของโควิดสายพันธุ์ S (Serine) และโควิดสายพันธุ์ L (Leucine) เป็นสายพันธุ์ดั้งเดิมที่พบ

ในเมืองอู่ฮั่น ประเทศจีน โดยโคโรนาไวรัสสายพันธุ์ S ระบาดระลอกแรกในไทย ในเดือนมีนาคม 2020 จากนั้นสายพันธุ์ L สามารถแพร่กระจายได้ดีกว่าสายพันธุ์ S ส่งผลให้สายพันธุ์ L แพร่ขยายไปอย่างรวดเร็วในยุโรป และกลายพันธุ์เป็นสายพันธุ์ V (Valine) และสายพันธุ์ G (Glycine) ต่อมาสายพันธุ์ G เมื่อสามารถแพร่ขยายต่อไป ได้กลายพันธุ์เป็นสายพันธุ์ GH, GR, GV ,GRY, GK, GKA และ GRK ในปัจจุบันการกลายพันธุ์ที่มีความสำคัญทางระบาดวิทยา คือ การกลายพันธุ์ของสารพันธุกรรมเพียงตำแหน่งเดียว (single nucleotide mutation) ยกตัวอย่างการกลายพันธุ์ในลำดับเบสของสายพันธุ์ไวรัสซาร์ส-โควี-2 ตรงตำแหน่งที่ 8752 ของสายพันธุ์ L มีการเปลี่ยนแปลงจาก C เป็น T ตรงตำแหน่งที่ 8752 ของสายพันธุ์ S หรือที่เรียกว่าการกลายพันธุ์แบบ C8752T โดยไวรัสซาร์ส-โควี-2

1.2 วัตถุประสงค์

1. ศึกษาการสร้างโมเดลจำลองการจำแนกพันธุกรรมของเชื้อไวรัส SARS-CoV-2
2. วิเคราะห์ความแตกต่างของพันธุกรรมของเชื้อไวรัส SARS-CoV-2 ในแต่ละสายพันธุ์
3. ศึกษาการแบ่งกลุ่มข้อมูล เพื่อหากกลุ่มผู้ป่วยที่สายพันธุ์พันธุกรรมใกล้เคียงกัน

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถจำลองการสร้างเครื่องมือในการจำแนกสายพันธุ์ของเชื้อไวรัส SARS-CoV-2
2. สามารถหาความแตกต่างของสายพันธุ์พันธุกรรมของแต่ละสายพันธุ์
3. สามารถแบ่งกลุ่มผู้ป่วยที่มีสายพันธุ์พันธุกรรมใกล้เคียงกัน

1.4 ขอบเขตและวิธีการดำเนินการ

1. ศึกษาความรู้พื้นฐาน
2. สร้างโมเดลจำแนกสายพันธุ์ไวรัส SARS-CoV-2 ที่เกิดขึ้นในประเทศไทย
3. จำลองสถานการณ์การติดเชื้อในกลุ่มคน พร้อมทั้งจำแนกกลุ่มผู้ป่วยที่มีสายพันธุ์พันธุกรรมใกล้เคียงกัน

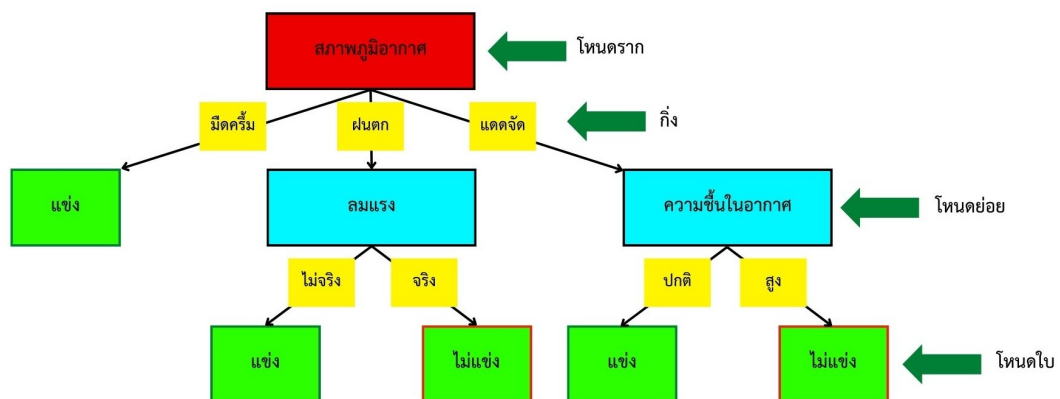
บทที่ 2

ความรู้พื้นฐาน (Preliminaries)

2.1 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree) เป็นการเรียนรู้โดยการจำแนกประเภท (Classification) ข้อมูลออกเป็นกลุ่ม (class) ต่าง ๆ โดยใช้คุณลักษณะ (attribute) ข้อมูลในการจำแนกประเภท ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่า คุณลักษณะใดเป็นตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวมีความสำคัญมากน้อยต่างกันอย่างไร

ต้นไม้ตัดสินใจเป็นอัลกอริทึมที่เหมาะสมนำมาใช้สำหรับการจำแนก ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ที่เป็นโหนดราก (Root Node) อยู่ด้านบนสุดและลำดับล่างลงมา จะเรียกว่าโหนดย่อย (Child Node) และโหนดย่อยที่อยู่ล่างสุดจะเรียกว่าโหนดใบ (Leaf Node) การทำงานจะเริ่มตั้งแต่โหนดราก โดยการนำข้อมูลมาตรวจสอบเงื่อนไขกับโหนดราก หากข้อมูลที่จะจำแนกเป็นไปตามเงื่อนไข ก็จะลงไปตามซ้ายของโหนดราก หากไม่เป็นไปตามเงื่อนไขก็จะลงไปตามขวาของโหนดราก จากนั้น ข้อมูลที่เข้ามาจะถูกตรวจสอบเงื่อนไขกับ โหนดย่อย ลำดับล่างลงมาเมื่อตรวจสอบจนถึงโหนดใบ ถือเป็นการสิ้นสุดข้อมูลที่ถูกจำแนกโดยโหนดใบ



รูปที่ 2.1: แสดงส่วนประกอบของผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจ

2.1.1. ส่วนประกอบของผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจ

- โหนดภายใน (internal node) คือ คุณลักษณะต่าง ๆ ของข้อมูล ซึ่งเมื่อข้อมูลใด ๆ ตกลงมาที่โหนด จะใช้คุณลักษณะนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่า โหนดราก

- กิ่ง (branch, link) เป็นค่าของคุณลักษณะในโหนดภายในที่แตกกิ่งนี้ออกมา ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าของคุณลักษณะในโหนดภายในนั้น
- โหนดใบ (leaf node) คือ กลุ่มต่าง ๆ ซึ่งเป็นผลลัพธ์ในการจำแนกประเภทข้อมูล

2.1.2. เครื่องมือที่ใช้ในการสร้างต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจ จะใช้การกระจายของข้อมูลเป็นตัววัด มี 2 วิธี คือ

วิธีที่ 1 : Gini Index

$$Gini_{split}(T) = \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i)$$

$$\text{โดยที่ } Gini(t_i) = 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2$$

T = โหนดภายใน หรือคุณลักษณะของข้อมูล

t_i = กิ่งที่ i หรือค่าของคุณลักษณะในโหนดภายในที่ i

M = จำนวนกิ่งของคุณลักษณะของข้อมูล

n_i = จำนวนข้อมูลในกิ่งที่ i

N = จำนวนกลุ่ม หรือผลลัพธ์ในการจำแนกประเภทข้อมูล

s_k = จำนวนข้อมูลในกลุ่มที่ k ของแต่ละกิ่ง

$Gini_{split}(T) = 0$ หมายถึง ความเท่าเทียมกันอย่างสมบูรณ์ และ $Gini_{split}(T) = 0.5$ หมายถึงความเหลื่อมล้ำอย่างสมบูรณ์

วิธีที่ 2 : Information Gain(IG)

$$IG(T) = Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i)$$

$$\text{โดยที่ } Entropy(T) = - \sum_{k=1}^N \left[\frac{\frac{r_k}{N}}{\sum_{l=1}^N r_l} \log_2 \left(\frac{\frac{r_k}{N}}{\sum_{l=1}^N r_l} \right) \right]$$

$$Entropy(t_i) = - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right]$$

T = โหนดภายใน หรือคุณลักษณะของข้อมูล

t_i = กิ่งที่ i หรือค่าของคุณลักษณะในโหนดภายในที่ i

M = จำนวนกิ่งของคุณลักษณะของข้อมูล

n_i = จำนวนข้อมูลในกิ่งที่ i

N = จำนวนกลุ่ม หรือผลลัพธ์ในการจำแนกประเภทข้อมูล

r_k = จำนวนข้อมูลในกลุ่มที่ k ของคุณลักษณะของข้อมูล

s_k = จำนวนข้อมูลในกลุ่มที่ k ของแต่ละกิ่ง

ในที่นี้ กำหนดให้ $0 \times \log_2 0 = 0$

$IG(T) = 0$ หมายถึง ความเหลื่อมล้ำอย่างสมบูรณ์ และ $IG(T) = 1$ หมายถึงความเท่าเทียมกันอย่างสมบูรณ์

2.1.3. ขั้นตอนวิธีทำต้นไม้ตัดสินใจ

1. ต้นไม้ตัดสินใจสร้างโดยวิธีแบบ top-down recursive
2. ในแต่ละคุณลักษณะควรอยู่ในรูปของ Categorical คือ ข้อมูลชนิดกลุ่ม หากเป็นข้อมูลที่อยู่ในรูป Continuous หรือ Numeric เป็นข้อมูลมีความต่อเนื่องกัน ควรทำแบ่งข้อมูลให้เป็นกลุ่มก่อน
3. คำนวณค่าของแต่ละคุณลักษณะเทียบกับกลุ่ม ตามเครื่องมือที่ใช้ในการวัด
4. เลือกโหนดภายในตามเงื่อนไขของเครื่องมือที่ใช้ในการวัด
5. พิจารณาแต่ละโหนดย่อยของโหนดภายในที่เลือก ถ้าภายในโหนดย่อยนั้นมีข้อมูลเป็นกลุ่มเดียวกัน จะถือว่าสิ้นสุดการคำนวณในโหนดย่อยนั้น จะไม่แตกกิ่งออกไปต่อ
6. ถ้าภายในโหนดย่อยของโหนดภายในที่เลือก มีข้อมูลของกลุ่มต่างกัน ให้วนกลับไปทำซ้ำในข้อ 3 - 5 จนกว่าจะไม่สามารถแตกกิ่งออกไปต่อได้

ตัวอย่าง 2.1 ตัวอย่าง การสร้างต้นไม้ตัดสินใจโดยใช้ Gini Index โดยสมมติมีข้อมูลของสภาพอากาศ 14 วันย้อนหลัง สำหรับการจัดแข่งขันกีฬา ดังตารางต่อไปนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
1	แดดจัด	ร้อน	สูง	ไม่จริง	ไม่แข่ง
2	แดดจัด	ร้อน	สูง	จริง	ไม่แข่ง
3	มีดครึ้ม	ร้อน	สูง	ไม่จริง	แข่ง
4	ฝนตก	อบอุ่น	สูง	ไม่จริง	แข่ง
5	ฝนตก	เย็น	ปกติ	ไม่จริง	แข่ง
6	ฝนตก	เย็น	ปกติ	จริง	ไม่แข่ง
7	มีดครึ้ม	เย็น	ปกติ	จริง	แข่ง
8	แดดจัด	อบอุ่น	สูง	ไม่จริง	ไม่แข่ง
9	แดดจัด	อบอุ่น	ปกติ	ไม่จริง	แข่ง
10	ฝนตก	อบอุ่น	ปกติ	ไม่จริง	แข่ง
11	แดดจัด	อบอุ่น	ปกติ	จริง	แข่ง
12	มีดครึ้ม	อบอุ่น	สูง	จริง	แข่ง
13	มีดครึ้ม	ร้อน	ปกติ	ไม่จริง	แข่ง
14	ฝนตก	อบอุ่น	สูง	จริง	ไม่แข่ง

ตารางที่ 2.1: แสดงข้อมูลสภาพอากาศ

จากตาราง ประกอบด้วย 5 คุณลักษณะ คือ

- สภาพภูมิอากาศ ประกอบด้วย 3 ค่า คือ แดดจัด, มีดครึ้ม, ฝนตก
- อุณหภูมิ ประกอบด้วย 3 ค่า คือ ร้อน, อบอุ่น, เย็น
- ความชื้นในอากาศ ประกอบด้วย 2 ค่า คือ สูง, ปกติ
- ลมแรง ประกอบด้วย 2 ค่า คือ จริง, ไม่จริง
- แข่งขัน ซึ่งเป็นกลุ่ม ประกอบด้วย 2 ค่า คือ แข่ง, ไม่แข่ง

ต่อไปจะแสดงการคำนวณค่าแต่ละคุณลักษณะเทียบกับกลุ่ม เพื่อหาคุณลักษณะที่มีค่า Gini Index น้อยที่สุด มาเป็นโหนดรากของต้นไม้ตัดสินใจ ดังนี้

1. คำนวณค่า Gini Index ของคุณลักษณะสภาพภูมิอากาศ

ให้ T = สภาพภูมิอากาศ

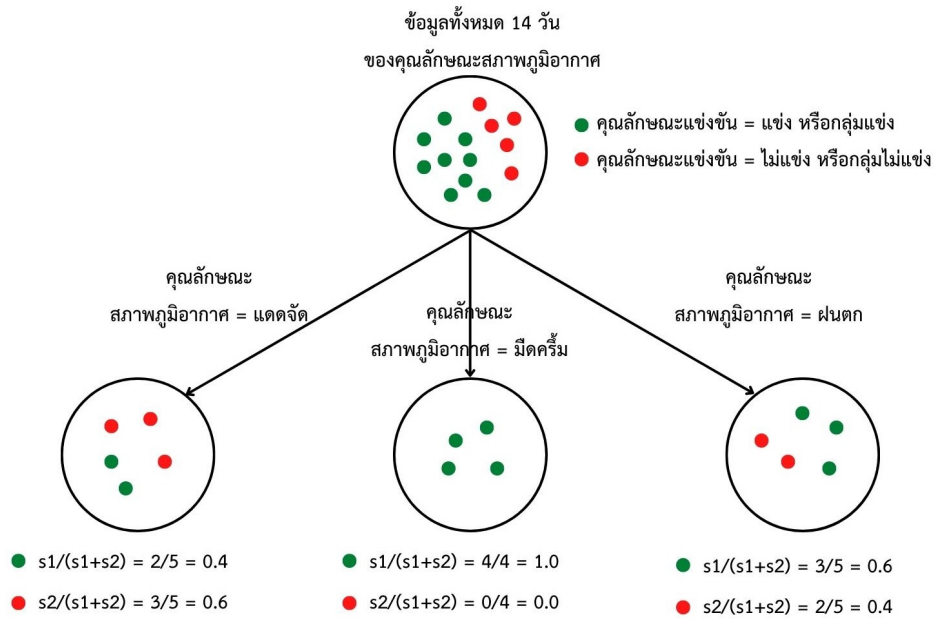
$$M = \text{จำนวนกิ่ง(แดดจัด, มีดครึ้ม, ฝนตก)} = 3$$

$$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (แดดจัด)} = 5$$

$$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (มีดครึ้ม)} = 4$$

$$n_3 = \text{จำนวนข้อมูลในกิ่งที่ 3 (ฝนตก)} = 5$$

$$N = \text{จำนวนกลุ่ม(แข่ง, ไม่แข่ง)} = 2$$



รูปที่ 2.2: แสดงข้อมูลของคุณลักษณะสภาพภูมิอากาศ

กิ่งที่ 1 (แดดจัด)

t_1 คือ สภาพภูมิอากาศ = แดดจัด

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด

$$\begin{aligned}
 Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{2 + 3} \right)^2 + \left(\frac{3}{2 + 3} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{5} \right)^2 + \left(\frac{3}{5} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{4}{25} \right) + \left(\frac{9}{25} \right) \right] \\
 &= 0.48
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) = 0.48$

กึ่งที่ 2 (มีดครี้ม)

t_2 คือ สภาพภูมิอากาศ = มีดครี้ม

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = มีดครี้ม

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = มีดครี้ม

$$\begin{aligned} Gini(\text{สภาพภูมิอากาศ} = \text{มีดครี้ม}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\ &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\ &= 1 - \left[\left(\frac{4}{4 + 0} \right)^2 + \left(\frac{0}{4 + 0} \right)^2 \right] \\ &= 1 - \left[\left(\frac{4}{0} \right)^2 + \left(\frac{0}{4} \right)^2 \right] \\ &= 1 - [1 + 0] \\ &= 0 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{มีดครี้ม}) = 0$

กึ่งที่ 3 (ฝนตก)

t_3 คือ สภาพภูมิอากาศ = ฝนตก

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก

$$\begin{aligned}
Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\
&= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{3}{3+2} \right)^2 + \left(\frac{2}{3+2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{3}{5} \right)^2 + \left(\frac{2}{5} \right)^2 \right] \\
&= 1 - \left[\left(\frac{9}{25} \right) + \left(\frac{4}{25} \right) \right] \\
&= 0.48
\end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) = 0.48$

จะได้ว่า

$$\begin{aligned}
Gini_{split}(\text{สภาพภูมิอากาศ}) &= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
&= \left(\frac{n_1}{n_1 + n_2 + n_3} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) \\
&\quad + \left(\frac{n_2}{n_1 + n_2 + n_3} \right) Gini(\text{สภาพภูมิอากาศ} = \text{มีดคร่ำม}) \\
&\quad + \left(\frac{n_3}{n_1 + n_2 + n_3} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) \\
&= \left(\frac{5}{5+4+5} \right) (0.48) + \left(\frac{4}{5+4+5} \right) (0) + \left(\frac{5}{5+4+5} \right) (0.48) \\
&= \left(\frac{5}{14} \right) (0.48) + \left(\frac{4}{14} \right) (0) + \left(\frac{5}{14} \right) (0.48) \\
&= 0.34
\end{aligned}$$

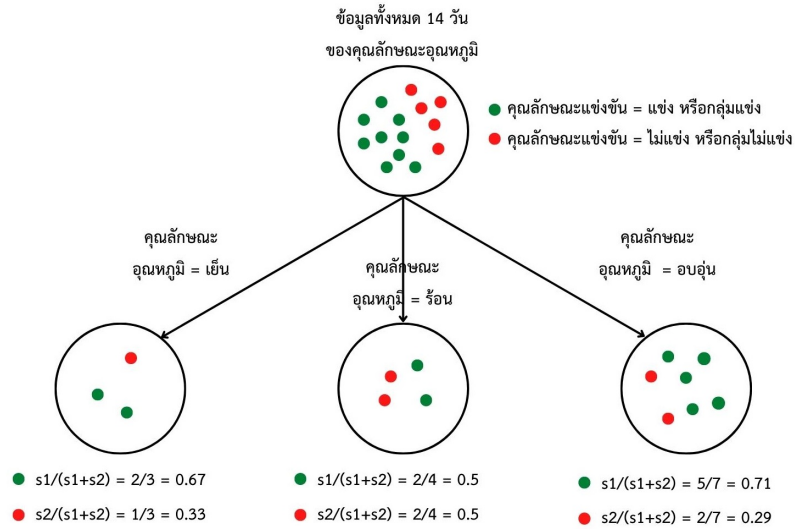
ดังนั้น $Gini_{split}(\text{สภาพภูมิอากาศ}) = 0.34$

2. คำนวณค่า Gini Index ของคุณลักษณะอุณหภูมิ

ให้ $T = \text{อุณหภูมิ}$

$M = \text{จำนวนกิ่ง(เย็น, ร้อน, อุ่น)} = 3$

n_1 = จำนวนข้อมูลในกิ่งที่ 1 (เย็น) = 3
 n_2 = จำนวนข้อมูลในกิ่งที่ 2 (ร้อน) = 4
 n_3 = จำนวนข้อมูลในกิ่งที่ 3 (อบอุ่น) = 7
 N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2



รูปที่ 2.3: แสดงข้อมูลของคุณลักษณะอุณหภูมิ

กิ่งที่ 1 (เย็น)

t_1 คือ อุณหภูมิ = เย็น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = เย็น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = เย็น

$$\begin{aligned}
 Gini(\text{อุณหภูมิ} = \text{เย็น}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{2+1} \right)^2 + \left(\frac{1}{2+1} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{3} \right)^2 + \left(\frac{1}{3} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{4}{9} \right) + \left(\frac{1}{9} \right) \right] \\
 &= 0.44
 \end{aligned}$$

ดังนั้น $Gini(\text{อุณหภูมิ} = \text{เย็น}) = 0.44$

กึ่งที่ 2 (ร้อน)

t_2 คือ อุณหภูมิ = ร้อน

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = ร้อน

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = ร้อน

$$\begin{aligned} Gini(\text{อุณหภูมิ} = \text{ร้อน}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\ &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\ &= 1 - \left[\left(\frac{2}{2+2} \right)^2 + \left(\frac{2}{2+2} \right)^2 \right] \\ &= 1 - \left[\left(\frac{2}{4} \right)^2 + \left(\frac{2}{4} \right)^2 \right] \\ &= 1 - \left[\frac{4}{16} + \frac{4}{16} \right] \\ &= 0.50 \end{aligned}$$

ดังนั้น $Gini(\text{อุณหภูมิ} = \text{ร้อน}) = 0.50$

กึ่งที่ 3 (อบอุ่น)

t_3 คือ อุณหภูมิ = อบอุ่น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = อบอุ่น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = อบอุ่น

$$\begin{aligned}
Gini(\text{อุณหภูมิ} = \text{อบอุ่น}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\
&= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{5}{5 + 2} \right)^2 + \left(\frac{2}{5 + 2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{5}{7} \right)^2 + \left(\frac{2}{7} \right)^2 \right] \\
&= 1 - \left[\frac{25}{49} + \frac{4}{49} \right] \\
&= 0.40
\end{aligned}$$

ดังนั้น $Gini(\text{อุณหภูมิ} = \text{อบอุ่น}) = 0.40$

จะได้ว่า

$$\begin{aligned}
Gini_{split}(\text{อุณหภูมิ}) &= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
&= \left(\frac{n_1}{n_1 + n_2 + n_3} \right) Gini(\text{อุณหภูมิ} = \text{เย็น}) \\
&\quad + \left(\frac{n_2}{n_1 + n_2 + n_3} \right) Gini(\text{อุณหภูมิ} = \text{ร้อน}) \\
&\quad + \left(\frac{n_3}{n_1 + n_2 + n_3} \right) Gini(\text{อุณหภูมิ} = \text{อบอุ่น}) \\
&= \left(\frac{3}{3 + 4 + 7} \right) (0.44) + \left(\frac{4}{3 + 4 + 7} \right) (0.50) + \left(\frac{7}{3 + 4 + 7} \right) (0.40) \\
&= \left(\frac{3}{14} \right) (0.44) + \left(\frac{4}{14} \right) (0.50) + \left(\frac{7}{14} \right) (0.40) \\
&= 0.44
\end{aligned}$$

ดังนั้น $Gini_{split}(\text{อุณหภูมิ}) = 0.44$

3. คำนวณค่า Gini Index ของคุณลักษณะความชื้นในอากาศ

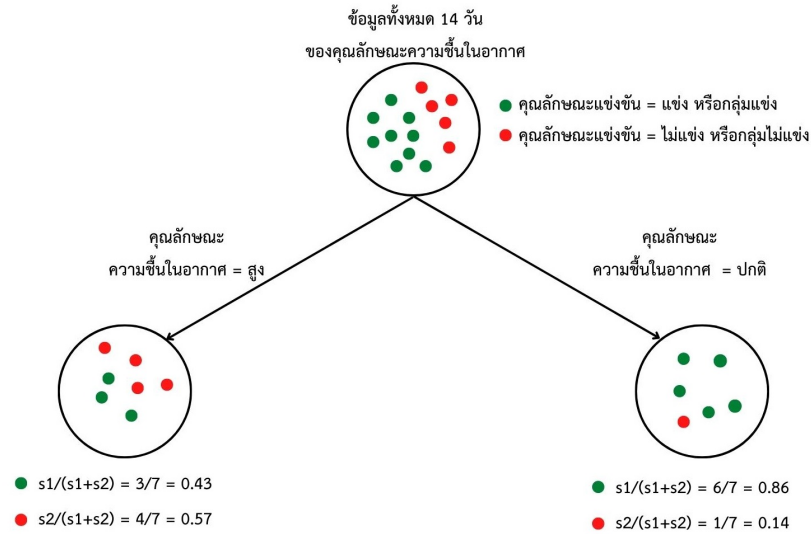
ให้ $T =$ ความชื้นในอากาศ

$M =$ จำนวนกิ่ง(สูง,ปกติ) $= 2$

n_1 = จำนวนข้อมูลในกิ่งที่ 1 (สูง) = 7

n_2 = จำนวนข้อมูลในกิ่งที่ 2 (ปกติ) = 7

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2



รูปที่ 2.4: แสดงข้อมูลของคุณลักษณะอุณหภูมิ

กิ่งที่ 1 (สูง)

t_1 คือ ความชื้นในอากาศ = สูง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดความชื้นในอากาศ = สูง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดความชื้นในอากาศ = สูง

$$\begin{aligned}
 Gini(\text{ความชื้นในอากาศ} = \text{สูง}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{3}{3 + 4} \right)^2 + \left(\frac{4}{3 + 4} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{3}{7} \right)^2 + \left(\frac{4}{7} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{9}{49} \right) + \left(\frac{16}{49} \right) \right] \\
 &= 0.49
 \end{aligned}$$

ดังนั้น $Gini(\text{ความขึ้นในอากาศ} = \text{สูง}) = 0.49$

กึ่งที่ 2 (ปกติ)

t_2 คือ ความขึ้นในอากาศ = ปกติ

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดความขึ้นในอากาศ = ปกติ

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดความขึ้นในอากาศ = ปกติ

$$\begin{aligned}
 Gini(\text{ความขึ้นในอากาศ} = \text{ปกติ}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{6}{6+1} \right)^2 + \left(\frac{1}{6+1} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{6}{7} \right)^2 + \left(\frac{1}{7} \right)^2 \right] \\
 &= 1 - \left[\frac{36}{49} + \frac{1}{49} \right] \\
 &= 0.24
 \end{aligned}$$

ดังนั้น $Gini(\text{ความขึ้นในอากาศ} = \text{ปกติ}) = 0.24$

จะได้ว่า

$$\begin{aligned}
 Gini_{split}(\text{ความขึ้นในอากาศ}) &= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{ความขึ้นในอากาศ} = \text{สูง}) \\
 &\quad + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{ความขึ้นในอากาศ} = \text{ปกติ}) \\
 &= \left(\frac{7}{7+7} \right) (0.49) + \left(\frac{7}{7+7} \right) (0.24) \\
 &= \left(\frac{7}{14} \right) (0.49) + \left(\frac{7}{14} \right) (0.24) \\
 &= 0.37
 \end{aligned}$$

ดังนั้น $Gini_{split}(\text{ความขึ้นในอากาศ}) = 0.37$

4. คำนวณค่า Gini Index ของคุณลักษณะลมแรง

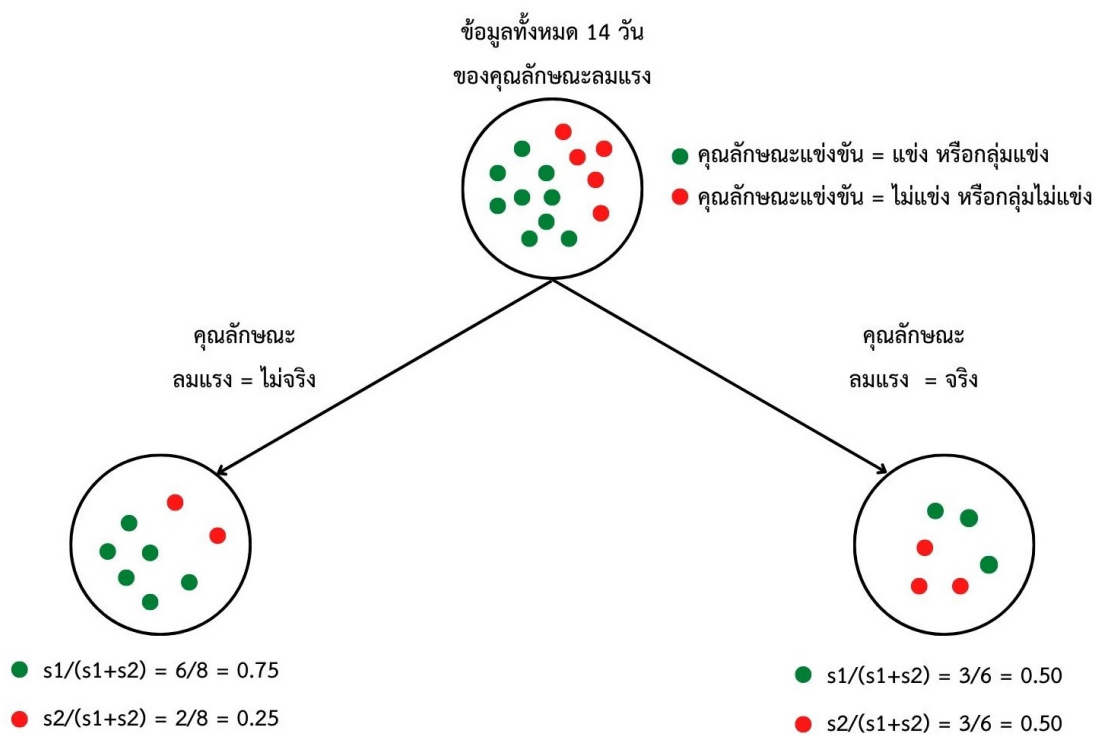
ให้ T = ลมแรง

M = จำนวนกิ่ง(ไม่จริง,จริง) = 3

n_1 = จำนวนข้อมูลในกิ่งที่ 1 (ไม่จริง) = 8

n_2 = จำนวนข้อมูลในกิ่งที่ 2 (จริง) = 6

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2



รูปที่ 2.5: แสดงข้อมูลของคุณลักษณะลมแรง

กิ่งที่ 1 (ไม่จริง)

t_1 คือ ลมแรง = ไม่จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดลมแรง = ไม่จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดลมแรง = ไม่จริง

$$\begin{aligned}
Gini(\text{ลมแรง} = \text{ไม่จริง}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
&= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{6}{6+2} \right)^2 + \left(\frac{2}{6+2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{6}{8} \right)^2 + \left(\frac{2}{8} \right)^2 \right] \\
&= 1 - \left[\left(\frac{36}{64} \right) + \left(\frac{4}{64} \right) \right] \\
&= 0.38
\end{aligned}$$

ดังนั้น $Gini(\text{ลมแรง} = \text{ไม่จริง}) = 0.38$

กึ่งที่ 2 (จริง)

t_2 คือ ลมแรง = จริง

n_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดลมแรง = จริง

n_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดลมแรง = จริง

$$\begin{aligned}
Gini(\text{ลมแรง} = \text{จริง}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
&= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{3}{3+3} \right)^2 + \left(\frac{3}{3+3} \right)^2 \right] \\
&= 1 - \left[\left(\frac{3}{6} \right)^2 + \left(\frac{3}{6} \right)^2 \right] \\
&= 1 - \left[\frac{9}{36} + \frac{9}{36} \right] \\
&= 0.50
\end{aligned}$$

ดังนั้น $Gini(\text{ลมแรง} = \text{จริง}) = 0.5$

จะได้ว่า

$$\begin{aligned}
 Gini_{split}(\text{ลมแรง}) &= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{ลมแรง} = \text{ไม่จริง}) + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{ลมแรง} = \text{จริง}) \\
 &= \left(\frac{8}{8 + 6} \right) (0.38) + \left(\frac{6}{8 + 6} \right) (0.50) \\
 &= \left(\frac{8}{14} \right) (0.38) + \left(\frac{6}{14} \right) (0.50) \\
 &= 0.43
 \end{aligned}$$

ดังนั้น $Gini_{split}(\text{ลมแรง}) = 0.43$

จากการคำนวณค่า $Gini_{split}$ ของคุณลักษณะต่าง ๆ จะได้ว่า

$$Gini_{split}(\text{สภาพภูมิอากาศ}) = 0.34$$

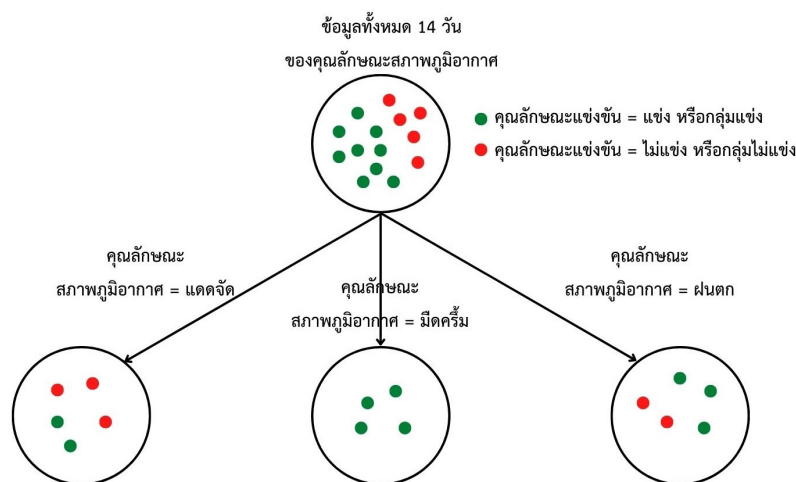
$$Gini_{split}(\text{อุณหภูมิ}) = 0.44$$

$$Gini_{split}(\text{ความชื้นในอากาศ}) = 0.37$$

$$Gini_{split}(\text{ลมแรง}) = 0.43$$

เนื่องจาก คุณลักษณะที่มีค่า Gini Index ที่น้อยที่สุดจะเป็นคุณลักษณะที่มีความหลากหลายของข้อมูลกลุ่มน้อยที่สุด ดังนั้น เราจะเลือกค่า Gini Index ที่น้อยที่สุดเป็นโหนดราก

จะเห็นได้ว่า ค่า Gini Index ของคุณลักษณะสภาพภูมิอากาศ มีค่าน้อยที่สุด คือ 0.34 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศเป็นโหนดราก



รูปที่ 2.6: แสดงข้อมูลของคุณลักษณะสภาพภูมิอากาศ

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = มีดครึ้ม มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = แข่ง ดังนั้น โหนดคุณลักษณะสภาพภูมิอากาศ = มีดครึ้ม จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ แต่โหนดคุณลักษณะสภาพ

ภูมิอากาศ = แดดจัด และโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก ต้องทำการแตกกิ่งจนกว่าข้อมูลในแต่ละโหนดจะมีข้อมูลเป็นกลุ่มเดียวกัน

