การเปรียบเทียบเทคนิคการจำแนกข้อมูลและการรวมกลุ่มข้อมูลในการคัดเลือกแม่พิมพ์โลหะแบบ Progressive Die

โดย
นายแสนศักดิ์ ชาวปากน้ำ

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
ภาควิชาคอมพิวเตอร์
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2553

การเปรียบเทียบเทคนิคการจำแนกข้อมูลและการรวมกลุ่มข้อมูลในการคัดเลือกแม่พิมพ์โลหะแบบ Progressive Die

โดย

นายแสนศักดิ์ ชาวปากน้ำ

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

ภาควิชาคอมพิวเตอร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2553

COMPARISON OF CLASSIFICATION AND CLUSTERING TECHNIQUE FOR PROGRESSIVE DIE SELECTION

By

Saensak Chaopaknam

An Independent Study Submitted in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

Department of Computing

Graduate School

SILPAKORN UNIVERSITY

2010

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร อนุมัติให้การค้นคว้าอิสระเรื่อง " การเปรียบเทียบเทคนิคการจำแนกข้อมูลและการรวมกลุ่มข้อมูลในการคัดเลือกแม่พิมพ์โลหะแบบ Progressive Die " เสนอโดย นายแสนศักดิ์ ชาวปากน้ำ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

…….....

(ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์)

คณบดีบัณฑิตวิทยาลัย

วันที่.....เดือน..... พ.ศ.....

อาจารย์ที่ปรึกษาการค้นคว้าอิสระ

ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์

คณะกรรมการตรวจสอบการค้นคว้าอิสระ

..... ประธานกรรมการ

(อาจารย์ ดร.สุนีย์ พงษ์พินิจภิญโญ)

...../...../.....

..... กรรมการ

(อาจารย์ ดร.วัสรา รอดเหตุภัย)

...../...../.....

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์)

...../...../.....

49309335 : สาขาวิชาเทคโนโลยีสารสนเทศ
คำสำคัญ : เหมืองข้อมูล / ต้นไม้ตัดสินใจ / การจัดกลุ่มแบบลำดับชั้น / ดัชนีวัดผลการปฏิบัติงาน

แสนศักดิ์ ชาวปากน้ำ : การเปรียบเทียบเทคนิคการจำแนกข้อมูลและการรวมกลุ่มข้อมูลในการคัดเลือกแม่พิมพ์โลหะแบบ Progressive Die. อาจารย์ที่ปรึกษาการค้นคว้าอิสระ : ผศ.ดร.ปานใจ ธารทัศนวงศ์. 119 หน้า.

การค้นคว้าอิสระนี้มีวัตถุประสงค์เพื่อศึกษาและคัดเลือกเทคนิคการทำเหมืองข้อมูลที่เหมาะสมกับการคัดเลือกกลุ่มแม่พิมพ์โลหะแบบ Progressive Die ซึ่งเป็นแม่พิมพ์ที่มีราคาสูงและออกแบบยาก โดยศึกษาเปรียบเทียบระหว่างเทคนิคการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) จากอัลกอริทึม C4.5 และเทคนิคการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) จากอัลกอริทึมการเชื่อมโยงเฉลี่ย (Average Link) ข้อมูลที่ใช้ในการค้นคว้าอิสระนี้นำมาจากฐานข้อมูลของบริษัท อาพิค ยามาดะ (ประเทศไทย) จำกัด ระหว่างปี พ.ศ. 2550 ถึงปี พ.ศ. 2552 และได้ประยุกต์ใช้ข้อมูลดัชนีวัดผลการปฏิบัติงาน (KPI) เป็นเกณฑ์ในการจัดกลุ่มแม่พิมพ์ร่วมด้วย จากนั้นทำการเปรียบเทียบผลการจัดกลุ่มแม่พิมพ์ของทั้งสองเทคนิคด้วยวิธีวัดค่าความเคลื่อน (Error Measurement) จำนวน 3 วิธี คือ Root Mean Squared Error, Mean Absolute Error และ Relative Absolute Error

ผลการประเมินและเปรียบเทียบค่าความคลาดเคลื่อนพบว่าเทคนิคการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) มีค่าความคลาดเคลื่อนต่ำกว่าเทคนิคการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) อย่างชัดเจนทั้ง 3 วิธี โดยเฉพาะค่า Relative Absolute Error ที่แตกต่างกันมาก กล่าวคือ อัลกอริทึมโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) มีค่า Relative Absolute Error ที่ 0.0802% และเทคนิคการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) มีค่า Relative Absolute Error ที่ 19.384% สำหรับค่าความคลาดเคลื่อนที่เหลือมีความแตกต่างกันเล็กน้อย

จากผลการเปรียบเทียบค่าความเคลื่อนสามารถสรุปได้ว่า เทคนิคเหมืองข้อมูลที่เหมาะสมกับการคัดเลือกกลุ่มแม่พิมพ์โลหะแบบ Progressive Die คือ อัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจแบบ C4.5

ภาควิชาคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ปีการศึกษา 2553

ลายมือชื่อนักศึกษา .....

ลายมือชื่ออาจารย์ที่ปรึกษาการค้นคว้าอิสระ .....

49309335 : MAJOR : INFORMATION TECHNOLOGY
KEY WORDS : DATA MINING/ DECISION TREE/ HIERARCHICAL CLUSTERING/ KPI
SAENSAK CHAOPAKNAM : COMPARISON OF CLASSIFICATION AND CLUSTERING TECHNIQUE FOR PROGRESSIVE DIE SELECTION. INDEPENDENT STUDY ADVISOR : ASST.PROF.PANJAI TANTATSANAWONG, Ph.D., 119 pp.

The purpose of this independent study is to study and to choose the appropriate data mining technique for Progressive Die selection which is high value and hard to design. This study is to compare between Classification with C4.5 Decision Tree algorithm and Hierarchical Clustering with Average Link algorithm. The sample data used in this study comes from transactional database of APIC Yamada (Thailand) Co., Ltd. during 2007 and 2009 and the Key Performance Indicator (KPI) is applied to classify data for Progressive Die selection. Three methods of error measurement; Root Mean Squared Error, Mean Absolute Error and Relative Absolute Error are used to compare between Decision Tree algorithm and Hierarchical Clustering for Progressive Die selection .

By comparing the value of the above error measurement, it can be concluded that the C4.5 Decision Tree algorithm has lower value than that of the Hierarchical Clustering algorithm. Especially, the value of Relative Absolute Error of Decision Tree is 0.0802% and that of Hierarchical Clustering is 19.384%. For the other methods, comparison of the value is slightly different.

By comparing the value of error measurement, it can be concluded that the classification with C4.5 Decision Tree algorithm is the appropriate data mining technique for Progressive Die selection.

Department of Computing Graduate School, Silpakorn University Academic Year 2010
Student's signature .....
Independent Study Advisor's signature .....

กิตติกรรมประกาศ

การค้นคว้าอิสระฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณาอย่างสูงจาก ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์ อาจารย์ที่ปรึกษาในการค้นคว้าอิสระ ที่สละเวลาให้คำปรึกษา คำแนะนำในการแก้ไขข้อบกพร่องต่างๆ ด้วยความเมตตาและเอาใจใส่เป็นอย่างดียิ่ง ขอกราบขอบพระคุณในความกรุณาของอาจารย์มา ณ โอกาสนี้

ขอกราบขอบพระคุณ ดร.สุนีย์ พงษ์พินิจภิญโญ ประธานกรรมการสอบการค้นคว้าอิสระ และดร.วัสรา รอดเหตุภัย กรรมการผู้ทรงคุณวุฒิ รวมถึงคณะกรรมการสอบทุกท่านที่กรุณาให้คำแนะนำเพื่อแก้ไขให้การค้นคว้าอิสระฉบับนี้สมบูรณ์ยิ่งขึ้น ตลอดจนอาจารย์ทุกท่านที่ประสิทธิ์ประสาทวิชาความรู้ให้แก่ข้าพเจ้า

ขอขอบคุณคุณประวิม เหลืองสมานกุล ตลอดจนบุคลากรภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์ทุกท่าน ที่อำนวยความสะดวกและให้คำแนะนำเกี่ยวกับการศึกษาด้วยดีมาโดยตลอด และขอขอบคุณเพื่อนๆ บัณฑิตศึกษาทุกท่านที่ให้ความช่วยเหลือ เป็นกำลังใจและมีมิตรภาพที่ดีเสมอมา

สุดท้ายนี้ขอกราบขอบพระคุณบิดา มารดา ญาติ พี่น้อง และบุคคลที่ใกล้ชิดที่ให้กำลังใจ ความห่วงใย และสนับสนุนการศึกษา ส่งผลให้ข้าพเจ้าประสบความสำเร็จในวันนี้

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1
บทนำ

การแข่งขันทางธุรกิจในอุตสาหกรรมและการค้านับวันยิ่งทวีความรุนแรงเพิ่มมากขึ้น รวมถึงวิธีการและรูปแบบในการดำเนินธุรกิจได้มีการพัฒนาไปอย่างต่อเนื่องและหลากหลาย ในยุคที่ข้อมูลสารสนเทศเป็นตัวแปรสำคัญในการดำเนินธุรกิจ องค์กรต่างๆจึงให้ความสำคัญและต้องการใช้ข้อมูลทั้งจากภายในองค์กรและภายนอกองค์กร เพื่อชิงความได้เปรียบในการดำเนินธุรกิจให้ได้มากที่สุด ข้อมูลในการดำเนินธุรกิจเหล่านี้ถูกสะสมและเก็บรวมรวบไว้เป็นจำนวนมากจากหลายหน่วยงานภายในองค์กรเป็นเวลาหลายปี โดยที่ข้อมูลเหล่านี้สามารถนำมาใช้ให้เกิดประโยชน์ได้อย่างมากแต่ยังไม่ได้ถูกนำมาใช้อย่างจริงจังหรือยังไม่ได้ถูกนำมาประยุกต์ใช้ร่วมกัน

1. ความเป็นมาและความสำคัญของปัญหา

ในอุตสาหกรรมผลิตชิ้นส่วนอิเลคทรอนิคส์โดยเฉพาะการผลิต IC Lead Frame ซึ่งเป็นชิ้นส่วนเริ่มต้นในการผลิต Semiconductor ชิ้นส่วนชนิดนี้ผลิตจากแม่พิมพ์โลหะแบบ Progressive Die ซึ่งเป็นแม่พิมพ์ที่มีราคาสูงและออกแบบยาก การคัดเลือกแม่พิมพ์หรือการปรับแต่งแม่พิมพ์ให้เหมาะสมกับชิ้นงานจึงเป็นตัวแปรสำคัญในเรื่องของคุณภาพชิ้นงาน ประสิทธิภาพในการผลิต และผลกำไรจากการดำเนินงาน แม้จะมีขั้นตอนและกระบวนการคัดเลือกแม่พิมพ์ก่อนการตัดสินใจผลิตแล้วก็ตาม แต่ก็พบว่าข้อมูลเหล่านั้นมาจากหลายหน่วยงานและใช้ข้อมูลในปัจจุบันเป็นหลัก อีกทั้งอาศัยประสบการณ์เฉพาะบุคคลของหน่วยงานที่เกี่ยวข้องเป็นตัวชี้วัด โดยขาดการประยุกต์ใช้ข้อมูลในอดีตร่วมกันเพื่อสนับสนุนกระบวนการคัดเลือกแม่พิมพ์ดังกล่าว ดังนั้นเมื่อมีการปรับเปลี่ยนบุคคลที่เกี่ยวข้อง ประสิทธิภาพในการคัดเลือกแม่พิมพ์ก็จะได้รับผลกระทบไปด้วย

มาตรฐานสากลที่ใช้วัดประสิทธิภาพในการผลิตหรือการดำเนินงานก็เป็นอีกปัจจัยหนึ่งที่สร้างความซับซ้อนในการคัดเลือกแม่พิมพ์ เช่น KPI (Key Performance Indicator) ซึ่งเป็นดัชนีชี้วัดผลการปฏิบัติงานของแต่ละหน่วยงาน ตัวอย่างเช่น แผนกขายตั้ง KPI ยอดขายที่ 80 ล้านบาทต่อเดือน และแผนกผลิตตั้ง KPI จำนวนของเสียจากการผลิตน้อยกว่า 2% ต่อเดือน หากแม่พิมพ์ที่ได้รับจากลูกค้าสามารถทำกำไรได้มากแต่ผลิตยาก และทำให้เกิดของเสียในกระบวนการผลิตมาก กรณีนี้จะสร้างเงื่อนไขในกระบวนการคัดเลือกแม่พิมพ์ให้กับทั้ง 2 แผนก ดังนั้นหากแม่พิมพ์ใดให้ KPI ด้านบวกกับหน่วยงานหนึ่ง แต่ก่อให้เกิด KPI ด้านลบกับหน่วยงานอื่น อาจทำให้เกิดข้อขัดแย้งและทำให้กระบวนการคัดเลือกแม่พิมพ์นั้นซับซ้อนและยาวนานขึ้น อย่างไรก็ดีหากสามารถวิเคราะห์

ข้อมูลในหลายมิติจะช่วยลดระยะเวลาและข้อขัดแย้งในกระบวนการคัดเลือกแม่พิมพ์ลงได้ อีกทั้งทำให้ผู้บริหารสามารถวางแผนการผลิตได้เร็วขึ้น

จากเหตุผลดังกล่าวผู้วิจัยได้ใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นการสำรวจและวิเคราะห์ข้อมูลขนาดใหญ่เพื่อใช้เป็นแนวทางในการคัดเลือกแม่พิมพ์ ซึ่งเทคนิคการทำเหมืองข้อมูลนั้นมีอัลกอริทึมให้เลือกใช้มากมาย ได้แก่ อัลกอริทึมที่ใช้หลักการของโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมที่ใช้หลักการของเครือข่ายใยประสาท (Neural Network) อัลกอริทึมที่ใช้หลักการหาความสัมพันธ์ของข้อมูล (Association Rule) และอัลกอริทึมอื่นๆ อีกมาก ซึ่งไม่มีอัลกอริทึมใดหรือเทคนิคใดที่ทำงานได้ดีที่สุดกับข้อมูลทุกประเภท ทั้งนี้เนื่องจากข้อมูลแต่ละประเภทมีลักษณะเฉพาะตัวที่แตกต่างกันตามลักษณะเฉพาะของงานหรือของธุรกิจนั้น ดังนั้นจึงกล่าวได้ว่าไม่มีอัลกอริทึมใดที่ดีที่สุดสำหรับข้อมูลทุกประเภท

การค้นคว้าอิสระนี้จึงเสนอขึ้นเพื่อศึกษาลักษณะของอัลกอริทึมที่เหมาะสมกับข้อมูลด้านการคัดเลือกแม่พิมพ์โลหะแบบ Progressive Die โดยเน้นข้อมูลที่สอดคล้องกับดัชนีชี้วัดผลการปฏิบัติงาน (KPI) ของแผนกที่เกี่ยวข้องกับการคัดเลือกแม่พิมพ์จำนวน 3 แผนก คือ แผนกผลิต แผนกขาย และแผนกซ่อมบำรุงแม่พิมพ์ จากนั้นจะทำการทดสอบอัลกอริทึมที่แตกต่างกัน 2 เทคนิค และเปรียบเทียบผลลัพธ์ที่ได้เพื่อวิเคราะห์และประเมินผลอัลกอริทึมที่เหมาะสมกับกลุ่มข้อมูลมากที่สุด เพื่อใช้เป็นแนวทางประกอบการพิจารณาในการคัดเลือกแม่พิมพ์ โดยอัลกอริทึมที่ใช้ทดสอบคือ อัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

2. วัตถุประสงค์การวิจัย

2.1 เพื่อศึกษาและเปรียบเทียบเทคนิคการทำเหมืองข้อมูลระหว่างอัลกอริทึมการจำแนกข้อมูล (Classification) ด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการรวมกลุ่มข้อมูล (Clustering) ด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

2.2 เพื่อเป็นแนวทางในการคัดเลือกแม่พิมพ์โลหะเพื่อการผลิต

3. ประโยชน์ที่คาดว่าจะได้รับ

3.1 แนวทางหรือวิธีการในการคัดเลือกตัวแทนกลุ่มเพื่อใช้ในการคัดเลือกแม่พิมพ์

3.2 เข้าใจถึงอัลกอริทึมที่เหมาะสมกับกลุ่มข้อมูลที่ใช้ในการคัดเลือกแม่พิมพ์

3.3 สามารถลดระยะเวลาในการคัดเลือกแม่พิมพ์ได้

4. ขอบเขตการวิจัย

การค้นคว้าอิสระนี้ใช้เทคนิคการทำเหมืองข้อมูลเพื่อพัฒนาตัวแบบ และเปรียบเทียบผลลัพธ์ที่ได้ระหว่างอัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) เพื่อใช้เป็นแนวทางในการคัดเลือกแม่พิมพ์ ตัวอย่างข้อมูลที่ใช้ในการศึกษานำมาจากฐานข้อมูลของบริษัท อาพิค ยามาดะ (ประเทศไทย) จำกัด โดยมีขอบเขตการศึกษาดังนี้

4.1 คัดเลือกข้อมูลการผลิตและการขายระหว่างปี พ.ศ. 2550-2552 จากฐานข้อมูล Progress 9.1C ของระบบ ERP โดยคัดเลือกข้อมูลที่มีสถานะพร้อมใช้งาน (Active) เท่านั้น

4.2 คัดเลือกข้อมูลการบำรุงรักษาแม่พิมพ์ระหว่างปี พ.ศ. 2550-2552 จากฐานข้อมูล Microsoft SQL Server 2005 ของระบบซ่อมบำรุงรักษาแม่พิมพ์ โดยคัดเลือกข้อมูลที่มีสถานะพร้อมใช้งาน (Active) เท่านั้น

4.3 คัดเลือกดัชนีชี้วัดผลการปฏิบัติงาน (KPI) ที่เป็นดัชนีชี้วัดหลักจากแผนกที่เกี่ยวข้องกับการคัดเลือกแม่พิมพ์จำนวน 3 แผนก คือ แผนกผลิต แผนกขาย และแผนกซ่อมบำรุงแม่พิมพ์ เพื่อใช้เป็นตัวกำหนดคุณลักษณะ (Attribute) ของแต่ละโหนดในการสร้างตัวแบบการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และเพื่อใช้เป็นตัวแปรในการสร้างตัวแบบการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

