บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหาที่นำไปสู่การค้นคว้าวิจัย

ข้อมูลเป็นส่วนสำคัญในการทำงานของหลากหลายสาขาวิชาชีพ ส่วนมากข้อมูลปฐมภูมิ (Primary data) ซึ่งเป็นข้อมูลที่ถูกรวบรวมจากแหล่งที่มาโดยตรง การวิเคราะห์และจำแนกข้อมูลอย่างเป็นระบบและสอดคล้องกับรูปแบบของข้อมูลแต่ละกลุ่มที่นำไปสู่ข้อมูลทุติยภูมิ (Secondary Data) มีคุณภาพเป็นสิ่งที่จำเป็นอย่างมากที่จะทำให้การประมวลผลต้นไม้ตัดสินใจมีคุณภาพ ในปัจจุบันมีหลากหลายวิธีการซึ่งประสิทธิภาพของการวิเคราะห์และจำแนกข้อมูลแต่ละระบบมีความแตกต่างกันขึ้นอยู่กับวัตถุประสงค์ของผู้ใช้งานเอง บ่อยครั้งที่มีการใช้งานระบบวิเคราะห์ข้อมูลที่เกินความจำเป็นหรือการวิเคราะห์ข้อมูลที่ขาดประสิทธิภาพทำให้การวิเคราะห์ข้อมูลจึงไม่บรรลุตามวัตถุประสงค์และขาดประสิทธิภาพ รวมถึงการใช้งานด้านการวิเคราะห์และจำแนกข้อมูลแต่ละรูปแบบนั้นมีความซับซ้อนของกระบวนการทำงานอีกทั้งยังมีการคำนวณที่ต้องใช้ความเข้าใจในทฤษฎีเฉพาะของรูปแบบของการวิเคราะห์และจำแนกข้อมูลนั้นๆ ซึ่งยากลำบากมากที่ผู้ที่ต้องการใช้งานการวิเคราะห์และจำแนกข้อมูลจะสามารถใช้งานได้อย่างเต็มประสิทธิภาพ การวิเคราะห์และจำแนกข้อมูลมีหลากหลายรูปแบบ เช่น วิธีโครงข่ายประสาทเทียม (Neural network) และ ต้นไม้ตัดสินใจ (Decision Tree) เป็นต้นต้นไม้ตัดสินใจเป็นรูปแบบหนึ่งที่มีการใช้งานอย่างแพร่หลายโดยให้ความถูกต้องและประสิทธิภาพดีพอสมควร แต่ผู้ที่ใช้งานต้นไม้ตัดสินใจจำเป็น ต้องมีความรู้ความเข้าใจในทฤษฎีมากพอสมควรและผู้ที่จำเป็นต้องใช้งานการวิเคราะห์และจำแนกข้อมูลในรูปแบบต้นไม้ตัดสินใจจำนวนมากที่ไม่มีความชำนาญหรือไม่เข้าใจในการทำงานของต้นไม้ตัดสินใจมากพอทำให้เกิดความยากลำบากที่จะเรียนรู้ภายใต้ข้อจำกัดหลายอย่าง เช่น ข้อจำกัดด้านเวลา พื้นฐานของความรู้ เป็นต้น รวมไปถึงการเปรียบเทียบต้นไม้ตัดสินใจหารูปแบบที่มีประสิทธิภาพมากที่สุดทั้งนี้ขึ้นอยู่กับการเตรียมข้อมูลก่อนประมวลผล (Pre-processing) อาชีพที่เกี่ยวเนื่องกับข้อมูลที่ช่วยในการตัดสินใจนั้นมีหลากหลายสาขา หากแต่ไม่สามารถนำรูปแบบต้นไม้ตัดสินใจเพื่อใช้ในการวิเคราะห์และจำแนกข้อมูลได้อย่างมีประสิทธิภาพ หรือต้องการที่จะลดระยะเวลาการทำงานให้สั้นลง อาทิเช่น

- ผู้ที่มีหน้าที่อนุมัติสินเชื่อ โดยใช้ข้อมูลของลูกค้าในการวิเคราะห์ประกอบการตัดสินใจเพื่ออนุมัติสินเชื่อ
- ผู้ที่มีหน้าที่อนุมัติงบประมาณ โดยใช้ข้อมูลโครงการประกอบการวิเคราะห์การตัดสินใจอนุมัติงบประมาณอย่างเหมาะสม

