บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การศึกษาและทำวิจัยการพัฒนาระบบสารสนเทศเพื่อการพยากรณ์ผู้เข้าศึกษาโดยผ่านเครือข่ายอินเทอร์เน็ต ผู้วิจัยได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องดังนี้

2.1 การทำเหมืองข้อมูล

2.2 กฎการจำแนก

2.3 ต้นไม้ตัดสินใจ

2.4 อินเทอร์เน็ต

2.5 ภาษาเอชทีเอ็มแอล

2.6 วิเคราะห์และออกแบบด้วยยูเอ็มแอล

2.7 ฐานข้อมูลมายเอสคิวแอล

2.8 ภาษาพีเอชพี

2.9 งานวิจัยที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) เป็นการกลั่นกรองข้อมูลและสืบค้นความรู้ที่เป็นประโยชน์จากฐานข้อมูลขนาดใหญ่ (Large Information) ขบวนการในการทำเหมืองข้อมูลเป็นขั้นตอนที่สำคัญในการค้นหาองค์ความรู้ในฐานข้อมูล (Knowledge Discovery in Database : KDD) เป็นการนำข้อมูลจำนวนมากที่มีอยู่มาวิเคราะห์และสืบค้นความรู้ สารสนเทศหรือสิ่งที่สำคัญที่ซ่อนอยู่ เพื่อให้ได้สารสนเทศที่สามารถนำไปใช้งานได้ (Howard, H., et al., 2010) สำหรับประกอบการตัดสินใจต่าง ๆ เช่น การวิเคราะห์พฤติกรรมผู้บริโภคในการเลือกซื้อสินค้าของห้างสรรพสินค้าเพื่อนำไปใช้ในจัดรายการโปรโมชั่นเพื่อการส่งเสริมการขายสินค้าในการดำเนินงานของห้างสรรพสินค้า เป็นต้น

2.1.1 ขั้นตอนการทำเหมืองข้อมูล การทำเหมืองข้อมูลเป็นวิธีการที่ให้ได้มาซึ่งองค์ความรู้ที่ซ้อนอยู่ในฐานข้อมูลขนาดใหญ่ โดยมีขั้นตอนการทำเหมืองข้อมูลดังนี้

(1) การทำความสะอาดข้อมูล (Data Cleaning and Pre-Processing) ข้อมูลที่นำมาใช้ในการทำเหมืองข้อมูลมีปริมาณมาก จำเป็นต้องมีขั้นตอนในการกรองข้อมูลเพื่อเลือกข้อมูลที่สามารถนำไปสู่ความรู้ที่ต้องการได้ตรงประเด็นหรือตรงความต้องการ โดยไม่นำข้อมูลที่ไม่ครบสมบูรณ์มาใช้ในการประมวลผลเพื่อจัดทำเหมืองข้อมูล ทำให้ได้ข้อมูลที่มีคุณภาพ จากนั้นจะนำ

ข้อมูลที่ผ่านการเลือกมาทำความสะอาดข้อมูล โดยการแก้ไขข้อมูลให้ถูกต้องสมบูรณ์ด้วยการปรับค่าให้อยู่ในขอบเขตที่ถูกต้องหรือกำจัดค่าข้อมูลที่ไม่สมบูรณ์ด้วยวิธีการเติมข้อมูลตามที่ควรจะเป็นลงไปเพิ่มเติม เช่น การแก้ไขค่าว่าง (Null) ของข้อมูลโดยอาจใส่ค่าศูนย์กรณีนำไปรวมค่า เพื่อไม่ให้มีผลกระทบต่อการรวมค่าข้อมูลหรือใส่ค่าหนึ่งแทนกรณีที่ต้องนำไปคำนวณค่าทางคณิตศาสตร์คูณหรือหาร เพื่อไม่ให้มีผลกระทบต่อการคำนวณค่าข้อมูล เป็นต้น ในขั้นตอนนี้ช่วยให้ได้ข้อมูลที่มีคุณภาพมากยิ่งขึ้นสำหรับนำไปวิเคราะห์

(2) การรวบรวมข้อมูล (Data Integration) การจัดทำเหมืองข้อมูลต้องใช้ข้อมูลปริมาณมาก อาจจำเป็นต้องมีการรวบรวมข้อมูลทั้งหมดที่ต้องการจากหลาย ๆ แหล่งข้อมูล ซึ่งข้อมูลอาจจัดเก็บอยู่ในรูปแบบที่แตกต่างกันออกไปหรืออยู่ในหลายฐานข้อมูล ซึ่งต้องมีการรวบรวมข้อมูลแล้วทำให้ข้อมูลอยู่ในรูปแบบเดียวกันก่อน ซึ่งในขั้นตอนนี้ผลลัพธ์ที่ได้มักถูกจัดเก็บไว้ในคลังข้อมูล (Data Warehouses)

(3) การคัดเลือกข้อมูล (Data Selection) เป็นวิธีการเลือกและดึงเอาเฉพาะข้อมูลที่เกี่ยวข้องหรือที่ต้องการจะนำไปวิเคราะห์มาใช้ (Task-relevant Data) โดยระบุถึงแหล่งของข้อมูลที่จะนำมาทำเหมืองข้อมูลและแนวทางการนำข้อมูลที่ต้องการออกจากฐานข้อมูล เพื่อนำไปสร้างกลุ่มของข้อมูลสำหรับใช้ในการพิจารณาหรือค้นหาองค์ความรู้ต่อไป

(4) การแปลงข้อมูล (Data Transformation) เป็นวิธีการปรับเปลี่ยนข้อมูลให้มีค่าที่เหมาะสมสำหรับการทำเหมืองข้อมูลในด้านการคำนวณและการสืบค้นในกระบวนการทำเหมืองข้อมูล ตลอดจนให้ข้อมูลนั้นสื่อความหมายในการประกอบการตัดสินใจ เช่น ข้อมูลของสินค้าเป็นข้อมูลที่มีค่า “ใกล้ชิด” และ “คอลเกต” มีการเปลี่ยนค่าให้เป็น “ยาสีฟัน” หรือการแบ่งกลุ่มรายได้ของลูกค้าออกเป็นช่วงรายได้ จะช่วยให้การวิเคราะห์และสืบค้นความสัมพันธ์ที่เกิดขึ้นในฐานข้อมูลทำได้โดยง่ายและสร้างความเหมาะสมในการตัดสินใจมากขึ้น

(5) การจัดรูปแบบข้อมูล (Data Transaction Identification) เป็นวิธีการจัดข้อมูลให้อยู่ในรูปแบบที่เหมาะสมและถูกต้องเพื่อให้สามารถนำไปเข้าสู่กระบวนการจัดทำเหมืองข้อมูล วิธีนี้นิยมใช้ในการทำข้อมูลให้อยู่ในรูปแบบตาราง (Table) มีลักษณะเป็นแถวและสดมภ์หรือคอลัมน์ที่มีความสัมพันธ์กัน

(6) การค้นหารูปแบบ (Pattern Discovery) เป็นวิธีการกำหนดรูปแบบในการวิเคราะห์และสืบค้นเพื่อให้ได้ผลลัพธ์ที่ต้องการ สามารถแบ่งเป็น รูปแบบการวิเคราะห์ (Pattern Analysis) กฎจำแนกเชิงความสัมพันธ์ (Association Rules) การจำแนกประเภทข้อมูลและการพยากรณ์ (Classification and Prediction) การแบ่งกลุ่มข้อมูลแบบอัตโนมัติ (Clustering) และรูปแบบการทำงานตามลำดับ (Sequential Patterns) เป็นต้น

(7) การวิเคราะห์รูปแบบ (Pattern Analysis) เป็นการนำผลลัพธ์จากการสืบค้นข้อมูลที่สัมพันธ์กันที่ผ่านกระบวนการจัดทำเหมืองข้อมูลมาทำการวิเคราะห์เพื่อช่วยประกอบการตัดสินใจหรือการวางแผนการดำเนินงานต่าง ๆ ในอนาคต