4.4 พัฒนาโปรแกรมเพื่อใช้ในการคัดเลือกและจัดเตรียมข้อมูล และสร้างระบบฐานข้อมูลเพื่อใช้จัดเก็บข้อมูลที่ถูกจัดเตรียมไว้สำหรับใช้ในการทดสอบอัลกอริทึม

4.5 วิเคราะห์และประเมินความเหมาะสมของตัวแบบ โดยเปรียบเทียบจากค่าความคลาดเคลื่อน (Error Rate) จำนวน 3 วิธี คือ Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) และ Relative Absolute Error (RAE)

4.6 สรุปผลและคัดเลือกอัลกอริทึมที่เหมาะสมกับการคัดเลือกแม่พิมพ์

5. ขั้นตอนการศึกษา

5.1 ศึกษาข้อมูลและเก็บรวบรวมข้อมูลที่ใช้ในงานวิจัย

5.2 ศึกษาเทคนิคการทำเหมืองข้อมูลและงานวิจัยที่เกี่ยวข้อง

5.3 พัฒนาโปรแกรมจัดเตรียมข้อมูล

5.4 สร้างตัวแบบและทดสอบตัวแบบ

5.5 เปรียบเทียบผลลัพธ์และประเมินผลการดำเนินงานวิจัย

5.6 สรุปผลและจัดทำรายงานการค้นคว้าอิสระ

6. เครื่องมือและอุปกรณ์ที่ใช้

6.1 ฮาร์ดแวร์

- CPU : Intel Core Duo 1.6 GHz
- RAM : 1 GB
- Hard disk : 80 GB

6.2 ซอฟต์แวร์

- ระบบปฏิบัติการ : Microsoft Windows XP Professional
- เครื่องมือที่ใช้พัฒนา : Microsoft Visual Basic 6.0
- ฐานข้อมูล : Microsoft SQL Server 2005
- โปรแกรมเหมืองข้อมูล : Weka 3.7.3

7. คำนิยามศัพท์เฉพาะ

7.1 Tool หมายถึง อุปกรณ์ที่ใช้ในการผลิต เพื่อให้ได้ชิ้นงานตามที่ต้องการ

7.2 Die หมายถึง แม่พิมพ์โลหะ

7.3 UPS (Unit per stroke) หมายถึง หน่วยนับจำนวนชิ้นงานที่ได้จากแม่พิมพ์เมื่อทำการปั๊มชิ้นงานในหนึ่งครั้ง

7.4 Progressive Die หมายถึง แม่พิมพ์โลหะแบบต่อเนื่องที่มีการวางตำแหน่งงานหลายตำแหน่งงานและกำหนดการทำงานที่มีความหลากหลายไว้ในแม่พิมพ์หนึ่งตัว การทำงานของแม่พิมพ์ชนิดนี้จะทำงานอย่างต่อเนื่องจนสำเร็จเป็นชิ้นงาน และเป็นแม่พิมพ์ที่มีต้นทุนสูง

7.5 KPI หมายถึง เครื่องมือที่ใช้วัดผลการดำเนินงานหรือประเมินผลการดำเนินงานในด้านต่างๆ ขององค์กร ซึ่งสามารถแสดงผลการวัดหรือการประเมินในรูปเชิงปริมาณ เพื่อสะท้อนประสิทธิภาพ ประสิทธิผลในการปฏิบัติงานขององค์กรหรือหน่วยงานภายในองค์กร

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ศึกษาเอกสาร แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการทำเหมืองข้อมูล โดยเน้นศึกษาเทคนิคการจำแนกข้อมูลด้วยโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) รวมถึงข้อมูลแม่พิมพ์โลหะ ข้อมูล IC Lead Frame และดัชนีชี้วัดผลการปฏิบัติงาน (KPI) หัวข้อสุดท้ายจะกล่าวถึงโปรแกรมวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลที่ใช้ในงานวิจัยครั้งนี้

1. การทำเหมืองข้อมูล

การทำเหมืองข้อมูล หมายถึง การค้นหาความสัมพันธ์และรูปแบบทั้งหมดซึ่งมีอยู่จริงในฐานข้อมูลแต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก โดย Data Mining จะเหมาะกับการแก้ปัญหาบางชนิดเท่านั้น เช่น ปัญหาที่ต้องใช้เหตุผลในการแก้ หรือปัญหาที่เกี่ยวข้องกับเศรษฐศาสตร์และการเงิน หรือปัญหาด้านการศึกษา เป็นต้น Data Mining มีเทคนิคต่างๆ ที่ใช้ในการแก้ปัญหาหลายเทคนิค แต่ไม่มีเทคนิคใดเลยที่สามารถแก้ปัญหาของ Data Mining ได้ทุกปัญหา ดังนั้นความหลากหลายของเทคนิคเป็นสิ่งที่จำเป็นที่จะนำไปสู่วิธีการแก้ปัญหาที่ดีที่สุดของการทำเหมืองข้อมูล (Berry and Linnoff 2004 : 7)

1.1 รูปแบบการทำเหมืองข้อมูล

การทำเหมืองข้อมูลมีได้หลายวิธีการ เช่น วิธีการทางสถิติ วิธีการทางคณิตศาสตร์ ฯลฯ รูปแบบวิธีการที่ได้รับความนิยมแบ่งออกได้หลายลักษณะ เช่น

1.1.1 การอธิบายข้อมูล (Description) เป็นการวิเคราะห์ข้อมูลที่ไม่ซับซ้อนมากนัก มักจะเป็นการอธิบายรูปแบบและแนวโน้มที่มีอยู่ในข้อมูลหรือเป็นการสรุปข้อมูล โดยทั่วไปเทคนิคที่นิยมใช้ เช่น Correlation Analysis, Data Visualization เป็นต้น

1.1.2 การจำแนกกลุ่มข้อมูล (Classification) เป็นการจำแนกกลุ่มข้อมูลที่เป็นประเภทเดียวกัน เพื่อวิเคราะห์คุณสมบัติข้อมูลให้เข้ากลุ่มตามจำนวนกลุ่มที่กำหนดไว้ โดยแบ่งออกเป็นกลุ่มที่ชัดเจน เช่น “ใช่” “ไม่ใช่” หรือ “น้อย” “ปานกลาง” “มาก” เป็นต้น เทคนิคที่ใช้ เช่น Decision Tree, Neural Network เป็นต้น

1.1.3 การรวมกลุ่มข้อมูล (Clustering) เป็นการรวมกลุ่มข้อมูลที่คล้ายคลึงกันไว้ในกลุ่มเดียวกัน โดยไม่มีข้อสมมุติเกี่ยวกับจำนวนกลุ่มที่แน่นอน เช่น การรวมกลุ่มพฤติกรรมการซื้อของลูกค้าที่มีลักษณะคล้ายคลึงกัน เทคนิคที่ใช้ เช่น เทคนิคการรวมกลุ่มแบบลำดับชั้น

1.1.4 การประมาณค่า (Estimation) เป็นการประมาณค่าหรือการคาดคะเนค่าที่สนใจจากข้อมูลที่มีอยู่ เช่น การประเมินราคาน้ำมันในอีก 2 ปีข้างหน้า

1.1.5 การทำนาย (Prediction) จะมีลักษณะคล้ายกับการประมาณค่า แต่ตัวแบบในการทำนายจะมุ่งเน้นการศึกษาพฤติกรรมในอนาคตมากกว่าในปัจจุบัน โดยการนำข้อมูลในอดีตหรือปัจจุบันที่มีอยู่มาสร้างตัวแบบเพื่อทำนายสิ่งที่จะเกิดขึ้นในอนาคต เช่น การทำนายยอดขายของบริษัทในปีหน้า

1.1.6 การหาความสัมพันธ์ของข้อมูล (Association Rule หรือ Affinity Grouping) เป็นการวิเคราะห์หาความสัมพันธ์หรือความเกี่ยวเนื่องกันของข้อมูล โดยอาศัยหลักหรือกฎ ซึ่งจะอยู่ในรูปแบบ “ถ้า *สิ่งใดเกิดขึ้นก่อน* แล้ว *สิ่งใดจะเกิดตามมา* (if *antecedent* then *consequent*)“ (Larose 2005 : 17) เช่น การวิเคราะห์ Market basket analysis

1.2 ขั้นตอนการทำเหมืองข้อมูล

หลักการทำเหมืองข้อมูลตามตัวแบบ CRISP-DM (Cross-Industry Standard Process for Data Mining) แสดงดังภาพที่ 1 ตัวแบบนี้เกิดขึ้นจากความร่วมมือของ DaimlerChrysler, SPSS และ NCR เพื่อพัฒนาตัวแบบการทำเหมืองข้อมูล โดยมีวัตถุประสงค์เพื่อให้การทำเหมืองข้อมูลในอุตสาหกรรมซอร์ฟแวร์เป็นไปในทิศทางเดียวกัน ซึ่งมีลักษณะการทำงานเป็นวงจรชีวิต 6 ระยะด้วยกัน คือ

ระยะที่ 1 การทำความเข้าใจธุรกิจ (Business Understanding Phase) เป็นขั้นตอนแรกที่จะต้องทำความเข้าใจในวัตถุประสงค์และความต้องการของธุรกิจให้ชัดเจน เป็นการเริ่มต้นกำหนดปัญหาที่จะใช้ในเหมืองข้อมูล และเป็นข้อมูลเริ่มต้นเพื่อวางแผนการทำเหมืองข้อมูล ประกอบด้วยกระบวนการย่อยดังนี้

- กำหนดเป้าหมายทางธุรกิจ (Determine Business Objectives)
- ประเมินสถานการณ์ (Assess Situation)
- กำหนดเป้าหมายในการทำเหมืองข้อมูล (Determine Data Mining Goals)
- วางแผนการทำเหมืองข้อมูล (Produce Project Plan)

ระยะที่ 2 การทำความเข้าใจข้อมูล (Data Understanding Phase) เป็นการเก็บรวบรวมข้อมูลและพิจารณาชุดข้อมูลที่จะใช้ในการทำเหมืองข้อมูล รวมถึงแหล่งที่มาและองค์ประกอบของข้อมูล ประกอบด้วยกระบวนการย่อยดังนี้

- รวบรวมข้อมูลเบื้องต้น (Collect Initial Data)
- อธิบายและนิยามข้อมูล (Describe Data)
- สำรวจข้อมูล (Explore Data)
- ตรวจสอบความถูกต้องและความสมบูรณ์ของข้อมูล (Verify Data Quality)

ระยะที่ 3 การเตรียมข้อมูล (Data Preparation Phase) เป็นการคัดเลือกข้อมูลที่จะใช้ในการวิเคราะห์รวมถึงการเปลี่ยนแปลงข้อมูลให้ถูกต้องและพร้อมที่จะนำไปใช้ในการทำเหมืองข้อมูล ประกอบด้วยกระบวนการย่อยดังนี้

- คัดเลือกข้อมูล (Select Data)
- ทำความสะอาดข้อมูล (Clean Data)
- วิเคราะห์ลักษณะข้อมูล (Construct Data)
- รวบรวมข้อมูล (Integrate Data)
- การแปลงข้อมูล (Format Data)

ระยะที่ 4 การสร้างตัวแบบ (Modeling Phase) เป็นการเลือกและประยุกต์ใช้เทคนิคที่จะนำมาทำเหมืองข้อมูล เพื่อให้เหมาะสมกับข้อมูลที่เตรียมไว้ ประกอบด้วยกระบวนการย่อยดังนี้

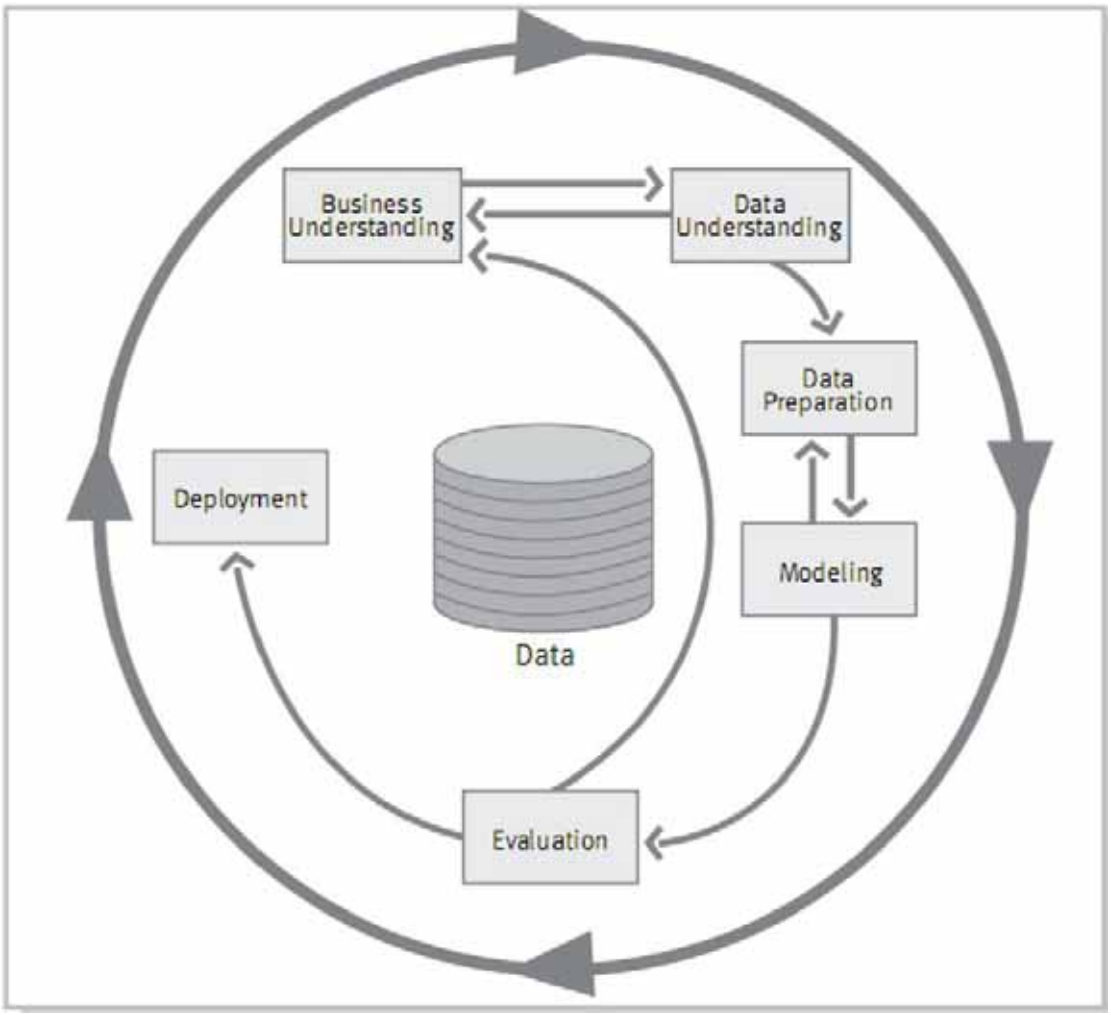
- คัดเลือกเทคนิคที่เหมาะสม (Select Modeling Techniques)
- กำหนดรูปแบบการทดสอบผลลัพธ์ (Generate Test Design)
- สร้างตัวแบบ (Build Model)
- ประเมินตัวแบบ (Assess Model)

ระยะที่ 5 การประเมินผล (Evaluation Phase) เป็นการทดสอบตัวแบบและวิเคราะห์ผลลัพธ์ที่ได้ว่าเป็นไปตามวัตถุประสงค์ที่กำหนดไว้หรือไม่ ซึ่งเป็นขั้นตอนที่สำคัญก่อนนำตัวแบบไปใช้งานจริง หากได้ผลตรงกับวัตถุประสงค์ที่กำหนดก็สามารถนำตัวแบบนี้ไปใช้งานได้ แต่ถ้าทดสอบแล้วผลที่ได้ไม่ตรงกับที่ต้องการจะต้องย้อนกลับไปเริ่มที่กระบวนการแรกใหม่ ประกอบด้วยกระบวนการย่อยดังนี้

- ประเมินผลลัพธ์ (Evaluate Results)
- ตรวจสอบกระบวนการ (Review Process)
- กำหนดแผนการในขั้นต่อไป (Determine Next Steps)

ระยะที่ 6 การนำไปใช้ (Deployment Phase) เป็นการรวบรวมและสรุปความรู้ที่ได้จากตัวแบบที่สร้างขึ้น รวมถึงการวางแผนนำตัวแบบเหมืองข้อมูลไปประยุกต์ใช้งาน ประกอบด้วยกระบวนการย่อยดังนี้

- วางแผนการนำไปใช้งาน (Plan Deployment)
- วางแผนตรวจสอบและบำรุงรักษาตัวแบบ (Plan Monitoring and Maintenance)
- สรุปผลที่ได้จากการใช้ตัวแบบ (Produce Final Report)
- ทบทวนและประเมินผลโครงการ (Review Project)



ภาพที่ 1 ตัวแบบ CRISP-DM 1.0

ที่มา : Daniel T. Larose, Discovering Knowledge in Data an Introduction to Data Mining (New Jersey : John Wiley & Sons Inc, 2005), 6.