- ผู้ที่มีหน้าที่วิเคราะห์ความต้องการของลูกค้าในองค์กรที่ผลิตซอฟต์แวร์ (Software) เพื่อจัดกลุ่มความต้องการของลูกค้าได้เหมาะสมกับบริการหรือสินค้าขององค์กร
- แพทย์และเภสัชกรผู้วินิจฉัยอาการของโรค ใช้ข้อมูลการรักษามาวิเคราะห์และจัดกลุ่มอาการของโรคแต่ละชนิดเพื่อง่ายต่อการรักษาและจ่ายยาแก่คนไข้
- นักกฎหมาย ผู้วินิจฉัยคดีความ ใช้รูปแบบข้อมูลประมวลกฎหมายชนิดต่างๆและข้อมูลการกระทำผิดในคดีความต่างๆเพื่อวิเคราะห์และจัดกลุ่มของประเภทความผิดของคดี
- นักวิเคราะห์และคาดการณ์ วิเคราะห์ข้อมูลเพื่อคาดการแนวโน้มที่จะเกิดขึ้นในอนาคต ในวัตถุประสงค์ที่แตกต่างกันไป เช่น นักลงทุนตลาดหุ้น พยากรณ์ อากาศ ผู้มีหน้าที่เกี่ยวเนื่องกับการป้องกันภัยพิบัติเป็นต้น
- นักการตลาด / ผู้ค้ารายย่อย ที่ต้องการคาดการแนวโน้มหรือความนิยมในการซื้อขายสินค้าและบริการในอนาคต
- นักศึกษา / วิจัย ที่ไม่มีความรู้ในทฤษฎีของต้นไม้ตัดสินใจ แต่อยากใช้ในการ วิเคราะห์ข้อมูลวิจัยเพื่อสรุปผลการวิจัยหรือเพื่อช่วยในการลดเวลาการทำงานให้อยู่ในระยะเวลาการดำเนินการวิจัย

1.2 **แนวทางการแก้ปัญหา**

การวิจัยนี้ได้ศึกษาการสร้างการสร้างต้นไม้ตัดสินใจที่มีประสิทธิภาพในด้านการจัดกลุ่มข้อมูล โดยอาศัยการฝึกระบบที่ใช้ข้อมูลที่ต้องการจัดกลุ่ม (75% จากข้อมูลทั้งหมด) และการวัดประสิทธิภาพจะวัดจากผลการทดสอบของข้อมูลชุดเดียวกันกับข้อมูลที่ใช้ฝึกระบบ (25%จากข้อมูลทั้งหมด) มีการใช้ทฤษฎีต้นไม้ตัดสินใจในวิธี C 4.5 เป็นวิธีที่เหมาะสมกับข้อมูลทั้งต่อเนื่องและไม่ต่อเนื่องซึ่งมีประสิทธิภาพในการประมวลผลดีกว่า แบบ ID3 ที่สามารถประมวลผลได้ในข้อมูลที่ไม่ต่อเนื่องเท่านั้น อีกทั้งการเตรียมข้อมูลก่อนประมวลผล จะทำให้ข้อมูลที่เข้าสู่กระบวนการโครงสร้างข้อต้นไม้ตัดสินใจเป็นข้อมูลที่มีประสิทธิภาพและทำให้ผลที่ออกมามีคุณภาพ อีกทั้งระบบยังถูกพัฒนาในรูปแบบเว็บแอพพลิเคชั่นทำใช้สามารถใช้งานได้โดยที่ผู้ต้องการใช้งานไม่จำเป็นต้องมีความรู้ความเชี่ยวชาญการวิเคราะห์และการจำแนกข้อมูล ระบบที่พัฒนาขึ้นมาจะสามารถจัดการข้อมูลที่ต้องการจำแนกได้ อีกทั้งยังลดระยะเวลาในการเปรียบเทียบโครงของต้นไม้ตัดสินใจในกรณีที่มีการเตรียมข้อมูลก่อนประมวลผลที่ต่างกัน เพื่อหาโครงสร้างของต้นไม้ตัดสินใจที่มีประสิทธิภาพและความถูกต้องมากที่สุด อีกทั้งระบบถูกพัฒนาขึ้นทำงานในรูปแบบเว็บไซต์ระบบประยุกต์ (Web Application) ทำให้มีความสะดวกสบายในการใช้งานนอกสถานที่และมีส่วนต่อผสานให้แก่ผู้ใช้งานได้ง่าย

1.3 **สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง**

จีรยุทธ ไชยจารุวณิช และดาวรุ่ง กังวานพงศ์ [2] กล่าวถึงการประยุกต์ใช้เทคนิค Decision Tree induction ในการจัดหมวดหมู่ของข้อมูลสารพันธุกรรมประเภท Microsatellites ที่ใช้ในการตรวจหาความใกล้เคียง หรือความแปรผันทางพันธุกรรมของชาวเขาทั้ง 4 เผ่า ได้แก่ ลีซอ อีก้อ กะเหรี่ยง และ ม้ง ซึ่งต่างจากการจำแนกแบบเดิมที่มีความละเอียดและใช้เทคนิคขั้นสูงทางการแพทย์ ทำให้มีความล่าช้าในการที่จะบ่งบอกความใกล้เคียงทางสายพันธุกรรม โดยผลของการทดลองของการประยุกต์ใช้เทคนิค Decision Tree induction จะเห็นได้ว่าเป็นที่น่าพึงพอใจเป็นอย่างมาก และสามารถจำแนกสายพันธุ์ โดยทดสอบจากข้อมูลความยาวของสายพันธุกรรม Microsatellites15 ตำแหน่ง ในโครโมโซม Y ของชาวเขา 51 คน

