บทที่ 3
วิธีการดำเนินงานโครงงาน

โครงงานเรื่อง การวิเคราะห์ข้อมูลการแยกประเภทของเห็ดในอเมริกาเหนือ เพื่อเผยแพร่ข้อมูลบนเว็บไซต์ ในบทนี้จะเป็นการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้าไมนิ่ง ซึ่งมีกระบวนการวิเคราะห์ที่สำคัญหลายขั้นตอน เมื่อเสร็จสิ้นจากกระบวนการวิเคราะห์ข้อมูลแล้วจะเป็นการออกแบบเว็บไซต์ และออกแบบรูปแบบการแสดงผลและบทสรุปจากวิธีการดำเนินงาน

3.1 การวิเคราะห์ข้อมูลด้วย CRISP-DM

3.2 การออกแบบเว็บไซต์

3.3 บทสรุป

3.1 การวิเคราะห์ข้อมูลด้วย CRISP-DM

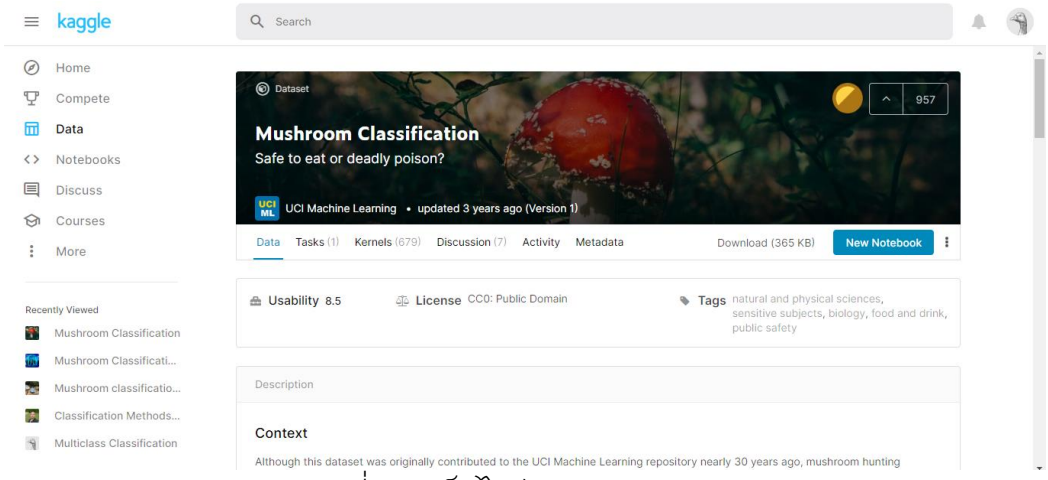
กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM หรือ Cross Industry Standard Process for Data Mining พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือของ 3 บริษัทคือ Daimler Chrysler, SPSS และ NCR ที่มีการพัฒนาเป็น Workflow มาตรฐานสำหรับการทำเหมืองข้อมูล ประกอบด้วย 6 ขั้นตอนหลัก ดังนี้

3.1.1 รู้จักและเข้าใจในธุรกิจ (Business Understanding) เป็นขั้นตอนแรกของกระบวนการที่มุ่งเน้นไปที่การทำความเข้าใจกระบวนการทางธุรกิจโดยรวม

ผู้วิเคราะห์ข้อมูลทำความเข้าใจกับปัญหาให้อยู่ในรูปของการวิเคราะห์ข้อมูลทาง Data Mining โดยการวิเคราะห์ข้อมูลในประเด็นนี้ คือ การวิเคราะห์ข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ ซึ่งมีข้อมูลมาก จำนวนรายการทั้งหมด 8,125 รายการ ทำให้ไม่สามารถทำความเข้าใจกับข้อมูลที่มีความซับซ้อนได้อย่างรวดเร็ว เช่น ต้องการทราบว่าลักษณะคุณสมบัติพื้นผิวเห็ดแบบใดที่บ่งบอกถึงเห็ดพิษมากที่สุด

3.1.2 จัดเก็บและรวบรวมข้อมูลให้ครบ (Data Understanding) ขั้นตอนการจัดเก็บและรวบรวมข้อมูล ตลอดจนการพิจารณาตรวจสอบความถูกต้องของข้อมูลที่ได้รับ โดยเลือกว่าจะใช้ข้อมูลทั้งหมดหรือบางส่วนในการวิเคราะห์ให้สอดรับกับวัตถุประสงค์ที่กำหนดไว้

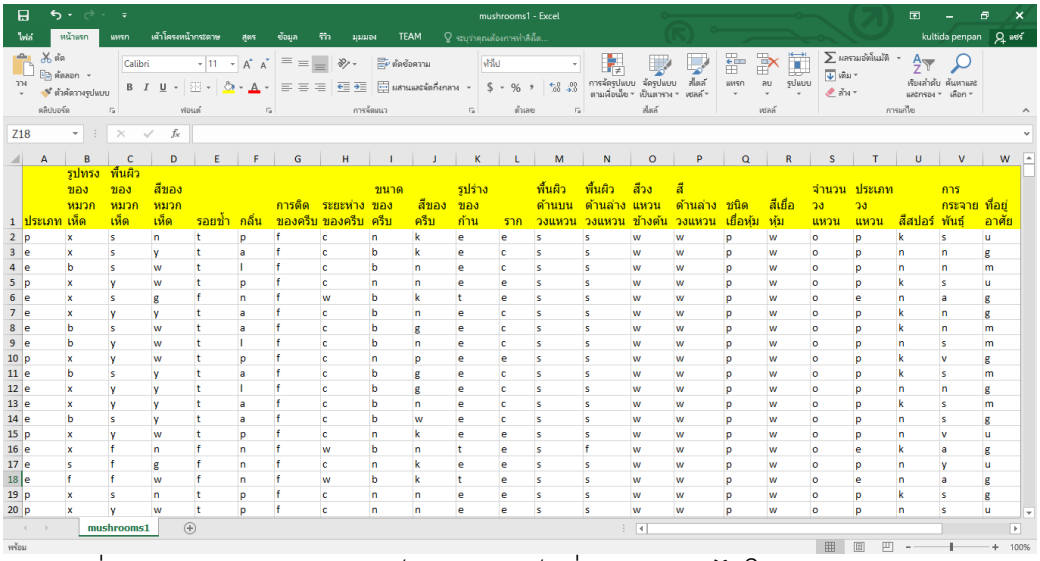
ผู้วิเคราะห์ข้อมูลทำการรวบรวมข้อมูล เพื่อตรวจสอบรายละเอียด ปริมาณ และ ความน่าเชื่อถือของข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ ที่ได้จากเว็บไซต์ kaggle.com ซึ่งเป็นเว็บไซต์ที่เก็บรวบรวมชุดข้อมูลต่าง ๆ เป็นแหล่งรวม Datasets หรือ ชุดข้อมูลสำหรับฝึกสอน Machine Learning ที่ใหญ่ที่สุดในโลกแห่งหนึ่ง มีข้อมูลทุกประเภท ไม่ว่าจะเป็น Datasets ในหมวดหมู่ Finance, Business, Physics, Biology, Sports, News ซึ่งเป็นข้อมูลที่เปิดเผยได้ เพื่อให้ผู้ใช้บริการสามารถนำชุดข้อมูลไปศึกษาหรือวิเคราะห์ให้เกิดประโยชน์ต่อไป



ภาพที่ 3. 1 เว็บไซต์ kaggle.com

(ที่มา : https:www.kaggle.com)

ซึ่งข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ มีจำนวนข้อมูล 8,125 รายการ ประกอบด้วย 23 แอตทริบิวท์ ประกอบด้วย ประเภท รูปทรงของหมวกเห็ด พื้นผิวของหมวกเห็ด สีของหมวกเห็ด รอยช้ำ กลิ่น การติดของครีบ ระยะห่างของครีบ ขนาดของครีบ สีของครีบ รูปร่างของก้าน ราก พื้นผิวด้านบนวงแหวน พื้นผิวด้านล่างวงแหวน สีวงแหวนข้างต้น สีด้านล่างวงแหวน ชนิดเยื่อหุ้ม สีเยื่อหุ้ม จำนวนวงแหวน ประเภทวงแหวน สีสปอร์ จำนวน การกระจายพันธุ์ ที่อยู่อาศัย



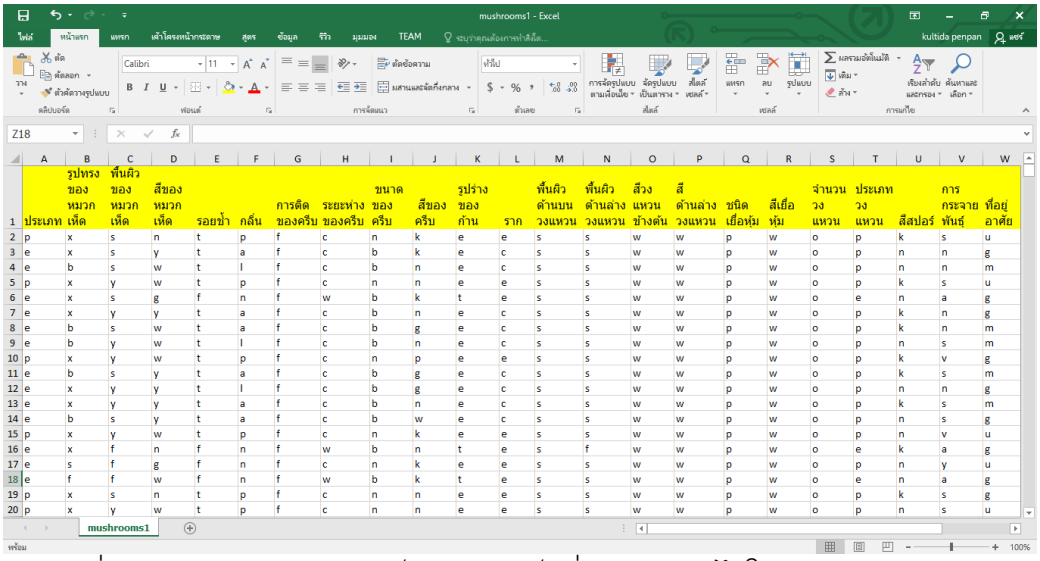
ประเภท	รูปทรงของหมวกเห็ด	พื้นผิวของหมวกเห็ด	สีของหมวกเห็ด	รอยช้ำ	กลิ่น	การติดของครีบ	ระยะห่างของครีบ	ขนาดของครีบ	สีของครีบ	รูปร่างของก้าน	ราก	พื้นผิวด้านบนวงแหวน	พื้นผิวด้านล่างวงแหวน	สีวงแหวนข้างต้น	สีด้านล่างวงแหวน	ชนิดเยื่อหุ้ม	สีเยื่อหุ้ม	จำนวนวงแหวน	ประเภทวงแหวน	สีสปอร์	การกระจายพันธุ์	ที่อยู่อาศัย
p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n	g
e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n	m
p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	n	g
e	b	s	w	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	n	m
e	b	y	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	s	m
p	x	y	w	t	p	f	c	n	p	e	e	s	s	w	w	p	w	o	p	k	v	g
e	b	s	y	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	s	m
e	x	y	y	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	n	n	g
e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	s	m
e	b	s	y	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n	s	g
p	x	y	w	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	v	u
e	x	f	n	f	n	f	w	b	n	t	e	s	f	w	w	p	w	o	e	k	a	g
e	s	f	g	f	n	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	y	u
e	f	f	w	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
p	x	s	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	g
p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n	s	u

ภาพที่ 3. 2 ข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ

3.1.3 เตรียมข้อมูลให้พร้อมใช้งาน (Data Preparation) ขั้นตอนการแปลงข้อมูลที่ได้รวบรวมมาและเลือกไว้ ให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปวิเคราะห์ในขั้นตอนต่อไปได้ โดยการทำให้เป็นข้อมูลที่ถูกต้อง (Data cleaning) มักใช้เวลาค่อนข้างมาก โดยมีขั้นตอนดังนี้

3.1.3.1 ทำการคัดเลือกข้อมูล (Data Selection) คือการคัดเลือกข้อมูลที่เหมาะสมเพื่อนำมาใช้ในการวิเคราะห์ข้อมูล

ผู้วิเคราะห์ข้อมูลทำการคัดเลือกข้อมูล และทำการ Data Cleaning ข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ โดยตัดส่วนที่ไม่จำเป็นออกให้เหลือเฉพาะข้อมูลที่จำเป็นในการวิเคราะห์ในภาพรวม จำนวน 5 แอตทริบิวท์ ได้แก่ ประเภทของเห็ด รูปทรงของหมวกเห็ด พื้นผิวของหมวกเห็ด รอยช้ำ และที่อยู่อาศัย จำนวน 3,377 รายการ เนื่องจากการศึกษาจากแหล่งอ้างอิงถึงคุณลักษณะของเห็ดพิษและเห็ดที่รับประทานได้ มักจะสังเกตได้จากรูปทรงของหมวกเห็ด พื้นผิวของหมวกเห็ด รอยช้ำ และที่อยู่อาศัย ซึ่งเป็นข้อมูลที่จำเป็นในการนำไปวิเคราะห์ข้อมูลต่อไป



