บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 โรคหลอดเลือดสมอง

โรคหลอดเลือดสมอง(Stroke) [1] คือภาวะที่สมองขาดเลือดไปเลี้ยงเนื่องจากหลอดเลือดตีบ หลอดเลือดอุดตันหรือหลอดเลือดแตกส่งผลให้เนื้อเยื่อในสมองถูกทำลายการทำงานของสมองหยุดชะงัก เป็นภาวะสมองขาดเลือด ทางการแพทย์เรียกว่า CardioVascular Disease (CVD) เกิดจากความผิดปกติของระบบประสาทเนื่องจากมีการเปลี่ยนแปลงของการไหลเวียนของเลือดภายในสมอง หรือภาวะที่มี

ความบกพร่องของการทำงานของสมอง เนื่องจากหลอดเลือดสมองตีบหรืออุดตัน (Ischemic stroke) หรือหลอดเลือดสมองแตก (Hemorrhagic stroke) สมองจึงขาดเลือดไปเลี้ยง ทำให้เนื้อเยื่อในสมองถูกทำลาย การทำงานของสมองหยุดชะงัก โรคหลอดเลือดสมองเป็นโรคที่คุกคามต่อชีวิตและความเป็นอยู่ของคนทั่วโลกโดยทั่วไปโรคหลอดเลือดสมองสมองอาจขาดเลือดทันทีภายในระยะเวลาเป็นนาทีหรือชั่วโมงแต่ไม่ใช่แบบค่อยเป็นค่อยไป โดยมีอาการที่เห็นได้ชัด คือ อ่อนแรงครึ่งซีก ชาครึ่งซีก เดินเซ พูดไม่ชัดหรือมองเห็นภาพซ้อนร่วมกับอาการต่างๆ ขึ้นอยู่กับบริเวณของสมองที่ขาดเลือด ความผิดปกติของหลอดเลือดที่ทำให้สมองขาดเลือด มีสาเหตุสำคัญที่ควรคำนึงอยู่ 2 ประการ คือ หลอดเลือดสมองอุดตันและหลอดเลือดสมองแตก [3]

1. หลอดเลือดสมองอุดตัน (Ischemic stroke) พบได้ 70% ของโรคหลอดเลือดสมอง เกิดจากการที่เลือดไปเลี้ยงสมองไม่เพียงพอ ซึ่งเกิดจากสาเหตุสำคัญ 3 ประการ คือ

1.1 การอุดตันของหลอดเลือดจากการเสื่อมหรือการแข็งตัวของหลอดเลือด (Atherosclerosis) เป็นสาเหตุของหลอดเลือดอุดตันที่พบบ่อยที่สุด เกิดจากการที่ผู้ป่วยมีปัจจัยเสี่ยง เช่น สูงอายุ ความดันโลหิตสูง เบาหวาน สูบบุหรี่หรือไขมันในเลือดสูง เป็นต้น หลอดเลือดของผู้ป่วยจะค่อยๆ แข็งตัวและตีบลงเรื่อยๆ จากการที่มีไขมันใยไฟบรินและแคลเซียมมาสะสมที่ผนังหลอดเลือดที่เรียกว่า พลัค (plaque) เมื่อพลัคมีขนาดใหญ่ขึ้นจนเหลือช่องในหลอดเลือดเล็กลงเกิดการอุดตันทำให้ขาดเลือดไปเลี้ยงสมอง สมองหยุดทำงานและเกิดอาการของโรคหลอดเลือดสมองขึ้น

1.2 ก้อนเลือดจากหัวใจหรือตะกอนเลือดจากผนังหลอดเลือดแดงที่คอด้านหน้าหลุดเข้าไปอุดตันหลอดเลือดในสมอง มักเกิดในคนที่มีการเต้นหัวใจไม่สม่ำเสมอชนิดหัวใจห้องซ้ายบนเต้นพลิ้ว (Atrial Fibrillation หรือ AF) การเต้นของหัวใจที่บีบตัวไม่พร้อมกันทั้งห้อง ทำให้เลือดค้างในห้องหัวใจ เลือดจะเกิดการแข็งตัวเป็นก้อนเลือดหลุดเข้าไปในสมอง นอกจากนี้ตะกอนเลือดที่อยู่ที่ผิวของ plaque ในผนังหลอดเลือดใหญ่ที่คอสามารถหลุดเข้าไปติดในหลอดเลือดสมองทำให้เกิดการอุดตันของหลอดเลือดสมองได้เช่นกัน

1.3 ความดันเลือดลดลงมากจนไปเลี้ยงสมองไม่พอเป็นสาเหตุที่พบน้อยมาก

2. หลอดเลือดสมองแตก (Hemorrhagic stroke) พบได้ประมาณ 30% เกิดจากหลอดเลือดมีความเปราะบางร่วมกับภาวะความดันโลหิตสูง ทำให้หลอดเลือดบริเวณที่เปราะบางโป่งพองและแตกออกหรือสูญเสียความยืดหยุ่นจากการสะสมของไขมันในหลอดเลือดทำให้หลอดเลือดบริเวณนั้นปริแตกได้ง่าย ส่งผลให้ปริมาณเลือดที่ไปเลี้ยงสมองลดลงในทันทีและเกิดเลือดออกในสมอง เป็นสาเหตุให้ผู้ป่วยเสียชีวิตในเวลาอันรวดเร็วได้ พบได้บ่อยในผู้ป่วยโรคความดันโลหิตสูงและโรคหลอดเลือดสมองโป่งพอง (Aneurysm) นอกจากนี้ยังเกิดจากความเครียด การดื่มแอลกอฮอล์รวมทั้งยาบางชนิด

จากรายงานขององค์การอัมพาตโลก (WSO) ในปี 2562 [4] พบว่าโรคหลอดเลือดสมองเป็นสาเหตุการเสียชีวิตอันดับ 2 ของโลก โดยมีผู้ป่วยที่เป็นโรคหลอดเหลือดสมองจำนวน 80 ล้านคน และมีผู้เสียชีวิตประมาณ 5.5 ล้านคน นอกจากนี้ผู้ป่วยโรคหลอดเหลือดสมองมีแนวโน้มจะสูงขึ้นและยังพบผู้ป่วยใหม่ถึง 13.7 ล้านคนต่อปี โดย 1 ใน 4 เป็นผู้ป่วยที่มีอายุ 25 ปีขึ้นไป และร้อยละ 60 เสียชีวิตก่อนวัยอันควร

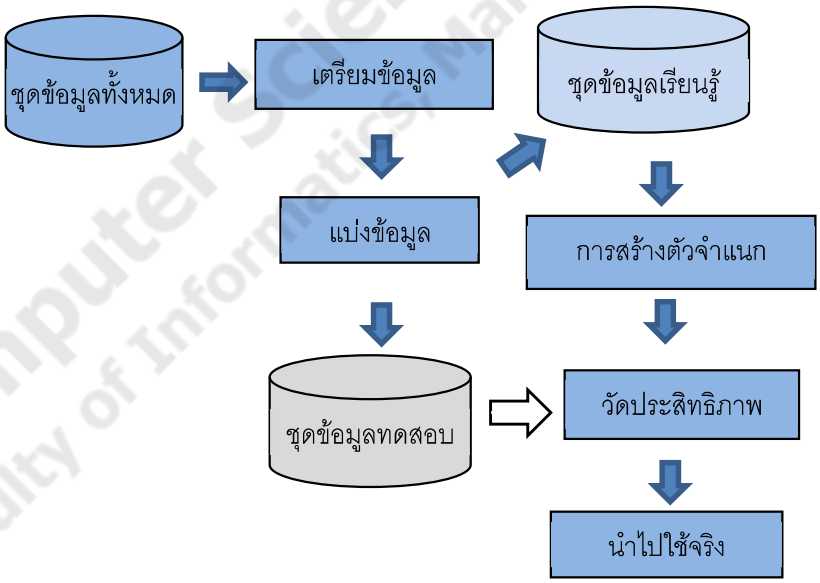
ภาพประกอบที่ 2.1 สถิติโรคหลอดเลือดสมองทั่วโลก