2. การสร้างต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเป็นแบบจำลองที่มีลักษณะคล้ายกับต้นไม้ โดยจะสร้างกฎต่างๆขึ้นเพื่อใช้ในการตัดสินใจ ต้นไม้ตัดสินใจเป็นวิธีที่ได้รับความนิยม เนื่องจากความไม่ซับซ้อนของกระบวนการ การเรียนรู้แบบนี้เป็นการเรียนรู้โดยการจำแนก (Classification) ข้อมูลออกเป็นกลุ่ม (Class) ต่างๆ โดยใช้คุณสมบัติ (Attribute) ของข้อมูลในการแยกแยะ ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่าคุณสมบัติใดของข้อมูลที่เป็นตัวกำหนดการจำแนก และคุณสมบัติแต่ละตัวของข้อมูลมีความสำคัญมากน้อยต่างกันอย่างไร ซึ่งเป็นประโยชน์ช่วยให้ผู้ใช้สามารถวิเคราะห์ข้อมูลและตัดสินได้ถูกต้องยิ่งขึ้น (บุญเสริม กิจศิริกุล 2546 : 19-27)

2.1 การแทนต้นไม้ตัดสินใจ

ผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจประกอบด้วย

- โหนดภายใน (Internal node) คือ แอตทริบิวท์ต่างๆของข้อมูล ซึ่งเมื่อข้อมูลใดๆตกลงที่โหนดจะใช้แอตทริบิวท์นี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่า โหนดราก (Root node)

- กิ่ง (Branch link) คือ ค่าแอตทริบิวท์ของโหนด (Node) ภายในที่แตกกิ่งนี้ออกมา ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าแอตทริบิวท์ของโหนดภายในนั้น

- โหนดใบ (Leaf node) คือ กลุ่ม (Class) ต่างๆ ซึ่งเป็นผลลัพธ์ในการจำแนกข้อมูล

2.2 ลักษณะการเรียนรู้ของต้นไม้ตัดสินใจ

- ผลการเรียนรู้แสดงอยู่ในรูปที่เข้าใจง่าย ทำให้ง่ายต่อการวิเคราะห์แอตทริบิวท์ที่มีผลต่อการจำแนกกลุ่มต่างๆ

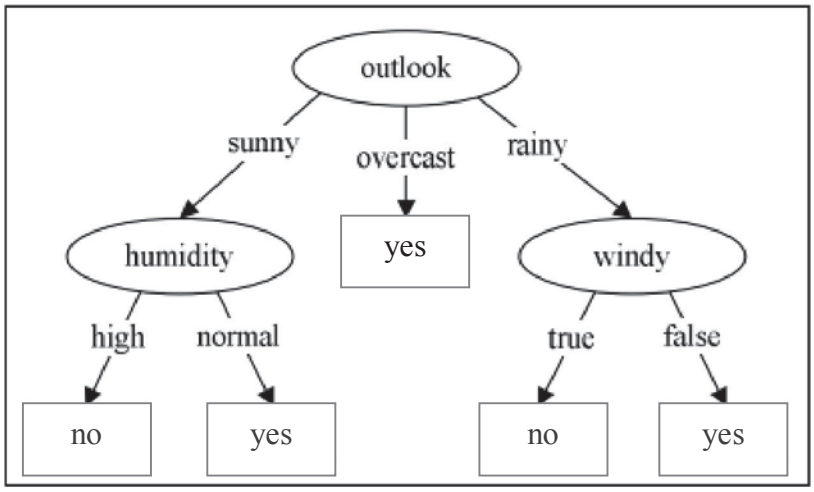
- แต่ละเส้นทางจากโหนดราก (Root node) ถึงโหนดใบ (Leaf node) สามารถแสดงให้อยู่ในรูปกฎ IF-THEN ได้

- มีความทนทานต่อข้อมูลที่มีสัญญาณรบกวน (Noisy data) เช่น แอตทริบิวท์ที่ไม่เกี่ยวข้องและค่าแอตทริบิวท์ที่ผิดพลาดหรือขาดหาย

- การเรียนรู้มีความรวดเร็วเมื่อเทียบกับอัลกอริทึมสำหรับการจำแนกชนิดอื่น

- สามารถนำไปใช้ในการวิเคราะห์ความเสี่ยงของลูกหนี้ การวิเคราะห์กลุ่มดาว การวิเคราะห์งานทางด้านธุรกิจและวิทยาศาสตร์อื่นๆ

ตัวอย่างต้นไม้ตัดสินใจแสดงได้ดังภาพที่ 2 ซึ่งแสดงถึงต้นไม้ที่ใช้ในการตัดสินใจว่าจะออกไปเล่นกอล์ฟหรือไม่ (Quinlan 1986) โดยพิจารณาจากสภาพอากาศต่างๆ ประกอบการตัดสินใจ โดยโหนดที่แสดงในรูปวงรีแสดงถึงการทดสอบค่าที่เป็นไปได้ของแอตทริบิวท์นั้นๆ และใบที่แสดงในรูปสี่เหลี่ยมจะแสดงผลการจำแนกกลุ่ม ซึ่งเป็นผลลัพธ์จากการทำนายตามเส้นทางของต้นไม้ตัดสินใจว่าจะออกไปเล่นกอล์ฟ (Yes) หรือไม่ออกไปเล่นกอล์ฟ (No)



ภาพที่ 2 ต้นไม้ตัดสินใจที่ใช้ในการตัดสินใจของการออกไปเล่นกอล์ฟ
ที่มา : Ian H. Witten and Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques (San Francisco : Morgan Kaufmann Publishers, 2005), 101.

โดยหลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ จะเป็นการสร้างในลักษณะจากบนลงล่าง (Top-down) กล่าวคือ จะเริ่มสร้างจากราก (Root node) ของต้นไม้ก่อน จากนั้นจึงทำการแตกกิ่งไปจนถึงโหนดใบ (Leaf node) โดยจะแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้

1. ต้นไม้จะเริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึก (Training set)
2. ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อตามกลุ่มของข้อมูลนั้น
3. ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ ให้ทำการวัดค่าเกน (Gain) ของแอตทริบิวท์แต่ละตัวเพื่อใช้เป็นเกณฑ์ในการคัดเลือกแอตทริบิวท์ที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุด โดยแอตทริบิวท์ที่มีค่าเกนมากที่สุดจะถูกเลือกเป็นตัวทดสอบหรือแอตทริบิวท์ในการตัดสินใจ และแสดงในรูปของโหนดบนต้นไม้

4. กิ่งของต้นไม้จะถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่างๆ ที่สร้างขึ้น

5. ทำการวนซ้ำเพื่อหาแอตทริบิวท์ที่มีค่าเกนมากที่สุดจากข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำแอตทริบิวท์นี้มาสร้างเป็นโหนดตัดสินใจตัวต่อไป แอตทริบิวท์ที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาเป็นโหนดในระดับต่อไปอีก

6. ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งออกไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้เป็นจริง

- ถ้าข้อมูลทุกตัวในโหนดนั้นอยู่ในกลุ่มเดียวกันให้สร้างใบตามกลุ่มของข้อมูลนั้น
- ถ้าไม่เหลือแอตทริบิวท์ใดสำหรับใช้ในการแบ่งข้อมูลแล้ว ซึ่งในกรณีนี้จะใช้กลุ่มที่มีข้อมูลสนับสนุนมากที่สุดเป็นใบ
- ถ้าไม่มีข้อมูลสนับสนุนสำหรับกิ่งนั้นๆ แล้ว ให้สร้างใบตามกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

ในขั้นตอนการสร้างต้นไม้ตัดสินใจ อัลกอริทึม C4.5 เป็นอัลกอริทึมที่มีชื่อเสียงและเป็นที่รู้จักกันอย่างแพร่หลาย ซึ่งพัฒนาโดย Quinlan (1993) ที่ได้พัฒนาต่อมาจากอัลกอริทึม ID3 ที่เขาได้พัฒนาขึ้น (Quinlan 1986) โดยเป็นวิธีการเรียนรู้จากกลุ่มตัวอย่างที่เรียกว่า ชุดข้อมูลฝึก (Training set) ที่อาศัยวิธีการจัดหมวดหมู่เพื่อสร้างต้นไม้ตัดสินใจ

ชุดข้อมูลฝึกจะมีลักษณะคล้ายกับข้อมูลในฐานข้อมูลเชิงสัมพันธ์ (Relational database) แสดงในรูปของตารางที่ประกอบไปด้วย “แถว” ซึ่งแสดงข้อมูลหรือตัวอย่าง และ “คอลัมน์” ซึ่งแสดงแอตทริบิวท์ของข้อมูล โดยแบ่งออกเป็น 2 ชนิด คือ

1. แอตทริบิวท์เป้าหมาย (Target attribute) เป็นแอตทริบิวท์ที่กำหนดว่าตัวอย่างนั้นๆ จะถูกจำแนกอยู่ในกลุ่มใด ข้อมูลแต่ละชุดจะมีเพียงแอตทริบิวท์เดียวและจะเป็นชนิดข้อความเท่านั้น

2. แอตทริบิวท์ประกอบการทำนาย (Predicting attribute) เป็นแอตทริบิวท์ที่บอกถึงคุณสมบัติต่างๆ ของตัวอย่างแต่ละตัว โดยแต่ละแอตทริบิวท์อาจมีข้อมูลเป็นชนิดข้อความหรือตัวเลขก็ได้

ตารางที่ 1 ชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจเล่นกอล์ฟ

ID	Attribute				Class
	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	hot	high	false	yes
5	rainy	mild	high	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

จากตารางที่ 1 เป็นตัวอย่างชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจออกไปเล่นกอล์ฟ โดยพิจารณาจากสภาพอากาศต่างๆ ประกอบการตัดสินใจ (Quinlan 1986) เมื่อนำมาสร้างเป็นต้นไม้ตัดสินใจสามารถแสดงโครงสร้างต้นไม้ตัดสินใจดังภาพที่ 2 ชุดข้อมูลฝึกนี้มีแอตทริบิวท์ Class เป็นแอตทริบิวท์เป้าหมาย โดยมีค่าที่เป็นไปได้ 2 ค่า คือ yes (เล่น) หรือ no (ไม่เล่น) สำหรับแอตทริบิวท์ outlook, temperature, humidity และ windy จะเป็นแอตทริบิวท์ประกอบการทำนาย

ประสิทธิภาพของต้นไม้ตัดสินใจไม่ได้อยู่ที่การสร้างต้นไม้ตัดสินใจ เพื่อให้จำแนกกลุ่มชุดข้อมูลฝึกได้อย่างถูกต้องเท่านั้น แต่ต้องสามารถจำแนกกลุ่มข้อมูลตัวอย่างใหม่ๆ ที่นอกเหนือจากชุดข้อมูลฝึกได้อย่างถูกต้องด้วย ดังนั้นการสร้างต้นไม้ตัดสินใจจึงควรจะมีชุดข้อมูลทดสอบ (Testing set) ที่ใช้ตรวจสอบความถูกต้องของต้นไม้ตัดสินใจที่ได้ด้วย

2.3 การคัดเลือกแอตทริบิวท์เพื่อจำแนกกลุ่มข้อมูล

ในการสร้างต้นไม้ตัดสินใจ ปัญหาสำคัญที่ต้องพิจารณา คือ การตัดสินใจเลือกแอตทริบิวท์ใดมาทำหน้าที่เป็นโหนดรากและในโหนดต่อๆ ไปเพื่อสร้างเป็นต้นไม้ย่อย (Subtree) ในลำดับถัดไป เกณฑ์ในการคัดเลือกแอตทริบิวท์ คือ การคำนวณค่าเกน (Gain) ซึ่งเป็นค่าที่บอกว่าแอตทริบิวท์นั้นจะสามารถจำแนกกลุ่มข้อมูลได้ดีเพียงใด โดยทดลองเลือกแต่ละแอตทริบิวท์ที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าแอตทริบิวท์ใดให้ค่าเกนสูงที่สุด แสดงว่าแอตทริบิวท์นั้นสามารถจำแนกกลุ่มข้อมูลได้ดีที่สุด หรือเป็นแอตทริบิวท์ที่จำแนกกลุ่มข้อมูลแล้วได้ข้อมูลในแต่ละใบของต้นไม้ตัดสินใจเป็นกลุ่มเดียวกันทั้งหมด หรือมีข้อมูลต่างกลุ่มปะปนมาบ้างเพียงเล็กน้อยเท่านั้น โดยค่าเกนสำหรับการคัดเลือกแอตทริบิวท์ที่สำคัญมีดังนี้

2.3.1 ค่ามาตรฐานเกน (Gain criterion)

วิธีการสร้างต้นไม้ตัดสินใจโดยใช้อัลกอริทึม ID3 จะใช้ค่ามาตรฐานเกนในการตัดสินใจเลือกแอตทริบิวท์ที่จะใช้เป็นโหนดรากของต้นไม้หรือของต้นไม้ย่อย โดยการคำนวณค่าเกนของแต่ละแอตทริบิวท์เพื่อใช้แบ่งกลุ่มตัวอย่าง และเลือกแอตทริบิวท์ที่มีค่าเกนสูงที่สุดมาเป็นโหนดราก ซึ่งแอตทริบิวท์นี้จะมีความสามารถในการจำแนกกลุ่มข้อมูลสูง การคัดเลือกแอตทริบิวท์นี้ทำให้สามารถแบ่งข้อมูลออกมาโดยมีการปะปนของกลุ่มข้อมูลที่ต่างกันเกิดขึ้นน้อยอีกด้วย ค่าเกนนี้คำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ (Information theory) ซึ่งมีสาระสำคัญ คือ ค่าสารสนเทศของข้อมูลจะขึ้นอยู่กับความน่าจะเป็นของข้อมูล และสามารถวัดอยู่ในรูปของบิต (bits) โดยเขียนสมการได้ดังนี้

$$\text{ค่าสารสนเทศของข้อมูล} = -\log_2(\text{ความน่าจะเป็นของข้อมูล}) \qquad (2\text{-}1)$$

การใช้ค่า Information gain จะช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูล อีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจที่ได้จะไม่มีความซับซ้อนมากจนเกินไป ซึ่งค่า Information gain นั้นสามารถคำนวณได้จากสมการดังต่อไปนี้ (Han and Kamber 2001)

โดยกำหนดให้

S เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s ระเบียน

m เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

C_i แทนกลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง m

S_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i

s_{ij} แทนจำนวนข้อมูลที่เป็นสมาชิกของ S ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นได้ j ของแอตทริบิวท์ A โดยที่ j มีค่าระหว่าง 1 ถึง v

s_i / s แทนค่าความน่าจะเป็นที่ข้อมูลจะอยู่ในกลุ่ม C_i

ค่า Information gain ที่ต้องการสำหรับจำแนกข้อมูลออกเป็นแต่ละกลุ่มหาได้โดย

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} \frac{s_i}{S} \log_2 \frac{s_i}{S} \quad (2\text{-}2)$$

ค่า Entropy ของแอตทริบิวท์ A ซึ่งมีค่าของแอตทริบิวท์เป็น $(a_1, a_2, a_3, ..., a_v)$ หาได้โดย

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j}, ..., s_{mj}) \quad (2\text{-}3)$$

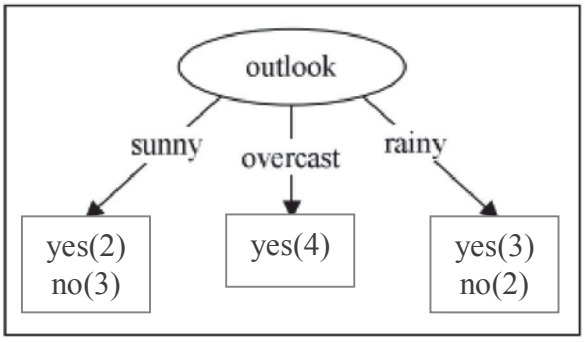
ค่ามาตรฐานเกนที่จะใช้มาพิจารณาในการเลือกแอตทริบิวท์ A มาเป็นโหนดของต้นไม้มีค่าเท่ากับปริมาณข้อมูลที่ต้องการเพื่อให้สามารถจำแนกกลุ่มของข้อมูลได้ ลบด้วยปริมาณข้อมูลที่ต้องการเพื่อจำแนกกลุ่มของข้อมูลโดยใช้แอตทริบิวท์ A เป็นตัวตรวจสอบเพื่อจำแนกกลุ่มของข้อมูล สามารถเขียนเป็นสมการได้ดังนี้

$$Gain(A) - I(s_1, s_2, ..s_m) - E(A) \quad (2\text{-}4)$$

จากตัวอย่างข้อมูลสภาพอากาศประกอบการตัดสินใจในการเล่นกอล์ฟในตารางที่ 1 เซตของข้อมูลฝึก T ประกอบด้วยข้อมูลจำนวน 14 ระเบียน แบ่งข้อมูลออกเป็น 2 กลุ่ม คือ ข้อมูลที่ตัดสินใจออกไปเล่นกอล์ฟ (Class = yes) จำนวน 9 ระเบียน และตัดสินใจไม่ออกไปเล่นกอล์ฟ (Class = no) จำนวน 5 ระเบียน การจะระบุว่าข้อมูลหนึ่งระเบียนอยู่ในกลุ่ม yes หรือ no ต้องการปริมาณข้อมูลประกอบการตัดสินใจ เพื่อจำแนกกลุ่มโดยใช้สมการที่ 2-2 ดังนี้

$$I(T) = -(9/14)x \log_2(9/14) - (5/14)x \log_2(5/14)$$
$$= 0.940 \text{ บิต}$$

การจะจำแนกกลุ่มของข้อมูลเพื่อตัดสินใจออกไปเล่นกอล์ฟหรือไม่นั้น ต้องใช้ข้อมูลจากแอตทริบิวท์อื่นประกอบการตัดสินใจด้วย ถ้าแบ่งข้อมูลชุดนี้โดยใช้แอตทริบิวท์ outlook จะสามารถจำแนกกลุ่มของข้อมูลได้ดังภาพที่ 3 โดยจะแสดงจำนวนระเบียนของแต่ละกลุ่มข้อมูลไว้ในวงเล็บ เมื่อแบ่งตามค่าที่เป็นไปได้จะต้องการปริมาณข้อมูลเพิ่มเพื่อประกอบการเลือกกลุ่ม และสามารถคำนวณค่า Entropy ของแอตทริบิวท์ได้โดยสมการที่ 2-3



ภาพที่ 3 แสดงการจำแนกกลุ่มข้อมูลโดยใช้แอตทริบิวท์ outlook

$$\begin{aligned} E(outlook) = {} & (5/14)x(-(2/5)x\log_2(2/5) - (3/5)x\log_2(3/5)) + \\ & (4/14)x(-(4/4)x\log_2(4/4) - (0/4)x\log_2(0/4)) + \\ & (5/14)x(-(3/5)x\log_2(3/5) - (2/5)x\log_2(2/5)) \\ = {} & 0.693 \text{ บิต} \end{aligned}$$

นั่นก็คือถ้าต้องการจำแนกกลุ่มของข้อมูลใหม่ โดยใช้แอตทริบิวท์ outlook เป็นตัวตรวจสอบเพื่อจำแนกกลุ่มของข้อมูล การพิจารณาจากค่า outlook ของข้อมูลใหม่นี้จะต้องใช้ข้อมูลเพิ่มอีก 0.693 บิต จึงจะบอกกลุ่มที่ถูกต้องของข้อมูลใหม่นี้ได้ ดังนั้นสามารถคำนวณค่าเกนจากการเลือกแอตทริบิวท์ outlook เป็นแอตทริบิวท์เพื่อใช้แบ่งข้อมูลได้จากสมการที่ 2-4 ดังนี้

$$\begin{aligned} Gain(outlook) &= I(T) - E(outlook) \\ &= 0.940 - 0.693 \\ &= 0.247 \text{ บิต} \end{aligned}$$

แอตทริบิวท์ที่เหลือที่สามารถถูกเลือกมาเป็นตัวทดสอบเพื่อจัดกลุ่มของข้อมูลฝึก คือ แอตทริบิวท์ temperature, humidity และ windy สามารถนำมาคำนวณค่าเกนของแต่ละแอตทริบิวท์ได้ดังนี้

