บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยเรื่อง การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อการพยากรณ์ผลผลิตมันสำปะหลังในเขตภาคเหนือตอนล่าง ผู้วิจัยได้นำแนวคิดและงานวิจัยที่เกี่ยวข้องมาประยุกต์ใช้ ดังนี้

1. แนวคิดเกี่ยวกับมันสำปะหลัง
2. แนวคิดและทฤษฎีเกี่ยวกับการพยากรณ์
3. แนวคิดและทฤษฎีเกี่ยวกับการทำเหมืองข้อมูล
4. แนวคิ์ดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบต้นไม้ตัดสินใจ
5. แนวคิดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบเบย์
6. แนวคิดและทฤษฎีเทคนิคการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ
7. แนวคิดและทฤษฎีตัววัดประสิทธิภาพตัวแบบพยากรณ์
8. แนวคิดเกี่ยวกับโปรแกรมเวกา (WEKA)
9. แนวคิดเกี่ยวกับหลักการพัฒนาระบบ
10. งานวิจัยที่เกี่ยวข้อง

แนวคิดเกี่ยวกับมันสำปะหลัง

การศึกษาแนวคิดเกี่ยวกับมันสำปะหลังประกอบด้วย ลักษณะทั่วไปและพันธุ์มันสำปะหลัง วิธีการปลูกและดูแลมันสำปะหลัง โรคและแมลงของมันสำปะหลัง และประโยชน์และการแปรรูปมันสำปะหลัง ซึ่งจากข้อมูลของสำนักงานเศรษฐกิจการเกษตร (สำนักงานเศรษฐกิจการเกษตร, 2552) สรุปว่าพืชเศรษฐกิจที่มีการเพาะปลูกและได้รับผลิตอยู่ใน 5 อันดับแรกของภาคเหนือตอนล่างรวมถึงมันสำปะหลังด้วย ซึ่งมีรายละเอียดดังนี้

ลักษณะทั่วไปและพันธุ์มันสำปะหลัง

มันสำปะหลังเป็นพืชมีดอกเป็นไม้พุ่มขนาดเล็กขยายพันธุ์ โดยปักชำลำต้นซึ่งตัดเป็นท่อนสั้น ๆ มีรากพองออกเป็นหัวเพื่อเก็บสะสมอาหารจำพวกแป้งขึ้นในที่ซึ่งมีอากาศร้อนเป็นที่ดอนน้ำท่วมได้ โดยทั่วไปเกษตรกรจะเก็บหัวมันสำปะหลังเมื่ออายุ 1 ปี มีผลผลิตประมาณ 4 - 5 ตันต่อไร่ ปลูกได้ตลอดปีทนความแห้งแล้งได้และให้ผลผลิตเร็ว (กรมวิชาการเกษตร, ม.ป.ป.)

พันธุ์ของมันสำปะหลัง ในประเทศไทยมีพันธุ์อยู่ 3 ประเภท ซึ่งมีรายละเอียดดังนี้

1. พันธุ์ที่ใช้หัวเป็นอาหารของมนุษย์ นิยมปลูกกันตามบ้านเพื่อใช้เป็นอาหารในครอบครัว เรียกกันว่าพันธุ์ห้านาที และมีรสดี ใช้เป็นอาหาร โดยปิ้ง เผา ต้ม และเชื่อม เป็นต้น

2. พันธุ์ที่ใช้หัวในอุตสาหกรรม ปลูกเป็นการค้าจำนวนมากนับล้านไร่ เพื่อส่งเข้าโรงงานทำแป้งหรือโรงงานมันเส้น เรียกว่า พันธุ์พื้นเมืองหรือพันธุ์ระยองเป็นพันธุ์ที่มีปริมาณแป้งในหัวมาก เจริญเติบโตได้ดีและผลผลิตสูงจึงปลูกเพื่อส่งเข้าโรงงาน

3. พันธุ์ที่ใช้เป็นไม้ประดับ นิยมปลูกกันตามบ้านเพื่อความสวยงาม เรียกว่า มันด่าง ใบจะมีสีเขียว ปนเขียวอ่อนและขาวสีต่างของใบ (สารานุกรมไทยสำหรับเยาวชนเล่ม 5, 2523)

ตาราง 1 พันธุ์มั่นสำปะหลังและลักษณะทั่วไป

พันธุ์	ลักษณะทั่วไป	การดูแลและเก็บเกี่ยว
ระยอง 1	ยอดสีม่วงใบที่เจริญเต็มที่สีเขียวปนม่วง ก้านใบสีเขียวปนม่วง แผ่นใบเป็นแบบใบหอกปลายมนมีแฉก	อายุเก็บเกี่ยว 12 เดือน ฤดูปลูกช่วงต้นฤดูฝนพฤษภาคมถึงมิถุนายน ปลายฤดูฝนกันยายนถึงตุลาคม
ระยอง 3	ลำต้นค่อนข้างเตี้ยหัวรวมกันแน่น เหมาะกับอุตสาหกรรมทำแป้งและอาหารสัตว์	อายุเก็บเกี่ยว 12 เดือน ไม่ควรปลูกช่วงฝนตกหนักหรือแล้งจัดจะมีโอกาสตายและผลผลิตต่ำ
ระยอง 5	ผลผลิตหัวสด แป้งและมันแห้งสูง ปรับตัวเข้ากับสภาพแวดล้อมได้ดี	ฤดูปลูกช่วงต้นฤดูฝนพฤษภาคมถึงมิถุนายน ปลายฤดูฝนกันยายนถึงตุลาคม ผลผลิตหัวสด 4.4 ตันต่อไร่
ระยอง 60	สะสมน้ำหนักหัวสดได้เร็ว มียอดอ่อนสีเขียวปนม่วง ก้านใบสีเขียว	อายุเก็บเกี่ยว 8 เดือน ผลผลิตหัวสด 3.15 ตันต่อไร่
ระยอง 90	เป็นพันธุ์ผสมระหว่างพันธุ์ CMC76 และพันธุ์ V43 ลำต้นมีลักษณะโค้ง	ฤดูปลูกช่วงต้นฤดูฝนพฤษภาคมถึงมิถุนายน ผลผลิตหัวสด 3.8 ตันต่อไร่
ระยอง 72	เป็นพันธุ์ที่เหมาะสมที่จะปลูกในภาคตะวันออกเฉียงเหนือ	ฤดูปลูกช่วงต้นฤดูฝนพฤษภาคมถึงมิถุนายน ปลายฤดูฝนกันยายนถึงตุลาคม ผลผลิตหัวสด 5.09 ตันต่อไร่

ตาราง 1 (ต่อ)

พันธุ์	ลักษณะทั่วไป	การดูแลและเก็บเกี่ยว
ระยอง 7	เป็นพันธุ์ที่มีผลผลิตสูงเหมาะสม สำหรับปลูกทั้งต้นและปลายฤดูฝน	ผลผลิตหัวสด 6.1 ตันต่อไร่ ปริมาณแป้งสูง ต้นพันธุ์คุณภาพดี ทนแล้ง
ระยอง 9	เป็นพันธุ์ที่มีผลผลิตแป้งและมันแห้งสูง และเหมาะในการผลิตเอทานอล	อายุเก็บเกี่ยว 8 เดือน ผลผลิตแป้ง 1.24 ตันต่อไร่ มันแห้ง 2.11 ตันต่อไร่
เกษตรศาสตร์ 50	เป็นพันธุ์ของม.เกษตรศาสตร์ ลำต้นโค้งเล็กน้อย สีเขียวเงิน	ผลผลิตหัวสด 4.4 ตันต่อไร่ ต้นพันธุ์เก็บได้นาน 30 วันหลังตัด
ห้วยบง 60	เป็นพันธุ์ผสมระยอง5 กับพันธุ์เกษตรศาสตร์ 50 ต้นสีเขียวเงิน	ผลผลิตหัวสดและแป้งสูง มีผลผลิต 5.8 ตันต่อไร่

ที่มา: กรมวิชาการเกษตร, ม.ป.ป.

วิธีการปลูกและดูแลมันสำปะหลัง

สภาพแวดล้อมที่เหมาะสมคือไม่มีน้ำท่วมขัง ดินร่วนปนทรายมีความอุดมสมบูรณ์ปานกลางปลูกได้ตลอดปีหลังเก็บเกี่ยวแล้วทำได้ โดยปลูกมากในเดือนเมษายนและพฤษภาคม ตามผลการทดลองของกรมวิชาการเกษตรพบว่าปลูกในเดือนมิถุนายนถึงกรกฎาคมจะให้ผลผลิตสูงและแต่ละภาคมีช่วงเวลาในการปลูกคือ ภาคเหนือตอนบนช่วงปลายมิถุนายน โดยภาคเหนือตอนล่างช่วงต้น-กลางกรกฎาคม มีระยะปลูก 80x80 เซนติเมตร จำนวนต้น 1,600 - 2,500 ต้นต่อไร่ ท่อนพันธุ์ยาว 20 เซนติเมตร มีจำนวนตาไม่น้อยกว่า 5 ตา หลังจากปลูกประมาณ 1 เดือนต้องกำจัดวัชพืชเป็นครั้งแรก ควรให้ปุ๋ยมีสูตร 15-7-18 อัตรา 70 กิโลกรัมต่อไร่ เมื่ออายุ 1 ปี ผลผลิตประมาณ 4 - 5 ตันต่อไร่ ถ้าปลูกติดต่อกันหลายปี โดยไม่ใส่ปุ๋ยผลผลิตในปีหลังจะลดลงเหลือประมาณ 2 ตันต่อไร่ ดังนั้นเพื่อให้ได้ผลผลิตจึงควรใส่ปุ๋ยทุกครั้งที่ปลูกและเก็บเกี่ยวตั้งแต่อายุ 8 เดือน แต่อายุที่เหมาะสมคือ 12 เดือน และหลังปลูกไม่ควรเก็บเกี่ยวในช่วงที่มีฝนชุก หลังจากเก็บเกี่ยวควรปล่อยให้ใบและยอดมันสำปะหลังคลุมดิน เพื่อเป็นปุ๋ยพืชสดและควรนำผลผลิตหัวมันสดส่งโรงงานทันที ไม่ควรเก็บไว้เกิน 2 วัน เพราะจะเน่าเสีย (กรมวิชาการเกษตร, ม.ป.ป.)

โรคและแมลงของมันสำปะหลัง

โรคของมันสำปะหลังมีด้วยกันหลายประเภท เช่น โรคใบจุดเกิดจากเชื้อราพบอยู่ทั่วไปในต้นที่ปลูกเกิดขึ้นที่ใบจะเป็นแผลจุดกลม ๆ โรคใบไหม้อาการเริ่มแรกใบจุดเหลี่ยม ฉ่ำน้ำ ใบไหม้ โรคลำต้นเน่าที่เกิดจากเชื้อรา เกษตรกรนิยมเก็บเกี่ยวผลผลิตในช่วงฤดูแล้งทำให้ต้องเก็บต้นพันธุ์ไว้รอเวลาปลูกที่เหมาะสมเป็นเวลานาน ทำให้เกิดต้นเน่าได้หรือบางปีสภาพอากาศแห้งแล้งมาก มันสำปะหลังทิ้งใบเป็นเวลานาน ทำให้พบอาการต้นแห้งจากปลายลงมามีอาการยืนตายส่วนแมลงที่สำคัญ เช่น ไรแดงระบาดอยู่ทั่วไปในประเทศ แต่ไม่ทำความเสียหายมากจะระบาดในระยะฤดูแล้งและอากาศร้อน ส่วนแมลงศัตรูชนิดอื่น ๆ เช่น เพลี้ยแป้ง เพลี้ยหอย และปลวก จะพบแต่ทำความเสียหายให้มันสำปะหลังในบางท้องที่เท่านั้น (สารานุกรมไทยสำหรับเยาวชนเล่ม 5, 2523)

ประโยชน์และการแปรรูปมันสำปะหลัง

หัวมันสำปะหลังเป็นที่สะสมแป้งจึงเป็นอาหารประเภทแป้งที่ให้พลังงานสำหรับมนุษย์และสัตว์ได้เป็นดี โดยแป้งมันสำปะหลังใช้เป็นอาหารของมนุษย์โดยตรงและใช้ในอุตสาหกรรม เช่น การทำกาว การทำกระดาษ การทอผ้า การผลิตน้ำตาลกลูโคส เป็นต้น ใช้หมักทำแอลกอฮอล์ เบียร์ และขนมปัง ใช้เป็นอาหารสัตว์ โดยทำเป็นมันเส้น และกากมันสำปะหลังซึ่งใช้เป็นแหล่งพลังงานผสมในอาหารสัตว์และนำมาแปรรูปทำแป้งและแปรรูปเป็นอาหารสัตว์ โรงงานมันสำปะหลังเส้น ใช้หัวมันสำปะหลังสดเป็นวัตถุดิบส่วนน้อยเท่านั้นที่ส่งออกจำหน่ายต่างประเทศในรูปมันสำปะหลังเส้น โรงงานมันสำปะหลังอัดเม็ด ใช้มันเส้นที่ซื้อมาหรือผลิตขึ้นเป็นวัตถุดิบในการผลิตมันอัดเม็ด ผลิตภัณฑ์มันสำปะหลังที่ส่งจำหน่ายต่างประเทศส่วนใหญ่คือ มันสำปะหลังอัดเม็ด (สารานุกรมไทยสำหรับเยาวชนเล่ม 5, 2523)

จากการศึกษาแนวคิดเกี่ยวกับมันสำปะหลัง ทำให้ทราบถึงลักษณะต่างๆ ของมันสำปะหลัง ซึ่งผู้วิจัยได้นำข้อมูลมันสำปะหลังที่ได้นี้มาทำการศึกษาเพื่อที่จะนำไปพยากรณ์ข้อมูลผลผลิตมันสำปะหลังในเขตภาคเหนือตอนล่างได้และเพื่อพัฒนาเป็นความรู้ที่ก่อให้เกิดประโยชน์กับเจ้าหน้าที่สำนักงานเศรษฐกิจการเกษตรเขต 2 และผู้ที่สนใจได้นำไปใช้ให้เกิดประโยชน์ต่อไป

แนวคิดและทฤษฎีเกี่ยวกับการพยากรณ์

การศึกษาแนวคิดและทฤษฎีเกี่ยวกับการพยากรณ์ประกอบด้วย ความหมายของการพยากรณ์ การจัดหาข้อมูลในการพยากรณ์ องค์ประกอบของการพยากรณ์ ข้อคำนึงเกี่ยวกับข้อมูลในการพยากรณ์ และขั้นตอนในการพยากรณ์ ซึ่งมีรายละเอียดดังนี้

ความหมายของการพยากรณ์

ดุสิต ดวงมาตย์พล (2550) ได้อธิบายความหมายของการพยากรณ์ว่า เป็นการคาดคะเนลักษณะต่าง ๆ หรือเป็นศิลปะของการะประเมินความต้องการในอนาคตด้วยการคาดการณ์ล่วงหน้าโดยกำหนดเงื่อนไขหรือสภาวะเป็นการใช้ศาสตร์และศิลปะที่จะทำนายเหตุการณ์ในอนาคต การพยากรณ์จะทำได้โดยใช้ประวัติและข้อมูลที่ผ่านมาเพื่อนำมาคำนวณ โดยใช้การทำนายเพื่อช่วยในการตัดสินใจ

การจัดหาข้อมูลในการพยากรณ์

แหล่งที่มาของข้อมูลและจำนวนข้อมูลเป็นสิ่งที่จำเป็นในการพยากรณ์ในแต่ละตัวแบบการพยากรณ์ โดยสามารถจำแนกได้เป็น 2 ประเภท ดังนี้

1. แหล่งข้อมูลปฐมภูมิ (Primary Data) เป็นข้อมูลที่ยังไม่มีผู้ใดเก็บรวบรวมไว้เป็นระบบผู้ใช้จะต้องดำเนินการเพื่อรวบรวมข้อมูลเอง โดยทั่วไปการรวบรวมใช้วิธีทำการวิจัยอาจใช้เทคนิควิจัยเชิงคุณภาพ เพื่อทำการรวบรวมข้อมูล เช่น ใช้เทคนิคสนทนา เป็นการสอบถามเพื่อให้ได้ข้อมูลจากกลุ่มย่อยในคำตอบที่ต้องการ ส่วนการใช้เทคนิคเชิงปริมาณ เช่น การทำการวิจัยเชิงสำรวจโดยใช้แบบสอบถามเป็นเครื่องมือในการเก็บข้อมูล โดยการสัมภาษณ์ส่วนตัวหรือส่งแบบสอบถามไปทางไปรษณีย์

2. แหล่งข้อมูลทุติยภูมิ (Secondary data) เป็นแหล่งข้อมูลที่ได้จากทั้งภายในและภายนอกองค์กร ข้อมูลภายในเช่น ข้อมูลการขาย ข้อมูลกำไร และสินค้าคงคลัง เป็นต้น แต่การพยากรณ์ข้อมูลภายในแล้วข้อมูลสภาวะแวดล้อมภายนอก เช่น เศรษฐกิจ การเมือง และสังคม เป็นต้น สิ่งจำเป็นเพื่อใช้ในการพยากรณ์แหล่งข้อมูลทุติยภูมิมีทั้งจากหน่วยงานภาครัฐยังมีจากองค์กรต่าง ๆ รวมถึงองค์กรที่ไม่หวังผลกำไร

องค์ประกอบของการพยากรณ์

องค์ประกอบของการพยากรณ์ที่ได้ผลที่แม่นยำมีลักษณะคือ ต้องระบุวัตถุประสงค์ในการนำผลการพยากรณ์ไปใช้และช่วงเวลาที่การพยากรณ์จะคลอบคลุม เพื่อจะเลือกใช้วิธีการในการพยากรณ์ได้เหมาะสมและรวบรวมข้อมูลอย่างมีระบบถูกต้องตามความเป็นจริง เพราะคุณภาพของข้อมูลมีผลอย่างยิ่งต่อการพยากรณ์ นอกจากนี้ควรบอกข้อจำกัดและสมมติฐานที่ตั้งไว้ในการพยากรณ์ เพื่อผู้ที่จะนำผลการพยากรณ์ไปใช้จะได้ทราบถึงเงื่อนไขที่มีผลต่อค่าพยากรณ์และต้อง

ตรวจสอบความถูกต้องแม่นยำค่าพยากรณ์ที่ได้กับค่าจริงที่เกิดขึ้นเป็นระยะ เพื่อปรับวิธีการที่ใช้ในการคำนวณให้เหมาะสมเมื่อเวลาเปลี่ยนไป

ข้อคำนึงเกี่ยวกับข้อมูลในการพยากรณ์

ข้อคำนึงในการพยากรณ์จะประกอบไปด้วย ความถี่ของข้อมูลซึ่งจะขึ้นกับระยะเวลาของตัวแปรที่เก็บ โดยจะพิจารณาถึงความต้องการของข้อมูลว่าต้องการละเอียดแค่ไหนรวมถึงหน่วยในการเก็บข้อมูลและระดับความแม่นยำที่ต้องการเพราะมีผลต่อเวลาและค่าใช้จ่ายในการเก็บรวบรวมข้อมูล