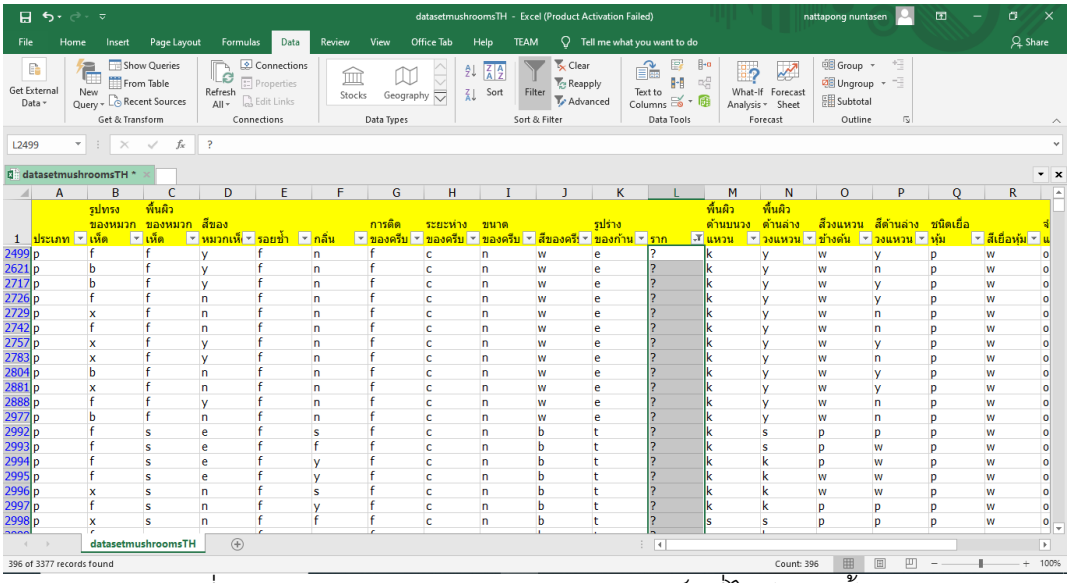
ประเภท	รูปทรงของหมวกเห็ด	พื้นผิวของหมวกเห็ด	สีของหมวกเห็ด	รอยช้ำ	กลิ่น	การติดของครีบ	ระยะห่างของครีบ	ขนาดของครีบ	สีของครีบ	รูปร่างของก้าน	ราก	พื้นผิวด้านบนวงแหวน	พื้นผิวด้านล่างวงแหวน	สีวงแหวนข้างต้น	สีด้านล่างวงแหวน	ชนิดเยื่อหุ้ม	สีเยื่อหุ้ม	จำนวนวงแหวน	ประเภทวงแหวน	สีสปอร์	การกระจายพันธุ์	ที่อยู่อาศัย
p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n	g
e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n	m
p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	n	g
e	b	s	w	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	n	m
e	b	y	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	s	m
p	x	y	w	t	p	f	c	n	p	e	e	s	s	w	w	p	w	o	p	k	v	g
e	b	s	y	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	s	m
e	x	y	y	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	n	n	g
e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	s	m
e	b	s	y	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n	s	g
p	x	y	w	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	v	u
e	x	f	n	f	n	f	w	b	n	t	e	s	f	w	w	p	w	o	e	k	a	g
e	s	f	g	f	n	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	y	u
e	f	f	w	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
p	x	s	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	g
p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n	s	u

ภาพที่ 3. 3 ข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ

3.1.3.2 ทำการกลั่นกรองข้อมูล (Data Cleaning) คือการทำความสะอาดข้อมูลเป็นกระบวนการตรวจสอบและการแก้ไข (หรือลบ) รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูลตารางหรือฐานข้อมูล ซึ่งเป็นหลักสำคัญของฐานข้อมูล ทางผู้วิเคราะห์ข้อมูลได้ดำเนินการดังนี้

1) ข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ ผู้วิเคราะห์ข้อมูลทำการแก้ไขและเปลี่ยนแปลงที่ข้อมูลที่ผิดพลาดและไม่ชัดเจนทิ้ง ซึ่งผู้วิเคราะห์ข้อมูลพบว่า บางข้อมูลราก นั้นมีค่าที่ไม่ชัดเจน จึงดูข้อมูลตามหลักความเป็นจริงโดยเปรียบเทียบข้อมูลจากส่วนใหญ่ที่ใช้ หรือข้อมูลอื่น ๆ ที่มีข้อผิดพลาด เนื่องด้วยอาจจะมาจากการคีย์ข้อมูลที่ต้นทาง ดังนั้นผู้วิเคราะห์ข้อมูลได้ทำการเปลี่ยนแปลงข้อมูลที่ไม่ชัดเจนได้ ดังนี้

– ราก. ? เป็นชนิดของรากเห็ดที่ไม่รู้จักหรืออาจจะเป็นข้อผิดพลาดของข้อมูล ผู้วิเคราะห์ข้อมูลพบว่าจะทำการเปลี่ยนแปลงชนิดของราก จึงได้ทำการตรวจสอบข้อมูลและทำการลบเปลี่ยนแปลงข้อมูลที่เป็น ? เป็นตัว n “ไม่มีราก” ทั้งหมด



ภาพที่ 3. 4 ข้อมูลผิดพลาดของชนิดของรากเห็ดที่ไม่ชัดเจนทั้งหมด

3.1.4 สร้างแบบจำลอง (Modeling) ขั้นตอนการสร้างตัวแบบทางคณิตศาสตร์ และสถิติเพื่อการวิเคราะห์ข้อมูล โดยสามารถใช้เทคนิควิธีการต่าง ๆ อาทิ การจำแนก (Classification) การแบ่งกลุ่ม (Clustering) และการสร้างความสัมพันธ์ (Association rule)

ผู้วิเคราะห์ข้อมูลวิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูลแบบ Data Classification เพื่อใช้ทำนายแนวโน้มการเกิดขึ้นของปัจจัยที่สามารถทำให้รู้ได้ว่าเห็ดชนิดไหนเป็นพิษและไม่เป็นพิษรับประทานได้หรือรับประทานไม่ได้ จากเทคนิคการจำแนกประเภทข้อมูล ด้วยการสร้างโมเดล Decision Tree เพื่อจัดกลุ่มข้อมูลที่มีความสัมพันธ์กัน โดยใช้โปรแกรมที่ใช้ทำเหมืองข้อมูลด้วยชุดข้อมูลที่คัดเลือก ดังนี้

	A	B	C	D	E
1	ประเภท	รูปทรง	พื้นผิว	รอยช้ำ	ที่อาศัย
2	p	x	s	t	u
3	e	x	s	t	g
4	e	x	s	f	g
5	e	b	s	t	g
6	e	x	f	f	g
7	e	f	f	f	g
8	p	x	s	t	g
9	p	x	s	t	u
10	p	f	s	t	g
11	e	f	f	f	u
12	e	x	s	t	d
13	e	x	f	t	d
14	e	x	f	t	d
15	e	x	f	f	u
16	e	x	s	t	g
17	p	x	s	t	u
18	e	f	f	f	g
19	e	b	s	t	g
20	e	x	s	t	g
21	e	f	s	f	g
22	e	x	s	t	g

ภาพที่ 3. 5 ชุดข้อมูลที่คัดเลือกมาวิเคราะห์ข้อมูล

จากรูปภาพที่ 3.5 ประกอบด้วย 5 แอตทริบิวต์ คือ

- ประเภท แสดงประเภทเห็ด ซึ่งเป็นค่าประกอบด้วย 2 ค่า คือ e"รับประทานได้" p" รับประทานไม่ได้"
- รูปทรง แสดงรูปทรงของเห็ด ประกอบด้วย 3 ค่า คือ b"ระฆัง" f"แบน" x"นูน"
- พื้นผิว แสดงพื้นผิวของเห็ด ประกอบด้วย 2 ค่า คือ f"เส้นใย" s"เรียบ"
- รอยฟกช้ำ แสดงรอยฟกช้ำของเห็ด ประกอบด้วย 2 ค่า คือ t"มี" f"ไม่มี"
- ที่อาศัย แสดงที่อาศัยของเห็ด ประกอบด้วย 3 ค่า คือ d"ป่า" g"ทุ่งหญ้า" u"เมือง"

การสร้างโมเดล decision tree จะทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของ tree (root node) หลังจากนั้นก็จะหาแอตทริบิวต์ถัดไปเรื่อย ๆ ในการหาความสัมพันธ์ของแอตทริบิวต์นี้จะใช้ตัววัด ที่เรียกว่า Information Gain (IG) ค่านี้คำนวณได้จากสมการดังนี้

- IG (parent, child) = entropy(parent) – [p(c1) × entropy(c1) + p(c2) × entropy(c2) + ...]
  - โดยที่ entropy(c1) = –p(c1) log p(c1) และ p(c2) คือ ค่าความน่าจะเป็นของ c1

การคำนวณค่าแต่ละแอตทริบิวต์เทียบกับคลาสเพื่อหาแอตทริบิวต์ที่มีค่า IG มากที่สุดมาเป็น root ของ decision tree กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้ ดังนี้

1) คำนวณค่า IG ของแอตทริบิวต์ รูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.63 \times \log_2(0.63) + 0.37 \times \log_2(0.37)]$

$= -[0.63 \times -0.67 + 0.37 \times -1.43]$

$= -[0.42 + 0.53]$

$= 0.95$

entropy(ผล = b) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.93 \times \log_2(0.93) + 0.07 \times \log_2(0.07)]$

$= -[0.93 \times -0.10 + 0.07 \times -3.84]$

$= -[0.09 + 0.27]$

= 0.36

entropy(ผล =f) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.63 × $\log_2$(0.63) + 0.37 × $\log_2$(0.37)]

= –[0.63 × –0.67 + 0.37 × –1.43]

= –[0.42 + 0.53]

= 0.95

entropy(ผล =x) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.60 × $\log_2$(0.60) + 0.40 × $\log_2$(0.40)]

= –[0.60 × –0.74 + 0.40 × –1.32]

= –[0.44 + 0.53]

= 0.97

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f)+ p(ผล = x) × entropy(ผล = x)]

= 0.95 – [0.05 × 0.36 + 0.42 × 0.95 + 0.53 × 0.97]

= 0.95 – [0.02 + 0.40 + 0.51]

= 0.95 – 0.93

= 0.02

2) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิว จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.63 × $\log_2$(0.63) + 0.37 × $\log_2$(0.37)]

= –[0.63 × –0.67 + 0.37 × –1.43]

= –[0.42 + 0.53]

= 0.95

entropy(ผล = f) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.73 × $\log_2$(0.73) + 0.27 × $\log_2$(0.27)]

= −[0.73 × −0.45 + 0.27 × −1.89]

= −[0.33 + 0.51]

= 0.84

entropy(ผล =s) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.48 × $\log_2$(0.48) + 0.52 × $\log_2$(0.52)]

= −[0.48 × −1.06 + 0.52 × −0.94]

= −[0.51 + 0.49]

= 1

IG(parent, child) = entropy (parent) – [p(ผล =f) × entropy(ผล = f) + p(ผล = s) × entropy(ผล = s)]

= 0.95 – [0.59 × 0.84 + 0.41 × 1]

= 0.95 – [0.50 + 0.41]

= 0.95 – 0.91

= 0.04

3) คำนวณค่า IG ของแอตทริบิวต์ รอยฟกช้ำ จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.63 × $\log_2$(0.63) + 0.37 × $\log_2$(0.37)]

= −[0.63 × −0.67 + 0.37 × −1.43]

= −[0.42 + 0.53]

= 0.95

entropy(ผล = f) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.55 × $\log_2$(0.55) + 0.45 × $\log_2$(0.45)]

= –[0.55 × –0.86 + 0.45 × –1.15]

= –[0.47 + 0.52]

= 0.99

entropy(ผล =t) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.71 × $\log_2$(0.73) + 0.29 × $\log_2$(0.27)]

= –[0.71 × –0.49 + 0.29 × –1.79]

= –[0.35 + 0.52]

= 0.87

IG(parent, child) = entropy (parent) – [p(ผล =f) × entropy(ผล = f) + p(ผล = t) × entropy(ผล = t)]

= 0.95 – [0.55 × 0.99 + 0.45 × 0.87]

= 0.95 – [0.54 + 0.39]

= 0.95 – 0.93

= 0.02

4) คำนวณค่า IG ของแอตทริบิวต์ ที่อาศัย จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.63 × $\log_2$(0.63) + 0.37 × $\log_2$(0.37)]

= –[0.63 × –0.67 + 0.37 × –1.43]

= –[0.42 + 0.53]

= 0.95

entropy(ผล = d) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.61 × $\log_2$(0.61) + 0.39 × $\log_2$(0.39)]

= –[0.61 × –0.71 + 0.39 × –1.36]

= –[0.43 + 0.53]

= 0.96

entropy(ผล =g) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.71 × $\log_2$(0.71) + 0.29 × $\log_2$(0.29)]

= –[0.71 × –0.49 + 0.36 × –1.79]

= –[0.35 + 0.64]

= 0.99

entropy(ผล =u) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.24 × $\log_2$(0.24) + 0.76 × $\log_2$(0.76)]

= –[0.24 × –2.06 + 0.76 × –0.40]

= –[0.49 + 0.30]

= 0.79

IG(parent, child) = entropy (parent) – [p(ผล =d) × entropy(ผล = d) + p(ผล = g) × entropy(ผล = g) + p(ผล = u) × entropy(ผล = u)]

= 0.95 – [0.47 × 0.96 + 0.45 × 0.99 + 0.08 × 0.79]

= 0.95 – [0.45 + 0.45 + 0.06]

= 0.95 – 0.96

= –0.01

จากการคำนวณค่า IG ของทุกแอตทริบิวต์พบว่าค่า IG ของแอตทริบิวต์ พื้นผิว มีค่ามากที่สุด (0.04) ดังนั้นจึงเลือกแอตทริบิวต์ พื้นผิว ขึ้นมาเป็นโหนด root และจะต้องทำการแตกกิ่งจาก โหนด root ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกัน และผู้วิเคราะห์จึงทำการสร้างโหนดในระดับถัดไป

การคำนวณค่าแต่ละแอตทริบิวต์ พื้นผิว ในฝั่งของ f “เส้นใย” ในระดับที่ 2 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.73 \times \log_2(0.73) + 0.27 \times \log_2(0.37)]$

$= -[0.73 \times -0.45 + 0.27 \times -1.89]$

$= -[0.33 + 0.51]$

$= 0.84$

entropy(ผล = b) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.92 \times \log_2(0.93) + 0.08 \times \log_2(0.07)]$

$= -[0.92 \times -0.12 + 0.08 \times -3.64]$

$= -[0.11 + 0.29]$

$= 0.40$

entropy(ผล =f) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.76 \times \log_2(0.76) + 0.24 \times \log_2(0.24)]$

$= -[0.76 \times -0.40 + 0.24 \times -2.06]$

$= -[0.30 + 0.49]$

$= 0.79$

entropy(ผล =x) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.70 \times \log_2(0.70) + 0.30 \times \log_2(0.30)]$