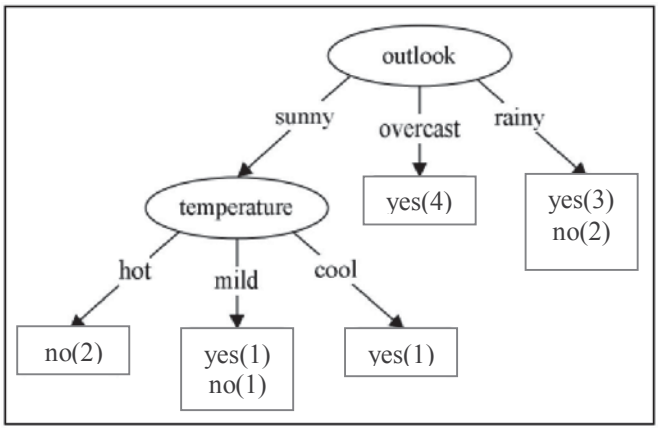
$$Gain(temperature) = I(T) - E(temperature)$$
$$= 0.940 - 0.911$$
$$= 0.029 \text{ บิต}$$

$$Gain(humidity) = I(T) - E(humidity)$$
$$= 0.940 - 0.788$$
$$= 0.152 \text{ บิต}$$

$$Gain(windy) = I(T) - E(windy)$$
$$= 0.940 - 0.892$$
$$= 0.048 \text{ บิต}$$

จะเห็นว่าแอตทริบิวท์ที่ให้ค่าเกนสูงที่สุด คือ outlook ดังนั้นแอตทริบิวท์ outlook จึงถูกเลือกเป็นโหนดรากของต้นไม้ตัดสินใจ แต่เนื่องจากยังไม่สามารถจัดกลุ่มข้อมูลให้เป็นกลุ่มเดียวกันทั้งหมดจึงต้องสร้างต้นไม้ตัดสินใจต่อไป โดยพิจารณาเลือกแอตทริบิวท์ที่จะมาเป็นโหนดในระดับที่ 2 ต่อจากโหนดราก ในกรณี outlook = overcast ไม่ต้องสร้างโหนดเพิ่มเติม เนื่องจากสามารถจัดกลุ่มข้อมูลที่เป็นกลุ่ม yes ได้ทั้งหมดแล้ว (ข้อมูลในกลุ่มเป็น yes ทุกตัว)

แอตทริบิวท์ที่สามารถถูกเลือกเป็นโหนดในระดับที่ 2 ประกอบด้วย temperature, humidity และ windy โดยแอตทริบิวท์ outlook จะไม่ถูกเลือกในระดับถัดไปอีกแล้ว เมื่อพิจารณาการสร้างโหนดลูกทางด้านซ้ายมือ (outlook = sunny) ถ้าเลือกแอตทริบิวท์ temperature จะสามารถจำแนกกลุ่มข้อมูลได้ดังภาพที่ 4 และสามารถคำนวณค่าเกนได้ดังนี้



ภาพที่ 4 แสดงการจำแนกกลุ่มข้อมูลโดยใช้แอตทริบิวท์ temperature เป็นโหนดระดับที่ 2

$$I(outlook = sunny) = -(2/5)x\log_2(2/5) - (3/5)x\log_2(3/5)$$
$$= 0.971 \text{ บิต}$$

$$E_{temperature}(outlook = sunny) = (2/5)x(-(0/2)x\log_2(0/2) - (2/2)x\log_2(2/2)) +$$
$$(2/5)x(-(1/2)x\log_2(1/2) - (1/2)x\log_2(1/2)) +$$
$$(1/5)x(-(1/1)x\log_2(1/1) - (0/1)x\log_2(0/1))$$
$$= 0.4 \text{ บิต}$$

$$Gain(temperature) = I(outlook = sunny) - E_{temperature}(outlook = sunny)$$
$$= 0.971 - 0.4$$
$$= 0.571 \text{ บิต}$$

แอตทริบิวท์ที่เหลือที่สามารถถูกเลือกมาเป็นตัวทดสอบเพื่อจัดกลุ่มของข้อมูลฝึก คือ แอตทริบิวท์ humidity และ windy สามารถนำมาคำนวณค่าเกนของแต่ละแอตทริบิวท์ได้ดังนี้

$$Gain(humidity) = I(outlook = sunny) - E_{humidity}(outlook = sunny)$$
$$= 0.971 - 0$$
$$= 0.971 \text{ บิต}$$

$$Gain(windy) = I(outlook = sunny) - E_{windy}(outlook = sunny)$$
$$= 0.971 - 0.951$$
$$= 0.020 \text{ บิต}$$

จะเห็นว่าแอตทริบิวท์ที่ให้ค่าเกนสูงที่สุด คือ humidity ดังนั้นแอตทริบิวท์นี้จึงถูกเลือกเป็นโหนดระดับที่ 2 ต่อจาก outlook = sunny และยังคงเหลือโหนดลูกทางด้านขวาของโหนด outlook (outlook = rainy) ที่ต้องพิจารณาอีก และจากวิธีการคำนวณค่าเกนที่อธิบายในข้างต้น สามารถเลือกได้ว่าแอตทริบิวท์ windy จะให้ค่าเกนสูงที่สุด ดังนั้นจึงถูกเลือกเป็นโหนดระดับที่ 2 ต่อจาก outlook = rainy กระบวนการสร้างต้นไม้ตัดสินใจจะสิ้นสุดเมื่อโหนดใบเป็นกลุ่มข้อมูลเดียวกันทั้งหมด และจะได้โครงสร้างต้นไม้ตัดสินใจดังภาพที่ 2

2.3.2 ค่ามาตรฐานอัตราส่วนเกน (Gain ratio criterion)

ในอัลกอริทึม ID3 จะใช้ค่ามาตรฐานเกนเป็นหลักในการเลือกแอตทริบิวท์ที่จะใช้เป็นโหนดรากของต้นไม้ตัดสินใจหรือของต้นไม้ย่อย แต่ในอัลกอริทึม C4.5 ได้เพิ่มการใช้ค่ามาตรฐานอัตราส่วนเกนในการตัดสินใจเลือกแอตทริบิวท์ที่จะใช้เป็นโหนดรากเข้ามาด้วย เนื่องจากค่ามาตรฐานเกนจะมีความลำเอียงอย่างมากกับข้อมูลที่ประกอบด้วยแอตทริบิวท์ที่มีค่าที่เป็นไปได้จำนวนมากๆ เช่น ชุดข้อมูลที่ประกอบด้วยแอตทริบิวท์หมายเลขประจำตัว ซึ่งมีค่าไม่ซ้ำกันในแต่ละตัวอย่าง ถ้าแบ่งข้อมูลตามแอตทริบิวท์นี้จะทำให้ได้จำนวนตัวอย่างเพียง 1 ตัวอย่างต่อ 1 กิ่งของต้นไม้ และเมื่อคำนวณค่า Entropy จากการแบ่งตัวอย่างบนแอตทริบิวท์นี้จะได้เท่ากับศูนย์ ทำให้ค่าเกนที่ได้ของแอตทริบิวท์นี้มีค่าสูงที่สุด (ก้องศักดิ์ จงเกษมวงศ์ 2543)

จากตัวอย่างข้อมูลการตัดสินใจเล่นกอล์ฟในตารางที่ 1 ถ้าใช้แอตทริบิวท์ ID ในการจัดกลุ่มข้อมูลจะต้องการปริมาณข้อมูลประกอบการตัดสินใจเพื่อจำแนกกลุ่มดังนี้

$$\begin{aligned} E(ID) &= (1/14)x(-(0/1)x\log_2(0/1) - (1/1)x\log_2(1/1)) + \ldots + \\ &\quad (1/14)x(-(0/1)x\log_2(0/1) - (1/1)x\log_2(1/1)) \\ &= 0 \text{ บิต} \end{aligned}$$

เมื่อแบ่งตัวอย่างบนแอตทริบิวท์นี้จะได้ค่า Entropy เท่ากับศูนย์ ดังนั้นค่ามาตรฐานเกนของแอตทริบิวท์นี้จะเท่ากับปริมาณข้อมูลที่ต้องการระบุว่าข้อมูลหนึ่งระเบียนอยู่ในกลุ่ม yes หรือ no ที่โหนดราก ซึ่งมีค่าเท่ากับ 0.940 บิต ทำให้ค่ามาตรฐานเกนนี้มีค่าสูงกว่าแอตทริบิวท์อื่นๆ ดังนั้นแอตทริบิวท์ ID นี้จะถูกเลือกมาเป็นตัวทดสอบเพื่อจัดกลุ่มข้อมูลฝึก

ดังนั้นจะเห็นว่า การวัดค่ามาตรฐานเกนจะได้ค่าสูงเมื่อแอตทริบิวท์นั้นมีค่าที่เป็นไปได้จำนวนมากๆ ซึ่งไม่สามารถนำมาใช้เป็นโหนดของต้นไม้เพื่อทำนายกลุ่มของข้อมูลใหม่ที่ไม่เคยเห็นได้อย่างถูกต้อง จึงต้องแก้ไขความลำเอียงนี้โดยการปรับค่าเกนให้ถูกต้อง ด้วยการใช้ค่าสารสนเทศการแบ่งแยก (Split information) ของแต่ละแอตทริบิวท์เพื่อใช้คำนวณค่ามาตรฐานอัตราส่วนเกน (Witten and Frank 2005)

ถ้ากำหนดให้ T แทนชุดข้อมูลฝึก เมื่อแบ่งตัวอย่างโดยใช้แอตทริบิวท์ A จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่งเป็น $\{t_1, t_2, \ldots, t_v\}$ จำนวน v ชุด ตามค่าที่เป็นไปได้ของแอตทริบิวท์ A และสามารถคำนวณค่าสารสนเทศการแบ่งแยกได้ดังนี้

$$\text{ค่าสารสนเทศการแบ่งแยก} = -\sum_{i=1}^{v}\frac{|t_i|}{|T|}\log_2\left(\frac{|t_i|}{|T|}\right) \qquad (2\text{-}5)$$

ค่าสารสนเทศการแบ่งแยกนี้จะแสดงถึงระดับการกระจายของข้อมูล เมื่อแบ่งข้อมูลตัวอย่าง T เป็น v ชุดย่อยตามค่าที่เป็นไปได้ของแอตทริบิวท์ A โดยค่านี้จะมีค่าสูงสุดเมื่อ $|t_i|$ เป็น 1 เท่ากันในทุกกิ่ง และจะลดลงเมื่อค่า $|t_i|$ เพิ่มขึ้น เมื่อนำค่านี้ไปหารค่ามาตรฐานเกนจะได้ค่ามาตรฐานอัตราส่วนเกน ซึ่งช่วยแก้ไขความลำเอียงที่เกิดขึ้นของค่ามาตรฐานเกนได้ โดยทำให้ค่ามาตรฐานอัตราส่วนเกนของแอตทริบิวท์ที่มีค่าที่เป็นไปได้จำนวนมากถูกปรับลดลง (ก้องศักดิ์ จงเกษมวงศ์ 2543)

ค่ามาตรฐานอัตราส่วนเกน = ค่ามาตรฐานเกน / ค่าสารสนเทศการแบ่งแยก

จากตัวอย่างข้อมูลการตัดสินใจเล่นกอล์ฟในตารางที่ 1 สามารถคำนวณค่า Gain ratio ของแอตทริบิวท์ outlook ได้ดังนี้

$$\begin{aligned}\text{ค่าสารสนเทศการแบ่งแยก}(outlook) &= -(5/14)x\log_2(5/14)\\ &\quad -(4/14)x\log_2(4/14)\\ &\quad -(5/14)x\log_2(5/14)\\ &= 1.577 \text{ บิต}\end{aligned}$$

$$\begin{aligned}Gain\ ratio(outlook) &= 0.247/1.577\\ &= 0.156\end{aligned}$$

เมื่อแบ่งข้อมูลตัวอย่างด้วยแอตทริบิวท์ temperature, humidity และ windy สามารถคำนวณค่า Gain ratio ได้ดังนี้

$$\begin{aligned}Gain\ ratio(temperature) &= 0.029/1.362\\ &= 0.021\end{aligned}$$

$$Gain\ ratio(humidity) = 0.152/1.000$$
$$= 0.152$$

$$Gain\ ratio(windy) = 0.048/0.985$$
$$= 0.049$$

จะเห็นว่าแอตทริบิวท์ที่ให้ค่า Gain ratio สูงสุด คือ outlook เช่นเดียวกับการคำนวณค่า Information gain ดังนั้นแอตทริบิวท์ outlook จึงถูกเลือกเป็นโหนดรากของต้นไม้ตัดสินใจ และจะสร้างต้นไม้ตัดสินใจต่อไปจนกระทั่งสามารถจัดกลุ่มข้อมูลให้เป็นกลุ่มเดียวกันได้ทั้งหมด

2.4 งานวิจัยที่เกี่ยวข้องกับการสร้างต้นไม้ตัดสินใจ

เกรียงไกร พิพิธหิรัญการ (2550) เสนอการเปรียบเทียบประสิทธิภาพความแม่นยำในการทำนายข้อมูลการตัดสินใจซื้อหนังสือของผู้ใช้บริการศูนย์หนังสือมหาวิทยาลัยเกษตรศาสตร์ โดยทำการเปรียบเทียบความแม่นยำจากการทำนายระหว่างเทคนิคโครงสร้างต้นไม้ตัดสินใจจากอัลกอริทึม C4.5 และเทคนิคกฎความสัมพันธ์สำหรับจำแนก (Associative Classification Rule) จากอัลกอริทึม Apriori ผลการทดลองพบว่าเทคนิคโครงสร้างต้นไม้ตัดสินใจสามารถทำนายการตัดสินใจซื้อหนังสือได้ถูกต้องถึง 92% ส่วนเทคนิคแบบกฎความสัมพันธ์สำหรับจำแนกสามารถทำนายการตัดสินใจซื้อหนังสือได้ถูกต้องเพียง 61% และเมื่อพิจารณาค่า TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative), Precision และค่า Recall จะเห็นว่าเทคนิคโครงสร้างต้นไม้ตัดสินใจมีค่าทางสถิติดีกว่าในทุกด้าน ดังนั้นสามารถสรุปได้ว่าเทคนิคโครงสร้างต้นไม้ตัดสินใจมีความแม่นยำ (Accuracy) ในการทำนายมากกว่าเทคนิคแบบกฎความสัมพันธ์สำหรับจำแนก

มลธิดา ฤทธิ์สมบูรณ์ และ สุชา สมานชาติ (2551 : 8-14) ได้ประยุกต์ใช้เทคนิคการแบ่งกลุ่มแบบโครงสร้างต้นไม้ตัดสินใจและใช้อัลกอริทึมในการเรียนรู้แบบ ID3 โดยใช้โปรแกรม Weka ในการทดลองเพื่อสร้างตัวแบบโครงสร้างต้นไม้สำหรับการตัดสินใจ เพื่อสร้างระบบสนับสนุนการพิจารณาอนุมัติให้สินเชื่อเพื่อการเช่าซื้อสินค้า โดยนำข้อมูลของลูกค้าในอดีตที่ได้ยื่นใบสมัครขอสินเชื่อเพื่อการเช่าซื้อสินค้า 14,332 รายการ โดยแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลสำหรับการสอน (Training data) จำนวน 90% หรือเท่ากับ 12,900 รายการ และข้อมูลสำหรับการทดสอบ (Testing data) จำนวน 10% หรือเท่ากับ 1,530 รายการ โดยมีการสลับตำแหน่งของรายการเพื่อให้เกิดการกระจายตัวของข้อมูลเพื่อเพิ่มประสิทธิภาพการเรียนรู้ พบว่าตัวแบบที่สร้างขึ้นมีค่าเฉลี่ยเปอร์เซ็นต์ความถูกต้องเท่ากับ 67.78% ปัญหาที่พบในการทดลองนี้ คือ ต้องใช้เวลานานใน

การจัดเก็บรวบรวมและจัดเตรียมข้อมูลใบสมัครสินเชื่อเพื่อใช้ในการสร้างตัวแบบ เนื่องจากข้อมูลที่ใช้ให้ตัวแบบเรียนรู้นั้นจะต้องอยู่ในรูปแบบที่ถูกต้องตามที่โปรแกรม Weka 3.4 ต้องการ หากมีระเบียนใดผิดพลาดแม้เพียงระเบียนเดียว โปรแกรมจะไม่สามารถสร้างตัวแบบออกมาได้

ณัฐมณฑ์ สิริวัฒนานันท์ (2551) เสนอการสร้างแบบจำลองและพัฒนาระบบตรวจสอบความเหมาะสมในการขนส่งสินค้าด้วยเทคนิคโครงสร้างต้นไม้ตัดสินใจจากอัลกอริทึม C4.5 โดยมีวัตถุประสงค์เพื่อลดต้นทุนค่าขนส่งสินค้าให้ต่ำที่สุด ข้อมูลกลุ่มตัวอย่างที่ใช้ในการศึกษา คือ ข้อมูลการขนส่งในเขตกรุงเทพฯและปริมณฑลปี 2551 จำนวน 11,089 รายการ โดยแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลสำหรับเรียนรู้ (Training data) จำนวน 8,316 รายการ และข้อมูลสำหรับทดสอบ (Testing data) จำนวน 2,773 รายการ ผลจากการทดสอบแบบจำลองด้วยโปรแกรม Weka ให้ค่าความถูกต้อง 97.23% จากนั้นพัฒนาระบบตรวจสอบความเหมาะสมในการขนส่งสินค้าด้วยโปรแกรม VB.NET 2003 และประเมินผลการทำงานของระบบจากแบบสอบถามความพึงพอใจโดยผู้ใช้งานระบบ ผลการประเมินโดยรวมเฉลี่ยเท่ากับ 4.67 ซึ่งแสดงว่าประสิทธิภาพของระบบอยู่ในระดับดี