ในประเทศไทย จากรายงานของกองยุทธศาสตร์และแผนงาน กระทรวงสาธารณสุข พบว่าตั้งแต่ปี 2556-2560 อัตราการเป็นโรคหลอดเหลือดสมองมีแนวโน้มสูงขึ้น โดยผู้ป่วยโรคหลอดเลือดสมองในปี 2559 มีจำนวน 293,463 ราย ผู้ป่วยโรคหลอดเลือดสมองในปี 2560 มีจำนวน 304,807 ราย และมีจำนวนผู้เสียชีวิตจากโรคหลอดเลือดสมองปีละประมาณ 30,000 ราย จากสถิติดังกล่าวแสดงให้เห็นว่า โรคหลอดเลือดสมองเป็นสาเหตุการเสียชีวิตอันดับ 1 ของประเทศไทย และสามารถเกิดได้กับประชาชนทุกกลุ่มวัย กระทรวงสาธารณสุข โดยกรมควบคุมโรค จึงได้กำหนดคำขวัญการรณรงค์วันอัมพาตโลกในวันที่ 29 ตุลาคม 2562 คือ “อย่าให้ อัมพฤกษ์ อัมพาต...เป็นส่วนหนึ่งในชีวิตคุณ” เพื่อให้ทุกคนตระหนักถึงความร้ายแรงของการเกิดโรคหลอดเลือดสมอง และมีความตระหนักในการป้องกันโรคหลอดเหลือดสมอง โดยรณรงค์ให้คนไทยทราบถึงสัญญาณเตือนของโรคหลอดเหลือดสมอง คือ “F.A.S.T” F (Face) เวลายิ้มแล้วพบว่ามุมปากข้างหนึ่งตก, A (Arms) ยกแขนข้างใดข้าง

หนึ่งไม่ขึ้น, S (Speech) มีปัญหาด้านการพูด แม้แต่ประโยคง่ายๆ, และ T (Time) ถ้ามีอาการดังกล่าว ควรนำส่งตัวที่โรงพยาบาลโดยด่วนภายใน 4 ชั่วโมงครึ่งรวมการรักษา เพื่อจะได้รับการรักษาให้ทันเวลาและสามารถฟื้นฟูให้กลับมาได้เป็นปกติมากที่สุดจะเห็นได้ว่าโรคหลอดเลือดสมองเป็นโรคที่มีความสำคัญมาก เป็นสาเหตุการณ์การตายของผู้สูงอายุในประเทศไทย ดังนั้นการศึกษาปัจจัยที่นำไปสู่โรคหลอดเลือดสมองจึงเป็นงานที่ท้าทาย ซึ่งจะทำให้ทราบถึงปัจจัยที่นำไปสู่โรคหลอดเลือดสมอง เพื่อหาทางป้องกันหรือดูแลสุขภาพไม่ให้เป็นโรคหลอดเลือดสมองได้

2.2 การจำแนกข้อมูล

การจำแนกข้อมูล เป็นการจำแนกประเภทข้อมูล ซึ่งในเหมืองข้อมูลจะประกอบไปด้วย การสร้างตัวจำแนก และใช้ตัวจำแนกดังกล่าวในการทำนายประเภทของข้อมูล โดยตัวจำแนกถูกสร้างขึ้นจากข้อมูลเรียนรู้ (Training set) ซึ่งเป็นข้อมูลที่ได้จำแนกประเภทไว้แล้ว เมื่อได้สร้างตัวจำแนกเสร็จแล้วจะทำการวัดประสิทธิภาพตัวจำแนกก่อนนำไปใช้จริง โดยนำข้อมูลทดสอบ (Test-ing) ซึ่งได้ระบุประเภทไว้แล้วไปทดสอบกับตัวจำแนก เพื่อดูผลการทำนายของตัวจำแนกว่ามีประสิทธิภาพมากน้อยเพียงใด เมื่อได้ตัวจำแนกที่มีประสิทธิภาพก็จะนำตัวจำแนกไปใช้ในการทำนายข้อมูลจริง สิ่งที่ต้องคำนึงในการจำแนกข้อมูลมี 2 ปัจจัย คือ การเตรียมข้อมูลและการวัดประสิทธิภาพ



ภาพประกอบที่ 2.2 กระบวนการจำแนกข้อมูล

2.2.1 การเตรียมข้อมูล

การเตรียมข้อมูลถือว่าเป็นสิ่งที่สำคัญ เนื่องจากการเตรียมข้อมูลที่ดีจะทำให้ตัวจำแนกมีประสิทธิภาพที่ดี โดยการเตรียมข้อมูลประกอบไปด้วยข้อดังต่อไปนี้

1. การทำความสะอาดข้อมูล ซึ่งเป็นการลดหรือลบข้อมูลรบกวนออกไป หรือทำการแทนค่าที่ขาดหายไปเพื่อให้ข้อมูลที่นำมาใช้มีความสมบูรณ์

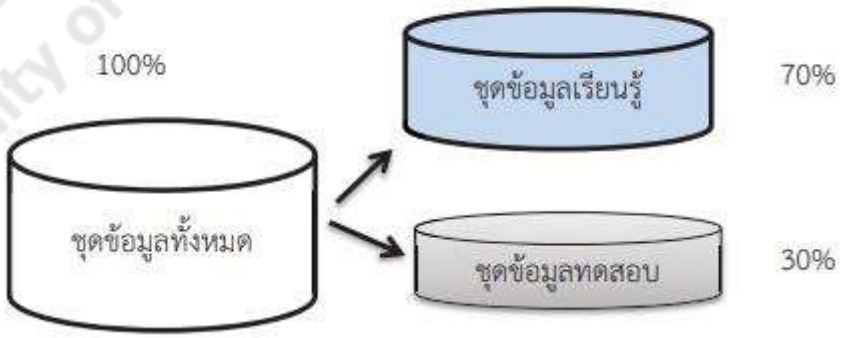
2. การเกี่ยวเนื่องของข้อมูล เป็นการตรวจสอบข้อมูลแต่ละคุณลักษณะว่ามีความเกี่ยวเนื่องหรือซ้ำซ้อนกันมากน้อยเพียงไหน ซึ่งสามารถประยุกต์ใช้การวิเคราะห์สหสัมพันธ์ (Correlation analysis) เพื่อตรวจสอบความเกี่ยวเนื่องของคุณลักษณะ

3. การเปลี่ยนรูปข้อมูลและการลดจำนวนข้อมูล การเปลี่ยนรูปข้อมูล ช่วยในเรื่องของการลดความหลากหลายของข้อมูลและทำให้ข้อมูลมีความละเอียดน้อยลง

4. การแปลงข้อมูล (Data Transformation)การแปลงข้อมูลเป็นขั้นตอนการแปลงให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมเพื่อนำไปใช้กับขั้นตอนวิธีต่าง ๆ ได้ การแปลงข้อมูลในงานวิจัยนี้มีรายละเอียดดังนี้

2.2.2 การแบ่งข้อมูล

วิธีการแบ่งข้อมูล Hold-out validation คือ การแบ่งข้อมูลออกเป็น 2 ส่วนตามอัตราส่วนที่กำหนด โดยจะต้องทำการสุ่มข้อมูลก่อนทำการแบ่ง เช่น แบ่งชุดข้อมูลเรียนรู้เป็น 70% จากข้อมูลทั้งหมดและแบ่งข้อมูลทดสอบเป็น 30% จากข้อมูลทั้งหมด ภาพประกอบที่ 2.3 เป็นต้น อัตราส่วนในการแบ่งชุดข้อมูลเรียนรู้และข้อมูลทดสอบขึ้นอยู่กับผู้ใช้หรือนักวิจัยที่จะกำหนดอัตราส่วนเองหรืออาจจะได้จากการทดลอง แต่นิยมแบ่งอัตราส่วนของชุดข้อมูลเรียนรู้มากกว่าชุดข้อมูลทดสอบ เพื่อให้ตัวจำแนกมีข้อมูลสำหรับการเรียนรู้ที่เพียงพอ การแบ่งข้อมูงโดยใช้วิธีนี้จะต้องทำการทดสอบหลาย ๆ ครั้งและแต่ละครั้งจะต้องสุ่มชุดข้อมูลทดสอบและชุดข้อมูลทดสอบและชุดข้อมูลเรียนรู้ต่างกันไป