ต่อไปจะแสดงการคำนวณการแตกกิ่งของโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด โดยสร้างตารางใหม่ที่สภาพภูมิอากาศ = แดดจัด ดังนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
1	แดดจัด	ร้อน	สูง	ไม่จริง	ไม่แข่ง
2	แดดจัด	ร้อน	สูง	จริง	ไม่แข่ง
8	แดดจัด	อบอุ่น	สูง	ไม่จริง	ไม่แข่ง
9	แดดจัด	อบอุ่น	ปกติ	ไม่จริง	แข่ง
11	แดดจัด	อบอุ่น	ปกติ	จริง	แข่ง

ตารางที่ 2.2: แสดงข้อมูลสภาพอากาศที่สภาพภูมิอากาศ = แดดจัด

1. คำนวณค่า Gini Index ของคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

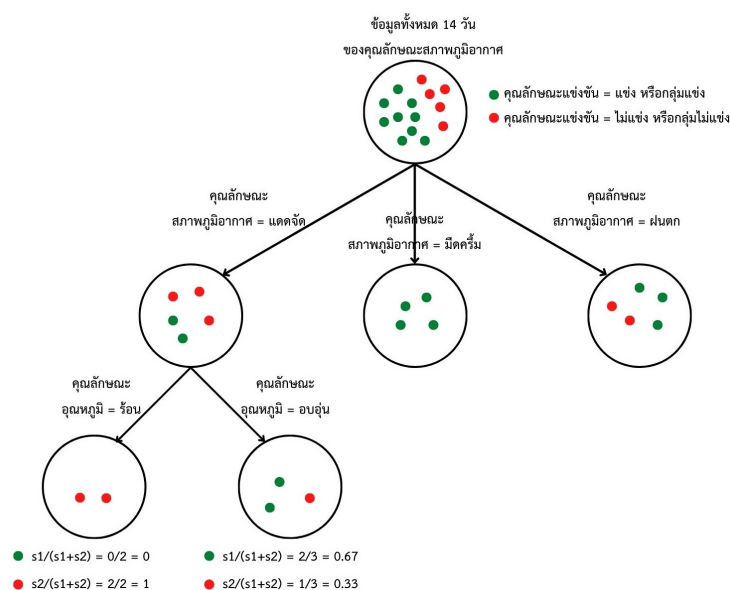
ให้ T = (สภาพภูมิอากาศ = แดดจัด) และอุณหภูมิ

M = จำนวนกิ่ง(ร้อน,อบอุ่น) = 2

n_1 = จำนวนข้อมูลในกิ่งที่ 1 (ร้อน) = 2

n_2 = จำนวนข้อมูลในกิ่งที่ 2 (อบอุ่น) = 3

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2



รูปที่ 2.7: แสดงข้อมูลคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

กึ่งที่ 1 (ร้อน)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

$$\begin{aligned} Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{ร้อน}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\ &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\ &= 1 - \left[\left(\frac{0}{0 + 2} \right)^2 + \left(\frac{2}{0 + 2} \right)^2 \right] \\ &= 1 - \left[\left(\frac{0}{2} \right)^2 + \left(\frac{2}{2} \right)^2 \right] \\ &= 1 - \left[\frac{0}{4} + \frac{4}{4} \right] \\ &= 0 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{ร้อน}) = 0$

กึ่งที่ 2 (อบอุ่น)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

$$\begin{aligned}
Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{อบอุ่น}) &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
&= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
&= 1 - \left[\left(\frac{2}{2+1} \right)^2 + \left(\frac{1}{2+1} \right)^2 \right] \\
&= 1 - \left[\left(\frac{2}{3} \right)^2 + \left(\frac{1}{3} \right)^2 \right] \\
&= 1 - \left[\frac{4}{9} + \frac{1}{9} \right] \\
&= 0.44
\end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{อบอุ่น}) = 0.44$

จะได้ว่า

$$\begin{aligned}
Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) \\
&= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
&= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{ร้อน}) \\
&\quad + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{อบอุ่น}) \\
&= \left(\frac{2}{2+3} \right) (0) + \left(\frac{3}{2+3} \right) (0.44) \\
&= \left(\frac{2}{5} \right) (0) + \left(\frac{3}{5} \right) (0.44) \\
&= 0.26
\end{aligned}$$

ดังนั้น $Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) = 0.26$

2. คำนวณค่า Gini Index ของคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

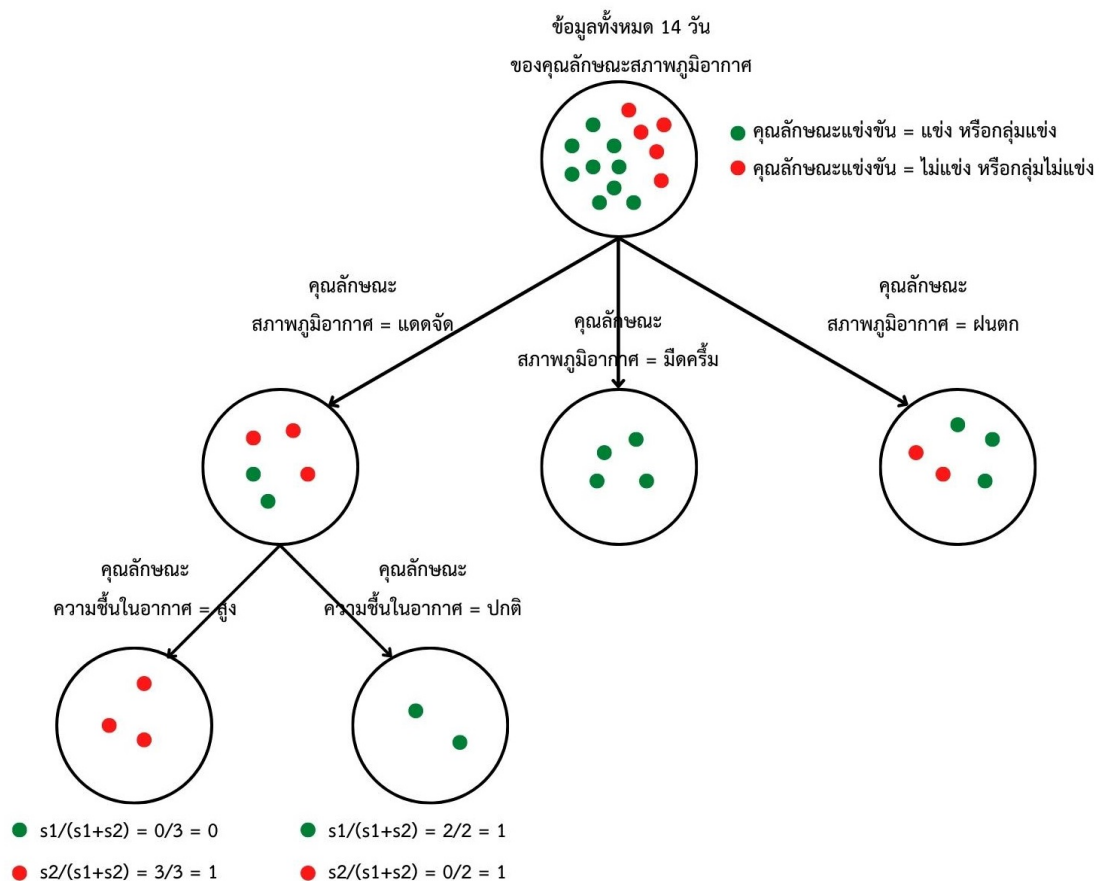
ให้ $T = (\text{สภาพภูมิอากาศ} = \text{แดดจัด})$ และความชื้นในอากาศ

$M = \text{จำนวนกิ่ง(สูง,ปกติ)} = 2$

$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (สูง)} = 3$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (ปกติ)} = 2$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$



รูปที่ 2.8: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

กิ่งที่ 1 (สูง)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

$Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด} \text{ และ } \text{ความชื้นในอากาศ} = \text{สูง})$

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{0 + 3} \right)^2 + \left(\frac{3}{0 + 3} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{3} \right)^2 + \left(\frac{3}{3} \right)^2 \right] \\
 &= 1 - \left[\frac{0}{9} + \frac{9}{9} \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด} \text{ และ } \text{ความชื้นในอากาศ} = \text{สูง}) = 0$

กึ่งที่ 2 (ปกติ)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และ ความชื้นในอากาศ = ปกติ

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และ ความชื้นในอากาศ = ปกติ

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และ ความชื้นในอากาศ = ปกติ

$Gini$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{2+0} \right)^2 + \left(\frac{0}{2+0} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{2} \right)^2 + \left(\frac{0}{2} \right)^2 \right] \\
 &= 1 - \left[\frac{4}{4} + \frac{0}{4} \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini$ (ภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ) = 0

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ)

$$\begin{aligned}
 &= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
 &= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ}) \\
 &\quad + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ} = \text{สูง}) \\
 &= \left(\frac{3}{3+2} \right) (0) + \left(\frac{2}{3+2} \right) (0) \\
 &= \left(\frac{3}{5} \right) (0) + \left(\frac{2}{5} \right) (0) \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ) = 0

3. คำนวณค่า Gini Index ของคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

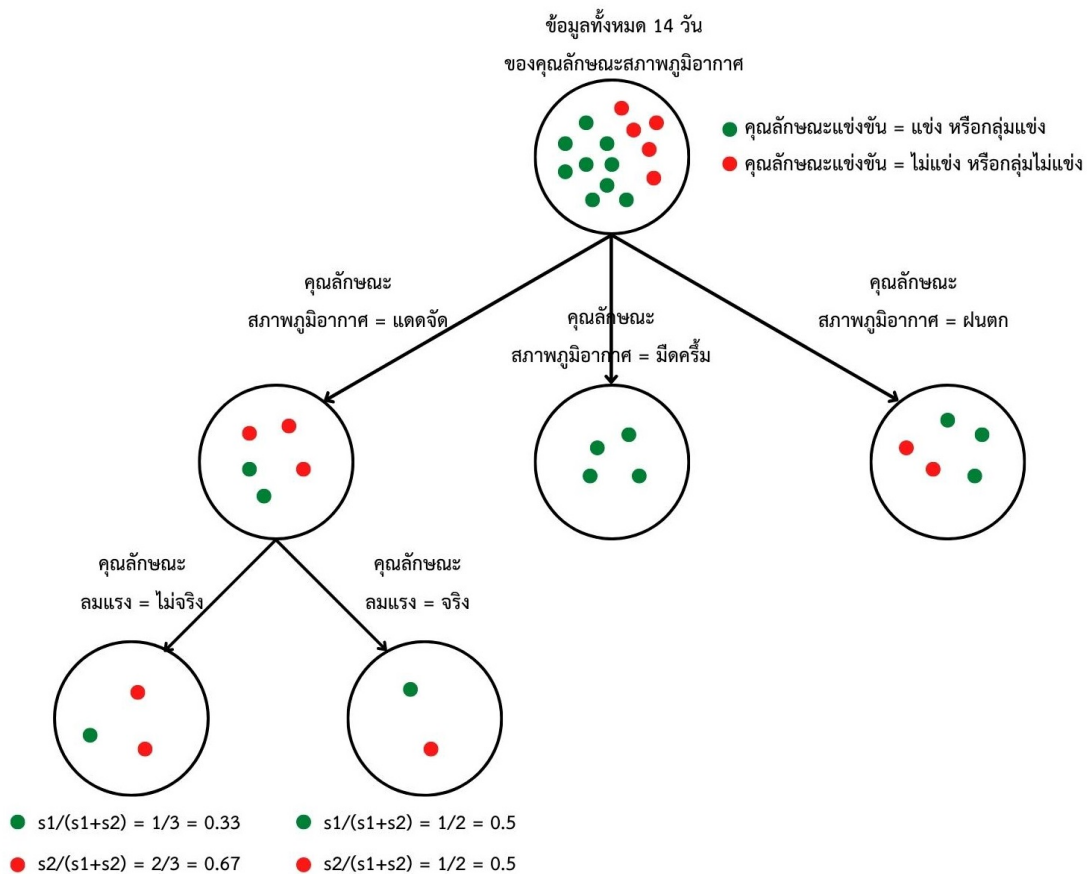
ให้ $T = (\text{สภาพภูมิอากาศ} = \text{แดดจัด})$ และลมแรง

$M = \text{จำนวนกิ่ง(ไม่จริง,จริง)} = 2$

$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (ไม่จริง)} = 3$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (จริง)} = 2$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$



รูปที่ 2.9: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

กิ่งที่ 1 (ไม่จริง)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

$Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{ไม่จริง})$

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{1+2} \right)^2 + \left(\frac{2}{1+2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{9} \right) + \left(\frac{4}{9} \right) \right] \\
 &= 0.44
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{ไม่จริง}) = 0.44$

กึ่งที่ 2 (จริง)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

$Gini$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{1+1} \right)^2 + \left(\frac{1}{1+1} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{4} \right) + \left(\frac{1}{4} \right) \right] \\
 &= 0.50
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง) = 0.50

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และลมแรง)

$$\begin{aligned}
 &= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
 &= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{ไม่จริง}) \\
 &\quad + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{จริง}) \\
 &= \left(\frac{3}{3+2} \right) (0.44) + \left(\frac{2}{3+2} \right) (0.50) \\
 &= \left(\frac{3}{5} \right) (0.44) + \left(\frac{2}{5} \right) (0.50) \\
 &= 0.46
 \end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และลมแรง) = 0.46

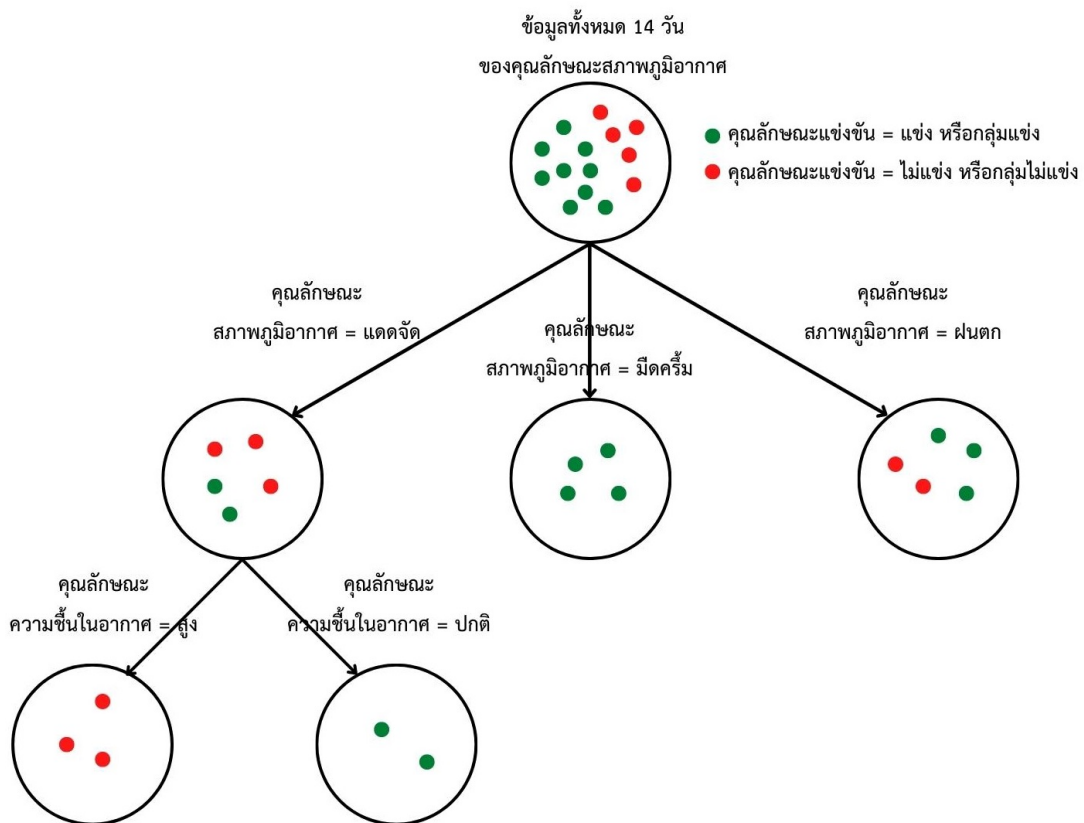
จากการคำนวณค่า $Gini_{split}$ ของคุณลักษณะต่าง ๆ จะได้ว่า

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) = 0.26$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ}) = 0$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง}) = 0.46$$

จะเห็นได้ว่า ค่า Gini Index ของคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ มีค่าน้อยที่สุด คือ 0 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ



รูปที่ 2.10: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = ไม่แข่ง และโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = แข่ง ดังนั้น จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ

ต่อไปจะแสดงการคำนวณการแตกกิ่งของโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก โดยสร้างตารางใหม่ที่สภาพภูมิอากาศ = ฝนตก ดังนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
4	ฝนตก	อบอุ่น	สูง	ไม่จริง	แข่ง
5	ฝนตก	เย็น	ปกติ	ไม่จริง	แข่ง
6	ฝนตก	เย็น	ปกติ	จริง	ไม่แข่ง
10	ฝนตก	อบอุ่น	ปกติ	ไม่จริง	แข่ง
14	ฝนตก	อบอุ่น	สูง	จริง	ไม่แข่ง

ตารางที่ 2.3: แสดงข้อมูลสภาพอากาศที่สภาพภูมิอากาศ = ฝนตก

1. คำนวณค่า Gini Index ของคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

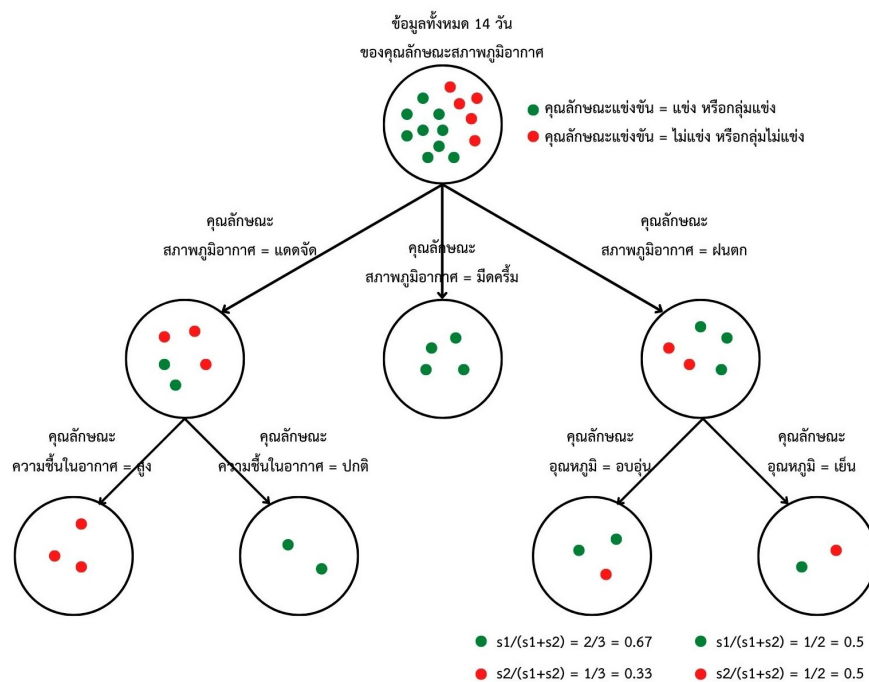
ให้ T = (สภาพภูมิอากาศ = ฝนตก) และอุณหภูมิ

M = จำนวนกิ่ง(อบอุ่น,เย็น) = 2

n_1 = จำนวนข้อมูลในกิ่งที่ 1 (อบอุ่น) = 3

n_2 = จำนวนข้อมูลในกิ่งที่ 2 (เย็น) = 2

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2



รูปที่ 2.11: แสดงข้อมูลคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

กึ่งที่ 1 (อบอุ้น)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

$Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{อบอุ่น})$

$$\begin{aligned} &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\ &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\ &= 1 - \left[\left(\frac{2}{2+1} \right)^2 + \left(\frac{1}{2+1} \right)^2 \right] \\ &= 1 - \left[\left(\frac{2}{3} \right)^2 + \left(\frac{1}{3} \right)^2 \right] \\ &= 1 - \left[\left(\frac{4}{9} \right) + \left(\frac{1}{9} \right) \right] \\ &= 0.44 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{อบอุ่น}) = 0.44$

กึ่งที่ 2 (เย็น)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

$Gini$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{1+1} \right)^2 + \left(\frac{1}{1+1} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{4} \right) + \left(\frac{1}{4} \right) \right] \\
 &= 0.50
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น) = 0.50

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ)

$$\begin{aligned}
 &= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
 &= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{อบอุ่น}) \\
 &\quad + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{เย็น}) \\
 &= \left(\frac{3}{3+2} \right) (0.44) + \left(\frac{2}{3+2} \right) (0.5) \\
 &= \left(\frac{3}{5} \right) (0.44) + \left(\frac{2}{5} \right) (0.5) \\
 &= 0.46
 \end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ) = 0.46

2. คำนวณค่า Gini Index ของคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

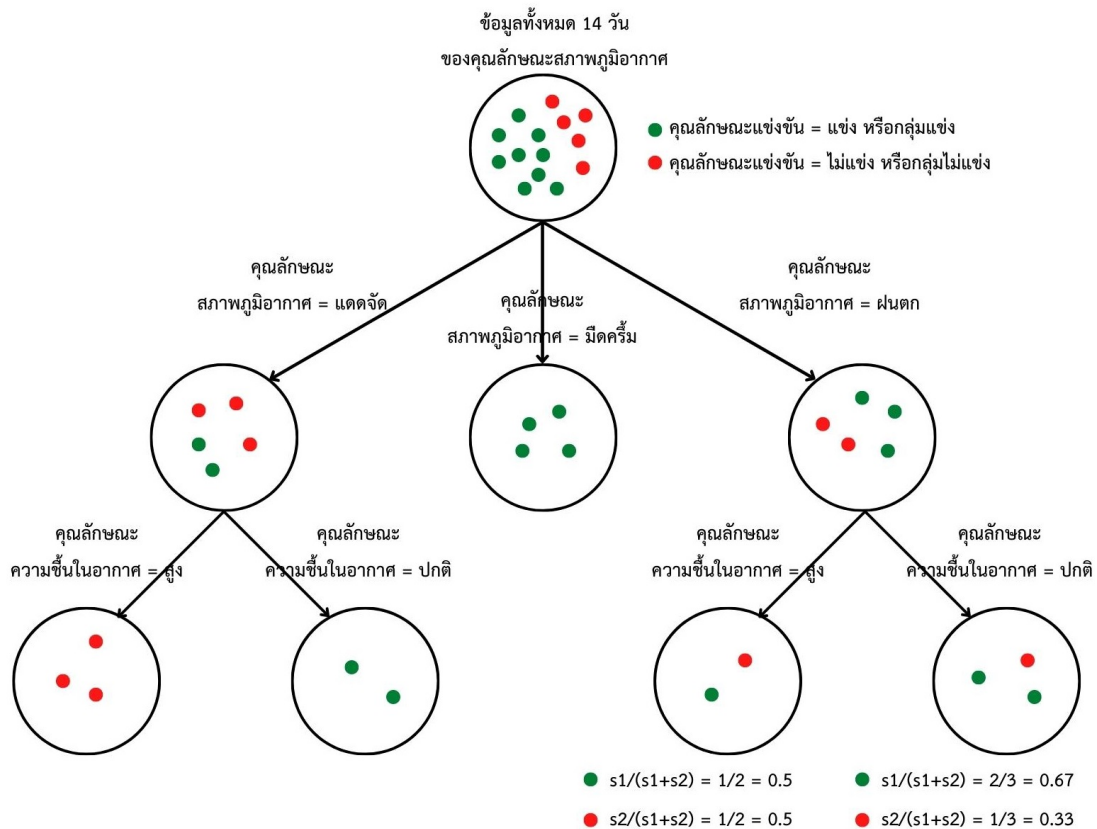
ให้ T = (สภาพภูมิอากาศ = ฝนตก) และความชื้นในอากาศ

M = จำนวนกิ่ง(สูง,ปกติ) = 2

n_1 = จำนวนข้อมูลในกิ่งที่ 1 (สูง) = 2

n_2 = จำนวนข้อมูลในกิ่งที่ 2 (ปกติ) = 3

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2



รูปที่ 2.12: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

กิ่งที่ 1 (สูง)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

$Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และ ความชื้นในอากาศ} = \text{สูง})$

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{1+1} \right)^2 + \left(\frac{1}{1+1} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1}{4} \right) + \left(\frac{1}{4} \right) \right] \\
 &= 0.50
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และ ความชื้นในอากาศ} = \text{สูง}) = 0.50$

กึ่งที่ 2 (ปกติ)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และ ความชื้นในอากาศ = ปกติ

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และ ความชื้นในอากาศ = ปกติ

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และ ความชื้นในอากาศ = ปกติ

$Gini$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{2+1} \right)^2 + \left(\frac{1}{2+1} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2}{3} \right)^2 + \left(\frac{1}{3} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{4}{9} \right) + \left(\frac{1}{9} \right) \right] \\
 &= 0.44
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ) = 0.44

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ)

$$\begin{aligned}
 &= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
 &= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ} = \text{สูง}) \\
 &\quad + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ} = \text{ปกติ}) \\
 &= \left(\frac{2}{2+3} \right) (0.50) + \left(\frac{3}{2+3} \right) (0.44) \\
 &= \left(\frac{2}{5} \right) (0.50) + \left(\frac{3}{5} \right) (0.44) \\
 &= 0.46
 \end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ) = 0.46

3. คำนวณค่า Gini Index ของคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก
= ฝนตก

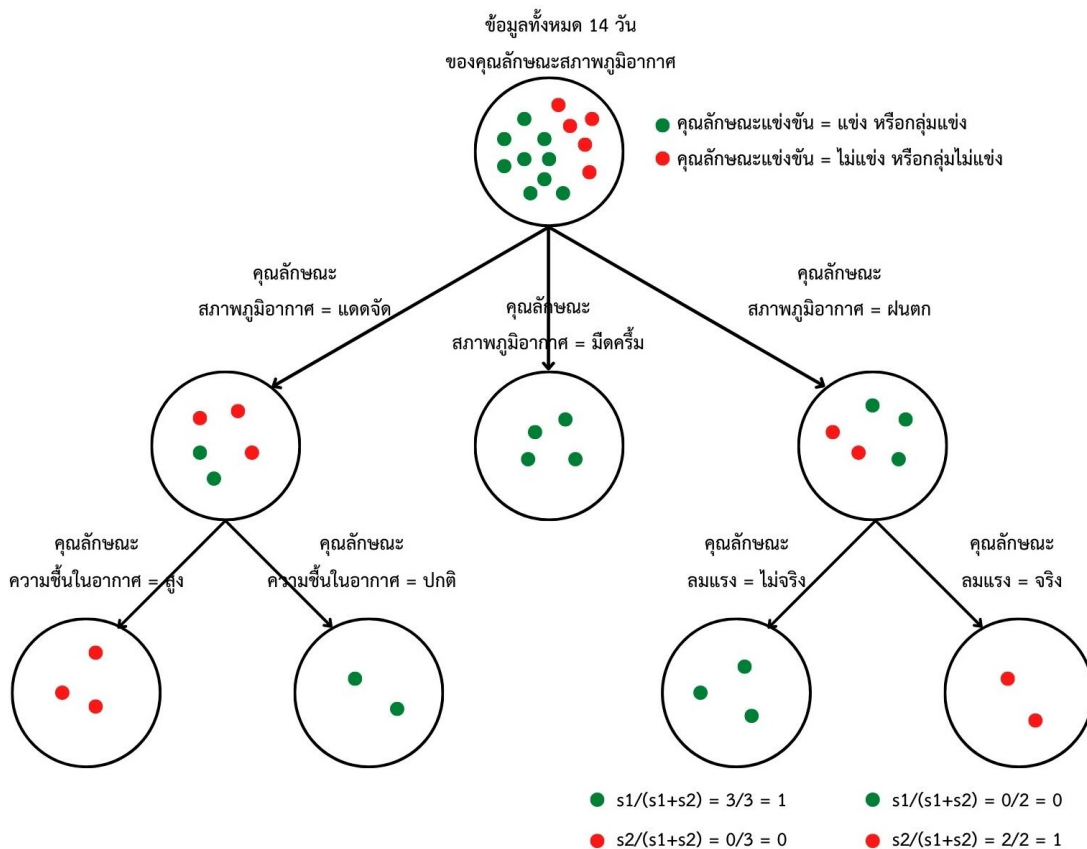
ให้ $T = (\text{สภาพภูมิอากาศ} = \text{ฝนตก})$ และลมแรง

$M = \text{จำนวนกิ่ง(ไม่จริง,จริง)} = 2$

$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (เย็น)} = 3$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (จริง)} = 2$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$



รูปที่ 2.13: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

กิ่งที่ 1 (ไม่จริง)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

$Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{ไม่จริง})$

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{3}{3+0} \right)^2 + \left(\frac{0}{3+0} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{3}{3} \right)^2 + \left(\frac{0}{3} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{9}{9} \right) + \left(\frac{0}{9} \right) \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{ไม่จริง}) = 0$

กึ่งที่ 2 (จริง)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

$Gini$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N s_l} \right]^2 \\
 &= 1 - \left[\left(\frac{s_1}{s_1 + s_2} \right)^2 + \left(\frac{s_2}{s_1 + s_2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{0 + 2} \right)^2 + \left(\frac{2}{0 + 2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{2} \right)^2 + \left(\frac{2}{2} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{4} \right) + \left(\frac{4}{4} \right) \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง) = 0

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และลมแรง)

$$\begin{aligned}
 &= \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i) \\
 &= \left(\frac{n_1}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{ไม่จริง}) \\
 &\quad + \left(\frac{n_2}{n_1 + n_2} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{จริง}) \\
 &= \left(\frac{3}{3 + 2} \right) (0) + \left(\frac{2}{3 + 2} \right) (0) \\
 &= \left(\frac{3}{5} \right) (0) + \left(\frac{2}{5} \right) (0) \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และลมแรง) = 0

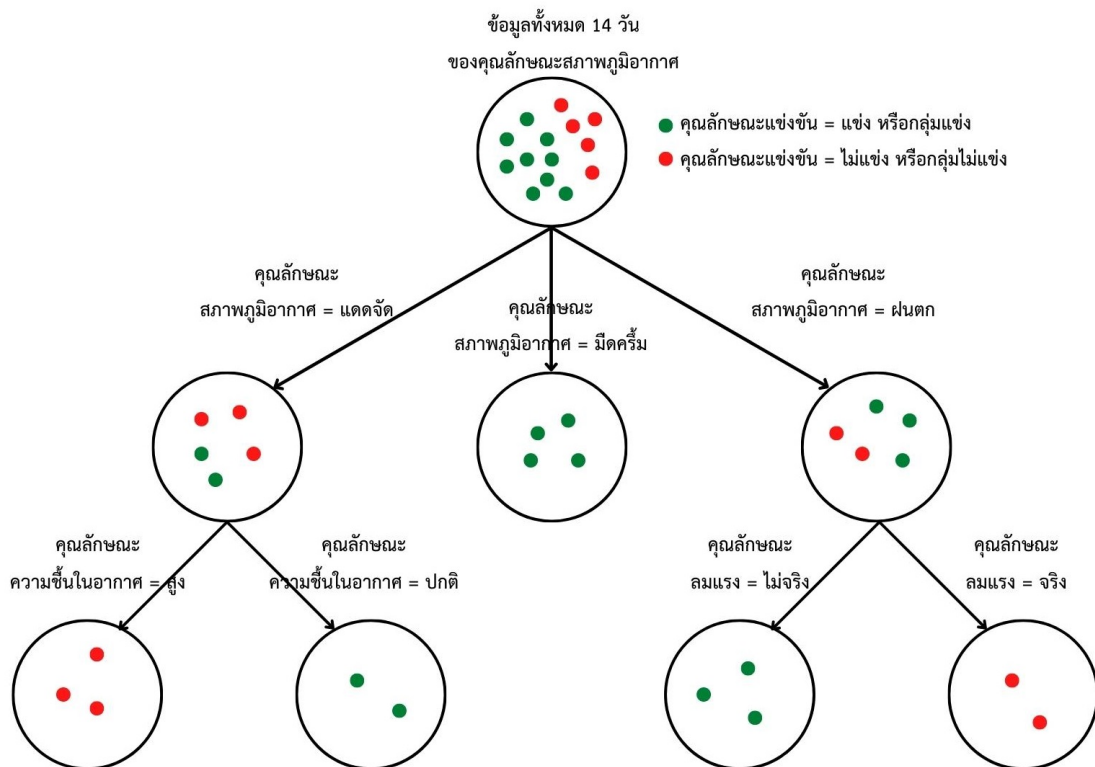
จากการคำนวณค่า $Gini_{split}$ ของคุณลักษณะต่าง ๆ เมื่อแตกกิ่งออกจากโหนดสภาพภูมิอากาศ = ฝนตก จะได้ว่า

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ}) = 0.46$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ}) = 0.46$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง}) = 0$$

จะเห็นได้ว่า ค่า Gini Index ของคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง มีค่าน้อยที่สุด คือ 0 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง

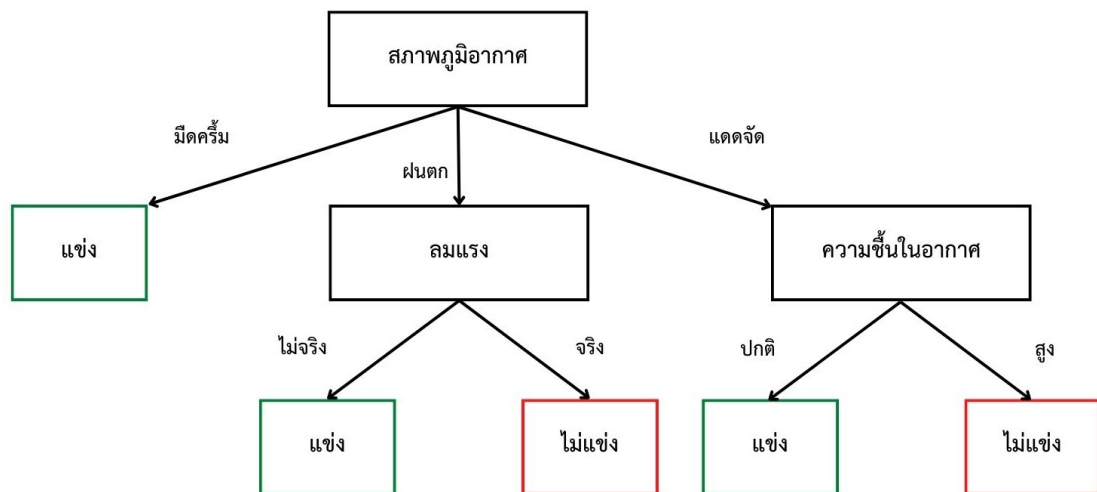


รูปที่ 2.14: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = แข่ง และโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = ไม่แข่ง ดังนั้น จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ

เนื่องจากในแต่ละโหนดมีข้อมูลเป็นกลุ่มเดียวกัน ฉะนั้น จะเสร็จสิ้นในขั้นตอนการคำนวณ

ดังนั้น สามารถสร้างโมเดลต้นไม้ตัดสินใจ ได้ดังนี้



รูปที่ 2.15: แสดงโมเดลต้นไม้ตัดสินใจของข้อมูลสภาพอากาศ 14 วันย้อนหลัง สำหรับการแข่งขันกีฬา

ตัวอย่าง 2.2 ตัวอย่าง การสร้างต้นไม้ตัดสินใจโดยใช้ Information Gain(IG) โดยสมมติมีข้อมูลของสภาพอากาศ 14 วันย้อนหลัง สำหรับการแข่งขันกีฬา ดังตารางต่อไปนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
1	แดดจัด	ร้อน	สูง	ไม่จริง	ไม่แข่ง
2	แดดจัด	ร้อน	สูง	จริง	ไม่แข่ง
3	มีดคร่ำ	ร้อน	สูง	ไม่จริง	แข่ง
4	ฝนตก	อบอุ่น	สูง	ไม่จริง	แข่ง
5	ฝนตก	เย็น	ปกติ	ไม่จริง	แข่ง
6	ฝนตก	เย็น	ปกติ	จริง	ไม่แข่ง
7	มีดคร่ำ	เย็น	ปกติ	จริง	แข่ง
8	แดดจัด	อบอุ่น	สูง	ไม่จริง	ไม่แข่ง
9	แดดจัด	อบอุ่น	ปกติ	ไม่จริง	แข่ง
10	ฝนตก	อบอุ่น	ปกติ	ไม่จริง	แข่ง
11	แดดจัด	อบอุ่น	ปกติ	จริง	แข่ง
12	มีดคร่ำ	อบอุ่น	สูง	จริง	แข่ง
13	มีดคร่ำ	ร้อน	ปกติ	ไม่จริง	แข่ง
14	ฝนตก	อบอุ่น	สูง	จริง	ไม่แข่ง

ตารางที่ 2.4: แสดงข้อมูลสภาพอากาศ

จากตาราง ประกอบด้วย 5 คุณลักษณะ คือ

- สภาพภูมิอากาศ ประกอบด้วย 3 ค่า คือ แดดจัด, มีดครีมี, ฝนตก
- อุณหภูมิ ประกอบด้วย 3 ค่า คือ ร้อน, อุ่น, เย็น
- ความชื้นในอากาศ ประกอบด้วย 2 ค่า คือ สูง, ปกติ
- ลมแรง ประกอบด้วย 2 ค่า คือ จริง, ไม่จริง
- แข่งขัน ซึ่งเป็นกลุ่ม ประกอบด้วย 2 ค่า คือ แข่ง, ไม่แข่ง

ต่อไปจะแสดงการคำนวณค่าแต่ละคุณลักษณะเทียบกับกลุ่ม เพื่อหาคุณลักษณะที่มีค่า IG มากที่สุดมาเป็น โหนดรากของต้นไม้ตัดสินใจ ดังนี้

1. คำนวณค่า IG ของคุณลักษณะสภาพภูมิอากาศ

ให้ T = สภาพภูมิอากาศ

$$M = \text{จำนวนกิ่ง(แดดจัด, มีดครีมี, ฝนตก)} = 3$$

$$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (แดดจัด)} = 5$$

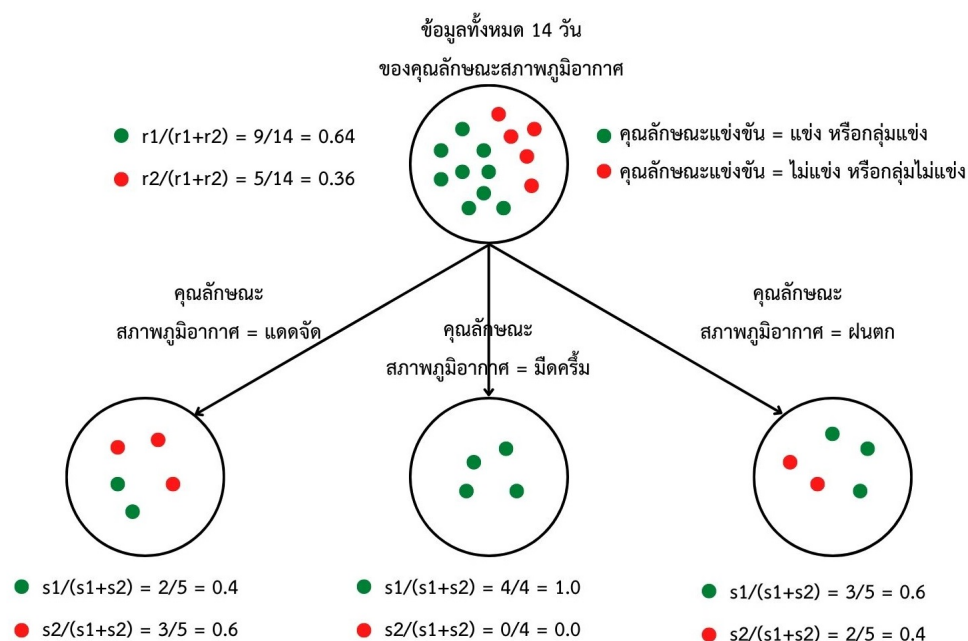
$$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (มีดครีมี)} = 4$$

$$n_3 = \text{จำนวนข้อมูลในกิ่งที่ 3 (ฝนตก)} = 5$$

$$N = \text{จำนวนกลุ่ม(แข่ง, ไม่แข่ง)} = 2$$

$$r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะสภาพภูมิอากาศ} = 9$$

$$r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะสภาพภูมิอากาศ} = 5$$



รูปที่ 2.16: แสดงข้อมูลของคุณลักษณะสภาพภูมิอากาศ

$$\begin{aligned}
Entropy(\text{สภาพภูมิอากาศ}) &= - \sum_{k=1}^N \left[\frac{\frac{r_k}{N}}{\sum_{l=1}^N \frac{r_l}{N}} \log_2 \left(\frac{\frac{r_k}{N}}{\sum_{l=1}^N \frac{r_l}{N}} \right) \right] \\
&= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
&= - \left[\left(\frac{9}{9+5} \log_2 \frac{9}{9+5} \right) + \left(\frac{5}{9+5} \log_2 \frac{5}{9+5} \right) \right] \\
&= - \left[\left(\frac{9}{14} \log_2 \frac{9}{14} \right) + \left(\frac{5}{14} \log_2 \frac{5}{14} \right) \right] \\
&= - [(0.64 \log_2 0.64) + (0.36 \log_2 0.36)] \\
&= - [(0.64 \times -0.64) + (0.36 \times -1.47)] \\
&= - [(-0.41) + (-0.53)] \\
&= 0.94
\end{aligned}$$

ดังนั้น $Entropy(\text{สภาพภูมิอากาศ}) = 0.94$

กึ่งที่ 1 (แดดจัด)

t_1 คือ สภาพภูมิอากาศ = แดดจัด

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด

$$\begin{aligned}
Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) &= - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N \frac{s_l}{N}} \log_2 \left(\frac{\frac{s_k}{N}}{\sum_{l=1}^N \frac{s_l}{N}} \right) \right] \\
&= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
&= - \left[\left(\frac{2}{2+3} \log_2 \frac{2}{2+3} \right) + \left(\frac{3}{2+3} \log_2 \frac{3}{2+3} \right) \right] \\
&= - \left[\left(\frac{2}{5} \log_2 \frac{2}{5} \right) + \left(\frac{3}{5} \log_2 \frac{3}{5} \right) \right] \\
&= - [(0.4 \log_2 0.4) + (0.6 \log_2 0.6)] \\
&= - [(0.4 \times -1.32) + (0.6 \times -0.74)] \\
&= - [(-0.53) + (-0.44)] \\
&= 0.97
\end{aligned}$$

ดังนั้น $Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) = 0.97$

กึ่งที่ 2 (มีดครี้ม)

t_2 คือ สภาพภูมิอากาศ = มีดครี้ม

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = มีดครี้ม

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = มีดครี้ม

$$\begin{aligned} Entropy(\text{สภาพภูมิอากาศ} = \text{มีดครี้ม}) &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\ &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\ &= - \left[\left(\frac{4}{4 + 0} \log_2 \frac{4}{4 + 0} \right) + \left(\frac{0}{4 + 0} \log_2 \frac{0}{4 + 0} \right) \right] \\ &= - \left[\left(\frac{4}{4} \log_2 \frac{4}{4} \right) + \left(\frac{0}{4} \log_2 \frac{0}{4} \right) \right] \\ &= - [(1 \log_2 1) + (0 \log_2 0)] \\ &= - [(1 \times 0) + 0] \\ &= - [0 + 0] \\ &= 0 \end{aligned}$$

ดังนั้น $Entropy(\text{สภาพภูมิอากาศ} = \text{มีดครี้ม}) = 0$

กึ่งที่ 3 (ฝนตก)

t_3 คือ สภาพภูมิอากาศ = ฝนตก

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก

$$\begin{aligned}
Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
&= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
&= - \left[\left(\frac{3}{3+2} \log_2 \frac{3}{3+2} \right) + \left(\frac{2}{3+2} \log_2 \frac{2}{3+2} \right) \right] \\
&= - \left[\left(\frac{3}{5} \log_2 \frac{3}{5} \right) + \left(\frac{2}{5} \log_2 \frac{2}{5} \right) \right] \\
&= - [(0.6 \log_2 0.6) + (0.4 \log_2 0.4)] \\
&= - [(0.6 \times -0.74) + (0.4 \times -1.32)] \\
&= - [(-0.44) + (-0.53)] \\
&= 0.97
\end{aligned}$$

ดังนั้น $Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) = 0.97$

จะได้ว่า

$$\begin{aligned}
IG(\text{สภาพภูมิอากาศ}) &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
&= Entropy(\text{สภาพภูมิอากาศ}) - \frac{n_1}{n_1 + n_2 + n_3} Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) \\
&\quad - \frac{n_2}{n_1 + n_2 + n_3} Entropy(\text{สภาพภูมิอากาศ} = \text{มีดครึ้ม}) \\
&\quad - \frac{n_3}{n_1 + n_2 + n_3} Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) \\
&= 0.94 - \left(\frac{5}{5+4+5} \times 0.97 \right) - \left(\frac{4}{5+4+5} \times 0 \right) - \left(\frac{5}{5+4+5} \times 0.97 \right) \\
&= 0.94 - \left(\frac{5}{14} \times 0.97 \right) - \left(\frac{4}{14} \times 0 \right) - \left(\frac{5}{14} \times 0.97 \right) \\
&= 0.94 - (0.35 \times 0.97) - (0.30 \times 0) - (0.35 \times 0.97) \\
&= 0.94 - 0.34 - 0 - 0.34 \\
&= 0.26
\end{aligned}$$

ดังนั้น $IG(\text{สภาพภูมิอากาศ}) = 0.26$

2. คำนวณค่า IG ของคุณลักษณะอุณหภูมิ

ให้ $T =$ อุณหภูมิ

$M =$ จำนวนกิ่ง(เย็น,ร้อน,อบอุ่น) = 3

$n_1 =$ จำนวนข้อมูลในกิ่งที่ 1 (เย็น) = 3

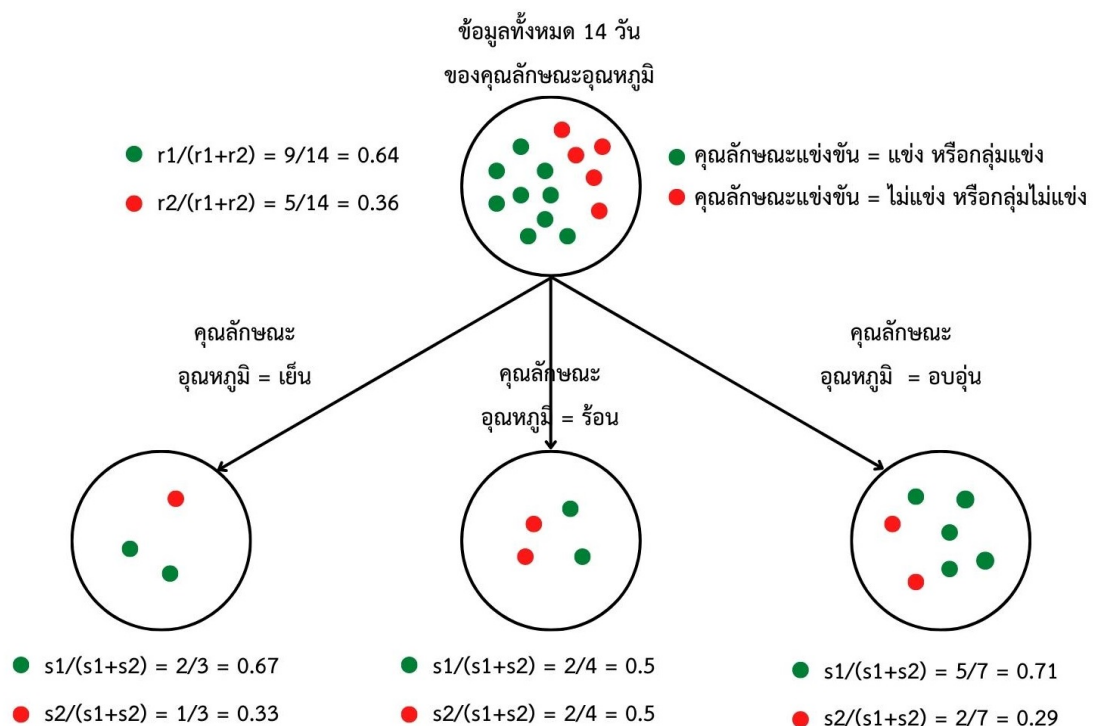
$n_2 =$ จำนวนข้อมูลในกิ่งที่ 2 (ร้อน) = 4

$n_3 =$ จำนวนข้อมูลในกิ่งที่ 3 (อบอุ่น) = 7

$N =$ จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

$r_1 =$ จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะอุณหภูมิ = 9

$r_2 =$ จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะอุณหภูมิ = 5



รูปที่ 2.17: แสดงข้อมูลของคุณลักษณะอุณหภูมิ

$$\begin{aligned}
Entropy(\text{อุณหภูมิ}) &= - \sum_{k=1}^N \left[\frac{\frac{r_k}{N}}{\sum_{l=1}^N \frac{r_l}{N}} \log_2 \left(\frac{\frac{r_k}{N}}{\sum_{l=1}^N \frac{r_l}{N}} \right) \right] \\
&= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
&= - \left[\left(\frac{9}{9 + 5} \log_2 \frac{9}{9 + 5} \right) + \left(\frac{5}{9 + 5} \log_2 \frac{5}{9 + 5} \right) \right] \\
&= - \left[\left(\frac{9}{14} \log_2 \frac{9}{14} \right) + \left(\frac{5}{14} \log_2 \frac{5}{14} \right) \right] \\
&= - [(0.64 \log_2 0.64) + (0.36 \log_2 0.36)] \\
&= - [(0.64 \times -0.64) + (0.36 \times -1.47)] \\
&= - [(-0.41) + (-0.53)] \\
&= 0.94
\end{aligned}$$

ดังนั้น $Entropy(\text{อุณหภูมิ}) = 0.94$

กึ่งที่ 1 (เย็น)

t_1 คือ อุณหภูมิ = เย็น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = เย็น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = เย็น

$$\begin{aligned}
Entropy(\text{อุณหภูมิ} = \text{เย็น}) &= - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N \frac{s_l}{N}} \log_2 \left(\frac{\frac{s_k}{N}}{\sum_{l=1}^N \frac{s_l}{N}} \right) \right] \\
&= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
&= - \left[\left(\frac{2}{2 + 1} \log_2 \frac{2}{2 + 1} \right) + \left(\frac{1}{2 + 1} \log_2 \frac{1}{2 + 1} \right) \right] \\
&= - \left[\left(\frac{2}{3} \log_2 \frac{2}{3} \right) + \left(\frac{1}{3} \log_2 \frac{1}{3} \right) \right] \\
&= - [(0.67 \log_2 0.67) + (0.33 \log_2 0.33)] \\
&= - [(0.67 \times -0.58) + (0.33 \times -1.60)] \\
&= - [(-0.39) + (-0.53)] \\
&= 0.92
\end{aligned}$$

ดังนั้น $Entropy(\text{อุณหภูมิ} = \text{เย็น}) = 0.92$

กึ่งที่ 2 (ร้อน)

t_2 คือ อุณหภูมิ = ร้อน

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = ร้อน

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = ร้อน

$$\begin{aligned} Entropy(\text{อุณหภูมิ} = \text{ร้อน}) &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\ &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\ &= - \left[\left(\frac{2}{2+2} \log_2 \frac{2}{2+2} \right) + \left(\frac{2}{2+2} \log_2 \frac{2}{2+2} \right) \right] \\ &= - \left[\left(\frac{2}{4} \log_2 \frac{2}{4} \right) + \left(\frac{2}{4} \log_2 \frac{2}{4} \right) \right] \\ &= - [(0.5 \log_2 0.5) + (0.5 \log_2 0.5)] \\ &= - [(0.5 \times -1) + (0.5 \times -1)] \\ &= - [(-0.5) + (-0.5)] \\ &= 1 \end{aligned}$$

ดังนั้น $Entropy(\text{อุณหภูมิ} = \text{ร้อน}) = 1$

กึ่งที่ 3 (อบอุ่น)

t_3 คือ อุณหภูมิ = อบอุ่น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = อบอุ่น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = อบอุ่น

$$\begin{aligned}
Entropy(\text{อุณหภูมิ} = \text{อบอุ่น}) &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
&= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
&= - \left[\left(\frac{5}{5+2} \log_2 \frac{5}{5+2} \right) + \left(\frac{2}{5+2} \log_2 \frac{2}{5+2} \right) \right] \\
&= - \left[\left(\frac{5}{7} \log_2 \frac{5}{7} \right) + \left(\frac{2}{7} \log_2 \frac{2}{7} \right) \right] \\
&= - [(0.71 \log_2 0.71) + (0.29 \log_2 0.29)] \\
&= - [(0.71 \times -0.49) + (0.29 \times -1.79)] \\
&= - [(-0.35) + (-0.52)] \\
&= 0.87
\end{aligned}$$

ดังนั้น $Entropy(\text{อุณหภูมิ} = \text{อบอุ่น}) = 0.87$

จะได้ว่า

$$\begin{aligned}
IG(\text{อุณหภูมิ}) &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
&= Entropy(\text{อุณหภูมิ}) - \frac{n_1}{n_1 + n_2 + n_3} Entropy(\text{อุณหภูมิ} = \text{เย็น}) \\
&\quad - \frac{n_2}{n_1 + n_2 + n_3} Entropy(\text{อุณหภูมิ} = \text{ร้อน}) \\
&\quad - \frac{n_3}{n_1 + n_2 + n_3} Entropy(\text{อุณหภูมิ} = \text{อบอุ่น}) \\
&= 0.94 - \left(\frac{3}{3+4+7} \times 0.92 \right) - \left(\frac{4}{3+4+7} \times 1 \right) - \left(\frac{7}{3+4+7} \times 0.87 \right) \\
&= 0.94 - \left(\frac{3}{14} \times 0.92 \right) - \left(\frac{4}{14} \times 1 \right) - \left(\frac{7}{14} \times 0.87 \right) \\
&= 0.94 - (0.21 \times 0.92) - (0.29 \times 1) - (0.50 \times 0.87) \\
&= 0.94 - 0.19 - 0.29 - 0.43 \\
&= 0.03
\end{aligned}$$

ดังนั้น $IG(\text{อุณหภูมิ}) = 0.03$

3. คำนวณค่า IG ของคุณลักษณะความชื้นในอากาศ

ให้ T = ความชื้นในอากาศ

M = จำนวนกิ่ง(สูง,ปกติ) = 2

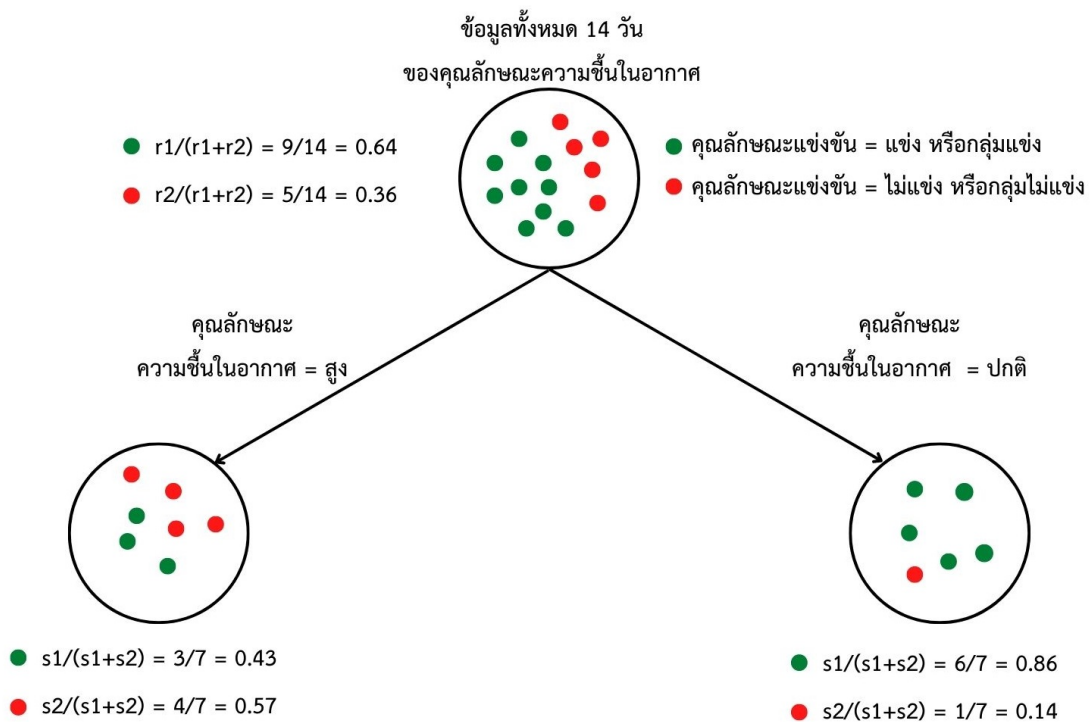
n_1 = จำนวนข้อมูลในกิ่งที่ 1 (สูง) = 7

n_2 = จำนวนข้อมูลในกิ่งที่ 2 (ปกติ) = 7

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

r_1 = จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะความชื้นในอากาศ = 9

r_2 = จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะความชื้นในอากาศ = 5



รูปที่ 2.18: แสดงข้อมูลของคุณลักษณะความชื้นในอากาศ

$$\begin{aligned}
Entropy(\text{ความชื้นในอากาศ}) &= - \sum_{k=1}^N \left[\frac{\frac{r_k}{N}}{\sum_{l=1}^N \frac{r_l}{N}} \log_2 \left(\frac{\frac{r_k}{N}}{\sum_{l=1}^N \frac{r_l}{N}} \right) \right] \\
&= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
&= - \left[\left(\frac{9}{9+5} \log_2 \frac{9}{9+5} \right) + \left(\frac{5}{9+5} \log_2 \frac{5}{9+5} \right) \right] \\
&= - \left[\left(\frac{9}{14} \log_2 \frac{9}{14} \right) + \left(\frac{5}{14} \log_2 \frac{5}{14} \right) \right] \\
&= - [(0.64 \log_2 0.64) + (0.36 \log_2 0.36)] \\
&= - [(0.64 \times -0.64) + (0.36 \times -1.47)] \\
&= - [(-0.41) + (-0.53)] \\
&= 0.94
\end{aligned}$$

ดังนั้น $Entropy(\text{ความชื้นในอากาศ}) = 0.94$

กึ่งที่ 1 (สูง)

t_1 คือ ความชื้นในอากาศ = สูง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดความชื้นในอากาศ = สูง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดความชื้นในอากาศ = สูง

$$\begin{aligned}
Entropy(\text{ความชื้นในอากาศ} = \text{สูง}) &= - \sum_{k=1}^N \left[\frac{\frac{s_k}{N}}{\sum_{l=1}^N \frac{s_l}{N}} \log_2 \left(\frac{\frac{s_k}{N}}{\sum_{l=1}^N \frac{s_l}{N}} \right) \right] \\
&= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
&= - \left[\left(\frac{3}{3+4} \log_2 \frac{3}{3+4} \right) + \left(\frac{4}{3+4} \log_2 \frac{4}{3+4} \right) \right] \\
&= - \left[\left(\frac{3}{7} \log_2 \frac{3}{7} \right) + \left(\frac{4}{7} \log_2 \frac{4}{7} \right) \right] \\
&= - [(0.43 \log_2 0.43) + (0.57 \log_2 0.57)] \\
&= - [(0.43 \times -1.22) + (0.57 \times -0.81)] \\
&= - [(-0.53) + (-0.46)] \\
&= 0.99
\end{aligned}$$

ดังนั้น $Entropy(\text{ความชื้นในอากาศ} = \text{สูง}) = 0.99$

กึ่งที่ 2 (ปกติ)

t_2 คือ ความชื้นในอากาศ = ปกติ

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดความชื้นในอากาศ = ปกติ

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดความชื้นในอากาศ = ปกติ

$$\begin{aligned}
 Entropy(\text{ความชื้นในอากาศ} = \text{ปกติ}) &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{6}{6 + 1} \log_2 \frac{6}{6 + 1} \right) + \left(\frac{1}{6 + 1} \log_2 \frac{1}{6 + 1} \right) \right] \\
 &= - \left[\left(\frac{6}{7} \log_2 \frac{6}{7} \right) + \left(\frac{1}{7} \log_2 \frac{1}{7} \right) \right] \\
 &= - [(0.86 \log_2 0.86) + (0.14 \log_2 0.14)] \\
 &= - [(0.86 \times -0.22) + (0.14 \times -2.84)] \\
 &= - [(-0.19) + (-0.39)] \\
 &= 0.58
 \end{aligned}$$

ดังนั้น $Entropy(\text{ความชื้นในอากาศ} = \text{ปกติ}) = 0.58$

จะได้ว่า

$$\begin{aligned}
 IG(\text{ความชื้นในอากาศ}) &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
 &= Entropy(\text{ความชื้นในอากาศ}) \\
 &\quad - \frac{n_1}{n_1 + n_2} Entropy(\text{ความชื้นในอากาศ} = \text{เย็น}) \\
 &\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{ความชื้นในอากาศ} = \text{ร้อน}) \\
 &= 0.94 - \left(\frac{7}{7+7} \times 0.99 \right) - \left(\frac{7}{7+7} \times 0.58 \right) \\
 &= 0.94 - \left(\frac{7}{14} \times 0.99 \right) - \left(\frac{7}{14} \times 0.58 \right) \\
 &= 0.94 - (0.5 \times 0.99) - (0.5 \times 0.58) \\
 &= 0.94 - 0.50 - 0.29 \\
 &= 0.15
 \end{aligned}$$

ดังนั้น $IG(\text{ความชื้นในอากาศ}) = 0.15$

4. คำนวณค่า IG ของคุณลักษณะลมแรง

ให้ $T = \text{ลมแรง}$

$M = \text{จำนวนกิ่ง(ไม่จริง,จริง)} = 3$

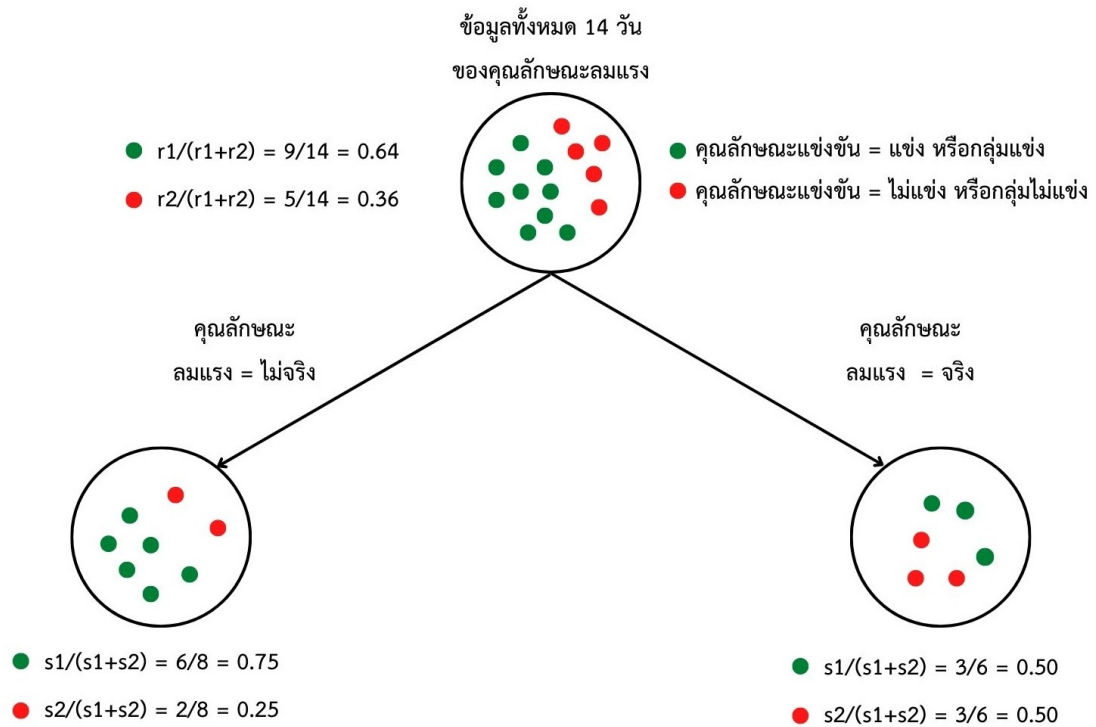
$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (ไม่จริง)} = 8$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (จริง)} = 6$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$

$r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะลมแรง} = 9$

$r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะลมแรง} = 5$



รูปที่ 2.19: แสดงข้อมูลของคุณลักษณะลมแรง

$$\begin{aligned}
 Entropy(\text{ลมแรง}) &= - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right] \\
 &= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
 &= - \left[\left(\frac{9}{9 + 5} \log_2 \frac{9}{9 + 5} \right) + \left(\frac{5}{9 + 5} \log_2 \frac{5}{9 + 5} \right) \right] \\
 &= - \left[\left(\frac{9}{14} \log_2 \frac{9}{14} \right) + \left(\frac{5}{14} \log_2 \frac{5}{14} \right) \right] \\
 &= - [(0.64 \log_2 0.64) + (0.36 \log_2 0.36)] \\
 &= - [(0.64 \times -0.64) + (0.36 \times -1.47)] \\
 &= - [(-0.41) + (-0.53)] \\
 &= 0.94
 \end{aligned}$$

ดังนั้น $Entropy(\text{ลมแรง}) = 0.94$

กึ่งที่ 1 (ไม่จริง)

t_1 คือ ลมแรง = ไม่จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดลมแรง = ไม่จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดลมแรง = ไม่จริง

$$\begin{aligned} Entropy(\text{ลมแรง} = \text{ไม่จริง}) &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\ &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\ &= - \left[\left(\frac{6}{6+2} \log_2 \frac{6}{6+2} \right) + \left(\frac{2}{6+2} \log_2 \frac{2}{6+2} \right) \right] \\ &= - \left[\left(\frac{6}{8} \log_2 \frac{6}{8} \right) + \left(\frac{2}{8} \log_2 \frac{2}{8} \right) \right] \\ &= - [(0.75 \log_2 0.75) + (0.25 \log_2 0.25)] \\ &= - [(0.75 \times -0.42) + (0.25 \times -2)] \\ &= - [(-0.31) + (-0.5)] \\ &= 0.81 \end{aligned}$$

ดังนั้น $Entropy(\text{ลมแรง} = \text{ไม่จริง}) = 0.81$

กึ่งที่ 2 (จริง)

t_2 คือ ลมแรง = จริง

n_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดลมแรง = จริง

n_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดลมแรง = จริง

$$\begin{aligned}
Entropy(\text{ลมแรง} = \text{จริง}) &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
&= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
&= - \left[\left(\frac{3}{3+3} \log_2 \frac{3}{3+3} \right) + \left(\frac{3}{3+3} \log_2 \frac{3}{3+3} \right) \right] \\
&= - \left[\left(\frac{3}{6} \log_2 \frac{3}{6} \right) + \left(\frac{3}{6} \log_2 \frac{3}{6} \right) \right] \\
&= - [(0.5 \log_2 0.5) + (0.5 \log_2 0.5)] \\
&= - [(0.5 \times -1) + (0.5 \times -1)] \\
&= - [(-0.5) + (-0.5)] \\
&= 1
\end{aligned}$$

ดังนั้น $Entropy(\text{ลมแรง} = \text{จริง}) = 1$

จะได้ว่า

$$\begin{aligned}
IG(\text{ลมแรง}) &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
&= Entropy(\text{ลมแรง}) - \frac{n_1}{n_1 + n_2} Entropy(\text{ลมแรง} = \text{ไม่จริง}) \\
&\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{ลมแรง} = \text{จริง}) \\
&= 0.94 - \left(\frac{8}{8+6} \times 0.81 \right) - \left(\frac{6}{8+6} \times 1 \right) \\
&= 0.94 - \left(\frac{8}{14} \times 0.81 \right) - \left(\frac{6}{14} \times 1 \right) \\
&= 0.94 - (0.57 \times 0.81) - (0.43 \times 1) \\
&= 0.94 - 0.46 - 0.43 \\
&= 0.05
\end{aligned}$$