อัญชลิสา แต้ตระกูล (2552) นำเสนอการออกแบบอัลกอริทึมสำหรับวิเคราะห์ความผิดพลาดในการผลิตฮาร์ดดิสก์ด้วยเทคนิคโครงสร้างต้นไม้ตัดสินใจแบบขนานด้วยอัลกอริทึม C4.5 โดยมุ่งเน้นในการผลิตชิ้นส่วน Head Gimbal Assembly (HGA) และรวมไปถึงข้อมูลในการผลิต WAFER และ SLIDER ซึ่งเป็นส่วนผลิตก่อนหน้าที่จะส่งผลต่อการผลิตชิ้นส่วน HGA ในการทดลองได้ทำการแบ่งข้อมูลเรียนรู้ (Training data) จำนวน 8,000 ระเบียน และข้อมูลสำหรับทดสอบ (Testing data) จำนวน 2,000 ระเบียน การทดลองจะทำการวิเคราะห์พารามิเตอร์ที่ก่อให้เกิดปัญหาในการผลิตและทำให้ผลผลิตตกต่ำ ผลการทดลองเบื้องต้นพบว่าอัลกอริทึม C4.5 ใช้เวลาในการคำนวณสูง ดังนั้นจึงได้พัฒนาโปรแกรมจากภาษา C เพื่อทำการทดลองสร้างต้นไม้ตัดสินใจแบบขนานใน 2 วิธีการ คือ Synchronous Tree Construction (STC) และ Partition Tree Construction (PTC) เพื่อทดสอบประสิทธิภาพระหว่างการสร้างต้นไม้ตัดสินใจแบบเดิมกับแบบขนาน โดยทำการสุ่มจำนวนระเบียนที่ใช้ในการประมวลผลอัลกอริทึม เช่น 2,000 ระเบียน 4,000 ระเบียน 6,000 ระเบียน และ 8,000 ระเบียน จากนั้นวัดประสิทธิภาพของอัลกอริทึมจากค่าการใช้ทรัพยากรคอมพิวเตอร์และระยะเวลาที่ใช้ในการประมวลผล จากผลการประเมินพบว่าต้นไม้ตัดสินใจแบบขนานด้วยวิธีการ Synchronous Tree Construction (STC) มีประสิทธิภาพดีที่สุด โดยสามารถลดเวลาในการประมวลลงได้ประมาณ 300% เมื่อทำงานแบบขนานบนเครื่องคอมพิวเตอร์ที่ใช้หน่วยประมวลผล (CPU) จำนวน 4 หน่วย (Processors)

3. การรวมกลุ่มข้อมูล (Cluster Analysis)

กัลยา วานิชย์บัญชา (2546 : 125) กล่าวว่า Cluster Analysis เป็นเทคนิคที่ใช้จำแนกหรือแบ่งเป็นกรณี (Case) เช่น คน สัตว์ สิ่งของ หรือองค์กร ฯลฯ หรือแบ่งตัวแปรออกเป็นกลุ่มย่อยๆ ตั้งแต่ 2 กลุ่มขึ้นไป กรณีที่อยู่กลุ่มเดียวกันจะมีลักษณะที่เหมือนกันหรือคล้ายกัน ส่วนกรณีที่อยู่ต่างกลุ่มกันจะมีลักษณะที่แตกต่างกัน ดังนั้นการพิจารณาเลือกลักษณะตัวแปรที่จะนำมาใช้แบ่งกลุ่มกรณีจึงมีความสำคัญ

โดยหลักการแล้วเทคนิคเหล่านี้จำเป็นต้องมีฟังก์ชันความเหมือน (Similarity function) ที่ชัดเจน ซึ่งจะมีผลกับกลุ่มข้อมูลที่แบ่งออกมาอย่างมาก ส่วนใหญ่แล้วฟังก์ชันเหล่านี้จะถูกใช้ในรูปของฟังก์ชันระยะห่าง (Distance function) ซึ่งเป็นการให้ค่าความต่างกันของข้อมูลมากกว่า โดยจะแตกต่างกันไปตามชนิดของข้อมูล เช่น ข้อมูลที่เป็นตัวเลข (Numerical) หรือข้อมูลแยกประเภท (Categorical) ฟังก์ชันระยะห่างที่นิยมกับข้อมูลที่เป็นตัวเลข เช่น ฟังก์ชันระยะห่างแมนฮัตตัน (Manhattan distance) ฟังก์ชันระยะห่างยูคลีเดียน (Euclidean function) และฟังก์ชันที่นิยมใช้กับข้อมูลแยกประเภท เช่น ฟังก์ชันความเหมือนแจ็คการ์ด (Jaccard similarity function) (คมกริช อุดมมณีธนกิจ 2548)

เทคนิคการจัดกลุ่มสามารถแบ่งเป็น 4 กลุ่มย่อยดังนี้

1. Partitioning เป็นการแบ่ง Cluster โดยทำการ partition criterion ซึ่งเมื่อได้กลุ่มของ cluster ก็จะแบ่งตามความเหมือนต่อ partition criterion ดังกล่าวออกเป็น cluster ย่อยๆ ตัวอย่างอัลกอริทึมที่ใช้ เช่น อัลกอริทึม K-means

2. Hierarchy เป็นการลำดับของ partition ซึ่งก็คือการที่แต่ละวัตถุทำการรวม (merge) กันไปเรื่อยๆ จนกระทั่งได้ cluster ใหญ่เพียงหนึ่ง cluster

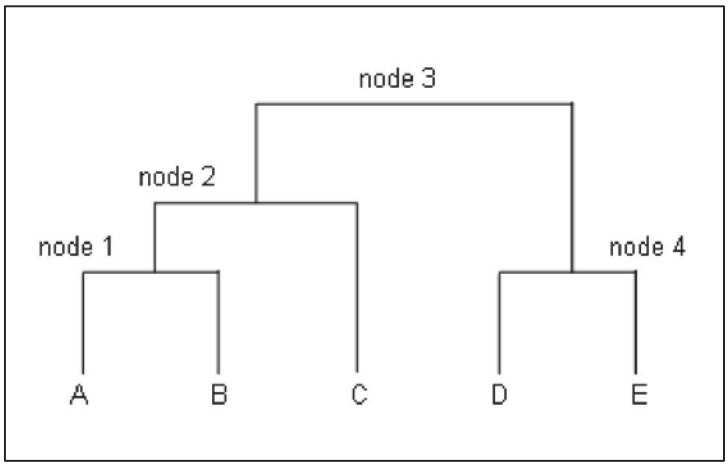
3. Density-based เป็นการหาพื้นที่ของข้อมูลที่มีความหนาแน่นมากกว่า threshold ที่กำหนดจากแต่ละ cluster คือ เป็นการแบ่งความหนาแน่นของข้อมูล

4. Grid-based เป็นการแบ่งพื้นที่ของข้อมูลออกเป็นเซล (cell) ย่อยแล้วทำการแบ่งข้อมูลตาม cell data space นั้นๆ

3.1 การจัดกลุ่มแบบลำดับชั้น

การจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) เป็นกรรมวิธีในการเชื่อมโยงเอกสารคล้ายโครงสร้างของต้นไม้ ซึ่งกรรมวิธีในการจัดกลุ่มแบบลำดับชั้นแบ่งออกได้เป็นอีก 2 แบบ คือ การรวมกลุ่มย่อยขึ้นไปเป็นกลุ่มใหญ่ (Agglomerative) หรือเป็นการประมวลผลแบบล่างขึ้นบน (Bottom up) และการแบ่งกลุ่มลำดับชั้นแบบย่อย (Division) หรือเป็นการประมวลผลแบบบนลงล่าง (Top down)

การจัดกลุ่มแบบลำดับชั้นนั้นจะเริ่มจากทุกออบเจ็คเดี่ยว (Single object) ในคลัสเตอร์เดี่ยว (Single cluster) จากนั้นในแต่ละรอบของการจัดกลุ่ม แต่ละออบเจ็คเดี่ยวจะถูกรวม (merge) เข้ากับออบเจ็คอื่นที่ใกล้กันเป็นคลัสเตอร์เดียวกัน โดยดูจากความคล้ายคลึงกันของออบเจ็ค ทำแบบนี้จนกระทั่งข้อมูลทั้งหมดกลายเป็นคลัสเตอร์เดียวเท่านั้น ตัวอย่างของการจัดกลุ่มข้อมูลแบบลำดับชั้นแสดงได้ดังภาพที่ 5



ภาพที่ 5 ตัวอย่างการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

จากภาพที่ 5 แสดงรายละเอียดได้ดังนี้

- คลัสเตอร์ Node 1 จะประกอบด้วยออบเจ็ค A และ B
- คลัสเตอร์ Node 2 จะประกอบด้วยออบเจ็ค A, B และ C
- คลัสเตอร์ Node 3 จะประกอบด้วยออบเจ็ค A, B, C, D และ E
- คลัสเตอร์ Node 4 จะประกอบด้วยออบเจ็ค D และ E
- คลัสเตอร์ Node 1 จะถูกรวมออบเจ็คที่ใกล้กันที่สุด (Closet) เข้าด้วยกันเป็นคลัสเตอร์แรก ถัดมาจะเป็นคลัสเตอร์ Node 4 ที่ถูกรวม จากนั้นเป็นคลัสเตอร์ Node 2 และ 3 ตามลำดับ
- จะเห็นว่าคลัสเตอร์ Node 3 จะมีสมาชิกเป็นออบเจ็คทั้งหมดจากข้อมูลทั้งหมดที่มี ซึ่งจะแทนคลัสเตอร์ที่มีระยะทางระหว่างสมาชิกแต่ละตัวของมันใหญ่ที่สุด (Largest distance)

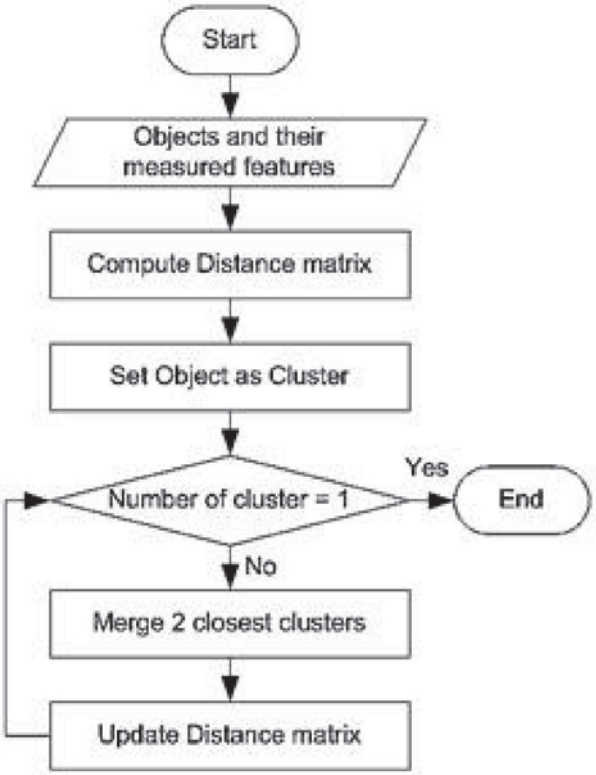
ลำดับชั้น (Hierarchy) ภายในคลัสเตอร์สุดท้าย (Final cluster) จะมีคุณสมบัติ ดังนี้

- คลัสเตอร์ที่ถูกสร้างในขั้นตอนก่อนหน้า (Early stages) จะถูกรวมเข้าเป็นคลัสเตอร์ใหม่ในขั้นตอนถัดมา (Later stages) ด้วย

- คลัสเตอร์ที่มีขนาดแตกต่างกันในต้นไม้ (Tree) นั้นจะมีค่าในการที่จะทำการวิเคราะห์หาความหมายต่อ

ขั้นตอนการจัดกลุ่มข้อมูลแบบลำดับชั้นแบบ Agglomerative มีดังนี้

- กำหนดให้แต่ละออบเจ็คเป็นคลัสเตอร์ที่แยกจากกัน (Separate cluster)
- ประเมินระยะทางระหว่างคลัสเตอร์แต่ละคู่ของคลัสเตอร์ทั้งหมด
- สร้างเมทริกซ์ระยะทาง (Distance matrix) โดยใช้ค่าระยะทาง (Distance values)
- มองหาคู่ของคลัสเตอร์ที่มีระยะทางสั้นที่สุด (Shortest distance) จากนั้นลบออกจากเมทริกซ์แล้วรวมคลัสเตอร์คู่นั้นเข้าด้วยกัน
- ประเมินระยะทางทั้งหมดจากคลัสเตอร์ใหม่กับคลัสเตอร์ที่มีอยู่ทั้งหมด แล้วทำการปรับปรุง (Update) เมทริกซ์ใหม่
- ทำตามขั้นตอนข้างบนซ้ำจนกระทั่งเมทริกซ์ระยะทางเหลือสมาชิกเพียงแค่หนึ่งตัวเท่านั้น (Single element)

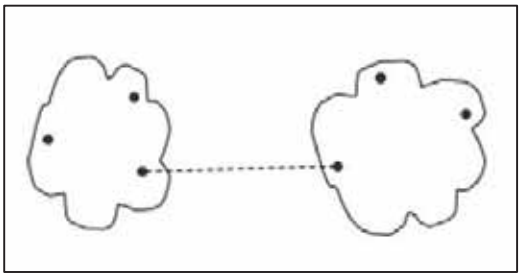


ภาพที่ 6 แสดงขั้นตอนการจัดกลุ่มแบบ Agglomerative

ที่มา : Kardi Teknomo, Hierarchical Clustering Tutorial [Online], accessed 9 August 2010. Available from http://people.revoledu.com/kardi/tutorial/Clustering/index.html

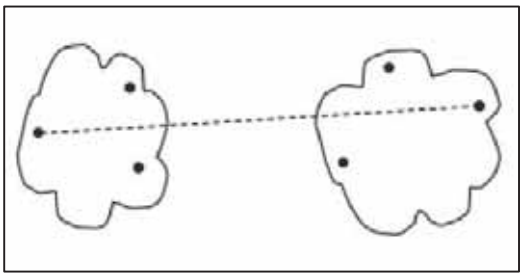
3.2 วิธีการจัดกลุ่มแบบลำดับชั้น

3.2.1 การเชื่อมโยงแบบ Single Link หรือเพื่อนบ้านที่ใกล้ที่สุด (Nearest neighbor) ในแต่ละขั้นตอนวิธีของการจัดกลุ่ม เลือกหาออบเจ็คที่มีค่าความเหมือนกับกลุ่มออบเจ็คและออบเจ็คที่จัดเข้ากลุ่มต้องมีความสัมพันธ์กับสมาชิกในกลุ่ม โดยมีระยะห่างที่น้อยที่สุดระหว่าง Cluster ขั้นตอนวิธีในการประยุกต์การจัดกลุ่ม Single Link คือ หลักการของ Minimum Spanning Tree



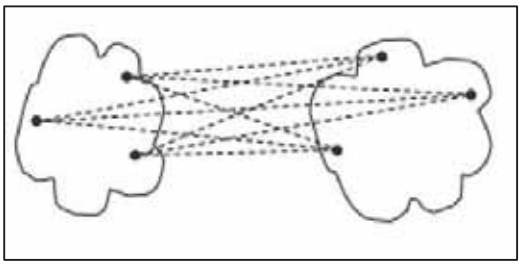
ภาพที่ 7 Single Link หาระยะห่างของออบเจ็คระหว่าง Cluster ที่น้อยที่สุด
ที่มา : Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining (U.S.A. : Pearson Addison Wesley, 2003), 517.

3.2.2 การเชื่อมโยงแบบ Complete Link ขั้นตอนและวิธีดังกล่าวจะตรงกันข้ามกับการเชื่อมโยงแบบ Single Link การเชื่อมโยงแบบ Complete Link ทำการหาค่าระยะห่างที่มากที่สุดระหว่าง Cluster



ภาพที่ 8 Complete Link หาระยะห่างของออบเจ็คระหว่าง Cluster ที่มากที่สุด
ที่มา : Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining (U.S.A. : Pearson Addison Wesley, 2003), 517.

3.2.3 การเชื่อมโยงแบบ Group Average Link ค่าความเหมือนของออบเจ็คกลุ่มใหม่ที่เกิดขึ้นจะใช้ค่าเฉลี่ยความเหมือนของออบเจ็คในกลุ่ม



ภาพที่ 9 Group Average Link หาระยะเฉลี่ยของออบเจ็คระหว่าง Cluster

ที่มา : Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining (U.S.A. : Pearson Addison Wesley, 2003), 517.