ภาพประกอบที่ 2.3 การแบ่งข้อมูลแบบ Hold-out validation

2.2.3 การสร้างตัวจำแนก

ตัวจำแนกถูกสร้างขึ้นจากชุดชุดข้อมูลเรียนรู้ ซึ่งเป็นข้อมูลที่มีการระบุกลุ่มข้อมูลไว้แล้ว ตัวจำแนกที่สร้างขึ้นได้จากการเรียนรู้ลักษณะหรือรูปแบบของข้อมูลที่มีการระบุกลุ่มข้อมูลไว้แล้ว ตัวจำแนกถูกสร้างขั้นด้วยวิธีการที่หลากหลาย เช่น นาอีฟเบส์ ต้นไม้ตัดสินใจ การค้นหาเพื่อนบ้านที่ใกล้ที่สุด k ตัว ซัพพอร์ตเวกเตอร์แมชชีน และ การจำแนกเชิงความสัมพันธ์ เป็นต้น ซึ่งแต่ละวิธีมีกระบวนการแตกต่างกันในการสร้างตัวจำแนก เช่น นาอีฟเบส์ใช้พื้นฐานทฤษฎีเบส์เพื่อสร้างตัวจำแนก ซึ่งพิจารณาจากความเป็นไปได้ของข้อมูลและกลุ่ม การค้นหาเพื่อนบ้านที่ใกล้ที่สุด k ตัวเป็นจำแนกข้อมูลโดยคำนวณระยะความห่างระหว่างแอตทริบิวต์ ทำการเลือกแค่ k กลุ่มที่มีระยะห่างใกล้ที่สุด แล้วกำหนดกลุ่มจากกลุ่มที่มีจำนวนมากที่สุดใน k กลุ่ม ส่วนซัพพอร์ตเวกเตอร์แมชชีนจำแนกข้อมูลด้วยการหาเส้นแบ่งที่เหมาะสม การจำแนกเชิงความสัมพันธ์ใช้กฎความสัมพันธ์ในการจำแนกข้อมูล ซึ่งเป็นวิธีการที่มีประสิทธิภาพและง่ายต่อการเข้าใจ

2.2.4 การวัดประสิทธิภาพ

การเปรียบเทียบประสิทธิภาพในการจำแนก โดยส่วนใหญ่การจำแนกข้อมูลจะต้องทำการวัดประสิทธิภาพของตัวจำแนกก่อนนำไปใช้ ซึ่งการวัดประสิทธิภาพตัวจำแนกสามารถวัดได้ดังนี้

1. ความถูกต้อง เป็นความสามารถในการทำนายข้อมูลของตัวจำแนกว่าสามารถทำนายข้อมูลได้ถูกต้องมากน้อยเพียงใด โดยการประเมินความถูกต้องอาจจะใช้ข้อมูลชุดเดียวหรือหลายๆชุดก็ได้

2. ความเร็ว เป็นความเร็วในการทำนายข้อมูลของตัวจำแนก

3.ความทนทาน เป็นความสามารถของตัวจำแนกว่าสามารถทำนายข้อมูลที่มีสิ่งรบกวนหรือการขาดหายไปของข้อมูลได้หรือไม่

4. ความยืดหยุ่นต่อปริมาณข้อมูล เป็นความสามารถของตัวจำแนกในการทำนายข้อมูลที่มีปริมาณมหาศาล

5. ความสามารถในการเข้าใจ เป็นความสามารถที่ตัวจำแนกสามารถเข้าใจได้ง่ายจากผู้ใช้งาน วิธีการที่ได้รับความนิยมในการวัดประสิทธิภาพในการจำแนก คือ ความสามารถในการทำนาย ซึ่งนิยมวัดด้วย ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าระลึก (Recall) ค่าประสิทธิภาพโดยรวม (F-measure) ซึ่งจะกล่าวโดยละเอียดในหัวข้อถัดไป

2.3 ต้นไม้ตัดสินใจ

การเรียนรู้ของต้นไม้ตัดสินใจ(Decision Tree) เป็นการเรียนรู้โดยการจำแนกประเภท (Classification)ข้อมูลออกเป็นกลุ่ม(Class) ต่างๆ โดยใช้คุณลักษณะ(attribute) ข้อมูลในการจำแนก

ประเภท ต้อนไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่า คุณลักษณะใดเป็นตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวมีความสำคัญมากน้อยต่างกันอย่างไร เพราะฉะนั้น การจำแนกประเภทมีประโยชน์ช่วยให้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องยิ่งขึ้น

ส่วนประกอบของผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจ

- โนดภายใน (internal node)คือ คุณลักษณะต่าง ๆ ของข้อมูล ซึ่งเมื่อข้อมูลใดๆ ตกลงมาที่โนด จะใช้คุณลักษณะนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโนดภายในทที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่าโนดราก
- กิ่ง (branch, link) เป็นค่าของคุณลักษณะในโนดภายในที่แตกกิ่งนี้ออกมา ซึ่งโนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าของคุณลักษณะในโนดภายในนั้น
- โนดใบ (leaf node) คือกลุ่มต่าง ๆ ซึ่งเป็นผลลัพธ์ในการจำแนกประเภทข้อมูล

ขั้นตอนวิธีทำ Decision Tree

- ต้นไม้ตัดสินใจสร้างโดยวิธีแบบ top-down recursive
- เริ่มต้นด้วยนำตัวอย่างการสอน มาสร้างเป็นราก
- Attribute ควรอยู่ในรูปของ Categorical คือข้อมูลชนิดกลุ่ม หากเป็นข้อมูลที่อยู่ในรูป Continuous หรือ Numeric เป็นข้อมูลมีความต่อเนื่องกันควรทำการแบ่งข้อมูลให้ เป็นกลุ่มก่อน
- การสร้างต้นไม้ตัดสินใจมีพื้นฐานมาจากวิธีการเลือก Attribute
- เมื่อไหร่ถึงจะหยุดการสร้างต้นไม้
    - เมื่อทุกข้อมูลใน node นั้นเป็น Class เดียวกัน
    - เมื่อทุกข้อมูลใน node นั้นมีค่าของ Attribute เหมือนกัน

ตัวอย่างเช่นตารางต่อไปนี้บอกปัจจัยในการตัดสินใจเล่นเทนนิสนอกบ้านในช่วง 14 วันที่ผ่านมา

ตารางที่ 2.1 ตัวอย่างข้อมูลปัจจัย

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes

ตารางที่ 2.1 ตัวอย่างข้อมูลปัจจัย(ต่อ)

Day	Outlook	Temp.	Humidity	Wind	Decision
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

สามารถสรุปอัลกอริทึม ID3 ได้ดังภาพประกอบด้านล่าง

Entropy(S) = ∑ – p(I) . log2p(I)

Gain (S, A) = Entropy(S) – ∑ [p(S|A) . Entropy(S|A)]

Entropy

เราต้องคำนวณ Entropy ก่อน คอลัมน์การตัดสินใจประกอบด้วย 14 ตัวอย่าง และมีสองป้ายกำกับ: ใช่ และไม่ใช่ มีการตัดสินใจ 9 ข้อระบุว่าใช่ และการตัดสินใจ 5 ข้อระบุว่าไม่ใช่

Entropy (Decision) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy (Decision) = – (9/14) . log2(9/14) – (5/14) . log2(5/14) = 0.940

ในขั้นตอนต่อไป จะทำการคำนวนหาปัจจัยของคอลัมน์แรก คือ Outlook

เราต้องคำนวณ (Decision| Outlook = Sunny), (Decision| Outlook = Overcast) and (Decision| Outlook = Rain) ก่อนเป็นอันดับแรก

ตารางที่ 2.2 ตัวอย่างข้อมูลสำหรับ Sunny

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes

ตารางที่ 2.2 ตัวอย่างข้อมูลสำหรับ Sunny(ต่อ)

Day	Outlook	Temp.	Humidity	Wind	Decision
11	Sunny	Mild	Normal	Strong	Yes

มี 5 กรณีสำหรับ Sunny การตัดสินของ 3 ข้อคือไม่ใช่ และ 2 ข้อคือใช่

Entropy(Decision| Outlook = Sunny) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Outlook = Sunny) = – (2/5) . log2(2/5) – (3/5) . log2(3/5) = 0.971

ตารางที่ 2.3 ตัวอย่างข้อมูล สำหรับ Overcast

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes

มี 4 กรณีสำหรับ Overcast การตัดสินของ 0 ข้อคือไม่ใช่ และ 4 ข้อคือใช่

Entropy(Decision| Outlook = Overcast) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Outlook = Overcast) = – (4/4) . log2(4/4) – (0/4) . log2(0/4) = 0

ตารางที่ 2.4 ตัวอย่างข้อมูล สำหรับ Rain

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

มี 5 กรณีสำหรับ Rain การตัดสินของ 3 ข้อคือไม่ใช่ และ 2 ข้อคือใช่

Entropy(Decision| Outlook = Rain) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Outlook = Rain) = – (3/5) . log2(3/5) – (2/5) . log2(2/5) = 0.971