ดังนั้น $IG(\text{ลมแรง}) = 0.05$

จากการคำนวณค่า IG ของคุณลักษณะต่าง ๆ จะได้ว่า

$$IG(\text{สภาพภูมิอากาศ}) = 0.26$$

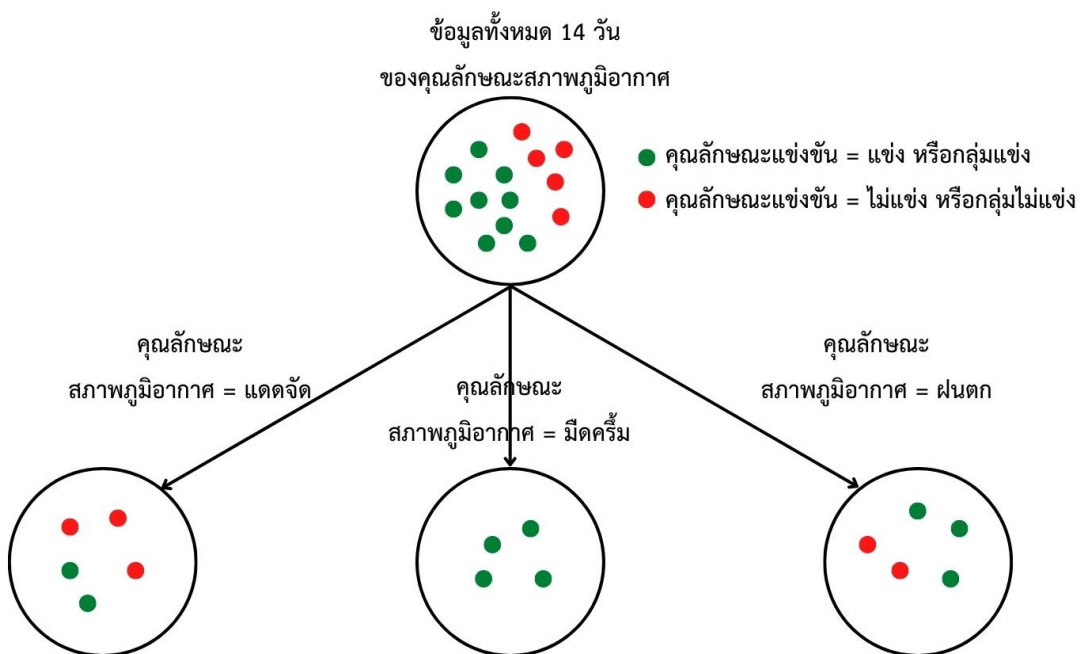
$$IG(\text{อุณหภูมิ}) = 0.03$$

$$IG(\text{ความชื้นในอากาศ}) = 0.15$$

$$IG(\text{ลมแรง}) = 0.05$$

เนื่องจาก คุณลักษณะที่มีค่า IG ที่มากที่สุดจะเป็นคุณลักษณะที่มีความหลากหลายของข้อมูลกลุ่มน้อยที่สุด ดังนั้น เราจะหาค่า IG ที่มากที่สุดเป็นไหนดราก

จะเห็นได้ว่า ค่า IG ของคุณลักษณะสภาพภูมิอากาศ มีค่ามากที่สุด คือ 0.26 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศเป็นไหนดราก



รูปที่ 2.20: แสดงข้อมูลของคุณลักษณะสภาพภูมิอากาศ

จะเห็นว่า ในไหนดคุณลักษณะสภาพภูมิอากาศ = มีดครึ้ม มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = แข่ง ดังนั้น ไหนดคุณลักษณะสภาพภูมิอากาศ = มีดครึ้ม จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ แต่ไหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และไหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก ต้องทำการแตกกิ่งจนกว่าข้อมูลในแต่ละไหนดจะมีข้อมูลเป็นกลุ่มเดียวกัน

ต่อไปจะแสดงการคำนวณการแตกกิ่งของโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด โดยสร้างตารางใหม่ที่สภาพภูมิอากาศ = แดดจัด ได้ดังนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
1	แดดจัด	ร้อน	สูง	ไม่จริง	ไม่แข่ง
2	แดดจัด	ร้อน	สูง	จริง	ไม่แข่ง
8	แดดจัด	อบอุ่น	สูง	ไม่จริง	ไม่แข่ง
9	แดดจัด	อบอุ่น	ปกติ	ไม่จริง	แข่ง
11	แดดจัด	อบอุ่น	ปกติ	จริง	แข่ง

ตารางที่ 2.5: แสดงข้อมูลสภาพอากาศที่สภาพภูมิอากาศ = แดดจัด

1. คำนวณค่า IG ของคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

ให้ $T = (\text{สภาพภูมิอากาศ} = \text{แดดจัด})$ และอุณหภูมิ

$M = \text{จำนวนกิ่ง(ร้อน,อบอุ่น)} = 2$

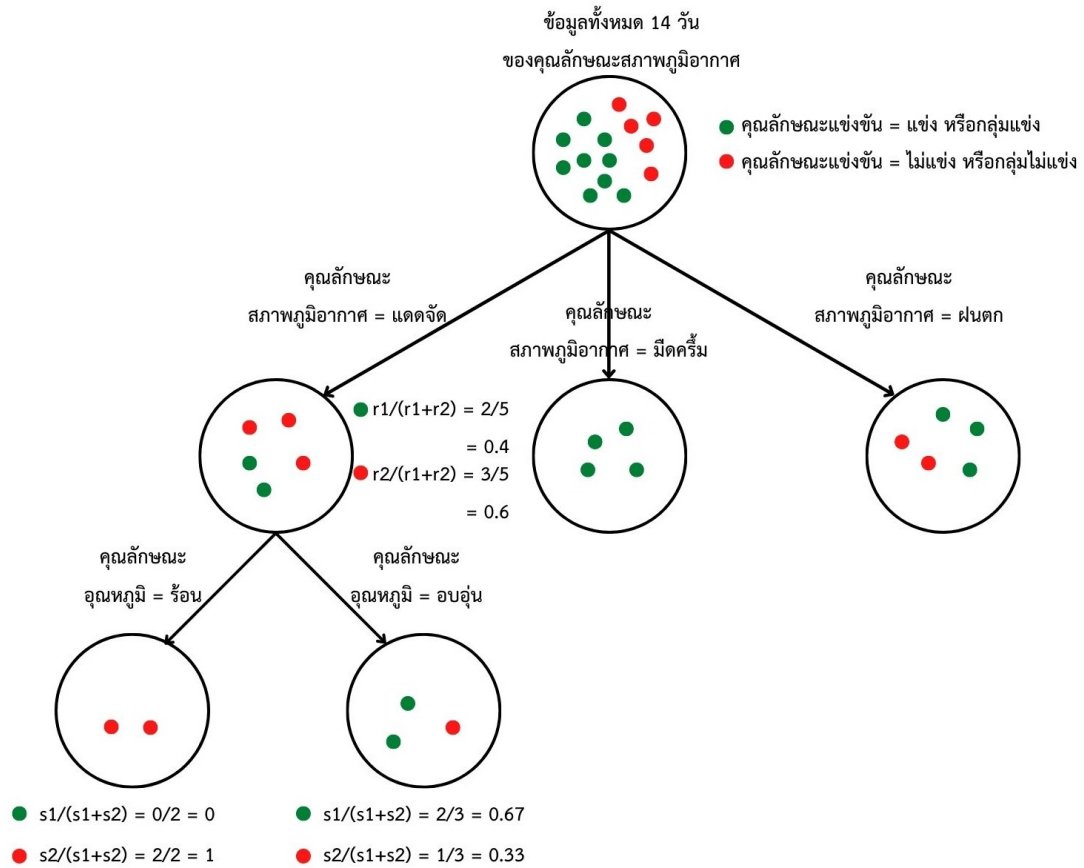
$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (ร้อน)} = 2$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (อบอุ่น)} = 3$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$

$r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = แดดจัด) และอุณหภูมิ} = 2$

$r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = แดดจัด) และอุณหภูมิ} = 3$



รูปที่ 2.21: แสดงข้อมูลคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

$$\begin{aligned}
 Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) &= - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right] \\
 &= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
 &= - \left[\left(\frac{2}{2+3} \log_2 \frac{2}{2+3} \right) + \left(\frac{3}{2+3} \log_2 \frac{3}{2+3} \right) \right] \\
 &= - \left[\left(\frac{2}{5} \log_2 \frac{2}{5} \right) + \left(\frac{3}{5} \log_2 \frac{3}{5} \right) \right] \\
 &= - [(0.4 \log_2 0.4) + (0.6 \log_2 0.6)] \\
 &= - [(0.4 \times -1.32) + (0.6 \times -0.74)] \\
 &= - [(-0.53) + (-0.44)] \\
 &= 0.97
 \end{aligned}$$

ดังนั้น $Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) = 0.97$

กึ่งที่ 1 (ร้อน)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

$Entropy$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน)

$$\begin{aligned} &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\ &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\ &= - \left[\left(\frac{0}{0 + 2} \log_2 \frac{0}{0 + 2} \right) + \left(\frac{2}{0 + 2} \log_2 \frac{2}{0 + 2} \right) \right] \\ &= - \left[\left(\frac{0}{2} \log_2 \frac{0}{2} \right) + \left(\frac{2}{2} \log_2 \frac{2}{2} \right) \right] \\ &= - [(0 \log_2 0) + (1 \log_2 1)] \\ &= - [0 + (1 \times 0)] \\ &= - [0 + 0] \\ &= 0 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน) = 0

กึ่งที่ 2 (อบอุ่น)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

$Entropy$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{2}{2+1} \log_2 \frac{2}{2+1} \right) + \left(\frac{1}{2+1} \log_2 \frac{1}{2+1} \right) \right] \\
 &= - \left[\left(\frac{2}{3} \log_2 \frac{2}{3} \right) + \left(\frac{1}{3} \log_2 \frac{1}{3} \right) \right] \\
 &= - [(0.67 \log_2 0.67) + (0.33 \log_2 0.33)] \\
 &= - [(0.67 \times -0.58) + (0.33 \times -1.60)] \\
 &= - [(-0.39) + (-0.53)] \\
 &= 0.92
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น) = 0.92

จะได้ว่า

IG (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ)

$$\begin{aligned}
 &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
 &= Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) \\
 &\quad - \frac{n_1}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{ร้อน}) \\
 &\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{อบอุ่น}) \\
 &= 0.97 - \left(\frac{2}{2+3} \times 0 \right) - \left(\frac{3}{2+3} \times 0.92 \right) \\
 &= 0.97 - \left(\frac{2}{5} \times 0 \right) - \left(\frac{3}{5} \times 0.92 \right) \\
 &= 0.42
 \end{aligned}$$

ดังนั้น IG (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ) = 0.42

2. คำนวณค่า IG ของคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

ให้ $T = (\text{สภาพภูมิอากาศ} = \text{แดดจัด})$ และความชื้นในอากาศ

$M = \text{จำนวนกิ่ง(สูง,ปกติ)} = 2$

$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (สูง)} = 3$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (ปกติ)} = 2$

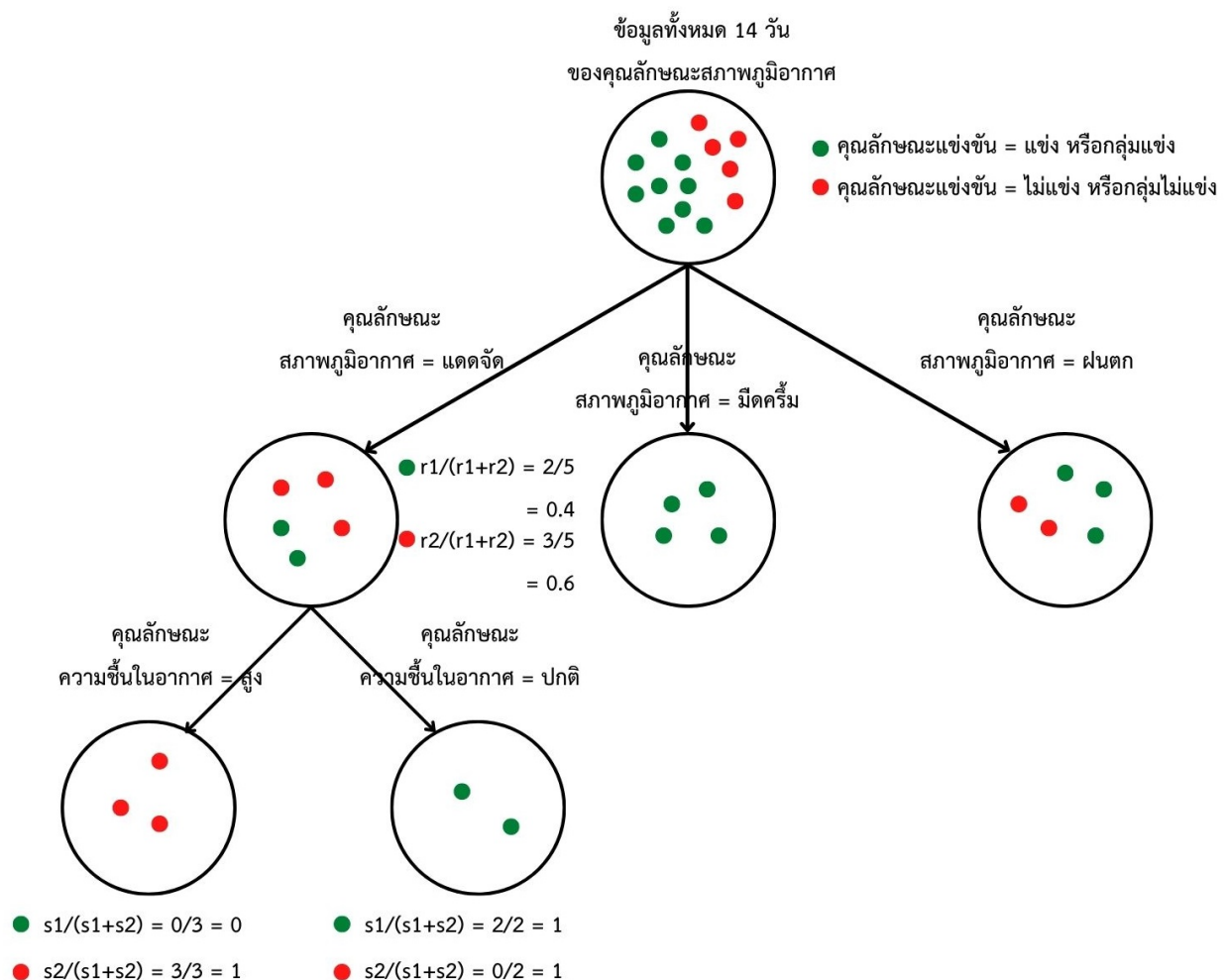
$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$

$r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = แดดจัด)}$

และความชื้นในอากาศ = 2

$r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = แดดจัด)}$

และความชื้นในอากาศ = 3



รูปที่ 2.22: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

$Entropy$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right] \\
 &= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
 &= - \left[\left(\frac{2}{2+3} \log_2 \frac{2}{2+3} \right) + \left(\frac{3}{2+3} \log_2 \frac{3}{2+3} \right) \right] \\
 &= - \left[\left(\frac{2}{5} \log_2 \frac{2}{5} \right) + \left(\frac{3}{5} \log_2 \frac{3}{5} \right) \right] \\
 &= - [(0.4 \log_2 0.4) + (0.6 \log_2 0.6)] \\
 &= - [(0.4 \times -1.32) + (0.6 \times -0.74)] \\
 &= - [(-0.53) + (-0.44)] \\
 &= 0.97
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ) = 0.97

กึ่งที่ 1 (สูง)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

$Entropy$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{0}{0 + 3} \log_2 \frac{0}{0 + 3} \right) + \left(\frac{3}{0 + 3} \log_2 \frac{3}{0 + 3} \right) \right] \\
 &= - \left[\left(\frac{0}{3} \log_2 \frac{0}{3} \right) + \left(\frac{3}{3} \log_2 \frac{3}{3} \right) \right] \\
 &= - [(0 \log_2 0) + (1 \log_2 1)] \\
 &= - [0 + (1 \times 0)] \\
 &= - [0 + 0] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง) = 0

กึ่งที่ 2 (ปกติ)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ

$Entropy$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{2}{2+0} \log_2 \frac{2}{2+0} \right) + \left(\frac{0}{2+0} \log_2 \frac{0}{2+0} \right) \right] \\
 &= - \left[\left(\frac{2}{2} \log_2 \frac{2}{2} \right) + \left(\frac{0}{2} \log_2 \frac{0}{2} \right) \right] \\
 &= - [(1 \log_2 1) + (0 \log_2 0)] \\
 &= - [(1 \times 0) + 0] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ) = 0

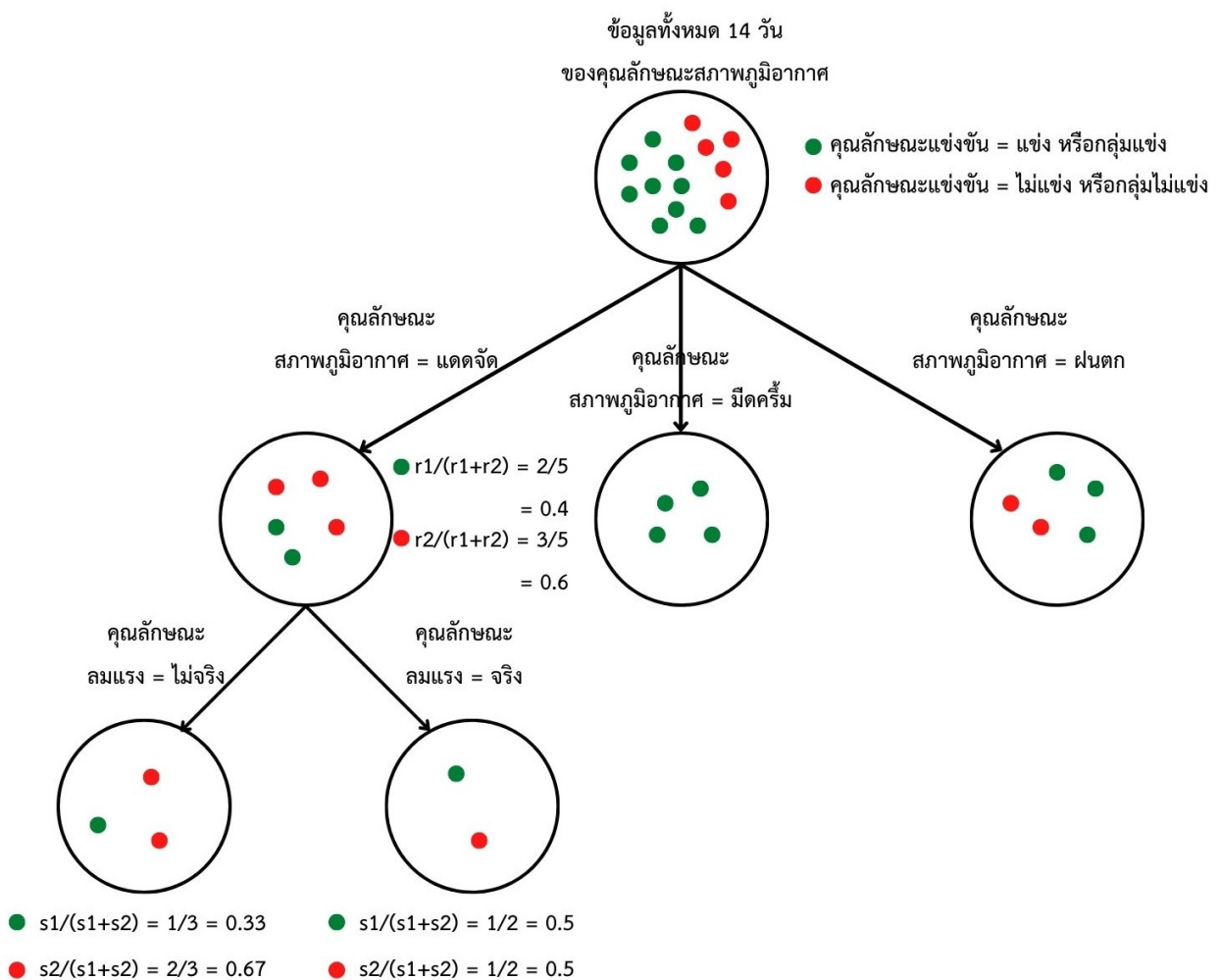
จะได้ว่า

IG (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ)

$$\begin{aligned}
 &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
 &= Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ}) \\
 &\quad - \frac{n_1}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ} = \text{สูง}) \\
 &\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ} = \text{ปกติ}) \\
 &= 0.97 - \left(\frac{3}{3+2} \times 0 \right) - \left(\frac{2}{3+2} \times 0 \right) \\
 &= 0.97 - \left(\frac{3}{5} \times 0 \right) - \left(\frac{2}{5} \times 0 \right) \\
 &= 0.97 - (0.6 \times 0) - (0.4 \times 0) \\
 &= 0.97
 \end{aligned}$$

ดังนั้น IG (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ) = 0.97

3. คำนวณค่า IG ของคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด
- ให้ $T = (\text{สภาพภูมิอากาศ} = \text{แดดจัด})$ และลมแรง
- $M = \text{จำนวนกิ่ง(ไม่จริง,จริง)} = 2$
- $n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (ไม่จริง)} = 3$
- $n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (จริง)} = 2$
- $N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$
- $r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = แดดจัด) และลมแรง} = 2$
- $r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = แดดจัด) และลมแรง} = 3$



รูปที่ 2.23: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

Entropy(สภาพภูมิอากาศ = แดดจัด และลมแรง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right] \\
 &= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
 &= - \left[\left(\frac{2}{2+3} \log_2 \frac{2}{2+3} \right) + \left(\frac{3}{2+3} \log_2 \frac{3}{2+3} \right) \right] \\
 &= - \left[\left(\frac{2}{5} \log_2 \frac{2}{5} \right) + \left(\frac{3}{5} \log_2 \frac{3}{5} \right) \right] \\
 &= - [(0.4 \log_2 0.4) + (0.6 \log_2 0.6)] \\
 &= - [(0.4 \times -1.32) + (0.6 \times -0.74)] \\
 &= - [(-0.53) + (-0.44)] \\
 &= 0.97
 \end{aligned}$$

ดังนั้น *Entropy* (สภาพภูมิอากาศ = แดดจัด และลมแรง) = 0.97

กึ่งที่ 1 (ไม่จริง)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

Entropy(สภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{1}{1+2} \log_2 \frac{1}{1+2} \right) + \left(\frac{2}{1+2} \log_2 \frac{2}{1+2} \right) \right] \\
 &= - \left[\left(\frac{1}{3} \log_2 \frac{1}{3} \right) + \left(\frac{2}{3} \log_2 \frac{2}{3} \right) \right] \\
 &= - [(0.33 \log_2 0.33) + (0.67 \log_2 0.67)] \\
 &= - [(0.33 \times -1.60) + (0.67 \times -0.58)] \\
 &= - [(-0.53) + (-0.39)] \\
 &= 0.92
 \end{aligned}$$

ดังนั้น *Entropy* (สภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง) = 0.92

กึ่งที่ 2 (จริง)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

$Entropy$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{1}{1+1} \log_2 \frac{1}{1+1} \right) + \left(\frac{1}{1+1} \log_2 \frac{1}{1+1} \right) \right] \\
 &= - \left[\left(\frac{1}{2} \log_2 \frac{1}{2} \right) + \left(\frac{1}{2} \log_2 \frac{1}{2} \right) \right] \\
 &= - [(0.5 \log_2 0.5) + (0.5 \log_2 0.5)] \\
 &= - [(0.5 \times -1) + (0.5 \times -1)] \\
 &= - [(-0.5) + (-0.5)] \\
 &= 1
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง) = 1
 จะได้ว่า

IG (สภาพภูมิอากาศ = แดดจัด และลมแรง)

$$\begin{aligned}
 &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
 &= Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง}) \\
 &\quad - \frac{n_1}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{ไม่จริง}) \\
 &\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{จริง}) \\
 &= 0.97 - \left(\frac{3}{3+2} \times 0.92 \right) - \left(\frac{2}{3+2} \times 1 \right) \\
 &= 0.97 - \left(\frac{3}{5} \times 0.92 \right) - \left(\frac{2}{5} \times 1 \right) \\
 &= 0.97 - (0.6 \times 0.92) - (0.4 \times 1) \\
 &= 0.97 - 0.55 - 0.4 \\
 &= 0.02
 \end{aligned}$$

ดังนั้น $IG(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง}) = 0.02$

จากการคำนวณค่า IG ของคุณลักษณะต่าง ๆ เมื่อแตกกิ่งออกจากโหนดสภาพภูมิอากาศ = แดดจัด จะได้ว่า

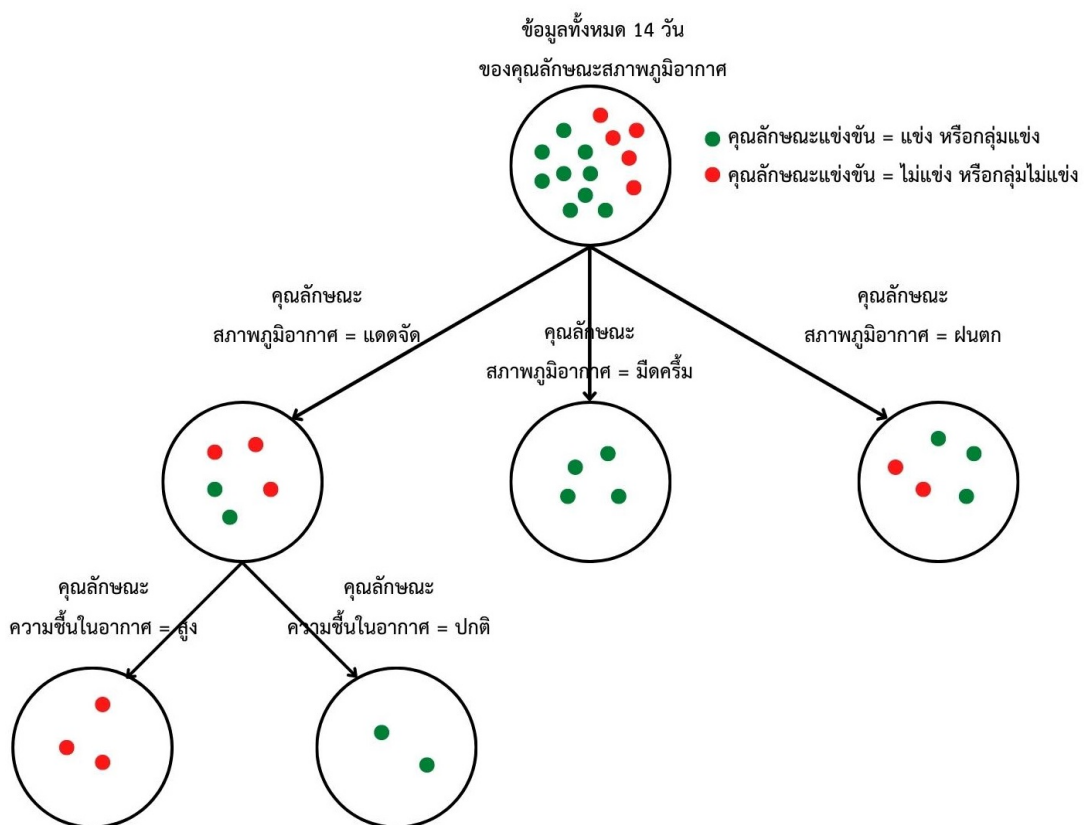
$$IG(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) = 0.42$$

$$IG(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ}) = 0.97$$

$$IG(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง}) = 0.02$$

จะเห็นว่า ค่า IG ของคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ มีค่ามากที่สุด คือ 0.97

ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ



รูปที่ 2.24: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = ไม่แข่ง และโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = แข่ง ดังนั้น จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ

ต่อไปจะแสดงการคำนวณการแตกกิ่งของโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก โดยสร้างตารางใหม่ที่สภาพภูมิอากาศ = ฝนตก ดังนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
4	ฝนตก	อบอุ่น	สูง	ไม่จริง	แข่ง
5	ฝนตก	เย็น	ปกติ	ไม่จริง	แข่ง
6	ฝนตก	เย็น	ปกติ	จริง	ไม่แข่ง
10	ฝนตก	อบอุ่น	ปกติ	ไม่จริง	แข่ง
14	ฝนตก	อบอุ่น	สูง	จริง	ไม่แข่ง

ตารางที่ 2.6: แสดงข้อมูลสภาพอากาศที่สภาพภูมิอากาศ = ฝนตก

1. คำนวณค่า IG ของคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

ให้ $T = (\text{สภาพภูมิอากาศ} = \text{ฝนตก})$ และอุณหภูมิ

$$M = \text{จำนวนกิ่ง(อบอุ่น,เย็น)} = 2$$

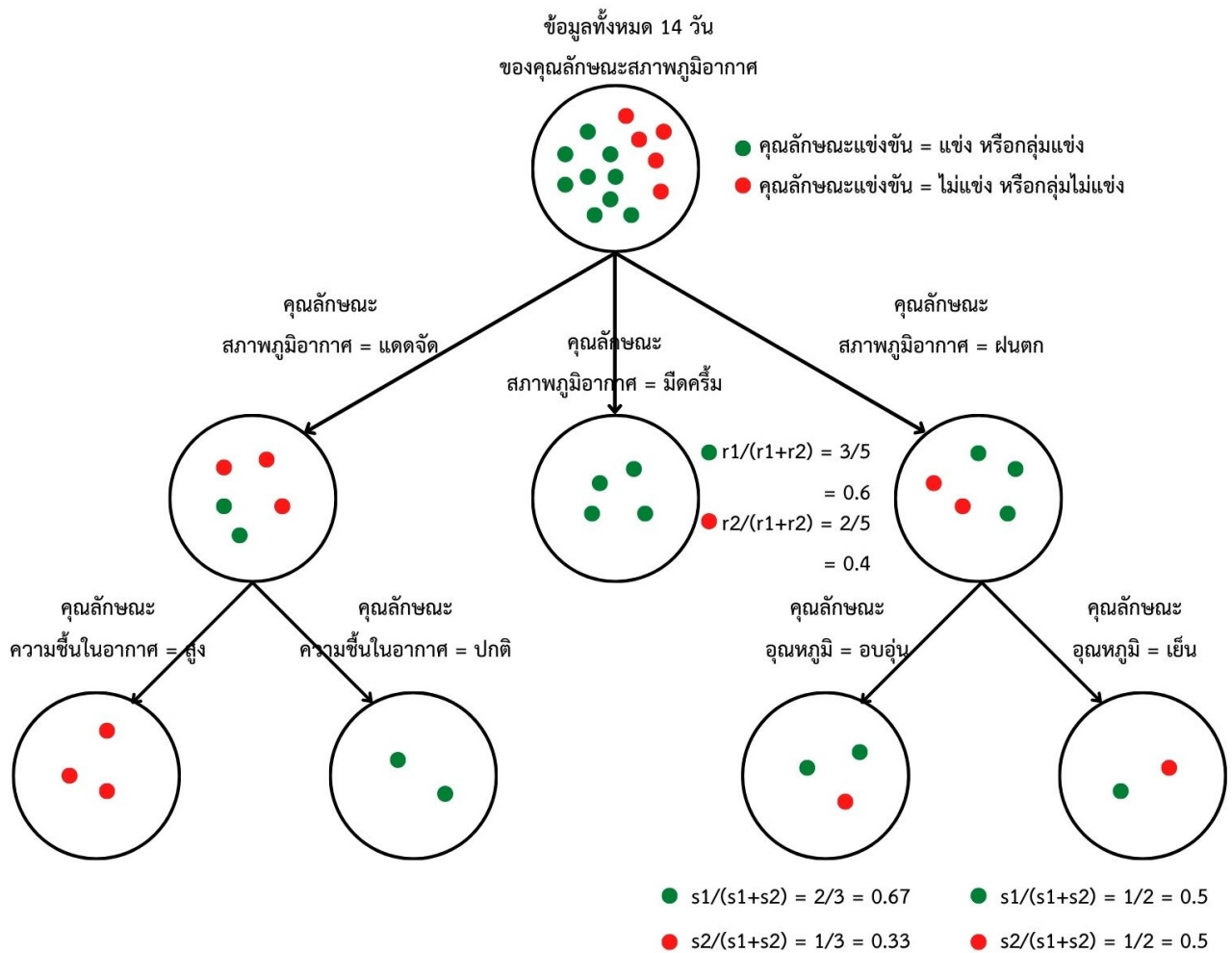
$$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (อบอุ่น)} = 3$$

$$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (เย็น)} = 2$$

$$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$$

$$r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = ฝนตก) และอุณหภูมิ} = 3$$

$$r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = ฝนตก) และอุณหภูมิ} = 2$$



รูปที่ 2.25: แสดงข้อมูลคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

Entropy(สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right] \\
 &= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
 &= - \left[\left(\frac{3}{3 + 2} \log_2 \frac{3}{3 + 2} \right) + \left(\frac{2}{3 + 2} \log_2 \frac{2}{3 + 2} \right) \right] \\
 &= - \left[\left(\frac{3}{5} \log_2 \frac{3}{5} \right) + \left(\frac{2}{5} \log_2 \frac{2}{5} \right) \right] \\
 &= - [(0.6 \log_2 0.6) + (0.4 \log_2 0.4)] \\
 &= - [(0.6 \times -0.74) + (0.4 \times -1.32)] \\
 &= - [(-0.44) + (-0.53)] \\
 &= 0.97
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = 0.97

กึ่งที่ 1 (อบอุ่น)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น)

$$\begin{aligned} &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\ &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\ &= - \left[\left(\frac{2}{2+1} \log_2 \frac{2}{2+1} \right) + \left(\frac{1}{2+1} \log_2 \frac{1}{2+1} \right) \right] \\ &= - \left[\left(\frac{2}{3} \log_2 \frac{2}{3} \right) + \left(\frac{1}{3} \log_2 \frac{1}{3} \right) \right] \\ &= - [(0.67 \log_2 0.67) + (0.33 \log_2 0.33)] \\ &= - [(0.67 \times -0.58) + (0.33 \times -1.60)] \\ &= - [(-0.39) + (-0.53)] \\ &= 0.92 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น) = 0.92

กึ่งที่ 2 (เย็น)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{1}{1+1} \log_2 \frac{1}{1+1} \right) + \left(\frac{1}{1+1} \log_2 \frac{1}{1+1} \right) \right] \\
 &= - \left[\left(\frac{1}{2} \log_2 \frac{1}{2} \right) + \left(\frac{1}{2} \log_2 \frac{1}{2} \right) \right] \\
 &= - [(0.5 \log_2 0.5) + (0.5 \log_2 0.5)] \\
 &= - [(0.5 \times -1) + (0.5 \times -1)] \\
 &= - [(-0.5) + (-0.5)] \\
 &= 1
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น) = 1
 จะได้ว่า

IG (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ)

$$\begin{aligned}
 &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
 &= Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ}) \\
 &\quad - \frac{n_1}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{อบอุ่น}) \\
 &\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{เย็น}) \\
 &= 0.97 - \left(\frac{3}{3+2} \times 0.92 \right) - \left(\frac{2}{3+2} \times 1 \right) \\
 &= 0.97 - \left(\frac{3}{5} \times 0.92 \right) - \left(\frac{2}{5} \times 1 \right) \\
 &= 0.97 - (0.6 \times 0.92) - (0.4 \times 1) \\
 &= 0.97 - 0.55 - 0.4 \\
 &= 0.02
 \end{aligned}$$