ตารางที่ 2 ตัวอย่างชุดข้อมูลสำหรับจัดกลุ่มอาหาร

Food item #	*Protein content, P*	*Fat content, F*
Food item #1	1.1	60
Food item #2	8.2	20
Food item #3	4.2	35
Food item #4	1.5	21
Food item #5	7.6	15
Food item #6	2.0	55
Food item #7	3.9	39

จากตารางที่ 2 เป็นตัวอย่างชุดข้อมูลอาหารที่ใช้ในการจัดกลุ่มอาหารด้วยการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) โดยวิธีวัดค่าความเหมือนแบบ Group Average Link และใช้ฟังก์ชันระยะห่างยูคลีเดียน (Euclidean function) ในการสร้างจะพิจารณาค่า Distance matrix ซึ่งเป็นวิธีการหาระยะทาง (Distance measure) แบบดั้งเดิม (Classical measure) ที่นิยมใช้กันมากในงานวิจัยต่างๆ โดยพิจารณาจากค่า Protein content (P) และ Fat content (F)

ในขั้นตอนที่ 1 จะเป็นการหา Centroid ของข้อมูล ซึ่งจะสมมติฐานว่าทุกๆ ข้อมูลเป็น Centroid อยู่ ขั้นตอนที่ 2 จะเป็นการหาระยะทางระหว่างข้อมูลทุกๆ ความเป็นไปได้ตามฟังก์ชันระยะห่างยูคลีเดียน (Euclidean function) ซึ่งเขียนเป็นสมการได้ดังนี้

$$D = \sqrt{\sum_{i=1}^{k} (p_i - q_i)^2} \qquad (2\text{-}6)$$

โดย D คือ ระยะทาง

p และ q คือ ข้อมูลของทั้งสองจุด ในทุกๆ แง่มุมจำนวนแง่มุม k

ซึ่งในขั้นตอนที่สองจะนำสมการที่ 2-6 มาใช้ในการหาระยะทางของทุกๆ คู่ ผลลัพธ์จากการหาระยะทางของทุกคู่แสดงดังตารางที่ 3 ตัวอย่างการหาค่าระยะทาง เช่น ระยะทางระหว่าง Food item #1 กับ Food item #2 แสดงได้ดังนี้

$$\begin{aligned} D_{c1,c2} &= \sqrt{(1.1-8.2)+(60-20)^2} \\ &= \sqrt{(-7.1)^2+(40)^2} \\ &= 40.62 \end{aligned}$$

ตารางที่ 3 ผลลัพธ์การหาระยะทางชุดข้อมูลอาหารในรอบที่ 1

Cluster number	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*
C1	0	40.62	25.19	39.00	45.46	5.08	21.18
C2	known	0	15.52	6.77	5.03	35.54	19.48
C3	known	known	0	14.25	20.28	20.12	4.01
C4	known	known	known	0	8.55	34.00	18.19
C5	known	known	known	known	0	40.39	24.28
C6	known	known	known	known	known	0	16.11
C7	known	known	known	known	known	known	0

จากตารางที่ 3 จะเห็นว่าจุดที่มีระยะทางน้อยที่สุด คือ C3 และ C7 โดยมีค่าระยะทางที่ 4.01 ดังนั้นจึงทำการรวมทั้ง 2 จุดเข้าด้วยกัน โดย centroid คือ P = (4.2+3.9)/2 และ F = (35+39)/2 ซึ่งผลลัพธ์จากการรวมกันของทั้ง 2 จุดแสดงได้ดังตารางที่ 4

ตารางที่ 4 ผลลัพธ์ที่ได้จากการรวม Cluster 3 และ Cluster 7

Cluster number	*Protein content, P*	*Fat content, F*
C1	1.1	60
C2	8.2	20
C3,C7	4.05	37
C4	1.5	21
C5	7.6	15
C6	2.0	55

ในขั้นตอนที่ 3 จะเป็นการทำซ้ำในขั้นตอนที่ 2 ไปเรื่อยๆ จนกระทั่งผลการรวมเหลือเพียงหนึ่งคลัสเตอร์ ในตัวอย่างนี้จะกำหนดกลุ่มที่ต้องการไว้ที่ 4 กลุ่ม (Cluster) ดังนั้นผลการรวมกลุ่มในรอบต่อไปจะแสดงในตารางที่ 5 – 7

ตารางที่ 5 ผลลัพธ์การหาระยะทางชุดข้อมูลอาหารในรอบที่ 2

Cluster number	*C1*	*C2*	*C3,C7*	*C4*	*C5*	*C6*
C1	0	40.62	25.19	39.00	45.46	5.08
C2	known	0	17.49	6.77	5.03	35.54
C3,C7	known	known	0	16.20	22.28	18.11
C4	known	known	known	0	8.55	34.00
C5	known	known	known	known	0	40.26
C6	known	known	known	known	known	0
C7	known	known	known	known	known	known

จากตารางที่ 5 จะเห็นว่าจุดที่มีระยะทางน้อยที่สุด คือ C2 และ C5 โดยมีค่าระยะทางที่ 5.03 ดังนั้นจึงทำการรวมทั้ง 2 จุดเข้าด้วยกัน โดย centroid คือ P = (8.2+7.6)/2 และ F = (20+15)/2 ซึ่งผลลัพธ์จากการรวมกันของทั้ง 2 จุดแสดงได้ดังตารางที่ 6

ตารางที่ 6 ผลลัพธ์ที่ได้จากการรวม Cluster 2 และ Cluster 5

Cluster number	*Protein content, P*	*Fat content, F*
C1	1.1	60
C2,C5	7.9	17.5
C3,C7	4.05	37
C4	1.5	21
C6	2.0	55

และเมื่อทำไปเรื่อยๆ จนกระทั่งได้จำนวนกลุ่มตามที่กำหนด คือ 4 กลุ่ม (Cluster) จะได้ ได้ผลลัพธ์การจัดกลุ่มดังตารางที่ 7

ตารางที่ 7 ผลลัพธ์สุดท้ายจากการทำ Clustering จำนวน 4 กลุ่ม

Cluster number	*Protein content, P*	*Fat content, F*
C1,C6	1.55	57.50
C2,C5	7.9	17.5
C3,C7	4.05	37
C4	1.5	21

3.3 งานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มแบบลำดับชั้น

อรนุช ชัยหมื่น (2548) ได้ประยุกต์ใช้เทคนิคการจัดกลุ่ม (Clustering) แบบ 2 ขั้นตอนใน 2 วิธีการ เพื่อศึกษาลักษณะและพฤติกรรมการสั่งซื้อสินค้าของลูกค้าสินค้าหัตถกรรมไทย เพื่อนำเสนอทางเลือกในการวิเคราะห์ข้อมูลในเชิงธุรกิจแบบใหม่ที่มีประสิทธิภาพสูง โดยทำการเปรียบเทียบวิธีการแบ่งกลุ่มแบบ 2 ขั้นตอน 2 วิธีการคือ SOM (Self's Organizing Map) กับ K-Means เปรียบเทียบกับ Hierarchical Clustering กับ K-Means โดยทำการเปรียบเทียบจากค่าส่วนเบี่ยงเบนมาตรฐานของแต่ละกลุ่ม ที่แต่ละวิธีแบ่งกลุ่มได้ โดยใช้ชุดทดลอง 4 ชุด โดยพิจารณาจาก 7 ปัจจัย ผลการทดลองพบว่าวิธีการที่ 1 คือ SOM กับ K-Means สามารถแบ่งกลุ่มได้ดีกว่าวิธีการที่ 2 คือ Hierarchical Clustering กับ K-Means จากผลการทดลองยังพบอีกว่าถ้ากลุ่มข้อมูลที่ใช้ทดสอบมีจำนวนน้อยวิธีการแบบ 2 ขั้นตอนที่เหมาะสมคือ Hierarchical Clustering กับ K-Means เนื่องจากเป็นวิธีการแบ่งกลุ่มที่คลอบคลุมลูกค้าและให้ค่าการแบ่งกลุ่มที่ไม่มากนัก แต่หากข้อมูลทดลองมี

จำนวนมากประมาณ 15,000 รายการขึ้นไปควรใช้วิธีการ SOM กับ K-Means ซึ่งจะให้จำนวนการแบ่งกลุ่มที่ละเอียดเหมาะกับข้อมูลที่มีจำนวนมาก

บุญทัน คำพาศัย (2549) ได้ศึกษาการจัดกลุ่มแขวงต่างๆ ในประเทศสาธารณรัฐประชาธิปไตยประชาชนลาว ซึ่งมีทั้งหมด 17 แขวงด้วยเทคนิคการวิเคราะห์การจัดกลุ่ม (Cluster Analysis) 3 วิธี ได้แก่ การจัดกลุ่มแบบ 2 ขั้น (Two-steps) การจัดกลุ่มแบบไม่มีขั้นตอน (K-means) และการจัดกลุ่มแบบเป็นขั้นตอน (Hierarchy) โดยอาศัยข้อมูลทางเศรษฐกิจ 9 ตัวแปร และข้อมูลทางสังคม 13 ตัวแปรของปี พ.ศ. 2547 ผลการวิจัยสรุปได้ว่า เมื่อเปรียบเทียบเทคนิคการจัดกลุ่มทั้ง 3 วิธีตามตัวแปรด้านเศรษฐกิจพบว่าการจัดกลุ่มโดยวิธี Agglomerative Hierarchical Methods และใช้การเชื่อมโยงแบบ Single หรือแบบ Complete หรือแบบ Ward สามารถจัดกลุ่มแขวงได้เหมาะสมที่สุดโดยจัดได้ 7 กลุ่ม กรณีใช้ตัวแปรด้านสังคมพบว่าการจัดกลุ่มโดยวิธี Agglomerative Hierarchical Methods และใช้การเชื่อมโยงแบบ Complete สามารถจัดกลุ่มได้ดีที่สุดโดยจัดได้ 8 กลุ่ม และเมื่อพิจารณาทั้งตัวแปรด้านเศรษฐิจและสังคมรวมกันพบว่าการจัดกลุ่มแขวงโดยวิธี Agglomerative Hierarchical Methods และใช้การเชื่อมโยงแบบ Centroid สามารถจัดกลุ่มแขวงได้เหมาะสมที่สุดโดยจัดได้ 12 กลุ่ม

สิทธา ดอร์จิ (2552) นำเสนอการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูลแบบผสมผสานระหว่าง Agglomerative Hierarchical Clustering และ Fuzzy C-means Clustering เพื่อนำเสนออัลกอริทึมใหม่สำหรับการจัดกลุ่มข้อมูลลูกค้าที่ใช้งานโทรศัพท์ทางไกล โดยนำข้อดีจากทั้งสองอัลกอริทึมมาประยุกต์ใช้ร่วมกันเพื่อแก้ปัญหาข้อด้อย โดยการเลือกพารามิเตอร์เริ่มต้นที่เหมาะสมเพื่อให้ได้ผลลัพธ์จากการจัดกลุ่มที่เหมาะสมขึ้น กล่าวคือ แม้ว่าผลลัพธ์จากการจัดกลุ่มด้วย Fuzzy C-means จะยอมให้ข้อมูลแต่ละระเบียนสามารถเป็นสมาชิกของกลุ่มได้มากกว่าหนึ่งกลุ่มด้วยความน่าจะเป็นที่แตกต่างกัน แต่ก็ต้องขึ้นอยู่กับจำนวนกลุ่มที่ผู้ใช้ต้องระบุล่วงหน้า ค่าตัวอ้างอิงกลุ่มและค่าความเป็นสมาชิกเริ่มต้น ในทางกลับกันการจัดกลุ่มด้วย Agglomerative Hierarchical ไม่ต้องการจำนวนกลุ่มที่ระบุล่วงหน้า ข้อมูลที่ใช้ในการทดลองได้มาจากบริษัท Bhutan Telecom Ltd. โดยใช้ข้อมูลระเบียนรายละเอียดการใช้โทรศัพท์เคลื่อนที่ของลูกค้า (Call Details Records : CDR) วิธีการทดลองจะเริ่มจากการจัดกลุ่มด้วย Agglomerative Hierarchical ก่อนจากนั้นจึงส่งต่อให้วิธีการจัดกลุ่มแบบ Fuzzy C-means ผลการประเมินจากการทดสอบกับข้อมูลเปรียบเทียบจากแหล่งข้อมูล UCI พบว่า อัลกอริทึมที่นำเสนอนี้มีประสิทธิภาพการจัดกลุ่มได้ดีทัดเทียมกับวิธี Fuzzy C-means หรือวิธี K-means แต่มีข้อดีกว่าวิธีทั้งสองในด้านผลการจัดกลุ่มที่ไม่เปลี่ยนแปลงตำแหน่งอ้างอิงของกลุ่ม

4. แม่พิมพ์โลหะ

แม่พิมพ์โลหะเป็นแม่พิมพ์ที่นำมาใช้ในการผลิตสินค้าหรือผลิตภัณฑ์ที่เป็นโลหะ เช่น น๊อต ตัวถังรถยนต์ ถาดอลูมิเนียม ช้อนส้อม ฯลฯ รวมไปถึงชิ้นส่วนขนาดเล็กในอุปกรณ์อิเล็คทรอนิกส์ เช่น ส่วนประกอบของฮาร์ดดิสก์ (Hard Disk) ในเครื่องคอมพิวเตอร์ เป็นต้น เราสามารถจำแนกชนิดของแม่พิมพ์ได้ 2 วิธี (เอกสารการอบรมการบำรุงรักษาแม่พิมพ์โลหะ สถาบันไทย-เยอรมัน 2550, ห้องสมุดกรมส่งเสริมอุตสาหกรรม 2551) คือ

4.1 การแบ่งชนิดแม่พิมพ์ตามขบวนการหรือกระบวนการที่ใช้ปฏิบัติงาน

4.1.1 แม่พิมพ์ปั๊ม (Stamping) เป็นวิธีการนำแผ่นเหล็ก (Strip) เข้ามาสู่ยังเครื่องปั๊มที่มีแม่พิมพ์ประกอบติดอยู่กับแท่นปั๊ม เมื่อแผ่นสตริปเข้ามายังแท่นปั๊มในตำแหน่งที่ต้องการแล้ว แท่นปั๊มจะกดลงมายังแผ่นสตริปเพื่อให้ได้ชิ้นงานตามแบบแม่พิมพ์ ชิ้นงานที่ได้จากแม่พิมพ์ปั๊ม เช่น ชิ้นส่วนยานยนต์ ชิ้นส่วนเครื่องใช้ไฟฟ้าและอิเล็กทรอนิกส์ เป็นต้น

4.1.2 แม่พิมพ์ขึ้นรูป (Forming) เป็นการเปลี่ยนรูปทรงของแผ่นเหล็กให้เป็นไปตามรูปทรงของพันซ์ (Punch) และดาย (Die) โดยไม่มีการเปลี่ยนแปลงความหนาของเหล็ก แม่พิมพ์ขึ้นรูปมักจะนำใช้ในการผลิตชิ้นส่วนยานยนต์

4.1.3 แม่พิมพ์ดึงขึ้นรูปลึก (Deep draw die) เป็นการควบคุมการใช้แรงกดดันหรือแรงที่กดลงบนแผ่นงาน (Blank) หรือชิ้นงาน (Work piece) ดันผ่านแม่พิมพ์ ด้วยพั้นช์ (Punch) ให้มีรูปร่างเป็นหลุมหรือโพรงลงไป โดยที่ความหนาของชิ้นงานมีความหนาเท่าวัสดุเดิม ผลิตภัณฑ์ที่ได้จากแม่พิมพ์ดึงขึ้นรูปลึก เช่น ชิ้นส่วนยานยนต์ ชิ้นส่วนไฟฟ้าและอิเล็กทรอนิกส์ เป็นต้น

4.1.4 แม่พิมพ์ตีขึ้นรูป (Forging) เป็นแม่พิมพ์ที่ใช้ในการแปรรูปโลหะให้ได้รูปร่างตามที่กำหนดเป็นจำนวนมาก เช่น น๊อต สกรู เพลา เครื่องมือช่าง ชิ้นส่วนยานยนต์ เป็นต้น ชิ้นงานที่ผ่านการตีขึ้นรูปความร้อนจะมีความหนาแน่นและคุณสมบัติทางกายภาพที่ดีขึ้น เช่น ความแข็งแรง ความสามารถในการรับแรงกระแทก (Toughness) ทั้งนี้การตีขึ้นรูปสามารถแบ่งได้ตามลักษณะอุณหภูมิที่ใช้ ได้แก่ Cold Forging ซึ่งเป็นการตีขึ้นรูปที่อุณหภูมิห้อง Warm Forging เป็นการตีขึ้นรูปชิ้นงานที่อุณหภูมิที่ยังไม่มีการเปลี่ยนแปลงโครงสร้างในทางโลหะวิทยา และ Hot Forging เป็นการตีขึ้นรูปที่อุณหภูมิสูง โลหะมีการเปลี่ยนแปลงรูปได้ง่าย กระบวนการตีขึ้นรูปเริ่มจากการเตรียมวัตถุดิบที่อยู่ในรูปของเหล็กเส้นรีดร้อน หรือในกรณีที่ต้องการให้มีขนาดหน้าตัดที่แน่นอนและคุณภาพผิวที่สูงขึ้นจะทำการดึงเย็นเพื่อลดขนาดเหล็กเส้นรีดร้อนให้อยู่ในรูปเพลาขาว แล้วจึงเอาเหล็กเข้าไปตีขึ้นรูปในแม่พิมพ์ที่ได้จัดเตรียมไว้เพื่อให้ได้ชิ้นงานตามต้องการ

4.1.5 แม่พิมพ์ฉีดหล่อ (Die casting) เป็นวิธีการหล่อที่ใช้ความดันสูงอัดน้ำโลหะเข้าสู่แม่พิมพ์ โดยน้ำโลหะนั้นจะนำเอาวัตถุดิบ เช่น เหล็ก อลูมิเนียม เป็นต้น ผ่านเข้าเตาหล่อเพื่อ

หลอมโลหะให้กลายเป็นน้ำโลหะ จากนั้นน้ำโลหะจะวิ่งเข้าสู่แม่พิมพ์โดยผ่านทางรูเข้าของแม่พิมพ์ รูเข้าจะต้องออกแบบให้อยู่ในลักษณะที่ทำให้น้ำโลหะวิ่งเข้าแม่พิมพ์ได้สะดวก โดยอาศัยความดันเข้าช่วย ทิ้งไว้สักครู่แล้วจึงทำการแกะชิ้นงานออกจากแบบ ข้อดีของแม่พิมพ์ฉีดหล่อ คือ สามารถผลิตชิ้นงานที่มีความซับซ้อน ผลิตชิ้นงานบางได้ อัตราการผลิตสูง และมีความเที่ยงตรงสูง ส่วนข้อเสีย คือ ไม่สามารถผลิตชิ้นงานที่มีขนาดใหญ่ได้ แม่พิมพ์มีราคาแพง โลหะที่ใช้ต้องมีจุดหลอมเหลวต่ำ ปัญหาด้านสิ่งแวดล้อม ผลิตภัณฑ์ที่ได้จากแม่พิมพ์ฉีดหล่อ เช่น ชิ้นส่วนยานยนต์ ชิ้นส่วนเครื่องจักร เครื่องใช้ภายในบ้าน ท่อน้ำ เป็นต้น

4.2 การแบ่งชนิดแม่พิมพ์ตามลักษณะโครงสร้างของแม่พิมพ์

4.2.1 แม่พิมพ์เดี่ยว (Single Die) เป็นแม่พิมพ์ที่มีการทำงานเพียงการทำงานเดียวในหนึ่งแม่พิมพ์ มักใช้สำหรับผลิตชิ้นงานที่มีจำนวนไม่มากนักและใช้ในการผลิตชิ้นงานใหญ่ๆ และยาก โดยทั่วๆไปแล้ว แม่พิมพ์เดี่ยวจะมีเงื่อนไขเบื้องต้นว่าเป็นงานที่ใช้คนทำเป็นส่วนใหญ่ และมีลักษณะการทำงานอย่างใดอย่างหนึ่ง และอาจกล่าวได้ว่าการแปรรูปชิ้นงานโดยใช้แม่พิมพ์เดี่ยวเป็นขั้นตอนสำคัญของการทำงานในแม่พิมพ์ตัวต่อๆไป ตัวอย่างแม่พิมพ์ในกลุ่มนี้ เช่น แม่พิมพ์ตัดผ่านชิ้นงาน แม่พิมพ์ดัดชิ้นงาน แม่พิมพ์เจาะรู เป็นต้น

4.2.2 แม่พิมพ์ผสมรวม (Compound Die) เป็นแม่พิมพ์ที่มีขั้นตอนในการขึ้นรูปชิ้นงานโลหะแผ่นจำนวน 2-3 ขั้นตอนขึ้นไปผสมรวมกันอยู่เป็นหนึ่งสถานีงานบนแม่พิมพ์ตัวเดียว (Single Station Die) โดยจะมีการขึ้นรูปชิ้นงานเพียงหนึ่งตำแหน่ง แต่ถ้าชิ้นงานที่ถูกขึ้นรูปไม่มีศูนย์กลางร่วมกันจะมีการขึ้นรูปสองตำแหน่งในหนึ่งสถานี แม่พิมพ์ผสมรวมจะประหยัดกว่าแม่พิมพ์แบบเดี่ยวและให้ความถูกต้องแม่นยำของชิ้นงานสูงกว่า ตัวอย่างแม่พิมพ์ในกลุ่มนี้ เช่น แม่พิมพ์ตัดแผ่นเปล่าพร้อมเจาะรู