2.1.2 สถาปัตยกรรมระบบการทำเหมืองข้อมูล การทำเหมืองข้อมูลมีสถาปัตยกรรมของระบบที่มีลักษณะลูกข่ายแม่ข่าย (Client-Server) ประกอบด้วย ฐานข้อมูลและแหล่งข้อมูลอื่นที่จัดเป็นคลังข้อมูล ข้อมูลในคลังข้อมูลเป็นข้อมูลเริ่มต้นสำหรับนำไปใช้ในการทำเหมืองข้อมูล โดยมีฐานข้อมูลแม่ข่ายซึ่งก็คือคลังข้อมูลแม่ข่ายทำหน้าที่นำเข้าข้อมูลจากแหล่งข้อมูลต่าง ๆ ที่ผ่านขั้นตอนการทำความสะอาด การรวบรวมข้อมูลและการเลือกข้อมูลมาเรียบร้อยแล้ว (John, G. Hendricks, 2000) เพื่อนำไปให้เอนจิ้นของเหมืองข้อมูล (Data Mining Engine) ทำการประมวลผลในส่วนของการหากฎความสัมพันธ์ การจำแนกประเภทและการจัดกลุ่มต่อไป เป็นต้น ทั้งนี้อาจจะอาศัยองค์ความรู้ที่มีอยู่หรือจากการสังเคราะห์ความรู้ได้ใหม่ในขั้นตอนนี้ไปช่วยประมวลผลตามรูปแบบของผลลัพธ์ที่ได้ แล้วนำเสนอความรู้ที่ได้ผ่านส่วนติดต่อผู้ใช้งาน เพื่อให้ผู้ใช้สามารถนำไปประกอบการตัดสินใจจากความรู้ที่ได้จากการสังเคราะห์

2.1.3 ประเภทข้อมูลที่ใช้ทำเหมืองข้อมูล จากสถาปัตยกรรมระบบการทำเหมืองข้อมูลจำเป็นต้องนำข้อมูลจากแหล่งข้อมูลต่าง ๆ มารวบรวมไว้ก่อนจะนำไปสู่กระบวนการประมวลผลในการสืบค้นความรู้โดยส่วนมากข้อมูลจะจัดเก็บอยู่ในรูปแบบดังนี้

(1) ฐานข้อมูลรายการ (Transactional Database) คือ ฐานข้อมูลที่ประกอบด้วยข้อมูลที่จัดเก็บได้จากการปฏิบัติการประจำวันในขั้นตอนการดำเนินงานของผู้ใช้งานระบบ โดยที่แต่รายการจะเป็นเหตุการณ์ในขณะใดขณะหนึ่งในการปฏิบัติงานประจำวัน เช่น รายการใบเสร็จรับเงิน จะเก็บข้อมูลที่ประกอบไปด้วยรหัสใบเสร็จรับเงิน รหัสลูกค้า ชื่อลูกค้า วัน เวลาที่ซื้อสินค้าและรายการสินค้า เป็นต้น

(2) ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) คือ ฐานข้อมูลที่จัดเก็บข้อมูลไว้ในรูปแบบของตารางที่มีความสัมพันธ์กัน โดยแต่ละตารางจะประกอบไปด้วยแถว (Row) และสดมภ์ (Column) ที่มีความสัมพันธ์กัน ข้อมูลที่จัดเก็บในฐานข้อมูลเชิงสัมพันธ์ทั้งหมดจะถูกแสดงไว้ภายใต้แบบจำลองเชิงสัมพันธ์ (Entity Relationship Model)

(3) ฐานข้อมูลขั้นสูง (Advanced Database) คือ ฐานข้อมูลที่จัดเก็บข้อมูลในรูปแบบอื่น ๆ เช่น ข้อมูลเชิงวัตถุ (Object-Oriented) ข้อมูลที่อยู่ในแฟ้มข้อความ (Text File) ข้อมูลมัลติมีเดีย เช่น เสียง รูปภาพและภาพเคลื่อนไหว เป็นต้น

(4) คลังข้อมูล (Data Warehouse) คือ ฐานข้อมูลขนาดใหญ่ที่รวบรวมฐานข้อมูลจากหลายแหล่งหลายช่วงเวลา ซึ่งอาจมีโครงสร้างการจัดเก็บที่แตกต่างกันมารวมไว้ที่เดียวกันและเป็น

ที่สำหรับจัดเก็บและรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ภายใต้รูปแบบข้อมูลเดียวกัน ทำให้สะดวกต่อการนำไปใช้งาน ส่วนมากในการรวบรวมฐานข้อมูลนิยมกำหนดความถี่ในการจัดเก็บและรวบรวมข้อมูลไว้ในคลังข้อมูล เช่น วันละครั้ง สัปดาห์ละครั้งหรือเดือนละครั้ง เป็นต้น

2.1.4 การขุดค้นข้อมูลจากเหมืองข้อมูล การทำเหมืองข้อมูลมีฟังก์ชันการทำงานสำหรับการขุดค้นข้อมูลดังนี้

(1) การวิเคราะห์คุณสมบัติ (Characterization) และการแยกแยะ (Discrimination) ข้อมูล มีรายละเอียดดังนี้

(ก) การวิเคราะห์คุณสมบัติ เป็นการค้นหาคุณลักษณะทั่วไปหรือภาพรวมของข้อมูล เช่น การค้นลักษณะทั่วไปของการซื้อแพ็คเกจการตรวจสุขภาพของเพศหญิงที่มีบุตรยากที่มียอดเพิ่มขึ้นมา 20%

(ข) การแยกแยะข้อมูล เป็นการเปรียบเทียบคุณลักษณะทั่วไปหรือเปรียบเทียบภาพรวมระหว่างข้อมูลตั้งแต่สองชุดขึ้นไป เช่น การซื้อแพ็คเกจการตรวจสุขภาพของเพศหญิงที่มีบุตรยากที่มียอดเพิ่มขึ้นมา 20% กับการซื้อแพ็คเกจการตรวจสุขภาพของเพศหญิงแบบอื่นที่มียอดขายลดลง 10% เป็นต้น

(2) การหาความสัมพันธ์ของข้อมูล (Association) เป็นการจัดแบ่งข้อมูลโดยระบุตามความสัมพันธ์ของข้อมูลที่มีผลต่อกัน ซึ่งข้อมูลจะผ่านการกลั่นกรองข้อมูลก่อนแล้วจึงนำมาจัดรวมเป็นกลุ่มของข้อมูลจำพวกเดียวกัน ส่วนมากนิยมใช้ในการค้นหาความสัมพันธ์ของข้อมูลที่มีขนาดใหญ่เพื่อจัดกลุ่มข้อมูลที่มีความสัมพันธ์กันให้สามารถเห็นข้อมูลเป็นกลุ่มที่มีขนาดเล็กลง เพื่อให้ง่ายต่อการนำข้อมูลที่จัดกลุ่มแล้วไปวิเคราะห์หรือพยากรณ์ปรากฏการณ์ต่าง ๆ เช่น การวิเคราะห์การซื้อสินค้าของลูกค้า (Market Basket Analysis) ในห้างสรรพสินค้า

(ก) การจัดหมวดหมู่ (Classification) ข้อมูลและการวิเคราะห์การถดถอย (Regression) เป็นการจัดกลุ่มรายการอ้างอิงที่มีการจำแนกประเภทและจัดกลุ่มของรายการ เช่น บริษัทแห่งหนึ่งต้องการทราบเหตุผลใดที่ลูกค้าบางกลุ่มยังซื่อสัตย์และภักดีต่อยี่ห้อสินค้า (Brand Loyalty) ของบริษัทและในขณะที่ลูกค้าอีกกลุ่มกลับเปลี่ยนใจไปซื้อสินค้าของคู่แข่งขัน การค้นหาคำตอบนี้ นักวิเคราะห์ต้องค้นหาลักษณะนิสัยของลูกค้าที่เปลี่ยนใจไปซื้อสินค้าของคู่แข่งหรือค้นหาลักษณะของลูกค้าที่ซื่อสัตย์และภักดีต่อยี่ห้อสินค้า โดยอาศัยข้อมูลในการซื้อสินค้าของลูกค้าในอดีตมาทดสอบกับแบบจำลองเพื่อวิเคราะห์ผลว่า สาเหตุใดที่ลูกค้าบางกลุ่มซื่อสัตย์และบางกลุ่มไม่ซื่อสัตย์ต่อยี่ห้อสินค้าขององค์กร เพื่อไม่ให้บริษัทต้องสูญเสียลูกค้าไปให้แก่คู่แข่งขัน