ดังนั้น $IG(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ}) = 0.02$

2. คำนวณค่า IG ของคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

ให้ $T = (\text{สภาพภูมิอากาศ} = \text{ฝนตก})$ และความชื้นในอากาศ

$M = \text{จำนวนกิ่ง(สูง,ปกติ)} = 2$

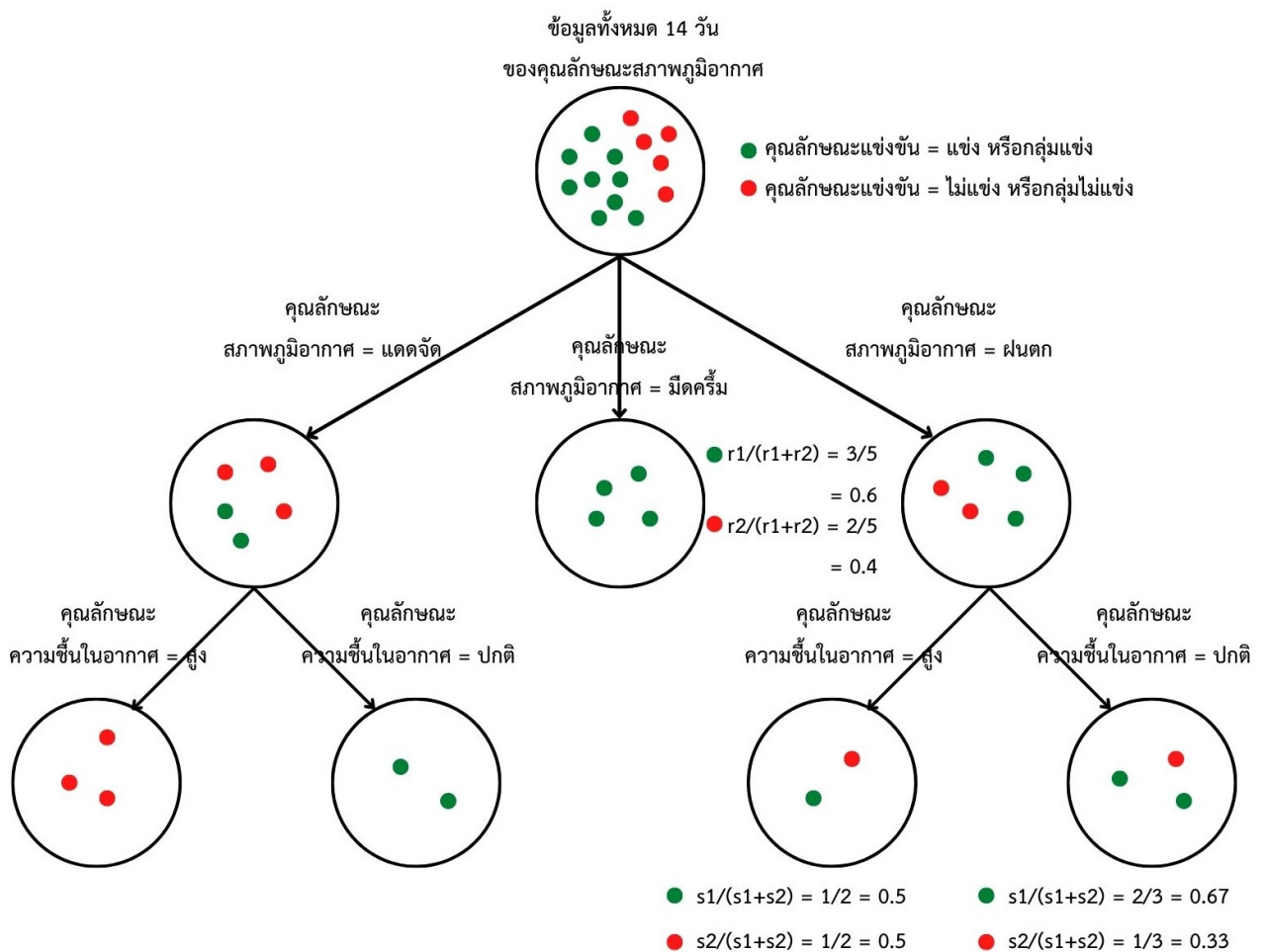
$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (สูง)} = 2$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (ปกติ)} = 3$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$

$r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = ฝนตก) และความชื้นในอากาศ} = 3$

$r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = ฝนตก) และความชื้นในอากาศ} = 2$



รูปที่ 2.26: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right] \\
 &= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
 &= - \left[\left(\frac{3}{3+2} \log_2 \frac{3}{3+2} \right) + \left(\frac{2}{3+2} \log_2 \frac{2}{3+2} \right) \right] \\
 &= - \left[\left(\frac{3}{5} \log_2 \frac{3}{5} \right) + \left(\frac{2}{5} \log_2 \frac{2}{5} \right) \right] \\
 &= - [(0.6 \log_2 0.6) + (0.4 \log_2 0.4)] \\
 &= - [(0.6 \times -0.74) + (0.4 \times -1.32)] \\
 &= - [(-0.44) + (-0.53)] \\
 &= 0.97
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ) = 0.97

กึ่งที่ 1 (สูง)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{1}{2} \log_2 \frac{1}{2} \right) + \left(\frac{1}{2} \log_2 \frac{1}{2} \right) \right] \\
 &= - \left[\left(\frac{1}{1+1} \log_2 \frac{1}{1+1} \right) + \left(\frac{1}{1+1} \log_2 \frac{1}{1+1} \right) \right] \\
 &= - [(0.5 \log_2 0.5) + (0.5 \log_2 0.5)] \\
 &= - [(0.5 \times -1) + (0.5 \times -1)] \\
 &= - [(-0.5) + (-0.5)] \\
 &= 1
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง) = 1

กึ่งที่ 2 (ปกติ)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{2}{2+1} \log_2 \frac{2}{2+1} \right) + \left(\frac{1}{2+1} \log_2 \frac{1}{2+1} \right) \right] \\
 &= - \left[\left(\frac{2}{3} \log_2 \frac{2}{3} \right) + \left(\frac{1}{3} \log_2 \frac{1}{3} \right) \right] \\
 &= - [(0.67 \log_2 0.67) + (0.33 \log_2 0.33)] \\
 &= - [(0.67 \times -0.58) + (0.33 \times -1.60)] \\
 &= - [(-0.39) + (-0.53)] \\
 &= 0.92
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ) = 0.92
 จะได้ว่า

IG (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ)

$$\begin{aligned}
 &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
 &= Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ}) \\
 &\quad - \frac{n_1}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ} = \text{สูง}) \\
 &\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ} = \text{ปกติ}) \\
 &= 0.97 - \left(\frac{2}{3+2} \times 1 \right) - \left(\frac{3}{3+2} \times 0.92 \right) \\
 &= 0.97 - \left(\frac{2}{5} \times 1 \right) - \left(\frac{3}{5} \times 0.92 \right) \\
 &= 0.97 - (0.4 \times 1) - (0.6 \times 0.92) \\
 &= 0.97 - 0.4 - 0.55 \\
 &= 0.02
 \end{aligned}$$

ดังนั้น $IG(\text{สภาพภูมิอากาศ} = \text{ฝนตก} \text{ และ } \text{ความชื้นในอากาศ}) = 0.02$

3. คำนวณค่า IG ของคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

ให้ $T = (\text{สภาพภูมิอากาศ} = \text{ฝนตก})$ และลมแรง

$M = \text{จำนวนกิ่ง(ไม่จริง,จริง)} = 2$

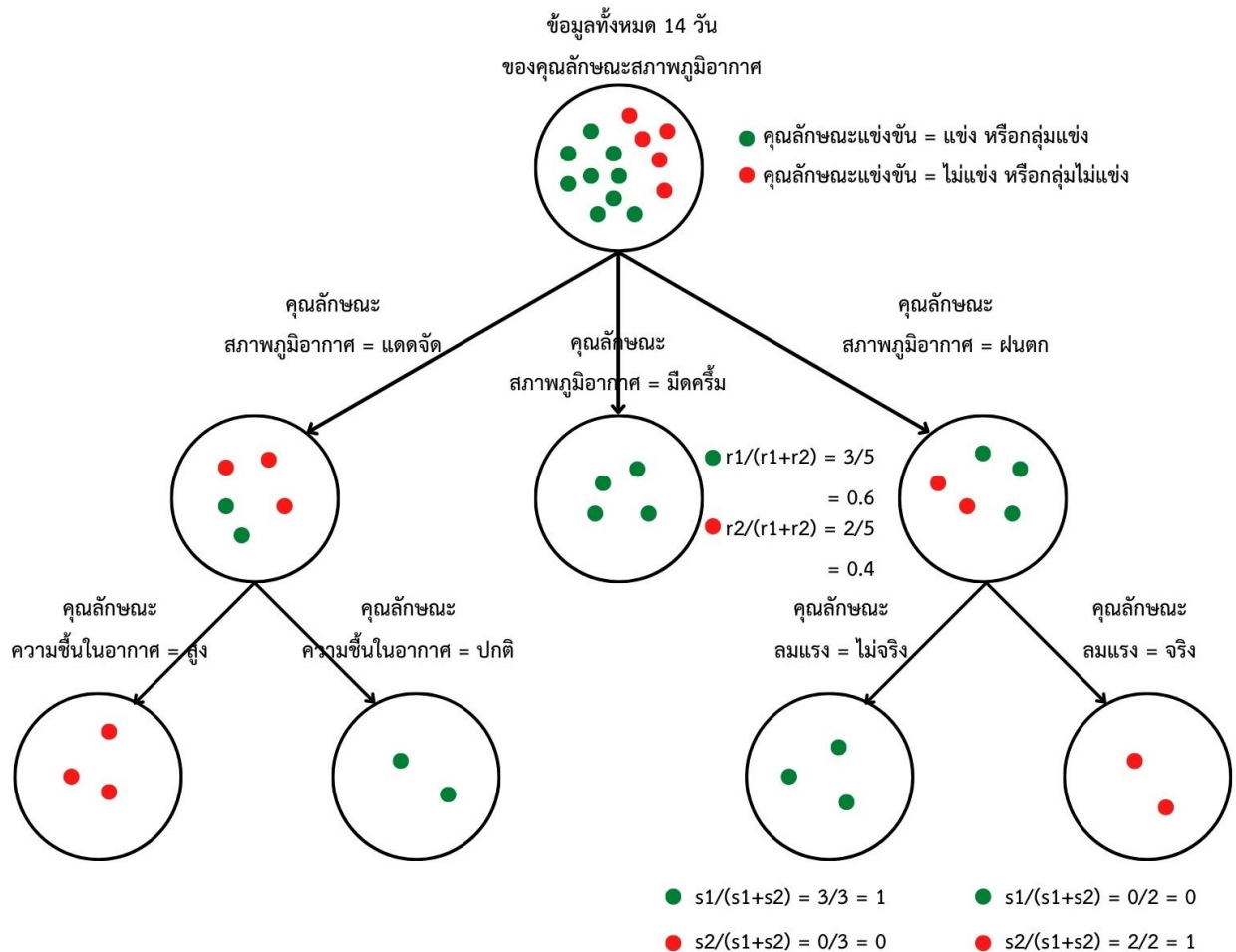
$n_1 = \text{จำนวนข้อมูลในกิ่งที่ 1 (ไม่จริง)} = 3$

$n_2 = \text{จำนวนข้อมูลในกิ่งที่ 2 (จริง)} = 2$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$

$r_1 = \text{จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = ฝนตก) และลมแรง} = 3$

$r_2 = \text{จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) ของคุณลักษณะ(สภาพภูมิอากาศ = ฝนตก) และลมแรง} = 2$



รูปที่ 2.27: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และลมแรง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right] \\
 &= - \left[\left(\frac{r_1}{r_1 + r_2} \log_2 \frac{r_1}{r_1 + r_2} \right) + \left(\frac{r_2}{r_1 + r_2} \log_2 \frac{r_2}{r_1 + r_2} \right) \right] \\
 &= - \left[\left(\frac{3}{3+2} \log_2 \frac{3}{3+2} \right) + \left(\frac{2}{3+2} \log_2 \frac{2}{3+2} \right) \right] \\
 &= - \left[\left(\frac{3}{5} \log_2 \frac{3}{5} \right) + \left(\frac{2}{5} \log_2 \frac{2}{5} \right) \right] \\
 &= - [(0.6 \log_2 0.6) + (0.4 \log_2 0.4)] \\
 &= - [(0.6 \times -0.74) + (0.4 \times -1.32)] \\
 &= - [(-0.44) + (-0.53)] \\
 &= 0.97
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และลมแรง) = 0.97

กึ่งที่ 1 (ไม่จริง)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{3}{3+0} \log_2 \frac{3}{3+0} \right) + \left(\frac{0}{3+0} \log_2 \frac{0}{3+0} \right) \right] \\
 &= - \left[\left(\frac{3}{3} \log_2 \frac{3}{3} \right) + \left(\frac{0}{3} \log_2 \frac{0}{3} \right) \right] \\
 &= - [(1 \log_2 1) + (0 \log_2 0)] \\
 &= - [(1 \times 0) + 0] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง) = 0

กึ่งที่ 2 (จริง)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

s_1 คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

s_2 คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

$Entropy$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง)

$$\begin{aligned}
 &= - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right] \\
 &= - \left[\left(\frac{s_1}{s_1 + s_2} \log_2 \frac{s_1}{s_1 + s_2} \right) + \left(\frac{s_2}{s_1 + s_2} \log_2 \frac{s_2}{s_1 + s_2} \right) \right] \\
 &= - \left[\left(\frac{0}{0 + 2} \log_2 \frac{0}{0 + 2} \right) + \left(\frac{2}{0 + 2} \log_2 \frac{2}{0 + 2} \right) \right] \\
 &= - \left[\left(\frac{0}{2} \log_2 \frac{0}{2} \right) + \left(\frac{2}{2} \log_2 \frac{2}{2} \right) \right] \\
 &= - [(0 \log_2 0) + (1 \log_2 1)] \\
 &= - [0 + (1 \times 0)] \\
 &= - [0 + 0] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Entropy$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง) = 0

จะได้ว่า

IG (สภาพภูมิอากาศ = ฝนตก และลมแรง)

$$\begin{aligned}
 &= Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i) \\
 &= Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง}) \\
 &\quad - \frac{n_1}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{ไม่จริง}) \\
 &\quad - \frac{n_2}{n_1 + n_2} Entropy(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{จริง}) \\
 &= 0.97 - \left(\frac{3}{3 + 2} \times 0 \right) - \left(\frac{2}{3 + 2} \times 0 \right) \\
 &= 0.97 - \left(\frac{3}{5} \times 0 \right) - \left(\frac{2}{5} \times 0 \right) \\
 &= 0.97 - (0.6 \times 0) - (0.4 \times 0) \\
 &= 0.97 - 0 - 0 \\
 &= 0.97
 \end{aligned}$$

ดังนั้น $IG(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง}) = 0.97$

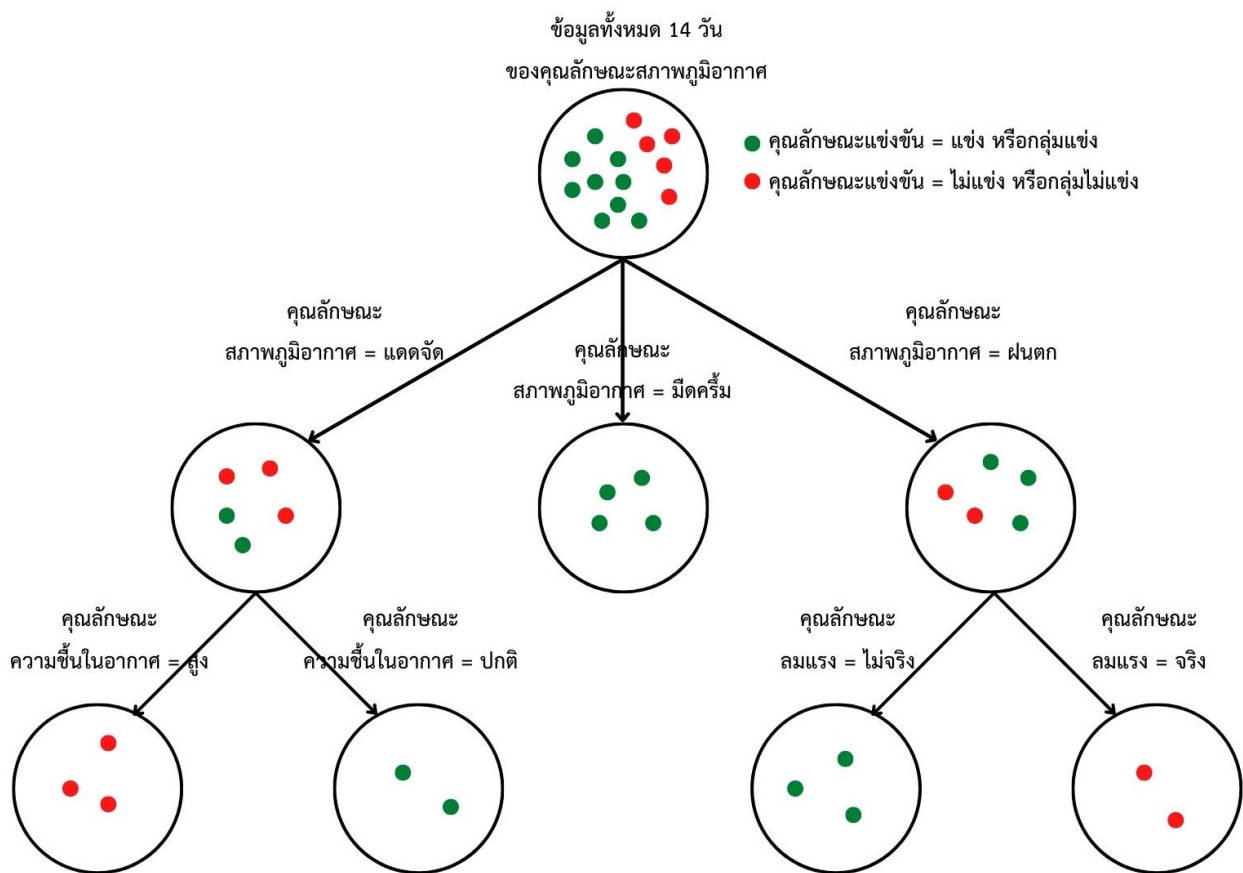
จากการคำนวณค่า IG ของคุณลักษณะต่าง ๆ เมื่อแตกกิ่งออกจากโหนดสภาพภูมิอากาศ = ฝนตก จะได้ว่า

$$IG(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ}) = 0.02$$

$$IG(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ}) = 0.02$$

$$IG(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง}) = 0.97$$

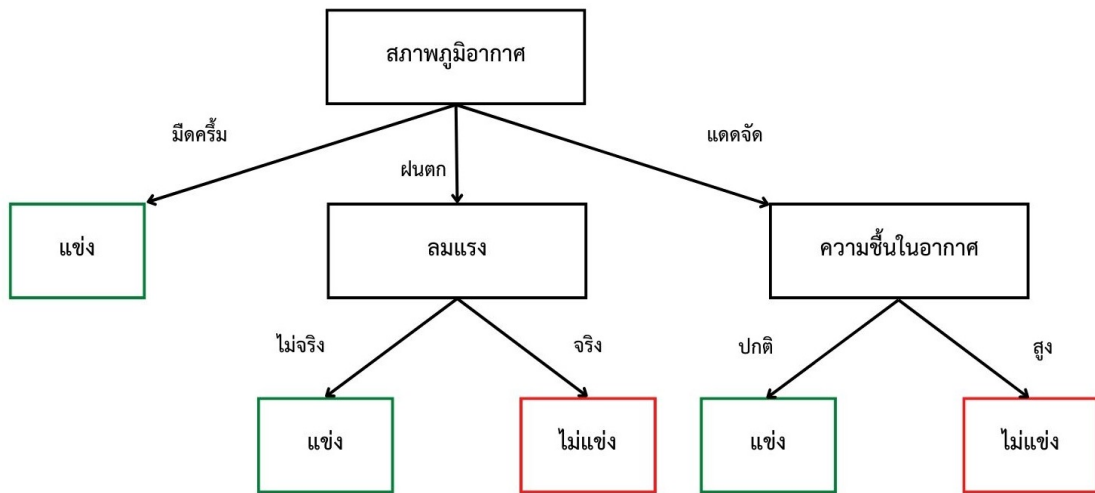
จะเห็นได้ว่า ค่า IG ของคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง มีค่ามากที่สุด คือ 0.97 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง



รูปที่ 2.28: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = แข่ง และโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = ไม่แข่ง ดังนั้น จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ

เนื่องจากในแต่ละโหนดใบมีข้อมูลเป็นกลุ่มเดียวกัน ฉะนั้น จะเสร็จสิ้นในขั้นตอนการคำนวณ ดังนั้น สามารถสร้างโมเดลต้นไม้ตัดสินใจ ได้ดังนี้



รูปที่ 2.29: แสดงโมเดลต้นไม้ตัดสินใจของข้อมูลสภาพอากาศ 14 วันย้อนหลัง สำหรับการแข่งขันกีฬา

เนื่องจากชุดข้อมูลที่นำมาใช้ประกอบการทำการค้นคว้าอิสระมีจำนวนข้อมูลในแต่ละกลุ่มไม่เท่ากัน และจำนวนข้อมูลในแต่ละกลุ่มแตกต่างกันมาก จึงจำเป็นต้องใช้ class weight เข้ามาช่วย เพื่อให้แต่ละกลุ่มมีความสมดุลกัน และให้ความสำคัญแต่ละกลุ่มเท่ากัน

- Gini Index แบบ class weight

$$Gini_{split}(T) = \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i)$$

$$\text{โดยที่ } Gini(t_i) = 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{\sum_{l=1}^N w_l s_{li}}}{\sum_{l=1}^N \frac{w_l s_{li}}{\sum_{l=1}^N w_l s_{li}}} \right]^2$$

$$w_i = \text{ค่า weight ของกลุ่มที่ } i = \frac{\sum_{k=1}^N r_k}{N} \frac{1}{r_i}$$

N = จำนวนกลุ่ม หรือผลลัพธ์ในการจำแนกประเภทข้อมูล

r_k = จำนวนข้อมูลในกลุ่มที่ k

T = โหนดภายใน หรือคุณลักษณะของข้อมูล

t_i = กิ่งที่ i หรือค่าของคุณลักษณะในโหนดภายในที่ i

M = จำนวนกิ่งของคุณลักษณะของข้อมูล

s_{ki} = จำนวนข้อมูลในกลุ่มที่ k ของกิ่งที่ i

- Information Gain(IG) แบบ class weight

$$IG(T) = Entropy(T) - \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Entropy(t_i)$$

$$\text{โดยที่ } Entropy(T) = - \sum_{k=1}^N \left[\frac{\frac{w_k r_k}{\sum_{l=1}^N w_l r_l} \log_2 \left(\frac{\frac{w_k r_k}{\sum_{l=1}^N w_l r_l}}{\sum_{l=1}^N w_l r_l} \right)}{\sum_{l=1}^N w_l r_l} \right]$$

$$Entropy(t_i) = - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{\sum_{l=1}^N w_l s_{li}} \log_2 \left(\frac{\frac{w_k s_{ki}}{\sum_{l=1}^N w_l s_{li}}}{\sum_{l=1}^N w_l s_{li}} \right)}{\sum_{l=1}^N w_l s_{li}} \right]$$

$$w_i = \text{ค่า weight ของกลุ่มที่ } i = \frac{\sum_{k=1}^N r_k}{N} \frac{1}{r_i}$$

N = จำนวนกลุ่ม หรือผลลัพธ์ในการจำแนกประเภทข้อมูล

r_k = จำนวนข้อมูลในกลุ่มที่ k

T = โหนดภายใน หรือคุณลักษณะของข้อมูล

t_i = กิ่งที่ i หรือค่าของคุณลักษณะในโหนดภายในที่ i

M = จำนวนกิ่งของคุณลักษณะของข้อมูล

s_{ki} = จำนวนข้อมูลในกลุ่มที่ k ของกิ่งที่ i

ตัวอย่าง 2.3 ตัวอย่าง การสร้างต้นไม้ตัดสินใจโดยใช้ Gini Index แบบ class weight โดยสมมติมีข้อมูลของสภาพอากาศ 14 วันย้อนหลัง สำหรับการจัดแข่งขันกีฬา ดังตารางต่อไปนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
1	แดดจัด	ร้อน	สูง	ไม่จริง	ไม่แข่ง
2	แดดจัด	ร้อน	สูง	จริง	ไม่แข่ง
3	มีดครึ้ม	ร้อน	สูง	ไม่จริง	แข่ง
4	ฝนตก	อบอุ่น	สูง	ไม่จริง	แข่ง
5	ฝนตก	เย็น	ปกติ	ไม่จริง	แข่ง
6	ฝนตก	เย็น	ปกติ	จริง	ไม่แข่ง
7	มีดครึ้ม	เย็น	ปกติ	จริง	แข่ง
8	แดดจัด	อบอุ่น	สูง	ไม่จริง	ไม่แข่ง
9	แดดจัด	อบอุ่น	ปกติ	ไม่จริง	แข่ง
10	ฝนตก	อบอุ่น	ปกติ	ไม่จริง	แข่ง
11	แดดจัด	อบอุ่น	ปกติ	จริง	แข่ง
12	มีดครึ้ม	อบอุ่น	สูง	จริง	แข่ง
13	มีดครึ้ม	ร้อน	ปกติ	ไม่จริง	แข่ง
14	ฝนตก	อบอุ่น	สูง	จริง	ไม่แข่ง

ตารางที่ 2.7: แสดงข้อมูลสภาพอากาศ

จากตาราง ประกอบด้วย 5 คุณลักษณะ คือ

- สภาพภูมิอากาศ ประกอบด้วย 3 ค่า คือ แดดจัด, มีดครึ้ม, ฝนตก
- อุณหภูมิ ประกอบด้วย 3 ค่า คือ ร้อน, อบอุ่น, เย็น
- ความชื้นในอากาศ ประกอบด้วย 2 ค่า คือ สูง, ปกติ
- ลมแรง ประกอบด้วย 2 ค่า คือ จริง, ไม่จริง
- แข่งขัน ซึ่งเป็นกลุ่ม ประกอบด้วย 2 ค่า คือ แข่ง, ไม่แข่ง

ต่อไปจะแสดงการคำนวณค่าแต่ละคุณลักษณะเทียบกับกลุ่ม เพื่อหาคุณลักษณะที่มีค่า Gini Index น้อยที่สุด มาเป็นโหนดรากของต้นไม้ตัดสินใจ ดังนี้

จาก N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

r_1 = จำนวนข้อมูลในกลุ่มที่ 1 (แข่ง) = 9

r_2 = จำนวนข้อมูลในกลุ่มที่ 2 (ไม่แข่ง) = 5

$$w_i = \text{ค่า weight ของกลุ่มที่ } i = \frac{\sum_{k=1}^N r_k}{N} \frac{1}{r_i}$$

จะได้ $w_1 = \frac{r_1 + r_2}{N} \frac{1}{r_1} = \frac{9 + 5}{2} \times \frac{1}{9} = \frac{14}{2} \times \frac{1}{9} = 0.78$ และ

$$w_2 = \frac{r_1 + r_2}{N} \frac{1}{r_2} = \frac{9 + 5}{2} \times \frac{1}{5} = \frac{14}{2} \times \frac{1}{5} = 1.40$$

1. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะสภาพภูมิอากาศ

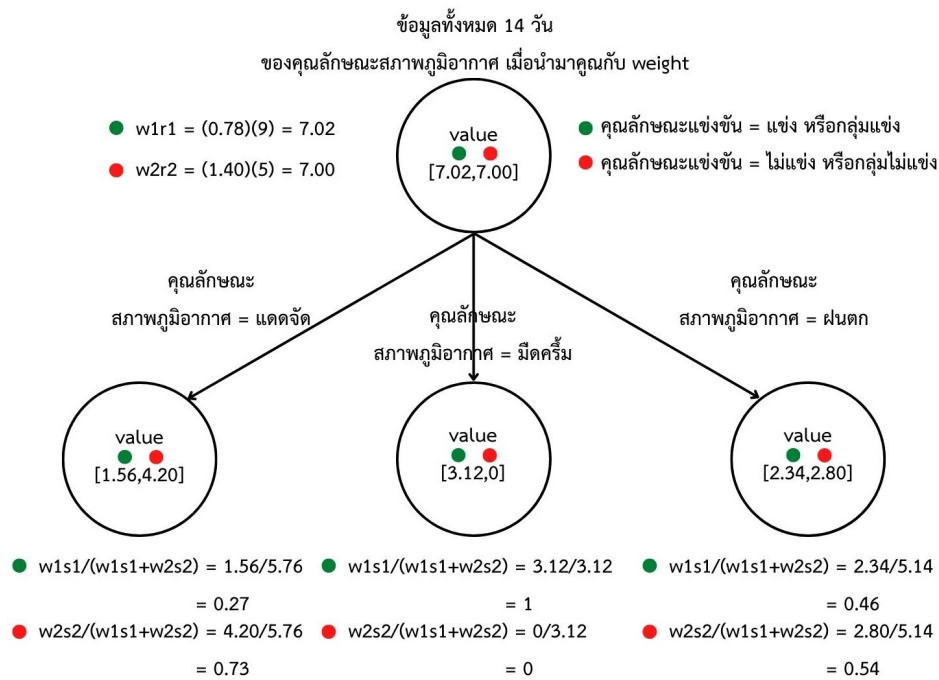
ให้ T = สภาพภูมิอากาศ

M = จำนวนกิ่ง(แดดจัด, มีดครีมี, ฝนตก) = 3

N = จำนวนกลุ่ม(แข่ง, ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.30: แสดงข้อมูลของคุณลักษณะสภาพภูมิอากาศ

กึ่งที่ 1 (แดดจัด)

t_1 คือ สภาพภูมิอากาศ = แดดจัด

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด

$Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด})$

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 2}{(0.78 \times 2) + (1.40 \times 3)} \right)^2 + \left(\frac{1.40 \times 3}{(0.78 \times 2) + (1.40 \times 3)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1.56}{5.76} \right)^2 + \left(\frac{4.20}{5.76} \right)^2 \right] \\
 &= 1 - \left[(0.27)^2 + (0.73)^2 \right] \\
 &= 0.39
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) = 0.39$

กึ่งที่ 2 (มีดครี้ม)

t_2 คือ สภาพภูมิอากาศ = มีดครี้ม

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = มีดครี้ม

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = มีดครี้ม

$Gini(\text{สภาพภูมิอากาศ} = \text{มีดครี้ม})$

$$\begin{aligned} &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\ &= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\ &= 1 - \left[\left(\frac{0.78 \times 4}{(0.78 \times 4) + (1.40 \times 0)} \right)^2 + \left(\frac{1.40 \times 0}{(0.78 \times 4) + (1.40 \times 0)} \right)^2 \right] \\ &= 1 - \left[\left(\frac{3.12}{3.12} \right)^2 + \left(\frac{0}{3.12} \right)^2 \right] \\ &= 1 - \left[(1)^2 + (0)^2 \right] \\ &= 0 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{มีดครี้ม}) = 0$

กึ่งที่ 3 (ฝนตก)

t_3 คือ สภาพภูมิอากาศ = ฝนตก

s_{13} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก

s_{23} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก

$$\begin{aligned}
 Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) &= 1 - \sum_{k=1}^N \left[\frac{w_k s_{ki}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{13}}{w_1 s_{13} + w_2 s_{23}} \right)^2 + \left(\frac{w_2 s_{23}}{w_1 s_{13} + w_2 s_{23}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 3}{(0.78 \times 3) + (1.40 \times 2)} \right)^2 + \left(\frac{1.40 \times 2}{(0.78 \times 3) + (1.40 \times 2)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2.34}{5.14} \right)^2 + \left(\frac{2.80}{5.14} \right)^2 \right] \\
 &= 1 - \left[(0.46)^2 + (0.54)^2 \right] \\
 &= 0.50
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) = 0.50$

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ)

$$\begin{aligned}
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22} + w_1 s_{13} + w_2 s_{23}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22} + w_1 s_{13} + w_2 s_{23}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{มีดคร่ำม}) \\
&\quad + \left(\frac{w_1 s_{13} + w_2 s_{23}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22} + w_1 s_{13} + w_2 s_{23}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก}) \\
&= \left(\frac{(0.78 \times 2) + (1.40 \times 3)}{(0.78 \times 2) + (1.40 \times 3) + (0.78 \times 4) + (1.40 \times 0) + (0.78 \times 3) + (1.40 \times 2)} \right) (0.39) \\
&\quad + \left(\frac{(0.78 \times 4) + (1.40 \times 0)}{(0.78 \times 2) + (1.40 \times 3) + (0.78 \times 4) + (1.40 \times 0) + (0.78 \times 3) + (1.40 \times 2)} \right) (0) \\
&\quad + \left(\frac{(0.78 \times 3) + (1.40 \times 2)}{(0.78 \times 2) + (1.40 \times 3) + (0.78 \times 4) + (1.40 \times 0) + (0.78 \times 3) + (1.40 \times 2)} \right) (0.50) \\
&= \left(\frac{5.76}{14.02} \right) (0.39) + \left(\frac{3.12}{14.02} \right) (0) + \left(\frac{5.14}{14.02} \right) (0.50) \\
&= 0.34
\end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ) = 0.34

2. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะอุณหภูมิ

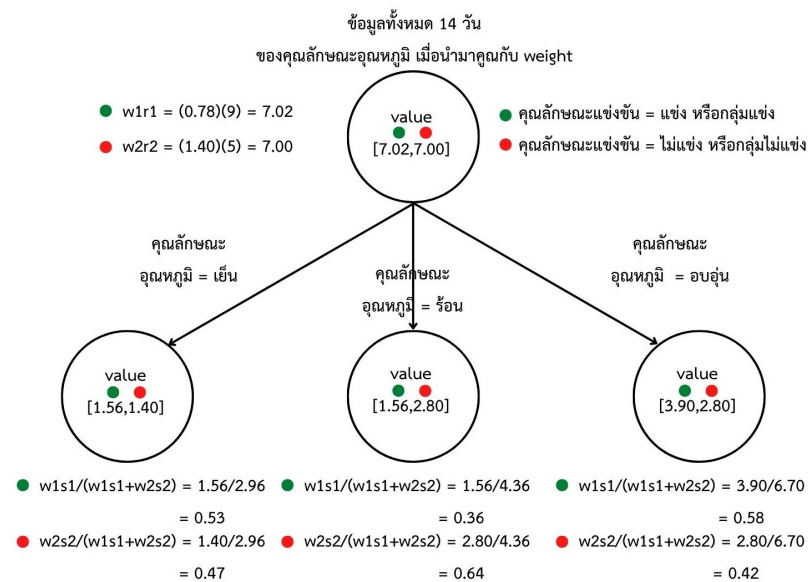
ให้ T = อุณหภูมิ

M = จำนวนกิ่ง(เย็น,ร้อน,อบอุ่น) = 3

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.31: แสดงข้อมูลของคุณลักษณะอุณหภูมิ