4.2.3 แม่พิมพ์แบบผสมแยกส่วน (Combination Die) เป็นแม่พิมพ์ที่มีรูปแบบการทำงานที่ผิดปกติไปจากแม่พิมพ์แบบอื่นๆ แม่พิมพ์แบบนี้สามารถใช้ในการขึ้นรูปชิ้นงานที่มีรูปร่างแตกต่างกันจำนวน 2 ชิ้นได้ในเวลาเดียวกัน เช่น ตัดแผ่นชิ้นงานและดัดงอขึ้นรูปชิ้นงาน โดยแม่พิมพ์ชุดหนึ่งจะถูกทำเป็นสองสถานีงานหรือมากกว่า ซึ่งนี่ก็คือจุดเริ่มต้นของแม่พิมพ์หลายสถานีงาน (Multistation Die) แม่พิมพ์แบบนี้จะแตกต่างจากแม่พิมพ์แบบต่อเนื่อง (Progressive Die) ในกรณีที่แผ่นป้อนชิ้นงานจะถูกขึ้นรูปจากสถานีที่หนึ่งไปยังสถานีที่สองโดยใช้คนงานหรือเครื่องมือกลป้อนชิ้นงาน และแม่พิมพ์แบบนี้จะไม่มีการป้อนชิ้นงานเข้าสู่แต่ละสถานีงานอย่างต่อเนื่อง ทั้งนี้เพราะว่าแต่ละสถานีงานบนแม่พิมพ์แบบผสมแยกส่วนจะมีการทำงานที่แตกต่างกัน

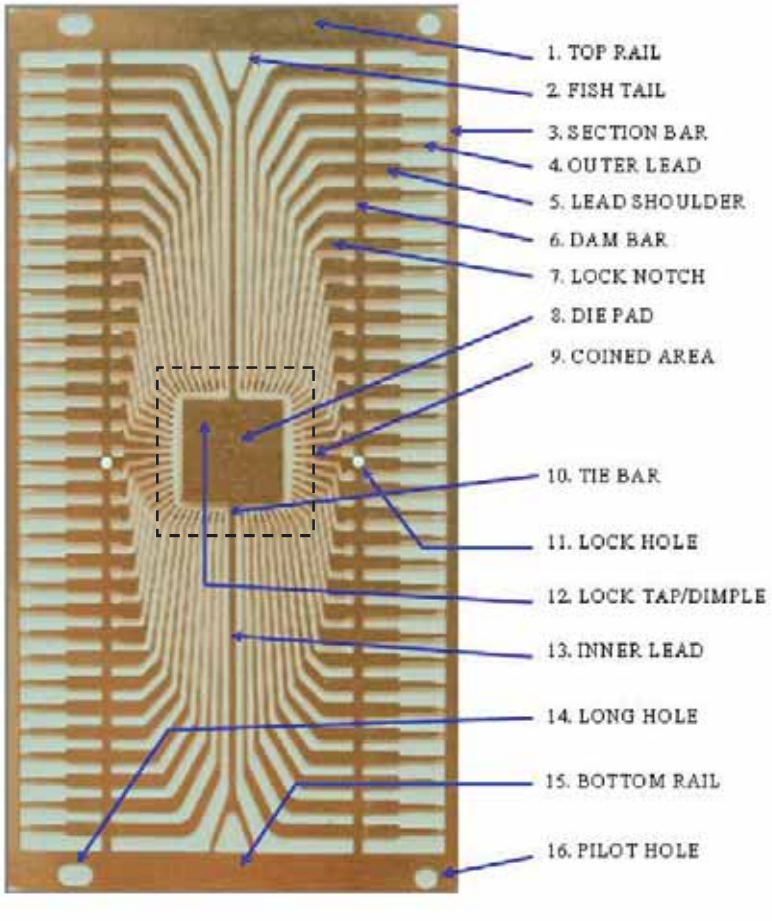
4.2.4 แม่พิมพ์แบบต่อเนื่อง (Progressive Die) เป็นแม่พิมพ์ที่มีสถานีการทำงานมากกว่า 3 สถานีขึ้นไป โดยแต่ละสถานีจะเรียงตัวอยู่ในแนวเดียวกัน แม่พิมพ์แบบนี้สามารถป้อน

แผ่นชิ้นงานด้วยมือคน โดยแผ่นป้อนชิ้นงานจะต้องถูกตัดซอยออกมาเป็นแถบยาวๆ และมีการกำหนดตำแหน่งชิ้นงานแบบสลัก หรืออาจป้อนด้วยแผ่นป้อนชิ้นงานแบบม้วน (Coil Stock) ซึ่งจะเป็นการป้อนอัติโนมัติ แผ่นป้อนชิ้นงานจะถูกป้อนชิ้นงานไปยังสถานีแรกเพื่อเจาะรูก่อน หลังจากผ่านการเจาะรูแล้วแผ่นป้อนชิ้นงานจะถูกป้อนเคลื่อนที่ไปยังสถานีอื่น โดยจะใช้ Pilot เป็นตัวกำหนดตำแหน่ง ตัวชิ้นงานจะติดอยู่กับโครงร่างของเศษโลหะตลอดเวลาจนถึงสถานีสุดท้ายจนสำเร็จเป็นชิ้นงาน โดยชิ้นงานอาจถูกตัดเป็นแผ่นให้หลุดออกมากลายเป็นชิ้นงาน หรือถูกป้อนให้กลับเข้าม้วน (Coil Stock) ที่ปลายทางสถานีสุดท้าย แม่พิมพ์แบบนี้เหมาะกับการผลิตที่มีปริมาณการผลิตจำนวนมากๆ และมีต้นทุนค่อนข้างสูง

4.2.5 แม่พิมพ์แบบส่งถ่าย (Transfer Die) เป็นแม่พิมพ์ที่มีหลายสถานีการทำงาน ซึ่งชิ้นงานจะถูกป้อนแบบอัติโนมัติเข้าสู่แม่พิมพ์ อุปกรณ์ป้อนอัติโนมัติจะเป็นแผ่นโลหะ 2 แผ่นยึดจับชิ้นงานจากสถานีหนึ่งไปใส่ในอีกสถานีหนึ่งเพื่อขึ้นรูปต่อเนื่องกันไปจนกระทั่งเสร็จ โดยที่ชิ้นงานจะแยกออกอิสระเป็นชิ้นๆ ไม่ได้ยึดติดกับโครงร่างของเศษโลหะเหมือนกับแม่พิมพ์แบบต่อเนื่อง โดยทั่วไปเพื่อความเหมาะสมควรป้อนแผ่นชิ้นงานสำเร็จรูปที่ถูกตัดมาแล้วเข้าสู่แม่พิมพ์แบบส่งถ่าย และแผ่นป้อนชิ้นงานจากม้วนโลหะ (Coil Stock) ควรป้อนเข้าสู่แม่พิมพ์แบบต่อเนื่องจะเหมาะสมกว่า ข้อดีของแม่พิมพ์ส่งถ่าย คือ แต่ละสถานีงานจะทำงานเป็นอิสระต่อกัน ทำให้เกิดความสะดวกในการทำงานบนสถานีนั้น แต่ในแม่พิมพ์แบบต่อเนื่องแผ่นชิ้นงานที่ถูกตัดจะยึดติดกับโครงร่างของเศษโลหะ ซึ่งมีผลทำให้โครงร่างของเศษโลหะยึดเข้าหรือหดตัวตามกรรมวิธีการขึ้นรูปของสถานีงานนั้น อาจทำให้โครงร่างของเศษโลหะเกิดการแตกได้

5. IC Lead Frame

Lead Frame คือ ฐานรองตัว Semiconductor Chip โดยใช้ Resin เป็นตัวยึดกับ Pad ของ Lead Frame และใช้ Wire Bonding เป็นสื่อนำกระแสไฟฟ้าสำหรับติดต่อให้วงจรกระแสไฟฟ้าจากภายในตัว IC ออกสู่ภายนอก โดยที่ใช้แต่ละขาของ Lead Frame ทำหน้าที่เป็นสื่อในการนำกระแสไฟฟ้าออกไปสู่ระบบภายนอก ตัวอย่างโครงสร้างของ Lead Frame แสดงได้ดังภาพที่ 10



ภาพที่ 10 ตัวอย่างโครงสร้างของ Lead Frame

โครงสร้างของ Lead Frame

รูปแบบและโครงสร้างของ Lead Frame จะมีความแตกต่างกันบ้างตาม Model การออกแบบของแต่ละผลิตภัณฑ์ โดยทั่วไปจะมีโครงสร้างหลักๆ ตามตัวอย่างในภาพที่ 10 ดังต่อไปนี้

1. TOP RAIL คือ ขอบบนของ Lead Frame (จะถูกตัดทิ้งในกระบวนการท้ายๆของการผลิต Semiconductor)
2. FISH TAIL คือ ส่วนของ Tie Bar ที่เชื่อมติดกับขอบของ Unit (เป็นตัวป้องกัน Die Pad ไม่ให้ขยับ)
3. SECTION BAR คือ ส่วนที่ตัดแยกแต่ละ Unit ออกจากกัน
4. OUTER LEAD คือ ส่วนที่อยู่ถัดจาก Dam Bar เป็นตัวเชื่อม Circuit Boards
5. LEAD SHOULDER คือ ตัวยึดขา IC ให้หยุดนิ่ง ไม่ให้หลุดไปจากตำแหน่ง

6. DAM BAR คือ ตัวแบ่ง Inner Lead กับ Outer Lead ออกจากกัน คล้ายเขื่อนที่กัน Resin ไม่ให้ไปเชื่อมกับ Outer Lead ใน Process Flow

7. LOCK NOTCH คือ ตัวหยักที่เป็นตัวยึดไม่ให้ตัว Unit เคลื่อนที่ไปมา

8. DIE PAD คือ พื้นที่ที่ให้ตัว Semicon Chip เข้าไปตั้งวาง

9. COINED AREA คือ บริเวณที่ราบของปลายขา Inner Lead ที่เชื่อม Wire

10. TIE BAR คือ ตัวยึดระหว่าง Die Pad กับแต่ละขอบของ Lead Frame และทำหน้าที่นำความร้อนจาก IC สู่ภายนอก ดังนั้น Tie Bar หน้าตัดกว้างยิ่งดีเพราะพลังงานความร้อนจะสามารถกระจายออกจาก IC ได้มาก

11. LOCK HOLE คือ ตัวยึดที่ทำหน้าที่ยึด Resin (Compound) เมื่อเวลา Molding ไม่ให้เคลื่อนที่

12. LOCK TAP/DIMPLE คือ ตัวยึดที่ไม่ให้ Resin บน Die Pad เคลื่อนที่ไปมา

13. INNER LEAD คือ บริเวณขา Lead ที่อยู่ถัดจาก Dam Bar เข้ามาของแต่ละขอบ หรือบางครั้งจะอ้างอิงจาก Coined Area

14. LONG HOLE คือ ตัวยึดไม่ให้ Feed เลื่อนไปเลยกว่าความเป็นจริง Unit by Unit (ตัวจัดลำดับ)

15. BOTTOM RAIL คือ ตัวที่ทำหน้าที่คล้าย Top Rail แต่อยู่ด้านล่าง (คล้ายเปลือกแตงโมที่ทำหน้าที่ยึดเนื้อภายในไว้)

16. PILOT HOLE คือ ตัววัดระยะระหว่างแต่ละความกว้างให้เท่าๆ กันของแต่ละที่ในตัว Lead Frame

6. KPI (Key Performance Indicator)

6.1 ความหมายของ KPI

KPI หมายถึง ดัชนีชี้วัดผลสำเร็จทางธุรกิจ เป็นความแนบเนียนในการนำปัจจัยวัดผลความสำเร็จทางธุรกิจที่นอกเหนือจากด้านการเงิน (Financial Perspective) ด้านลูกค้า (Customers Perspective) ด้านกระบวนการภายใน (Internal Business Process Perspective) และด้านการเรียนรู้และการเติบโต (Learning and Growth Perspective) เข้ามา ทำให้เกิดความร่วมมือและสนับสนุนกลยุทธ์ให้ดำเนินไปสู่ภาคปฏิบัติจนประสบความสำเร็จ (ดนัย เทียนพุฒ 2544 : 23)

ดังนั้น KPI จึงเป็นปัจจัยหรือตัวชี้วัดที่อยู่ในขอบเขตที่บอกว่าจะทำธุรกิจอย่างไรจึงจะบรรลุเป้าหมายตามที่องค์กรตั้งไว้ ในการเลือก KPI ของแต่ละองค์กรจะต้องสะท้อนให้ได้ว่าปัจจัยอะไรที่สำคัญที่ทำให้องค์กรประสบความสำเร็จทางธุรกิจ เช่น ภาพรวมของคุณภาพการผลิต โดย

อาจกำหนดตัวชี้วัดในเรื่องของปริมาณงานที่ได้จากการผลิตหรืองานที่ต้องทำใหม่ (Rework) เพื่อให้ได้มาตรฐาน หรือภาพรวมของคุณภาพในการให้บริการ ตัวอย่างเช่น บริษัทประกันอาจเปิดศูนย์การประเมินความเสียหายโดยไม่ต้องผ่านขั้นตอนการเคลม (Claims) เป็นต้น

6.2 ลักษณะของ KPI ที่ดี

- สอดคล้องกับวิสัยทัศน์ ภารกิจ และกลยุทธ์ขององค์กร
- ควรแสดงถึงสิ่งที่มีความสำคัญต่อองค์กรและหน่วยงานเท่านั้น ซึ่งดัชนีชี้วัดที่มีความสำคัญต่อองค์กรและหน่วยงานมี 2 ลักษณะ คือ ดัชนีชี้วัดที่แสดงผลการดำเนินงานที่สำคัญต่อองค์กร และดัชนีชี้วัดกิจกรรมหรืองานที่สำคัญ ซึ่งหากผิดพลาดจะก่อให้เกิดปัญหาร้ายแรงในองค์กรหรือหน่วยงาน
- ประกอบด้วยดัชนีชี้วัดที่เป็นเหตุและดัชนีชี้วัดที่เป็นผล
- ดัชนีชี้วัดที่สร้างขึ้นควรเป็นดัชนีชี้วัดที่องค์กรหรือหน่วยงานสามารถควบคุมผลงานได้ไม่น้อยกว่าร้อยละ 80
- เป็นดัชนีชี้วัดที่สามารถวัดผลได้ และบุคคลทั่วไปเข้าใจ ไม่ใช่มีเพียงผู้จัดทำเท่านั้นที่เข้าใจ
- ต้องช่วยให้ผู้บริหารและพนักงานสามารถติดตามการเปลี่ยนแปลงที่สำคัญขององค์กรได้ นอกเหนือจากการใช้ดัชนีชี้วัดเพื่อการประเมินผลงาน

6.3 ขั้นตอนการสร้าง KPI

- กำหนดวัตถุประสงค์หรือผลลัพธ์ที่องค์กรต้องการ (What to measure?) ซึ่งควรสะท้อนถึงกลยุทธ์ที่องค์กรมุ่งเน้น
- กำหนดปัจจัยสู่ความสำเร็จหรือปัจจัยวิกฤต (Key Success Factor or Critical Success Factor) ที่สัมพันธ์กับวัตถุประสงค์หรือผลลัพธ์ที่องค์กรต้องการ
- กำหนดตัวชี้วัดที่สามารถบ่งชี้ความสำเร็จ ประสิทธิภาพ ประสิทธิผลจากการดำเนินการตามวัตถุประสงค์หรือผลลัพธ์ที่องค์กรต้องการ (How to measure?) ซึ่งสามารถแสดงผลเป็นข้อมูลในเชิงปริมาณ และกำหนดสูตรคำนวณรวมทั้งหน่วยของดัชนีชี้วัดแต่ละตัว
- กลั่นกรองดัชนีชี้วัดเพื่อหาดัชนีชี้วัดหลัก โดยจัดลำดับและกำหนดน้ำหนักความสำคัญของดัชนีชี้วัดแต่ละตัว
- กระจายดัชนีชี้วัดสู่หน่วยงานที่เกี่ยวข้อง
- จัดทำ KPI Dictionary โดยระบุรายละเอียดที่สำคัญของดัชนีชี้วัดแต่ละตัว เช่น ชื่อของดัชนี คำจำกัดความหรือนิยามของดัชนีชี้วัด สูตรในการคำนวณ หน่วยของดัชนีชี้วัด ผู้เก็บ

ข้อมูล ความถี่ในการรายงานผลเพื่อสร้างความเข้าใจร่วมกันของผู้ที่เกี่ยวข้องในการนำดัชนีชี้วัดไปใช้ในการปฏิบัติงาน

6.4 ปัจจัยที่ทำให้การใช้ KPI ล้มเหลว

หลังจากที่นำ KPI มาประยุกต์ใช้ในองค์กร ผู้บริหารหรือผู้ใช้ KPI ในหลายองค์กรที่ใช้KPI แล้วไม่บรรลุผลตามเป้าหมายที่ตั้งไว้ มักจะเกิดคำถามต่างๆ ต่อไปนี้ (เสาวคนธ์ วิทวัสโอฬาร 2550)

- “ใช้ KPI มาสองปีแล้ว ไม่เห็นจะช่วยเพิ่มผลประกอบการเลย”
- “คนยังทำงานเหมือนเดิม ทั้งๆ ที่ทำเรื่อง KPI และ Balanced scorecard”
- “พอใช้ KPI แล้ว คนในองค์กรเห็นแก่ตัว กลัวว่าคะแนนประเมินของตัวเองตก”
- “คนจำไม่ได้เลยด้วยซ้ำว่า KPI ของตนคืออะไร”