(ข) การวิเคราะห์การรวมกลุ่มหรือการแบ่งแยกข้อมูล (Cluster Analysis or Segmentation) เป็นการนำข้อมูลที่มีลักษณะรูปแบบหรือแนวโน้มที่เหมือนกันมารวมเป็นกลุ่มเดียวกัน เพื่อช่วยในการลดขนาดข้อมูลที่มีจำนวนมาก การวิเคราะห์การรวมกลุ่มหรือการแบ่งแยกข้อมูลจะไม่มีผลลัพธ์ (Output) ตัวแปรอิสระ (Independent Variable) และไม่มีการจัดโครงร่างของวัตถุ เป็นเทคนิคที่เรียนรู้จากข้อมูลโดยไม่ต้องอาศัยการสอน (Unsupervised Learning) อาศัยพื้นฐานของข้อมูลในอดีตเป็นฐานในการวิเคราะห์และแบ่งแยก เช่น เมื่อบริษัทต้องการทราบลักษณะความเหมือนที่มีในกลุ่มของลูกค้า โดยค้นหาลักษณะเฉพาะของลูกค้ากลุ่มเป้าหมาย แล้วนำมาจัดรวมกลุ่มเพื่อสร้างเป็นกลุ่มของลูกค้าที่มีลักษณะเดียวกัน ทำให้บริษัทสามารถทำการแยกกลุ่มของข้อมูลลูกค้าออกเป็นกลุ่ม ๆ เพื่อนำเสนอรายการสินค้าในการส่งเสริมการขายสินค้าในอนาคตให้กับกลุ่มลูกค้าที่ได้จัดแบ่งไว้

(3) การประเมิน (Estimation) และการพยากรณ์ (Prediction) มีรายละเอียดดังนี้

(ก) การประเมิน เป็นการประเมินข้อมูลที่ไม่สามารถกำหนดค่าหรือคุณสมบัติที่ชัดเจนได้จากข้อมูลจำนวนมาก เพื่อใช้จัดการกับข้อมูลที่มีผลแบบต่อเนื่อง เช่น การประเมินจำนวนเด็กชายต่อหนึ่งครอบครัว การประเมินรายได้ของบิดาต่อครัวเรือนและการประเมินน้ำหนักของบุคคลในครอบครัว เป็นต้น

(ข) การพยากรณ์ มีลักษณะคล้ายกับการจัดหมวดหมู่ข้อมูลและการประเมิน แต่จะแตกต่างกันตรงที่ข้อมูลจะถูกแยกจัดลำดับในการพยากรณ์ค่าในอนาคต โดยอาศัยข้อมูลในอดีตมาสร้างเป็นแบบจำลอง เพื่อนำใช้ไปใช้ในการพยากรณ์สิ่งที่จะเกิดขึ้นในอนาคต เช่น การพยากรณ์ว่าลูกค้าลักษณะแบบใดหรือกลุ่มใดที่บริษัทอาจจะสูญเสียไปในอีกหกเดือนข้างหน้า หรือการพยากรณ์ยอดซื้อสินค้าจากลูกค้าเมื่อบริษัทมีการลดราคาสินค้าลงอีก 20% เป็นต้น

(4) การบรรยาย (Description) และการแสดงภาพของข้อมูล (Visualization) มีรายละเอียดดังนี้

(ก) การบรรยาย เป็นการหาคำอธิบายถึงสิ่งที่จะเกิดขึ้น โดยอาศัยข้อมูลจากฐานข้อมูลเดิมที่มีอยู่มาใช้ประกอบการอธิบาย เช่น กลุ่มลูกค้าที่มีการศึกษาหรือรายได้สูงจะเลือกซื้อสินค้าในห้างสรรพสินค้ามากกว่าซื้อสินค้าในร้านขายของชำใกล้บ้าน

(ข) การแสดงภาพของข้อมูล เป็นการนำเสนอข้อมูลที่อยู่ในรูปแบบกราฟิก (Graphic) หรือนำเสนอในลักษณะภาพ 2 มิติ โดยแสดงข้อมูลให้เป็นรูปธรรมเพื่อช่วยให้เกิดความเข้าใจได้ง่ายขึ้น เช่น บริษัทมีความต้องการที่จะขยายสาขาใหม่ในเขตพื้นที่ภาคตะวันออกของประเทศไทย เพื่อพิจารณาสถานที่ตั้งที่เหมาะสมที่สุดในการแข่งขันกับบริษัทคู่แข่ง บริษัท

จึงใช้แผนที่แสดงที่ตั้งของคู่แข่งขันทางธุรกิจที่มีสาขาอยู่ในเขตพื้นที่ดังกล่าวและเขตพื้นที่ใกล้เคียง เป็นต้น

2.2 กฎการจำแนก

กฎการจำแนก (Classification Rule) เป็นเทคนิคหนึ่งในการจำแนกประเภทข้อมูล ที่ใช้ในการค้นหาความรู้บนฐานข้อมูลขนาดใหญ่ โดยมีกระบวนการสร้างแบบจำลองเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดให้มีความแตกต่างระหว่างข้อมูล โดยกฎที่ได้จะนำไปใช้ในการพยากรณ์ต่อไป กฎการจำแนกจึงเป็นกระบวนการจัดแบ่งข้อมูลตามลักษณะของวัตถุประสงค์ โดยมีกระบวนการวิเคราะห์เซตของกลุ่มข้อมูล (Data Object) ที่ยังไม่ได้มีการจัดแบ่งประเภท เพื่อสร้างแบบจำลองจัดการข้อมูลให้อยู่ในรูปของชุดข้อมูล (Class) หรือประเภทที่กำหนด โดยจะนำข้อมูลส่วนหนึ่งจากข้อมูลทั้งหมดมาเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ (Training Data) เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ ผลลัพธ์ที่ได้จากระบบการเรียนรู้คือแบบจำลองที่สามารถนำมาจัดประเภทข้อมูล (Classifier Model) และจะนำข้อมูลอีกส่วนหนึ่งมาใช้เป็นข้อมูลสำหรับทดสอบ (Testing Data) ความถูกต้องของแบบจำลองที่ได้ ผลลัพธ์ที่ได้จะนำไปใช้เปรียบเทียบกับกลุ่มที่หามาได้จากการผ่านกระบวนการเรียนรู้จากแบบจำลองเพื่อทดสอบความถูกต้อง (ปรีชา ยามันสะบีดี และคณะ, 2548) โดยจะมีการปรับปรุงแบบจำลองจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ ในการพยากรณ์จะนำข้อมูลใหม่เข้ามาผ่านกฎการจำแนกที่ได้จากแบบจำลอง เพื่อให้แบบจำลองสามารถพยากรณ์กลุ่มของข้อมูลใหม่ได้อย่างถูกต้อง

กฎการจำแนกในการทำเหมืองข้อมูลเป็นการวิเคราะห์เซตของกลุ่มข้อมูล (Data Object) ที่ยังไม่จัดแบ่งประเภท เพื่อสร้างแบบจำลองออกเป็นชุดข้อมูล (Class) ซึ่งลักษณะของคลาสถูกอธิบายโดยกลุ่มของคุณสมบัติ (Attribute) และกลุ่มของข้อมูล (Training Data Set) ที่ใช้ในการสร้างกฎการจำแนกจากแบบจำลอง ซึ่งมีรูปแบบดังนี้

IF <Conditions> THEN <Class> หรือ “ถ้า <เงื่อนไข> แล้ว <คลาส>”