โดยแทนค่าในสูตรดังต่อไปนี้

Gain(Decision, Outlook) = Entropy(Decision) – [p(Decision| Outlook = Sunny) . Entropy(Decision| Outlook = Sunny)] – [p(Decision| Outlook = Overcast) . Entropy(Decision| Outlook = Overcast) – [p(Decision| Outlook = Rain) . Entropy(Decision| Outlook = Rain)]

Gain(Decision, Outlook) = 0.94 – (5/4)(0.971) – (4/5)(0) –(5/14)(0.971) = 0.2464

เราต้องใช้การคำนวณแบบเดียวกันสำหรับคอลัมน์อื่นๆ เพื่อค้นหาปัจจัยที่สำคัญที่สุดในการตัดสินใจ ในขั้นตอนต่อไป จะทำการคำนวนหาปัจจัยของคอลัมน์แรก คือ Temp โดยต้องคำนวณ (Decision| Temp = Hot), (Decision| Temp = Mild) and (Decision| Temp = Cool)

ตารางที่ 2.5 ตัวอย่างข้อมูลสำหรับ Hot

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
13	Overcast	Hot	Normal	Weak	Yes

มี 4 กรณีสำหรับ Hot การตัดสินของ 2 ข้อคือไม่ใช่ และ 2 ข้อคือใช่

Entropy(Decision| Temp = Hot) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Temp = Hot) = – (2/4) . log2(2/4) – (2/4) . log2(2/4) = 1.0

ตารางที่ 2.6 ตัวอย่างข้อมูลสำหรับ Mild

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
8	Sunny	Mild	High	Weak	No
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
14	Rain	Mild	High	Strong	No

มี 6 กรณีสำหรับ Mild การตัดสินของ 2 ข้อคือไม่ใช่ และ 4 ข้อคือใช่

Entropy(Decision| Temp = Mild) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Temp = Mild) = – (4/6) . log2(4/6) – (2/6) . log2(2/6) = 0.9183

ตารางที่ 2.7 ตัวอย่างข้อมูลสำหรับ Cool

Day	Outlook	Temp.	Humidity	Wind	Decision
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
9	Sunny	Cool	Normal	Weak	Yes

มี 4 กรณีสำหรับ Cool การตัดสินของ 1 ข้อคือไม่ใช่ และ 3 ข้อคือใช่

Entropy(Decision| Temp = Cool) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Temp = Cool) = – (3/4) . log2(3/4) – (1/4) . log2(1/4) = 0.8113

โดยแทนค่าในสูตรดังต่อไปนี้

Gain(Decision, Temp) = Entropy(Decision) – [p(Decision| Temp = Hot) . Entropy(Decision| Temp = Hot)] – [p(Decision| Temp = Mild) . Entropy(Decision| Temp = Mild) – [p(Decision| Temp = Cool) . Entropy(Decision| Temp = Cool)]

Gain(Decision, Temp) = 0.94 – (4/14)(1.0) – (6/14)(0.9183) –(4/14)(0.8113) = 0.0289

ในขั้นตอนต่อไป จะทำการคำนวนหาปัจจัยของคอลัมน์แรก คือ Humidity โดยต้องคำนวณ (Decision| Humidity = High) and (Decision| Humidity = Normal)

ตารางที่ 2.8 ตัวอย่างข้อมูลสำหรับ High

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
8	Sunny	Mild	High	Weak	No
12	Overcast	Mild	High	Strong	Yes
14	Rain	Mild	High	Strong	No

มี 7 กรณีสำหรับ High การตัดสินของ 4 ข้อคือไม่ใช่ และ 3 ข้อคือใช่

Entropy(Decision| Humidity = High) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Humidity = High) = – (3/7) . log2(3/7) – (4/7) . log2(4/7) = 0.9852

ตารางที่ 2.9 ตัวอย่างข้อมูลสำหรับ Normal

Day	Outlook	Temp.	Humidity	Wind	Decision
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

มี 7 กรณีสำหรับ Cool การตัดสินของ 4 ข้อคือไม่ใช่ และ 3 ข้อคือใช่

Entropy(Decision| Humidity = Normal) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision| Humidity = Normal) = – (6/7) . log2(6/7) – (1/7) . log2(1/7) = 0.5916

โดยแทนค่าในสูตรดังต่อไปนี้

Gain(Decision, Humidity) = Entropy(Decision) – [p(Decision| Humidity = Hot) .

Entropy(Decision| Humidity = Hot)] – [p(Decision| Humidity = Mild) . Entropy(Decision| Humidity = Mild)]

Gain(Decision, Humidity) = 0.94 – (7/14)(0.9852) – (7/14)(0.5916) = 0.1516

ในขั้นตอนต่อไป จะทำการคำนวนหาปัจจัยของคอลัมน์แรก คือ Wind โดยต้องคำนวณ (Decision| Wind = Weak) and (Decision| Wind = Strong)

ตารางที่ 2.10 ตัวอย่างข้อมูลสำหรับ Weak

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes

ตารางที่ 2.10 ตัวอย่างข้อมูลสำหรับ Weak(ต่อ)

Day	Outlook	Temp.	Humidity	Wind	Deciion
5	Rain	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
13	Overcast	Hot	Normal	Weak	Yes

มี 8 กรณีสำหรับ Weak การตัดสินของ 2 ข้อคือไม่ใช่ และ 6 ข้อคือใช่

Entropy(Decision|Wind=Weak) = – p(Yes) . log2p(Yes)– p(No) . log2p(No)

Entropy(Decision|Wind=Weak) = – (6/8) . log2(6/8) – (2/8) . log2(2/8) = 0.8113

ตารางที่ 2.11 ตัวอย่างข้อมูลสำหรับ Strong

Day	Outlook	Temp.	Humidity	Wind	Decision
2	Sunny	Hot	High	Strong	No
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
14	Rain	Mild	High	Strong	No

มี 6 กรณีสำหรับ Strong การตัดสินของ 3 ข้อคือไม่ใช่ และ 3 ข้อคือใช่

Entropy(Decision|Wind=Strong) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy(Decision|Wind=Strong) = – (3/6) . log2(3/6) – (3/6) . log2(3/6) = 1

โดยแทนค่าในสูตรดังต่อไปนี้

Gain(Decision, Wind) = Entropy(Decision) – [p(Decision| Wind = Weak) . Entropy(Decision| Wind = Weak)] – [p(Decision| Wind = Strong) . Entropy(Decision| Wind = Strong)]

Gain(Decision, Wind) = 0.94 – (8/14)(0.8113) – (6/14)(1) = 0.0478

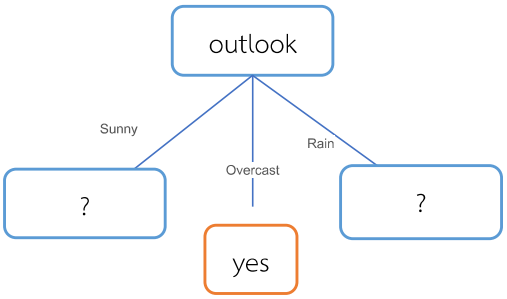
ดังที่เห็น ปัจจัยแนวโน้มในการตัดสินใจทำให้เกิดคะแนนสูงสุด นั่นเป็นเหตุผลที่การตัดสินใจของ outlook จะปรากฏในโหนดรูทของทรี

Gain(Decision, Outlook) = 0.246

Gain(Decision, Temperature) = 0.029

Gain(Decision, Humidity) = 0.151

Gain(Decision, Wind) = 0.048



ภาพประกอบที่ 2.4 โหนด decision tree

ทดสอบชุดข้อมูลสำหรับชุดย่อยที่กำหนดเองของแอตทริบิวต์ outlook ในรอบที่2 Overcast outlook on decision

ตารางที่ 2.12 ตัวอย่างข้อมูล Overcast outlook on decision

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

โดยพื้นฐานแล้ว การตัดสินใจมักจะใช่เสมอหากทัศนะมืดครึ้ม Sunny outlook on decision ในที่นี้ มี 5 กรณีสำหรับมุมมองที่มีแดดจ้า การตัดสินใจน่าจะเป็น 3/5 เปอร์เซ็นต์ไม่ใช่ 2/5 เปอร์เซ็นต์ใช่

ตารางที่ 2.13 ตัวอย่างข้อมูล Outlook ของ sunny

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

โดยเริ่มจากการหา Entropy ของ sunny

Entropy (Decision, Outlook=Sunny) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy (Decision, Outlook=Sunny) = – (2/5) . log2(2/5) – (3/5) . log2(3/5) = 0.97