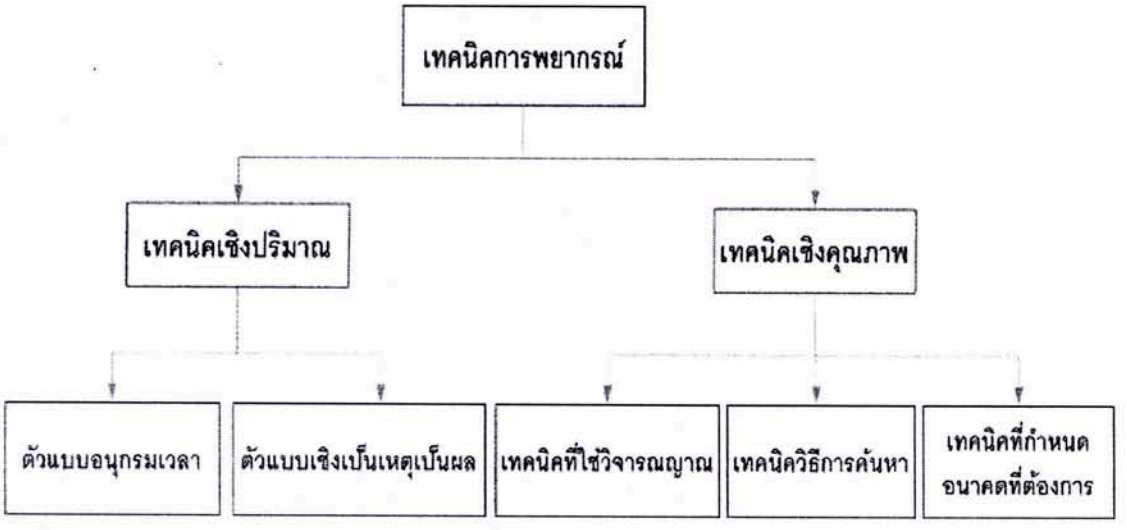
ขั้นตอนในการพยากรณ์

ขั้นตอนของระบบพยากรณ์จะมีการนำโมเดลมาใช้เพื่อหาเทคนิคและโมเดลที่เหมาะสมในการพยากรณ์ซึ่งมีขั้นตอนสำหรับการสร้างโมเดล โดยการเลือกโมเดลขึ้นมาเพื่อจัดการข้อมูลที่ได้จัดการรวบรวมไว้จากนั้นนำไปสู่การทำต้นแบบการประมาณเพื่อค้นหารูปแบบโมเดล แล้วตรวจสอบโมเดลว่ามีความเหมาะสมเพียงพอหรือไม่ ถ้ายังไม่เหมาะสมก็จะต้องกลับไปในขั้นตอนการระบุโมเดลใหม่อีกครั้ง แต่ถ้าโมเดลมีความเหมาะสมแล้วก็จะเข้าสู่ขั้นตอนในการพยากรณ์มีการนำข้อสังเกตใหม่เข้ามาใช้ในการตรวจสอบว่าเป็นโมเดลที่เหมาะสมหรือยัง สำหรับการพยากรณ์กับข้อสังเกตใหม่ ๆ ที่มีผลต่อโมเดลนั้นจะต้องกลับไปสู่ขั้นตอนในการเลือกโมเดลใหม่อีกครั้ง แต่ถ้าเป็นโมเดลที่เหมาะสมแล้วก็จะได้ระบบการพยากรณ์ที่สามารถนำมาใช้งานได้จริง

โดยวิธีการพยากรณ์ ออกได้เป็น 2 ประเภทใหญ่ ๆ ดังนี้

1. เทคนิคการพยากรณ์ทางด้านสถิติ ข้อมูลที่จะนำมาใช้ในกระบวนการพยากรณ์จะเป็นข้อมูลตัวเลข ข้อมูลที่ถูกรวบรวมไว้เป็นสถิติ

2. เทคนิคการพยากรณ์ทางด้านการเรียนรู้ ในการพยากรณ์มีหลายวิธี โดยจัดประเภทการพยากรณ์ได้ 2 ประเภท การพยากรณ์เชิงคุณภาพ การพยากรณ์เชิงปริมาณ ซึ่งทั้งสองวิธีนี้จะสามารถแบ่งได้อีกหลายวิธี ดังภาพ 2



ภาพ 2 เทคนิคการพยากรณ์ประเภทต่าง ๆ

เทคนิคการพยากรณ์ แบ่งได้เป็น 2 กลุ่มใหญ่ ๆ ดังนี้

1. การพยากรณ์เชิงคุณภาพ (Qualitative Forecasting Methods) ใช้คุณสมบัติเกี่ยวกับคุณภาพโดยใช้ความคิดเห็นของผู้เชี่ยวชาญไม่ใช้ตัวแบบทางคณิตศาสตร์ ข้อมูลที่นำมาประกอบการพยากรณ์อาจเป็นทั้งข้อมูลเชิงคุณภาพและเชิงปริมาณต้องอาศัยวิจารณญาณของผู้เชี่ยวชาญมาเป็นเครื่องมือเพื่อช่วยในการพยากรณ์ให้ถูกต้องมากยิ่งขึ้น

2. การพยากรณ์เชิงปริมาณ (Quantitative Forecasting Methods) เป็นการพยากรณ์ซึ่งใช้โมเดลทางคณิตศาสตร์หนึ่งอย่างหรือมากกว่าขึ้นอยู่กับข้อมูลในอดีตหรือตัวแปรด้านเหตุผลความต้องการพยากรณ์หรือเป็นวิธีการทางคณิตศาสตร์โดยการวิเคราะห์ตัวเลขในอดีตเพื่อพิจารณารูปแบบซึ่งใช้ในการคาดคะเนเหตุการณ์ในอนาคต

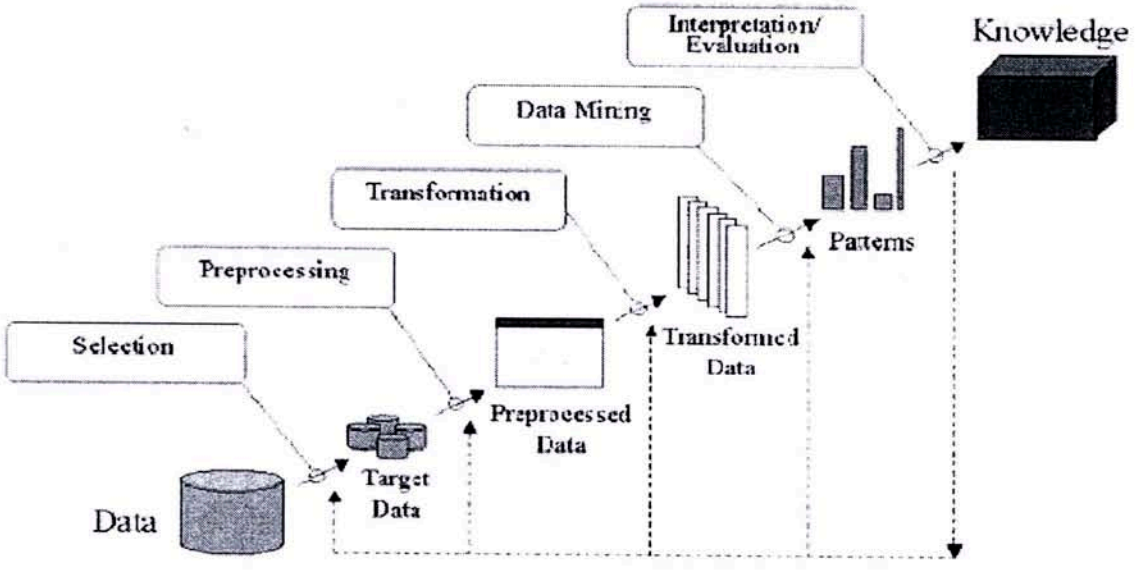
จากการศึกษาแนวคิดและทฤษฎีเกี่ยวกับการพยากรณ์ ผู้วิจัยทราบถึงความรู้ด้านการพยากรณ์ ซึ่งผู้วิจัยได้ใช้เทคนิคการพยากรณ์เชิงปริมาณ โดยใช้ตัวแบบเชิงเป็นเหตุเป็นผลแล้วนำแนวคิดดังกล่าวมาประยุกต์ใช้กับงานวิจัยในส่วนของการพยากรณ์ว่าจะต้องมีองค์ประกอบ เทคนิคการพยากรณ์ และมีขั้นตอนการพยากรณ์เป็นอย่างไร เป็นต้น เพื่อนำไปสู่ผลลัพธ์ที่มีความถูกต้อง

แนวคิดและทฤษฎีเกี่ยวกับการทำเหมืองข้อมูล

การศึกษาแนวคิดและทฤษฎีเกี่ยวกับการทำเหมืองข้อมูลประกอบด้วย ความหมายของเหมืองข้อมูล ขั้นตอนการทำเหมืองข้อมูล และเทคนิคเหมืองข้อมูล ซึ่งมีรายละเอียดดังนี้

ความหมายของเหมืองข้อมูล

ไพฑูรย์ จันทร์เรือง (2550) ได้กล่าวว่า การทำเหมืองข้อมูลคือ ขบวนการทำงานในการกลั่นกรองข้อมูลจากฐานข้อมูลขนาดใหญ่เพื่อให้ได้สารสนเทศที่ยังไม่รู้เป็นสารสนเทศที่สามารถนำไปใช้งานได้ซึ่งเป็นสิ่งสำคัญที่ช่วยตัดสินใจในการทำธุรกิจ โดยการทำเหมืองข้อมูลเป็นขั้นตอนที่สำคัญในการค้นหาองค์ความรู้ในฐานข้อมูล (Knowledge Discovery in Database) หรือ KDD เช่น การวิเคราะห์พฤติกรรมในการซื้อสินค้าของผู้บริโภค การพยากรณ์อากาศ ซึ่ง KDD ประกอบด้วยขั้นตอนต่าง ดังภาพ 3



ภาพ 3 ขั้นตอนของกระบวนการ KDD

ที่มา: Fayyad, 1996

ขั้นตอนการทำเหมืองข้อมูล

การทำเหมืองข้อมูลเป็นขั้นตอนหนึ่งที่สำคัญในกระบวนการค้นหาลักษณะแฝงของข้อมูลที่มีประโยชน์ในฐานข้อมูล ซึ่งกระบวนการ KDD ประกอบด้วยลำดับขั้นตอนดังต่อไปนี้

1. ขั้นตอนการคัดเลือกข้อมูล (Data Selection) คือ การเลือกเฉพาะข้อมูลที่ต้องการและเป็นประโยชน์ไปใช้ โดยการระบุแหล่งที่มาของข้อมูลและทำการดึงข้อมูลออกมาสำหรับการ

วิเคราะห์ซึ่งจะแตกต่างไปตามวัตถุประสงค์ที่ได้กำหนดและลักษณะงานที่จะนำไปใช้ โดย (ลลิตา ศรีชัยญา, 2553) ได้กล่าวถึงรูปแบบของข้อมูลว่ามี 2 ประเภท ดังนี้

1.1 ข้อมูลเชิงคุณภาพ (Qualitative) แบ่งเป็น 2 ชนิด

1.1.1 ข้อมูลลักษณะนามบัญญัติ (Nominal Attribute) เป็นข้อมูลที่ค่าเป็นไปได้ไม่มีลำดับ เช่น สถานการณ์แต่งงาน (โสด แต่งงาน หย่า) เพศ (ชาย หญิง) เป็นต้น

1.1.2 ข้อมูลชนิดลำดับ (Ordinal Attribute) เป็นข้อมูลที่ค่าเป็นไปได้มีลำดับ เช่น ลำดับลูกค้า (ดี ปานกลาง ไม่ดี) ระดับอุณหภูมิ (ต่ำ กลาง สูง) ระดับการศึกษา (ประถมกาศึกษา มัธยมศึกษา ปริญญาตรี ปริญญาโท) เป็นต้น

1.2 ข้อมูลเชิงปริมาณ (Quantitative) เป็นข้อมูลที่อยู่ในรูปตัวเลขที่แสดงถึงปริมาณของข้อมูลแบ่งเป็น 2 ชนิด ได้แก่

1.2.1 ข้อมูลชนิดตัวเลขต่อเนื่อง (Continuous) เช่น รายได้ เป็นต้น

1.2.2 ตัวเลขไม่ต่อเนื่อง (Discrete) เช่น จำนวนรถยนต์ จำนวนพนักงาน เป็นต้น

2. ขั้นตอนการเตรียมข้อมูล (Pre-processing) คือ การเตรียมข้อมูลให้เหมาะสมและให้อยู่ในรูปที่สามารถนำไปใช้งานได้ ในที่นี้จะรวมขั้นตอนการทำความสะอาดข้อมูลไว้ด้วย

การทำความสะอาดข้อมูล (Data Cleaning) เป็นการขั้นตอนที่แยกข้อมูลที่ไม่เกี่ยวข้อง ซ้ำซ้อน ออกไปทำให้ได้ข้อมูลที่มีคุณภาพนำไปวิเคราะห์ได้และเป็นการแก้ไขข้อมูลให้ถูกต้องสมบูรณ์ เช่น การแก้ไขค่าว่างของข้อมูลโดยอาจใส่ค่า 0 ลงไปหรืออาจไม่นำข้อมูลแถวนั้นมาใช้ในการประมวลผล ขึ้นอยู่กับการตัดสินใจของผู้ดูแลระบบ

3. ขั้นตอนการแปลงข้อมูล (Data Transformation) คือ การเปลี่ยนแปลงรูปแบบหรือปรับค่าของข้อมูล โดยการลดรูปและจัดข้อมูลให้อยู่ในรูปแบบเดียวกันเพื่อให้เหมาะสมสำหรับการนำไปใช้งานตามคุณสมบัติเฉพาะของแต่ละขั้นตอนวิธีที่ใช้ในการทำเหมืองข้อมูล เพื่อวิเคราะห์ตามอัลกอริทึมที่ใช้ในการทำเหมืองข้อมูลต่อไป

การจัดข้อมูลให้เป็นรูปแบบมาตรฐานเดียวกัน (Normalization) เป็นการแปลงข้อมูล โดยจะเปลี่ยนค่าข้อมูลให้อยู่ในรูปมาตรฐานอันดับที่กำหนด โดยการแปลงข้อมูลที่ผู้วิจัยจะกล่าวถึงในที่นี้ด้วยกัน 2 วิธี

3.1 การหาค่าแบบ Min-Max Normalization ซึ่งเป็นเทคนิคที่ต้องรู้ค่าสูงสุดและค่าต่ำสุดของข้อมูลจะทำให้ข้อมูลที่ได้นั้นอยู่ในช่วง 0 และ 1 ดังสมการ (1)

$$X^* = \frac{X - \min}{\max - \min} \quad (1)$$

จากสมการแทนค่าดังนี้

X^* คือค่าที่ได้จากการแปลงค่าที่ได้อยู่ในรูปมาตรฐาน

X คือค่าปัจจุบันที่นำมาทำเป็นค่าในรูปมาตรฐาน

min คือค่าข้อมูลที่มีค่าน้อยที่สุดในชุดข้อมูล

max คือค่าข้อมูลที่มีค่ามากที่สุดในชุดข้อมูล

3.2 วิธีที่ทำให้อยู่ให้รูปคะแนนมาตรฐาน Z (Z-scores) โดยต้องทราบค่าเฉลี่ย (μ) และค่าส่วนเบี่ยงเบนมาตรฐาน (σ) ดังสมการ (2)

$$X^* = \frac{X - \mu}{\sigma} \quad (2)$$

ซึ่งขั้นตอนต่าง ๆ ที่กล่าวมาจะเป็นตัวกำหนดความถูกต้องและขอบเขตวัตถุประสงค์ที่ต้องการของฐานความรู้ได้ เนื่องจากถ้าเลือกข้อมูลไม่ดี ข้อมูลมีขยะมาปะปนมีการแปลงข้อมูลไม่ถูกต้องก็จะทำให้ฐานความรู้ที่ได้ไม่ถูกต้องตามไปด้วย (ดิษฐพล มั่นธรรม, 2553)

4. ขั้นตอนเหมืองข้อมูล (Data Mining) คือ การเลือกเทคนิคที่เหมาะสมสำหรับงานที่ต้องการ โดยสามารถใช้เทคนิคมากกว่าหนึ่งเทคนิคมาประมวลผลได้ เพื่อดึงความรู้หรือสิ่งที่สนใจจากข้อมูล โดยผลลัพธ์ที่ได้จากขั้นตอนนี้คือ ฐานความรู้ นอกจากนี้ประเภทของงานตามลักษณะของแบบจำลองที่ใช้ในการทำเหมืองข้อมูลสามารถแบ่งกลุ่มได้เป็น 2 ประเภทใหญ่ ๆ คือ

4.1 การคาดคะเนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น โดยใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต

4.2 การหาแบบจำลองเพื่ออธิบายลักษณะบางอย่างของข้อมูลที่มีอยู่ซึ่ง โดยส่วนมากจะเป็นลักษณะการแบ่งกลุ่มให้กับข้อมูล

5. ขั้นตอนการแปลความ (Interpretation) คือการนำฐานความรู้ที่ได้จากขั้นตอนเหมืองข้อมูลมาทดสอบ พิจารณาว่าถูกต้องตามความต้องการและมีประโยชน์หรือไม่ โดยอาจต้องทำการ

ปรับแก้ค่าตัวแปรเพื่อกลับสู่ขั้นตอนเหมืองข้อมูลใหม่ หรือต้องเลือกข้อมูลชุดใหม่ แม้กระทั่งทำความสะอาดข้อมูลใหม่อีกครั้ง หลังจากตรวจสอบความรู้จนถูกต้องแล้วจึงนำความรู้นี้ไปใช้ในการสืบค้นความรู้ การวิเคราะห์ หรือการทำนายข้อมูลในอนาคตได้ต่อไป

ในการศึกษานี้ผู้วิจัยสนใจที่จะเลือกประยุกต์ใช้เทคนิคการจำแนกข้อมูล (Data Classification) เพื่อพยากรณ์ผลผลิตมันสำปะหลัง โดยค่าพยากรณ์ที่ได้นั้นจะอยู่ในลักษณะของค่าระดับ ซึ่งการจำแนกข้อมูลเป็นเทคนิคหนึ่งที่นำมาใช้สร้างตัวจำแนกกลุ่มหรือคลาส (Class) ของข้อมูล ซึ่งตัวจำแนกที่ได้สามารถนำไปทำนายกลุ่มของข้อมูลต่อไป โดยหลักการทำงานของเทคนิคนี้ จะนำข้อมูลส่วนหนึ่งมาสอนระบบให้สามารถเรียนรู้ (Training) เพื่อให้ระบบจำแนกข้อมูลนั้นออกมาเป็นกลุ่มต่าง ๆ ตามที่กำหนดไว้ผลที่ได้จากการเรียนรู้คือตัวจำแนกกลุ่มข้อมูล (Classifier) เมื่อมีข้อมูลใหม่เข้ามาสู่ระบบ ตัวจำแนกกลุ่มข้อมูลจะสามารถทำนายกลุ่มของข้อมูลที่ควรจะเป็นอัตโนมัติ