ในขั้นตอนการจำแนกข้อมูลตามเทคนิคกฎการจำแนก สามารถแบ่งได้เป็น 2 ขั้นตอน

ขั้นตอนที่ 1 การสร้างแบบจำลองตัวแบบ (Classifier Model) เป็นการนำข้อมูลส่วนหนึ่งมาเข้าสู่ระบบเรียนรู้ผ่านกระบวนการของอัลกอริทึมการจำแนก (Classification Algorithm) ซึ่งผลลัพธ์ที่ได้จะอยู่ในรูปของแบบจำลองกฎการจำแนก

ขั้นตอนที่ 2 การใช้แบบจำลองกฎการจำแนกเพื่อการพยากรณ์ (Prediction) สิ่งที่ต้องการรู้ในอนาคตโดยมีจุดมุ่งหมายในการแก้ไขปัญหา จากแบบจำลองจะนำข้อมูลใหม่เข้ามาเพื่อให้จำแนกข้อมูล โดยการนำผลลัพธ์ที่ได้ทำการเปรียบเทียบกับแบบจำลองการจำแนกและวิเคราะห์เพื่อตัดสินใจความเป็นไปได้ของข้อมูลนั้น

2.3 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคในการสร้างแบบจำลองวิธีหนึ่งที่นิยมใช้ในการพยากรณ์ (Prediction) หรือการจำแนกข้อมูล (Classification) ที่มีลักษณะการทำงานเหมือนโครงสร้างต้นไม้ และมีการสร้างกฎต่าง ๆ เพื่อนำไปช่วยใช้ในการประกอบการตัดสินใจ ซึ่งต้นไม้ตัดสินใจเป็นแบบจำลองที่ง่ายต่อการเข้าใจและง่ายต่อการปรับเปลี่ยนเป็นกฎการจำแนก (Classification Rules) โดยทั่วไปอัลกอริทึมพื้นฐานของการสร้างต้นไม้ตัดสินใจ คือ อัลกอริทึมละโมบ (Greedy Algorithm) โดยจะสร้างต้นไม้จากบนลงล่างแบบวนซ้ำ (Recursive) ด้วยวิธีการแบ่งปัญหาใหญ่เป็นปัญหาย่อย (Divide-and-Conquer) รูปแบบของต้นไม้ตัดสินใจจะประกอบด้วยโหนดแรกสุดที่เรียกว่าโหนดราก (Root node) จากโหนดรากจะแตกออกเป็นโหนดลูกและที่โหนดลูกก็จะมีลูกของตัวเองซึ่งโหนดในระดับสุดท้ายจะเรียกว่าใบหรือลีฟโหนด (Leaf Node) แต่ละโหนดแสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล แต่ละกิ่งแสดงผลลัพธ์ในการทดสอบตามเงื่อนไขและลีฟโหนดแสดงกลุ่มข้อมูลหรือคลาส (Class) ที่กำหนดไว้

การจำแนกกลุ่มของข้อมูลได้มาจากการทดสอบค่าคุณลักษณะต่าง ๆ ของข้อมูลเหล่านั้นตามหลักการพื้นฐานของโครงสร้างต้นไม้ตัดสินใจที่ประกอบไปด้วยโหนดรากของต้นไม้แล้วแตกกิ่งไปจนถึงใบในลักษณะจากบนลงล่าง (Top-Down) (บุญมา เพ่งช่วน, 2548) โดยแต่ละโหนดก็คือเกณฑ์ในการตัดสินใจ การสร้างต้นไม้ตัดสินใจมีขั้นตอนดังนี้

2.3.1 สร้างโหนดรากซึ่งเป็นโหนดแรกสุดและมีเพียงโหนดเดียวแสดงถึงข้อมูลชุดสอนการเรียนรู้ (Training Set)

2.3.2 ตรวจสอบว่าข้อมูลทั้งหมดนั้นอยู่ในกลุ่มเดียวกันหรือไม่ หากอยู่ในกลุ่มเดียวกันให้กำหนดโหนดนั้นเป็นใบแล้วตั้งชื่อใบตามกลุ่มข้อมูลนั้น

2.3.3 หากตรวจสอบแล้วพบว่าข้อมูลทั้งหมดนั้นประกอบด้วยข้อมูลหลายกลุ่มปะปนกัน ให้ทำการวัดค่าเกน (Gain) จากค่าคุณลักษณะสำหรับใช้เป็นเกณฑ์ในการคัดเลือกและจัดกลุ่มข้อมูลตามค่าคุณลักษณะ

ค่าที่ใช้ในการตัดสินใจว่า จะใช้ตัวแปรใดสำหรับการแบ่งข้อมูลโดยวิธีการกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นการเลือกข้อมูลตามลำดับของตัววัดคือค่าเกนที่สูงที่สุดจะถือว่า

เป็นข้อมูลเริ่มต้นและข้อมูลถัดไปที่มีค่าลดหลั่นกันตามลำดับ เช่น กำหนดให้มีการพิจารณาจากกลุ่มข้อมูล 2 คลาส คือ P และ Nโดยจำนวนตัวอย่างในคลาส P คือ p ตัว และจำนวนตัวอย่างในคลาส N คือ n ตัว

ค่าของกลุ่มข้อมูล คือ ค่าคาดคะเนที่กลุ่มตัวอย่างต้องใช้จำนวนบิตในการแยกคลาส P และ N โดยนิยามตามสมการที่ 2-1

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \tag{2-1}$$

ค่าคาดคะเนข้อมูล (Entropy) การใช้ลักษณะประจำ A ซึ่งกำหนด A คือ ลักษณะประจำที่แบ่ง S ออกเป็น {S1, S2, ,Sv} โดยให้ S1 มีตัวอย่างจากคลาส P จำนวน p1 และตัวอย่างจากคลาส N จำนวน n1 ดังสมการที่ 2-2

$$E(A) = \sum_{i=1}^{v} \frac{p_1 + n_1}{p+n} I(p_1, n_1) \tag{2-2}$$

ดังนั้นค่าเกนของข้อมูล (Data Gain) ที่ได้จากการแยกข้อมูลด้วยลักษณะประจำ A เป็นดังสมการที่ 2-3

$$Gain(A) = I(p,n) - E(A) \tag{2-3}$$

ค่าที่เป็นรูปแบบหนึ่งของการใช้ต้นไม้ตัดสินใจที่พัฒนาเพิ่มเติมโดยไม่ได้ใช้ค่าเกนเป็นตัวแบ่ง แต่ใช้อัตราค่าเกน (Gain ratio) เป็นตัวแบ่งดังสมการที่ 2-4

$$Gain\ ratio\ =\ \frac{Gain}{SplitInfo} \tag{2-4}$$

2.4 อินเทอร์เน็ต

อินเทอร์เน็ต (Internet) มาจากคำเต็มว่า International Network คือ การเชื่อมต่อเครือข่ายคอมพิวเตอร์ขนาดใหญ่ โดยการเชื่อมอินเทอร์เน็ตต้องใช้บริการของไอเอสพี (ISP) ย่อมาจาก Internet Service Provider ซึ่งเป็นองค์กรที่ทำหน้าที่เป็นผู้ที่คอยให้บริการเชื่อมต่ออินเทอร์เน็ตและ

โปรโตคอล ทีซีพีสแลชไอพี (TCP/IP) ย่อมาจาก Transmission Control Protocol/ Internet Protocol เป็นมาตรฐานที่ใช้ในการสื่อสารและรับส่งข้อมูลระหว่างกันได้ทั่วโลก อินเทอร์เน็ตได้ถูกนำมาใช้ทั้งภาครัฐบาล ภาคธุรกิจและภาคอุตสาหกรรมในงานด้านต่าง ๆ