$= -[0.70 \times -0.51 + 0.30 \times -1.74]$

$= -[0.36 + 0.52]$

$= 0.88$

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f)+ p(ผล = x) × entropy(ผล = x)]

= 0.84 – [0.03 × 0.40 + 0.45 × 0.79 + 0.52 × 0.88]

= 0.84 – [0.01 + 0.36 + 0.46]

= 0.84 – 0.83

= 0.01

2) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรอยฟกช้ำ จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.73 × $\log_2$(0.73) + 0.27 × $\log_2$(0.37)]

= –[0.73 × –0.45 + 0.27 × –1.89]

= –[0.33 + 0.51]

= 0.84

entropy(ผล = f) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.50 × $\log_2$(0.50) + 0.50 × $\log_2$(0.50)]

= –[0.50 × –1 + 0.50 × –1]

= –[0.50 + 0.50]

= 1

entropy(ผล =t) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= –[1.00 × –0.00 + 0.00 × –1]

= –[0.00 + 0.00]

= 0

IG(parent, child) = entropy (parent) – [p(ผล =f) × entropy(ผล = f) + p(ผล = t) × entropy(ผล = t)]

= 0.84 – [0.54 × 1 + 0.46 × 0]

= 0.84 – [0.54 + 0]

= 0.84 – 0.54

= 0.30

3) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัย จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.73 × $\log_2$(0.73) + 0.27 × $\log_2$(0.37)]

= –[0.73 × –0.45 + 0.27 × –1.89]

= –[0.33 + 0.51]

= 0.84

entropy(ผล = d) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.74 × $\log_2$(0.74) + 0.26 × $\log_2$(0.39)]

= –[0.74 × –0.43 + 0.26 × –1.94]

= –[0.32 + 0.50]

= 0.82

entropy(ผล =g) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.69 × $\log_2$(0.69) + 0.31 × $\log_2$(0.31)]

= –[0.69 × –0.54 + 0.31 × –1.69]

= –[0.37 + 0.52]

= 0.89

entropy(ผล =u) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= –[1.00 × –0.00 + 0.00 × –1.00]

= –[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =d) × entropy(ผล = d) + p(ผล = g) × entropy(ผล = g) + p(ผล = u) × entropy(ผล = u)]

= 0.84 – [0.62 × 0.82 + 0.35 × 0.89 + 0.03 × 0.00]

= 0.84 – [0.51 + 0.31 + 0.00]

= 0.84 – 0.82

= 0.02

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรูปทรง รอยฟกช้ำและที่อาศัยพบว่าค่า IG ของแอตทริบิวต์ พื้นผิวและรอยฟกช้ำ มีค่ามากที่สุด (0.30) ดังนั้นจึงเลือกแอตทริบิวต์ รอยฟกช้ำ ขึ้นมาเป็นโหนดในระดับที่ 2 ต่อจากโหนด root และทำการแตกกิ่งจากโหนดในระดับที่ 2 ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันในฝั่งของ เส้นใย หรือ f

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 3 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรอยฟกช้ำและที่อาศัย จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.50 × $\log_2$(0.50) + 0.50 × $\log_2$(0.50)]

= –[0.50 × –1 + 0.50 × –1]

= –[0.50 + 0.50]

= 1

entropy(ผล = d) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.00 × $\log_2$(0.00) + 1.00 × $\log_2$(1.00)]

= –[0.00 × –1.00 + 1.00 × –0.00]

= –[0.00 + 0.00]

= 0.00

entropy(ผล =g) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.69 × $\log_2$(0.69) + 0.31 × $\log_2$(0.31)]

= –[0.69 × –0.54 + 0.31 × –1.69]

= –[0.37 + 0.52]

= 0.89

entropy(ผล =u) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= –[1.00 × –0.00 + 0.00 × –1.00]

= –[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =d) × entropy(ผล = d) + p(ผล = g) × entropy(ผล = g) + p(ผล = u) × entropy(ผล = u)]

= 1 – [0.30 × 0.00 + 0.64 × 0.89 + 0.06 × 0.00]

= 1 – [0.00 + 0.57 + 0.00]

= 1 – 0.57

= 0.43

2) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรอยฟกช้ำและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.73 × $\log_2$(0.73) + 0.27 × $\log_2$(0.37)]

= –[0.73 × –0.45 + 0.27 × –1.89]

= –[0.33 + 0.51]

= 0.84

entropy(ผล = b) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= −[1.00 × −0.00 + 0.00 × −1.00]

= −[0.00 + 0.00]

= 0.00

entropy(ผล =f) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.00 × $\log_2$(0.00) + 0.00 × $\log_2$(0.00)]

= −[0.00 × −1.00 + 0.00 × −1.00]

= −[0.00 + 0.00]

= 0.00

entropy(ผล =x) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.00 × $\log_2$(0.00) + 0.00 × $\log_2$(0.00)]

= −[0.00 × −1.00 + 0.00 × −1.00]

= −[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f)+ p(ผล = x) × entropy(ผล = x)]

= 0.84 – [1.00 × 0.00 + 0.00 × 0.00 + 0.00 × 0.00]

= 0.84 – [0.00 + 0.00 + 0.00]

= 0.84 – 0.00

= 0.84

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรอยฟกช้ำ “เส้นใย” และที่อาศัย พบว่าข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันแล้ว คือ ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้ และผู้วิเคราะห์ข้อมูลพบว่าข้อมูลในการวิเคราะห์ พื้นผิวและรอยฟกช้ำและรูปทรง นั้นผลลัพธ์ออกมาได้เป็น รับประทานได้ทั้งหมด

การคำนวณค่าแต่ละแอตทริบิวต์ พื้นผิว ในฝั่งของ s “เรียบ” ในระดับที่ 2 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.48 \times \log_2(0.48) + 0.52 \times \log_2(0.52)]$

$= -[0.48 \times -1.06 + 0.52 \times -0.94]$

$= -[0.51 + 0.49]$

$= 1.00$

entropy(ผล = b) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.93 \times \log_2(0.93) + 0.07 \times \log_2(0.07)]$

$= -[0.92 \times -0.10 + 0.07 \times -3.84]$

$= -[0.09 + 0.27]$

$= 0.36$

entropy(ผล = f) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.41 \times \log_2(0.41) + 0.59 \times \log_2(0.59)]$

$= -[0.41 \times -1.29 + 0.59 \times -0.76]$

$= -[0.53 + 0.45]$

$= 0.98$

entropy(ผล = x) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.45 \times \log_2(0.45) + 0.55 \times \log_2(0.55)]$

$= -[0.45 \times -1.15 + 0.55 \times -0.86]$

$= -[0.52 + 0.47]$

$= 0.99$

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f)+ p(ผล = x) × entropy(ผล = x)]

= 1.00 – [0.09 × 0.36 + 0.38 × 0.98 + 0.53 × 0.99]

= 1.00 – [0.03 + 0.37 + 0.52]

= 1.00 – 0.92

= 0.08

2) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรอยฟกช้ำ จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.48 × $\log_2$(0.48) + 0.52 × $\log_2$(0.52)]

= –[0.48 × –1.06 + 0.52 × –0.94]

= –[0.51 + 0.49]

= 1.00

entropy(ผล = f) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.63 × $\log_2$(0.63) + 0.37 × $\log_2$(0.37)]

= –[0.63 × –0.67 + 0.37 × –1.43]

= –[0.42 + 0.53]

= 0.95

entropy(ผล =t) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.29 × $\log_2$(0.29) + 0.71 × $\log_2$(0.71)]

= –[0.29 × –1.79 + 0.71 × –0.49]

= –[0.52 + 0.35]

= 0.87

IG(parent, child) = entropy (parent) – [p(ผล =f) × entropy(ผล = f) + p(ผล = t) × entropy(ผล = t)]

= 1.00 – [0.56 × 0.95 + 0.44 × 0.87]

= 1.00 – [0.53 + 0.38]

= 1.00 – 0.91

= 0.09

3) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัย จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.48 × $\log_2$(0.48) + 0.52 × $\log_2$(0.52)]

= –[0.48 × –1.06 + 0.52 × –0.94]

= –[0.51 + 0.49]

= 1.00

entropy(ผล = d) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.15 × $\log_2$(0.15) + 0.85 × $\log_2$(0.85)]

= –[0.15 × –2.74 + 0.85 × –0.23]

= –[0.41 + 0.20]

= 0.61

entropy(ผล =g) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.73 × $\log_2$(0.73) + 0.27 × $\log_2$(0.27)]

= –[0.73 × –0.45 + 0.27 × –1.89]

= –[0.33 + 0.51]

= 0.84

entropy(ผล =u) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.00 × $\log_2$(0.00) + 1.00 × $\log_2$(1.00)]

= −[0.00 × −1.00 + 1.00 × −0.00]

= −[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =d) × entropy(ผล = d) + p(ผล = g) × entropy(ผล = g) + p(ผล = u) × entropy(ผล = u)]

= 1.00 – [0.25 × 0.82 + 0.60 × 0.89 + 0.15 × 0.00]

= 1.00 – [0.21 + 0.53 + 0.00]

= 1.00 – 0.74

= 0.26

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและรูปทรง รอยฟกช้ำและที่อาศัยพบว่าค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัย มีค่ามากที่สุด (0.26) ดังนั้นจึงเลือกแอตทริบิวต์ ที่อาศัยขึ้นมาเป็นโหนดในระดับที่ 2 ต่อจากโหนด root และทำการแตกกิ่งจากโหนดในระดับที่ 2 ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันในฝั่งของ เรียบ หรือ s

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 3 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้ ของ กิ่ง d “ในป่า”

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.15 × $\log_2$(0.15) + 0.85 × $\log_2$(0.85)]

= −[0.15 × −2.74 + 0.85 × −0.23]

= −[0.41 + 0.20]

= 0.61

entropy(ผล = b) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

$= -[1.00 \times \log_2(1.00) + 0.00 \times \log_2(0.00)]$

$= -[1.00 \times -0.00 + 0.00 \times -1.00]$

$= -[0.00 + 0.00]$

$= 0.00$

entropy(ผล =f) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.21 \times \log_2(0.21) + 0.79 \times \log_2(0.79)]$

$= -[0.21 \times -2.25 + 0.79 \times -0.34]$

$= -[0.47 + 0.27]$

$= 0.74$

entropy(ผล =x) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.12 \times \log_2(0.12) + 0.88 \times \log_2(0.88)]$

$= -[0.12 \times -3.06 + 0.88 \times -0.18]$

$= -[0.37 + 0.16]$

$= 0.53$

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f) + p(ผล = x) × entropy(ผล = x)]

$= 0.61 - [0.00 \times 0.00 + 0.74 \times 0.89 + 0.64 \times 0.53]$

$= 0.61 - [0.00 + 0.66 + 0.34]$

$= 0.61 - 1.00$

$= -0.39$

2) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำ จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.15 \times \log_2(0.15) + 0.85 \times \log_2(0.85)]$

$= -[0.15 \times -2.74 + 0.85 \times -0.23]$

= –[0.41 + 0.20]

= 0.61

entropy(ผล = f) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.01 × $\log_2$(0.63) + 0.99 × $\log_2$(0.99)]

= –[0.01 × –6.64 + 0.99 × –0.01]

= –[0.07 + 0.01]

= 0.08

entropy(ผล =t) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= –[1.00 × –0.00 + 0.00 × –1.00]

= –[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =f) × entropy(ผล = f) + p(ผล = t) × entropy(ผล = t)]

= 0.61 – [0.86 × 0.08 + 0.14 × 0.00]

= 0.61 – [0.07 + 0.00]

= 0.61 – 0.07

= 0.54

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรูปทรงและรอยฟกช้ำในระดับที่2 พบว่าค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำ มีค่ามากที่สุด (0.54) ดังนั้นจึงเลือกแอตทริบิวต์ รอยฟกช้ำ ขึ้นมาเป็นโหนดในระดับที่ 3 ต่อจากโหนด root ของ กิ่ง d และทำการแตกกิ่งจากโหนดในระดับที่ 3 ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันในฝั่งของ เรียบ หรือ s

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 3 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้ ของ กิ่ง g “ในหญ้า”

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.73 \times \log_2(0.73) + 0.27 \times \log_2(0.27)]$

$= -[0.73 \times -0.45 + 0.27 \times -1.89]$

$= -[0.33 + 0.51]$

$= 0.84$

entropy(ผล = b) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.93 \times \log_2(0.93) + 0.07 \times \log_2(0.00)]$

$= -[0.93 \times -0.10 + 0.07 \times -3.84]$

$= -[0.09 + 0.27]$

$= 0.36$

entropy(ผล =f) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.63 \times \log_2(0.63) + 0.37 \times \log_2(0.37)]$

$= -[0.63 \times -0.67 + 0.37 \times -1.43]$

$= -[0.42 + 0.53]$

$= 0.95$

entropy(ผล =x) = $-p(\text{รับประทานได้}) \times \log_2 p(\text{รับประทานได้}) + p(\text{รับประทานไม่ได้}) \times \log_2 p(\text{รับประทานไม่ได้})$

$= -[0.75 \times \log_2(0.75) + 0.25 \times \log_2(0.25)]$

$= -[0.75 \times -0.42 + 0.25 \times -2.00]$

$= -[0.32 + 0.50]$

$= 0.82$

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f) + p(ผล = x) × entropy(ผล = x)]

= 0.84 – [0.15 × 0.36 + 0.37 × 0.95 + 0.49 × 0.82]

= 0.84 – [0.05 + 0.35 + 0.40]

= 0.84 – 0.80

= 0.04

2) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำ จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.73 × $\log_2$(0.73) + 0.27 × $\log_2$(0.27)]

= –[0.73 × –0.45 + 0.27 × –1.89]

= –[0.33 + 0.51]

= 0.84

entropy(ผล = f) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= –[1.00 × –0.00 + 0.00 × –1.00]

= –[0.00 + 0.00]

= 0.00

entropy(ผล =t) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[0.36 × $\log_2$(0.36) + 0.64 × $\log_2$(0.64)]