ชัดชัย แก้วตา และอัจฉรา มหาวีรวัฒน์ [3] ได้แสดงให้เห็นถึงการนำต้นไม้ตัดสินใจแบบ C 4.5 และ ID 3 มาใช้ในการวิเคราะห์และประมวลผลคดีโดยใช้ข้อมูลประมวลกฎหมายอาญาลักษณะความผิด 12 กรณีที่เกี่ยวกับทรัพย์สินในการสร้างต้นไม้ตัดสินใจทั้งสองแบบผ่านระบบ Weka version 3.6.2 โดยในแบบเดิมการจำแนกลักษณะความผิดทั้ง 12 กรณีนั้นต้องพิจารณาด้วยการเทียบคำหรือวลีในคดีเก่าว่าสอดคล้องกับคดีที่กำลังพิจารณาอยู่หรือไม่ ซึ่งผลการทดสอบจากการใช้ข้อมูลการทดลองจำนวนความผิด 4026 ความผิดและเลือกวิธี Cross-validation แสดงผลออกมาว่าต้นไม้ตัดสินใจแบบ C 4.5 และ ID 3 ประมาณ 94% ซึ่งแสดงให้เห็นว่า ต้นไม้ตัดสินใจสามารถใช้งานกับข้อมูลได้อย่างมีประสิทธิภาพ

พรทิพย์ พงษ์สวัสดิ์ และศิฬาณี นุชิตประสิทธิ์ชัย [4] นำข้อมูลงบประมาณที่เคยตัดสินในอดีตมาใช้ในการช่วยตัดสินงบประมาณในอนาคตโดยการประยุกต์ใช้รูปแบบการวิเคราะห์และจำแนกข้อมูลแบบต้นไม้ตัดสินใจ วิธีแบบ ID3 ซึ่งแตกต่างจากระบบเดิมคือการประชุมเพื่อจัดสรรค่าใช้จ่ายของแต่ละหน่วยงาน มีความยุ่งยากและใช้เวลานานทั้งที่มีความใกล้เคียงกับข้อมูลงบประมาณที่เคยติดสินในอดีตก็ตาม แต่จากการทดสอบโดยที่นำข้อมูลงบประมาณของแต่ละ

หน่วยงาน พบว่าระบบมีความถูกต้องในการตัดสินงบประมาณและมีประสิทธิในการทำงานเป็นอย่างมาก อีกทั้งสามารถการติดตามการใช้งบประมาณและเรียกดูรายงายในแบบต่างๆได้ตามที่ต้องการได้

ศักดิ์ชาย ตั้งประเสริฐ และพุธษดี ศิริแสงตระกูล [5] ได้นำการวิเคราะห์และจำแนกข้อมูล ในรูปแบบต้นไม้ตัดสินใจมาประยุกต์ใช้ในการพัฒนาระบบการจำแนกประเภทแบบทดสอบของผู้ทดสอบสุขภาพจิตรูปแบบออนไลน์ โดยแบบเดิมจะใช้เอกสารการสอบหรือโปรแกรมประยุกต์ บ่อยครั้งที่ผู้สอบมักหลีกเลี่ยงข้อสอบที่คำตอบเสี่ยงจะทำให้ผลชี้ว่าตนมีสภาวะผิดปกติ จึงทำให้การทดสอบไม่สามารถประมวลผลได้อย่างตรงจุดประสงค์เท่าที่ควร ดังนั้นการพัฒนาระบบจำแนกประเภทแบบทดสอบของผู้ทดสอบสุขภาพจิตในรูปแบบออนไลน์ที่ใช้การจำแนกข้อมูลในรูปแบบต้นไม้ตัดสินใจสามารถวิเคราะห์และจัดประเภทแบบทดสอบโดยใช้ข้อมูลของกลุ่มตัวอย่าง (Simple Size) ที่ใช้สูตรของ Taro Yamane ในการคำนวณจากข้อมูลประชากรทั้งหมด ทำให้ได้ขนาดกลุ่มตัวอย่างข้อมูลประชากรไทยประมาณ 400 คน มาทำการทดสอบในระบบ ซึ่งผลการทดสอบที่ออกมาอยู่ในเกณฑ์ค่อนข้างดี