อินเทอร์เน็ตเริ่มต้นจากเครือข่ายที่ใช้ในกิจการทางทหารของสหรัฐอเมริกา ชื่อ อาร์พาเน็ต (ARPANET : Advanced Research Projects Agency Network) ซึ่งเริ่มใช้ในกิจการเมื่อประมาณ พ.ศ. 2512 ภายหลังมีมหาวิทยาลัยหลายแห่งขอร่วมเครือข่าย โดยเชื่อมต่อระบบคอมพิวเตอร์ของมหาวิทยาลัยกับเครือข่ายดังกล่าว เพื่อใช้ประโยชน์ในการศึกษาและการวิจัยต่อมาจึงมีการนำมาใช้ในเชิงพาณิชย์อย่างกว้างขวาง

อินเทอร์เน็ตในประเทศไทยเริ่มขึ้นตั้งแต่ พ.ศ. 2530 โดยมหาวิทยาลัยสงขลานครินทร์ (วิทยาเขตหาดใหญ่) และสถาบันเทคโนโลยีแห่งเอเชีย นับเป็นที่อยู่ของอินเทอร์เน็ตแห่งแรกของประเทศไทย โดยได้รับที่อยู่ (Address) ชื่อ sritrang.psu.th ต่อมาใน พ.ศ. 2534 มีการนำอินเทอร์เน็ตเข้ามาอยู่ในประเทศไทย โดยจุฬาลงกรณ์มหาวิทยาลัยได้ดำเนินการเช่าสายซึ่งเป็นสายอินเทอร์เน็ตความเร็วสูงต่อเชื่อมกับเครือข่าย UUNET ของบริษัทเอกชนที่รัฐเวอร์จิเนีย ประเทศสหรัฐอเมริกา ต่อมามหาวิทยาลัยมหิดล มหาวิทยาลัยเชียงใหม่สถาบันเทคโนโลยีพระจอมเกล้าและมหาวิทยาลัยอัสสัมชัญได้ขอเชื่อมต่อผ่านจุฬาลงกรณ์มหาวิทยาลัยและเรียกเครือข่ายนี้ว่า ไทยเน็ต (THAInet) นับเป็นเกตเวย์ (Gateway) แรกสู่เครือข่ายอินเทอร์เน็ตสากลของประเทศไทย ในปี พ.ศ. 2535 ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC : National Electronic and Computer Technology Centre) ได้จัดตั้งกลุ่มเครือข่ายประกอบด้วย มหาวิทยาลัยอีกหลายแห่ง เรียกว่า เครือข่ายไทยสาร โดยเชื่อมต่อกับเครือข่าย UUNET ด้วยนับเป็นเกตเวย์สู่เครือข่ายอินเทอร์เน็ตแห่งที่สองและได้พัฒนาต่อจนมีการเชื่อมต่อกันอย่างแพร่หลายกลายเป็นเครือข่ายขนาดใหญ่และใช้กันอย่างแพร่หลายในปัจจุบัน

2.5 ภาษาเอชทีเอ็มแอล

ภาษาเอชทีเอ็มแอล (HTML) ย่อมาจาก Hypertext Markup Language เป็นภาษาเครื่องหมายซึ่งเป็นภาษาหลักที่ใช้ในการสร้างหรือพัฒนาเว็บเพจเพื่อแสดงผลข้อมูลบนเว็บบราวเซอร์ (Web Browser) ในเครือข่ายอินเทอร์เน็ต โดยเมื่อทำการจัดเก็บเอกสารเป็นแบบสกุล .html หรือ .htm หากต้องการเพิ่มประสิทธิภาพของเว็บเพจก็สามารถนำเอาภาษาสคริปต์ต่าง ๆ เช่น จาวา (JAVA) เอเอสพี (ASP) วีบี (VB) และพีเอชพี (PHP) เป็นต้น เข้ามาใช้ในเอกสารภาษาเอชทีเอ็มแอลเพื่อแสดงข้อมูลในรูปเอกสารภาษาเอชทีเอ็มแอล การสร้างเอกสารภาษาเอชทีเอ็มแอลที่สมบูรณ์ประกอบด้วย 2 ส่วนคือ ส่วนที่เป็นเนื้อหาและส่วนที่เป็นคำสั่งที่อยู่ในแท็ก (Tag) ซึ่ง

เมื่อมีการเปิดแท็กคำสั่ง <คำสั่ง> แล้วต้องทำการปิดแท็กคำสั่ง </คำสั่ง> การปิดคำสั่งที่ต่อเนื่องกันให้ปิดโดยเริ่มจากคำสั่งที่อยู่ด้านในก่อนแล้วค่อยขยายออกไปด้านนอกเพื่อผลลัพธ์ที่ถูกต้อง แสดงรูปแบบเอกสารเอชทีเอ็มแอลได้ดังนี้

```
<Html>
<Head>
<Title>หัวข้อเรื่อง</Title>
</Head>
<Body>ข้อมูลที่ต้องการแสดง </Body>
<Html>
```

สามารถอธิบายความหมายของแต่ละแท็กดังนี้

<Html> </Html> เป็นคำสั่งเริ่มต้นและสิ้นสุดของเอกสาร

<Head> </Head> เป็นคำสั่งที่กำหนดข้อความบรรยายคุณลักษณะของเอกสาร

<Title> </Title> เป็นคำสั่งที่แสดงชื่อเอกสารซึ่งจะแสดงผลที่ส่วนไตเติลบาร์ของวินโดว์

<Body> </Body> เป็นส่วนเนื้อหาที่จะแสดงผลบนบราวเซอร์

2.6 การวิเคราะห์และออกแบบด้วยยูเอ็มแอล

การวิเคราะห์และออกแบบด้วยยูเอ็มแอล เป็นการวิเคราะห์ความต้องการของระบบที่จะพัฒนาขึ้นมาใหม่ว่าควรมีสิ่งใดบ้างภายในระบบและมีผู้ใช้ที่เกี่ยวข้องกับระบบในส่วนใดด้วยแผนภาพ เมื่อได้ความต้องการของระบบครบถ้วนแล้วก็จะเข้าสู่การออกแบบระบบเพื่อนำไปสู่การพัฒนาฐานข้อมูลและการพัฒนาโปรแกรมของระบบ คำว่า ยูเอ็มแอล (UML) ย่อมาจาก Unified Modeling Language เป็นภาษาสัญลักษณ์รูปภาพมาตรฐานสำหรับใช้ในการสร้างแบบจำลองเชิงวัตถุ (กิตติ ภักดีวัฒนะกุล และกิตติพงษ์ กลมกล่อม, 2547) โดยยูเอ็มแอลเป็นภาษามาตรฐานสำหรับสร้างแบบพิมพ์เขียว (Blueprint) ให้แก่ระบบงาน ยูเอ็มแอลใช้แนวความคิดเชิงวัตถุเป็นพื้นฐานในการสร้างแบบจำลองในการวิเคราะห์และออกแบบ (Analysis and Design) เพื่อให้ระบบที่ซับซ้อนและเข้าใจได้ยากระหว่างผู้ใช้ระบบและนักพัฒนาระบบสามารถเห็นภาพระบบได้ชัดเจน โดยการสร้างแบบจำลองจะเปรียบเสมือนพิมพ์เขียวที่แสดงถึงภาพรวมของระบบทั้งหมด ยูเอ็มแอลเป็นภาษามาตรฐานที่นำมาใช้ในการพัฒนาระบบเชิงวัตถุเพื่อสนับสนุนในการเขียนโปรแกรมภาษาคอมพิวเตอร์เชิงวัตถุ แบบจำลองที่สร้างขึ้นจะต้องมีความสอดคล้องกับความต้องการของ

ระบบเป็นสำคัญ โดยในแต่ละแบบจำลองจะมีการเพิ่มในส่วนของรายละเอียดต่าง ๆ ลงไปในแบบจำลอง ในที่สุดแบบจำลองจะถูกนำไปพัฒนาขึ้นเป็นระบบจริง