การเรียนรู้ (Learning) แบ่งออกเป็น 2 แบบ ด้วยกันคือ

1. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการเรียนรู้เพื่อค้นหาลักษณะบางอย่างที่เหมือนกันของข้อมูลแต่ละรายการ โดยที่ไม่ทราบชัดเจนว่าข้อมูลรายการนั้นถูกจัดอยู่ในกลุ่มใดหรือมีทั้งหมดกี่กลุ่ม โดยกระบวนการเรียนรู้จะพยายามที่จะหารูปแบบความสัมพันธ์ภายในชุดข้อมูลตัวอย่างแล้วทำการวิเคราะห์ว่าควรจะจัดข้อมูลออกเป็นกี่กลุ่มและข้อมูลใดควรจะอยู่ในกลุ่มใดแล้วทำการจัดกลุ่มข้อมูล (Clustering) (Witten and Frank, 2005)

2. การเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการเรียนรู้แบบมีเป้าหมายชัดเจนว่าต้องการศึกษาหรือจำแนกกลุ่มข้อมูลออกเป็นอย่างไร โดยนำเทคนิคและอัลกอริทึมต่าง ๆ มาใช้เพื่อจำแนกข้อมูลประมาณค่าของข้อมูล หรือทำนายลักษณะข้อมูล จากความสัมพันธ์ของสิ่งต่าง ๆ ที่เกี่ยวข้องซึ่งกระบวนการเรียนรู้จะพยายามหาแบบจำลองของตัวจำแนก (Classifier) ในการแบ่งข้อมูลจากชุดของตัวอย่างออกเป็นกลุ่ม เพื่อให้เกิดความผิดพลาดน้อยที่สุด เมื่อนำไปใช้กับข้อมูลจริงที่ไม่ได้อยู่ในชุดข้อมูลตัวอย่าง โดยชุดข้อมูลตัวอย่างจำเป็นที่จะต้องทราบและได้รับการกำหนดว่าอยู่กลุ่มใดก่อนเข้ากระบวนการเรียนรู้สำหรับการจำแนกข้อมูล (Data Classification) (Witten and Frank, 2005)

กลไกการจำแนกจัดเป็นรูปแบบหนึ่งของเหมืองข้อมูลซึ่งมีด้วยกันหลายวิธีและหลายอัลกอริทึม ในที่นี้ขอยกตัวอย่างของกฎการจำแนก (Classification Rules) ที่เป็นกระบวนการในการจำแนกข้อมูลขึ้นอยู่กับลักษณะของวัตถุประสงค์ของการจำแนก ในด้านการทำเหมืองข้อมูลลักษณะนี้เป็นการวิเคราะห์เซตของกลุ่มข้อมูลที่ยังไม่จัดแบ่งประเภทเพื่อสร้างโมเดลหรือฟังก์ชัน

ออกเป็นชุดข้อมูล (Class) ซึ่งลักษณะของคลาสจะถูกอธิบายโดยกลุ่มของคุณสมบัติ (Attribute) และกลุ่มของข้อมูลเรียนรู้ (Training data set) ที่ใช้ในการสร้างกฎการจำแนก

รูปแบบการเขียนกฎจำแนก

P_1, P2 , ...Pr ⇨ Ci Pr = เงื่อนไขของกฎ

เงื่อนไขของคุณสมบัติ ชุดข้อมูล

C i = ค่าต่าง ๆ ตามชุดข้อมูล

หรือ

IF <เงื่อนไข> THEN <ชุดข้อมูล> "ถ้า <เงื่อนไข> แล้ว <คลาส>"

ตัวอย่างการสร้างกฎสำหรับการวิเคราะห์กลุ่มลูกค้าว่าควรได้รับเครดิตหรือไม่ ในที่นี้ มีค่าคำตอบที่เป็นไปได้ 2 คำตอบ คือ Yes คือกลุ่มลูกค้าที่ควรได้รับเครดิต และ No คือกลุ่มลูกค้าที่ไม่ควรได้รับเครดิตดังต่อไปนี้

R1: Age>25, YR_Work>5 ⇨ Yes

R2: Sex = Male, YR_Work>2 ⇨ Yes

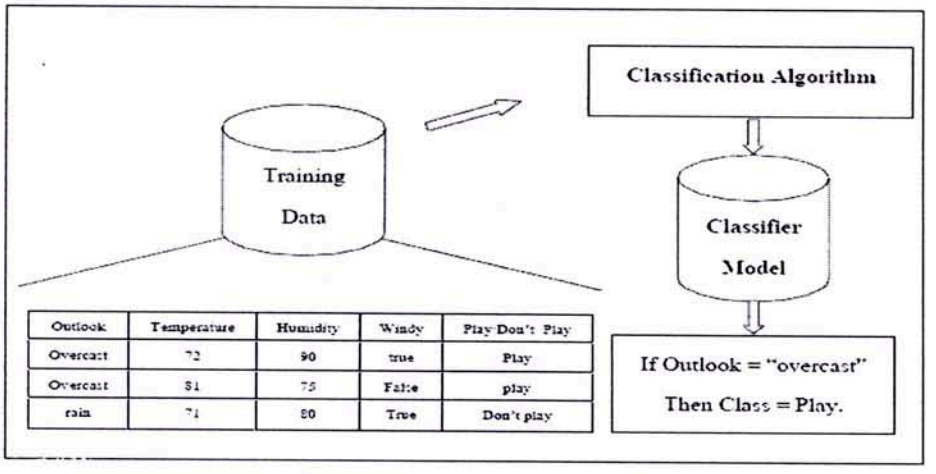
R3: Sex = Female, Age<=25 ⇨ No

R4: YR_Work<=2 ⇨ No

จากตัวอย่างกฎลูกค้าที่ควรได้รับเครดิตประกอบด้วยเงื่อนไข R1 คือคนที่อายุมากกว่า 25 ปี และทำงานมากกว่า 5 ปี และ R2 คือเพศชายและทำงานมากกว่า 2 ปี ส่วนลูกค้าที่ไม่ควรได้รับเครดิตประกอบด้วยเงื่อนไข R3 คือเพศหญิงและอายุน้อยกว่าหรือเท่ากับ 25 ปี และ R4 คืออายุการทำงานน้อยกว่าหรือเท่ากับ 2 ปี

ในการทำเหมืองข้อมูลแบบจำแนกจะมีขั้นตอน 2 ขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 การสร้างโมเดลต้นแบบ (Classifier Model) เป็นการนำชุดข้อมูลเรียนรู้ (Training Data) ผ่านกระบวนการของอัลกอริทึมการจำแนก (Classification Algorithm) ซึ่งผลลัพธ์ที่ได้จะอยู่ในรูปของโมเดลการจำแนก เช่น ต้นไม้ เป็นต้น จากนั้นสร้างเป็นกฎตัวจำแนกแบบเบย์ ซึ่งสามารถสรุปได้ ดังภาพ 4



ภาพ 4 ขั้นตอนการสร้างโมเดลการจำแนก

ที่มา: ไพฑูรย์ จันทร์เรือง, 2550

จากภาพ 4 เป็นตัวอย่างข้อมูลสำหรับการตัดสินใจเล่นกอล์ฟว่าสามารถเล่นได้ (Play) หรือเล่นไม่ได้ (Don't Play) มีรายละเอียดของข้อมูลคือ ทัศนียภาพ (Outlook) อุณหภูมิ (Temperature) ความชื้น (Humidity) และสถานะลม (Windy) ซึ่งเมื่อผ่านอัลกอริทึมการจำแนก ผลลัพธ์ที่ได้มาเป็นกฎคือ If Outlook = "overcast" Then Play หมายความว่า ถ้าทัศนียภาพมีเมฆครึ้ม แล้วสามารถเล่นกอล์ฟได้

ขั้นตอนที่ 2 การใช้โมเดลเพื่อการทำนาย (Prediction) ซึ่งจุดมุ่งหมายสูงสุดในการแก้ไขปัญหา คือการสร้างโมเดลเมื่อมีข้อมูลใหม่จะสามารถทำนายได้ โดยการนำข้อมูลที่ได้รับทำการเปรียบเทียบกับโมเดลการจำแนก และวิเคราะห์เพื่อตัดสินใจความเป็นไปได้ของข้อมูลนั้น ๆ

กลไกการทำเหมืองข้อมูลเป็นเทคนิคในการค้นความรู้จากข้อมูลขนาดใหญ่ ซึ่งวิธีการมากมาย เช่น วิธีทางสถิติ วิธีทางคณิตศาสตร์ วิธีทางคอมพิวเตอร์ เป็นต้น โดยเหมืองข้อมูลที่ได้รับความนิยมสามารถแบ่งตามลักษณะการใช้ในการแก้ปัญหาได้ ดังนี้

1. การอธิบายข้อมูล (Description) จะเป็นการวิเคราะห์ข้อมูลที่ไม่ซับซ้อนมาก จะเป็นการอธิบายรูปแบบและแนวโน้มที่อยู่ในข้อมูลหรือการสรุปข้อมูล เช่น การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) และการทดสอบสมมุติฐาน เป็นต้น

2. การจำแนกกลุ่ม (Classification) เป็นการจำแนกของวัตถุที่ต้องการ โดยอาศัยลักษณะที่คล้ายคลึงกันหรือแตกต่างกันเป็นการจัดวัตถุที่เข้ามาใหม่เข้ากลุ่มที่จัดไว้แล้วให้เข้ากลุ่มได้ถูกต้อง โดยต้องรู้ว่าวัตถุที่ต้องการจัดเข้ากลุ่มนั้นมีจำนวนกลุ่มที่แน่นอน โดยตัวแปรตามจะเป็น

ตัวแปรเชิงกลุ่มการจำแนกกลุ่มจะเน้นการสร้างตัวแบบที่กำหนดว่าวัตถุที่เข้ามาใหม่นั้นจะถูกจัดเข้ากลุ่มใดกลุ่มหนึ่งที่กำหนด เช่น การจำแนกลูกค้าที่ค้างชำระจะสามารถชำระหนี้ได้หรือไม่หรือการจำแนกลูกค้าของการให้เครดิตว่าจะมีความเสี่ยงมากปานกลางหรือน้อยเทคนิคที่ใช้ เช่น เทคนิคโรงข่ายประสาทเทียม เทคนิคต้นไม้ตัดสินใจ เป็นต้น

3. การรวมกลุ่ม (Clustering) เป็นการรวมวัตถุที่คล้ายคลึงกันไว้ในกลุ่มเดียวกัน โดยจะไม่มีข้อสมมุติเกี่ยวกับจำนวนกลุ่มว่ามีกี่กลุ่มแตกต่างจากการจำแนกกลุ่มตรงที่การจำแนกกลุ่มเป็นการจัดวัตถุใหม่เข้ากลุ่มใดกลุ่มหนึ่งจากกลุ่มที่มีอยู่การรวมกลุ่มจะไม่มีตัวแปรตามแต่จะมีแต่ตัวแปรอิสระที่ใช้วัดความคล้ายคลึงหรือใช้ในการคำนวณความคล้ายคลึง เช่น การรวมกลุ่มของลูกค้าที่มีพฤติกรรมการซื้อที่มีลักษณะคล้ายคลึงกันเทคนิคที่ใช้ เช่น เทคนิค K-Means เทคนิครวมกลุ่มแบบมีลำดับชั้น เป็นต้น

4. การประมาณค่า (Estimation) จุดประสงค์หลักในการประมาณค่านั้นจะเน้นการกำหนดค่าของตัวแปรตามที่ไม่ทราบค่าจากตัวแปรอิสระต่าง ๆ ซึ่งตัวแปรตามนั้นจะมีค่าเป็นตัวเลขหรือมีลักษณะเป็นข้อมูลชนิดต่อเนื่อง เช่น การประมาณรายได้ของพนักงานแต่ละคน

5. การทำนาย (Prediction) จะมีลักษณะคล้ายกับการประมาณค่าแต่ตัวแบบในการทำนายจะมุ่งเน้นเป็นการศึกษาพฤติกรรมในอนาคตมากกว่าในปัจจุบัน โดยตัวแปรตามที่จะทำนายนั้นจะเป็นข้อมูลชนิดต่อเนื่องหรือไม่ต่อเนื่องก็ได้ เช่น การทำนายยอดขายของบริษัทในเดือนหน้า

6. การหาความสัมพันธ์ (Association Rule) เป็นการหาความสัมพันธ์หรือความเกี่ยวเนื่องของข้อมูล โดยอาศัยหลักของกฎซึ่งจะอยู่ในรูปแบบ ถ้า สิ่งที่เกิดขึ้นแล้วผลที่จะตามมา เช่น การวิเคราะห์การซื้อสินค้าของลูกค้าซึ่งเป็นการวิเคราะห์พฤติกรรมการซื้อสินค้าซึ่งอาศัยวิธีการของการหาความสัมพันธ์เป็นหลัก (Larose, 2005)

เทคนิคเหมืองข้อมูล

เทคนิคเหมืองข้อมูลที่ใช้ในการจำแนก โดยทั่วไปจะประกอบด้วยองค์ประกอบหลัก 3 ส่วน ได้แก่ การสร้างตัวแบบ การประเมินความเหมาะสมของตัวแบบ และวิธีการค้นหาสาระในฐานข้อมูล ตัวแบบต้องแสดงถึงข้อสมมติฐานได้อย่างชัดเจนเพื่อให้สามารถค้นพบรูปแบบที่มีอยู่ได้และต้องมีความเหมาะสมและตรวจสอบในการพยากรณ์ได้ แม้กระทั่งกระบวนการค้นหาควรทำให้เกณฑ์การประเมินความเหมาะสมของตัวแบบมีความถูกต้องมากที่สุด (Cabema, 1998) เทคนิคต่าง ๆ ได้แก่

1. โครงข่ายประสาทเทียม (Neural Network) เป็นการเลียนแบบการทำงานของระบบประสาทเทียมซึ่งเลียนแบบการทำงานของระบบประสาทในสมองของมนุษย์ การทำงานของโครงข่ายประสาทเทียมแต่ละการทำงานจะรับข้อมูลเข้าไปคำนวณและสร้างผลลัพธ์ออกมาในลักษณะที่ไม่ใช่เป็นการทำงานแบบเชิงเส้นตรงเพราะว่าข้อมูลเข้าแต่ละตัวจะถูกให้ลำดับความสำคัญของค่าไม่เท่ากันค่าของผลลัพธ์ที่ได้จากการเชื่อมโยงกันนี้จะถูกนำมาเปรียบเทียบกับผลลัพธ์ที่ได้ตั้งเอาไว้ถ้าค่าที่ออกมาเกิดความคลาดเคลื่อนก็จะนำไปสู่การปรับค่าหรือน้ำหนักของค่าที่ใส่ไว้ให้แต่ละผลลัพธ์

2. ต้นไม้ตัดสินใจ (Decision Tree) เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของ ต้นไม้ตัดสินใจซึ่งต้นไม้ตัดสินใจมีการทำงานแบบมีเป้าหมายชัดเจน (Supervised Learning) คือ สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดได้ก่อนล่วงหน้าที่เรียกว่าข้อมูลเรียนรู้ (Training Set) ได้อัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้

3. วิธีการเข้าถึงหน่วยความจำพื้นฐานอย่างมีเหตุผล (Memory Based Reasoning: MBR) เปรียบเหมือนกับประสบการณ์การเรียนรู้ของมนุษย์ซึ่งอาศัยการสังเกตการณ์ที่เกิดขึ้นแล้วสร้างรูปแบบของสิ่งนั้นขึ้นมาในเหมืองข้อมูลใช้ MBR เพื่อทำการวิเคราะห์ฐานข้อมูลที่มีอยู่และกำหนดลักษณะพิเศษของข้อมูลที่อยู่ในนั้นจะต้องมีลักษณะสมบูรณ์ การสังเกตอย่างสมบูรณ์จะช่วยสร้างการทำนายอย่างแม่นยำขึ้น โมเดลจะถูกบอกคำตอบที่ถูกต้องจากกรณีศึกษา การทำงานแบบนี้วิธีนี้ถูกเรียกว่า การเรียนรู้แบบมีผู้สอน (Supervised Learning)

4. กฎการอุปนัย (Rule Induction) เป็นวิธีสำหรับการดึงเอาชุดของกฎเกณฑ์ต่างๆ มาเพื่อจัดแบ่งเงื่อนไขหรือกรณี ดังที่กล่าวข้างต้น โครงสร้างต้นไม้สามารถสร้างชุดของกฎต่างๆ และขณะที่บางครั้งเรียกวิธีการแบบนี้ว่า การสร้างกฎใหม่จากตัวอย่าง แต่วิธีการ หลังก็ยังมีความหมายที่แตกต่างกัน เนื่องจากวิธีการกฎการอุปนัยจะสร้างชุดของกฎที่เป็นอิสระซึ่งไม่จำเป็นต้องอยู่ในรูปโครงสร้างของต้นไม้ เพราะตัวสร้างกฎ (Rule Inducer) ไม่ได้บังคับการแตกข้อมูลเป็นแต่ละระดับแต่อาจจะสามารถค้นหา Pattern ที่แตกต่างกันได้และบางครั้งอาจดีกว่าสำหรับการจัดแบ่งระดับของผลลัพธ์

5. ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุด (K-nearest neighbor) เทคนิคของ (K-NN) ใช้วิธีการเดียวกันในการจัดแบ่งคลาสเทคนิคนี้จะตัดสินใจว่าคลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ได้ โดยการตรวจสอบจำนวนบางจำนวน (K ใน K-nearest neighbor) ของกรณีหรือเงื่อนไขที่เหมือนกัน

หรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวมของจำนวนเงื่อนไข หรือกรณีต่างๆสำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกับมันมากที่สุด

6. การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) เป็นการวิเคราะห์ความถดถอย ที่ใช้ในการพยากรณ์ผลลัพธ์ของสองตัวแปรเช่น Yes/No หรือ 0/1 แต่เนื่องจากตัวแปรตามมีค่าเพียงสองอย่างเท่านั้นจึงไม่สามารถสร้างแบบจำลองได้ด้วยวิธีการวิเคราะห์ความถดถอยแบบเส้นตรง

7. การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) เป็นวิธีทางคณิตศาสตร์ ซึ่งใช้ในการจำแนกและวิเคราะห์ วิธีนี้ได้รับการเผยแพร่ครั้งแรกในปี 1936 โดย R. A. Fisher เพื่อแยกต้น Iris ออกเป็น 3 พันธุ์ วิธีการนี้ทำให้ค้นพบพันธุ์ของต้นไม้ประเภทอื่น ๆ ผลลัพธ์ที่ได้จากแบบจำลองชนิดนี้ง่ายต่อการทำ ความเข้าใจ เพราะผู้ใช้งานทั่วๆ ไปก็สามารถพิจารณาได้ว่าผลลัพธ์จะอยู่ทางด้านใดของเส้นทางในแบบจำลองการเรียนรู้สามารถทำได้ง่ายวิธีการที่ใช้มีความไวต่อรูปแบบของข้อมูลวิธีนี้ถูกนำมาใช้มากในทางการแพทย์ สังคมวิทยา และชีววิทยา

จากการศึกษาแนวคิดและทฤษฎีเกี่ยวกับการทำเหมืองข้อมูล ผู้วิจัยได้นำความรู้ด้านการทำเหมืองข้อมูล โดยใช้การจำแนกแบบต้นไม้ตัดสินใจมาประยุกต์ใช้ เพื่อช่วยในการค้นหาสารสนเทศที่อยู่ในฐานข้อมูลขนาดใหญ่ทำได้ง่ายขึ้น ทำให้การจัดการข้อมูลสะดวกขึ้น และนำสารสนเทศที่ได้มาสร้างตัวแบบพยากรณ์ผลผลิตมันสำปะหลังในเขตภาคเหนือตอนล่าง

แนวคิดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบต้นไม้ตัดสินใจ

การศึกษาแนวคิดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบต้นไม้ตัดสินใจประกอบด้วยแนวคิดเกี่ยวกับต้นไม้ตัดสินใจ คุณลักษณะของต้นไม้ตัดสินใจ โครงสร้างต้นไม้ ขั้นตอนการสร้างต้นไม้ และอัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจ ซึ่งมีรายละเอียดดังนี้

แนวคิดเกี่ยวกับต้นไม้ตัดสินใจ

ในงานวิจัยนี้ได้ใช้เทคนิคต้นไม้ตัดสินใจในการจำแนกข้อมูล ซึ่งต้นไม้ตัดสินใจเป็นเครื่องมือที่ช่วยกำหนดขอบเขตของปัญหาและช่วยสร้างทางเลือกที่เป็นไปได้ในการแก้ปัญหา (Quinlan, 1993) ซึ่งต้นไม้ตัดสินใจนั้น เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของต้นไม้ตัดสินใจซึ่งต้นไม้ตัดสินใจนั้นมีการมีการเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยหลักการสร้างต้นแบบคือการแบ่งข้อมูลเป็น 2 ส่วน คือข้อมูลเรียนรู้ (Training Data) เป็นกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดไว้ก่อนล่วงหน้า และนำแบบจำลองที่ได้ไปใช้กับข้อมูลทดสอบ (Test Data) เพื่อทดสอบความแม่นยำของโมเดล โดยสามารถพยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้และในกระบวนการเหมืองข้อมูลใช้ในการพยากรณ์ (Prediction) หรือการจำแนก

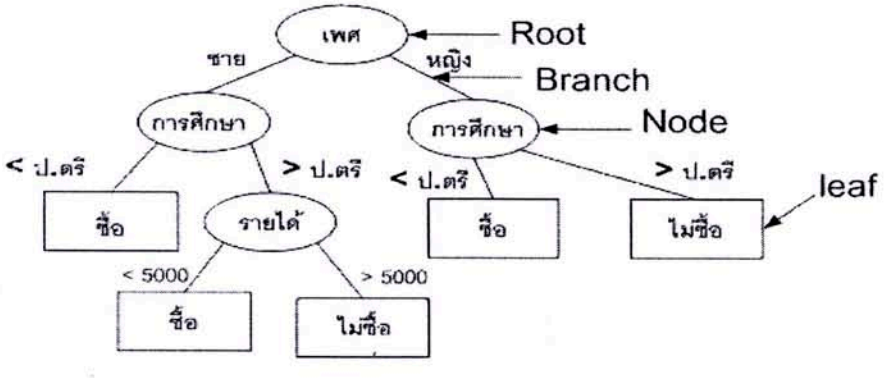
กลุ่ม (Classification) แผนผังต้นไม้ตัดสินใจในรูปแบบของโหนดและแสดงผลลัพธ์จากการกระทำหรือตัดสินใจในเงื่อนไขต่าง ๆ แล้วเชื่อมต่อกันเป็นเส้นที่แตกแขนงออกไป เทคนิคนี้จะช่วยให้ผู้ใช้เข้าใจถึงความสัมพันธ์และคุณสมบัติของข้อมูลได้ง่ายกว่าเทคนิคอื่น (ชลนิศา สาระ, 2550)

คุณลักษณะของต้นไม้ตัดสินใจ

คุณลักษณะของต้นไม้ตัดสินใจจะสามารถแสดงการเชื่อมโยงความสัมพันธ์ของปัญหาได้ชัดเจน และช่วยจัดการกับสถานการณ์ที่ซับซ้อนต่าง ๆ ให้อยู่ในรูปแบบที่กระชับขึ้น เพื่อช่วยให้เห็นภาพของปัญหาได้ชัดเจนขึ้น นอกจากนี้ยังมีโครงสร้างที่สามารถบอกผลลัพธ์ที่อาจเกิดขึ้นจากการคัดเลือกทางเลือกต่าง ๆ สำหรับการตัดสินใจและช่วยวิเคราะห์ลำดับการตัดสินใจแก้ไขปัญหาพร้อมทั้งวิเคราะห์ผลลัพธ์จากการตัดสินใจด้วยแนวทางต่าง ๆ

โครงสร้างต้นไม้

การวิเคราะห์โครงสร้างต้นไม้จะประกอบไปด้วยส่วนของโหนด (Node) แสดงถึงแอททริบิวต์ต่าง ๆ ส่วนกิ่ง (branch) ของต้นไม้คือค่าของแอททริบิวต์ในแต่ละโหนด ส่วนใบ (leaf) คือผลลัพธ์ที่ได้จากการจำแนกขั้นตอนในการสร้างต้นไม้ตัดสินใจจะเริ่มจากการนำข้อมูลแอททริบิวต์แต่ละตัวมาคำนวณหาค่าเกณฑ์ซึ่งโดยมากจะใช้เกณฑ์เอ็นโทรปี (Entropy) แล้วคำนวณหาค่า Information Gain เมื่อเริ่มต้นแอททริบิวต์ใดที่มีค่า Information Gain มากที่สุดก็จะได้รับเลือกเป็นโหนดราก (root) และในรอบถัดมา โหนดชั้นถัดลงมาก็จะได้กับแอททริบิวต์ที่มีค่า Information Gain มากที่สุด ซึ่งกลไกจะทำวนในลักษณะนี้จนกว่าพบเงื่อนไขการหยุดการแตกโหนด (บดินทร์ มิลินทางกูร และสุปิยา เจริญศิริวัฒน์, 2553), M.Mittechell, (1997) โดยต้นไม้ที่ได้ในแต่ละจุดจะมีลักษณะจำเพาะเพื่อใช้ในการตรวจสอบการจำแนกข้อมูล ดังภาพ 5



ภาพ 5 โครงสร้างต้นไม้ตัดสินใจ

การสร้างต้นไม้จะเริ่มสร้างจากบนลงล่าง โดยเริ่มจากโหนดราก (Root Node) คือ โหนดตัดสินใจแรก (Decision Node) การเลือกโหนดที่แตกต่างไปก็ทำให้โครงสร้างต้นไม้เปลี่ยนแปลงไปตามเงื่อนไขที่เลือก

ขั้นตอนการสร้างต้นไม้

ต้นไม้ตัดสินใจนั้นสามารถสร้างได้จากอัลกอริทึมที่แตกต่างกันได้หลายวิธี เช่น กรีดี อัลกอริทึมหรือขั้นตอนเชิงละโมบ ซึ่งเป็นการสร้างต้นไม้จากบนลงล่างแบบวนซ้ำด้วยวิธีการแบ่งปัญหาใหญ่เป็นปัญหาย่อย โดยส่วนใหญ่จะใช้กับข้อมูลที่เป็นแบบต่อเนื่อง และมีตัวแปรตามได้ 1 ตัวแปรต่อแบบจำลอง และใช้ค่าสัมประสิทธิ์จีนี (GINI) ในการตัดสินใจเลือกคุณสมบัติของโหนด (พิจิตรา จอมศรี, 2549) ในขณะที่อัลกอริทึม ID3 ที่เป็นพื้นฐานของการสร้างต้นไม้ตัดสินใจอื่นๆ เช่น C4.5 (Quinlan, 1993) โดยการทำงานจะใช้หลักการทฤษฎีข่าวสารในการค้นหาโหนดในต้นไม้ ซึ่ง (มลธิดา ฤทธิ์สมบูรณ์ และสุชา สมาชาติ, 2551) นอกจากนี้ยังมีต้นไม้ตัดสินใจ CART (Breiman, 1984), (ลลิตา ศรีชัยญา, 2553)

เมื่อต้องการทำนายหลาย ๆ ตัวแปรจะต้องสร้างแบบจำลองสำหรับตัวแปรตาม แต่ละอัลกอริทึมของเทคนิคต้นไม้ตัดสินใจ ส่วนใหญ่ไม่รองรับข้อมูลแบบต่อเนื่อง (Continuous Data) จะต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Data) ก่อน ตัวอย่างอัลกอริทึมเช่น ID3, C4.5 และ C5.0 อัลกอริทึม (เกรียงไกร พิพิธหิรัญการ, 2550)

การตัดเล็มกิ่งต้นไม้ (Pruning)

เป็นขั้นการตัดโหนดในต้นไม้ตัดสินใจ เพื่อลดความซ้ำซ้อนและการรู้จำจากข้อมูลฝึกมากเกินไป ทำให้สามารถจำแนกประเภทข้อมูลได้เร็วขึ้น รูปแบบของตัดเล็มต้นไม้ตัดสินใจแบ่งเป็น 2 รูปแบบดังนี้

1. การตัดเล็มก่อนการสร้างต้นไม้เสร็จเต็มรูป (Pre-Pruning) เป็นขั้นตอนการหยุดสร้างโหนดเพิ่ม เมื่อพบว่าคลาสที่พบในข้อมูลนั้นเป็นคลาสประเภทเดียวกันหมด หรือพบว่าค่าของแอททริบิวต์นั้นมีค่าเดียวกันทั้งหมด

2. การตัดเล็มหลังการสร้างต้นไม้เสร็จเต็มรูป (Post-Pruning) เป็นขั้นตอนการตัดเล็มโหนดจากล่างขึ้นบน โดยเลือกต้นไม้ย่อยที่มีความลึกมากที่สุดแล้วทำการตัดต้นไม้ย่อยที่เลือกออก จากนั้นแทนต้นไม้ย่อยด้วยโหนดของใบซึ่งการเลือกคลาสที่ใส่ในโหนดใบนั้นได้มาจากการหาคลาสที่ปรากฏอยู่ในต้นไม้ย่อยที่ตัดทิ้งไปที่มีค่ามากที่สุด

ต้นไม้ตัดสินใจจะแสดงผลลัพธ์ในรูปโครงสร้างต้นไม้ ซึ่งสามารถแปลงเป็นกฎที่เข้าใจได้ง่าย โดยที่แต่ละโหนดของต้นไม้แสดงถึงแอททริบิวต์ แต่ละกิ่งแสดงถึงเงื่อนไขในการทดสอบและโหนดใบแสดงถึงประเภทข้อมูลที่กำหนดไว้ โดยลักษณะของโครงสร้างต้นไม้ที่มีขนาดเล็ก ไม่

ซับซ้อน จะเป็นต้นไม้ตัดสินใจที่ดีที่สุดเหมาะสำหรับการนำไปใช้ในการตัดสินใจ (ลลิตา ศรีชัยญา, 2553)

อัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจ

ผู้วิจัยได้ทำการคัดเลือกอัลกอริทึมที่สนใจมาทำการวิจัยดังนี้

อัลกอริทึม ID3 (Iterative Dichotomiser 3) คือ วิธีการสร้างต้นไม้ตัดสินใจ โดยมีพื้นฐานจากเทคนิค Divide-and-conquer ที่ใช้ในการสร้างต้นไม้หรือที่เรียกว่า Top-Down Induction พัฒนาโดย J.Ross Quinlan (1975) การสร้างต้นไม้ด้วย ID3 ใช้หลักการของการใช้ Information Entropy ค่าที่วัดได้จะนำมาใช้ตัดสินใจเลือกตัวแปรในการสร้างเงื่อนไขในต้นไม้ตัดสินใจว่าจะใช้ตัวแปรใดในการแบ่งข้อมูล โดยวิธีการกำหนดโครงสร้างต้นไม้ตัดสินใจ จะเป็นการเลือกข้อมูลตามลำดับของตัวชี้วัดหรือค่าเกนสูงที่สุดเป็นข้อมูลเริ่มต้นและข้อมูลถัดไปที่มีค่าลดหลั่นกันตามลำดับ ตัวอย่างเช่น การพิจารณาจากกลุ่มข้อมูล 2 คลาสคือ P และ N โดยจำนวนตัวอย่างในคลาส P คือ p ตัว และจำนวนตัวอย่างในคลาส N คือ n ตัว ส่วนค่าของกลุ่มข้อมูลคือ ค่าคาดคะเนที่กลุ่มตัวอย่างต้องใช้จำนวนบิตในการแยกคลาส P และ N โดยนิยามตามสมการ (3) (ณัฐมณฑ์ สิริวัฒนานันท์, 2551)

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \qquad (3)$$

ค่าคาดคะเนของข้อมูล (Entropy) ที่แยกโดยการใช้ลักษณะประจำ A ซึ่งกำหนด A คือ ลักษณะ ประจำที่แบ่ง S ออกเป็น {S1, S2,..,SV} โดยให้ S1 มีตัวอย่างจากคลาส P จำนวน p1 และตัวอย่างจากคลาส N จำนวน n1 ดังสมการ (4)

$$E(A) = \sum_{i=1}^{v} \frac{p_1 + n_1}{p+n} I(p_1, n_1) \qquad (4)$$

ค่าเกนข้อมูล (Data Gain) ที่ได้จากการแยกข้อมูลด้วยลักษณะประจำ A ดังสมการ (5)

$$Gain(A) = I(p,n) - E(A) \qquad (5)$$

เมื่อแอททริบิวต์ตัวถัดไปถูกเลือกเป็นโหนดราก การคำนวณค่า Gain สำหรับแอททริบิวต์ก็จะคำนวณในลักษณะเดียวกับข้างต้น กระบวนการคำนวณค่า Gain ก็จะทำไปเรื่อย ๆ จนกว่าจะใช้แอททริบิวต์ครบทุกตัวแล้วเลือกแอททริบิวต์ที่ค่า Gain สูงสุดมาเป็นโหนดราก (ดิษฐพล มั่นธรรม, 2553)

ลักษณะของข้อมูลสำหรับอัลกอริทึม ID3 ต้องเป็นข้อมูลเชิงกลุ่มเท่านั้นและไม่สามารถใช้กับข้อมูลที่มีการสูญหายได้ หากจำเป็นต้องใช้ข้อมูลเชิงปริมาณต้องทำการจัดกลุ่มข้อมูลทั้งหมดก่อน เพื่อป้องกันปัญหาความอคติของข้อมูล (Lior and Oded, 2008)

อัลกอริทึม C4.5 เป็นอัลกอริทึมที่ผู้วิจัยได้เลือกมานำมาใช้ในการสร้างต้นไม้ตัดสินใจซึ่งพัฒนาโดย J. Ross Quinlan (1993) โดยนำเอา ID3 มาปรับปรุงให้มีความสามารถมากขึ้นใช้วิธีการ Information Gain เพิ่มเติมการจัดการกับข้อมูล ตัวเลข ข้อมูลที่ขาดไปและไม่สมบูรณ์ (Missing Values, Noisy Data) และการ Prune ด้วยการแทน Branch (กิ่ง) ที่ไม่ช่วยในการตัดสินใจด้วย Leaf Node ที่ตัดสินใจได้ดีกว่า (ณัฐมณฑ์ สิริวัฒนานันท์, 2551) โดยใช้ค่าอัตราส่วนเกน (Gain Ratio) ซึ่งจะใช้เป็นตัวจำแนกในการตัดสินใจเลือกแอททริบิวต์ที่จะใช้เป็นรากหรือโหนด ซึ่งในการทำงานขั้นตอนแรกคล้ายกับการทำงานด้วย ID3 คือต้องหาค่าสารสนเทศของข้อมูล คือ ค่า Gain จึงหาค่าสารสนเทศของการแบ่งหรือจำแนกข้อมูล (Split Information) ตามลักษณะของแอททริบิวต์แต่ละตัวเพื่อนำมาคำนวณค่าอัตราส่วนเกน (Gain Ratio) ซึ่งจะใช้เป็นหลักเกณฑ์ในการตัดสินใจสำหรับอัลกอริทึม C4.5 และถ้ายังไม่สามารถจัดกลุ่มข้อมูลให้เป็นกลุ่มเดียวกันทั้งหมดจะต้องสร้างต้นไม้ตัดสินใจต่อไป โดยพิจารณาเลือกแอททริบิวต์ที่จะมาเป็นโหนดต่อจากโหนดรากทำเช่นนี้ต่อไป โดยกระบวนการสร้างต้นไม้ตัดสินใจ ซึ่งจะสิ้นสุดลงเมื่อโหนดใบกลุ่มเป็นกลุ่มของข้อมูลเดียวกันทั้งหมด (ดิษฐพล มั่นธรรม, 2553)

โดยที่ C4.5 ต้องทำเพิ่มเติมคือ การหาค่าอัตราส่วนเกนซึ่งจะใช้ตัวนี้เป็นตัวแบ่งซึ่งหาได้จากสูตร ดังสมการ (6)

$$Gain\ ratio = \frac{Gain}{SplitInfo} \tag{6}$$

ซึ่ง Split info ใช้หลักการเดิมในการหา Information ออกมาเป็น 3 กลุ่ม โดยการนับจำนวนรายการของข้อมูลไม่ได้นับจำนวน ค่า Yes หรือ No จะได้ค่าอัตราส่วนเกนและทำการคำนวณในรายการอื่น ๆ จนครบ ซึ่งจะสามารถแสดงได้ว่าจะแบ่งการทำงานอยู่ค่าของรายการใด

C4.5 มีการพัฒนาอัลกอริทึมเพิ่มเติม (ชลนิศา สาระ, 2550) เพื่อแก้ปัญหาการเกิดความโน้มเอียง (Bias) ของข้อมูล โดยการหาค่าอัตราส่วนเกน ดังสมการ (7)

$$GainRatio(S,A) = \frac{Gain(S,A)}{IV(A)} \tag{7}$$

โดยที่ IV (A) คือ สัดส่วนของข้อมูลแอททริบิวต์ A ที่มีค่า i (P_i) ต่อจำนวนข้อมูลทั้งหมดของแอททริบิวต์ A ดังสมการ (8)

$$IV(A) = \sum_{v \varepsilon Vallues(A)} -\frac{pi}{N}\log_2\frac{p}{N}i \tag{8}$$