กฤษณะ ไวยมัย และชิดชนก ส่งศิริ [6] กล่าวถึงการวิเคราะห์ข้อมูลและการจำแนกข้อมูลในรูปแบบต้นไม้ตัดสินใจที่ช่วยในการตัดสินใจให้แก่นักศึกษาให้สามารถเลือกวิชาที่จะเรียนได้อย่างเหมาะสม โดยนำตัวอย่างข้อมูลของนักศึกษาที่มีผลการเรียนที่ดีทุกรายวิชามาทำการวิเคราะห์และสร้างต้นไม้ตัดสินใจ แต่ผลที่ออกมายังมีความถูกต้องไม่มากเท่าที่ควรหลังจากที่ได้มีการนำกลุ่มข้อมูลมาทดสอบ ทั้งนี้อาจอยู่ที่การเตรียมข้อมูลก่อนการประมวลผลที่ยังไม่มีความเหมาะสม ทางผู้วิจัยจึงเปลี่ยนเป็นการวิเคราะห์และสร้างต้นไม้ตัดสินใจเพื่อจำแนกประเภทข้อมูลของแต่ละสาขาโดยพิจารณาถึงความเหมาะสมของนักศึกษาว่าเหมาะสมในการเรียนกับสาขานั้นๆหรือไม่ และผลจากการทดสอบความถูกต้องของผลการเรียนปลายภาคที่แสดงถึงความแม่นยำของระบบในการคาดการ มีประสิทธิภาพและความถูกต้องถึง 84.58%

พรชัย คำเพิงใจ และสิรภัทร เชี่ยวชาญวัฒนา [7] จากอดีตการจัดสรรงบประมาณให้กับโรงพยาบาลแต่ละแห่งโดยการพิจารณาเปรียบเทียบความสอดคล้องตามมาตรฐานทั่วไปของข้อมูลการเงินและการรักษาในโรงพยาบาลรัฐบาล ซึ่งบางครั้งการรักษาโรคเดียวกันมีขั้นตอนหรือวิธีการที่ต่างกันทำให้ความสอดคล้องของการเงินและการรักษาต่างไปจากเดิม จะส่งผลต่อการจัดสรรงบประมาณที่ผิดพลาดได้ ดังนั้นการประยุกต์เอาวิธีการต้นไม้ตัดสินใจ และ K-mean มาใช้ในการ

จำแนกข้อมูลการรักษาพยาบาลผู้ป่วยอยู่รักษาของโรงพยาบาลศูนย์ประเทศไทยจึงเป็นสิ่งที่ทำให้การพิจารณาความสอดคล้องนั้นมีประสิทธิภาพมากยิ่งขึ้น โดยการนำข้อมูลจากสำนักงานหลักประกันสุขภาพแห่งชาติและสำนักกลางสารสนเทศบริการสุขภาพเฉพาะข้อมูลการให้บริการผู้ป่วยในของโรงพยาบาลศูนย์มาทำการเตรียมข้อมูลก่อนประมวลผลและจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจแล้วจึงนำข้อมูลในแต่ละกลุ่มมาทำการ Clustering โดยใช้หลักการของ K-mean ผลการทดลองพบว่าการที่ใช้ต้นไม้ตัดสินใจ และ K-mean ในการจำแนกกลุ่มมีประสิทธิภาพมากกว่าการจัดกลุ่มโดยใช้ข้อมูลทั้งหมดครั้งเดียว

1.4 **หลักการและทฤษฎี**

1.4.1 **การเตรียมข้อมูลก่อนการประมวลผล (Preprocessing)**

โดยส่วนมากข้อมูลปฐมภูมิที่รับมาจากแหล่งที่มาโดยตรงมักจะเป็นข้อมูลที่ไม่สะอาดหมายถึงข้อมูลที่มีสิ่งแปลกปลอมปะปนเข้ามา เช่น ข้อมูลที่ผิดรูปแบบ ข้อมูลขาดหายไม่ครบถ้วนสมบูรณ์เป็นต้นซึ่งทำให้ข้อมูลโดยรวมผิดเพี้ยนไปหรืออาจจะทำให้ไม่สามารถประมวลผลต่อได้ ดังนั้น การเตรียมข้อมูลก่อนการประมวลผลจึงเป็นสิ่งที่สำคัญเป็นอย่างยิ่ง

ส่วนมากการวิธีการเตรียมข้อมูลก่อนการประมวลผลมีขั้นตอนดังนี้

- **การบอกลักษณะโดยรวมของข้อมูล (Descriptive data summarization)**

คือขั้นตอนการทราบถึงลักษณะของข้อมูล เช่น แนมโน้มของข้อมูล การกระจายตัวของข้อมูลเป็นต้น ซึ่งส่วนมากในการจะทราบถึงนั้นต้องมีการคำนวณเพื่อวัดค่าของข้อมูล

- การเฉลี่ยข้อมูล (Mean) คือการหาค่าเฉลี่ยกลางของข้อมูลโดยการหาแบ่งออกเป็นสองวิธีดังนี้