= –[0.36 × –1.47 + 0.64 × –0.64]

= –[0.53 + 0.41]

= 0.94

IG(parent, child) = entropy (parent) – [p(ผล =f) × entropy(ผล = f) + p(ผล = t) × entropy(ผล = t)]

= 0.84 – [0.58 × 0.00 + 0.42 × 0.94]

= 0.84 – [0.00 + 0.00]

= 0.84 – 0.39

= 0.45

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรูปทรงและรอยฟกช้ำในระดับที่2 พบว่าค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำ มีค่ามากที่สุด (0.45) ดังนั้นจึงเลือกแอตทริบิวต์ รอยฟกช้ำ ขึ้นมาเป็นโหนดในระดับที่ 3 ต่อจากโหนด root ของ กิ่ง g และทำการแตกกิ่งจากโหนดในระดับที่ 3 ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันในฝั่งของ เรียบ หรือ s

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 4 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้ ของ กิ่ง f “ไม่มีรอยฟกช้ำ”

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p$(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= $-[1.00 \times \log_2(1.00) + 0.00 \times \log_2(0.00)]$

= $-[0.00 \times -0.00 + 0.00 \times -1.00]$

= $-[0.00 + 0.00]$

= 0.00

entropy(ผล = b) = $-p$(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= $-[1.00 \times \log_2(1.00) + 0.00 \times \log_2(0.00)]$

= $-[1.00 \times -0.00 + 0.00 \times -1.00]$

= $-[0.00 + 0.00]$

= 0.00

entropy(ผล =f) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= –[1.00 × –0.00 + 0.00 × –1.00]

= –[0.00 + 0.00]

= 0.00

entropy(ผล =x) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= –[1.00 × $\log_2$(1.00) + 0.00 × $\log_2$(0.00)]

= –[1.00 × –0.00 + 0.00 × –1.00]

= –[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f) + p(ผล = x) × entropy(ผล = x)]

= 0.00 – [0.10 × 0.00 + 0.40 × 0.00 + 0.50 × 0.00]

= 0.00 – [0.00 + 0.00 + 0.00]

= 0.00 – 0.00

= 0.00

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำและรูปทรง ฝั่ง f “ไม่มีรอยฟกช้ำ” พบว่าข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันแล้ว คือ ผลลัพธ์เป็น รับประทานได้ และรับประทานไม่ได้ และผู้วิเคราะห์ข้อมูลพบว่าข้อมูลในการวิเคราะห์ พื้นผิวและที่อาศัยและรอยฟกช้ำและรูปทรง นั้น ผลลัพธ์ออกมาได้เป็น รับประทานได้ทั้งหมด

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 4 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้ ของ กิ่ง t “มีรอยฟกช้ำ”

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

$= -[0.36 \times \log_2(0.36) + 0.64 \times \log_2(0.64)]$

$= -[0.36 \times -1.47 + 0.64 \times -0.64]$

$= -[0.53 + 0.41]$

$= 0.94$

entropy(ผล = b) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.88 \times \log_2(0.88) + 0.12 \times \log_2(0.12)]$

$= -[0.88 \times -0.18 + 0.12 \times -3.06]$

$= -[0.16 + 0.37]$

$= 0.53$

entropy(ผล =f) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.00 \times \log_2(0.00) + 1.00 \times \log_2(1.00)]$

$= -[0.00 \times -1.00 + 1.00 \times -0.00]$

$= -[0.00 + 0.00]$

$= 0.00$

entropy(ผล =x) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.38 \times \log_2(0.38) + 0.62 \times \log_2(0.62)]$

$= -[0.38 \times -1.40 + 0.62 \times -0.69]$

$= -[0.53 + 0.43]$

$= 0.96$

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f) + p(ผล = x) × entropy(ผล = x)]

$= 0.94 - [0.21 \times 0.53 + 0.32 \times 0.00 + 0.47 \times 0.96]$

$= 0.94 - [0.11 + 0.00 + 0.45]$

$= 0.94 - 0.56$

$= 0.38$

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรูปทรงและรอยฟกช้ำในระดับที่3 พบว่าค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำและรูปทรง มีค่ามากที่สุด (0.38) ดังนั้นจึงเลือกแอตทริบิวต์ รูปทรง ขึ้นมาเป็นโหนดในระดับที่ 4 ต่อจากโหนด root ของ กิ่ง t และ ผู้วิเคราะห์ข้อมูลพบว่าข้อมูลในการวิเคราะห์ พื้นผิวและที่อาศัยและรอยฟกช้ำและรูปทรง นั้นผลลัพธ์ออกมาได้เป็น รับประทานได้และรับประทานไม่ได้

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 3 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นรับประทานได้ และรับประทานไม่ได้ ของ กิ่ง u “ในเมือง”

1) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรูปทรง จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.00 \times \log_2(0.00) + 1.00 \times \log_2(1.00)]$

$= -[0.00 \times -1.00 + 1.00 \times -0.00]$

$= -[0.00 + 0.00]$

$= 0.00$

entropy(ผล = b) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.00 \times \log_2(0.00) + 0.00 \times \log_2(0.00)]$

$= -[0.00 \times -1.00 + 0.00 \times -1.00]$

$= -[0.00 + 0.00]$

$= 0.00$

entropy(ผล =f) = $-p$(รับประทานได้) $\times \log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) $\times \log_2$ p(รับประทานไม่ได้)

$= -[0.00 \times \log_2(0.00) + 1.00 \times \log_2(1.00)]$

$= -[0.00 \times -1.00 + 1.00 \times -0.00]$

$= -[0.00 + 0.00]$

$= 0.00$

entropy(ผล =x) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.00 × $\log_2$(0.00) + 1.00 × $\log_2$(1.00)]

= −[0.00 × −1.00 + 1.00 × −0.00]

= −[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =b) × entropy(ผล = b) + p(ผล = f) × entropy(ผล = f) + p(ผล = x) × entropy(ผล = x)]

= 0.00 – [0.00 × 0.00 + 0.50 × 0.00 + 0.50 × 0.00]

= 0.00 – [0.00 + 0.00 + 0.00]

= 0.00 – 0.00

= 0.00

2) คำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรอยฟกช้ำ จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.00 × $\log_2$(0.00) + 1.00 × $\log_2$(1.00)]

= −[0.00 × −1.00 + 1.00 × −0.00]

= −[0.00 + 0.00]

= 0.00

entropy(ผล = f) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.00 × $\log_2$(0.00) + 0.00 × $\log_2$(0.00)]

= −[0.00 × −1.00 + 0.00 × −1.00]

= −[0.00 + 0.00]

= 0.00

entropy(ผล =t) = −p(รับประทานได้) × $\log_2$ p(รับประทานได้) + p(รับประทานไม่ได้) × $\log_2$ p(รับประทานไม่ได้)

= −[0.00 × $\log_2$(0.00) + 1.00 × $\log_2$(1.00)]

= −[0.00 × −1.00 + 1.00 × −0.00]

= −[0.00 + 0.00]

= 0.00

IG(parent, child) = entropy (parent) – [p(ผล =f) × entropy(ผล = f) + p(ผล = t) × entropy(ผล = t)]

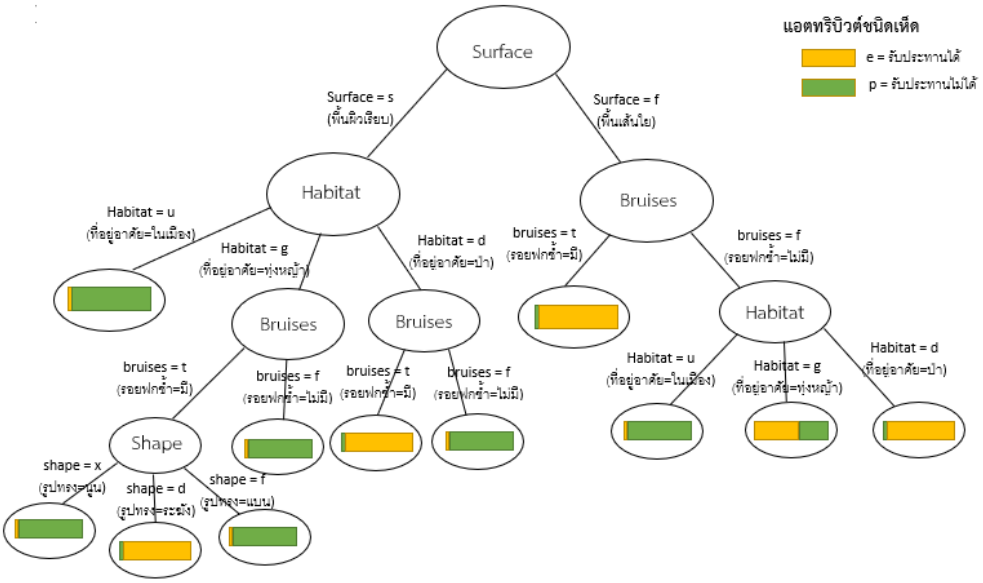
= 0.00 – [0.00 × 0.00 + 0.01 × 0.00]

= 0.00 – [0.00 + 0.00]

= 0.00 – 0.00

= 0.00

จากการคำนวณค่า IG ของแอตทริบิวต์ พื้นผิวและที่อาศัยและรูปทรงและรอยฟกช้ำในระดับที่3 พบว่าค่า IG ของแอตทริบิวต์ พื้นผิวและที่อยู่อาศัยไปรูปทรงกับรอยฟกช้ำ มีค่า IG (0.00) ซึ่งไม่สามารถแตกกิ่งได้แล้ว ดังนั้นจึงสรุปข้อมูลได้ว่า รับประทานไม่ได้ ทั้งหมด ของ กิ่ง u “ในเมือง” ตามดังรูป

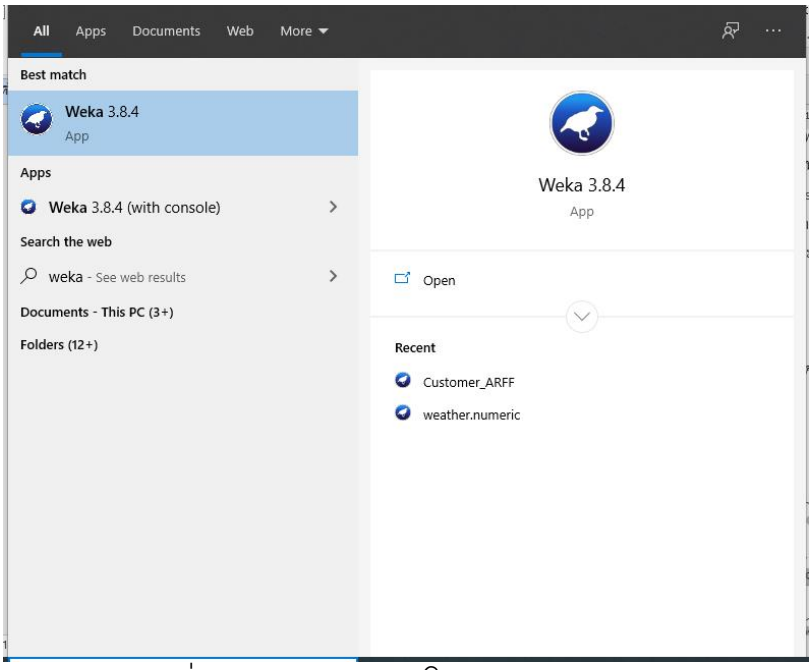


ภาพที่ 3. 6 แสดงรูปแบบโมเดล Decision Tree จากการคำนวณด้วยมือ

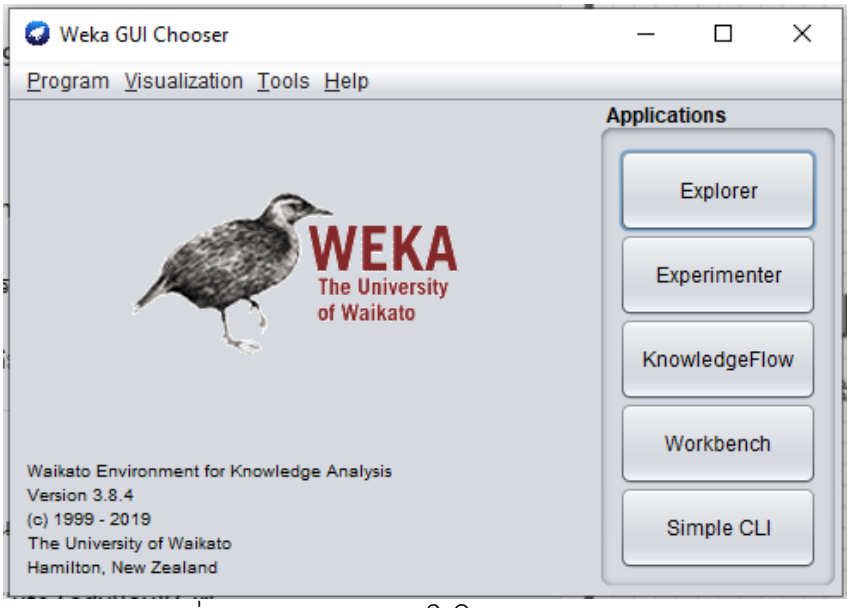
3.1.5 การประเมินผล (Evaluation) เป็นขั้นตอนก่อนนำผลลัพธ์ที่ได้จากขั้นตอนที่ 4 ไปใช้งาน ด้วยการวัดประสิทธิผลของผลลัพธ์ที่ได้กับวัตถุประสงค์ที่ตั้งไว้ในขั้นตอนแรก ว่ามีนัยสำคัญหรือความน่าเชื่อถือมากน้อยเพียงใด ด้วยการประเมินผลจากโปรแกรมว่าถูกต้องหรือไม่