ตารางที่ 2.14 ตัวอย่างข้อมูล Entropy ของ Temp

Day	Temp.	Humidity	Wind	Decision
1	Hot	High	Weak	No
2	Hot	High	Strong	No
8	Mild	High	Weak	No
9	Cool	Normal	Weak	Yes
11	Mild	Normal	Strong	Yes

Entropy (Decision, Outlook=Sunny|Temp= Hot) = – (0/2) . log2(0/2) – (2/2) . log2(2/2) = 0

Entropy (Decision, Outlook=Sunny|Temp= Mild) = – (1/2) . log2(1/2) – (1/2) . log2(1/2) = 1

Entropy (Decision, Outlook=Sunny|Temp= Cool) = – (1/1) . log2(1/1) – (0/1) . log2(0/1) = 0

Gain(Decision, Outlook=Sunny|Temp) = 0.97 – (2/5)(0) – (2/5)(1) – (1/5)(0) = 0.570

ตารางที่ 2.15 ตัวอย่างข้อมูล Outlook=Sunny | Humidity

Day	Temp.	Humidity	Wind	Decision
1	Hot	High	Weak	No
2	Hot	High	Strong	No
8	Mild	High	Weak	No
9	Cool	Normal	Weak	Yes
11	Mild	Normal	Strong	Yes

Entropy (Decision,Outlook=Sunny|Humidity = High) = – (0/3) . log2(0/3) – (3/3) . log2(3/3) = 0

Entropy (Decision,Outlook=Sunny|Humidity = Normal) = – (2/2). log2(2/2)–(0/2) . log2(0/2) = 0

Gain(Decision, Outlook=Sunny|Humidity) = 0.97 – (3/5)(0) – (2/5)(0) = 0.97

ตารางที่ 2.16 ตัวอย่างข้อมูล Outlook=Sunny |Wind

Day	Temp.	Humidity	Wind	Decision
1	Hot	High	Weak	No

ตารางที่ 2.16 ตัวอย่างข้อมูล Outlook=Sunny |Wind(ต่อ)

Day	Temp.	Humidity	Wind	Decision
2	Hot	High	Strong	No
1	Hot	High	Weak	No
2	Hot	High	Strong	No
8	Mild	High	Weak	No
9	Cool	Normal	Weak	Yes
11	Mild	Normal	Strong	Yes

Entropy (Decision, Outlook=Sunny|Wind = Weak) = – (1/3) . log2(1/3) – (2/3) . log2(2/3) = 0.9183

Entropy (Decision, Outlook=Sunny|Wind = Strong) = – (1/2) . log2(1/2) – (1/2) . log2(1/2) = 1.0

Gain(Decision, Outlook=Sunny|Wind) = 0.97 – (3/5)(0.9183) – (2/5)(1) = 0.0192

ปัจจัยแนวโน้มในการตัดสินใจทำให้เกิดคะแนนสูงสุด

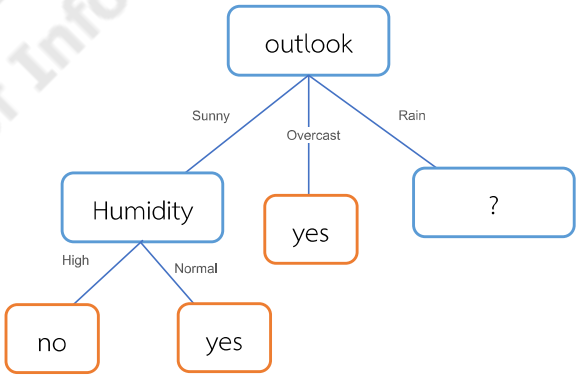
Gain(Outlook=Sunny|Temperature) = 0.570

Gain(Outlook=Sunny|Humidity) = 0.970

Gain(Outlook=Sunny|Wind) = 0.0192

Humidity คือการตัดสินใจ เพราะมันให้คะแนนสูงสุดหากแนวโน้มมี Sunny

ณ จุดนี้ การตัดสินใจจะไม่เกิดขึ้นหากมี Humidity สูง



ภาพประกอบที่ 2.5 โหนด

ตารางที่ 2.17 ตัวอย่างข้อมูล Outlook= Rain

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

โดยเริ่มจากการหา Entropy ของ Rain

Entropy (Decision, Outlook= Rain) = – p(Yes) . log2p(Yes) – p(No) . log2p(No)

Entropy (Decision, Outlook= Rain) = – (3/5) . log2(3/5) – (2/5) . log2(2/5) = 0.97

ตารางที่ 2.18 ตัวอย่างข้อมูล Outlook= Rain | Temp

Day	Temp.	Humidity	Wind	Decision
4	Mild	High	Weak	Yes
5	Cool	Normal	Weak	Yes
6	Cool	Normal	Strong	No
10	Mild	Normal	Weak	Yes
14	Mild	High	Strong	No

Entropy (Decision, Outlook= Rain |Temp= Hot) = – (0/0) . log2(0/0) – (0/0) . log2(0/0) = 0

Entropy (Decision, Outlook= Rain |Temp= Mild) = – (2/3) . log2(2/3) – (1/3) . log2(1/3) = 0.9183

Entropy (Decision, Outlook= Rain |Temp= Cool) = – (1/2) . log2(1/2) – (1/2) . log2(1/2) = 1

Gain(Decision, Outlook= Rain |Temp) = 0.97 – (0/0)(0) – (3/5)(0.9183) – (2/5)(1) = 0.0192

ตารางที่ 2.19 ตัวอย่างข้อมูล Outlook= Rain | Humidity

Day	Temp.	Humidity	Wind	Decision
4	Mild	High	Weak	Yes
5	Cool	Normal	Weak	Yes
6	Cool	Normal	Strong	No
10	Mild	Normal	Weak	Yes

ตารางที่ 2.19 ตัวอย่างข้อมูล Outlook= Rain | Humidity(ต่อ)

Day	Temp.	Humidity	Wind	Decision
14	Mild	High	Strong	No

Entropy (Decision, Outlook= Rain |Humidity = High) = – (1/2) . log2(1/2) – (1/2) . log2(1/2) =1

Entropy (Decision, Outlook= Rain |Humidity = Normal) = – (2/3) . log2(2/3) – (1/3) . log2(1/3) = 0.9183

Gain(Decision, Outlook= Rain |Humidity) = 0.97 – (2/5)(1) – (3/5)(0.9183) = 0.0192

ตารางที่ 2.20 ตัวอย่างข้อมูล Outlook= Rain | Wind

Day	Temp.	Humidity	Wind	Decision
4	Mild	High	Weak	Yes
5	Cool	Normal	Weak	Yes
6	Cool	Normal	Strong	No
10	Mild	Normal	Weak	Yes
14	Mild	High	Strong	No

Entropy (Decision, Outlook= Rain |Wind = Weak) = – (3/3) . log2(3/3) – (0/3) . log2(0/3) = 0

Entropy (Decision, Outlook= Rain |Wind = Strong) = – (0/2) . log2(0/2) – (2/2) . log2(2/2) = 0

Gain(Decision, Outlook= Rain |Wind) = 0.97 – (3/5)(0) – (2/5)(0) = 0.97

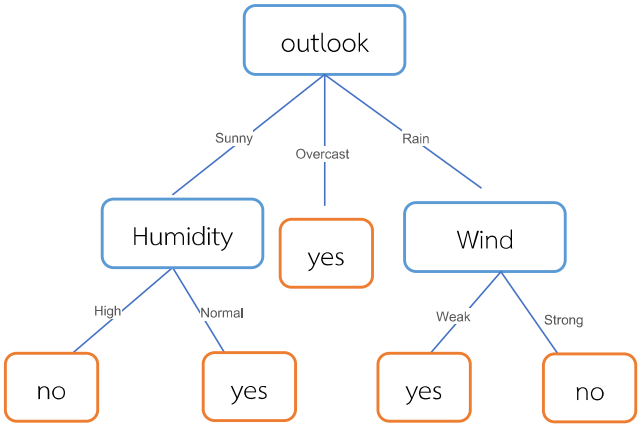
Wind จะสร้างคะแนนสูงสุดหากแนวโน้มเป็น Rain จึงต้องตรวจสอบแอตทริบิวต์ Wind ในระดับที่ 2 ว่าแนวโน้มมี Rain