  - การหาค่าเฉลี่ยเทียบกับจำนวนประชากร

    สูตรการคำนวณดังนี้การคำนวณคือ

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N} \quad (1)$$

$\bar{x}$ คือค่าเฉลี่ย

xi คือ ตัวแปรแทนค่าลำดับที่ i

N คือ จำนวนข้อมูลทั้งหมด

- การหาค่าเฉลี่ยเทียบกับน้ำหนักของข้อมูล

สูตรการคำนวณการคำนวณคือ

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} wi} \quad (2)$$

$\bar{x}$ คือ ค่าเฉลี่ย

Xi คือตัวแปรแทนค่าลำดับที่i

Wj คือ ตัวแปรแทนน้ำหนักข้อมูลลำดับที่ j

- การหามัธยฐาน (Median) คือการหาจุดกึ่งกลางของข้อมูลโดยการเรียงลำดับข้อมูลจากน้อยไปหามาก

สูตรการคำนวณการคำนวณคือ

$$median = L_1 + (\frac{\frac{n}{2} - (\sum f)l}{f_{median}}) \quad (3)$$

L_1 คือ ขีดจำกัดล่างแท้จริงของชั้นที่มัฐยฐานอยู่

n คือ จำนวนข้อมูลทั้งหมด

f_{median} คือ ความถี่ชั้นที่ชั้นมัธยฐานอยู่

f คือ ความถี่ชั้นที่ต่ำกว่าชั้นมัธยฐานอยู่

l คือ ความกว้างของอันตรภาคชั้น

- การหาค่าฐานนิยม (Mode) คือการหาความถี่สูงสุดของข้อมูลทั้งหมดหรือหาค่าที่เกิดขึ้นบ่อย

สูตรการคำนวณการคำนวณคือ

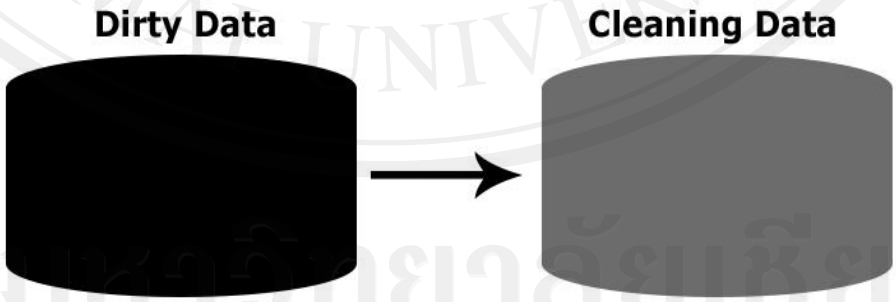
$$mean - mode = 3 \times (mean - median) \quad (4)$$

$mean$ คือ ค่าเฉลี่ย

$mode$ คือค่าฐานนิยม

- **การทำความสะอาดข้อมูล (Data Cleaning)**

ข้อมูลปฐมภูมิที่ได้จากแหล่งที่มาโดยตรง ส่วนมากข้อมูลเหล่านี้มักจะเป็นข้อมูลที่ไม่สะอาด กล่าวคือข้อมูลที่ไม่สมบูรณ์หรือเบี่ยงเบนจากความเป็นจริง ทั้งนี้อาจขึ้น อยู่กับผู้ที่ให้ข้อมูลเองที่พยายามให้ข้อมูลที่เกินจริง หรือ การเก็บข้อมูลที่มีชนิดที่ไม่สอด คล้องกัน(ต่างรูปแบบ เช่น ขนาดความยาวที่มีมาตรวัดที่ต่างกัน เป็นต้น) หรือ เป็นการผิด พลาดของผู้ที่มีหน้าที่เก็บข้อมูลเองที่รับข้อมูลที่ไม่มีความสอดคล้องกัน (เช่น เงินเดือนมีค่าติดลบ) ซึ่งสิ่งเหล่านี้จะเป็นปัจจัยสำคัญในการที่จะทำให้การประมวลผลผิดพลาดและอาจ จะทำให้การประมวลผลล้มเหลว



รูปที่ 1.1 ภาพที่แสดงถึงแนวคิดของการทำความสะอาดข้อมูล [1]

การทำความสะอาดข้อมูลมีวิธีดังนี้

- การสร้างประเภทข้อมูลที่ไม่รู้จัก (Class Unknown) สำหรับจัดกลุ่มประเภทข้อมูลที่ไม่สามารถจัดกลุ่มได้เนื่องจากข้อมูลไม่ครบถ้วน
- การหาค่าเฉลี่ยของข้อมูลในคุณลักษณะ (Attribute) เดียวกันเพื่อหาค่าเฉลี่ยที่ไม่มีผลต่อข้อมูลอื่นการคำนวณในการประมวลผล
- การกำจัดข้อมูลที่ซ้ำกันเพื่อป้องกันการประมวลผลซ้ำซ้อนและอาจจะทำให้เกิดการเบี่ยงเบนของผลลัพธ์ที่จะเกิดขึ้น
- การกำจัดข้อมูลรบกวน (Noisy) เป็นสิ่งที่จำเป็นอย่างมากในการทำความสะอาดข้อมูล โดยมีอยู่ด้วยกันหลากหลายวิธีด้วยกันดังนี้