ลักษณะของข้อมูลสำหรับอัลกอริทึม C4.5 สามารถใช้ได้ทั้งข้อมูลเชิงกลุ่มปละข้อมูลเชิงปริมาณ และสามารถจัดการกับกรณีที่ค่าสูญหายได้ เป็นส่วนที่พัฒนาเพิ่มเติมจาก ID3 (Lior and Oded, 2008)

อัลกอริทึม CART (Classification and Regression Tree) ผู้วิจัยได้เลือกอัลกอริทึมCART มาใช้ซึ่ง (Breiman, 1984) ได้อธิบายว่า CART เป็นสถิตินอนพาราเมตริกที่ใช้จำแนกกลุ่มออกเป็น 2 กลุ่ม โดยตัวแปรอิสระจะมีลักษณะเชิงกลุ่มหรือเชิงปริมาณก็ได้ ซึ่งหากตัวแปรตามเป็นตัวแปรเชิงกลุ่ม จะใช้เทคนิคการแบ่งกลุ่มแบบ Classification Tree แต่ถ้าตัวแปรตามเป็นตัวแปรเชิงปริมาณจะใช้เทคนิคการแบ่งกลุ่มแบบ Regression Tree ซึ่งหลักการของวิธี CART จะเริ่มต้นด้วยการถือว่าสิ่งของทุกชิ้นอยู่กลุ่มเดียวกันแล้วแบ่งกลุ่มออกเป็น 2 กลุ่ม โดยใช้ตัวแปรที่มีค่าสูงในกลุ่มหนึ่งและต่ำในกลุ่มหนึ่งเป็นตัวแบ่ง หากตัวแปรมีความเป็นไปได้มากกว่า 2 กลุ่ม จะพยายามจัดกลุ่มให้เหลือเพียง 2 กลุ่มเท่านั้น โดยจำนวนในแต่ละกลุ่มไม่จำเป็นต้องเท่ากัน จากนั้นก็แบ่งแต่ละกลุ่มย่อยออกเป็น 2 กลุ่ม แบบทวิภาพทำการแบ่งย่อยอย่างนี้ไปเรื่อย ๆ จนกระทั่งถึงจุดที่เหมาะสมจึงหยุดค่าของตัวแปรที่ใช้แบ่งกลุ่ม อาจเป็นตัวแปรเชิงกลุ่มที่เรียงลำดับหรือไม่ได้เรียงลำดับ ซึ่งเรียกแผนภาพที่ได้จากวิธีการนี้ว่า "ต้นไม้จำแนก" โดยหลักเกณฑ์ที่ใช้ในการจำแนกกลุ่มของอัลกอริทึมนี้คือค่า Goodness of Split ซึ่งคำนวณได้จากสูตร ดังสมการ (9)

$$\Delta i(s,t) = i(t) - p_L[i(t_L)] - p_R[i(t_R)] \quad (9)$$

$$= 1 - \sum_{j=1}^{k} p^2(j|t) - p_L[i(t_L)] - p_R[i(t_R)]$$

เมื่อ $i(t) = 1 - \sum_{j=1}^{k} p^2(j|t)$

และ $\Delta i(s,t)$ = ค่า Goodness of Split เมื่อเลือก Attribute s เป็นโหนดรากและค่า Attribute t เป็นโหนดลูก

j = จำนวนเหตุการณ์ที่สนใจ

t = จำนวนของชุดข้อมูล

k = จำนนวนของ class

p_L = สัดส่วนของ Case ที่ node t ใน Child Node ด้านชุดข้อมูล t_L

p_R = สัดส่วนของ Case ที่ node t ใน Child Node ด้านชุดข้อมูล t_R

$$i(t_L) = 1 - \sum_{j=1}^{k} p^2(j|t_L)$$

$$i(t_R) = 1 - \sum_{j=1}^{k} p^2(j|t_R)$$

t_L = ชุดข้อมูลทางซ้ายของโหนด t

t_R = ชุดข้อมูลทางขวาของโหนด t

โดยกระบวนการสร้างต้นไม้ตัดสินใจจะสิ้นสุดเมื่อโหนดใบเป็นข้อมูลกลุ่มเดียวกันทั้งหมด อัลกอริทึมนั้นจะใช้ Goodness of Split เป็นเกณฑ์ในการสร้างตัวจำแนก และมีข้อจำกัดในการสร้างต้นไม้การจำแนกซึ่งสามารถสร้างต้นไม้การจำแนกแบบทวิภาค (Binary Tree) ได้เท่านั้น (Lior and Oded, 2008)

จากการศึกษาแนวคิดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบต้นไม้ตัดสินใจผู้วิจัยได้นำวิธีการของต้นไม้ตัดสินใจและอัลกอริทึมไปประยุกต์ใช้ในการวิเคราะห์ข้อมูลเพื่อสร้างตัวแบบพยากรณ์จากต้นไม้ตัดสินใจเพื่อนำไปพยากรณ์ผลผลิตมันสำปะหลังต่อไป

แนวคิดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบเบย์

การศึกษาแนวคิดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบเบย์ประกอบด้วย แนวคิดเกี่ยวกับตัวแบบพยากรณ์แบบเบย์และการนำวิธีการเรียนรู้เบย์อย่างง่ายไปใช้ ซึ่งมีรายละเอียดดังนี้

แนวคิดเกี่ยวกับตัวแบบพยากรณ์แบบเบย์

ทฤษฎีของเบย์ (Naïve Bayesian) ผู้วิจัยได้ทำการเลือกอัลกอริทึมที่ใช้วิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติของตัวอย่างไม่ขึ้นต่อกันมีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่นเป็นวิธีการจำแนกข้อมูลที่มีประสิทธิภาพ และมีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่น ๆ (สุวนีย์ กุลกรินีธรรม, 2549)

กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม vj สำหรับข้อมูลที่มีคุณสมบัติ n ตัว X = { a1, a2, ..., an} หรือ ใช้สัญลักษณ์ว่า P(a1, a2, ..., an | vj) ดังสมการ (10)

$$P(a_1, a_2, \ a_n \Big| v_j) = \prod_{i=1}^{n} P\left(a_i \Big| v_j\right) \tag{10}$$

จะได้สมการหาค่าความน่าจะเป็นจากกฎของเบย์ โดยที่ $\prod$ หมายถึง ผลคูณของค่า $P(a_i|v_j)$ ทั้งหมด

i = 1,2,3,...n และ j = 1,2,3,...,n

การนำวิธีการเรียนรู้เบย์อย่างง่ายไปใช้

การนำวิธีการเรียนรู้เบย์อย่างง่ายไปใช้ มีวิธีการดังต่อไปนี้

1. หาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2, \ a_n|v_j)$ จากสมการที่ 10 มาคูณกับค่าความน่าจะเป็น ของกลุ่มนั้น ๆ คือ $P(v_j)$ ได้เท่ากับ v_{NB}

2. นำค่าที่ได้มาเปรียบเทียบกันกลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นจะได้ว่าวิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการ (11) จะได้สมการวิธีจำแนกประเภทแบบเบย์อย่างง่าย

$$v_{NB} = \ \mathrm{g} \ \underset{v_i \in V}{mag} P(v_j) \times \prod_{i=1}^{n} P\left(a_i \Big| v_j\right) \tag{11}$$

จากการศึกษาแนวคิดและทฤษฎีเกี่ยวกับตัวแบบพยากรณ์แบบเบย์ ผู้วิจัยได้นำวิธีการของเบย์และอัลกอริทึมไปประยุกต์ใช้ในการวิเคราะห์ข้อมูลเพื่อสร้างตัวแบบพยากรณ์จากต้นไม้ตัดสินใจ เพื่อนำไปพยากรณ์ผลผลิตมันสำปะหลัง

แนวคิดและทฤษฎีเทคนิคการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ

ในการวิจัยนี้ผู้วิจัยทำการศึกษาและใช้วิธีแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ

1. วิธีการแบ่งข้อมูลออกเป็นส่วน ๆ (Cross – Validation) เป็นวิธีการแบ่งข้อมูล เรียกว่า "Fold" เช่น 3 Fold Cross- validation จะทำการแบ่งข้อมูลออกเป็น 3 ส่วน ในครั้งแรกจะเก็บส่วนที่ 3 ไว้เพื่อเป็นชุดข้อมูลทดสอบ และส่วนที่ 1 และ 2 ใช้สำหรับเป็นชุดข้อมูลเรียนรู้ ครั้งที่สองจะเก็บข้อมูลส่วนที่ 2,ไว้เป็นชุดข้อมูลทดสอบ สำหรับส่วนที่ 1 และ 3 ใช้เป็นชุดข้อมูลเรียนรู้ และครั้งสุดท้ายจะใช้ข้อมูลส่วนที่ 1 ไว้เป็นชุดข้อมูลทดสอบ สำหรับข้อมูลส่วนที่ 2 และ 3 เป็นชุดข้อมูลเรียนรู้ ซึ่งวิธี Cross- validation ข้อมูลทุกส่วนจะถูกใช้เป็นทั้งชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบสลับกันไป (Witten and Frank, 2005)

2. วิธีแบ่งข้อมูลออกเป็น 2 ส่วนตามร้อยละที่กำหนด (Percentage Split) เป็นวิธีการแบ่งข้อมูล โดยแบ่งข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ เช่น วิธี 60% split จะแบ่งข้อมูลชุดเรียนรู้ เป็น 60% และที่เหลืออีก 40% จะใช้เป็นข้อมูลชุดทดสอบ (Witten and Frank, 2005)

3. วิธีการใช้ข้อมูลทั้งหมดเป็นข้อมูลเรียนรู้ (Use Training Set) เป็นวิธีที่ผู้ใช้กำหนดตัวเลือก Use Training Set เพื่อใช้ทุกตัวอย่างในการสร้างต้นไม้ โดยใช้ข้อมูลทั้งหมดเป็นข้อมูลเรียนรู้

จากการศึกษาแนวคิดและทฤษฎีเทคนิคการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ผู้วิจัยได้นำวิธีการ Cross - Validation และ Use Training Set ไปใช้ในการวิเคราะห์ข้อมูลจากตัวแบบพยากรณ์ต้นไม้ตัดสินใจเพื่อนำไปสร้างตัวแบบพยากรณ์ผลผลิตมันสำปะหลัง

แนวคิดและทฤษฎีตัววัดประสิทธิภาพตัวแบบพยากรณ์

วิธีการวัดประสิทธิภาพของตัวแบบพยากรณ์ โดยใช้ข้อมูลทดสอบนั้นจะพิจารณาจากขนาดของต้นไม้คือ จำนวนโหนดและใบของต้นไม้ ซึ่งประกอบไปด้วยค่าต่าง ๆ ดังนี้

1. ค่าความถูกต้องในการจำแนกกลุ่ม (Accuracy)
2. ค่าความระลึก (Recall)
3. ค่าความแม่นยำ (Precision)

4 ค่าความเหวี่ยง (F-Measure)

5. ค่าผลบวกจริง (True Positive Rate)

6. ค่าผลบวกเท็จ (False Positive Rate)

ตัวอย่าง การวัดประสิทธิภาพแบบผลลัพธ์ 3 ค่า คือ ค่าระดับ 1 ค่าระดับ 2 ค่าระดับ 3 แสดงในตาราง 2

ตาราง 2 ตัวอย่างการแสดงค่า Confusion Matrix ในโปรแกรมเวกา

		ค่าพยากรณ์		
	ระดับ	**1**	**2**	**3**
ค่าจริง	1	100	115	3
	2	44	1083	56
	3	7	308	86
	รวม	151	1506	145

เมื่อแทนค่าพารามิเตอร์ดังนี้

100 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 1 ผลการจำแนกเป็นเท่ากับระดับ 1

115 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 1 ผลการจำแนกเป็นเท่ากับระดับ 2

3 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 1 ผลการจำแนกเป็นเท่ากับระดับ 3

44 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 2 ผลการจำแนกเป็นเท่ากับระดับ 1

1083 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 2 ผลการจำแนกเป็นเท่ากับระดับ 2

56 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 2 ผลการจำแนกเป็นเท่ากับระดับ 3

7 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 3 ผลการจำแนกเป็นเท่ากับระดับ 1

308 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 3 ผลการจำแนกเป็นเท่ากับระดับ 2

86 คือ จำนวนผลผลิตที่มีจำนวนอยู่ในระดับที่ 3 ผลการจำแนกเป็นเท่ากับระดับ 3

จากนั้นนำผลลัพธ์ที่ได้จากตาราง 2 มาคำนวณเพื่อหาค่าความถูกต้องในการจำแนกกลุ่ม ค่าความระลึก ค่าความแม่นยำ ค่าความเหวี่ยง ค่าผลบวกจริง และค่าผลบวกเท็จ ดังนี้

ค่าความถูกต้องในการจำแนกกลุ่ม (Accuracy) คือ ผลที่ได้จากการใช้อัลกอริทึมต้นไม้ตัดสินใจในการหาค่าความถูกต้องในการจำแนกคำตอบระดับผลผลิต ดังสมการ (12)

ค่าความถูกต้องในการจำแนกกลุ่ม = $\frac{\text{จำนวนข้อมูลที่จำแนกได้ถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมด}}$ (12)

$$Accuracy = \frac{100+1083+86}{100+115+3+44+1083+56+7+308+86} = 0.70$$

ค่าความระลึก (Recall) คำนวณจากค่าของข้อมูลที่ผลลัพธ์ถูกต้อง โดยพิจารณาจากข้อมูลของผลลัพธ์เดียวกัน ดังสมการ (13)

ค่าความระลึก = $\frac{\text{จำนวนข้อมูลที่จำแนกได้ถูกต้องของกลุ่มนั้น}}{\text{จำนวนข้อมูลที่อยู่ในกลุ่มนั้นจริง}}$ (13)

ตัวอย่าง ค่าความระลึก ของระดับ 1 คือ

$$r = \frac{100}{100+115+3} = 0.45$$

ตัวอย่าง ค่าความระลึก ของระดับ 2 คือ

$$r = \frac{1083}{44+1083+56} = 0.91$$

ตัวอย่าง ค่าความระลึก ของระดับ 3 คือ

$$r = \frac{86}{7+308+86} = 0.21$$

ค่าความแม่นยำ (Precision) คำนวณจากค่าของข้อมูลที่ผลลัพธ์ถูกต้องโดยพิจารณาจากจำนวนข้อมูลทั้งหมดที่ถูกจำแนกมีผลลัพธ์เดียวกัน ดังสมการ (14)

ค่าความแม่นยำ = $\frac{\text{จำนวนข้อมูลที่จำแนกได้ถูกต้องของกลุ่มนั้น}}{\text{จำนวนข้อมูลที่ถูกจำแนกว่าเป็นกลุ่มนั้นทั้งหมด}}$ (14)

ตัวอย่าง ค่าความแม่นยำ ของระดับ 1 คือ

$$p = \frac{100}{100+44+7} = 0.66$$

ตัวอย่าง ค่าความแม่นยำ ของระดับ 2 คือ

$$p = \frac{1083}{115+1083+308} = 0.71$$

ตัวอย่าง ค่าความแม่นยำ ของระดับ 3 คือ

$$p = \frac{86}{3+56+86} \qquad = 0.59$$

ค่าความเหวี่ยง (F – Measure) คือค่าระหว่างค่าความแม่นยำและค่าความระลึก โดยจะคำนวณจากค่าความแม่นยำ (p) และค่าความระลึก (r) และนำค่ามาหาค่าเฉลี่ยระหว่างค่าทั้งสอง ดังสมการ (15)

$$F(r,p) = \frac{2pr}{p+r} \tag{15}$$

ตัวอย่าง ค่าความเหวี่ยง ของระดับ 1 คือ

$$F(r,p) = \frac{2(0.662)(0.459)}{0.662+0.459} \qquad = 0.542$$

ตัวอย่าง ค่าความเหวี่ยง ของระดับ 2 คือ

$$F(r,p) = \frac{2(0.719)(0.915)}{0.719+0.915} \qquad = 0.806$$

ตัวอย่าง ค่าความเหวี่ยง ของระดับ 3 คือ

$$F(r,p) = \frac{2(0.593)(0.214)}{0.593+0.214} \qquad = 0.315$$

ค่าผลบวกจริง (True Positive Rate: TP Rate) หมายถึง ผลการทดสอบที่ให้ผลบวกแสดงว่าเป็นจำนวนคำตอบระดับผลผลิตที่ตอบตรงค่าจริง เทียบกับจำนวนค่าคำตอบระดับผลผลิตที่ตอบค่าจริงทั้งหมดแสดงค่าและอัลกอริทึมที่ให้ค่าของผลบวกจริงสูงกว่าจะเป็นอัลกอริทึมที่ดีกว่า ดังสมการ (16)

$$\text{ค่าผลบวกจริง} = \frac{\text{จำนวนคำตอบระดับผลผลิตที่ตอบตรงค่าจริง}}{\text{จำนวนคำตอบระดับผลผลิตที่ตอบค่าจริงทั้งหมด}} \tag{16}$$

ตัวอย่าง ค่าผลบวกจริง ของระดับ 1 คือ

$$TPRate = \frac{100}{100+115+3} \qquad = 0.45$$

ตัวอย่าง ค่าผลบวกจริง ของระดับ 2 คือ

$$TPRate = \frac{1083}{44+1083+56} \qquad = 0.91$$

ตัวอย่าง ค่าผลบวกจริง ของระดับ 3 คือ

$$TPRate = \frac{86}{7+308+86} \qquad = 0.21$$

ค่าผลบวกเท็จ (False Positive Rate: FP Rate) หรือ 1 – ค่าความจำเพาะ หมายถึง ผลการทดสอบที่ให้ผลบวก คือ จำนวนคำตอบระดับผลผลิตไม่ได้ตอบค่าจริง เทียบกับจำนวนคำตอบระดับผลผลิตไม่ได้ตอบค่าจริงทั้งหมด แสดงค่าและอัลกอริทึมที่ให้ค่าของผลบวกเท็จต่ำกว่าจะเป็นอัลกอริทึมที่ดีกว่า ดังสมการ (17)

$$\text{ค่าผลบวกเท็จ} = \frac{\text{จำนวนคำตอบระดับผลผลิตไม่ได้ตอบค่าจริง}}{\text{จำนวนคำตอบระดับผลผลิตไม่ได้ตอบค่าจริงทั้งหมด}} \qquad (17)$$