กิ่งที่ 1 (เย็น)

t_1 คือ อุณหภูมิ = เย็น

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = เย็น

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = เย็น

$$\begin{aligned}
Gini(\text{อุณหภูมิ} = \text{เย็น}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
&= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
&= 1 - \left[\left(\frac{0.78 \times 2}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 \right] \\
&= 1 - \left[\left(\frac{1.56}{2.96} \right)^2 + \left(\frac{1.40}{2.96} \right)^2 \right] \\
&= 1 - \left[(0.53)^2 + (0.47)^2 \right] \\
&= 0.50
\end{aligned}$$

ดังนั้น $Gini(\text{อุณหภูมิ} = \text{เย็น}) = 0.50$

กึ่งที่ 2 (ร้อน)

t_2 คือ อุณหภูมิ = ร้อน

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดอุณหภูมิ = ร้อน

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดอุณหภูมิ = ร้อน

$$\begin{aligned}
Gini(\text{อุณหภูมิ} = \text{ร้อน}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
&= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\
&= 1 - \left[\left(\frac{0.78 \times 2}{(0.78 \times 2) + (1.40 \times 2)} \right)^2 + \left(\frac{1.40 \times 2}{(0.78 \times 2) + (1.40 \times 2)} \right)^2 \right] \\
&= 1 - \left[\left(\frac{1.56}{4.36} \right)^2 + \left(\frac{2.80}{4.36} \right)^2 \right] \\
&= 1 - \left[(0.36)^2 + (0.64)^2 \right] \\
&= 0.46
\end{aligned}$$

ดังนั้น $Gini(\text{อุณหภูมิ} = \text{ร้อน}) = 0.46$

กึ่งที่ 3 (อบอู่่น)

t_3 คือ อุณหภูมิ = อบอู่่น

s_{13} คือ จํานวนข้อมูลในกลุ่มน้ําของโหนดอุณหภูมิ = อบอู่่น

s_{23} คือ จํานวนข้อมูลในกลุ่มน้ําของโหนดอุณหภูมิ = อบอู่่น

$$\begin{aligned} Gini(\text{อุณหภูมิ} = \text{อบอู่่น}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\ &= 1 - \left[\left(\frac{w_1 s_{13}}{w_1 s_{13} + w_2 s_{23}} \right)^2 + \left(\frac{w_2 s_{23}}{w_1 s_{13} + w_2 s_{23}} \right)^2 \right] \\ &= 1 - \left[\left(\frac{0.78 \times 5}{(0.78 \times 5) + (1.40 \times 2)} \right)^2 + \left(\frac{1.40 \times 2}{(0.78 \times 5) + (1.40 \times 2)} \right)^2 \right] \\ &= 1 - \left[\left(\frac{3.90}{6.70} \right)^2 + \left(\frac{2.80}{6.70} \right)^2 \right] \\ &= 1 - \left[(0.58)^2 + (0.42)^2 \right] \\ &= 0.49 \end{aligned}$$

ดังนั้น $Gini(\text{อุณหภูมิ} = \text{อบอู่่น}) = 0.49$

จะได้ว่า

$Gini_{split}$ (อุณหภูมิ)

$$\begin{aligned}
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22} + w_1 s_{13} + w_2 s_{23}} \right) Gini(\text{อุณหภูมิ} = \text{เย็น}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22} + w_1 s_{13} + w_2 s_{23}} \right) Gini(\text{อุณหภูมิ} = \text{ร้อน}) \\
&\quad + \left(\frac{w_1 s_{13} + w_2 s_{23}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22} + w_1 s_{13} + w_2 s_{23}} \right) Gini(\text{อุณหภูมิ} = \text{อบอุ่น}) \\
&= \left(\frac{(0.78 \times 2) + (1.40 \times 1)}{(0.78 \times 2) + (1.40 \times 1) + (0.78 \times 2) + (1.40 \times 2) + (0.78 \times 5) + (1.40 \times 2)} \right) (0.50) \\
&\quad + \left(\frac{(0.78 \times 2) + (1.40 \times 2)}{(0.78 \times 2) + (1.40 \times 1) + (0.78 \times 2) + (1.40 \times 2) + (0.78 \times 5) + (1.40 \times 2)} \right) (0.46) \\
&\quad + \left(\frac{(0.78 \times 5) + (1.40 \times 2)}{(0.78 \times 2) + (1.40 \times 1) + (0.78 \times 2) + (1.40 \times 2) + (0.78 \times 5) + (1.40 \times 2)} \right) (0.49) \\
&= \left(\frac{2.96}{14.02} \right) (0.50) + \left(\frac{4.36}{14.02} \right) (0.46) + \left(\frac{6.70}{14.02} \right) (0.49) \\
&= 0.48
\end{aligned}$$

ดังนั้น $Gini_{split}$ (อุณหภูมิ) = 0.48

3. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะความชื้นในอากาศ

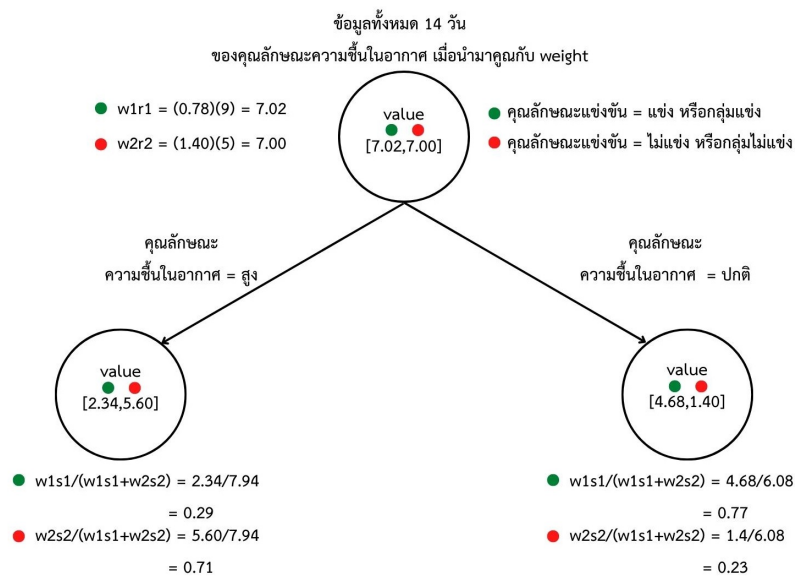
ให้ T = ความชื้นในอากาศ

M = จำนวนกิ่ง(สูง,ปกติ) = 2

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.32: แสดงข้อมูลของคุณลักษณะความชื้นในอากาศ

กิ่งที่ 1 (สูง)

t_1 คือ ความชื้นในอากาศ = สูง

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดความชื้นในอากาศ = สูง

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดความชื้นในอากาศ = สูง

$$\begin{aligned}
Gini(\text{ความขึ้นในอากาศ} = \text{สูง}) &= 1 - \sum_{k=1}^N \left[\frac{w_k s_{ki}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
&= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
&= 1 - \left[\left(\frac{0.78 \times 3}{(0.78 \times 3) + (1.40 \times 4)} \right)^2 + \left(\frac{1.40 \times 4}{(0.78 \times 3) + (1.40 \times 4)} \right)^2 \right] \\
&= 1 - \left[\left(\frac{2.34}{7.94} \right)^2 + \left(\frac{5.60}{7.94} \right)^2 \right] \\
&= 1 - \left[(0.29)^2 + (0.71)^2 \right] \\
&= 0.42
\end{aligned}$$

ดังนั้น $Gini(\text{ความขึ้นในอากาศ} = \text{สูง}) = 0.42$

กึ่งที่ 2 (ปกติ)

t_2 คือ ความขึ้นในอากาศ = ปกติ

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดความขึ้นในอากาศ = ปกติ

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดความขึ้นในอากาศ = ปกติ

$$\begin{aligned}
Gini(\text{ความขึ้นในอากาศ} = \text{ปกติ}) &= 1 - \sum_{k=1}^N \left[\frac{w_k s_{ki}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
&= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\
&= 1 - \left[\left(\frac{0.78 \times 6}{(0.78 \times 6) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 6) + (1.40 \times 1)} \right)^2 \right] \\
&= 1 - \left[\left(\frac{4.68}{6.08} \right)^2 + \left(\frac{1.40}{6.08} \right)^2 \right] \\
&= 1 - \left[(0.77)^2 + (0.23)^2 \right] \\
&= 0.35
\end{aligned}$$

ดังนั้น $Gini(\text{ความขึ้นในอากาศ} = \text{ปกติ}) = 0.35$

จะได้ว่า

$Gini_{split}$ (ความชื้นในอากาศ)

$$\begin{aligned}
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{ความชื้นในอากาศ} = \text{สูง}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{ความชื้นในอากาศ} = \text{ปกติ}) \\
&= \left(\frac{(0.78 \times 3) + (1.40 \times 4)}{(0.78 \times 3) + (1.40 \times 4) + (0.78 \times 6) + (1.40 \times 1)} \right) (0.42) \\
&\quad + \left(\frac{(0.78 \times 6) + (1.40 \times 1)}{(0.78 \times 3) + (1.40 \times 4) + (0.78 \times 6) + (1.40 \times 1)} \right) (0.35) \\
&= \left(\frac{7.94}{14.02} \right) (0.42) + \left(\frac{6.08}{14.02} \right) (0.35) \\
&= 0.39
\end{aligned}$$

ดังนั้น $Gini_{split}$ (ความชื้นในอากาศ) = 0.39

4. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะลมแรง

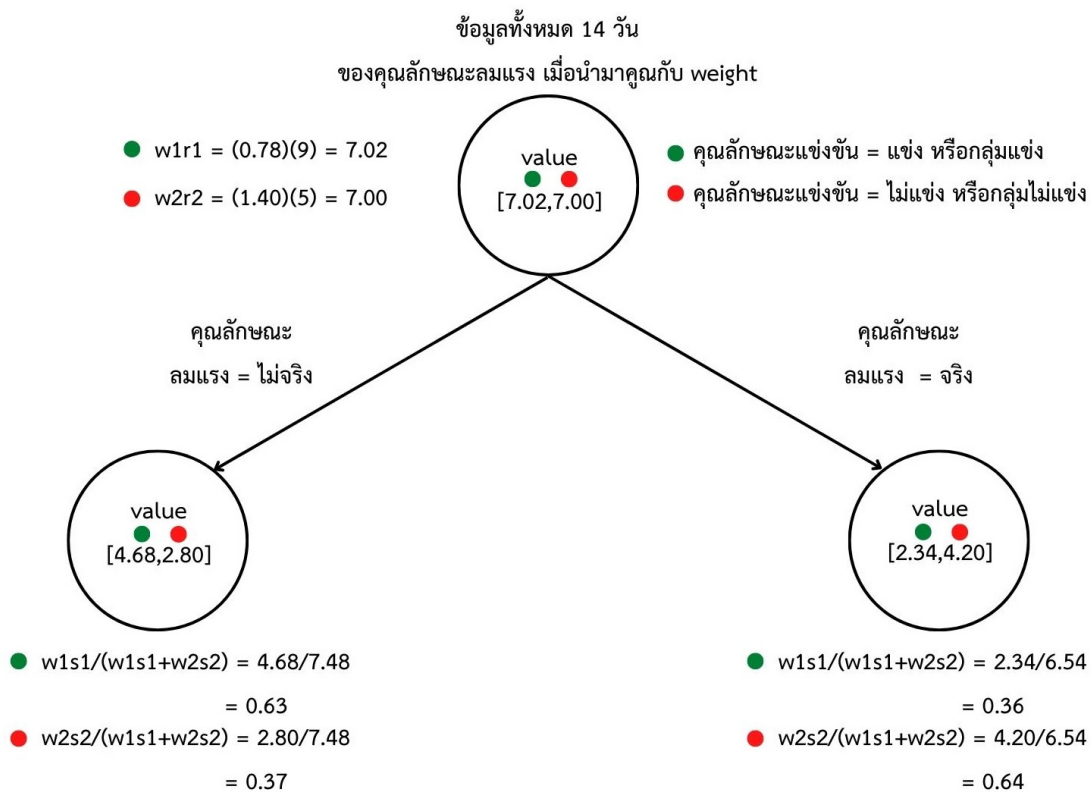
ให้ T = ลมแรง

M = จำนวนกิ่ง(ไม่จริง,จริง) = 3

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.33: แสดงข้อมูลของคุณลักษณะลมแรง

กิ่งที่ 1 (ไม่จริง)

t_1 คือ ลมแรง = ไม่จริง

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดลมแรง = ไม่จริง

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดลมแรง = ไม่จริง

$$\begin{aligned}
Gini(\text{ลมแรง} = \text{ไม่จริง}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
&= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
&= 1 - \left[\left(\frac{0.78 \times 6}{(0.78 \times 6) + (1.40 \times 2)} \right)^2 + \left(\frac{1.40 \times 2}{(0.78 \times 6) + (1.40 \times 2)} \right)^2 \right] \\
&= 1 - \left[\left(\frac{4.68}{7.48} \right)^2 + \left(\frac{2.80}{7.48} \right)^2 \right] \\
&= 1 - \left[(0.63)^2 + (0.37)^2 \right] \\
&= 0.47
\end{aligned}$$

ดังนั้น $Gini(\text{ลมแรง} = \text{ไม่จริง}) = 0.47$

กึ่งที่ 2 (จริง)

t_2 คือ ลมแรง = จริง

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดลมแรง = จริง

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดลมแรง = จริง

$$\begin{aligned}
Gini(\text{ลมแรง} = \text{จริง}) &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
&= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\
&= 1 - \left[\left(\frac{0.78 \times 3}{(0.78 \times 3) + (1.40 \times 3)} \right)^2 + \left(\frac{1.40 \times 3}{(0.78 \times 3) + (1.40 \times 3)} \right)^2 \right] \\
&= 1 - \left[\left(\frac{2.34}{6.54} \right)^2 + \left(\frac{4.20}{6.54} \right)^2 \right] \\
&= 1 - \left[(0.36)^2 + (0.64)^2 \right] \\
&= 0.46
\end{aligned}$$

ดังนั้น $Gini(\text{ลมแรง} = \text{จริง}) = 0.46$

จะได้ว่า

$$\begin{aligned}
& Gini_{split}(\text{ลมแรง}) \\
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{ลมแรง} = \text{ไม่จริง}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{ลมแรง} = \text{จริง}) \\
&= \left(\frac{(0.78 \times 6) + (1.40 \times 2)}{(0.78 \times 6) + (1.40 \times 2) + (0.78 \times 3) + (1.40 \times 3)} \right) (0.47) \\
&\quad + \left(\frac{(0.78 \times 3) + (1.40 \times 3)}{(0.78 \times 6) + (1.40 \times 2) + (0.78 \times 3) + (1.40 \times 3)} \right) (0.46) \\
&= \left(\frac{7.48}{14.02} \right) (0.47) + \left(\frac{6.54}{14.02} \right) (0.46) \\
&= 0.47
\end{aligned}$$

ดังนั้น $Gini_{split}(\text{ลมแรง}) = 0.47$

จากการคำนวณค่า $Gini_{split}$ ของคุณลักษณะต่าง ๆ จะได้ว่า

$$Gini_{split}(\text{สภาพภูมิอากาศ}) = 0.34$$

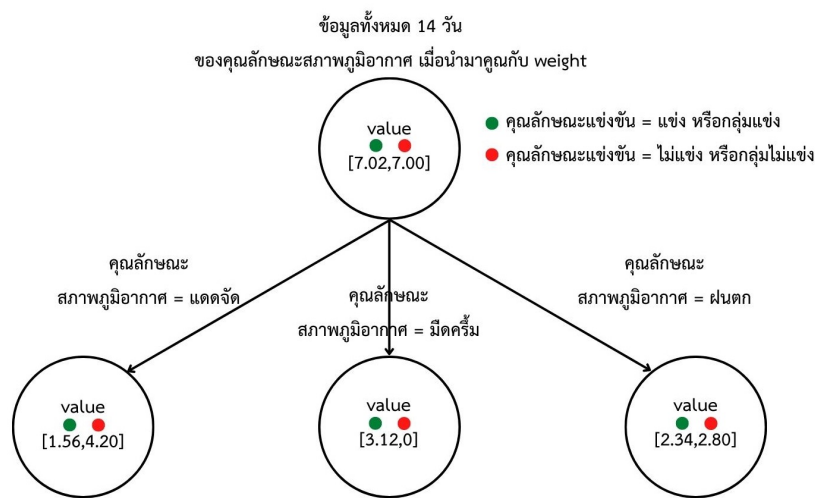
$$Gini_{split}(\text{อุณหภูมิ}) = 0.48$$

$$Gini_{split}(\text{ความชื้นในอากาศ}) = 0.39$$

$$Gini_{split}(\text{ลมแรง}) = 0.47$$

เนื่องจาก คุณลักษณะที่มีค่า Gini Index ที่น้อยที่สุดจะเป็นคุณลักษณะที่มีความหลากหลายของข้อมูลกลุ่มน้อยที่สุด ดังนั้น เราจะเลือกค่า Gini Index ที่น้อยที่สุดเป็นโหนดราก

จะเห็นได้ว่า ค่า Gini Index ของคุณลักษณะสภาพภูมิอากาศ มีค่าน้อยที่สุด คือ 0.34 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศเป็นโหนดราก



รูปที่ 2.34: แสดงข้อมูลของคุณลักษณะสภาพภูมิอากาศ

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = มีดครึ้ม มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = ชนะ ดังนั้น โหนดคุณลักษณะสภาพภูมิอากาศ = มีดครึ้ม จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ แต่โหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก ต้องทำการแตกกิ่งจนกว่าข้อมูลในแต่ละโหนดจะมีข้อมูลเป็นกลุ่มเดียวกัน

ต่อไปจะแสดงการคำนวณการแตกกิ่งของโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด โดยสร้างตารางใหม่ที่สภาพภูมิอากาศ = แดดจัด ดังนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
1	แดดจัด	ร้อน	สูง	ไม่จริง	ไม่แข่ง
2	แดดจัด	ร้อน	สูง	จริง	ไม่แข่ง
8	แดดจัด	อบอุ่น	สูง	ไม่จริง	ไม่แข่ง
9	แดดจัด	อบอุ่น	ปกติ	ไม่จริง	แข่ง
11	แดดจัด	อบอุ่น	ปกติ	จริง	แข่ง

ตารางที่ 2.8: แสดงข้อมูลสภาพอากาศที่สภาพภูมิอากาศ = แดดจัด

1. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

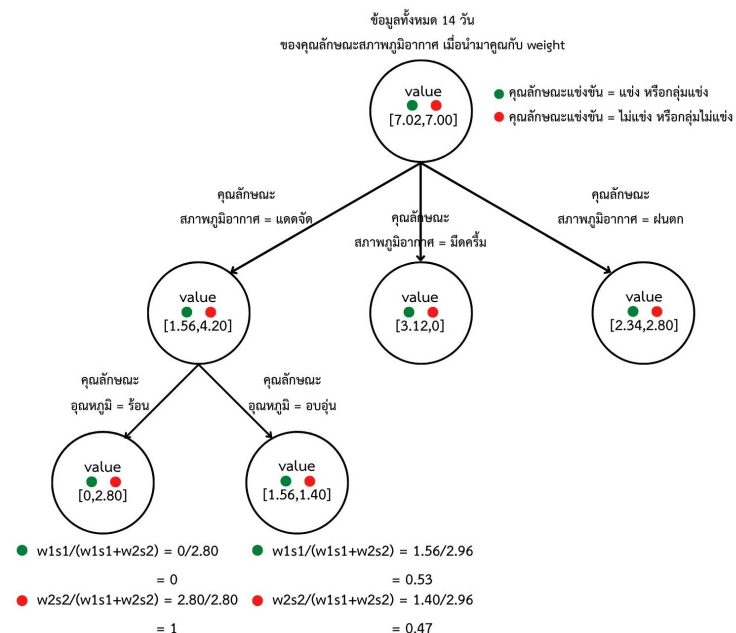
ให้ T = (สภาพภูมิอากาศ = แดดจัด) และอุณหภูมิ

M = จำนวนกิ่ง(ร้อน,อบอุ่น) = 2

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.35: แสดงข้อมูลคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

กึ่งที่ 1 (ร้อน)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = ร้อน

$Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{ร้อน})$

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{w_k s_{ki}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 0}{(0.78 \times 0) + (1.40 \times 2)} \right)^2 + \left(\frac{1.40 \times 2}{(0.78 \times 0) + (1.40 \times 2)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{2.80} \right)^2 + \left(\frac{2.80}{2.80} \right)^2 \right] \\
 &= 1 - [(0)^2 + (1)^2] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{ร้อน}) = 0$

กึ่งที่ 2 (อบอุ่น)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และอุณหภูมิ = อบอุ่น

$Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{อบอุ่น})$

$$\begin{aligned} &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\ &= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\ &= 1 - \left[\left(\frac{0.78 \times 2}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 \right] \\ &= 1 - \left[\left(\frac{1.56}{2.96} \right)^2 + \left(\frac{1.40}{2.96} \right)^2 \right] \\ &= 1 - \left[(0.53)^2 + (0.47)^2 \right] \\ &= 0.50 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{อบอุ่น}) = 0.50$

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ)

$$\begin{aligned}
 &= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
 &= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
 &= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{ร้อน}) \\
 &\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ} = \text{อบอุ่น}) \\
 &= \left(\frac{(0.78 \times 0) + (1.40 \times 2)}{(0.78 \times 0) + (1.40 \times 2) + (0.78 \times 2) + (1.40 \times 1)} \right) (0) \\
 &\quad + \left(\frac{(0.78 \times 2) + (1.40 \times 1)}{(0.78 \times 0) + (1.40 \times 2) + (0.78 \times 2) + (1.40 \times 1)} \right) (0.50) \\
 &= \left(\frac{2.80}{5.76} \right) (0) + \left(\frac{2.96}{5.76} \right) (0.50) \\
 &= 0.26
 \end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ) = 0.26

2. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนด ที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

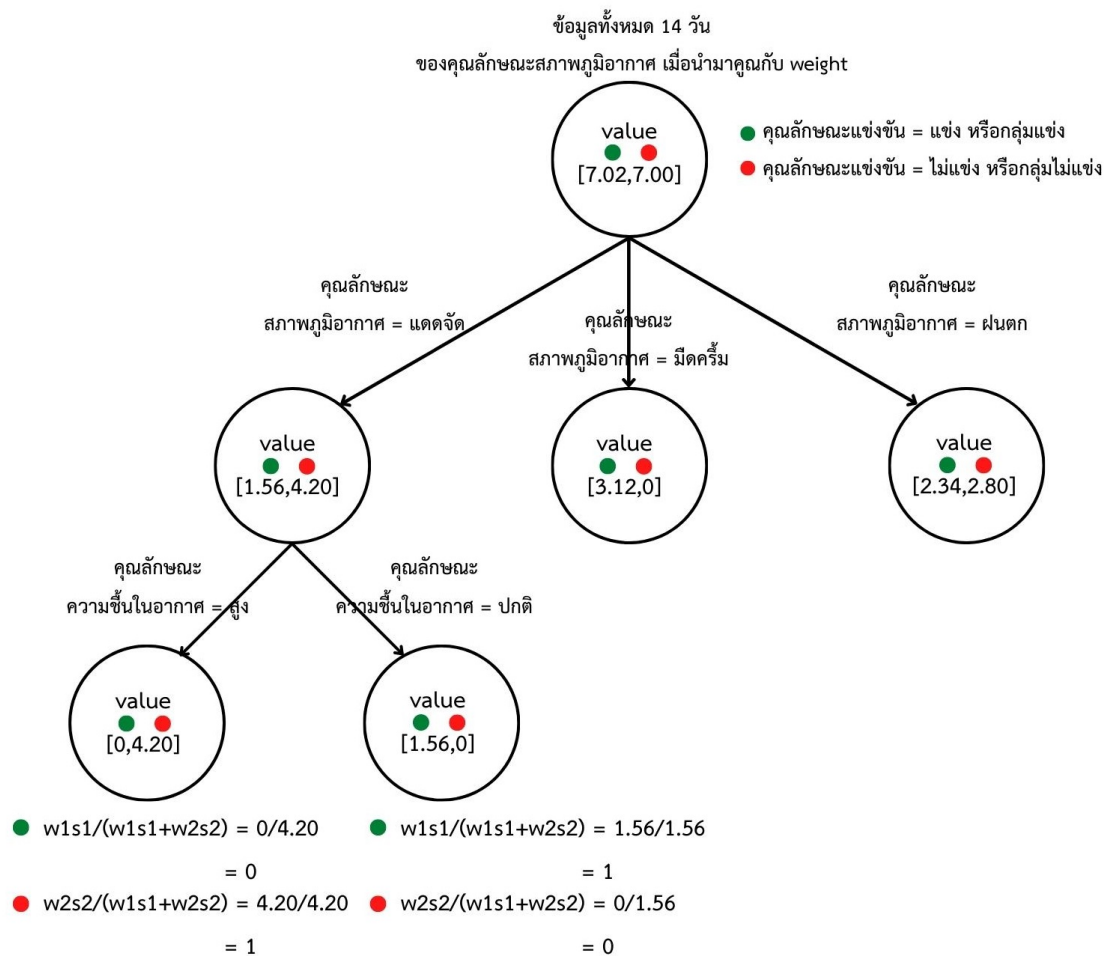
ให้ T = (สภาพภูมิอากาศ = แดดจัด) และความชื้นในอากาศ

M = จำนวนกิ่ง(สูง,ปกติ) = 2

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.36: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

กึ่งที่ 1 (สูง)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง

$Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และ ความชื้นในอากาศ} = \text{สูง})$

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 0}{(0.78 \times 0) + (1.40 \times 3)} \right)^2 + \left(\frac{1.40 \times 3}{(0.78 \times 0) + (1.40 \times 3)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{4.20} \right)^2 + \left(\frac{4.20}{4.20} \right)^2 \right] \\
 &= 1 - \left[(0)^2 + (1)^2 \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และ ความชื้นในอากาศ} = \text{สูง}) = 0$

กึ่งที่ 2 (ปกติ)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ =

ปกติ

$Gini$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 2}{(0.78 \times 2) + (1.40 \times 0)} \right)^2 + \left(\frac{1.40 \times 0}{(0.78 \times 2) + (1.40 \times 0)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1.56}{1.56} \right)^2 + \left(\frac{0}{1.56} \right)^2 \right] \\
 &= 1 - \left[(1)^2 + (0)^2 \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini$ (ภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ) = 0

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ)

$$\begin{aligned}
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ} = \text{สูง}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{ภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ} = \text{ปกติ}) \\
&= \left(\frac{(0.78 \times 0) + (1.40 \times 3)}{(0.78 \times 0) + (1.40 \times 3) + (0.78 \times 2) + (1.40 \times 0)} \right) (0) \\
&\quad + \left(\frac{(0.78 \times 2) + (1.40 \times 0)}{(0.78 \times 0) + (1.40 \times 3) + (0.78 \times 2) + (1.40 \times 0)} \right) (0) \\
&= \left(\frac{4.20}{5.76} \right) (0) + \left(\frac{1.56}{5.76} \right) (0) \\
&= 0
\end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ) = 0

3. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = แดดจัด

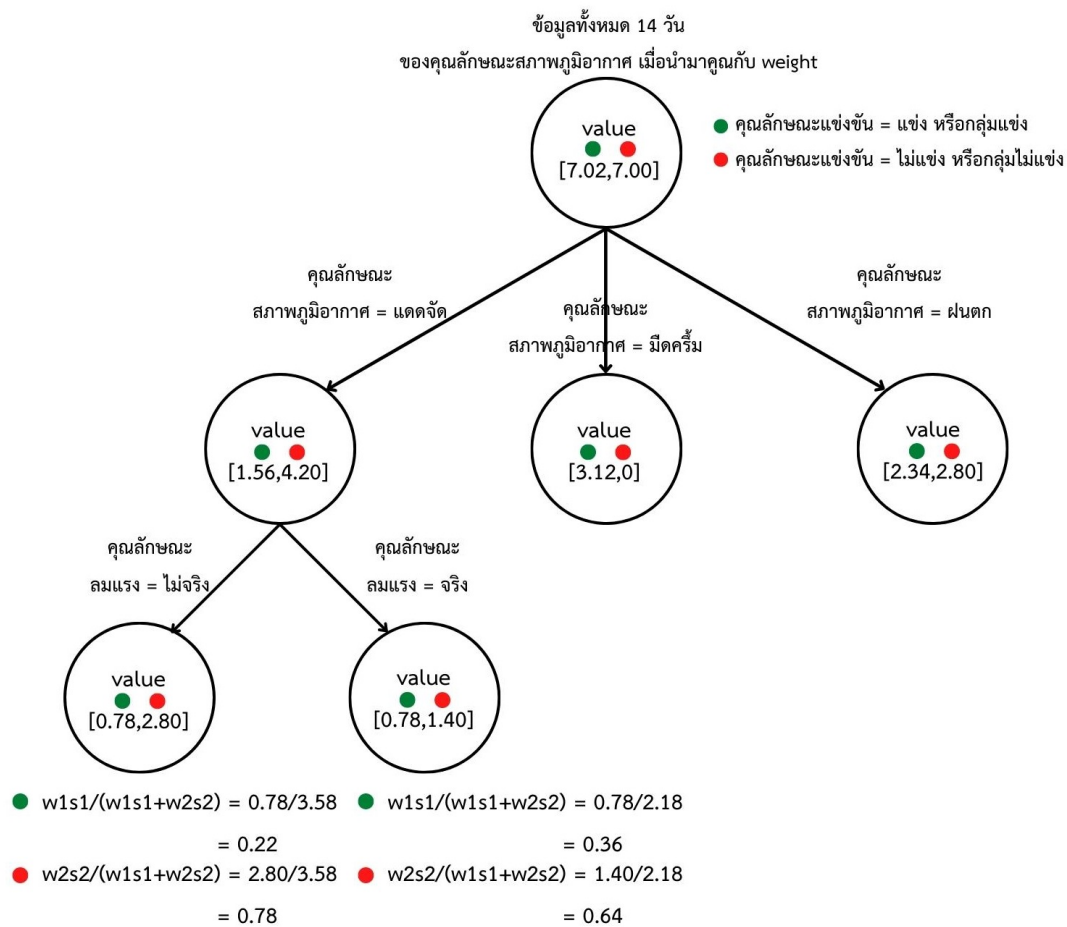
ให้ $T = (\text{สภาพภูมิอากาศ} = \text{แดดจัด})$ และลมแรง

$M = \text{จำนวนกิ่ง(ไม่จริง,จริง)} = 2$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$

$w_1 = \text{ค่า weight ของกลุ่มที่ 1 (แข่ง)} = 0.78$

$w_2 = \text{ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง)} = 1.40$



รูปที่ 2.37: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

กึ่งที่ 1 (ไม่จริง)

t_1 คือ สภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง

$Gini$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 1}{(0.78 \times 1) + (1.40 \times 2)} \right)^2 + \left(\frac{1.40 \times 2}{(0.78 \times 1) + (1.40 \times 2)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78}{3.58} \right)^2 + \left(\frac{2.80}{3.85} \right)^2 \right] \\
 &= 1 - \left[(0.22)^2 + (0.78)^2 \right] \\
 &= 0.34
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = ไม่จริง) = 0.34

กึ่งที่ 2 (จริง)

t_2 คือ สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = แดดจัด และลมแรง = จริง

$Gini$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง)

$$\begin{aligned} &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\ &= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\ &= 1 - \left[\left(\frac{0.78 \times 1}{(0.78 \times 1) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 1) + (1.40 \times 1)} \right)^2 \right] \\ &= 1 - \left[\left(\frac{0.78}{2.18} \right)^2 + \left(\frac{1.40}{2.18} \right)^2 \right] \\ &= 1 - \left[(0.36)^2 + (0.64)^2 \right] \\ &= 0.46 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = แดดจัด และลมแรง = จริง) = 0.46

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และลมแรง)

$$\begin{aligned}
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{ไม่จริง}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง} = \text{จริง}) \\
&= \left(\frac{(0.78 \times 1) + (1.40 \times 2)}{(0.78 \times 1) + (1.40 \times 2) + (0.78 \times 1) + (1.40 \times 1)} \right) (0.34) \\
&\quad + \left(\frac{(0.78 \times 1) + (1.40 \times 1)}{(0.78 \times 1) + (1.40 \times 2) + (0.78 \times 1) + (1.40 \times 1)} \right) (0.46) \\
&= \left(\frac{3.58}{5.76} \right) (0.34) + \left(\frac{2.18}{5.76} \right) (0.46) \\
&= 0.39
\end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และลมแรง) = 0.39

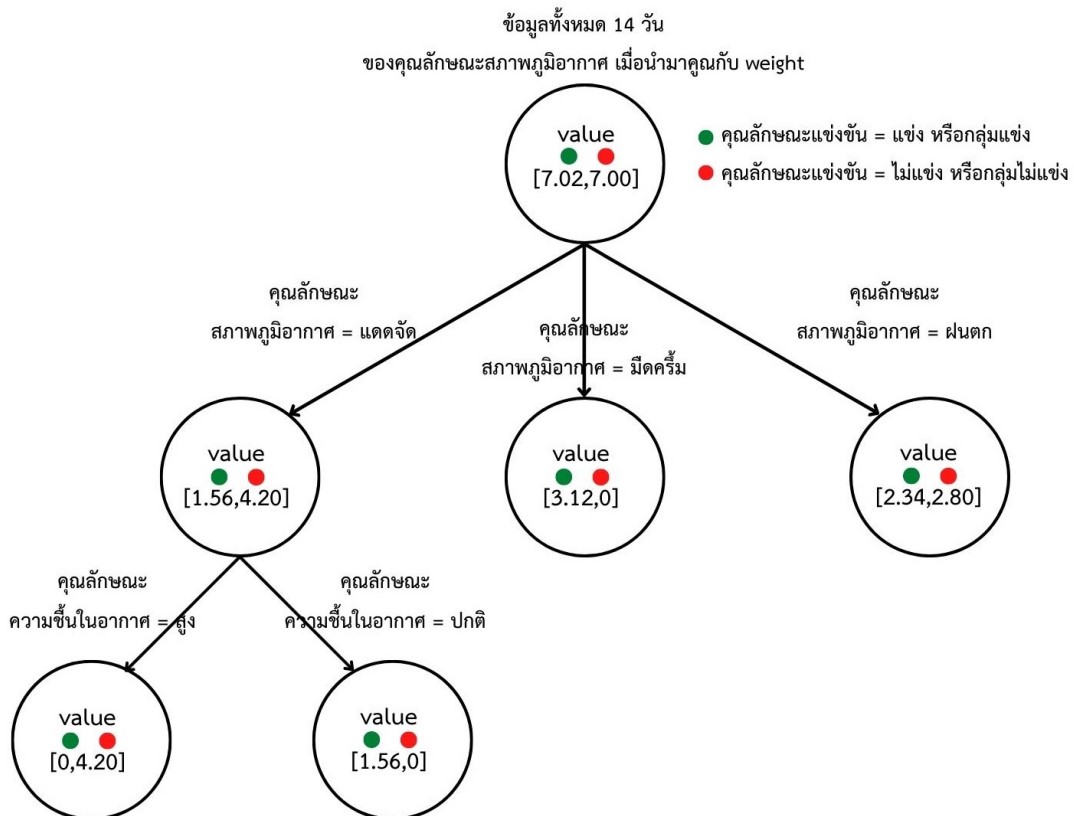
จากการคำนวณค่า $Gini_{split}$ ของคุณลักษณะต่าง ๆ จะได้ว่า