ผู้วิเคราะห์ได้ทำการทดลองโมเดล เพื่อวัดประสิทธิภาพที่ตรงกับความต้องการ ซึ่งการวัดประสิทธิภาพด้วยวิธี Self-Consistency Test เหมาะสำหรับใช้ในการทดสอบประสิทธิภาพ เพื่อดูแนวโน้มของโมเดลที่สร้างขึ้น และเมื่อนำข้อมูลมาทดสอบ (Testing data) กับโปรแกรมที่ผู้วิเคราะห์เลือกมาทดสอบกับข้อมูลที่ผ่านการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining จากการสร้างโมเดล Decision Tree จึงนำข้อมูลดังกล่าว มาทดสอบกับโปรแกรม Weka เวอร์ชัน 3.8.4 ซึ่งมีขั้นตอนการทำงาน ดังนี้

ขั้นตอนที่ 1 เปิดโปรแกรม Weka 3.8.4

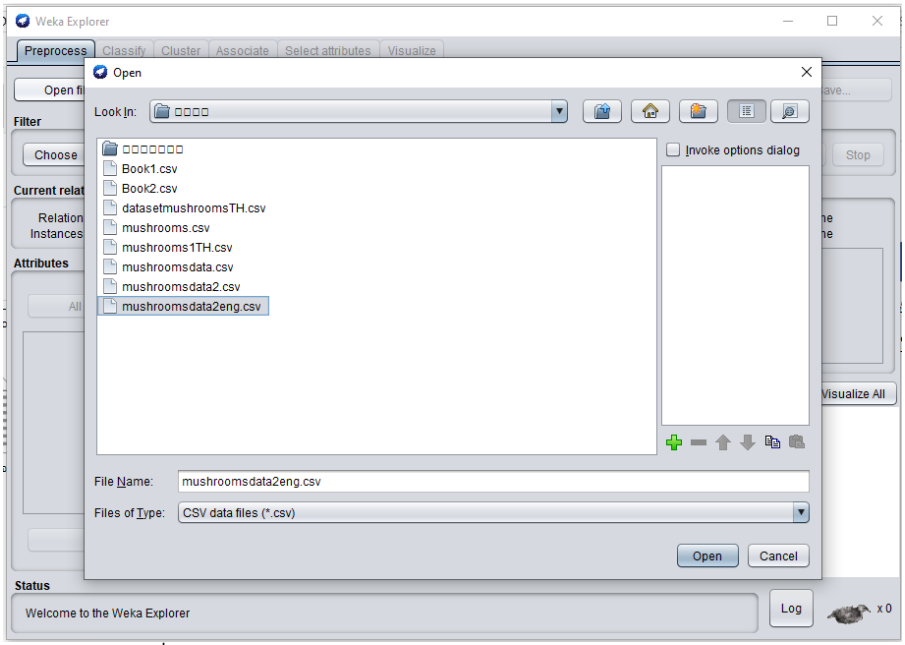


ภาพที่ 3. 7 แสดงการเปิดโปรแกรม Weka 3.8.4

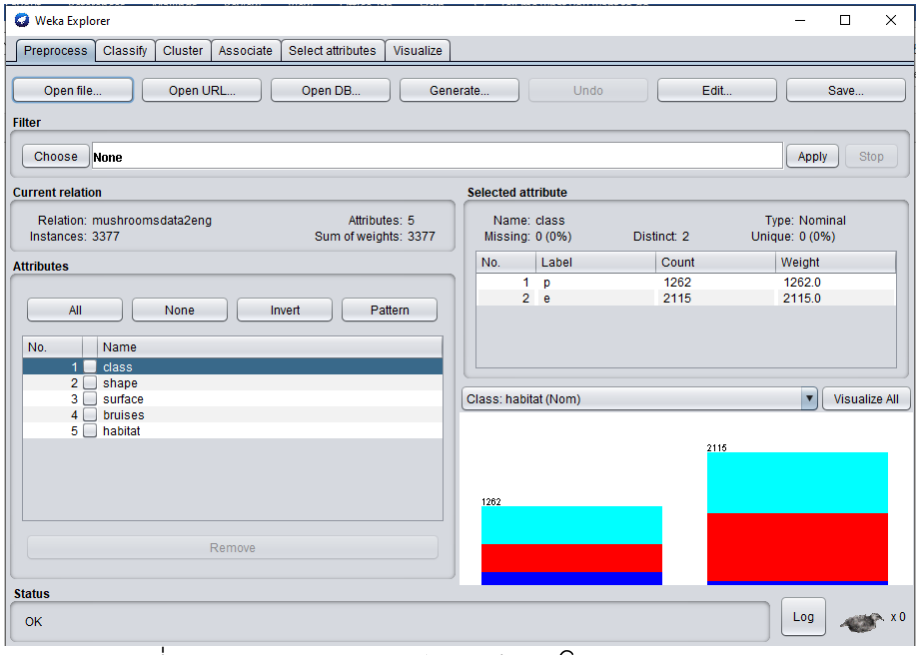


ภาพที่ 3. 8 แสดงการเข้าใช้โปรแกรม Weka 3.8.4

ขั้นตอนที่ 2 นำเข้าข้อมูลที่ได้จัดเตรียมไว้ โดยเลือกที่ Application>>Explorer>>Open file เลือกไฟล์ข้อมูลที่ต้องการนำมาทดสอบตามภาพที่ 3.7 และหลังจาก นั้นโปรแกรมแสดงหน้าจอข้อมูล ตามภาพที่ 3.8

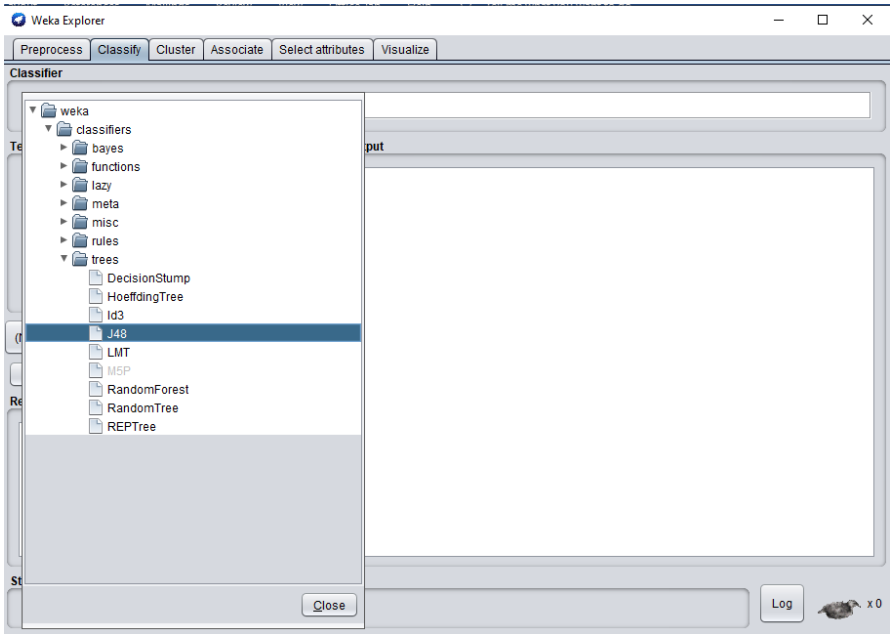


ภาพที่ 3. 9 แสดงการนำไฟล์ข้อมูลเข้าโปรแกรม Weka 3.8.4

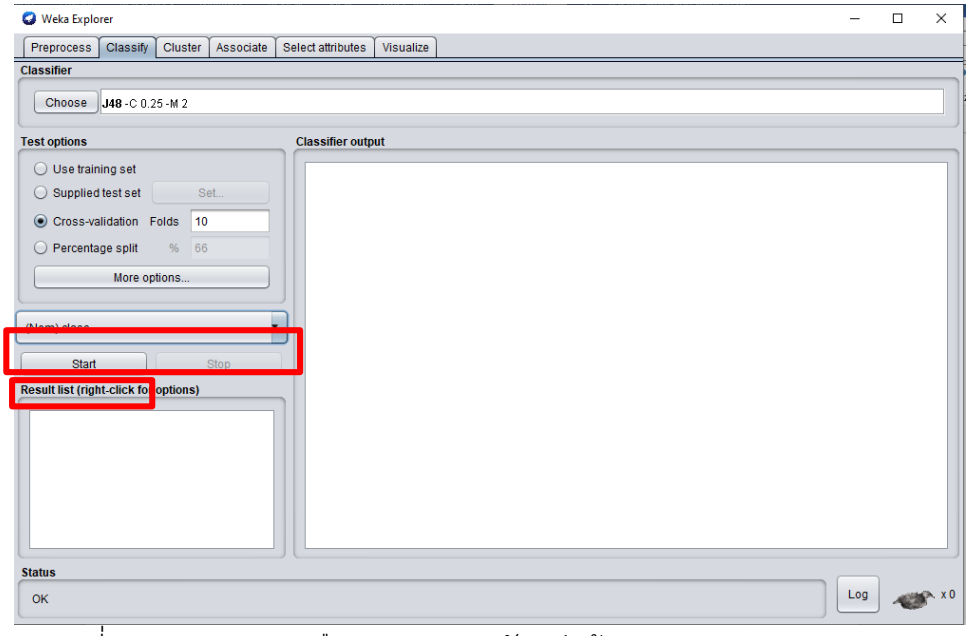


ภาพที่ 3. 10 แสดงข้อมูลหลังจากนำเข้าโปรแกรม Weka 3.8.4

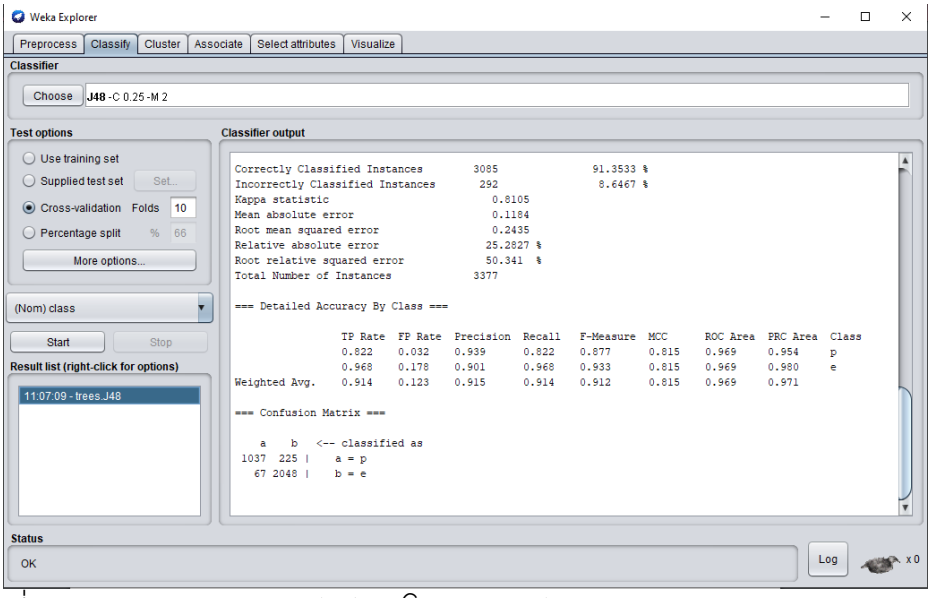
ขั้นตอนที่ 3 ดำเนินการเลือกเทคนิคการจัดกลุ่มข้อมูลแบบ Decision Tree โดย เลือกที่ Classification>>Choose>>tree และเลือกรูปแบบเป็น J48 ตามภาพที่ 3.9 จากนั้นทำการเลือก entropy (parent) แล้วกด ปุ่ม Start ตามภาพที่ 3.10 จะแสดงผลลัพธ์ที่ได้ตามภาพที่ 3.11



ภาพที่ 3. 11 แสดงการเลือกเทคนิคการจัดกลุ่มข้อมูลแบบ Decision Tree



ภาพที่ 3. 12 แสดงการเลือกเทคนิคการจัดกลุ่มข้อมูลแบบ Decision Tree: J48

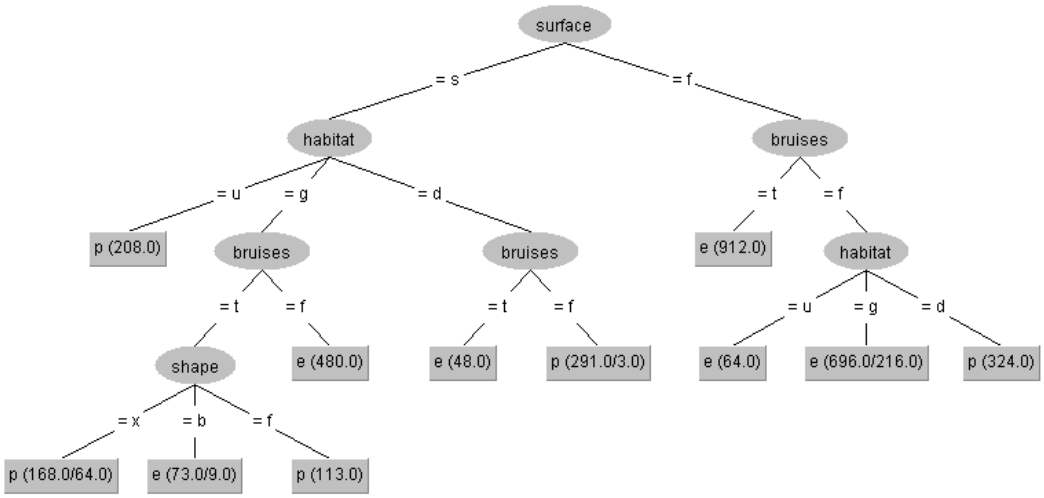


ภาพที่ 3. 13 แสดงหน้าจอผลลัพธ์ของโมเดลการจัดกลุ่มข้อมูลแบบ Decision Tree: J48

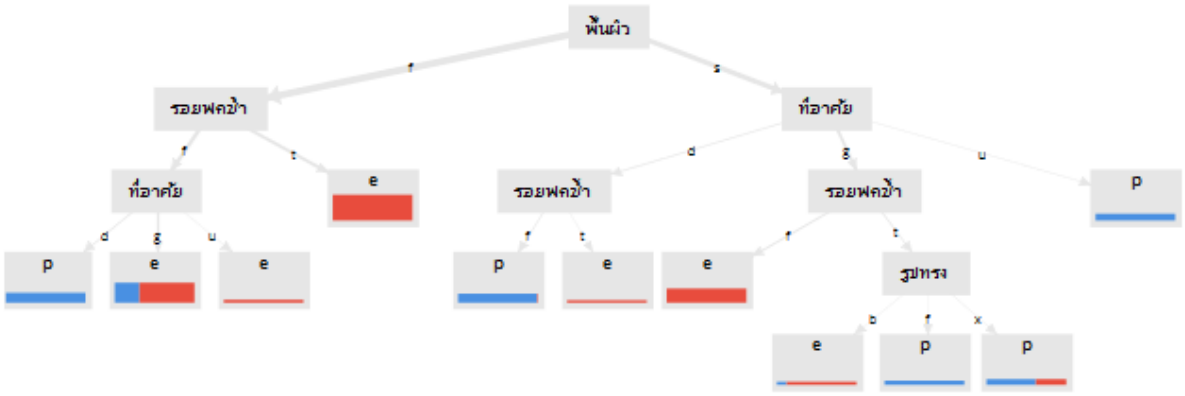
จากผลลัพธ์การทดลองพบว่าเทคนิค Decision Tree: J48 ให้ผลลัพธ์การจำแนกประเภทเห็ด รับประทานได้(e) และรับประทานไม่ได้(p) มีความถูกต้องถึง 91.35%