Gain(Outlook=Rain | Temperature) = 0.01997309402197489

Gain(Outlook=Rain | Humidity) = 0.01997309402197489

Gain(Outlook=Rain | Wind) = 0.9709505944546686

Wind คือการตัดสินใจ เพราะมันให้คะแนนสูงสุดหากแนวโน้มมี Rain ณ จุดนี้ การตัดสินใจจะไม่เกิดขึ้นหากมี Strong ดังนั้น การสร้างต้นไม้เพื่อการตัดสินใจจึงสิ้นสุดลง เราสามารถใช้กฎต่อไปนี้ในการตัดสินใจ



ภาพประกอบที่ 2.6 โหนด Outlook

กฎที่ได้จากต้นไม้ตัดสินใจ

If (outlook==sunny & Humidity==high){ การตัดสินใจเล่นเทนนิสนอกบ้าน = no}

If (outlook==sunny & Humidity==normal){ การตัดสินใจเล่นเทนนิสนอกบ้าน = yes}

If (outlook==rain & Wind== Weak) { การตัดสินใจเล่นเทนนิสนอกบ้าน = yes}

If (outlook==rain & Wind== Strong) { การตัดสินใจเล่นเทนนิสนอกบ้าน = no}

If (outlook== Overcast) { การตัดสินใจเล่นเทนนิสนอกบ้าน = yes}

2.4 การวัดประสิทธิภาพในการจำแนกข้อมูล

การวัดประสิทธิภาพในการจำแนกสามารถวัดได้หลายวิธี เช่น ความเร็วในการทำนายข้อมูลของตัวจำแนก ความทนทานต่อการทำนายข้อมูลที่มีสิ่งรบกวนหรือการขาดหายไป ความยืดหยุ่นต่อปริมาณข้อมูล ความสามารถที่ตัวจำแนกสามารถเข้าใจได้ง่ายจากผู้ใช้งาน และความสามารถในการทำนาย เป็นต้น โดยวิธีการที่ได้รับความนิยมในการวัดประสิทธิภาพในการจำแนก คือ ความสามารถในการทำนาย ซึ่งนิยมวัดด้วย ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าระลึก (Recall) ค่าประสิทธิภาพโดยรวม (F-measure) ถ้าค่าดังกล่าวมีค่าสูง แสดงว่าตัวจำแนกมีประสิทธิภาพในการทำนายสูง ซึ่งค่าดังกล่าวสามารถคำนวณได้จากการพิจารณาค่าที่อยู่ในตารางเมทริกซ์ความสับสน (Confusion matrix) โดยค่าที่อยู่ในตารางเมทริกซ์ความสับสนเป็นค่าที่แสดงผลลัพธ์ในการทำนายของแต่ละกลุ่มหรือคลาส

ตารางที่ 2.21 ตารางเมทริกซ์ความสับสน

		ผลการทำนาย			
		C_1	C_2	...	C_n
ค่าความจริง	C_1	p_{11}	p_{12}	...	p_{1n}
	C_2	p_{21}	p_{22}	...	p_{2n}
	...				
	C_n	p_{n1}	p_{n2}	...	p_{nn}

โดยที่ Pij คือ จำนวนข้อมูลทำนายว่าเป็นคลาส Cj แต่คำตอบจริงเป็นคลาส Ci
สมมติให้ข้อมูลมีทั้งหมด 3 คลาส ข้อมูลจริงในแต่ละคลาสมีทั้งหมด 10ข้อมูล ผลการทำนายของตัวจำแนกแสดงในตารางเมทริกซ์ความสับสนดังตารางที่ 2.21

ตารางที่ 2.22 ผลการทำนาย

		ผลการทำนาย			รวม
		C_1	C_2	C_3	
ค่าความจริง	C_1	p_{11} = 7	P_{12} = 1	P_{13}= 2	10
	C_2	p_{21} = 1	P_{22} = 8	P_{23}= 1	10
	C_3	P_{31} = 2	P_{32} = 3	P_{33} = 5	10

จากตารางที่ 2.22 ค่าแต่ละค่ามีความหมายดังนี้

p_{11} = 7 หมายความว่า ทำนายถูกต้องว่าเป็นคลาส C_1จำนวน 7 ข้อมูล

p_{12} = 1 หมายความว่า ทำนายว่าเป็นคลาส C_2 จำนวน 1 ข้อมูล แต่จริงแล้วข้อมูลดังกล่าวเป็นคลาส c (แสดงว่าทำนายผิด)

p_{13} = 2 หมายความว่า ทำนายว่าเป็นคลาส C_3 จำนวน 2 ข้อมูล แต่จริงแล้วข้อมูลดังกล่าวเป็นคลาส C_1 (แสดงว่าทำนายผิด)

p_{21} = 1 หมายความว่า ทำนายว่าเป็นคลาส C_1 จำนวน 1 ข้อมูล แต่จริงแล้วข้อมูลดังกล่าวเป็นคลาส C_2 (แสดงว่าทำนายผิด)

p_{22} = 8 หมายความว่า ทำนายถูกต้องว่าเป็นคลาส C_2 จำนวน 8 ข้อมูล

p_{23} = 1 หมายความว่า ทำนายว่าเป็นคลาส C_3 จำนวน 1 ข้อมูล แต่จริงแล้วข้อมูลดังกล่าวเป็นคลาส C_2 (แสดงว่าทำนายผิด)

p_{31} = 2 หมายความว่า ทำนายว่าเป็นคลาส C_1 จำนวน 2 ข้อมูล แต่จริงแล้วข้อมูลดังกล่าวเป็นคลาส C_3 (แสดงว่าทำนายผิด)

p_{32} = 3 หมายความว่า ทำนายว่าเป็นคลาส C_2 จำนวน 3 ข้อมูล แต่จริงแล้วข้อมูลดังกล่าวเป็นคลาส C_3 (แสดงว่าทำนายผิด)

p_{33} = 5 หมายความว่า ทำนายถูกต้องว่าเป็นคลาส C_3 จำนวน 6 ข้อมูล

นิยามที่ 2.1 ค่าความถูกต้องเป็นค่าที่บ่งบอกประสิทธิภาพในการทำนายโดยรวม สามารถคำนวณได้ดังสมการที่ (2.1) ซึ่งหาได้จากจำนวนข้อมูลที่ทำนายถูกต้องทั้งหมดหารด้วยจำนวนข้อมูลทั้งหมดที่ใช้ในการทดสอบ

$$AC = \frac{\sum_{i=1}^{n} p_{ii}}{N} \qquad (2.1)$$

โดยที่ P_{ii} คือ จำนวนข้อมูลที่ทำนายถูกต้องว่าเป็นคลาส C_i

n คือ จำนวนคลาส

N คือ จำนวนข้อมูลทั้งหมดที่ใช้ในการทดสอบ

ตัวอย่างที่ 2.1 จากตารางที่ 2.22 ค่าความถูกต้องสามารถคำนวณได้ดังนี้

$$AC = \frac{p_{11} + p_{22} + p_{33}}{p_{11} + p_{12} + p_{13} + p_{21} + p_{22} + p_{23} + p_{13} + p_{23} + p_{33}}$$

$$AC = \frac{7 + 8 + 5}{7 + 1 + 2 + 1 + 8 + 1 + 2 + 3 + 5}$$

$$AC = \frac{20}{30} = 0.67$$

ดังนั้นสามารถสรุปได้ว่าค่าความถูกต้องในการทำนายของตัวจำแนก คือ 0.67 หรือ 67% นั้นเอง