2.7 ฐานข้อมูลมายเอสคิวแอล

ฐานข้อมูลมายเอสคิวแอล (My Structured Query Language : MySQL) คือ โปรแกรมฐานข้อมูลที่มีการบริหารจัดการฐานข้อมูลที่เป็นระบบจัดการฐานข้อมูลเชิงสัมพันธ์ที่ใช้จัดการตารางข้อมูลในส่วนของการจัดการคอลัมภ์ การกำหนดคุณลักษณะของแต่ละคอลัมภ์ การบันทึกข้อมูลทั้งที่เป็นตัวเลข ตัวอักษร และค่าอื่น ๆ เป็นต้น ที่ประกอบกันเป็นแถวเพื่อบันทึกลงในฐานข้อมูล โดยแต่ละแถวจะไม่มีข้อมูลที่ซ้ำกันเนื่องจากจะมีคอลัมภ์หนึ่งที่เป็นกุญแจหลัก แต่ละตารางข้อมูลจะมีความสัมพันธ์เชื่อมโยงกัน (วรรณวิภา คิตถศิริ, 2545) ในการจัดการกับข้อมูลต้องอาศัยภาษาคอมพิวเตอร์ที่เรียกว่าเอสคิวแอล (Structured Query Language : SQL) เป็นเครื่องมือสำหรับจัดการข้อมูลในฐานข้อมูลที่ต้องใช้ร่วมกับเครื่องมืออื่นอย่างสอดคล้องเพื่อให้ได้ระบบที่รองรับความต้องการของผู้ใช้ เช่น เครื่องบริการเว็บ (Web Server) และโปรแกรมประมวลผลฝั่งเครื่องบริการ (Server-Side Script)

ฐานข้อมูลมายเอสคิวแอลเป็นฟรีแวร์ (Freeware) ทางด้านฐานข้อมูลจึงได้รับความนิยมอย่างมากในปัจจุบัน สามารถดาวน์โหลดซอร์สโค้ด (Source Code) ได้จากอินเทอร์เน็ตโดยไม่ต้องเสียค่าใช้จ่ายและสามารถแก้ไขได้ตามความต้องการ พร้อมทั้งยังสนับสนุนการใช้งานบนระบบปฏิบัติส่วนใหญ่ที่ใช้กันในปัจจุบันได้ เช่น ระบบปฏบัติการลีนุกส์ (Linux) ระบบปฏิบัติการไอโอเอส (iOS) และระบบปฏิบัติการวินโดว์ส (Windows) (สงกรานต์ ทองสว่าง, 2548) นอกจากนี้ฐานข้อมูลมายเอสคิวแอลสามารถทำงานร่วมกับภาษาคอมพิวเตอร์ได้หลายภาษา เช่น จาวา เอเอสพี และพีเอชพี เป็นต้น

ด้วยลักษณะเด่นของฐานข้อมูลมายเอสคิวแอลที่เร็ว ใช้งานง่าย มีความน่าเชื่อถือได้สูงและไม่ต้องเสียค่าใช้จ่ายเนื่องจากเป็นฐานข้อมูลที่อนุญาตให้สามารถใช้ได้ คุณสมบัติของฐานข้อมูลมายเอสคิวแอลมีดังนี้

2.7.1 การทำงานแบบแบ่งการทำงานเป็นส่วนย่อยแยกออกไป (Multi Threaded) คือ การมีการแบ่งงานในลักษณะต่างส่วนต่างทำ ทำให้สามารถทำงานได้เร็วและมีการทำงานที่เป็นอิสระในส่วนแต่ละส่วนย่อย

2.7.2 สามารถใช้กับภาษาคอมพิวเตอร์หรือภาษาสคริปต์ได้หลากหลายภาษา

2.7.3 สามารถทำงานกับฐานข้อมูลขนาดเล็กและขนาดใหญ่ที่มีจำนวนข้อมูลในตารางข้อมูลถึง 60,000 ตาราง มีจำนวนรายการข้อมูลถึง 5,000,000,000 รายการได้

2.7.4 สามารถรองรับชนิดข้อมูลที่หลากหลาย เช่น ตัวเลข ตัวอักษร บูลีน วันที่และเวลา เป็นต้น

2.7.5 สามารถรองรับภาษาเอสคิวแอลที่เป็นภาษามาตรฐานในการจัดการฐานข้อมูล

2.7.6 สามารถรองรับรูปแบบการเชื่อมต่อฐานข้อมูลได้หลากหลายวิธี

2.7.7 ใช้ได้กับระบบปฏิบัติการคอมพิวเตอร์หลากหลายระบบ

2.8 ภาษาพีเอชพี

ภาษาพีเอชพี (PHP) ย่อมาจาก Personal Home Page ซึ่งเป็นภาษาสคริปต์ที่ทำงานฝั่งเซิร์ฟเวอร์ที่เรียกว่าเซิร์ฟเวอร์ไซด์สคริปต์ (Server Side Script) โดยการทำงานของภาษาพีเอชพี จะประมวลผลฝั่งเซิร์ฟเวอร์แล้วส่งผลลัพธ์ไปยังฝั่งไคลเอนท์ผ่านเว็บเบราเซอร์ (นิรุช อำนวยศิลป์, 2545) ผู้ใช้จะส่งคำร้องขอเพื่อร้องขอข้อมูลที่ต้องการจากเครื่องคอมพิวเตอร์แม่ข่ายที่เป็นเครื่องคอมพิวเตอร์ให้บริการเว็บ (Web Server) เครื่องคอมพิวเตอร์ให้บริการเว็บจะประมวลภาษาภาษาพีเอชพี (อดิศักดิ์ จันทร์มิน , 2548) ซึ่งจะแสดงผลภายใต้ภาษาเอชทีเอ็มแอลเป็นภาษาที่ใช้เพื่อการแสดงผลบนเว็บ และแสดงผลไปยังหน้าจอโปรแกรมค้นหาข้อมูลของเครื่องผู้ใช้ทำให้การทำงานมีความปลอดภัยสูง

ภาษาพีเอชพี สามารถนำมาเขียนเป็นโปรแกรมที่ซับซ้อนบนอินเทอร์เน็ตเป็นการเขียนโปรแกรมเพิ่มเข้าไปในภาษาเอชทีเอ็มแอล (HTML) ซึ่งเป็นภาษาที่ใช้แสดงผลทางอินเทอร์เน็ต ในการเขียนภาษาพีเอชพีนั้นไม่จำเป็นต้องประกาศตัวแปรก่อนการใช้งาน นอกจากนั้นภาษาพีเอชพียังรองรับการเขียนโปรแกรมแบบเชิงวัตถุ (Object-Oriented) ซึ่งเป็นหลักการเขียนโปรแกรมคอมพิวเตอร์ที่สามารถนำโค้ดที่เขียนไว้มาใช้ใหม่ (พร้อมเลิศ หล่อวิจิตร, 2550) โดยนักพัฒนาระบบคนอื่นสามารถนำโค้ดนั้นไปใช้ได้โดยไม่ต้องเขียนโค้ดขึ้นมาใหม่ ทำให้ประหยัดเวลาและรวดเร็วในการพัฒนาระบบสารสนเทศ ในปัจจุบันภาษาพีเอชพีได้กลายมาเป็นภาษาคอมพิวเตอร์ที่ได้รับความนิยมและมีความสามารถนำมาใช้สำหรับงานด้านการพัฒนาระบบงานบนเครือข่ายอินเทอร์เน็ต (Web Application) ซึ่งมีความสามารถด้านการจัดการฐานข้อมูลที่มีความเร็วสูง มีประสิทธิภาพและพัฒนาได้ง่ายที่สุดภาษาหนึ่ง (กิตติ ภักดีวัฒนะกุล, 2548) นอกจากนี้ภาษาพีเอชพียังเป็นซอฟต์แวร์โอเพนซอส ซึ่งไม่ต้องเสียค่าใช้จ่ายในการใช้ซอฟต์แวร์

2.9 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องที่ใช้เทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้เพื่อการพยากรณ์ในด้านการศึกษาและด้านต่าง ๆ มีดังต่อไปนี้