ตัวอย่าง ค่าผลบวกเท็จ ของระดับ 1 คือ

$$FPRate = \frac{(44+7)}{(1183+401)} \qquad = 0.032$$

ตัวอย่าง ค่าผลบวกเท็จ ของระดับ 2 คือ

$$FPRate = \frac{(115+308)}{(218+401)} \qquad = 0.683$$

ตัวอย่าง ค่าผลบวกเท็จ ของระดับ 3 คือ

$$FPRate = \frac{(3+56)}{(218+1183)} \qquad = 0.042$$

จากการศึกษาแนวคิดและทฤษฎีตัววัดประสิทธิภาพตัวแบบพยากรณ์ ผู้วิจัยได้นำวิธีการวัดประสิทธิภาพตัวแบบพยากรณ์มาเป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของอัลกอริทึมของต้นไม้ตัดสินใจและนำมาวิเคราะห์ข้อมูลเพื่อนำไปสร้างระบบสารสนเทศการสืบค้นข้อมูลและโปรแกรมพยากรณ์ผลผลิตมันสำปะหลัง

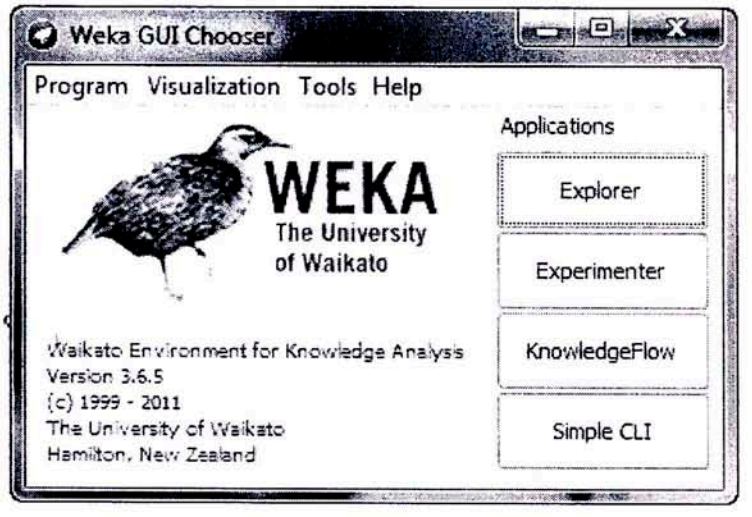
แนวคิดเกี่ยวกับโปรแกรมเวกา (WEKA)

การศึกษาแนวคิดเกี่ยวกับโปรแกรมเวกาประกอบด้วย ลักษณะทั่วไปของโปรแกรมเวกาและการใช้งานโปรแกรมเวกา ซึ่งมีรายละเอียดดังนี้

ลักษณะทั่วไปของโปรแกรมเวกา

เกรียงไกร พิพิธหิรัญการ (2550) ได้กล่าวถึงโปรแกรมเวกา (Waikato Environment for Knowledge Analysis: WEKA) ว่าเป็นโปรแกรมซึ่งเขียนโดยใช้ภาษาจาวาจะเน้นการเรียนรู้ด้วย

เครื่อง (Machine Learning) กับการทำเหมืองข้อมูล โปรแกรมเวกามีความสามารถด้านการประมวลผลในเรื่องของข้อมูลแบบกฎเชื่อมโยง ข้อมูลแบบจำแนกประเภท ข้อมูลแบบการวิเคราะห์การเกาะกลุ่ม และการแสดงการไหลของข้อมูล (Knowledge flow) ดังภาพ 6



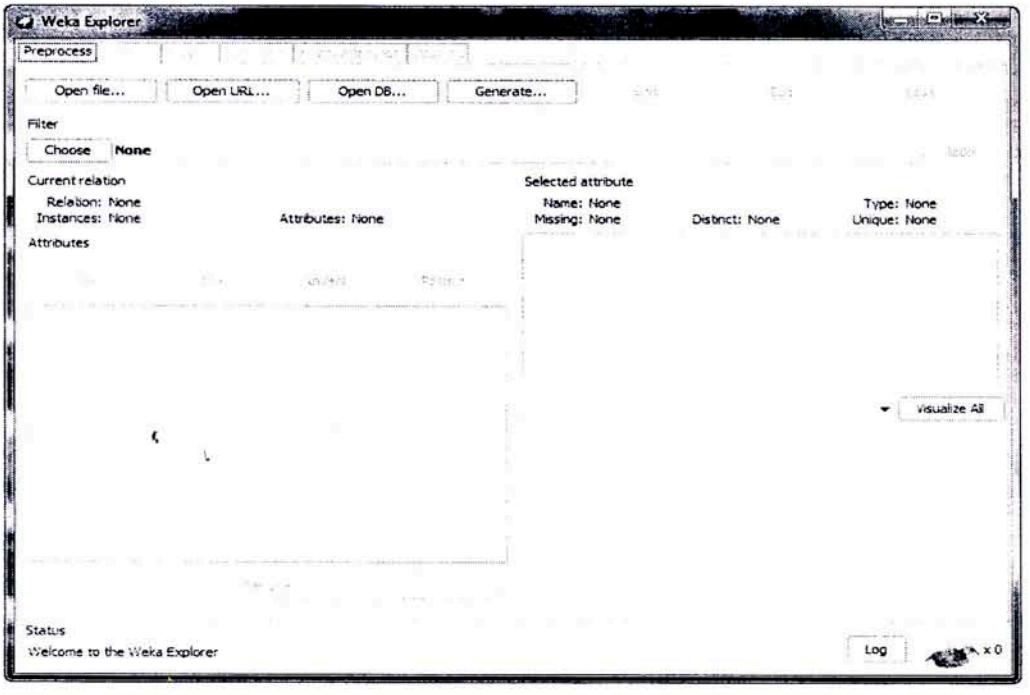
ภาพ 6 โปรแกรมเวกา

ที่มา: Waikato University, n.d.

การใช้งานโปรแกรมเวกา

ณัฐกรณ์ ตั้งพูนทรัพย์ (2551) ได้อธิบายถึงโปรแกรมหลักของซอฟต์แวร์ว่าประกอบไปด้วยส่วนต่าง ๆ ดังนี้

1. Simple CLI (Command Line Interface) เป็นโปรแกรมรับคำสั่งทำงานผ่านการพิมพ์
2. Explorer เป็นโปรแกรมที่ออกแบบในลักษณะ GUI
3. Experimenter เป็นโปรแกรมที่ออกแบบการทดลองและการทดสอบผล
4. Knowledge Flow เป็นโปรแกรมออกแบบผังการไหลของความรู้
5. ArffViewer เป็นโปรแกรมที่ใช้สำหรับแก้ไขแฟ้มประเภท Arff
6. Log เป็นโปรแกรมที่ใช้อ่านข้อความบันทึกเก็บระหว่างการทำงาน



ภาพ 7 เมนูหลักของการใช้งานโปรแกรมเวกา

จากภาพ 7 แสดงเมนูหลักของ Explorer ประกอบด้วย

1. Preprocess การเตรียมข้อมูล
2. Classify รวมโมดูลการทำเหมืองข้อมูลแบบจัดแบ่งประเภท
3. Cluster รวมโมดูลการทำเหมืองข้อมูลแบบการเกาะกลุ่ม
4. Associate รวมโมดูลการทำเหมืองข้อมูลแบบกฎเชื่อมโยง
5. Select attributes รวมโมดูลสำหรับการวิเคราะห์ความเกี่ยวพันของลักษณะประจำ
6. Visualize นำเสนอข้อมูลด้วยภาพนามธรรมสองมิติ

ภาพ 8 ขั้นตอนการใช้งาน Explorer ในโปรแกรมเวกา

จากภาพ 8 โปรแกรม Explorer แสดงขั้นตอนการทำงานซึ่งประเภทของแฟ้มข้อมูลที่รับได้ต้องอยู่ในรูปแบบ ASCII อาจเป็น Arff, CSV, C45 ในกรณีแฟ้มข้อมูลอยู่ในเครือข่ายผู้ใช้สามารถเรียกใช้โดยอาศัย URL หรือใช้ข้อมูลที่อยู่ในฐานข้อมูลที่เชื่อมโยงผ่าน JDBC

จากการศึกษาแนวคิดเกี่ยวกับโปรแกรมเวกา (WEKA) ผู้วิจัยได้นำวิธีการและเทคนิคต่าง ๆ ของโปรแกรมเวกามาประยุกต์ใช้ในการหาค่าคำตอบที่ต้องการจากข้อมูลที่มีและนำมาวิเคราะห์ข้อมูล เพื่อนำไปสร้างโปรแกรมพยากรณ์ผลผลิตมันสำปะหลังต่อไป

แนวคิดเกี่ยวกับหลักการพัฒนาระบบ

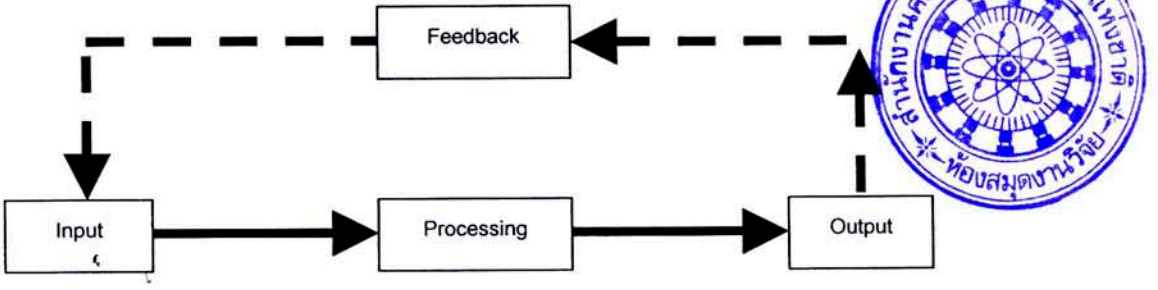
การศึกษาแนวคิดเกี่ยวกับพัฒนาระบบ ความหมายของระบบสารสนเทศ องค์ประกอบของระบบสารสนเทศ หน้าที่ของระบบสารสนเทศ และวงจรการพัฒนาระบบ ซึ่งมีรายละเอียดดังนี้

ความหมายของระบบสารสนเทศ มีผู้ให้ความหมายของคำว่าระบบสารสนเทศ ดังนี้

ไพลิน ผ่องใส (2536) ได้อธิบายว่าระบบสารสนเทศหมายถึง วิธีการรวบรวม จัดเก็บข่าวสารที่เกี่ยวกับการดำเนินงานทั้งภายในและภายนอก ทั้งในอดีตปัจจุบัน และคาดคะเนหรือพยากรณ์สำหรับอนาคตด้วย ข่าวสารต่างๆ ดังกล่าวเป็นส่วนสำคัญที่สนับสนุนในการวางแผน การควบคุม ตลอดจนการปฏิบัติงานขององค์การ ทั้งการจัดเตรียมข่าวสารข้อมูลนั้นต้องอยู่ในระยะเวลาที่เหมาะสม คือทันเวลาและทันเหตุการณ์จึงจะให้ผลด้านการใช้ประกอบการตัดสินใจของผู้บริหาร

Loudon and Loudon (1994) อธิบายว่า ระบบสารสนเทศ หมายถึง กลุ่มของกระบวนการที่มีองค์ประกอบของการเก็บรวบรวมข้อมูล การประมวลผลข้อมูล การจัดเก็บ และการเผยแพร่สารสนเทศ เพื่อใช้สนับสนุนการตัดสินใจในการปฏิบัติงาน

กิตติ ภัคดีวัฒนะกุล (2546) อธิบายว่าระบบสารสนเทศหมายถึง การรวบรวมองค์ประกอบต่าง ๆ เช่น ข้อมูล การประมวลผล การเชื่อมโยง เครือข่าย เพื่อนำเข้า (Input) สู่ระบบใดๆ แล้วนำมาผ่านกระบวนการ (Process) ที่อาจใช้คอมพิวเตอร์ช่วยเพื่อเรียบเรียง เปลี่ยนแปลง และจัดเก็บ เพื่อให้ได้ผลลัพธ์ (Output) ที่สามารถใช้สนับสนุนการตัดสินใจทางธุรกิจได้ ดังภาพ 9



ภาพ 9 กระบวนการทำงานของระบบสารสนเทศ

ข้อมูลนำเข้า (Input) คือ การเก็บรวบรวมสมาชิกหรือองค์ประกอบของระบบ เช่นข้อมูล (Data) หรือสารสนเทศ (Information) เพื่อนำไปทำการประมวลผลต่อไป

กระบวนการทำงาน (Processing) คือการเปลี่ยนแปลงหรือแปรสภาพข้อมูลที่นำเข้าสู่ระบบ (Input) เพื่อให้ได้ผลลัพธ์ (Output) ที่สามารถใช้ในการตัดสินใจได้ โดยการเปลี่ยนแปลงหรือแปรสภาพนั้นอาจจะเป็นการคำนวณ เปรียบเทียบหรือวิธีการอื่นๆ

ผลลัพธ์ (Output) คือ ผลลัพธ์ที่ได้เนื่องจากการประมวลผลข้อมูลหรือสารสนเทศ แสดงอยู่ในรูปแบบของรายงาน (Report) หรือเป็นแบบฟอร์มต่างเพื่อนำไปใช้ในการดำเนินงานทางธุรกิจต่อไป

ผลตอบรับ (Feedback) คือ ผลลัพธ์ที่ทำให้เกิดการปรับปรุง เปลี่ยนแปลงในการนำข้อมูลเข้า หรือการประมวลผลข้อมูล เช่น ข้อผิดพลาดที่พบรายงานต่างๆ นั้น ทำให้ทราบได้ว่า ในขณะนำข้อมูลเข้า หรือการประมวลผลนั้น อาจมีข้อผิดพลาดเกิดขึ้น ผลตอบรับจึงมีความสำคัญอย่างยิ่งในการทำงานเพื่อให้เกิดประสิทธิภาพและประสิทธิผลเป็นที่น่าพอใจ

องค์ประกอบของระบบสารสนเทศ

เป็นการนำเอาเทคโนโลยีสารสนเทศด้านเทคโนโลยีคอมพิวเตอร์เข้ามาช่วยในการจัดการภายในองค์กรด้านต่างๆ เพื่อให้องค์กร บรรลุ เป้าหมายตามที่องค์กรตั้งไว้ด้วยความสะดวก รวดเร็ว แม่นยำ โดยระบบสารสนเทศมีองค์ประกอบ 6 ส่วน ดังนี้

1. ฮาร์ดแวร์ (Hardware) ประกอบด้วยเครื่องคอมพิวเตอร์ และอุปกรณ์ต่างๆ ที่ใช้เชื่อมโยงคอมพิวเตอร์เข้าด้วยกันกับเครือข่าย

2. ซอฟต์แวร์ (Software) โปรแกรมเป็นชุดคำสั่งที่สั่งงานให้ฮาร์ดแวร์ทำงานตามลำดับ

3. ข้อมูล/สารสนเทศ (Data/Information) จะเป็นตัวชี้ให้เห็นความสำเร็จหรือความล้มเหลวขององค์กร

4. ผู้ใช้/บุคลากร (User/People) บุคลากรในระดับต่างๆ

5. เทคโนโลยีการติดต่อสื่อสาร (Network & Communication)

6. ขั้นตอนการปฏิบัติงาน (Procedure)

หน้าที่ของระบบสารสนเทศ มีดังต่อไปนี้

1. การจัดเก็บ บันทึกและประมวลผลเป็นหน้าที่หลักของระบบในการจัดเก็บข้อมูลต่างๆ เกี่ยวกับการดำเนินการทั้งหมดขององค์กร เช่น เมื่อลูกค้าสั่งสินค้ามา ระบบจะจัดการตรวจสอบว่ามีสินค้าพอจัดส่งหรือไม่ ถ้ามีจะทำการจัดส่งของ พิมพ์ใบส่งของ ใบ Invoice และบันทึกข้อมูลการสั่งซื้อสินค้านั้นไว้

2. การจัดการฐานข้อมูล ข้อมูลที่เกิดขึ้นในองค์กรจะถูกเก็บไว้ในฐานข้อมูล ซึ่งต้องใช้โปรแกรมจัดการฐานข้อมูล (Database Management System: DBMS) มาช่วยในการจัดการและอำนวยความสะดวกในการจัดเก็บข้อมูล เปลี่ยนแปลงแก้ไขข้อมูล ค้นหาข้อมูล ตลอดจนช่วยผู้ใช้ให้สามารถใช้ข้อมูลร่วมกันได้โดยไม่สับสน

3. การจัดทำรายงาน เป็นกลุ่มโปรแกรมที่มีหน้าที่ในการจัดทำรายงานต่าง ๆ เพื่อส่งให้แก่ผู้บริหารหรือผู้ใช้ระบบ

4. การสอบถามข้อมูล คือโปรแกรมที่อำนวยความสะดวกในการสอบถามข้อมูลต่าง ๆ ที่มีในฐานข้อมูลซึ่งการสอบถามมี 2 แบบคือ

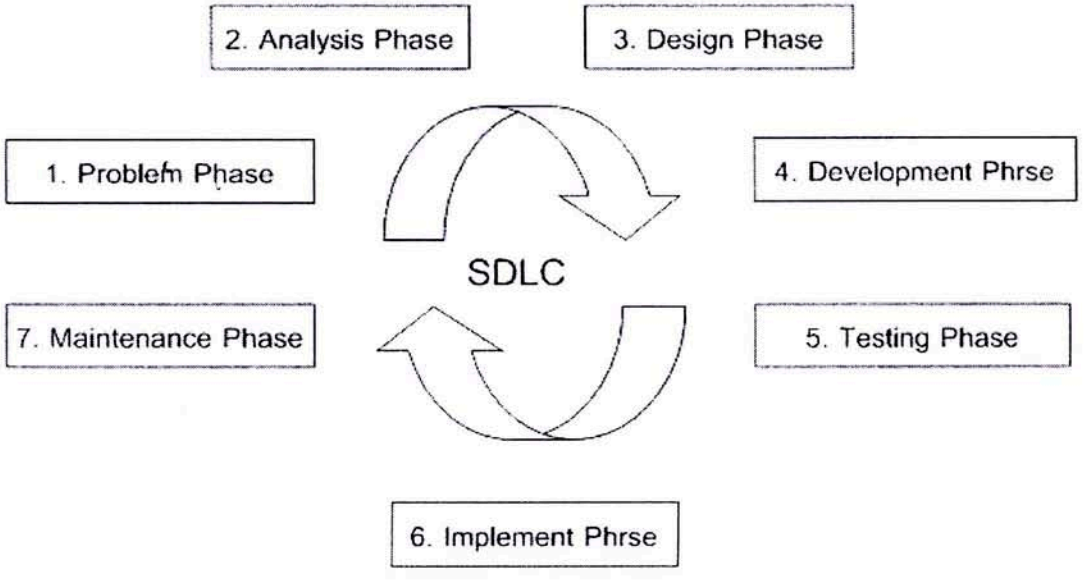
4.1 การสอบถามแบบประจำ ทำให้สามารถออกแบบโปรแกรมในการค้นหาไว้ล่วงหน้าได้เมื่อถึงเวลาก็เรียกมาใช้ได้ทันที