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

surface = s
|   habitat = u: p (208.0)
|   habitat = g
|   |   bruises = t
|   |   |   shape = x: p (168.0/64.0)
|   |   |   shape = b: e (73.0/9.0)
|   |   |   shape = f: p (113.0)
|   |   bruises = f: e (480.0)
|   habitat = d
|   |   bruises = t: e (48.0)
|   |   bruises = f: p (291.0/3.0)
surface = f
|   bruises = t: e (912.0)
|   bruises = f
|   |   habitat = u: e (64.0)
|   |   habitat = g: e (696.0/216.0)
|   |   habitat = d: p (324.0)

Number of Leaves  :     11

Size of the tree :      18
```

ภาพที่ 3. 14 แสดงผลลัพธ์กฎต้นไม้การตัดสินใจที่แตกกิ่งออกมา



ภาพที่ 3. 15 แสดงรูปแบบโมเดล Graph Decision Tree ในโปรแกรม Weka 3.8.4



ภาพที่ 3. 16 แสดงรูปแบบโมเดล Graph Decision Tree ในโปรแกรม RapidMiner

Tree

```
พื้นผิว = f
|   รอยฟกช้ำ = f
|   |   ที่อาศัย = d: p {p=324, e=0}
|   |   ที่อาศัย = g: e {p=216, e=480}
|   |   ที่อาศัย = u: e {p=0, e=64}
|   รอยฟกช้ำ = t: e {p=0, e=912}
พื้นผิว = s
|   ที่อาศัย = d
|   |   รอยฟกช้ำ = f: p {p=288, e=3}
|   |   รอยฟกช้ำ = t: e {p=0, e=48}
|   ที่อาศัย = g
|   |   รอยฟกช้ำ = f: e {p=0, e=480}
|   |   รอยฟกช้ำ = t
|   |   |   รูปทรง = b: e {p=9, e=64}
|   |   |   รูปทรง = f: p {p=113, e=0}
|   |   |   รูปทรง = x: p {p=104, e=64}
|   ที่อาศัย = u: p {p=208, e=0}
```

ภาพที่ 3. 17 คำบรรยายลักษณะงาน Decision Tree ในโปรแกรม RapidMiner

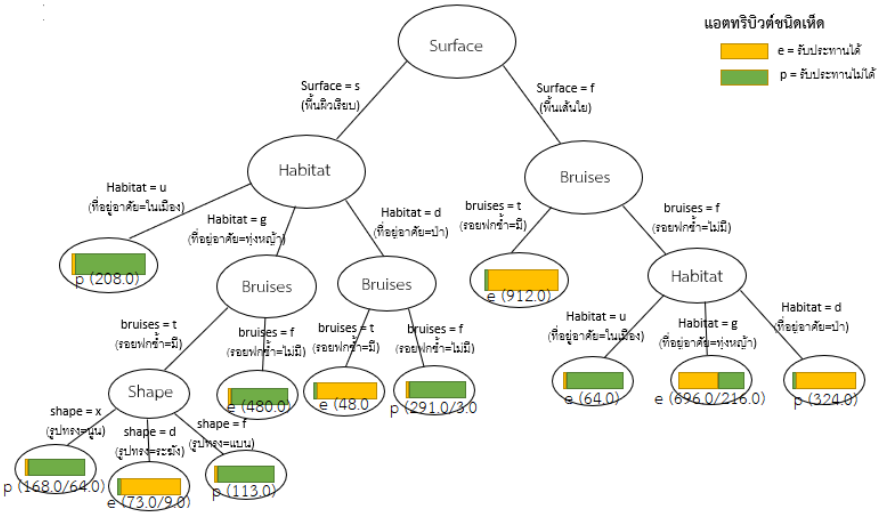
หลังจากผู้วิเคราะห์เลือกการทดสอบประสิทธิภาพของ Model ด้วยวิธี Self Consistency Test หรือเรียกว่า Use Training Set เป็นวิธีการที่นำข้อมูลที่ใช้ในการสร้างโมเดล (model) และข้อมูลที่ใช้ในการทดสอบโมเดลเป็นข้อมูลชุดเดียวกัน คือข้อมูลการจำแนกเห็ดพิษหรือเห็ดที่รับประทานได้ในอเมริกาเหนือ ที่ได้ทำการคัดเลือกมาทั้งหมด ซึ่งผู้วิเคราะห์เลือก

โปรแกรมที่ใช้นำเสนอ คือ โปรแกรม Weka 3.8.4 พบว่าการทดสอบประสิทธิภาพโมเดล Decision Tree (J48) พิจารณาได้ว่า โมเดลที่ถูกสร้างขึ้น มีค่าความถูกต้องเฉลี่ยในทุกโมเดล เท่ากับ 91.35% มีค่าการทำนายข้อมูลไม่ถูกต้องเท่ากับ 8.65% และมีค่าความคลาดเคลื่อน เท่ากับ 0.2435 และเมื่อพิจารณาส่วนค่า Confusion Matrix ในภาพที่ 3.17 พบว่าการหาค่า ของข้อมูลค่าจริง กับจำนวนข้อมูลจากการทำนาย แบ่งตามประเภทของผลการจำแนกเห็ด รับประทานได้(e) และรับประทานไม่ได้(p) และนำมาหาค่าเฉลี่ยรวมของทุก class ได้ค่าเฉลี่ยรวม เท่ากับ 0.895 มีผลลัพธ์ตรงกันอยู่ในระดับค่อนข้างดี สามารถนำโมเดลไปใช้งานได้

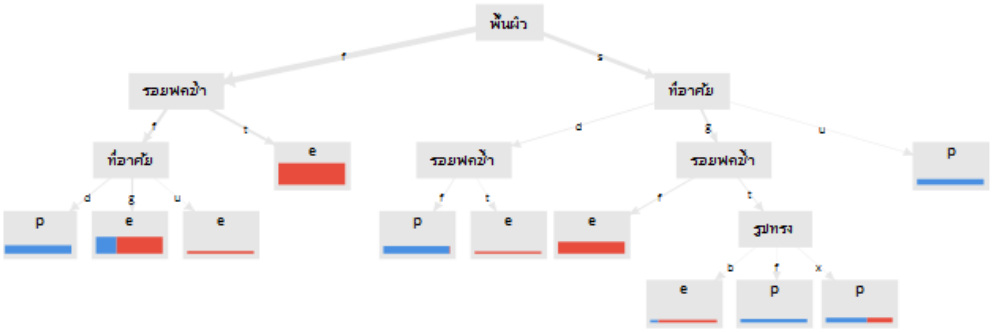
ดังนั้น ผู้วิเคราะห์จึงเลือกใช้โมเดล Decision Tree: J48 ในการนำไปหาแนวทางการวางแผนการจัดการและรับมือต่อปัจจัยเสี่ยงต่าง ๆ ของผู้ที่ประสบปัญหาด้านการจำแนกเห็ด และยังเป็นปัจจัยชี้นำประสิทธิภาพการทำงานขององค์กรการเกษตรและภาครัฐต่าง ๆ ได้อีกด้วย เพราะมีค่าความถูกต้องของโมเดล และค่าเฉลี่ย Confusion Matrix จากการจำแนกเห็ด รับประทานได้ (e) และรับประทานไม่ได้ (p) อยู่ในระดับที่ค่อนข้างดี

```
=== Summary ===

Correctly Classified Instances        3085               91.3533 %
Incorrectly Classified Instances       292                8.6467 %
Kappa statistic                          0.8105
Mean absolute error                      0.1184
Root mean squared error                  0.2435
Relative absolute error                 25.2827 %
Root relative squared error             50.341  %
Total Number of Instances             3377

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.822    0.032    0.939      0.822    0.877      0.815    0.969     0.954     p
                 0.968    0.178    0.901      0.968    0.933      0.815    0.969     0.980     e
Weighted Avg.    0.914    0.123    0.915      0.914    0.912      0.815    0.969     0.971