การทำเหมืองข้อมูลได้นำมาประยุกต์ใช้ในการสืบค้นสิ่งที่น่าสนใจจากข้อมูลนิสิต โดยใช้เทคนิคสำคัญ 3 ประการ ได้แก่ การค้นหากฎจำแนกเชิงความสัมพันธ์ (Association Rule Discovery) การจำแนกข้อมูล (Data Classification) การพยากรณ์ข้อมูล (Data Prediction) มาประยุกต์ในการช่วยนิสิตเลือกสาขาที่เหมาะสมและพยากรณ์เกรดแต่ละรายวิชาในภาคการศึกษาต่อไป การใช้ต้นไม้ตัดสินใจเพื่อสร้างแบบจำลองการจำแนกประเภทข้อมูล งานวิจัยสามารถพยากรณ์สาขาวิชาที่เหมาะสมที่สุดให้กับนักศึกษาได้ค่อนข้างเป็นที่น่าพอใจ โดยมีเปอร์เซ็นต์ความถูกต้องค่อนข้างสูง แต่มีปัญหาแบบจำลองจะพยากรณ์แนวโน้มโอนเอียงไปทางสาขาวิชาที่มีจำนวนนิสิตมากเป็นผลทำให้ความถูกต้องของแบบจำลองที่ได้ค่อนข้างต่ำ เมื่อจำนวนข้อมูลในบางสาขาวิชามีปริมาณน้อยทำให้แบบจำลองที่ได้ไม่แม่นยำเท่าที่ควร ในการพยากรณ์ถึงประสิทธิภาพการเรียนของนักศึกษา (กฤษณะ ไวยมัยและคณะ, 2554) การวิจัยจากการประยุกต์ใช้กับประวัตินักศึกษาและข้อมูลรายวิชาที่ได้รับจากมหาวิทยาลัยในไต้หวัน เพื่อที่จะพยากรณ์แนวโน้มที่จะผ่านหรือไม่ผ่านรายวิชาหนึ่ง ๆ เมื่อผ่านการสอบกลางภาคไปแล้วหนึ่งสัปดาห์ (Han, J. and Kamber, M., 2001) โดยการสร้างต้นไม้ที่ช่วยในการตัดสินใจที่มีความเหมาะสมและแม่นยำจากการดำเนินการร่วมมือกันของวิธีการค้นหากฎจำแนกเชิงความสัมพันธ์ (Association Rule) และวิธีทางพันธุกรรม (Genetic Algorithm) ซึ่งวิธีการค้นหากฎความสัมพันธ์จะใช้เทคนิคอัลกอริทึมอพริโอริ (Apriori Algorithm) และใช้วิธีทางพันธุกรรมในการหาต้นไม้ที่เหมาะสมที่สุดที่จะไปใช้ในการพยากรณ์รวมเรียกว่า เอจีเอ (Association base-GA :AGA) และได้ทดลองกระบวนการทั้งหมดโดยเปรียบเทียบกับเอสจีเอ (Simple Genetic Algorithm : SGA) และนำไปประยุกต์ใช้กับข้อมูลของนักศึกษาเพื่อหาประสิทธิภาพการเรียนของนักศึกษา จากการเปรียบเทียบกับเอสจีเอ พบว่าเอจีเอนั้นมีความแม่นยำในการพยากรณ์สูงกว่าและใช้เวลาในการคำนวณน้อยกว่าและชุดข้อมูลที่จะพยากรณ์ประสิทธิภาพนักศึกษา ซึ่งในการพยากรณ์ประสิทธิภาพของนักศึกษาสามารถช่วยเหลือในขั้นต้นที่จะให้นักศึกษาได้เข้าใจถึงอนาคตที่ได้ประมาณไว้ก่อนที่จะสายเกินไป เอจีเอได้พิสูจน์ให้เห็นได้ว่ามีความแม่นยำในการพยากรณ์ถึง 80%

การนำเทคนิคแบบจำลองต้นไม้ตัดสินใจมาช่วยในการจำแนกกลุ่มสถานภาพการสำเร็จการศึกษาของนักศึกษาในแต่ละปี ได้ทำการเรียนรู้จากข้อมูลส่วนตัวของนักศึกษา ข้อมูลการลงทะเบียนตั้งแต่ภาคเรียนที่ 1 และสถานภาพของการศึกษาในปีสุดท้ายตามหลักสูตรที่มหาวิทยาลัยกำหนด เพื่อพยากรณ์โอกาสการสำเร็จการศึกษาและทราบถึงข้อมูลและความสัมพันธ์ของข้อมูลที่

มีผลต่อการสำเร็จการศึกษาของนักศึกษา เพื่อใช้ในการดูแลนักศึกษาเพื่อให้เรียนจบภายในหลักสูตรและลดจำนวนนักศึกษาที่ไม่สามารถสำเร็จการศึกษา (ชลนิศา สาระ, 2550) ผลของงานวิจัยทำให้สามารถทราบปัจจัยที่มีผลต่อการสำเร็จการศึกษาและสามารถคาดเดาระยะเวลาในการสำเร็จการศึกษาได้อย่างแม่นยำ แต่เนื่องด้วยการหาปัจจัยที่มีผลต่อการสำเร็จการศึกษาอาจต้องมีการเก็บข้อมูลที่ละเอียดกว่าเดิมเพื่อให้สามารถทราบปัจจัยที่มีผลต่อการสำเร็จการศึกษาได้อย่างครบถ้วน การศึกษาเพื่อการเปรียบเทียบตัวแปรที่เกี่ยวข้องกับการสำเร็จการศึกษาและไม่สำเร็จการศึกษาของนักศึกษามหาวิทยาลัยรามคำแหง (สุมาลี ชาญกาญจน์, 2551) โดยได้ทำการศึกษาจากการศึกษาของนักศึกษาที่ผ่านมานั้น ผลสำเร็จในการศึกษาเป็นอย่างไร พบว่ามีหลากหลายปัจจัยอันอาจทำให้ยังไม่ทราบอย่างแน่ชัดว่ามีปัจจัยใดที่มีผลต่อการสำเร็จมากที่สุด การศึกษาวิจัยระบบการวิเคราะห์แนวโน้มการสำเร็จการศึกษาของนักศึกษาวิทยาลัยเทคนิคในรัฐเท็กซัส การศึกษาอัตราแนวโน้มการสำเร็จการศึกษาที่ลดลงเรื่อย ๆ โดยทำการศึกษาข้อมูลที่เกี่ยวข้องกับนักศึกษา (John, G. Hendricks, 2000) ซึ่งข้อมูลที่ศึกษาดังกล่าวเป็นข้อมูลที่เก็บเป็นฐานข้อมูลขนาดใหญ่ที่ได้จากวิทยาลัยเทคนิคตัวอย่าง 3 แห่ง จากนั้นนำข้อมูลมาวิเคราะห์โดยใช้โปรแกรมโนว์เลดจ์ซีเคอร์ไฟว์ทีเอ็ม (Knowledge SEEKER IV TM) ซึ่งเป็นโปรแกรมเหมืองข้อมูลที่มีการศึกษาตัวแปรอิสระและตัวแปรอื่น ๆ จากนั้นได้สร้างแบบจำลองในการพยากรณ์และทดสอบ เพื่อค้นหาและแสดงปัจจัยสำคัญที่มีผลต่อการเพิ่มอัตราการสำเร็จการศึกษาของนักศึกษาในสถาบันดังกล่าว การเปรียบเทียบหาเทคนิคการคัดเลือกคุณลักษณะที่จะให้ประสิทธิภาพการจำแนกข้อมูลที่ดีที่สุดเพื่อนำมาทำการจำแนกและหาปัจจัยที่มีผลต่อพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา โดยเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะ 3 วิธี ได้แก่ การเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ (Correlation-based Feature Selection) กับการประเมินกลุ่มย่อยพื้นฐานที่เหมาะสม (Consistency-based Subset Evaluation) และการประเมินกลุ่มย่อยแบบเวร์บเบอร์ (Wrapper Subset Evaluation) เพื่อมาทำการคัดเลือกคุณลักษณะที่เหมาะสมร่วมกับตัวจำแนกประเภทเบย์อย่างง่าย (Naive Bayes classifier) เปรียบเทียบกับข่ายงานความเชื่อเบย์เซียน (Baysian Belief Network) และสำหรับการวัดประสิทธิภาพได้ทำการทดสอบโดยใช้เทคนิคการตรวจสอบไขว้ (k-fold Cross Validation) จากการทดสอบพบว่าเทคนิคการคัดเลือกคุณลักษณะโดยใช้การประเมินกลุ่มย่อยแบบเวร์บเบอร์ร่วมกับแนวคิดข่ายงานความเชื่อเบย์ให้ค่าความถูกต้องสูงที่สุดถึง 82.42% และในทางเดียวกันสามารถลดจำนวนคุณลักษณะของข้อมูลลงได้ถึง 71.87%