นิยามที่ 2.2 ค่าความแม่นยำเป็นค่าที่แสดงให้เห็นถึงความแม่นยำในการทำนายแต่ละคลาส พิจารณาจากอัตราส่วนข้อมูลที่ทำนายถูกต้องว่าเป็นคลาส C_i ต่อจำนวนข้อมูลที่ทำนายว่าเป็นคลาส C_i ค่าความแม่นยำในการทำนายแต่ละคลาสสามารถคำนวณดังสมการที่ (2.2)

$$\text{precision}_{c_i} = \frac{P_{ii}}{\sum_{i=1}^{n} p_{ij}} \quad (2.2)$$

โดยที่ P_{ii} คือ จำนวนข้อมูลที่ทำนายถูกต้องว่าเป็นคลาส C_i

P_{ij} คือ จำนวนข้อมูลทำนายว่าเป็นคลาส C_j แต่คำตอบจริงเป็นคลาส C_i

n คือ จำนวนคลาส

ตัวอย่างที่ 2.2 จากตารางที่ 2.22 ค่าความแม่นยำของแต่ละคลาสสามารถคำนวณได้ดังนี้

$$precision_{C_1} = \frac{p_{11}}{p_{11} + p_{21} + p_{31}}$$

$$precision_{C_1} = \frac{7}{7+1+2} = \frac{7}{10} = 0.70$$

$$precision_{C_2} = \frac{p_{22}}{p_{12} + p_{22} + p_{32}}$$

$$precision_{C_2} = \frac{8}{1+8+3} = \frac{8}{11} = 0.73$$

$$precision_{C_3} = \frac{p_{33}}{p_{13} + p_{23} + p_{33}}$$

$$precision_{C_3} = \frac{5}{2+1+5} = \frac{5}{8} = 0.63$$

สามารถสรุปได้ว่า ค่าความแม่นยำในการทำนายคลาส C_1มีค่าเท่ากับ 0.70 หรือ 70%
ค่าความแม่นยำในการทำนายคลาส C_2 มีค่าเท่ากับ 0.73 หรือ 73%
ส่วนค่าความแม่นยำในการทำนายคลาส C_3 มีค่าเท่ากับ 0.63 หรือ 63%
แสดงให้เห็นว่าตัวจำแนกมีความแม่นยำในการทำนายคลาส C_2 มากที่สุด

นิยามที่ 2.3 ค่าระลึก หรือ ค่าความครบถ้วน แสดงถึงสามารถในการทำนายแต่ละคลาสว่ามีความถูกต้องเพียงใด โดยจะพิจารณาจากจำนวนข้อมูลที่ทำนายถูกต้องว่าเป็นคลาส C_i ต่อจำนวนข้อมูลจริงทั้งหมดที่เป็น C_i ค่าระลึกของแต่ละคลาสสามารถคำนวณดังสมการที่ (2.3)

$$\text{recall}_{c_i} = \frac{P_{ii}}{\sum_{j=1}^{n} p_{ij}} \quad (2.3)$$

โดยที่ P_{ii} คือ จำนวนข้อมูลที่ทำนายถูกต้องว่าเป็นคลาส C_i

P_{ij} คือ จำนวนข้อมูลทำนายว่าเป็นคลาส C_j แต่คำตอบจริงเป็นคลาส C_i

n คือ จำนวนคลาส

ตัวอย่างที่ 2.3 จากตารางที่ 2.22 ค่าระลึกของแต่ละคลาสสามารถคำนวณได้ดังนี้

$$recall_{C_1} = \frac{p_{11}}{p_{11} + p_{12} + p_{13}}$$

$$recall_{C_1} = \frac{7}{7+1+2} = \frac{7}{10} = 0.70$$

$$recall_{C_2} = \frac{p_{22}}{p_{21} + p_{22} + p_{23}}$$

$$recall_{C_2} = \frac{8}{1+8+1} = \frac{8}{10} = 0.80$$

$$recall_{C_3} = \frac{p_{22}}{p_{31} + p_{32} + p_{33}}$$

$$recall_{C_3} = \frac{5}{2+3+5} = \frac{5}{10} = 0.50$$

สามารถสรุปได้ว่า ค่าระลึกในการทำนายคลาส C_1 มีค่าเท่ากับ 0.70 หรือ 70%
ค่าระลึกในการทำนายคลาส C_2 มีค่าเท่ากับ 0.80 หรือ 80%
ส่วนค่าระลึกในการทำนายคลาส C_3 มีค่าเท่ากับ 0.50 หรือ 50%
แสดงให้เห็นว่าตัวจำแนกมีความสามารถในการทำนายคลาส C_2 ได้ดีที่สุด

นิยามที่ 2.4 ค่าประสิทธิภาพโดยรวมเป็นค่าที่แสดงภาพรวมของค่าความแม่นยำและความระลึก สามารถคำนวณได้ดังสมการที่ (2.4)

$$\text{F - Measure}_{C_i} = \frac{2x\ \text{Precision}_{C_i} \text{ x Recall}_{C_i}}{\left(\text{Precision}_{C_i} + \text{Recall}_{C_i}\right)} \qquad (2.4)$$

ตัวอย่างที่ 2.4 จากตารางที่ 2.22 ค่าประสิทธิภาพโดยรวมของแต่ละคลาสสามารถคำนวณได้ดังนี้

$$\text{F - Measure}_{C_1} = \frac{2x\ \text{Precision}_{C_1} \text{ x Recall}_{C_1}}{\left(\text{Precision}_{C_1} + \text{Recall}_{C_1}\right)}$$

$$F\text{-}\ Measure_{C_1} = \frac{2x\ 0.7x0.7}{(0.7 + 0.7)} = 0.70$$

$$F\text{-}\ Measure_{C_2} = \frac{2x\ Precision_{C_2}\ x\ Recall_{C_2}}{(Precision_{C_2} + Recall_{C_2})}$$

$$F\text{-}\ Measure_{C_2} = \frac{2x\ 0.73x0.80}{(0.73 + 0.80)} = 0.76$$

$$F\text{-}\ Measure_{C_3} = \frac{2x\ Precision_{C_3}\ x\ Recall_{C_3}}{(Precision_{C_3} + Recall_{C_3})}$$

$$F\text{-}\ Measure_{C_3} = \frac{2x\ 0.63x0.50}{(0.63 + 0.50)} = 0.56$$

สามารถสรุปได้ว่า ค่าประสิทธิภาพโดยรวม ในการทำนายคลาส C_1มีค่าเท่ากับ 0.70 หรือ 70%
ค่าประสิทธิภาพโดยรวมในการทำนายคลาส C_2 มีค่าเท่ากับ 0.76 หรือ 76%
ส่วนค่าประสิทธิภาพโดยรวมในการทำนายคลาส C_3 มีค่าเท่ากับ 0.56 หรือ 56%
แสดงให้เห็นว่าตัวจำแนกมีประสิทธิภาพโดยรวมในการทำนายคลาส C_2 มากที่สุด

2.5 ระบบงานที่เกี่ยวข้อง

2.5.1 Thai CV risk score

แอพพลิเคชันนี้ทำขึ้นเพื่อใช้ประเมินความเสี่ยงต่อการเกิดโรคหัวใจและหลอดเลือด โดยแสดงผลการประเมินเป็นความเสี่ยงต่อการเสียชีวิตหรือเจ็บป่วยจากโรคเส้น เลือดหัวใจตีบตัน และโรคเส้นเลือดสมองตีบตันในระยะเวลา 10 ปีข้างหน้า ซึ่งสามารถใช้ได้ทั้งในกรณีที่ท่านไม่มีผลเลือด โดยให้ใช้ขนาดรอบเอวหรือขนาด รอบเอวหารด้วยส่วนสูงแทน และในกรณีที่มีผลการตรวจระดับไขมันในเลือด แบบประเมินนี้สร้างขึ้นจากการติดตามศึกษาหาปัจจัยเสี่ยงต่อการเกิดโรคหัวใจ และหลอดเลือดในประชากรไทยภายใต้โครงการศึกษาพนักงานการไฟฟ้าฝ่ายผลิตแห่ง ประเทศไทย เป็นระยะเวลายาวนานกว่า 20 ปี แบบประเมินความเสี่ยงนี้จึงควรใช้เฉพาะในคนไทยที่มีอายุ 35-70 ปี ยังไม่มีโรคหัวใจและหลอดเลือด หากท่านมีข้อสงสัยหรือไม่แน่ใจแนะนำให้เข้ารับการประเมินโดยแพทย์ผู้เชี่ยวชาญ

ภาพประกอบที่ 2.7 แอพพลิเคชัน Thai CV risk score

2.5.2 การประยุกต์ใช้เทคนิคจำแนกข้อมูลแบบต้นไม้ตัดสินใจ

เพื่อการวินิจฉัยโรคในโคเบื้องต้นบนโทรศัพท์มือถือวัตถุประสงค์ของงานวิจัยนี้คือ

1) เพื่อพัฒนาโมเดลการวินิจฉัยโรคในโคเบื้องต้นโดยประยุกต์ใช้เทคนิคจำแนกข้อมูลแบบต้นไม้ตัดสินใจ
2) เพื่อพัฒนาแอพพลิเคชั่นการวินิจฉัยโรคในโคเบื้องต้นบนโทรศัพท์มือถือและ
3) เพื่อประเมินความพึงพอใจของผู้ใช้งานแอพพลิเคชั่นบนโทรศัพท์มือถือ โดยทำการรวบรวมข้อมูลปัจจัยที่เกี่ยวข้องกับการวินิจฉัยโรคโนโคจากกลุ่มเกษตรกรผู้เลี้ยงโคและผู้เชี่ยวชาญในเขตจังหวัดพิษณุโลกสร้างโมเดลวินิจฉัยโรคใช้เทคนิคต้นไม้ตัดสินใจเปรียบเทียบอัลกอริทึมจำนวน 3 อัลกอริทึม ได้แก่ J48 Random Tree และ REPTree แล้วทำการ ทดสอบประสิทธิภาพโมเดลด้วยวิธีการตรวจสอบแบบไขว้ เพื่อที่จะหาโมเดลการวินิจฉัยโรคที่มีประสิทธิภาพดีที่สุด