4.2 การสอบถามแบบไม่แน่นอน ไม่สามารถทราบล่วงหน้าได้ว่าผู้บริหารต้องการสารสนเทศแบบใด ทำให้ไม่สามารถเตรียมโปรแกรมการค้นหาล่วงหน้าได้

5. การช่วยสนับสนุนการตัดสินใจ เป็นโปรแกรมสำหรับอำนวยความสะดวกให้ผู้บริหารนำข้อมูลการดำเนินงานของบริษัทหรือหน่วยงานมาป้อนเข้าสู่แบบจำลองเพื่อทดสอบแนวคิดของตนหรือเพื่อหาแนวทางการตัดสินใจว่าแบบใดจะให้ผลดีที่สุด (กิตติ ภักดีวัฒนะกุล, 2546)

วงจรการพัฒนาระบบ (System Develop Life Circle: SDLC)

โอภาส เอี่ยมสิริวงศ์ (2548) ได้อธิบายวงจรการพัฒนาระบบว่าเป็น กระบวนการทางความคิดที่แก้ปัญหาและตอบสนองความต้องการของผู้ใช้ วงจรการพัฒนาระบบ (SDLC) เป็นวิธีการพัฒนาระบบเก่าหรือแบบเดิมที่มักนำมาปะยุกต์ใช้กับการพัฒนาระบบตั้งอดีตถึงปัจจุบัน ซึ่งมีกรอบการร่างานที่เป็นโครงสร้างชัดเจน โยมีลำดับของกิจกรรมในแต่ละระยะที่เป็นลำดับแน่นอนสำหรับระยะหรือเฟสต่าง ๆ ตามแบบแผนของวงจรการพัฒนา ประกอบด้วย 7 ระยะ ดังภาพ 10



ภาพ 10 วงจรการพัฒนาระบบ

1. การกำหนดปัญหาของระบบเดิม (Problem Definition) คือ ขั้นตอนนี้เป็นการกำหนดขอบเขตปัญหา สาเหตุของปัญหา รวมทั้งกลยุทธ์ในการแก้ปัญหา นักวิเคราะห์ระบบจะต้องศึกษาระบบงานเดิม โดยหาเป้าหมายที่ชัดเจนของงานต่าง ๆประกอบกับนำคอมพิวเตอร์เข้าไปใช้ในส่วนต่าง ๆ ของระบบจากการสุ่มตัวอย่าง การสอบถามความต้องการ เพื่อค้นหาและเก็บข้อมูลที่ต้องการของระบบจากผู้ใช้ กำหนดวัตถุประสงค์ที่สามารถวัดได้และขอบเขตการพัฒนาระบบ

2. การวิเคราะห์ระบบ (Analysis) การวิเคราะห์ระบบจะรวบรวมจ้อมูลต่าง ๆ ที่ได้จากขั้นตอนที่ 1 เขียนเป็น แผนภาพบริบท (Context Diagram) แผนภาพกระแสข้อมูล (Data Flow Diagram: DFD) พจนานุกรมข้อมูล (Data Dictionary) มาช่วยในการวิเคราะห์เพื่อแก้ไขปัญหาให้ถูกต้อง และนักวิเคราะห์ระบบต้องมีการทำงานร่วมกับผู้ใช้ระบบเพื่อได้ความต้องการจากผู้ใช้โดยแท้จริง นำผลการวิเคราะห์มาออกแบบและพัฒนาระบบต่อไป

3. การออกแบบระบบ (Design) หลังจากการเคราะห์ระบบแล้ว ขั้นตอนนี้จะต้องทำการวางโครงสร้างระบบงาน ทั้งในรูปลักษณะทั่วไปและเฉพาะ เพื่อแก้ไขปัญหาที่เกิด ซึ่งขั้นตอนนี้จะได้ระบบเพื่อทำการออกแบบ Input, Output, ER - Model และ Database เพื่อให้ได้ระบบงานที่สมบูรณ์ เพื่อส่งขั้นตอนนี้ไปยังโปรแกรมเมอร์ในการเขียนชุกคำสั่งต่อไป

4. การพัฒนาระบบ (Development) ขั้นตอนนี้จะเป็นการทำงานร่วมกันระหว่างโปรแกรมเมอร์และนักวิเคราะห์ระบบเพื่อพัฒนาซอฟต์แวร์ ซึ่งจะต้องนำส่วนที่ได้จากการวิเคราะห์ในขั้นตอนที่ 2 และการออกแบบในส่วนที่ 3 มาใช้โดยโปรแกรมเมอร์จะต้องเป็นผู้เขียนโปรแกรม ตรวจสอบข้อผิดพลาด กำหนดความปลอดภัยของระบบและทดสอบโปรแกรมรวมถึงทำเอกสารโปรแกรมสำหรับผู้ใช้ระบบอีกด้วย

5. การทดสอบระบบ (Testing) ก่อนที่จะนำระบบที่สร้างขึ้นไปใช้งานจริงจะต้องมีการทดสอบระบบก่อน ซึ่งบางครั้งผู้ทดสอบอาจเป็นโปรแกรมเมอร์เองหรือนักวิเคราะห์ระบบและผู้ใช้ระบบทดสอบ

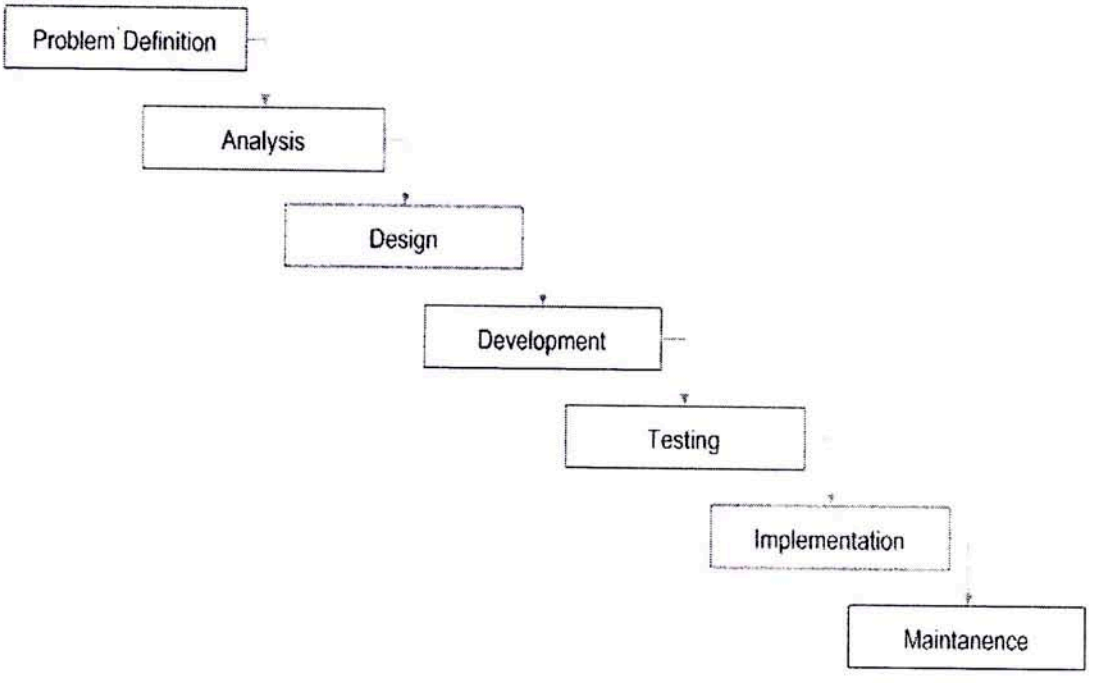
6. การนำไปใช้งานจริง (Implement) หลังจากทดสอบระบบเสร็จสิ้นจึงนำระบบมาติดตั้งให้แก่ผู้ใช้ระบบได้ทดลองใช้จริง ซึ่งถือว่าเป็นขั้นตอนสุดท้ายของระบบของนักวิเคราะห์ที่ต้องตรวจสอบระบบ

7. การบำรุงรักษาพัฒนาระบบต่อ (Maintenance) หลังจากนำระบบใหม่มาติดตั้งให้กับผู้ใช้ระบบ ซึ่งผู้ใช้ระบบบางส่วนที่ยังไม่คุ้นเคยกับการทำงานของระบบใหม่จะต้องมีการอบรมให้คำแนะนำอย่างต่อเนื่อง คอยดูแลบำรุงรักษาฐานข้อมูลและช่วยเหลือผู้ใช้ระบบในการปฏิบัติงาน

ปัจจุบันมีรูปแบบของวงจรการพัฒนาระบบที่หลากหลาย โดยผู้วิจัยได้ศึกษารูปแบบของวงจรการพัฒนาระบบ ดังนี้

1. **วงจรการพัฒนาระบบรูปแบบจำลองน้ำตก** (Waterfall)

มีหลักการเหมือนน้ำตกซึ่งไหลจากที่สูงลงสู่ที่ต่ำที่ไม่สามารถย้อนกลับได้ ซึ่งเมื่อเลือกพัฒนาระบบนี้แล้วจะไม่สามารถย้อนกลับมาที่ขั้นตอนก่อนหน้าได้อีก ดังภาพ 11



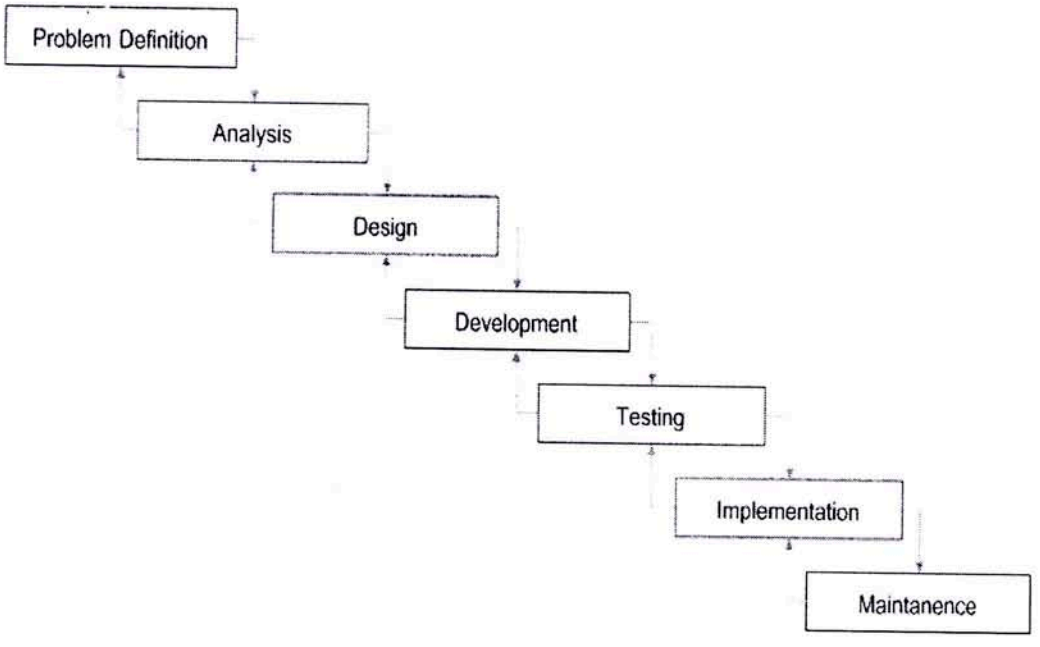
ภาพ 11 วงการพัฒนาระบบรูปแบบจำลองน้ำตก

ข้อจำกัดของรูปแบบการพัฒนาระบบด้วยรูปแบบจำลองน้ำตก

หากมีข้อผิดพลาดเกิดขึ้นในขั้นตอนก่อนหน้าแล้วจะไม่สามารถย้อนกลับมาได้ ดังนั้นการพัฒนาระบบด้วยรูปแบบ ระบบงานที่มีรูปแบบการพัฒนาระบบที่ดีและตายตัวอยู่แล้ว

2. **วงจรการพัฒนาระบบรูปแบบน้ำตกที่ย้อนกลับขั้นตอนได้** (Adapted Waterfall)

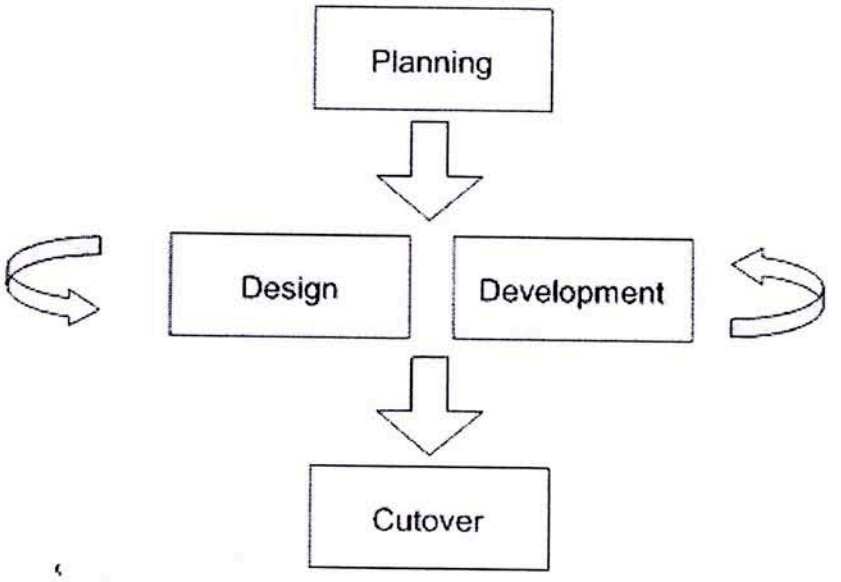
เป็นรูปแบบที่พัฒนามาจากรูปแบบ Waterfall ซึ่งแตกต่างจากรูปแบบเดิม ซึ่งในขณะดำเนินงานอยู่นั้นสามารถย้อนกลับมายังขั้นตอนก่อนหน้า เพื่อแก้ไขข้อผิดพลาดหรือย้อนข้ามขั้นได้ ดังภาพ 12



ภาพ 12 วงการพัฒนาระบบรูปแบบน้ำตกที่ย้อนกลับขั้นตอนได้

3. **การพัฒนาระบบแบบเร่งด่วน** (Rapid Application Development: RAD)

เป็นวิธีการพัฒนาระบบ (Methodology) วิธีการหนึ่งที่รวบรวมเทคนิค เครื่องมือ และเทคโนโลยี มาประยุกต์ใช้ในการสนับสนุนการพัฒนาระบบให้สำเร็จ โดยใช้เวลาน้อยที่สุด ทั้งนี้ขึ้นอยู่กับความพร้อมขององค์กรในขณะนั้น ไม่ว่าจะเป็นค่าใช้จ่าย บุคลากร รวมทั้งความต้องการที่แน่นอนของผู้ใช้ระบบ จากแนวความคิดของวิธีการพัฒนาระบบแบบเร่งด่วน ทำให้มีการผสมผสานและประยุกต์ใช้เทคนิค เครื่องมือ และเทคโนโลยีเข้าด้วยกันเพื่อพัฒนาระบบ โดยมีการแบ่งแยกขั้นตอนในวงจรการพัฒนาระบบให้น้อยลง เลือกใช้เทคนิคการเก็บรวบรวมข้อมูลด้วย JAD และเลือกใช้ตัวต้นแบบ ในการออกแบบ เป็นต้น ทำให้การพัฒนาดำเนินการได้อย่างรวดเร็ว ดังนั้นการพัฒนาระบบแบบเร่งด่วน (RAD) จึงเหมาะกับองค์กรที่มีความพร้อมในการพัฒนา เช่น ความพร้อมทางด้านการเงิน บุคลากร ประกอบกับผู้ใช้ความต้องการแน่นอนไม่เปลี่ยนแปลง และต้องการระบบใหม่โดยใช้เวลาพัฒนาไม่นาน ซึ่งมีขั้นตอนการทำงานของวงจรการพัฒนาระบบแบบเร่งด่วน (RAD) ดังภาพ 13



ภาพ 13 ขั้นตอนการทำงานของวงจรการพัฒนาแบบเร่งด่วน

ขั้นตอนการทำงานของวงจรการพัฒนาแบบเร่งด่วน ประกอบไปด้วย 4 ส่วน ดังนี้

1. การกำหนดความต้องการ เป็นการกำหนดหน้าที่และงานต่าง ๆ ภายในระบบ โดยผู้ใช้และผู้บริหารร่วมสัมมนา

2. การออกแบบโดยผู้ใช้ ผู้ใช้มีส่วนในการออกแบบระบบที่ไม่ใช่ทางเทคนิคคอมพิวเตอร์ เช่น ฟอร์ม หน้าจอ เป็นต้น

3. การสร้างระบบ โดยการใช้ตัวซอฟต์แวร์ประยุกต์อย่างเร็ว (RAD Software) ในการสร้างโปรแกรม

4. การเปลี่ยนระบบ ทำการทดสอบระบบให้เสร็จสิ้นก่อนฝึกอบรมแล้วจึงมีการเปลี่ยนแปลงเครื่องมือในการพัฒนาระบบงานอย่างเร็ว

การพัฒนาระบบแบบเร่งด่วน (RAD) มีวัตถุประสงค์ที่สำคัญคือ ต้องการรวบรวมกระบวนการสำคัญต่าง ๆ เพื่อพัฒนาระบบด้วยระยะเวลาอันสั้น โดยใช้เครื่องมือสำคัญอย่าง CASE Tools (Computer Aided Software Engineering) ซึ่งหมายถึง ซอฟต์แวร์พิเศษสำหรับช่วยในการเขียนโปรแกรม สามารถสร้างโปรแกรมต่าง ๆ จากข้อกำหนดเช่น โปรแกรมบันทึกข้อมูล โปรแกรมแสดงรายงาน โปรแกรมค้นหาฐานข้อมูล โปรแกรมคำนวณ เป็นต้น และการนำเทคนิคการพัฒนาระบบแบบเร่งด่วน (RAD) มาใช้ในการพัฒนาระบบนั้นเพื่อมุ่งหวังที่จะพัฒนาระบบให้สำเร็จอย่างรวดเร็ว

ข้อดีการพัฒนาระบบแบบเร่งด่วน (RAD)

1. สามารถพัฒนาระบบได้อย่างรวดเร็ว เนื่องจากมีการใช้เครื่องมือหรือซอฟต์แวร์ที่สนับสนุนการพัฒนาระบบ (CASE Tools)

2. เพิ่มคุณภาพ เนื่องจากมีการพัฒนาระบบเป็นไปตามความต้องการของผู้ใช้และผู้ใช้ได้มีส่วนร่วมในการวิเคราะห์และออกแบบรวมทั้งทดลองใช้ต้นแบบที่ได้สร้างขึ้น ทำให้ระบบที่ใช้งานนั้นเป็นไปตามความต้องการของผู้ใช้และเพิ่มคุณลักษณะในระหว่างการพัฒนาระบบได้

ข้อจำกัดของการพัฒนาระบบแบบเร่งด่วน (RAD)