ข้อมูลภาวะการมีงานทำของบัณฑิตได้นำมาวิเคราะห์โดยใช้เทคนิคเหมืองข้อมูลต้นไม้ตัดสินใจ เพื่อศึกษาและพัฒนาตัวแบบเพื่อใช้ในการพยากรณ์แนวโน้มเลือกอาชีพแรกหลังสำเร็จการศึกษาของนักศึกษาระดับปริญญาตรีและนำข้อมูลรายบุคคลนักศึกษาของสถาบันอุดมศึกษาของ

รัฐ ปีการศึกษา 2548 (บุญมา เพ่งช่วน, 2548) มาวิเคราะห์กับตัวแบบที่สร้างได้และนำเสนอในรูปแบบของรายงาน แบบตารางและกราฟ เพื่อนำไปสนับสนุนการตัดสินใจในด้านการผลิตบัณฑิต ผลการวิจัยพบว่า สามารถนำตัวแบบที่พัฒนาขึ้นมาหาแนวโน้มการเลือกอาชีพแรกหลังสำเร็จการศึกษาของนักศึกษาระดับปริญญาตรีได้ และนำข้อมูลรายบุคคลนักศึกษามาประมวลผลกับตัวแบบเพื่อนำไปสนับสนุนการตัดสินใจด้านการผลิตบัณฑิตระดับปริญญาตรีต่อไปได้

การเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรีได้มีการใช้เทคนิคต้นไม้ตัดสินใจเพื่อทำการวิจัยสร้างตัวแบบสำหรับพัฒนาระบบสนับสนุนการตัดสินใจเลือกสาขาการเรียนของนักศึกษาระดับปริญญาตรี (ไพฑูรย์ จันทร์เรือง, 2550) งานวิจัยนี้ได้แยกสร้างตัวแบบสำหรับแต่ละสาขาการเรียนเนื่องจากคุณสมบัติของผู้เรียนแต่ละสาขามีความแตกต่างกัน ทำให้ได้ตัวแบบที่สามารถพยากรณ์แนวโน้มของผลการเรียนที่เหมาะสมสำหรับแต่ละสาขา แต่เนื่องจากผลการเรียนของนักศึกษาที่นำมาพัฒนาตัวแบบ ส่วนใหญ่จะมีเกณฑ์คะแนนเกาะกลุ่มกันอยู่ในช่วงกลางของข้อมูล ทำให้ผลการตัดสินส่วนใหญ่จะโน้มเอียงไปในเกณฑ์พอใช้และปานกลาง

การวางแผนรับนักศึกษาของสถานศึกษา ได้จัดทำเว็บการทำเหมืองข้อมูลสำหรับซึ่งจะใช้การวิเคราะห์ระบบโดยใช้แนวทางวัตถุ (Object-Oriented Approach) ซึ่งจะใช้เทคนิคกรณีเหตุผลความสัมพันธ์ (Case Realization) ช่วยในการออกแบบ (แสงจันทร์ เรืองอ่อน และคณะ, 2545) จากนั้นนำมาสร้างคอมโพเนนท์ (Component) โดยใช้ภาษาเอเอสพีร่วมกับวีบีสคริปต์ (VBScript) เป็นการสร้างระบบแล้วนำมาทดสอบกับฐานข้อมูลเชิงสัมพันธ์ เพื่อจะเป็นการทดสอบระบบที่ได้ออกแบบจากการทดลองใช้เว็บการทำเหมืองข้อมูลบนเอกสารเอ็กเอ็มแอลเมื่อเปรียบเทียบกับวิธีเดิมที่ผู้ใช้ต้องแปลงข้อมูลในฐานข้อมูลให้อยู่ในรูปแบบของเอกสารไมโครซอฟต์เอ็กเซล (Microsoft Excel) แล้วทำการวิเคราะห์ เนื่องจากเป็นการทำงานผ่านเว็บบราวเซอร์จึงทำให้การทำงานของผู้ใช้ง่ายขึ้น สามารถเรียนรู้ได้ในเวลาอันสั้นและสามารถเข้าถึงข้อมูลได้ตลอดเวลา

การทำเหมืองข้อมูลได้นำไปช่วยในการแสดงพฤติกรรมของการเรียนร่วมมือกันแบบไม่มีโครงสร้าง (Talavera, L. and Gaudioso, E., 2004) ได้ทำใช้เทคนิคการจัดกลุ่มที่เป็นไปอย่างอัตโนมัติในการค้นหากลุ่มที่เป็นประโยชน์จากข้อมูลเพื่อให้ได้คำบรรยายลักษณะพฤติกรรมของผู้เรียน โดยมีขั้นตอน คือ การรวบรวมและเตรียมข้อมูล จะใช้ข้อมูลการปฏิสัมพันธ์ เช่น จำนวนของการเข้าสนทนา การส่งข้อความ การออกความเห็นและการสร้างแบบจำลอง โดยใช้อัลกอริทึมของการจัดกลุ่มที่จัดการกับข้อมูลที่ไม่ต่อเนื่องในการสร้างแบบจำลองการแปลความหมาย ได้กลุ่มย่อย 6 กลุ่มซึ่งอธิบายพฤติกรรมของผู้เรียนในด้านต่าง ๆ ซึ่งจะนำไปช่วยผู้สอนในการสร้างกลุ่มของผู้เรียนเพื่อแสดงการร่วมมือกันในกิจกรรมและการตัดสินใจต่างๆ รายละเอียดของยุทธศาสตร์การสอนที่จะใช้ให้เป็นประโยชน์ต่อไป ผลลัพธ์ของการนำเสนองานนี้ถึงแม้จะเป็นเบื้องต้น แต่ก็

เป็นการแสดงให้เห็นถึงประโยชน์ของเทคนิคการทำเหมืองข้อมูล ที่สนับสนุนการประเมินผลกิจกรรมแบบร่วมมือกันในชุมชนเสมือนจริงและยังทำให้พบงานวิจัยเพิ่มอย่างน้อยสองงาน คือ การสร้างกลุ่มจากระดับต่ำเพื่อให้ได้แนวทางกับผู้สอนเกี่ยวกับลักษณะระดับที่สูงกว่า สำหรับการวิเคราะห์ข้างหน้าและ การแบ่งกลุ่มที่สามารถนำไปใช้ได้โดยตรงกับข้อมูลที่มีรายละเอียดมากขึ้น คือ รูปแบบที่สามารถทำงานอย่างอัตโนมัติซึ่งแทนที่การสำรวจแต่ละอันด้วยมือหรือรายงานทั้งหมด

จากงานวิจัยดังกล่าวข้างต้น เป็นการนำเอาเทคนิคการทำเหมืองข้อมูล กฎการจำแนกแบบจำลองและต้นไม้ตัดสินใจ เข้ามาใช้ในการพยากรณ์ในงานด้านต่าง ๆ และจากทฤษฎีที่ได้กล่าวมาแล้วข้างต้นผู้วิจัยจึงมีแนวคิดที่จะนำทฤษฎีและแนวทางจากงานวิจัยดังกล่าวมาประยุกต์ใช้ในการพัฒนาระบบสารสนเทศเพื่อการพยากรณ์ผู้เข้าศึกษาโดยผ่านเครือข่ายอินเทอร์เน็ต