2.5.3 การศึกษาพฤติกรรมผู้ใช้งานตู้แช่เชิงพาณิชย์ด้วยเทคนิคต้นไม้ตัดสินใจ

เทคนิค Decision Tree Learning เพื่อศึกษาพฤติกรรมการใช้งานตู้แช่เชิงพาณิชย์ เพื่อพัฒนาระบบ FTC เดิมให้มีประสิทธิภาพมากขึ้น โดยการนำเอาระบบ FTCเดิมมาเปรียบเทียบกับระบบ FTC Behavior (อุปกรณ์ควบคุมอุณหภูมิตู้แช่อาดุยโน่โดยเพิ่มพฤติกรรมผู้ใช้งาน) และทำให้มีค่าใช้จ่ายในค่าไฟฟ้าลดลง โดยมีการเก็บข้อมูลแบบแยกหมวดหมู่เพื่อหาค่ามาตรฐานมาใช้

กำหนดเวลาและจะมีระบบเตือนผู้ใช้งานในกรณีที่เปิดตู้เย็นเกินเวลาที่กำหนด เมื่อนำระบบ FTC เปรียบเทียบกับระบบ FTC Behavior (อุปกรณ์ควบคุมตู้แช่แบบเพิ่มพฤติกรรมผู้ใช้งาน) จะพบว่าจากการทดลองจริงพฤติกรรมการใช้งานเปิดปิดตู้แช่จริงลดลง 18.75% และมีค่าไฟฟ้าที่ลดลง 4.4% ในการวิจัยเรื่องการศึกษาพฤติกรรมผู้ใช้งานตู้แช่เชิงพาณิชย์ผู้วิจัยได้กำหนดวัตถุประสงค์ของโครงการวิจัยดังนี้

1) เพื่อศึกษาพฤติกรรมการใช้งานการเปิดปิดเครื่องใช้ไฟฟ้าประเภทตู้แช่โดยใช้เทคนิค Decision Tree Learning

2) เพื่อพัฒนาระบบ FTC ให้เป็นแบบใหม่โดยบวกส่วนที่เป็นพฤติกรรมการใช้งาน เพื่อการประหยัดพลังงานไฟฟ้า

2.5.4 การพัฒนาโปรแกรมสำเร็จรูปเพื่อสนับสนุนการตัดสินใจสำหรับการประกอบการธุรกิจหอพัก

การพัฒนาโปรแกรมสำเร็จรูปมาใช้เพื่อสนับสนุนการตัดสินใจ (ประสงค์ ประณีตพลกรัง, 2548) ในการประกอบการธุรกิจเกี่ยวกับหอพักในเขตบริเวณรอบๆ มหาวิทยาลัยราชภัฏร้อยเอ็ด โดยใช้เทคนิคเหมืองข้อมูล (Data Mining) ในการสร้างแบบจำลองต้นไม้ตัดสินใจ (Decision Tree) เพื่อต้องการช่วยให้ผู้ประกอบการ/เจ้าของหอพักสามารถนำมาใช้ในการทำนายแนวโน้มหรือใช้ในการตัดสินใจ ในการประกอบการธุรกิจหอพักว่าจะประสบผลสำเร็จหรือไม่ งานวิจัยมีวัตถุประสงค์ดังนี้

1. เพื่อหาความสัมพันธ์ของปัจจัยต่างๆ ที่ทำให้การประกอบการธุรกิจหอพักประสบผลสำเร็จตามเป้าหมายของผู้ประกอบการธุรกิจหอพัก

2. เพื่อพัฒนาโปรแกรมสำเร็จรูปที่ใช้ในการสนับสนุนการตัดสินใจ โดยใช้แบบจำลองต้นไม้ตัดสินใจสำหรับการประกอบการธุรกิจหอพัก

2.5.5 ระบบสนับสนุนการตัดสินใจเพื่อวินิจฉัยโรคใบลำไยด้วยเทคนิคต้นไม้ตัดสินใจ

ระบบสนับสนุนการตัดสินใจเพื่อวินิจฉัยโรคใบลำไยด้วยเทคนิคต้นไม้ตัดสินใจ เป็นระบบที่พัฒนาขึ้นมีลักษณะเป็นเว็บเบสแอพพลิเคชัน โดยมีการสร้างและประเมินตัวแบบวินิจฉัย ซึ่งผลการประเมินพบว่าให้ค่าความถูกต้องในการประมวลผลเพื่อวินิจฉัยโรคใบลำไยมีค่าเท่ากับ ร้อยละ 85.3 หลังจากนั้นจะนำไปพัฒนาระบบสนับสนุนการตัดสินใจ โดยใช้ภาษา PHP ในการเขียนเว็บแอพพลิเคชั่น และใช้โปรแกรมฐานข้อมูล MySQL พบว่า ระบบที่พัฒนามีผลการประเมินประสิทธิภาพของระบบโดยผู้เชี่ยวชาญโดยรวมอยู่ในระดับดี (X = 4.10,S.D. = 0.51) และผู้ใช้งานระบบมีประสิทธิภาพโดยรวมอยู่ในระดับดี (X = 3.99, S.D. = 0.60) จากการสร้างตัวแบบวินิจฉัยโรคใบลำไยโดยใช้อัลกอรึทึม C5.0 พบว่า ให้ค่าความถูกต้องในการประมวลผลเพื่อวินิจฉัยโรคใบลำไยมีค่าเท่ากับ ร้อยละ 85.3 โดยมีผลที่ได้ค่อนข้างสูง ทั้งๆ ที่อัลกอริทึม C5.0 นี้เป็นอัลกอริทึมที่ดีและเหมาะสมที่สุดในการใช้เทคนิคต้นไม้ตัดสินใจ ไม่ว่าจะเป็นการสร้างกฎที่ทำได้อย่างรวดเร็วมีการใช้หน่วยจำที่มีขนาดเล็ก รวมทั้งมีค่าความผิดพลาดต่ำ (Upadhyay, Shukla, and Kumar, 2013) แต่

อาจเป็นเพราะปัจจัยที่ใช้ในการวินิจฉัยยังไม่ครบถ้วนสมบูรณ์ ดังนั้นควรจะมีการศึกษาข้อมูลปัจจัยที่มีผลต่อการวินิจฉัยโรคใบลำไยเพิ่มขึ้น เพื่อให้ค่าความถูกต้องในการวินิจฉัยโรคใบลำไยมีค่าความถูกต้องมากยิ่งขึ้นต่อไป ส่วนการพัฒนาระบบงาน พบว่าประสิทธิภาพโดยรวมอยู่ในระดับดี จากการประเมินโดยผู้เชี่ยวชาญและผู้ใช้งาน เนื่องจากมีการออกแบบระบบที่ง่ายต่อการใช้งาน และตรงตามความต้องการของผู้ใช้งาน ระบบที่พัฒนาขึ้นนี้มีความเหมาะสม สามารถนำไปใช้งานได้จริง สามารถช่วยวินิจฉัยโรคที่เกิดขึ้นกับใบลำไย และแนวทางการรักษาโรคใบได้

ตารางที่ 2.23 ตารางเปรียบเทียบระบบงานที่เกี่ยวข้อง

ฟังก์ชันการทำงาน	**Thai CV risk score มหาลัยมหิดล**	**ระบบที่พัฒนา**
ล็อคอินเข้าสู่ระบบ	X	/
เพิ่ม ลบ แก้ไข ผู้ดูแล	X	/
สามารถเพิ่มไฟล์โมเดล	X	/
ประชาสัมพันธ์	X	/
ดูโมเดลที่สร้างขึ้น	X	/
กรอกข้อมูลการทำนาย	/	/
ดูผลการทำนาย	/	/
ใช้ผลเลือดมาตรวจสอบ	/	X