  - บินดิ้ง (Binding)

    คือการแบ่งข้อมูลเป็นกลุ่มๆ โดยเริ่มจากการจัดเรียงข้อมูลจากน้อยไปหามากและแบ่งกลุ่มข้อมูลตามความถี่

    หลังจากนั้นเลือกวิธีการแบ่งกลุ่มดังนี้

    การหาค่าเฉลี่ย

    การหาจากค่ากลางของข้อมูล

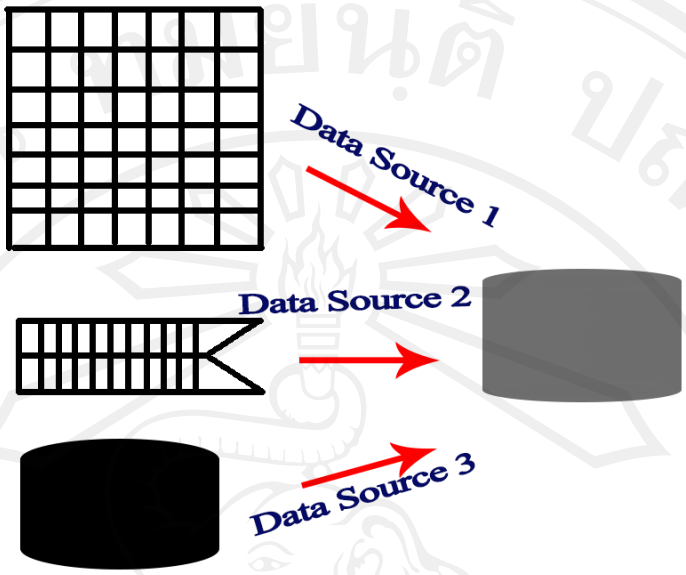
    การหาจากขอบของข้อมูล

  - รีเกรทชั่น (Regression)

    คือการทำให้ข้อมูลละเอียดขึ้นด้วยการกรองข้อมูลด้วย รีเกรทชั่น ฟังก์ชัน (Regression functions)

- การรวบรวมข้อมูลจากหลายแหล่งที่มาให้อยู่ในรูปเดียวกัน

  คือการแปลงข้อมูลที่มาจากหลายแหล่งที่มาให้อยู่ในรูปแบบที่เหมาะสมกันการใช้งาน (Data Integration and Data Transformation)



รูปที่ 1.2 ภาพที่แสดงถึงแนวคิดของการรวบรวมข้อมูลจากหลายแหล่ง [1]

การรวบรวมข้อมูลมีวิธีดังนี้

- การวิเคราะห์ความสัมพันธ์ (ใช้กับข้อมูลที่เป็นตัวเลข)

โดยใช้สูตรการคำนวณคือ

$$r_{A,B} = \frac{\sum(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum(AB)-n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B} \quad (5)$$

$r_{A,B}$ คือ ค่าการวิเคราะห์ความสัมพันธ์

A คือ ข้อมูล A

$\bar{A}$ คือ ค่าเฉลี่ย A

B คือ ข้อมูล B

$\bar{B}$ คือ ค่าเฉลี่ย B

n คือ จำนวนข้อมูลทั้งหมด

σ_A คือ ค่าเบี่ยงเบนมาตรฐาน คุณลักษณะ A

σ_B คือ ค่าเบี่ยงเบนมาตรฐาน คุณลักษณะ B

- การวิเคราะห์ความสัมพันธ์

โดยใช้สูตรการคำนวณคือ

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \quad (6)$$

χ^2 คือ ค่าการวิเคราะห์ความสัมพันธ์

$Observed$ คือ ค่าสังเกตการณ์

$Expected$ คือ ค่าที่คาดหวัง

การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกันการใช้งาน โดยการลดการซ้ำซ้อนของข้อมูลมีวิธีดังนี้