=== Confusion Matrix ===

    a    b   <-- classified as
 1037  225 |    a = p
   67 2048 |    b = e
```

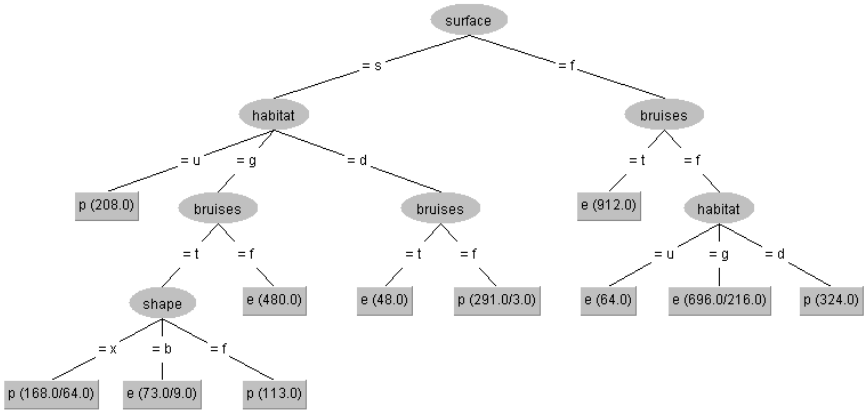
ภาพที่ 3. 18 แสดงผลลัพธ์จากการจำแนกกลุ่มแบบ Decision Tree: J48 ในโปรแกรม Weka 3.8.4



ภาพที่ 3. 19 แสดงรูปแบบโมเดล Decision Tree จากการคำนวณด้วยมือ



ภาพที่ 3. 20 แสดงรูปแบบโมเดล Graph Decision Tree ในโปรแกรม RapidMiner



ภาพที่ 3. 21 แสดงรูปแบบโมเดล Graph Decision Tree ในโปรแกรม Weka 3.8.4

ดังนั้นจะเห็นได้ว่าทางผู้วิเคราะห์ได้นำโมเดลของข้อมูลที่เลือกใช้ มาเปรียบเทียบกับโมเดลที่สร้างด้วย โปรแกรม RapidMiner Studio 9.5.1 ซึ่งผู้วิเคราะห์พบว่า ได้ผลลัพธ์จากการจำแนกกลุ่ม และรูปแบบแผนภาพโมเดลที่ตรงกันสูง ผู้วิเคราะห์จึงใช้เทคนิคของการจำแนกกลุ่มแบบ Decision Tree: J48 มาใช้ในการศึกษา เนื่องจากให้ผลลัพธ์ของกฎที่สามารถทำนายได้จำนวน 11 กฎ ซึ่งสามารถนำไปใช้ในการแบ่งกลุ่มได้ตามเงื่อนไขได้ชัดเจน และสามารถนำกฎที่ได้ไปวิเคราะห์กฎต่อไปได้ โดยสามารถจำแนกกฎได้ ดังนี้

กฎข้อที่ 1 IF surface=s Then habitat=u หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเรียบและอาศัยอยู่ในเมือง ผลการพิจารณาพบว่า เห็ดชนิดนี้รับประทานไม่ได้

กฎข้อที่ 2 IF surface=s Then habitat=g Then bruises=t Then shape=x หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเรียบและอาศัยอยู่ในทุ่งหญ้าและมีรอยฟกช้ำและมีรูปทรงนูน ผลการพิจารณาพบว่า เห็ดชนิดนี้รับประทานไม่ได้

กฎข้อที่ 3 IF surface=s Then habitat=g Then bruises=t Then shape=b หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเรียบและอาศัยอยู่ในทุ่งหญ้าและมีรอยฟกช้ำและมีรูปทรงระฆัง ผลการพิจารณาพบว่า เห็ดชนิดนี้รับประทานได้

กฎข้อที่ 4 IF surface=s Then habitat=g Then bruises=t Then shape=f หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเรียบและอาศัยอยู่ในทุ่งหญ้าและมีรอยฟกช้ำและมีรูปทรงแบน ผลการพิจารณาพบว่า เห็ดชนิดนี้รับประทานไม่ได้

กฎข้อที่ 5 IF surface=s Then habitat=g Then bruises=f หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเรียบและอาศัยอยู่ในทุ่งหญ้าและไม่มีรอยฟกช้ำ ผลการพิจารณาพบว่า เห็ดชนิดนี้รับประทานได้

กฎข้อที่ 6 IF surface=s Then habitat=d Then bruises=t หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเรียบและอาศัยอยู่ในป่าและมีรอยฟกช้ำ ผลการพิจารณาพบว่าเห็ดชนิดนี้รับประทานได้

กฎข้อที่ 7 IF surface=s Then habitat=d Then bruises=f หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเรียบและอาศัยอยู่ในป่าและไม่มีรอยฟกช้ำ ผลการพิจารณาพบว่าเห็ดชนิดนี้รับประทานไม่ได้

กฎข้อที่ 8 IF surface=f Then bruises=t หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเส้นใยและมีรอยฟกช้ำ ผลการพิจารณาพบว่าเห็ดชนิดนี้รับประทานได้

กฎข้อที่ 9 IF surface=f Then bruises=f Then habitat=u หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเส้นใยและไม่มีรอยฟกช้ำและอาศัยอยู่ในเมือง ผลการพิจารณาพบว่าเห็ดชนิดนี้รับประทานได้

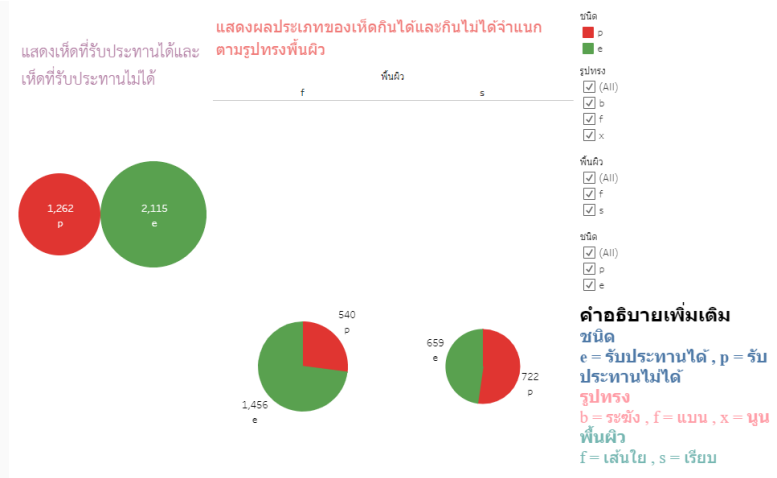
กฎข้อที่ 10 IF surface=f Then bruises=f Then habitat=g หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเส้นใยและไม่มีรอยฟกช้ำและอาศัยอยู่ในทุ่งหญ้า ผลการพิจารณาพบว่าเห็ดชนิดนี้รับประทานได้

กฎข้อที่ 11 IF surface=f Then bruises=f Then habitat=d หมายความว่า ถ้าพื้นผิวเป็นประเภทผิวเส้นใยและไม่มีรอยฟกช้ำและอาศัยอยู่ในป่า ผลการพิจารณาพบว่าเห็ดชนิดนี้รับประทานไม่ได้

3.1.6 เผยแพร่ผลวิเคราะห์ (Deployment) ขั้นตอนการนำผลลัพธ์ที่ได้ไปใช้งานเป็นการทั่วไป อาจจัดทำเป็นรูปแบบของรายงาน (Report) หรือแผนภาพ (Dashboard) ที่พร้อมให้ฝ่ายต่าง ๆ นำไปใช้ประโยชน์ในการวางแผน กำหนดกลยุทธ์ และดำเนินการต่าง ๆ ในทางธุรกิจ

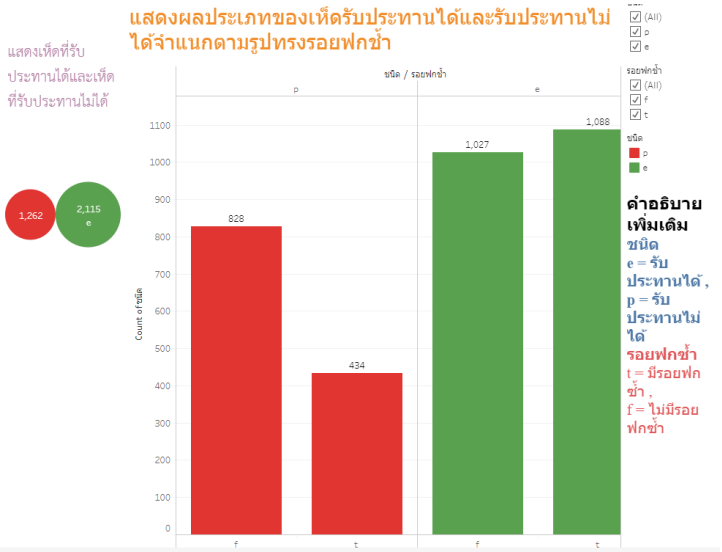
ผู้วิเคราะห์ข้อมูลนำผลข้อมูลที่ทำการวิเคราะห์แล้ว แสดงผลข้อมูลบน Web browser โดยใช้ชุดคำสั่ง HTML และ CSS3 ร่วมกับการนำเสนอข้อมูลแบบ visualization ด้วยการแสดงผลข้อมูลในรูปแบบของภาพโดยใช้โปรแกรม Tableau Public