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และอุณหภูมิ}) = 0.26$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และความชื้นในอากาศ}) = 0$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{แดดจัด และลมแรง}) = 0.39$$

จะเห็นว่า ค่า Gini Index ของคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ มีค่าน้อยที่สุด คือ 0 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ



รูปที่ 2.38: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = สูง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = ไม่แข่ง และโหนดคุณลักษณะสภาพภูมิอากาศ = แดดจัด และความชื้นในอากาศ = ปกติ มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งขัน = แข่ง ดังนั้น จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ

ต่อไปจะแสดงการคำนวณการแตกกิ่งของโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก โดยสร้างตารางใหม่ที่สภาพภูมิอากาศ = ฝนตก ดังนี้

วันที่	สภาพภูมิอากาศ	อุณหภูมิ	ความชื้นในอากาศ	ลมแรง	แข่งขัน
4	ฝนตก	อบอุ่น	สูง	ไม่จริง	แข่ง
5	ฝนตก	เย็น	ปกติ	ไม่จริง	แข่ง
6	ฝนตก	เย็น	ปกติ	จริง	ไม่แข่ง
10	ฝนตก	อบอุ่น	ปกติ	ไม่จริง	แข่ง
14	ฝนตก	อบอุ่น	สูง	จริง	ไม่แข่ง

ตารางที่ 2.9: แสดงข้อมูลสภาพอากาศที่สภาพภูมิอากาศ = ฝนตก

1. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

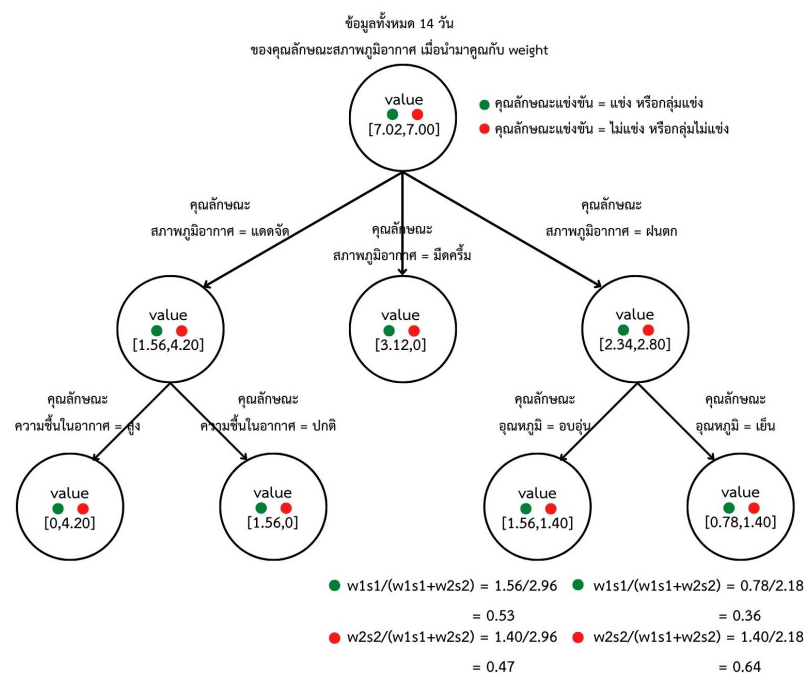
ให้ T = (สภาพภูมิอากาศ = ฝนตก) และอุณหภูมิ

M = จำนวนกิ่ง(อบอุ่น,เย็น) = 2

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.39: แสดงข้อมูลคุณลักษณะอุณหภูมิ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

กึ่งที่ 1 (อบอุ่น)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น

$Gini$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 2}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1.56}{2.96} \right)^2 + \left(\frac{1.40}{2.96} \right)^2 \right] \\
 &= 1 - \left[(0.67)^2 + (0.33)^2 \right] \\
 &= 0.50
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = อบอุ่น) = 0.50

กึ่งที่ 2 (เย็น)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น

$Gini$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 1}{(0.78 \times 1) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 1) + (1.40 \times 1)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78}{2.18} \right)^2 + \left(\frac{1.40}{2.18} \right)^2 \right] \\
 &= 1 - \left[(0.36)^2 + (0.64)^2 \right] \\
 &= 0.46
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ = เย็น) = 0.46

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และอุณหภูมิ)

$$\begin{aligned}
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{อบอุ่น}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ} = \text{เย็น}) \\
&= \left(\frac{(0.78 \times 2) + (1.40 \times 1)}{(0.78 \times 2) + (1.40 \times 1) + (0.78 \times 1) + (1.40 \times 1)} \right) (0.50) \\
&\quad + \left(\frac{(0.78 \times 1) + (1.40 \times 1)}{(0.78 \times 2) + (1.40 \times 1) + (0.78 \times 1) + (1.40 \times 1)} \right) (0.46) \\
&= \left(\frac{2.96}{5.14} \right) (0.50) + \left(\frac{2.18}{5.14} \right) (0.46) \\
&= 0.48
\end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = แดดจัด และอุณหภูมิ) = 0.48

2. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนด ที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

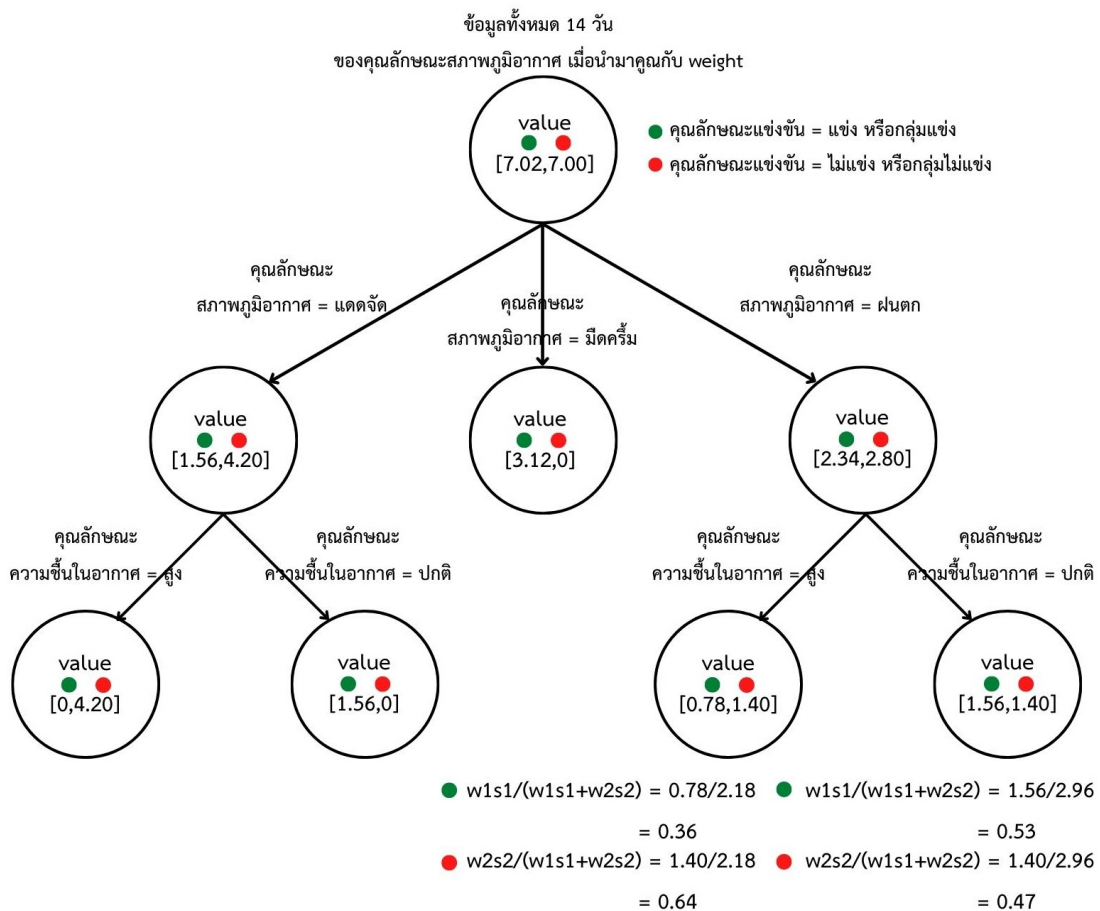
ให้ T = (สภาพภูมิอากาศ = ฝนตก) และความชื้นในอากาศ

M = จำนวนกิ่ง(สูง,ปกติ) = 2

N = จำนวนกลุ่ม(แข่ง,ไม่แข่ง) = 2

w_1 = ค่า weight ของกลุ่มที่ 1 (แข่ง) = 0.78

w_2 = ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง) = 1.40



รูปที่ 2.40: แสดงข้อมูลคุณลักษณะความชื้นในอากาศ เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

กึ่งที่ 1 (สูง)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

s_{12} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง

$Gini$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{12}} \right)^2 + \left(\frac{w_2 s_{12}}{w_1 s_{11} + w_2 s_{12}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 1}{(0.78 \times 1) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 1) + (1.40 \times 1)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78}{2.18} \right)^2 + \left(\frac{1.40}{2.18} \right)^2 \right] \\
 &= 1 - \left[(0.36)^2 + (0.64)^2 \right] \\
 &= 0.46
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = สูง) = 0.46

กึ่งที่ 2 (ปกติ)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ

$Gini$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 2}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 + \left(\frac{1.40 \times 1}{(0.78 \times 2) + (1.40 \times 1)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{1.56}{2.96} \right)^2 + \left(\frac{1.40}{2.96} \right)^2 \right] \\
 &= 1 - \left[(0.53)^2 + (0.47)^2 \right] \\
 &= 0.50
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ = ปกติ) = 0.50

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ)

$$\begin{aligned}
 &= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
 &= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
 &= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก} \text{ และความชื้นในอากาศ} = \text{สูง}) \\
 &\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก} \text{ และความชื้นในอากาศ} = \text{ปกติ}) \\
 &= \left(\frac{(0.78 \times 1) + (1.40 \times 1)}{(0.78 \times 1) + (1.40 \times 1) + (0.78 \times 2) + (1.40 \times 1)} \right) (0.46) \\
 &\quad + \left(\frac{(0.78 \times 2) + (1.40 \times 1)}{(0.78 \times 1) + (1.40 \times 1) + (0.78 \times 2) + (1.40 \times 1)} \right) (0.50) \\
 &= \left(\frac{2.18}{5.14} \right) (0.46) + \left(\frac{2.96}{5.14} \right) (0.50) \\
 &= 0.48
 \end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และความชื้นในอากาศ) = 0.48

3. คำนวณค่า Gini Index แบบ class weight ของคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดที่มีคุณลักษณะสภาพภูมิอากาศ = ฝนตก

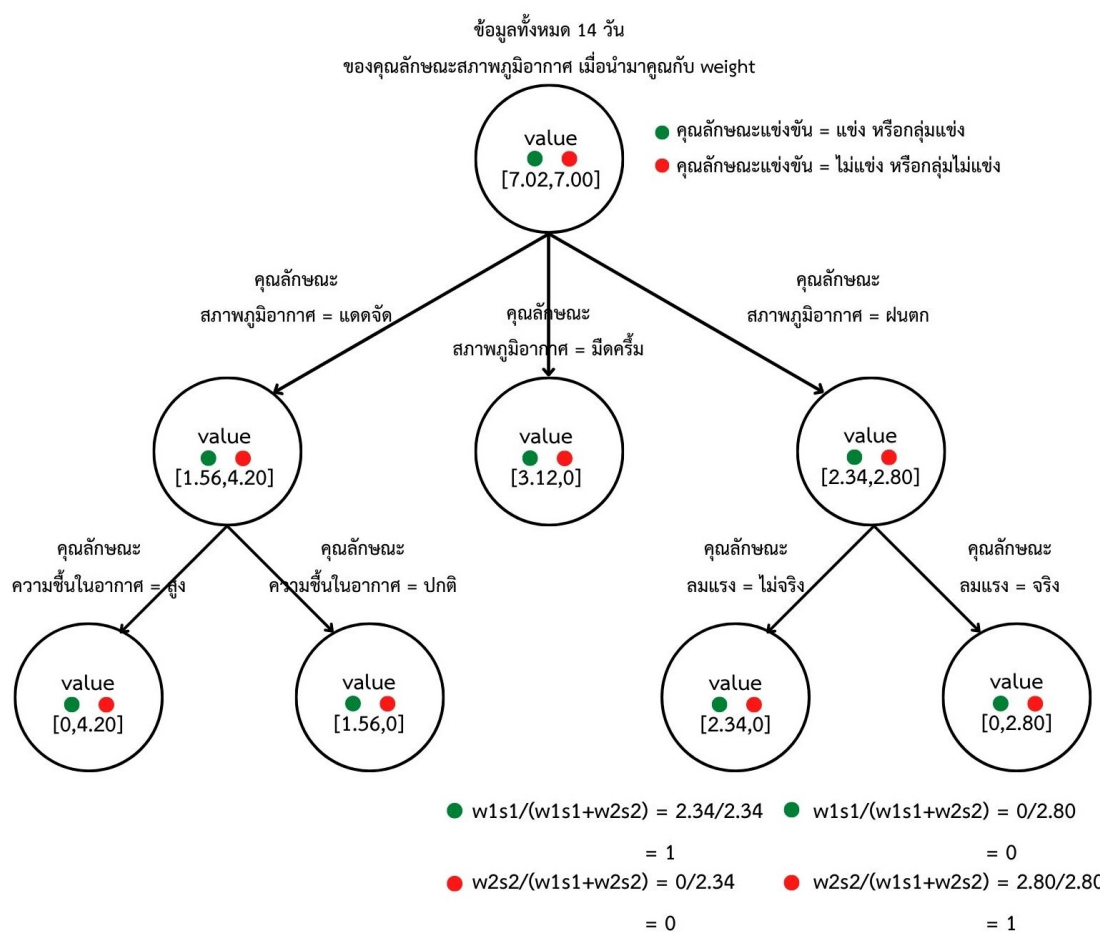
ให้ $T = (\text{สภาพภูมิอากาศ} = \text{ฝนตก})$ และลมแรง

$M = \text{จำนวนกิ่ง(ไม่จริง,จริง)} = 2$

$N = \text{จำนวนกลุ่ม(แข่ง,ไม่แข่ง)} = 2$

$w_1 = \text{ค่า weight ของกลุ่มที่ 1 (แข่ง)} = 0.78$

$w_2 = \text{ค่า weight ของกลุ่มที่ 2 (ไม่แข่ง)} = 1.40$



รูปที่ 2.41: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

กึ่งที่ 1 (ไม่จริง)

t_1 คือ สภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

s_{11} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

s_{21} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง

$Gini$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{11}}{w_1 s_{11} + w_2 s_{21}} \right)^2 + \left(\frac{w_2 s_{21}}{w_1 s_{11} + w_2 s_{21}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 3}{(0.78 \times 3) + (1.40 \times 0)} \right)^2 + \left(\frac{1.40 \times 0}{(0.78 \times 3) + (1.40 \times 0)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{2.34}{2.34} \right)^2 + \left(\frac{0}{2.34} \right)^2 \right] \\
 &= 1 - \left[(1)^2 + (0)^2 \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง) = 0

กึ่งที่ 2 (จริง)

t_2 คือ สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

s_{12} คือ จำนวนข้อมูลในกลุ่มแข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

s_{22} คือ จำนวนข้อมูลในกลุ่มไม่แข่งของโหนดสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง

$Gini$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง)

$$\begin{aligned}
 &= 1 - \sum_{k=1}^N \left[\frac{\frac{w_k s_{ki}}{N}}{\sum_{l=1}^N w_l s_{li}} \right]^2 \\
 &= 1 - \left[\left(\frac{w_1 s_{12}}{w_1 s_{12} + w_2 s_{22}} \right)^2 + \left(\frac{w_2 s_{22}}{w_1 s_{12} + w_2 s_{22}} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0.78 \times 0}{(0.78 \times 0) + (1.40 \times 2)} \right)^2 + \left(\frac{1.40 \times 2}{(0.78 \times 0) + (1.40 \times 2)} \right)^2 \right] \\
 &= 1 - \left[\left(\frac{0}{2.80} \right)^2 + \left(\frac{2.80}{2.80} \right)^2 \right] \\
 &= 1 - \left[(0)^2 + (1)^2 \right] \\
 &= 0
 \end{aligned}$$

ดังนั้น $Gini$ (สภาพภูมิอากาศ = ฝนตก และลมแรง = จริง) = 0

จะได้ว่า

$Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และลมแรง)

$$\begin{aligned}
&= \sum_{i=1}^M \frac{\sum_{k=1}^N w_k s_{ki}}{\sum_{j=1}^M \sum_{k=1}^N w_k s_{kj}} Gini(t_i) \\
&= \sum_{i=1}^M \frac{w_1 s_{1i} + w_2 s_{2i}}{\sum_{j=1}^M w_1 s_{1j} + w_2 s_{2j}} Gini(t_i) \\
&= \left(\frac{w_1 s_{11} + w_2 s_{21}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{ไม่จริง}) \\
&\quad + \left(\frac{w_1 s_{12} + w_2 s_{22}}{w_1 s_{11} + w_2 s_{21} + w_1 s_{12} + w_2 s_{22}} \right) Gini(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง} = \text{จริง}) \\
&= \left(\frac{(0.78 \times 3) + (1.40 \times 0)}{(0.78 \times 3) + (1.40 \times 0) + (0.78 \times 0) + (1.40 \times 2)} \right) (0) \\
&\quad + \left(\frac{(0.78 \times 0) + (1.40 \times 2)}{(0.78 \times 3) + (1.40 \times 0) + (0.78 \times 0) + (1.40 \times 2)} \right) (0) \\
&= \left(\frac{2.34}{5.14} \right) (0) + \left(\frac{2.80}{5.14} \right) (0) \\
&= 0
\end{aligned}$$

ดังนั้น $Gini_{split}$ (สภาพภูมิอากาศ = ฝนตก และลมแรง) = 0

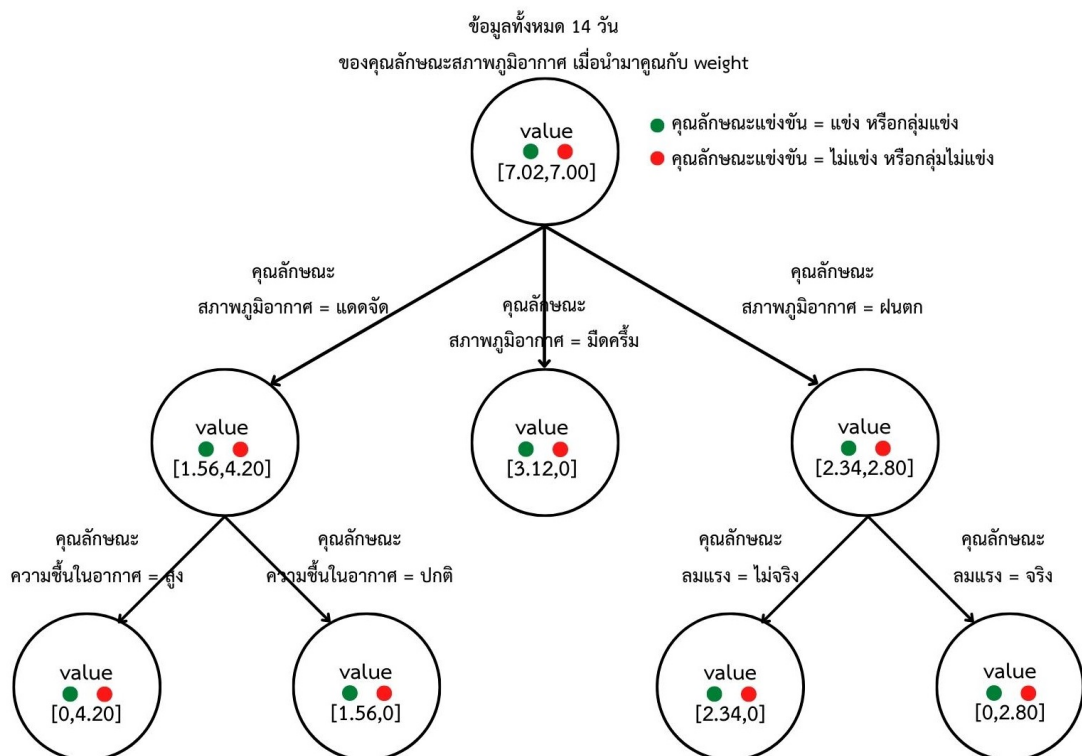
จากการคำนวณค่า $Gini_{split}$ ของคุณลักษณะต่าง ๆ เมื่อแตกกิ่งออกจากโหนดสภาพภูมิอากาศ = ฝนตก จะได้ว่า

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{ฝนตก และอุณหภูมิ}) = 0.48$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{ฝนตก และความชื้นในอากาศ}) = 0.48$$

$$Gini_{split}(\text{สภาพภูมิอากาศ} = \text{ฝนตก และลมแรง}) = 0$$

จะเห็นได้ว่า ค่า Gini Index ของคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง มีค่าน้อยที่สุด คือ 0 ดังนั้นจึงเลือกคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง

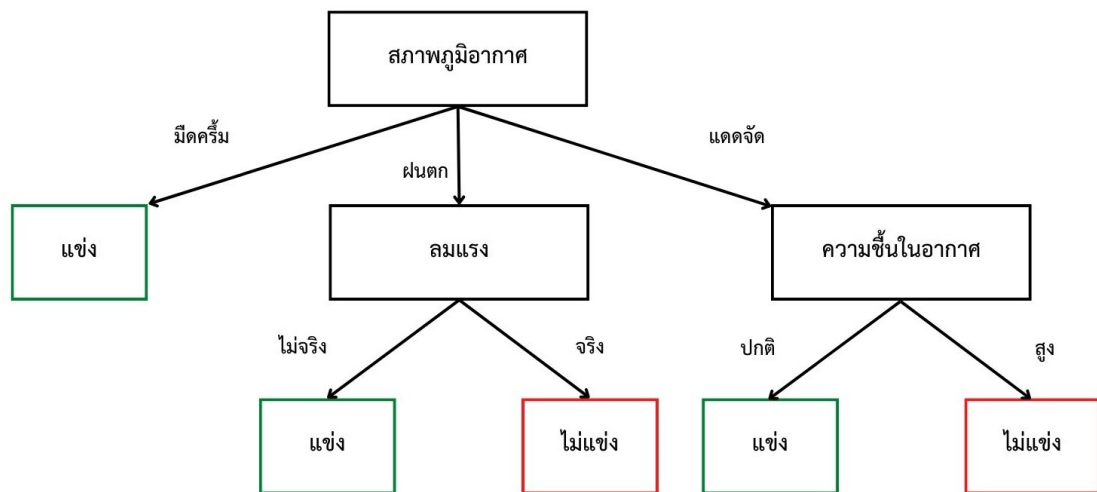


รูปที่ 2.42: แสดงข้อมูลคุณลักษณะลมแรง เมื่อแตกกิ่งออกจากโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก

จะเห็นว่า ในโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง = ไม่จริง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งชั้น = แข่ง และโหนดคุณลักษณะสภาพภูมิอากาศ = ฝนตก และลมแรง = จริง มีข้อมูลเป็นกลุ่มเดียวกัน คือ กลุ่มแข่งชั้น = ไม่แข่ง ดังนั้น จะหยุดแตกกิ่ง หรือสิ้นสุดการคำนวณ

เนื่องจากในแต่ละโหนดใบมีข้อมูลเป็นกลุ่มเดียวกัน ฉะนั้น จะเสร็จสิ้นในขั้นตอนการคำนวณ

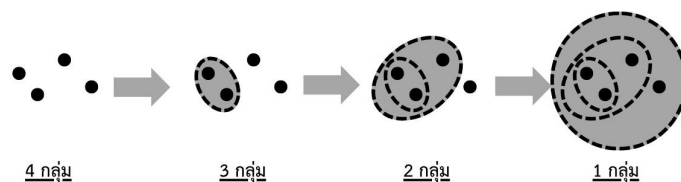
ดังนั้น สามารถสร้างโมเดลต้นไม้ตัดสินใจ ได้ดังนี้



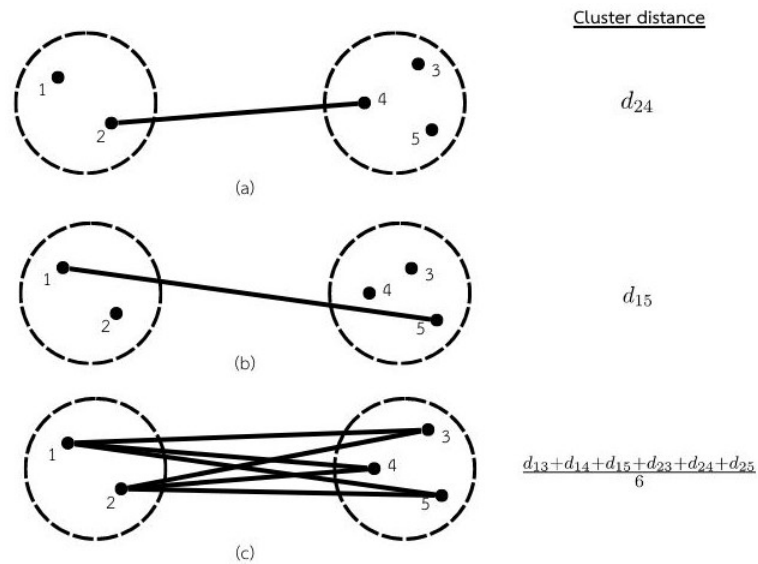
รูปที่ 2.43: แสดงโมเดลต้นไม้ตัดสินใจของข้อมูลสภาพอากาศ 14 วันย้อนหลัง สำหรับการแข่งขันกีฬา

2.2 Agglomerative Hierarchical Cluster Analysis

วิธีนี้ เริ่มต้นจะสมมติว่ามี n กลุ่มย่อย สิ่งของ หรือ item จากนั้นจะรวมกลุ่มที่มีระยะทางสั้นที่สุด หรือคล้ายกันมากที่สุดจะรวมเข้าด้วยกัน จึงเหลือ $n-1$ กลุ่มย่อย จากนั้นหาระยะทางหรือความคล้ายจาก $n-1$ กลุ่มย่อยใหม่ แล้วดูว่ากลุ่มย่อยใดมีระยะทางสั้นที่สุด หรือคล้ายกันมากที่สุดก็รวมกลุ่มย่อยนั้นเข้าด้วยกัน ทำเช่นนี้ต่อ ๆ ไป ในท้ายที่สุดแล้วจะมีเพียง 1 กลุ่มซึ่งประกอบด้วยสิ่งของ n สิ่ง



หลักการในการรวมกลุ่ม ที่เป็นที่รู้จักกันโดยทั่วไปมี 3 วิธีคือ single linkage (หรือเรียกว่า nearest neighbor) complete linkage (หรือ furthest neighbor) และ average linkage (หรือ average distance) แนวคิดทั้ง 3 วิธีอธิบายด้วยรูปดังนี้



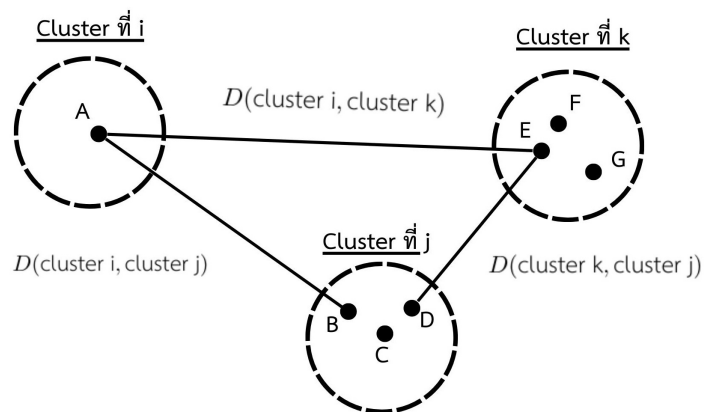
รูปที่ 2.44: แนวคิดทั้ง 3 วิธี

จากรูปจะเห็นว่า การจัดกลุ่มโดยวิธี single linkage ใช้ระยะทางที่สั้นที่สุด ระหว่างสิ่งของหรือ item การจัดกลุ่มโดยวิธี complete linkage ใช้ระยะทางที่ยาวที่สุด และกลุ่มที่ได้โดยวิธี average distance ใช้ระยะทางเฉลี่ย

สำหรับในการค้นคว้าอิสระนี้ เราเลือกใช้วิธี single linkage ในการรวมกลุ่ม โดยมีสูตรการหา ดังนี้

$$D(C, C') = \min_{x \in C, y \in C'} d(x, y)$$

เมื่อ $C \cap C' \neq \emptyset$ และ $d(x, y)$ คือ ระยะทางระหว่าง x และ y

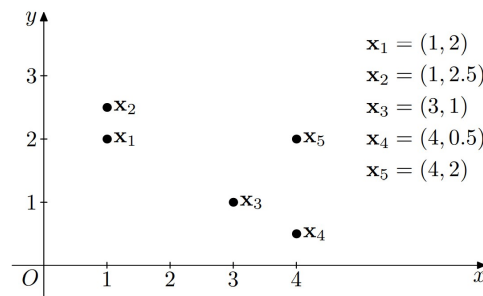


รูปที่ 2.45: Single linkage

จากรูปมี 3 cluster จะเห็นว่า ระยะทางระหว่าง cluster k และ j สั้นที่สุด ดังนั้น เราจะรวม cluster k และ j เข้าด้วยกันก่อน จากนั้นจึงรวมกับ cluster i

ตัวอย่าง 2.4 การทำ Clustering Model ด้วย Agglomerative Hierarchical Algorithm โดยใช้วิธี Single Linkage

สมมติมีชุดข้อมูล 2 มิติของข้อมูล 5 จุด



รูปที่ 2.46: ตัวอย่าง ชุดข้อมูล 2 มิติของข้อมูล 5 จุด

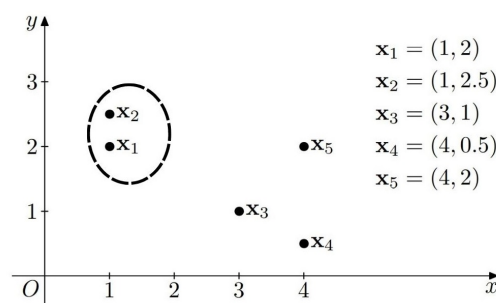
เริ่มต้น เราจะสร้างตารางเมตริกซ์ระยะทาง โดยที่ $d(X_1, X_2) = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2}$ เมื่อ $X_1 = (a_1, b_1)$ และ $X_2 = (a_2, b_2)$

	x_1	x_2	x_3	x_4	x_5
x_1	0	0.5	2.24	3.35	3
x_2	0.5	0	2.5	3.61	3.04
x_3	2.24	2.5	0	1.12	1.41
x_4	3.35	3.61	1.12	0	1.5
x_5	3	3.04	1.41	1.5	0

ตารางที่ 2.10: ตารางเมตริกซ์ระยะทางระหว่าง x_i และ x_j

จะเห็นได้ว่า ค่า distance ของ x_1 และ x_2 มีค่าน้อยที่สุด คือ 0.5

ดังนั้น เราจะทำการรวม x_1, x_2 เป็นกลุ่มเดียวกัน



จากนั้น ทหาระยะทางระหว่าง $\{x_1, x_2\}$ กับ x_3, x_4 และ x_5 จะได้ว่า

$$D(\{x_1, x_2\}, x_3) = \min \{d(x_1, x_3), d(x_2, x_3)\} = 2.24,$$

$$D(\{x_1, x_2\}, x_4) = \min \{d(x_1, x_4), d(x_2, x_4)\} = 3.35,$$

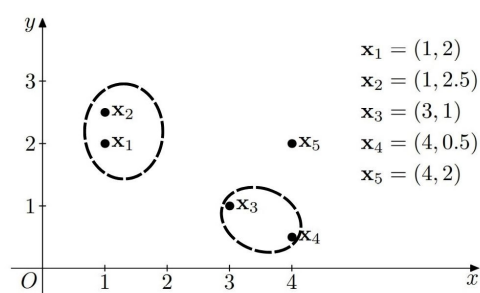
$$D(\{x_1, x_2\}, x_5) = \min \{d(x_1, x_5), d(x_2, x_5)\} = 3.$$

สร้างตารางเมตริกซ์ระยะทางได้ดังนี้

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.24	3.35	3
$\mathbf{x}_3$	2.24	0	1.12	1.41
$\mathbf{x}_4$	3.35	1.12	0	1.5
$\mathbf{x}_5$	3	1.41	1.5	0

จะเห็นว่า ค่า distance ของ x_3 และ x_4 มีค่าน้อยที่สุด คือ 1.12

ดังนั้น เราจะทำการรวม x_3, x_4 เป็นกลุ่มเดียวกัน



จากนั้น ทหาระยะทางระหว่าง $\{x_3, x_4\}$ กับ $\{x_1, x_2\}$ และ x_5 จะได้ว่า

$$\begin{aligned}
 D(\{x_3, x_4\}, \{x_1, x_2\}) &= \min \{d(x_1, x_3), d(x_2, x_3), d(x_1, x_4), d(x_2, x_4)\} \\
 &= \min \{D(\{x_1, x_2\}, x_3), D(\{x_1, x_2\}, x_4)\} \\
 &= 2.24
 \end{aligned}$$

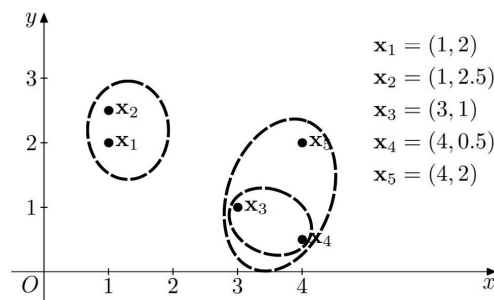
$$D(\{x_3, x_4\}, x_5) = \min \{d(x_3, x_5), d(x_4, x_5)\} = 1.41$$

สร้างตารางเมตริกซ์ระยะทางได้ดังนี้

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	2.24	3
$\{x_3, x_4\}$	2.24	0	1.41
x_5	3	1.41	0