- การแปลงข้อมูลโดยสัมพันธ์กับการหาค่าสูงสุดและต่ำสุดใหม่

โดยใช้สูตรการคำนวณคือ

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (7)$$

v' คือ ค่าคุณลักษณะใหม่

v คือ ค่าคุณลักษณะเดิม

min_A คือ ค่าต่ำสุดเดิมของคุณลักษณะ A

max_A คือ ค่าสูงสุดเดิมของคุณลักษณะ A

new_max_A คือ ค่าต่ำสุดใหม่ของคุณลักษณะ A

new_min_A คือ ค่าสูงสุดใหม่ของคุณลักษณะ A

- การแปลงข้อมูลโดยสัมพันธ์กับค่าเฉลี่ย-ส่วนเบี่ยงเบนมาตรฐาน

โดยใช้สูตรการคำนวณคือ

$$v' = \frac{v - \mu_A}{\sigma_A} \quad (8)$$

v' คือ ค่าคุณลักษณะใหม่

v คือ ค่าคุณลักษณะเดิม

μ_A คือ ค่าเฉลี่ย คุณลักษณะ A

σ_A คือ ค่าเบี่ยงเบนมาตรฐาน คุณลักษณะ A

- การแปลงข้อมูลให้อยู่ในรูปมาตราส่วนฐาน10

โดยใช้สูตรการคำนวณ คือ

$$v' = \frac{v}{10^j} \quad (9)$$

v' คือ ค่าคุณลักษณะใหม่

v คือ ค่าคุณลักษณะเดิม

j คือ ค่าที่ทำให้ Max(|v|) <1

1.4.2 **ต้นไม้ตัดสินใจ C 4.5 (Decision tree C 4.5)**

การสร้างโครงสร้างของต้นไม้ตัดสินใจมี ขั้นตอนในการทำดังนี้

- **การกำหนดข้อมูลชื่อกลุ่มที่ใช้จำแนกข้อมูลและระบุตัวย่อของกลุ่มในข้อมูล**

เช่น Class P : buy_Computer yes ,Class N : buy_Computer no เป็นต้น

- **การคำนวณหาค่าคาดการของกลุ่มการจำแนกข้อมูล**

โดยการใช้สูตรการคำนวณคือ

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j) \quad (10)$$

$Info_A(D)$ คือ ค่าคาดการของกลุ่มการจำแนกข้อมูล

D_j คือ ค่าจำนวนทางเลือกที่ j

D คือ ค่าจำนวนทางเลือกทั้งหมด

I คือ ค่าคาดการคาดการของคุณลักษณะย่อย

- **การคำนวณหาข้อมูลคาดการเพื่อหาโหนด (Node) แยกของแต่ละคุณลักษณะ**

โดยการใช้สูตรการคำนวณคือ

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (11)$$

$Info(D)$ คือ การคาดการคาดการแต่ละคุณลักษณะ

p_i คือ อัตราส่วนทางเลือกที่ i ต่อทางเลือกทั้งหมด

- **การคำนวณหาค่าได้เปรียบเพื่อเปรียบเทียบลำดับความสำคัญของแต่ละคุณลักษณะ**

โดยการใช้สูตรการคำนวณคือ

$$Gain(A) = Info(D) - Info_A(D) \quad (12)$$

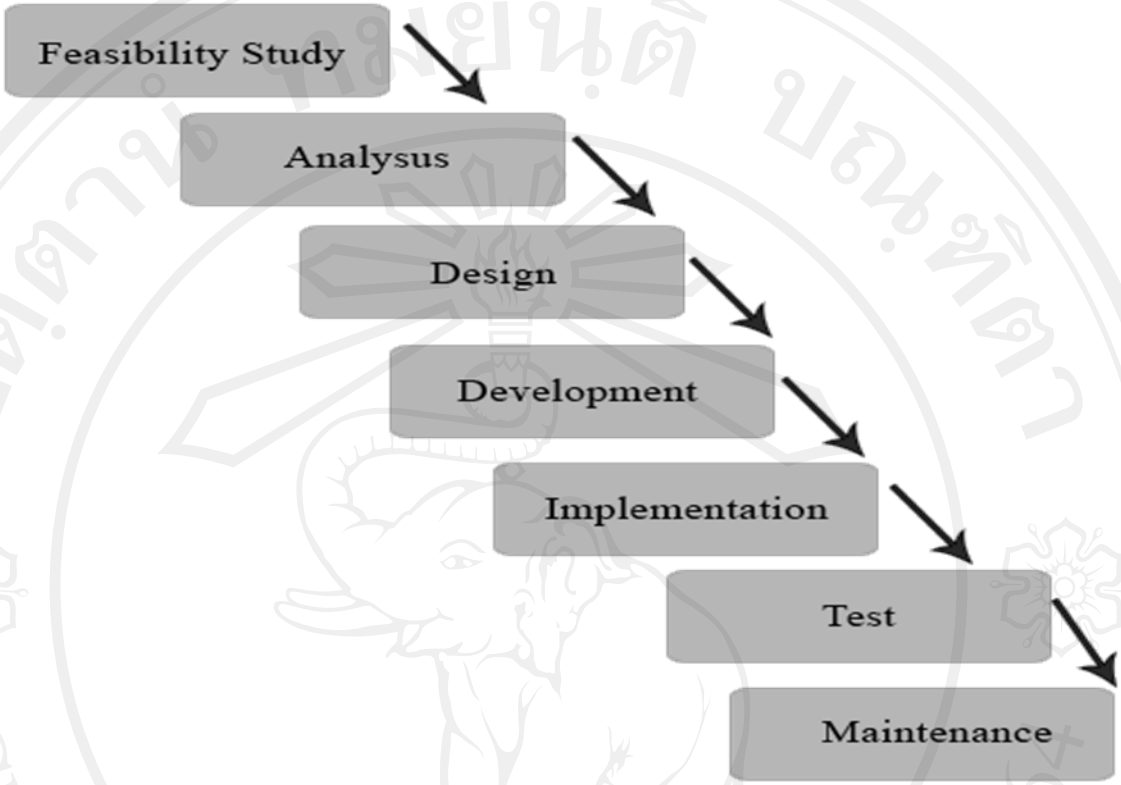
$Gain(A)$ คือ ค่าได้เปรียบ

$Info(D)$ คือ ค่าคาดการคาดการแต่ละคุณลักษณะ

$Info_A(D)$ คือ ค่าคาดการของกลุ่มการจำแนกข้อมูล