1) แสดงข้อมูลแสดงผลการวิเคราะห์ประเภทของเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามพื้นผิว ซึ่งเป็นการแสดงผลในรูปแบบแผนภูมิวงกลม ผู้ใช้สามารถเลือกดูการจำแนกประเภทของเห็ดได้



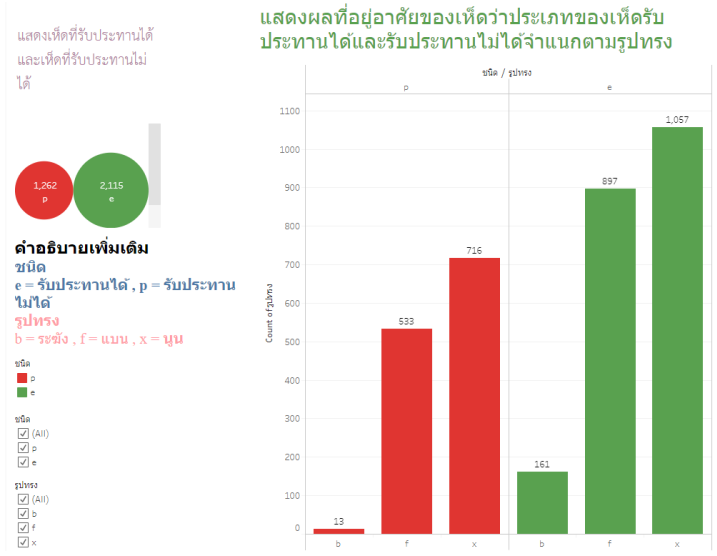
ภาพที่ 3. 22 แสดงข้อมูลแสดงผลการวิเคราะห์ประเภทของเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามพื้นผิวของเห็ด

2) แสดงข้อมูลประเภทของเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามรอยฟกช้ำ ซึ่งเป็นการแสดงผลในรูปแบบแผนภูมิแท่ง ผู้ใช้สามารถเลือกดูการจำแนกประเภทของเห็ดได้



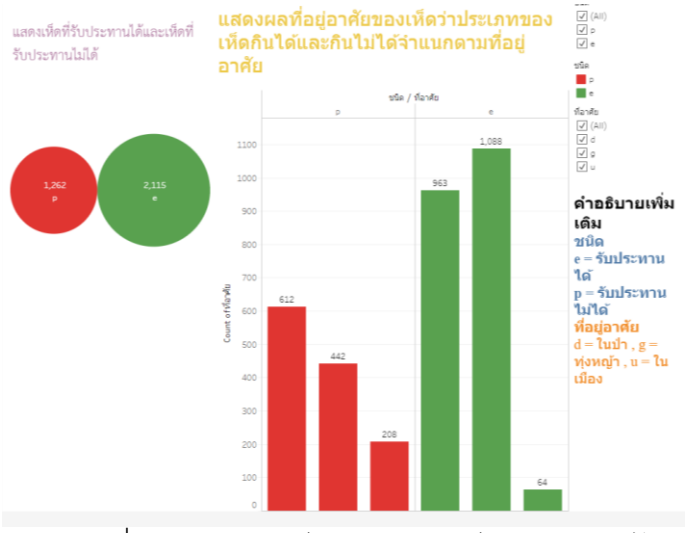
ภาพที่ 3. 23 แสดงผลประเภทของเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามรอยฟกช้ำ

3) แสดงข้อมูลประเภทของเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามรูปทรง ซึ่งเป็นการแสดงผลในรูปแบบแผนภูมิแท่ง ผู้ใช้สามารถเลือกดูการจำแนกประเภทของเห็ดได้



ภาพที่ 3. 24 แสดงข้อมูลประเภทของเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามรูปทรง

4) แสดงผลข้อมูลที่อยู่อาศัยของเห็ดว่าประเภทเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามที่อยู่อาศัย ซึ่งเป็นการแสดงผลในรูปแบบแผนภูมิแท่ง ผู้ใช้สามารถเลือกดูการจำแนกประเภทของเห็ดได้

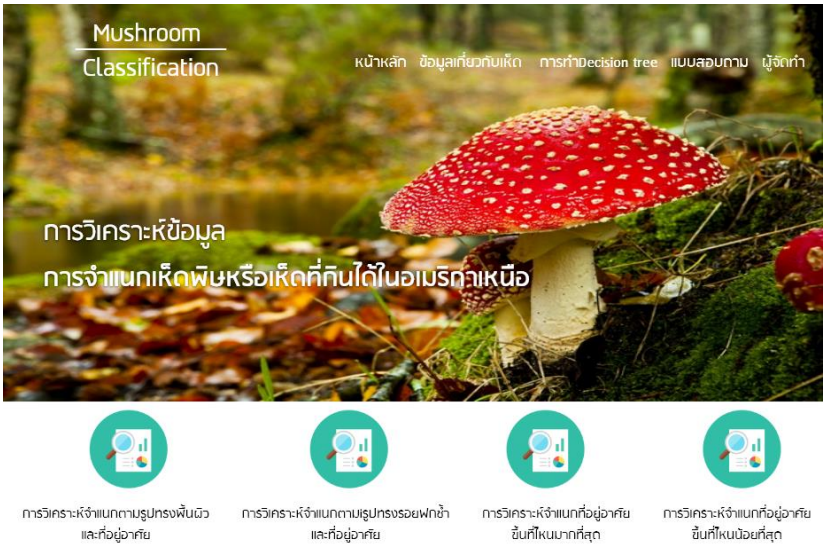


ภาพที่ 3. 25 แสดงผลข้อมูลที่อยู่อาศัยของเห็ดว่าประเภทเห็ดรับประทานได้และรับประทานไม่ได้จำแนกตามที่อยู่อาศัย

3.2 การออกแบบเว็บไซต์

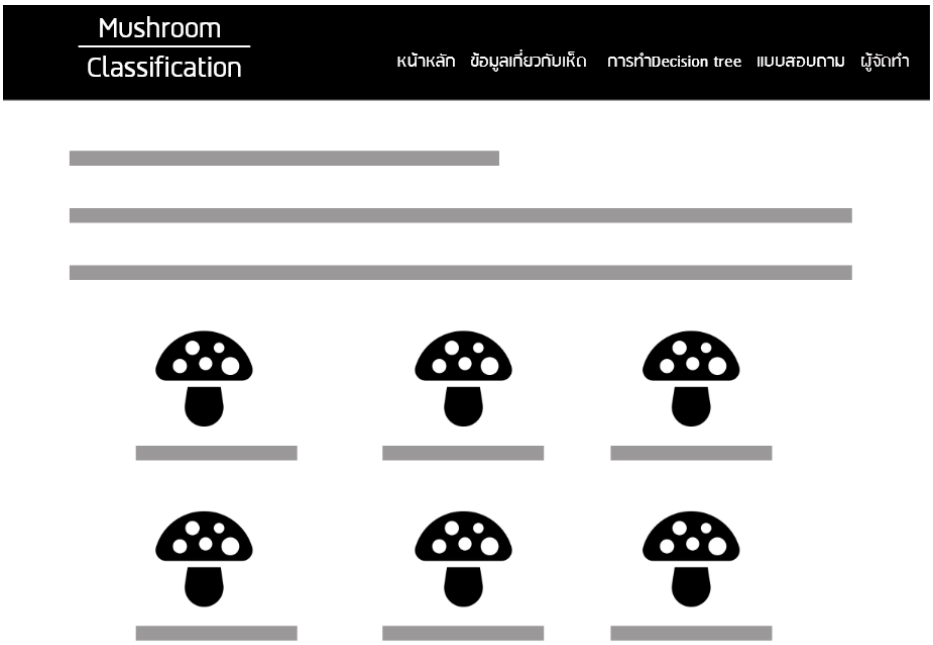
3.2.1 การออกแบบ Wireframe หน้าจอเว็บไซต์

1) หน้าแรกของเว็บไซต์ แสดงเมนูต่าง ๆ ของหน้าเว็บไซต์



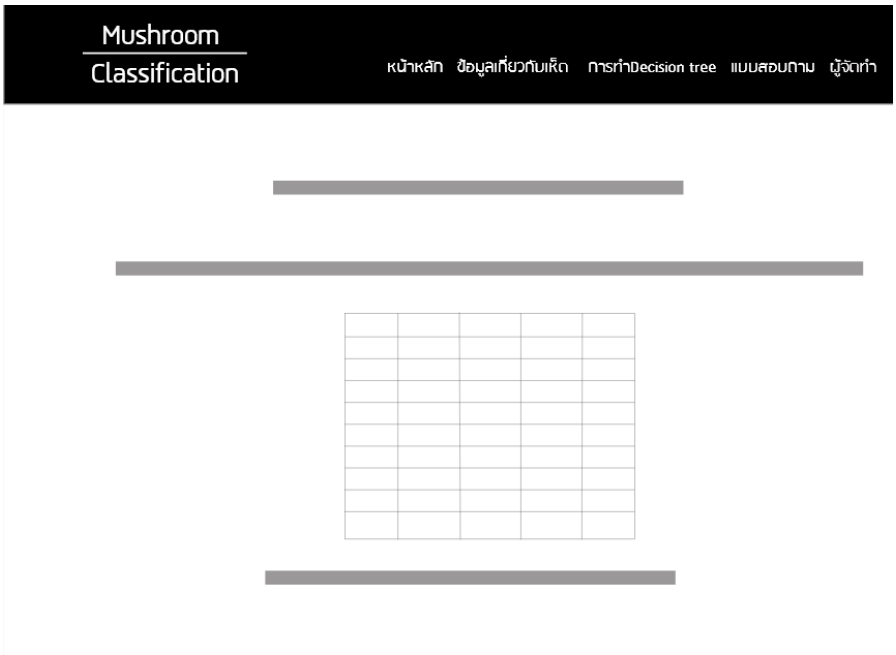
ภาพที่ 3. 26 แสดงหน้าแรกของเว็บไซต์ แสดงเมนูต่าง ๆ ของหน้าเว็บไซต์

2) หน้าข้อมูลเกี่ยวกับเห็ด



ภาพที่ 3. 27 แสดงหน้าข้อมูลเกี่ยวกับเห็ด

3) หน้าการทำการทำDecision tree

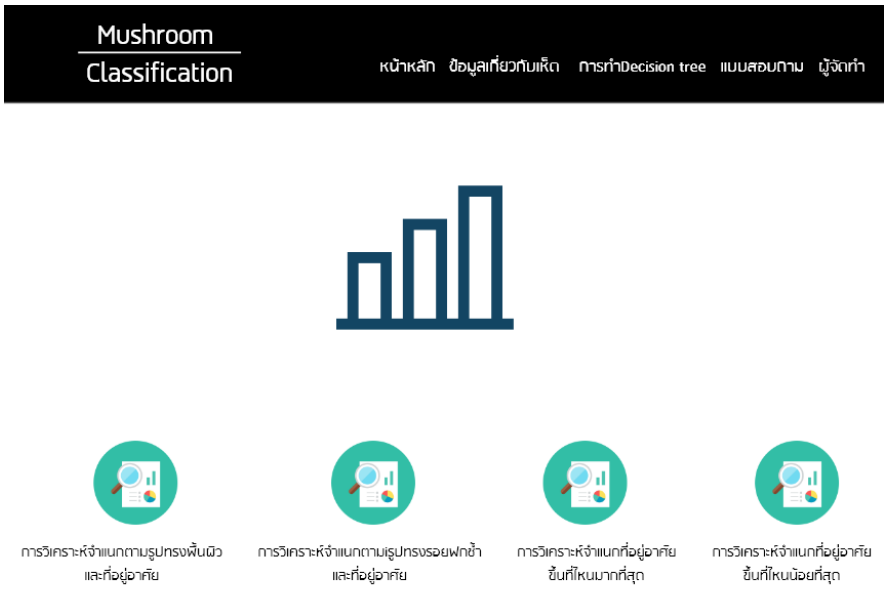


ภาพที่ 3. 28 แสดงหน้าการทำการทำDecision tree

4) หน้าแสดงผลข้อมูลการวิเคราะห์จำแนกตามรูปทรงพื้นผิวและที่อยู่อาศัย

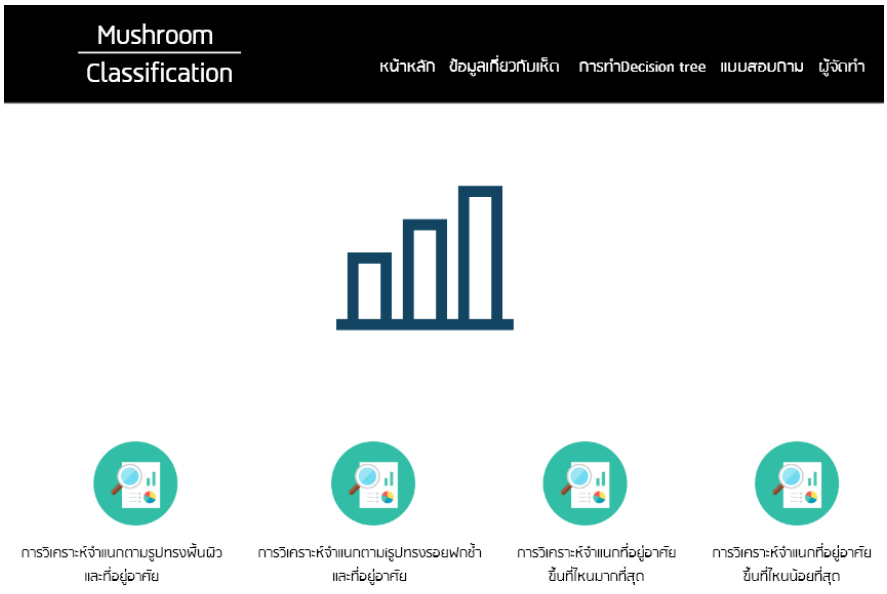
ภาพที่ 3. 29 แสดงหน้าแสดงผลข้อมูลการวิเคราะห์จำแนกตามรูปทรงพื้นผิวและที่อยู่อาศัย

5) หน้าแสดงผลข้อมูลการวิเคราะห์จำแนกตามรูปทรงรอยฟกช้ำและที่อยู่อาศัย



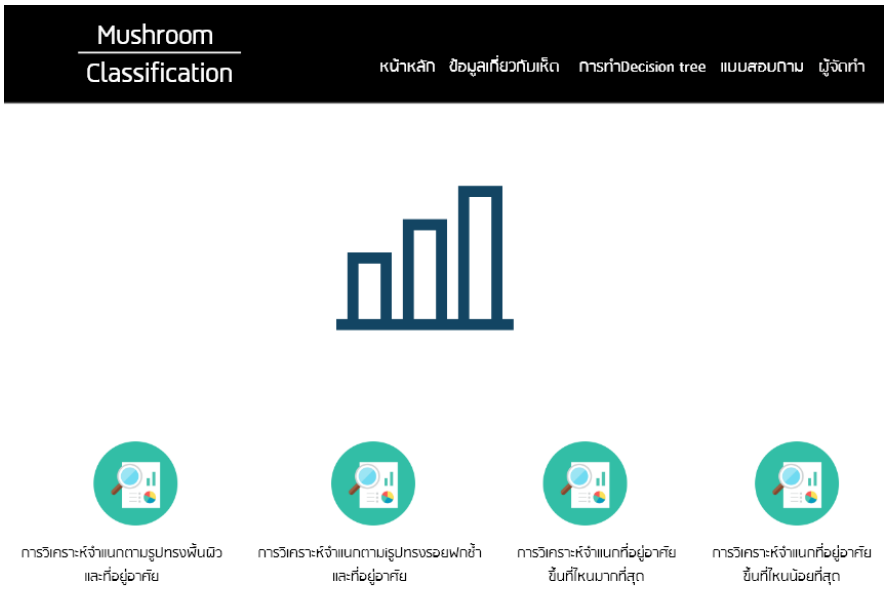
ภาพที่ 3. 30 แสดงหน้าแสดงผลข้อมูลการวิเคราะห์จำแนกตามรูปทรงรอยฟกช้ำและที่อยู่อาศัย

6) หน้าแสดงผลข้อมูลการวิเคราะห์จำแนกที่อยู่อาศัยขึ้นที่ไหนมากที่สุด



ภาพที่ 3. 31 แสดงหน้าแสดงผลข้อมูลการวิเคราะห์จำแนกที่อยู่อาศัยขึ้นที่ไหนมากที่สุด

7) หน้าแสดงผลข้อมูลการวิเคราะห์จำแนกที่อยู่อาศัยขึ้นที่ไหนน้อยที่สุด



ภาพที่ 3. 32 แสดงหน้าแสดงผลข้อมูลการวิเคราะห์จำแนกที่อยู่อาศัยขึ้นที่ไหนน้อยที่สุด

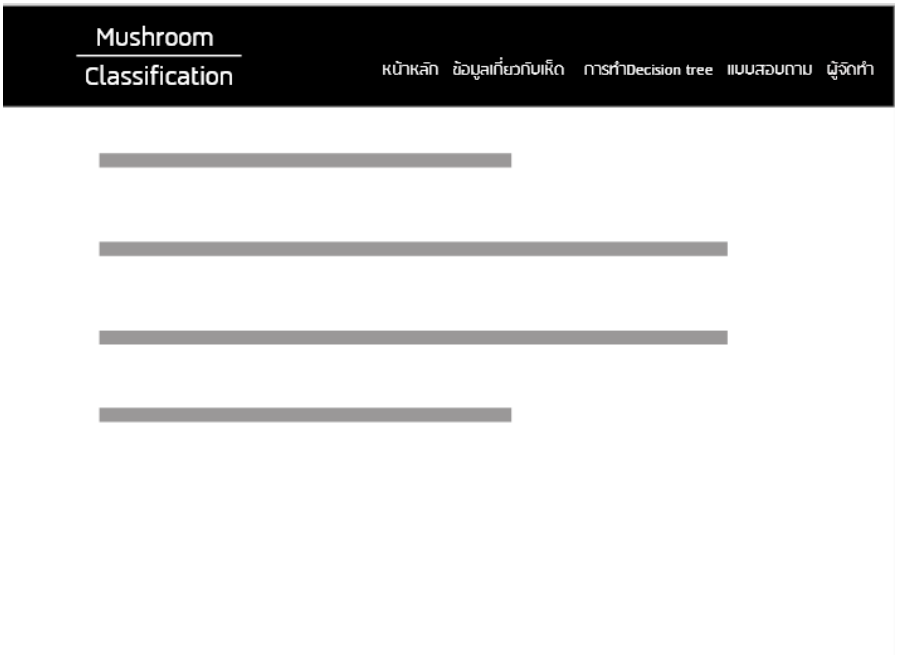
8) หน้าแบบทดสอบการให้ความรู้เกี่ยวกับเห็ด

ภาพที่ 3. 33 แสดงหน้าแบบทดสอบการให้ความรู้เกี่ยวกับเห็ด

9) หน้าแบบสอบถามการนำ Decision Treeมาประยุกต์

ภาพที่ 3. 34 แสดงหน้าแบบสอบถามการนำ Decision Treeมาประยุกต์

10) หน้าแบบสอบถามความพึ่งพอใจการใช้เว็บไซต์



ภาพที่ 3. 35 แสดงหน้าแบบสอบถามความพึ่งพอใจการใช้เว็บไซต์

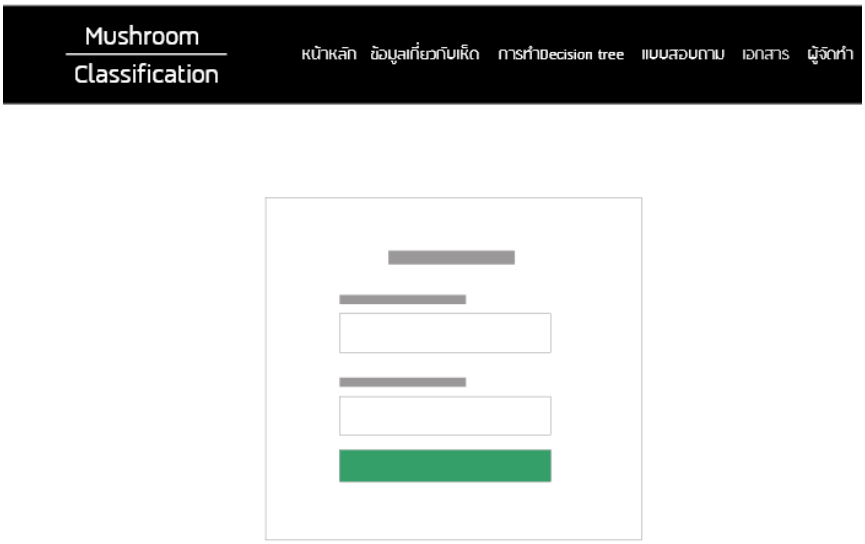
11) หน้าแสดงผลข้อมูลผู้จัดทำ

ภาพที่ 3. 36 แสดงหน้าแสดงผลข้อมูลผู้จัดทำ

12) หน้าเอกสาร

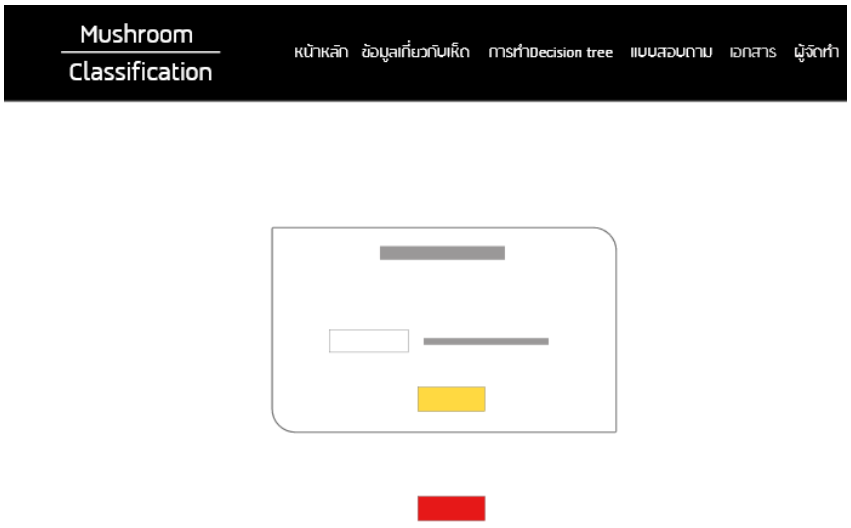
ภาพที่ 3. 37 แสดงหน้าเอกสาร

13) หน้าล็อกอินสำหรับผู้วิเคราะห์



ภาพที่ 3. 38 แสดงหน้าล็อกอินสำหรับผู้วิเคราะห์

14)หน้าอัปโหลดไฟล์



3.3 บทสรุป

จากวิธีการดำเนินงานโครงงานในข้างต้นทั้งหมดนี้ ผู้วิเคราะห์ข้อมูลการวิเคราะห์ข้อมูลการแยกประเภทของเห็ดในอเมริกาเหนือ เพื่อเผยแพร่ข้อมูลบนเว็บไซต์ ด้วยขั้นตอนกระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM อย่างละเอียด รวมถึงการสร้างโมเดล Decision Tree จากการคำนวณด้วยมือและโปรแกรมที่ใช้ทำเหมืองข้อมูล คือ, โปรแกรม Weka 3.8.4 และโปรแกรม RapidMiner Studio 9.5.1 ในการสร้างโมเดล Decision Tree ซึ่งได้ผลลัพธ์ที่เหมือนกัน และนำข้อมูลสารสนเทศมาทำการแสดงผลแบบ visualization ในรูปแบบของภาพโดยใช้โปรแกรม Tableau Public และออกแบบ Wireframe ของเว็บไซต์ที่จะเผยแพร่บน Web Browser ด้วยโปรแกรม Adobe XD