จะเห็นว่า ค่า distance ของ $\{x_3, x_4\}$ และ x_5 มีค่าน้อยที่สุด คือ 1.41

ดังนั้น เราจะทำการรวม $\{x_3, x_4\}, x_5$ เป็นกลุ่มเดียวกัน



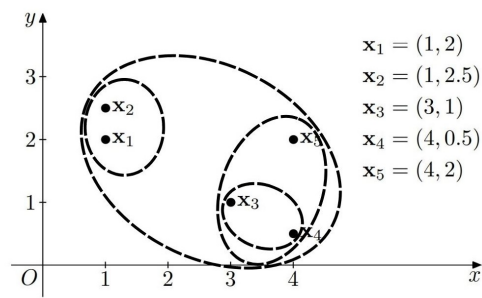
จากนั้น ทหาระยะทางระหว่าง $\{x_3, x_4, x_5\}$ และ $\{x_1, x_2\}$ จะได้ว่า

$$\begin{aligned}
 D(\{x_3, x_4, x_5\}, \{x_1, x_2\}) &= \min \{d(x_1, x_3), d(x_2, x_3), d(x_1, x_4), d(x_2, x_4), d(x_1, x_5), d(x_2, x_5)\} \\
 &= \min \{D(\{x_1, x_2\}, x_3), D(\{x_1, x_2\}, x_4), D(\{x_1, x_2\}, x_5)\} \\
 &= 2.24
 \end{aligned}$$

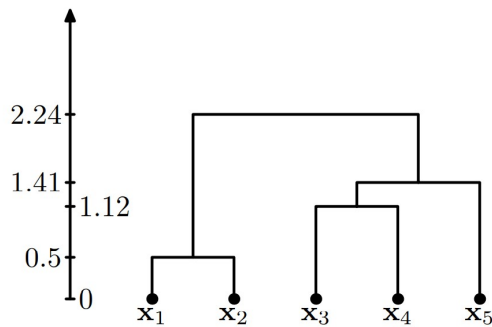
สร้างตารางเมตริกซ์ระยะทางได้ดังนี้

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	2.24
$\{x_3, x_4, x_5\}$	2.24	0

สุดท้าย รวม $\{x_1, x_2\}$ กับ $\{x_3, x_4, x_5\}$ เป็นกลุ่มเดียวกัน



เราสามารถสร้าง dendrogram เพื่อใช้ดูการจัดกลุ่มได้อย่างชัดเจน ดังนี้



รูปที่ 2.47: dendrogram โดยการใช้วิธี Single Linkage

ผลลัพธ์ของการรวมกลุ่มไม่ได้บอกค่าทางสถิติของจำนวนกลุ่มที่เหมาะสม เราจะต้องทำการวิเคราะห์จำนวนกลุ่มที่เหมาะสมเอง โดยอาจกำหนดตัวเลขของระยะห่าง หรือความคล้ายเป็นเกณฑ์ในการตัดสินใจ เมื่อกำหนดระยะห่างที่ 0.5 จะสามารถแบ่งได้ 4 กลุ่ม คือ กลุ่มที่ 1 $\{x_1, x_2\}$, กลุ่มที่ 2 $\{x_3\}$, กลุ่มที่ 3 $\{x_4\}$ และกลุ่มที่ 4 $\{x_5\}$ แต่ถ้าแบ่งระยะห่างที่ 1.5 จะสามารถแบ่งได้ 2 กลุ่ม คือ กลุ่มที่ 1 $\{x_1, x_2\}$ และกลุ่มที่ 2 $\{x_3, x_4, x_5\}$

บทที่ 3

ผลการศึกษา (Results)

ในงานค้นคว้าอิสระนี้เราได้ทำการ download file รหัสพันธุกรรมของผู้ติดเชื้อ SARS-CoV-2 ในประเทศไทย จำนวน 24,594 คน รหัสพันธุกรรมมีความยาว 29,903 เบส ซึ่งประกอบไปด้วย A T C G N -

โดย - คือ deletion และ N คือ ไม่ทราบค่า

reference genome ATCGGCCTTA

ผู้ป่วย 1 AT- -GTCTNN

จากตัวอย่างข้างต้น หมายความว่า ผู้ป่วย 1 มี deletion 2 ตำแหน่ง

และไม่ทราบค่า 2 ตำแหน่ง และมี mutation จาก C เป็น T

โดยเราจะเขียนสัญลักษณ์ แทนทุกตำแหน่งที่แตกต่างจาก reference genome

ดังนี้ C3-, G4-, C6T, T9N, A10N

สำหรับชื่อของสายพันธุกรรม ของไวรัส SARS-CoV-2 มีทั้งหมด 3 แบบ

1. Clade ประกอบด้วย S, L, V, G, GH, GR, GRY, GRA, GV และ GK
2. WHO name ประกอบด้วย Delta, Omicron, Alpha, Beta, Eta, Gamma, Kappa และ Others
3. Lineage ประกอบด้วย A, B, C, P, AY, BA, XM และ XAG

จากข้อมูลใน GISAID เราทราบตำแหน่งของ mutations ของแต่ละสายพันธุ์ ดังตาราง

clade	mutations
S	C8782T,T28144C
L	C241,C3037,A23403,C8782,G11083,G26144,T28144
V	G11083T,G26144T
G	C241T,C3037T,A23403G
GK	C241T,C3037T,A23403G,C22995A
GH	C241T,C3037T,A23403G,G25563T
GR	C241T,C3037T,A23403G,G28882A
GV	C241T,C3037T,A23403G,C22227T
GRY	C241T,C3037T

จากตาราง หากผู้ป่วยมีสายพันธุ์กรรมตำแหน่ง 8782 เป็น T และ 28144 เป็น C แสดงว่าผู้ป่วยคนนั้น ติดเชื้อสายพันธุ์ S แต่หากผู้ป่วยมีสายพันธุ์กรรมตำแหน่ง 241 เป็น C, 3037 เป็น C, 23403 เป็น A, 8782 เป็น C, 11083 เป็น G, 26144 เป็น G และ 28144 เป็น T แสดงว่าผู้ป่วยคนนั้นติดเชื้อสายพันธุ์ L ซึ่งเป็นสายพันธุ์เดียวกับ reference genome

เราจะทำการจำลองการแยกสายพันธุ์กรรม โดยอัลกอริทึมของต้นไม้ตัดสินใจ สำหรับ Clade

ให้ คุณลักษณะ คือ ตำแหน่งของสายพันธุ์กรรม (1-29903)

กิ่ง หรือค่าของคุณลักษณะ คือ ATCG – N

กลุ่ม คือ Clade

จากข้อมูลที่ download จะได้ว่า

ผู้ติดเชื้อ Clade S มี 219 คน

Clade L มี 18 คน

Clade G มี 296 คน

Clade GH มี 611 คน

Clade GR มี 1,418 คน

Clade GRY มี 796 คน

Clade GRA มี 12,754 คน

Clade GV มี 12 คน

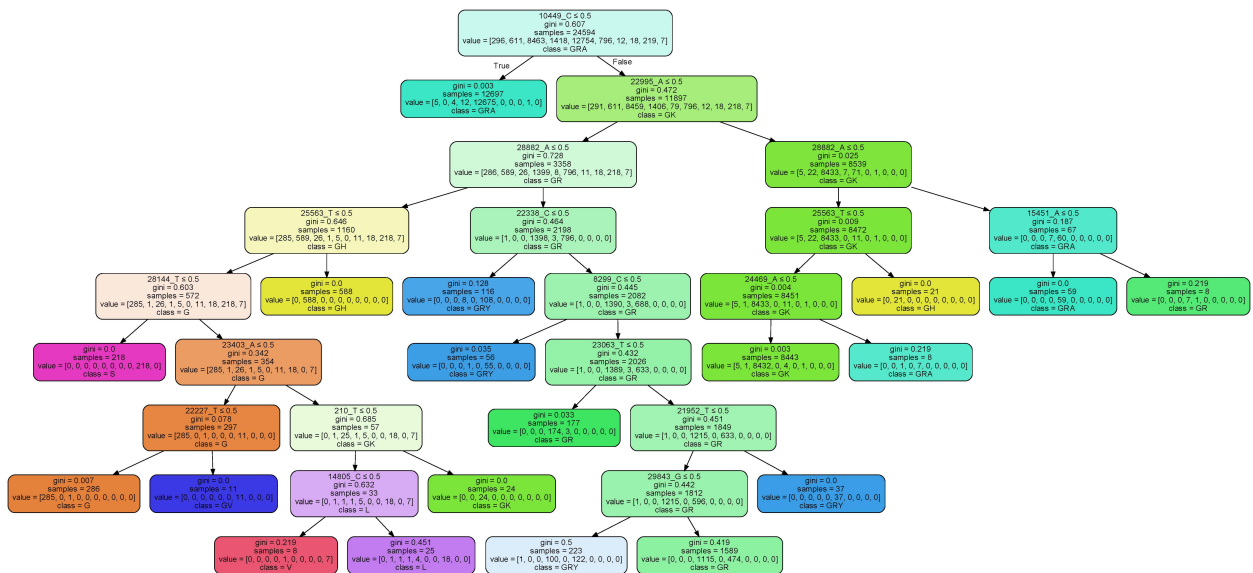
Clade GK มี 8,463 คน

Clade V มี 7 คน

เมื่อทำการ รันโปรแกรม python โดยใช้ Decision Tree Classifier ของ Scikit-Learn โดยไม่ใช้ class weight

```
DecisionTreeClassifier(criterion='gini', max_depth=8,min_samples_leaf=8,min_samples_split=30,max_leaf_nodes=19)
```

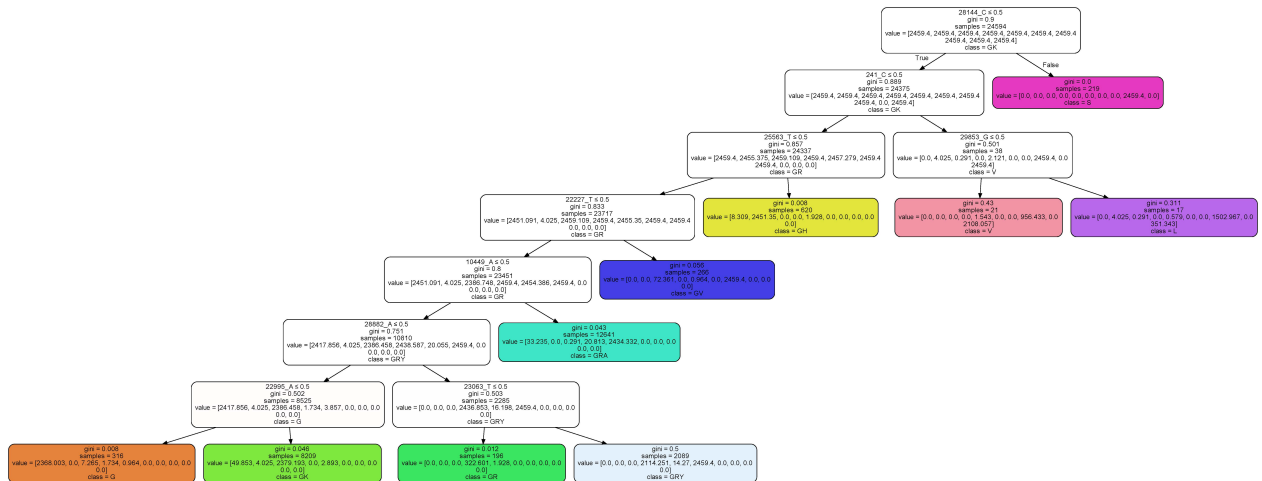
ได้ผลลัพธ์ คือ



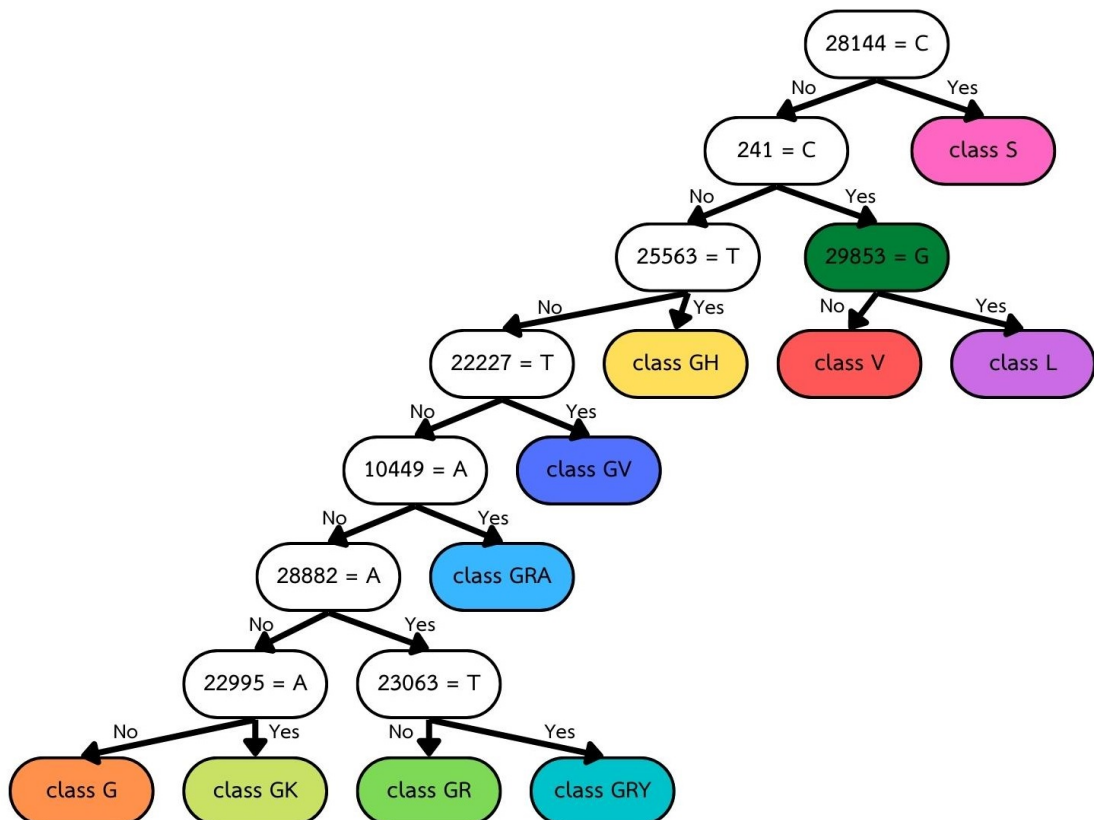
เมื่อทำการจำลองการแยกสายพันธุ์กรรม โดยอัลกอริทึมของต้นไม้ตัดสินใจ โดยไม่ใช้ class weight เข้ามาช่วย จะสามารถแยกสายพันธุ์กรรมได้ แต่ตำแหน่งของ mutations ของแต่ละสายพันธุ์ไม่สอดคล้องกับข้อมูลใน GDSIAD เนื่องจากชุดข้อมูลมีจำนวนข้อมูลในแต่ละสายพันธุ์ไม่เท่ากัน จะเห็นได้ว่า ในกลุ่มสายพันธุ์ของ Clade GRA มีจำนวนผู้ติดเชื้อ 12,754 คน แต่ในกลุ่มสายพันธุ์ของ Clade V มีจำนวนผู้ติดเชื้อ 7 คน ซึ่งมีจำนวนที่แตกต่างกันมาก ดังนั้นเราจึงจำเป็นต้องใช้ class weight เพื่อให้แต่ละกลุ่มมีความสมดุล และให้ความสำคัญของแต่ละสายพันธุ์เท่ากัน

เมื่อทำการ รันโปรแกรม python โดยใช้ Decision Tree Classifier ของ Scikit-Learn โดยใช้ class weight

ได้ผลลัพธ์ คือ



เขียนเป็นต้นไม้ตัดสินใจอย่างง่าย ได้ดังนี้



เมื่อทำการจำลองการแยกสายพันธุ์กรรม โดยอัลกอริทึมของต้นไม้ตัดสินใจ โดยใช้ class weight เข้ามาช่วย จะสามารถแยกสายพันธุ์กรรมได้ และมีตำแหน่งของ mutations ของแต่ละสายพันธุ์สอดคล้องกับข้อมูลใน GISAID และได้ต้นไม้ตัดสินใจที่สามารถดูและอ่านได้ง่าย

หากเราบอกว่า ผู้ป่วยมี mutations C241T, C3037T หมายความว่า ตำแหน่งอื่น ๆ ในสายพันธุ์กรรมเหมือน reference genome ยกเว้น ตำแหน่ง 241 ที่เปลี่ยนจาก C เป็น T และตำแหน่ง 3037 ที่เปลี่ยนจาก C เป็น T

สมมติผู้ป่วยมี mutations คือ C241T, C3037T, A23403G, G25563T ดังนั้น ตำแหน่งอื่นเหมือน reference genome นั่นคือ ตำแหน่ง 28144 เป็น T เมื่อนำไปเข้าในต้นไม้ตัดสินใจ จะได้ว่าผู้ป่วยคนนี้ติดเชื้อสายพันธุ์ GH

ทั้งนี้ ต้นไม้ตัดสินใจที่เราสร้างอาจไม่ตรงกับผลลัพธ์ของทั่วโลกเพราะขึ้นอยู่กับหลายปัจจัย เช่น จำนวนข้อมูล ลักษณะของสายพันธุ์กรรมที่มาจากคนละเชื้อชาติ หรือแม้กระทั่ง การกำหนดค่า พารามิเตอร์ใน model แต่อย่างน้อย ด้วยจำนวนข้อมูลที่มีมากเพียงพอ ก็สามารถได้ผลลัพธ์ที่น่าเชื่อถือในการจำแนกสายพันธุ์กรรม

ต่อไป เราจะทำการจำลองการแยกสายพันธุ์กรรม โดยอัลกอริทึมของต้นไม้ตัดสินใจ สำหรับ WHO name

ให้ คุณลักษณะ คือ ตำแหน่งของสายพันธุ์กรรม (1-29903)

กึ่ง หรือค่าของคุณลักษณะ คือ ATCG – N

กลุ่ม คือ WHO name

จากข้อมูลที่ download จะได้ว่า

ผู้ติดเชื้อ WHO name : Delta มี 8672 คน

WHO name : Omicron มี 12738 คน

WHO name : Alpha มี 2120 คน

WHO name : Beta มี 105 คน

WHO name : Eta มี 6 คน

WHO name : Gamma มี 1 คน

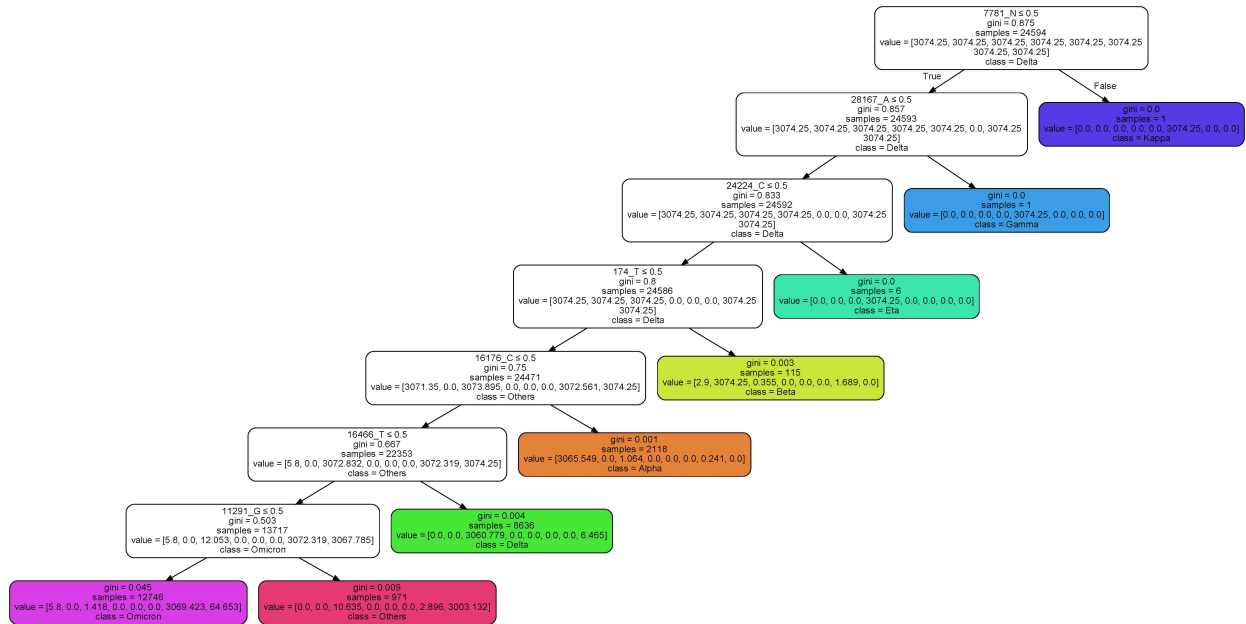
WHO name : Kappa มี 1 คน

WHO name : Others มี 951 คน

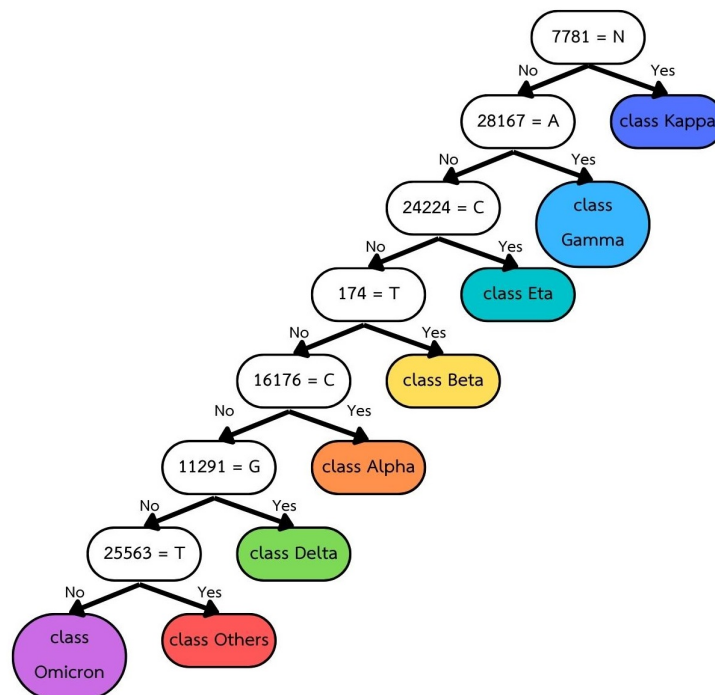
เมื่อทำการ รันโปรแกรม python โดยใช้ Decision Tree Classifier ของ Scikit-Learn โดยใช้ class weight

```
DecisionTreeClassifier(criterion='gini', class_weight='balanced',max_leaf_nodes=8)
```

ได้ผลลัพธ์ คือ



เขียนเป็นต้นไม้ตัดสินใจอย่างง่ายในการจำแนก WHO name เป็นดังนี้



ต่อไป เราจะสมมติ สถานการณ์การติดเชื้อในกลุ่มคน ซึ่งอาจเป็น ในโรงพยาบาล หรือในหอพัก ทั้งหมด 60 คน เป็นสายพันธุ์ Delta 30 คน และ Omicron 30 คน จะใช้ Agglomerative Hierarchical Clustering ในการหา กลุ่มคนที่มีสายพันธุ์กรรมที่คล้ายกัน

ในที่นี้ เราให้ $d(x, y)$ = hamming distance ระหว่าง x และ y

บทนิยาม 3.1 ระยะทางฮามมิง(Hamming Distance) กำหนดให้ $u, v \in \mathbb{F}^n$ ระยะทางฮามมิงของค่า u และ v คือ จำนวนตำแหน่งที่แตกต่างกัน ทั้ง u และ v สัญลักษณ์ที่ใช้ คือ

$$d_H(u, v) = |\{i \in \{1, 2, 3, \dots, n\}; u_i \neq v_i\}|$$

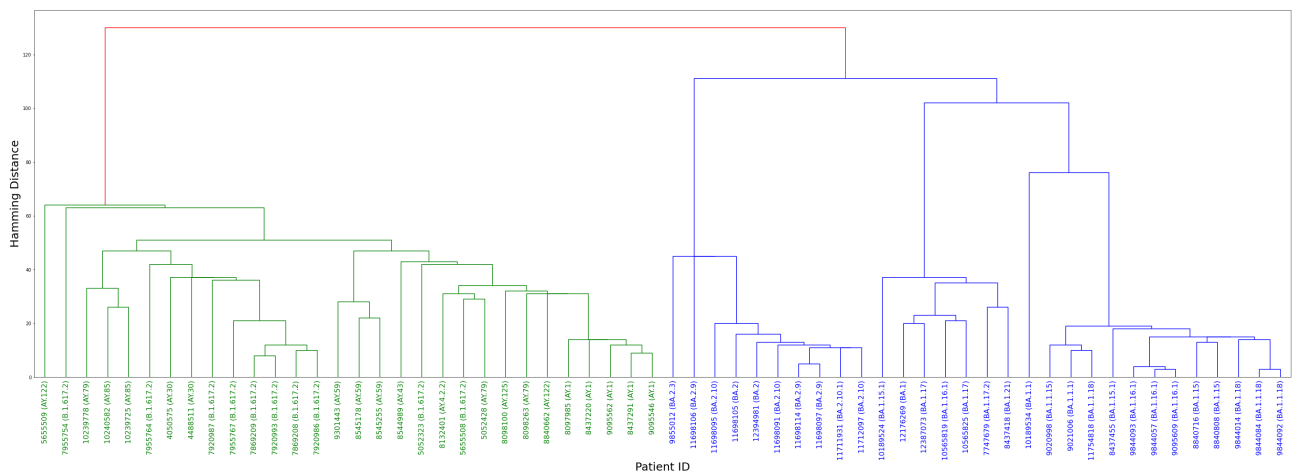
ตัวอย่าง 3.2 $d_H(11044, 55044) = |\{1, 2\}| = 2$

$$d_H(20314, 20314) = |\{\}| = 0$$

เมื่อ รันโปรแกรม

```
hierarchy.linkage(Y, method='single')
```

จะได้ dendrogram ดังนี้



เส้นสีเขียว คือ ผู้ป่วย สายพันธุ์ omicron

เส้นสีน้ำเงิน คือ ผู้ป่วย สายพันธุ์ delta

เราจะเห็นว่า Agglomerative Hierarchical Clustering สามารถแบ่งกลุ่ม ผู้ป่วยแต่ละสายพันธุ์ได้ นอกจากนี้ยังสามารถบอกได้ว่าผู้ป่วยคนใดมีสายพันธุ์กรรมใกล้เคียงกัน ซึ่งอาจคาดการณ์ได้ว่า ติดเชื้อมาจากที่เดียวกัน หรือ ติดจากกัน ซึ่งถ้าเป็นสถานการณ์จริงในโรงพยาบาล เราก็สามารถรู้ได้ว่า มีการติดเชื้อในโรงพยาบาล หรือไม่ หากมีการติดเชื้อในโรงพยาบาล ก็สามารถมีมาตรการจัดการได้ทันทั่วทั้ง

บทที่ 4

สรุปผลการศึกษา (Conclusion)

ในการค้นคว้าอิสระนี้ เราได้ศึกษาการสร้างโมเดลจำลองการจำแนกพันธุกรรมของเชื้อไวรัส SARS-CoV-2 สามารถจำแนกได้ โดยสร้างต้นไม้ตัดสินใจ เพื่อจำแนกผู้ติดเชื้อ SARS-CoV-2 ออกเป็นกลุ่มต่าง ๆ ตาม Clade และ WHO name โดยใช้การกระจายของข้อมูลเป็นตัววัด คือ Gini Index ซึ่งแบ่งออกเป็น Gini Index แบบไม่มี class weight และ Gini Index แบบ class weight

สำหรับการสร้างต้นไม้ตัดสินใจโดยใช้ Gini Index แบบไม่มี class weight จะสามารถแยกสายพันธุ์ได้ แต่ตำแหน่งของ mutations ของแต่ละสายพันธุ์ไม่สอดคล้องกับข้อมูลใน GDSIAD เนื่องจากชุดข้อมูลมีจำนวนข้อมูลในแต่ละสายพันธุ์ไม่เท่ากัน และมีจำนวนแตกต่างกันมาก ในทางกลับกัน การสร้างต้นไม้ตัดสินใจโดยใช้ Gini Index แบบ class weight จะสามารถแยกสายพันธุ์ได้ และมีตำแหน่งของ mutations ของแต่ละสายพันธุ์สอดคล้องกับข้อมูลใน GISAID

นอกจากนี้ ได้ศึกษาการแบ่งกลุ่มข้อมูล เพื่อหาผู้ป่วยที่สายพันธุ์เหมือนกัน โดยการทำ Clustering Model ด้วย Agglomerative Hierarchical Cluster Analysis โดยใช้วิธี Single Linkage สามารถหาผู้ป่วยที่มีสิทธิ์ติดเชื้อ SARS-CoV-2 มาจากที่เดียวกันได้ และสามารถคาดการณ์กลุ่มของการระบาดได้

บรรณานุกรม

- [1] Chris Ding and Xiaofeng He., Cluster merging and splitting in hierarchical clustering algorithms., IEEE Xplore, 2002
- [2] T. Daniya and Suresh Kumar K Dr., Classification and regression trees with gini index., Advances in Mathematics Scientific Journal, 2020
- [3] T. Suryakanthi., Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*, International Journal of Advanced Computer Science and Applications, 2020.
- [4] Carl Löndahl., Some Notes on Code-Based Cryptography., Lund University, Sweden, 2014
- [5] <https://calmbelt-demo.mtms.dev/about> (วันที่ 24 เดือน มิถุนายน ปี พ.ศ.2565)
- [6] <https://datacube.ai/decision-tree/> (วันที่ 11 เดือน มีนาคม ปี พ.ศ.2565)
- [7] <https://gisaid.org/resources/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/> (วันที่ 24 เดือน มิถุนายน ปี พ.ศ.2565)
- [8] <https://medium.com/tni-university/> (วันที่ 14 เดือน มิถุนายน ปี พ.ศ.2565)
- [9] [http://mitran-lab.amath.unc.edu/courses/MATH590/biblio/Clustering.ch7.Hierarchical Clustering.pdf](http://mitran-lab.amath.unc.edu/courses/MATH590/biblio/Clustering.ch7.Hierarchical%20Clustering.pdf) (วันที่ 14 เดือน มิถุนายน ปี พ.ศ.2565)
- [10] <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm> (วันที่ 9 เดือน ตุลาคม ปี พ.ศ.2565)
- [11] <http://www.pratya.nuankaew.com/wp-content/uploads/2017/10/cluster-analysis.pdf> (วันที่ 14 เดือน มิถุนายน ปี พ.ศ.2565)
- [12] <https://www.slideshare.net/NontawatB/05-classification-1-decision-tree-and-rule-based-classification> (วันที่ 11 เดือน มีนาคม ปี พ.ศ.2565)

[13] <https://quantdare.com/decision-trees-gini-vs-entropy> (วันที่ 20 เดือน ตุลาคม ปี พ.ศ.2565)

[14] <https://www.who.int/thailand/news/feature-stories/detail/tracking-SARS-CoV-2-variants> (วันที่ 24 เดือน มิถุนายน ปี พ.ศ.2565)

ภาคผนวก

Identifying genomic patterns across SARS-CoV-2 lineages

Mr. Yongyut Aunyo

Advisor : Asst. Prof. Dr. Hatairat Yingtaeewesittikul

Department of Mathematics, Faculty of Science, Chiang Mai University

Abstract

In this independent study, we study decision trees and single linkage agglomerative hierarchical algorithms for identifying genomic patterns across SARS-CoV-2 lineages and determining patient clusters based on genomic information.

Main Results

In this independent study, we downloaded genomic data of SARS-CoV-2 detected in COVID patients in Thailand from GISAID[5]. The dataset consists of 24,594 genomes. For each genome, there are 29,903 attributes and a class/label (i.e., GISAID clade).

We then constructed decision trees based on Gini Index to identify the best attribute values at each step (figure below). We also balanced the class weight to take into account different numbers of genomes per class.

Preliminaries

Decision Tree

A decision tree is a supervised learning technique that can be used for both classification and regression problems. It is a tree-structured classifier where internal nodes represent the attributes of a dataset, branches represent the decision rules, and each leaf node represents the classes. There are two widely used techniques for constructing a decision tree:

- Gini Index**

$$Gini_{split}(T) = \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Gini(t_i)$$

Where $Gini(t_i) = 1 - \sum_{k=1}^N \left[\frac{s_k}{s_i} \right]^2$

- Information Gain(IG)**

$$IG(T) = Entropy(T) - \sum_{i=1}^M \frac{n_i}{\sum_{j=1}^M n_j} Entropy(t_i)$$

Where

$$Entropy(T) = - \sum_{k=1}^N \left[\frac{r_k}{\sum_{l=1}^N r_l} \log_2 \left(\frac{r_k}{\sum_{l=1}^N r_l} \right) \right]$$

$$Entropy(t_i) = - \sum_{k=1}^N \left[\frac{s_k}{\sum_{l=1}^N s_l} \log_2 \left(\frac{s_k}{\sum_{l=1}^N s_l} \right) \right]$$

Agglomerative Hierarchical Cluster Analysis

An agglomerative hierarchical clustering begins with each object or observation in a separate cluster. In each subsequent step, two clusters that are the most similar are combined into a new aggregate cluster. The combining process is repeated until all objects are finally combined into a single cluster, i.e., from n clusters to 1. We select the single linkage method for clustering. The formula for finding it is as follows :

$$D(C, C') = \min_{x \in C, y \in C'} d(x, y)$$

Where $C \cap C' \neq \emptyset$, $d(x, y)$ is distance between x and y

Figure 1 : Decision Tree reveal each Clade

To evaluate whether we could determine the correct patient clusters, we created a synthetic dataset of 60 genomes from 30 and 30 patients infected with Omicron and Delta, respectively. The dendrogram figure below demonstrates the correct patient clusters based on agglomerative hierarchical clustering.

Figure 2 : Dendrogram that green represented Delta and blue represented Omicron

Reference

- [1] Chris Ding and Xiaofeng He., Cluster merging and splitting in hierarchical clustering algorithms., IEEE Xplore, 2002
- [2] T. Daniya and Suresh Kumar K Dr., Classification and regression trees with gini index., Advances in Mathematics Scientific Journal, 2020
- [3] T. Suryakanthi., Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*, International Journal of Advanced Computer Science and Applications, 2020.
- [4] Carl Löndahl., Some Notes on Code-Based Cryptography., Lund University, Sweden, 2014
- [5] "Clade and lineage nomenclature, " 2 March 2021. [Online]. Available: <https://gisaid.org/resources/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/>.

รูปที่ 4.1: โพสต์เตอร์