1.4.3 วงจรชีวิตการพัฒนาซอฟต์แวร์ (Software Development Life Cycle : SDLC)

การวางแผนการสร้างซอฟต์แวร์อย่างมีประสิทธิภาพตามแนวทางของวิศวกรรมซอฟต์แวร์ เพื่อให้การพัฒนาซอฟต์แวร์มีระบบและขั้นตอนอย่างแน่ชัดทำให้ง่ายต่อการจัดการทั้งด้านการแก้ไขหรือการเปลี่ยนแปลงเมื่อเกิดปัญหากับกาพัฒนา



รูปที่ 1.3 ภาพที่แสดงถึงวงจรชีวิตการพัฒนาซอฟต์แวร์ [12]

การพัฒนาซอฟต์แวร์ตามขั้นตอนของวงจรชีวิตการพัฒนาซอฟต์แวร์

- **การศึกษาความเป็นไปได้ของซอฟต์แวร์ที่พัฒนา (Feasibility Study)**

คือการระบุหัวหรือแนวทางของซอฟต์แวร์ที่ต้องการพัฒนารวมถึงการวางแผนเพื่อพัฒนาซอฟต์แวร์

- **การวิเคราะห์การพัฒนาซอฟต์แวร์ (Analysis)**

คือการวิเคราะห์ซอฟต์แวร์เพื่อเป็นแนวทางการออกแบบทั้งความต้องการของระบบและพฤติกรรมการใช้งาน

- **การออกแบบ (Design)**

คือการออกแบบซอฟต์แวร์ทั้งหมด ไม่ว่าจะเป็นส่วนต่อผสานกับผู้ใช้งานและส่วนการประมวลผลของซอฟต์แวร์ อีกทั้งส่วนของการบันทึกข้อมูล ให้เป็นไปตามข้อกำหนดของซอฟต์แวร์ที่ได้วิเคราะห์ไว้แล้ว

- **การพัฒนา (Development)**

คือการพัฒนาซอฟต์แวร์ตามที่ได้ออกแบบไว้เพื่อทำให้ซอฟต์แวร์ที่ออกแบบไว้สามารถประมวลได้ตรงตามวัตถุประสงค์

- **การดำเนินการ (Implementation)**

คือการดำเนินการที่จะทำให้การพัฒนาซอฟต์แวร์สมบูรณ์พร้อมใช้งาน

- **การทดสอบ (Test)**

คือการทดสอบซอฟต์แวร์เพื่อให้เป็นไปตามแผนที่วางไว้รวมถึงการทดสอบซอฟต์แวร์ให้สามารถใช้งานได้จริง

- **การบำรุงรักษา (Maintenance)**

คือการบำรุงรักษาซอฟต์แวร์หลังจากการติดตั้งใช้งานและการปรับปรุงเพื่อให้ซอฟต์แวร์สามารถทำงานได้อย่างมีประสิทธิ์ภาพมากขึ้น

1.4.4 กระบวนการพัฒนาโครงสร้างข้อมูล (Information Architecture Development Process)

โครงสร้างข้อมูลเป็นสิ่งที่สำคัญเป็นอย่างมากในการที่จะนำเสนอให้ผู้ที่ใช้งานเข้าใจตรงกันกับวัตถุประสงค์ของผู้พัฒนา บ่อยครั้งที่ความผิดพลาดเกิดจากการสื่อสารจึงทำให้เกิดปัญหาตามมามากมาย ซึ่งการจะพัฒนาเว็บไซต์ให้รองรับถึงการใช้งานได้ง่ายของผู้ใช้ (Uses ability) นั้นมีความจำเป็นอย่างมากที่ต้องศึกษาและพัฒนาโครงสร้างให้สอดคล้องกับผู้ใช้งานและรวมถึงการจัดกลุ่มข้อมูลที่เอื้อต่อผู้ใช้งานจะพบได้ง่าย

รูปที่ 1.4 ภาพที่แสดงถึงกระบวนการพัฒนาโครงสร้างข้อมูล [14]