1. คุณลักษณะลดลง เนื่องจากถูกจำกัดด้วยระยะเวลาที่ไม่นานสำหรับการพัฒนาระบบ อาจต้องใส่คุณลักษณะในระบบได้ไม่มากนัก เพื่อการพัฒนาระบบเสร็จทันตามกำหนดเวลาที่กำหนด

2. การเปลี่ยนแปลงความต้องการของผู้ใช้อยู่ตลอดเวลา เนื่องจากผู้ใช้ได้ทดลองใช้โปรแกรมต้นแบบที่สามารถสร้างและแก้ไขง่าย ผู้ใช้จึงอาจมีความต้องการที่เปลี่ยนแปลงได้ตลอดเวลา

ในการวิจัยนี้ได้นำแนวคิดเกี่ยวกับหลักการพัฒนาระบบ ซึ่งผู้วิจัยได้เลือกการพัฒนาระบบแบบเร่งด่วนมาประยุกต์ใช้ในการออกแบบและพัฒนาระบบให้มีประสิทธิภาพในการใช้งานสืบค้นข้อมูลและการพยากรณ์ เพื่อที่จะนำไปใช้ในการพัฒนาระบบสารสนเทศการสืบค้นข้อมูลและโปรแกรมพยากรณ์ผลผลิตมันสำปะหลังต่อไป

งานวิจัยที่เกี่ยวข้อง

บดินทร์ มิลินทางกูร และสุปิยา เจริญศิริวัฒน์. (2553) ได้ทำการวิจัยเรื่อง การวิเคราะห์หาความเสี่ยงต่อการเกิดโรคจากขนาดรูปร่างด้วยเทคนิควิธีเชิงพันธุกรรมร่วมกับเทคนิคต้นไม้ช่วยตัดสินใจ พบว่าผลเลือดไตรกลีเซอไรด์ มีผลต่อปัญหาสุขภาพในหลายด้าน ถ้ามีค่าไตรกลีเซอไรด์สูงกว่าค่ามาตรฐานมากอัตราความเสี่ยงต่อการเกิดโรคจึงสูงขึ้นตามไปด้วย ขณะนี้มีเครื่องตรวจวัดเรือนร่างสามมิติ ถือเป็นแนวทางหนึ่งในการหาวิธีการเฝ้าระวังตนเองจากความเสี่ยงต่อการเกิดโรคต่าง ๆ งานวิจัยนี้ใช้เทคนิคขั้นตอนเชิงพันธุกรรมและเทคนิคต้นไม้ตัดสินใจในการหาความสัมพันธ์ของขนาดรูปร่าง โดยนำมาสร้างเป็นหลักเกณฑ์สำหรับใช้วิเคราะห์ระดับไตรกลีเซอไรด์จากการทดสอบผลลัพธ์ที่ได้มีความแม่นยำสูงในการหาความเสี่ยงต่อการเกิดโรคที่มีโอกาสได้จากไตรกลีเซอไรด์

จากการศึกษางานวิจัยนี้ ผู้วิจัยได้นำแนวคิดเกี่ยวกับเทคนิคต้นไม้ตัดสินใจมาประยุกต์ใช้เป็นแนวทางในการทำเหมืองข้อมูลในการพยากรณ์ผลผลิตมันสำปะหลังต่อไป

ลลิตา ศรีชัยญา (2553) ได้ทำการวิจัยเรื่อง เทคนิคการจำแนกข้อมูลด้วยต้นไม้เชิงสัมพันธ์ พบว่างานวิจัยนี้เป็นการนำเสนอ วิธีการคัดเลือกลักษณะเด่นของข้อมูลเพื่อให้ได้คุณลักษณะของข้อมูลฝึกที่เหมาะสมที่สุดสำหรับการเรียนรู้ โดยใช้ต้นไม้ตัดสินใจ เพื่อให้ได้ต้นไม้ตัดสินใจที่มีประสิทธิภาพในการจำแนกข้อมูล โดยการนำเอาเทคนิคการหากฎความสัมพันธ์เข้ามาใช้ร่วมกันกับต้นไม่ตัดสินใจ จึงได้ทำการทดลองกับชุดข้อมูล 3 ชุด ข้อมูลที่ได้จากฐานข้อมูลมาตรฐาน UCL Machine Learning Repository โดยทำการทดลองวัดประสิทธิภาพเปรียบเทียบกับอัลกอริทึม C4.5, CBA, และ CMAR ซึ่งผลการทดลองพบว่าอัลกอริทึมที่นำเสนอมีประสิทธิภาพที่ดีกว่าใน 2 ชุดข้อมูลจาก 3 ชุดข้อมูล จากการศึกษางานวิจัยนี้ได้ศึกษาและใช้เป็นแนวทางในการวัดประสิทธิภาพของอัลกอริทึมที่ใช้ในการตัดสินใจในแบบต่าง ๆ เพื่อหาความแม่นยำและอัลกอริทึมที่มีความเหมาะสมที่สุด

จากการศึกษางานวิจัยนี้ ผู้วิจัยได้ศึกษาและใช้เป็นแนวทางในการวัดประสิทธิภาพของอัลกอริทึมที่ใช้ในการตัดสินใจในแบบต่าง ๆ เพื่อหาความแม่นยำและอัลกอริทึมที่มีความเหมาะสมที่สุด

มลธิดา ฤทธิ์สมบูรณ์ และสุชา สมาชาติ (2551)ได้ทำการวิจัยเรื่อง การพัฒนาระบบสนับสนุนการพิจารณาอนุมัติให้สินเชื่อเพื่อการเช่าซื้อสินค้าโดยใช้ต้นไม้ตัดสินใจ มีวัตถุประสงค์เพื่อสร้างระบบสนับสนุนการพิจารณาต้นไม้ตัดสินใจและใช้อัลกอริธึม ID3 และใช้โปรแกรมเวกาในการทดลองเพื่อสร้างโมเดลโครงสร้างต้นไม้สำหรับตัดสินใจ จากนั้นจึงนำโมเดลที่ได้ไปพัฒนาขึ้นในลักษณะเว็บแอพพลิเคชั่น หลังจากพัฒนาระบบเสร็จ ประเมินแบบสอบถามประเมินความพึงพอใจการใช้งานระบบ ผลการประเมินจากผู้เชี่ยวชาญได้ค่าเฉลี่ยเท่ากับ 4.21 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 0.63 และผู้ใช้งานทั่วไปได้ค่าเฉลี่ยเท่ากับ 4.21 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 0.65 ซึ่งสรุปได้ว่าระบบที่พัฒนาขึ้นมีความพึงพอใจในการใช้งานอยู่ในระดับดี

จากการศึกษางานวิจัยนี้ ผู้วิจัยได้นำแนวคิดในการพัฒนาเว็บแอปพลิเคชั่นด้วยการพัฒนาระบบ โดยใช้ภาษา PHP และฐานข้อมูล MySQL มาประยุกต์ใช้ในงานวิจัยนี้ โดยภาษา PHP และฐานข้อมูล MySQL เป็นภาษาและฐานข้อมูลที่เป็นลักษณะโอเพนซอร์ส ที่ทุกคนสามารถแก้ไขการเขียนโค้ดให้เหมาะสมกับงานที่ต้องการได้ ส่งผลให้สามารถใช้เป็นฐานข้อมูลสำหรับการสืบค้นข้อมูลค้นและใช้ตัวแบบพยากรณ์ที่ได้จากต้นไม้ตัดสินใจมาพัฒนาเป็นโปรแกรมพยากรณ์ผลผลิตมันสำปะหลังและนำไปใช้งานผ่านเว็บแอปพลิเคชั่นได้

ปฏิยากร โคผดุง (2551) ได้ทำการวิจัยเรื่อง ต้นไม้ตัดสินใจสำหรับการวิเคราะห์ภาวะตัวรับฮอร์โมนของผู้ป่วยมะเร็งเต้านมและอัตราการอยู่รอดชีวิตของผู้ป่วยมะเร็งเต้านมในโรงพยาบาล

ขอนแก่น พบว่าโรคมะเร็งเต้านมเป็นสาเหตุการตายมากเป็นอันดับสองในหญิงไทย ชุดข้อมูลจากข้อมูลผู้ป่วยเพศหญิง ที่มีผลการวินิจฉัยโรคเป็นมะเร็งเต้านมและเข้ารับการรักษาในโรงพยาบาลขอนแก่นตั้งแต่วันที่ 1 มกราคม พ.ศ. 2550 วันที่ 31 ธันวาคม พ.ศ. 2550 ซึ่งงานวิจัยนี้ยกเว้นข้อมูลผู้ป่วยที่ไม่สามารถติดตามการรักษาหรือส่งต่อการรักษาไปที่สถานพยาบาลอื่น วัตถุประสงค์ในการวิจัยเพื่อใช้ในการจำแนกระดับภาวะตัวรับฮอร์โมนสำหรับแนวทางการรักษาผู้ป่วยมะเร็งเต้านมและใช้เป็นระบบสนับสนุนการตัดสินใจด้านคลินิก โดยการประยุกต์ใช้อัลกอริทึม J48 ในโปรแกรมเวกา สร้างตัวแบบต้นไม้ตัดสินใจโดยใช้ทฤษฎีการทำเหมืองข้อมูลกับข้อมูลผู้ป่วยจำนวน 57 คน ซึ่งทำการวิเคราะห์ข้อมูลจำนวน 57 instance มีแอททริบิวต์จำนวน 6 แอททริบิวต์ สำหรับการประเมินความถูกต้องของตัวแบบที่ใช้ทำนายใช้วิธี 10-fold cross validation และวิธี Train test พบว่าวิธี 10-fold cross validation มีความถูกต้อง 57.89 % และ วิธี Train test มีความถูกต้อง 66.67% ซึ่งวิธีนี้จะมีประสิทธิภาพสูงสุด ความรู้ที่ได้จากการทำเหมืองข้อมูลนี้จะเป็นประโยชน์ต่อแพทย์ผู้เชี่ยวชาญในการพัฒนาตัวแบบการทำนายค่าและสนับสนุนกระบวนการตัดสินใจของแพทย์ได้อย่างถูกต้อง โดยเฉพาะการประยุกต์ใช้การจำแนกภาวะตัวรับฮอร์โมนในการรักษามะเร็งเต้านมซึ่งข้อมูลนี้สามารถทำนายการแพร่กระจายของมะเร็งเต้านมไปยังต่อมน้ำเหลือง

ผู้วิจัยได้นำแนวคิดเกี่ยวกับเทคนิคต้นไม้ตัดสินใจมาประยุกต์ใช้เป็นแนวทางในการทำเหมืองข้อมูลและได้มีการประยุกต์ใช้การแบ่งข้อมูลการเรียนรู้ออกเป็น 10 ส่วน เท่าๆ กัน (10 Fold Cross Validation) ในการสร้างต้นไม้ตัดสินใจขึ้นเพื่อนำไปใช้กับการพยากรณ์ผลผลิตมันสำปะหลังต่อไป

ไพฑูรย์ จันทร์เรือง (2550) ได้ทำการวิจัยเรื่อง ระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคต้นไม้ พบว่าการเลือกสาขาการเรียนของผู้ที่ต้องการสมัครเข้าศึกษาต่อในระดับปริญญาตรี ในสถาบันอุดมศึกษาต่าง ๆ เป็นสิ่งที่สำคัญเนื่องจากส่งผลโดยตรงกับผู้สมัครเองทางด้านคุณภาพหลังจบการศึกษา งานวิจัยนี้ได้พัฒนาระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคต้นไม้ตัดสินใจ ซึ่งจากการทดลองพบว่าในการสร้างตัวแบบสำหรับพัฒนาระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคต้นไม้ตัดสินใจนั้น ควรแยกสร้างตัวแบบสำหรับแต่ละสาขาการเรียน เนื่องจากคุณสมบัติของผู้เรียนแต่ละสาขามีความแตกต่างกันเพื่อให้ได้ตัวแบบที่สามารถทำนายแนวโน้มของผลการเรียนที่เหมาะสมสำหรับแต่ละสาขา แต่เนื่องจากคะแนนเฉลี่ยของนักศึกษาที่นำมาพัฒนาตัวแบบนั้น ส่วนใหญ่จะมีเกณฑ์คะแนนเกาะ

กลุ่มกันอยู่ในช่วงกลางของข้อมูล (2.00 - 3.00) ทำให้ผลการตัดสินส่วนใหญ่จะโน้มเอียงไปในเกณฑ์พอใช้ (ช่วงคะแนน 2.00 - 2.49) และปานกลาง (ช่วงคะแนน 2.50 - 2.99)

จากการศึกษางานวิจัยนี้ ผู้วิจัยได้ศึกษาการสร้างตัวแบบเพื่อนำไปพัฒนาระบบสารสนเทศการสืบค้นข้อมูลและโปรแกรมพยากรณ์ผลผลิตมันสำปะหลังและได้นำแนวคิดการทำนายผลแบบค่าเป็นระดับ (ค่าช่วง) มาประยุกต์ใช้ในการพยากรณ์ผลผลิตมันสำปะหลังในเขตภาคเหนือตอนล่างได้

พิจิตรา จอมศรี (2549) ได้ทำการวิจัยเรื่อง การทำนายเนื้อหาของเว็บ โดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปากร พบว่างานวิจัยนี้ได้นำเทคนิคกฎการค้นหาความสัมพันธ์ข้อมูลมาประยุกต์ใช้สร้างตัวแบบเพื่อทำนายข้อมูลการเรียกใช้เว็บในอนาคต โดยผู้วิจัยได้จัดเก็บข้อมูล 2 ส่วนคือการเรียกใช้เว็บภายในมหาวิทยาลัยศิลปากรจากระบบพร็อกซี เซิร์ฟเวอร์ และจัดทำฐานข้อมูลเว็บเพื่อจัดหมวดหมู่ของเว็บ แล้วนำข้อมูลทั้ง 2 ส่วนมาสร้างความสัมพันธ์ซึ่งงานวิจัยนี้นำข้อมูลวัน เวลา หมวดเว็บ และเว็บมาค้นหาความสัมพันธ์เพื่อสร้างตัวแบบและพิจารณาตัวแบบจากค่าความเชื่อมั่นและค่าสนับสนุนโดยผลของการศึกษาตัวแบบพบว่าโมเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ในวันถัดมาได้ ผลของการใช้เทคนิคเหมืองข้อมูลพบว่าตัวแบบที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ได้ โดยมีความถูกต้องร้อยละ 66.67 % นอกจากนี้ยังนำเสนอระบบการทำนายเนื้อหาของเว็บ โดยใช้เทคนิคเหมืองข้อมูลซึ่งสร้างอยู่บนขั้นตอนวิธีที่นำเสนอและอธิบายถึงผลการทดลองบนข้อมูลจริงด้วย

จากการศึกษางานวิจัยนี้ ผู้วิจัยได้ศึกษากฎการค้นหาความสัมพันธ์ ซึ่งเป็นเทคนิคหนึ่งในเทคนิคเหมืองข้อมูล สามารถนำมาประยุกต์ใช้เพื่อสร้างตัวแบบเพื่อพยากรณ์ผลผลิตได้

จารุวรรณ วีระเศรษฐกุล (2548) ได้ทำการวิจัยเรื่อง ศักยภาพและแนวโน้มการผลิตและการตลาดของมันสำปะหลังในภาคตะวันออกเฉียงเหนือตอนล่าง พบว่ามันสำปะหลังแปรรูปเป็นผลิตภัณฑ์ได้หลายประเภทและเกษตรที่มีพื้นที่ปลูกมันสำปะหลังจำกัดและได้รับผลผลผลิตต่อไร่น้อยซึ่งปัญหาหลักในด้านการผลิตคือ สภาพภูมิอากาศ และปัจจัยสำคัญที่มีผลกระทบต่อผลผลิตมันสำปะหลังหัวสดมากที่สุด คือราคามันสำปะหลังหัวสดและพื้นที่ในการปลูกมันสำปะหลัง ส่วนปัจจัยสำคัญที่มีผลต่อความต้องการซื้อมันสำปะหลังหัวสดของผู้ประกอบการลานมันและอุตสาหกรรม คือราคามันสำปะหลังหัวสดและราคาของผลิตภัณฑ์ซึ่งแนวโน้มความต้องการมันสำปะหลังหัวสดเพื่อนำไปแปรรูปมีปริมาณสูงขึ้นและคาดว่ายังคงมีความต้องการสูงขึ้นในอนาคต ขณะที่ผลผลิตมีแนวโน้มลดลงทำให้ผู้ประกอบการธุรกิจเกี่ยวกับมันสำปะหลังประสบกับสภาวะขาดแคลนวัตถุดิบ เกิดการแข่งขันด้านตลาดเพื่อให้ได้มาซึ่งวัตถุดิบ เกษตรกรและผู้ประกอบการ

ธุรกิจมันสำปะหลังต้องพัฒนาเทคโนโลยีการผลิตเพื่อให้การปลูกและการผลิตผลิตภัณฑ์มันสำปะหลังมีประสิทธิภาพสามารถแข่งขันได้ในภาวะตลาดที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา

จากการศึกษางานวิจัยนี้ ผู้วิจัยได้ศึกษาและรวบรวมความรู้ของมันสำปะหลัง เพื่อเป็นแนวทางในการศึกษาสภาพการเจริญเติบโตและประโยชน์ของมันสำปะหลังต่อไป

ดังนั้น จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่ายังไม่มีระบบการสืบค้นข้อมูลและโปรแกรมพยากรณ์ผลผลิตมันสำปะหลังในเขตภาคเหนือตอนล่างผ่านเว็บแอพพลิเคชั่น แต่จะมีการพยากรณ์ในรูปแบบอื่นแทน ดังนั้นผู้วิจัยจึงสนใจที่จะแก้ปัญหาโดยการประยุกต์ใช้เทคนิคเหมืองข้อมูล ในการวิเคราะห์ผลผลิตมันสำปะหลังในเขตภาคเหนือตอนล่าง โดยสร้างตัวแบบพยากรณ์ผลผลิตมันสำปะหลังด้วยการจำแนกแบบต้นไม้ตัดสินใจ โดยพิจารณาปัจจัยต่าง ๆ ที่มีส่วนเกี่ยวข้องเพื่อสร้างตัวแบบพยากรณ์ขึ้น แล้วนำตัวแบบพยากรณ์มาสร้างเป็นโปรแกรมพยากรณ์ผลผลิตมันสำปะหลังและใช้งานผ่านเว็บแอพพลิเคชั่น เพื่อเป็นประโยชน์แก่เจ้าหน้าที่สำนักงานเศรษฐกิจการเกษตรเขต 2 นำไปใช้เป็นแนวทางในการกำหนดนโยบาย เพื่อส่งเสริมและสนับสนุนการเพาะปลูกมันสำปะหลัง นอกจากนี้เกษตรกรหรือผู้สนใจทั่วไปยังสามารถใช้เป็นแนวทางในการศึกษาเพิ่มเติมได้ต่อไป