ตัวอย่างของปัจจัยที่ทำให้การนำ KPI ไปใช้แล้วไม่ประสบผลสำเร็จมีดังต่อไปนี้

- ผู้บริหารขาดความมุ่งมั่นในการสร้างดัชนีชี้วัดความสำเร็จของงาน
- การกำหนดดัชนีชี้วัดและค่าเป้าหมายที่มีความลำเอียง
- ดัชนีชี้วัดแต่ละตัวไม่อยู่บนพื้นฐานที่สามารถเปรียบเทียบกันได้
- ช่วงเวลาในการเก็บข้อมูลของดัชนีชี้วัดไม่เหมาะสม ทำให้ไม่สามารถใช้สำหรับการชี้นำ หรือบ่งบอกเหตุการณ์ที่อาจก่อให้เกิดขึ้นในอนาคต
- ไม่มีการนำข้อมูลที่ได้จากดัชนีชี้วัดมาประกอบการบริหาร เพื่อผลักดันให้เกิดการปรับปรุงอย่างต่อเนื่อง
- ในการสร้างดัชนีชี้วัดส่วนใหญ่เน้นที่ผลลัพธ์มากกว่ากระบวนการในการสร้างดัชนีชี้วัด

6.5 ปัจจัยที่ทำให้การใช้ KPI ประสบความสำเร็จ

- ความมุ่งมั่นของผู้บริหารในการสร้างดัชนีชี้วัด
- ใช้โปรแกรมคอมพิวเตอร์รวบรวม ประมวลผล วิเคราะห์ข้อมูล แสดงผล และกระตุ้นเตือนผู้รับผิดชอบดัชนีชี้วัดนั้นๆ
- กำหนดเงื่อนไขการให้คะแนนดัชนีชี้วัดแต่ละตัวให้อยู่บนพื้นฐานที่สามารถนำไปใช้ในการเปรียบเทียบผลงานที่เกิดขึ้นได้
- ประยุกต์ใช้ดัชนีชี้วัดในการบริหารเพื่อผลักดันให้เกิดการปรับปรุงอย่างต่อเนื่อง
- เชื่อมโยงผลงานที่ได้จากดัชนีชี้วัดกับการประเมินผลการปฏิบัติงาน

7. เครื่องมือที่ใช้ในการทำเหมืองข้อมูล

ภาพที่ 11 หน้าจอหลักของโปรแกรม Weka

Weka (Waikato Environment for Knowledge Analysis) เป็นโปรแกรมตัวหนึ่งที่นิยมใช้ในการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล โปรแกรม Weka ได้รวบรวมเทคนิคที่ใช้ในงาน Machine Learning และอัลกอริทึมการทำเหมืองข้อมูลไว้หลายแขนง เช่น Decision Tree, Neural Network, Naivebayes, EM algorithm, SimpleKmeans, Hierarchical Clustering, Association Rules เป็นต้น โปรแกรมนี้พัฒนาและเผยแพร่โดยทีมวิจัยจากมหาวิทยาลัย Waikato ประเทศนิวซีแลนด์ พัฒนาด้วยภาษา Java และเป็นโปรแกรมบนระบบเปิดเผยซอร์สโค้ด (Open source) สามารถใช้งานได้โดยไม่เสียค่าใช้จ่ายภายใต้ข้อตกลง GNU General Public License และนำไปใช้งานได้หลายแพล็ตฟอร์ม (Platform) เช่น Windows, Linux, MAC OS เป็นต้น (Witten and Frank 2005 : 366)

นอกจากนี้โปรแกรม Weka ได้จัดเตรียมเครื่องมือในการทำเหมืองข้อมูลไว้หลายอย่าง เช่น เครื่องมือในการจัดเตรียมข้อมูล เครื่องมือในการสร้างตัวแบบ และเครื่องมือในการประเมินผล จากภาพที่ 11 การทำงานของโปรแกรมแบ่งออกเป็น 4 เมนูหลักผ่านหน้าจอการใช้งานแบบ GUI (Graphical User Interface) รายละเอียดโดยสังเขปของเมนูต่างๆ มีดังนี้

1. เมนู Explorer เป็นส่วนที่เหมาะสำหรับผู้เริ่มต้นใช้งาน โดยสามารถเรียกใช้ฟังก์ชั่นการทำงานต่างๆ ของโปรแกรมผ่านทางหน้าจอ GUI และมีขั้นตอนการใช้งานที่ง่าย เช่น สามารถดูค่าทางสถิติของพารามิเตอร์ในชุดข้อมูลได้ทางหน้าจอ เป็นต้น

2. เมนู Experimenter เป็นส่วนที่ผู้ใช้สามารถทดลองปรับเปลี่ยนค่าพารามิเตอร์ต่างๆ เพื่อค้นหาคำตอบที่ดีที่สุดสำหรับคำถามหรือปัญหาที่ต้องการ โดยสามารถทำการเปรียบเทียบ

ผลลัพธ์ที่ได้ในแต่ละเทคนิคได้ เช่น เปรียบเทียบผลลัพธ์ระหว่างเทคนิค Classification กับ Regression เป็นต้น

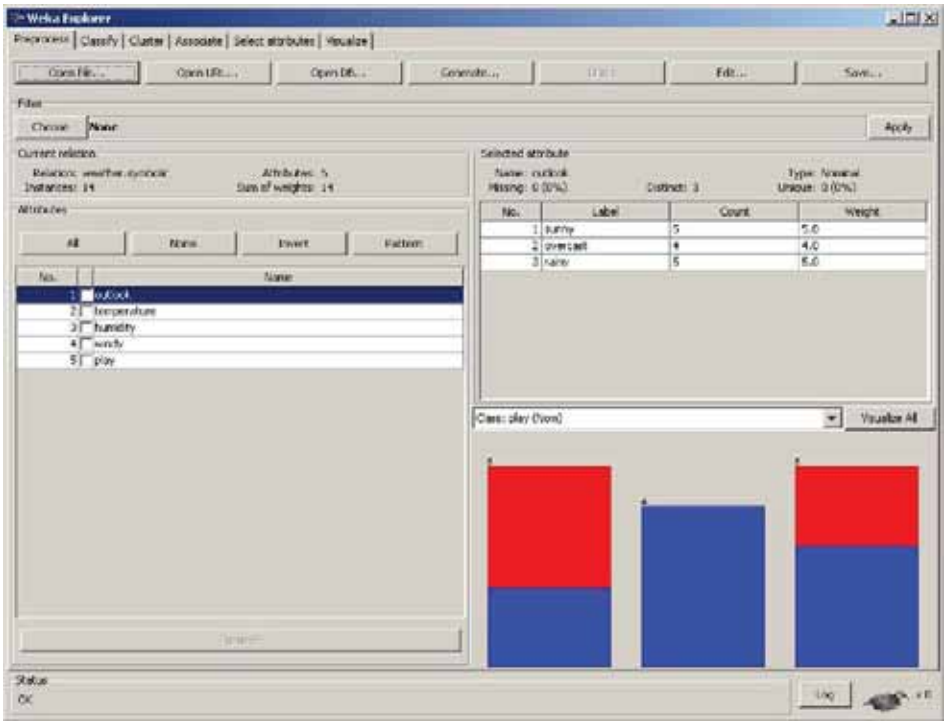
3. เมนู KnowledgeFlow เป็นส่วนที่ผู้ใช้สามารถออกแบบการไหลของข้อมูลร่วมกับเทคนิคเหมืองข้อมูลในส่วนที่เมนู Explorer ไม่ได้จัดเตรียมไว้ การออกแบบจะเป็นในลักษณะ drag and drop เครื่องมือ (Components) และนำมาผูกกันเป็นกระบวนการ (Process) ทำงานแบบอัติโนมัติได้

4. เมนู Simple CLI เป็นส่วนที่ผู้ใช้สามารถเรียกใช้ฟังก์ชั่นการทำงานผ่านทาง Command line ได้ ซึ่งช่วยให้ผู้ใช้สามารถเข้าใจการทำงานของฟังก์ชั่นที่อยู่เบื้องหลังหน้าจอ GUI และยังสามารถนำไปประยุกต์ใช้ในการเขียนโปรแกรมเพื่อเรียกใช้ฟังก์ชั่นจากโปรแกรม Weka ได้อีกด้วย

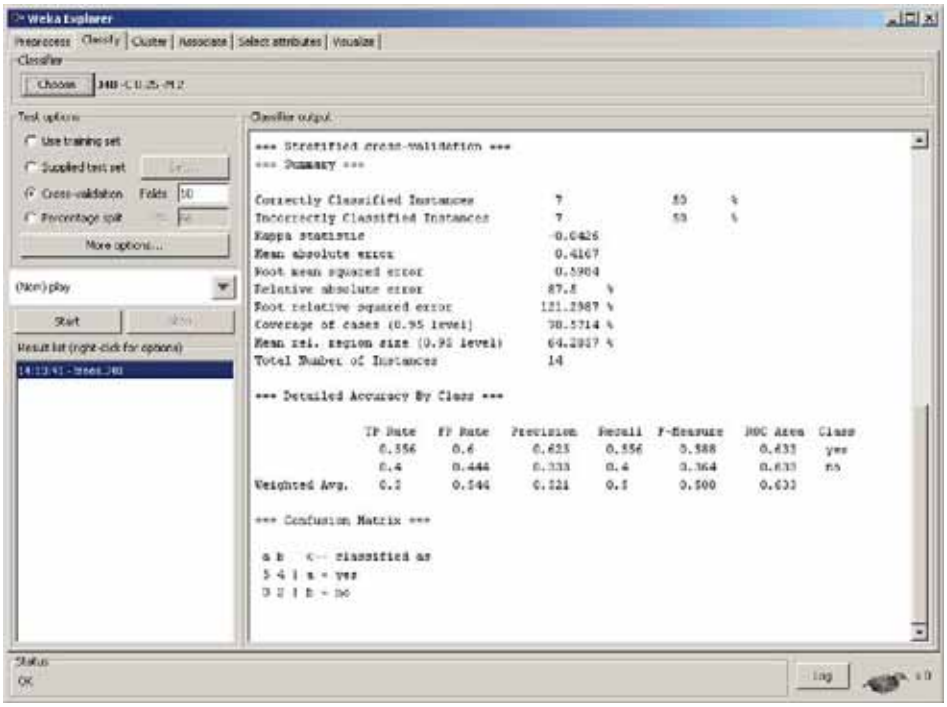
นอกจากนี้ผู้ใช้ยังสามารถดูค่าพารามิเตอร์หรือผลลัพธ์ในรูปแบบ Visualize ซึ่งเป็นการ Plot จุดข้อมูลในรูปแบบ 2 มิติได้ สำหรับประเภทของไฟล์ข้อมูลที่โปรแกรม Weka รองรับมีหลายรูปแบบ (Format) เช่น ARFF (Attribute-Relation File Format), CSV (Comma-Separated Value) ฯลฯ ตัวอย่างการใช้งานเมนู Explorer ในการวิเคราะห์ชุดข้อมูลประกอบการตัดสินในการเล่นกอล์ฟด้วยอัลกอริทึม C4.5 แสดงได้ดังภาพที่ 12 – 14 ในตัวอย่างภาพที่ 12 ที่ส่วนบนสุดของหน้าจอจะแสดงแท็บตามกลุ่มเทคนิคเหมืองข้อมูลที่สามารถใช้งานในโปรแกรม Weka

สำหรับแท็บ Preprocess จะแบ่งออกเป็น 5 ส่วน คือ

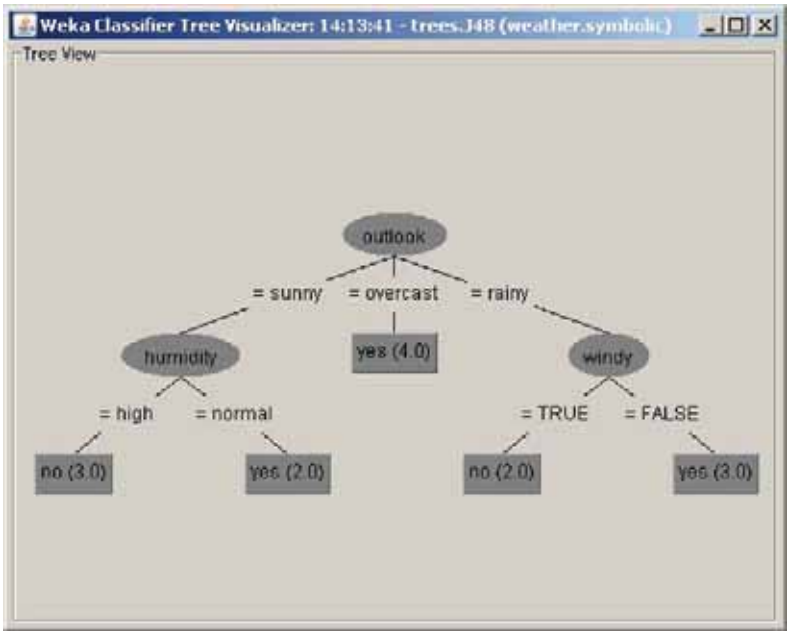
- ส่วนที่ 1 คือ Filter เป็นส่วนที่มีฟังก์ชั่นต่างๆ ที่ทำให้เราจัดการข้อมูลให้มีความถูกต้องมากขึ้น เช่น การแทนค่าข้อมูลที่หายไป (Missing value) ในชุดข้อมูล
- ส่วนที่ 2 คือ Current relation จะเป็นส่วนสรุปรายละเอียดของข้อมูลที่เรานำเข้าระบบ
- ส่วนที่ 3 คือ Attributes เป็นส่วนของการกำหนดแอตทริบิวท์ที่ต้องการใช้ในอัลกอริทึม
- ส่วนที่ 4 คือ Selected attribute เป็นส่วนที่แสดงรายละเอียดค่าสถิติของแอตทริบิวท์ เช่น ค่าเฉลี่ย (Mean)
- ส่วนที่ 5 คือ Visualization เป็นส่วนที่แสดงกราฟฮิสโทแกรมของแอตทริบิวท์



ภาพที่ 12 ตัวอย่างหน้าจอการนำเข้าข้อมูลในโปรแกรม Weka

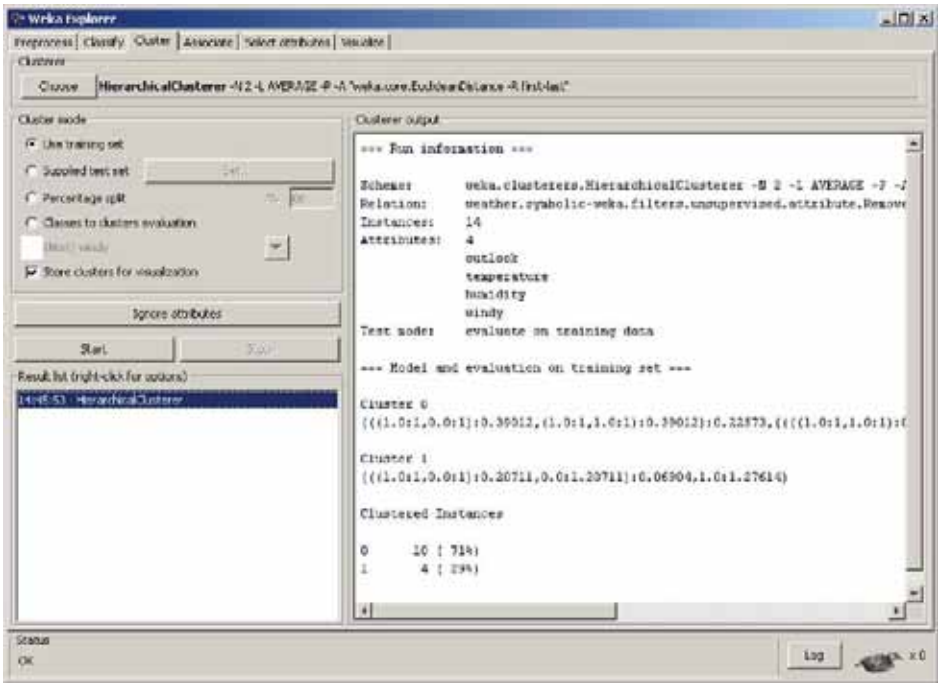


ภาพที่ 13 ตัวอย่างผลลัพธ์ที่ได้จากอัลกอริทึม C4.5



ภาพที่ 14 ตัวอย่างการแสดงผลลัพธ์ในรูปโครงสร้างต้นไม้ตัดสินใจ

นอกจากนี้เรายังสามารถจัดกลุ่มข้อมูลดังกล่าวด้วยเทคนิค Hierarchical Clustering ดังแสดงตัวอย่างผลลัพธ์ในภาพที่ 15



ภาพที่ 15 ตัวอย่างการแสดงผลลัพธ์การจัดกลุ่มด้วย Hierarchical Clustering

บทที่ 3
วิธีการดำเนินงานวิจัย

1. การเตรียมข้อมูล

ข้อมูลที่ใช้ในโครงงานนี้นำมาจากฐานข้อมูลของ บริษัท อาพิค ยามาดะ (ประเทศไทย) จำกัด โดยคัดเลือกเฉพาะข้อมูลที่สอดคล้องกับดัชนีชี้วัดผลการปฏิบัติงาน (KPI) จากแผนกที่เกี่ยวข้องกับการคัดเลือกแม่พิมพ์จำนวน 3 แผนก คือ แผนกผลิต แผนกขาย และแผนกซ่อมบำรุงแม่พิมพ์

1.1. แหล่งที่มาของข้อมูล

ตัวอย่างข้อมูลที่ใช้ในการศึกษานำมาจาก 2 แหล่งข้อมูล คือ

1.1.1 ข้อมูลการผลิตและการขาย นำมาจากฐานข้อมูล Progress 9.1C ของระบบ ERP ระหว่างเดือน ม.ค. 2550 – ธ.ค. 2552 จำนวน 705,376 ระเบียน

1.1.2 ข้อมูลการซ่อมบำรุงรักษาแม่พิมพ์ นำมาจากฐานข้อมูล Microsoft SQL Server 2005 ของระบบซ่อมบำรุงรักษาแม่พิมพ์ โดยคัดเลือกเฉพาะแม่พิมพ์ที่มีประวัติการซ่อมบำรุงระหว่างเดือน ม.ค. 2550 – ธ.ค. 2552 จำนวน 2,141 ระเบียน

1.2 การคัดเลือกดัชนีชี้วัดผลการปฏิบัติงาน (KPI)

คัดเลือกดัชนีชี้วัดผลการปฏิบัติงานที่เป็นดัชนีชี้วัดหลักจากแผนกที่เกี่ยวข้องกับการคัดเลือกแม่พิมพ์แผนกละหนึ่งดัชนีชี้วัด เพื่อใช้เป็นเกณฑ์กำหนดเงื่อนไขการจัดกลุ่มข้อมูลสำหรับทดสอบอัลกอริทึม ดัชนีชี้วัดหลักจากแผนกที่เกี่ยวข้องที่ถูกคัดเลือก คือ

1.2.1 แผนกผลิต ดัชนีชี้วัดหลัก คือ เปอร์เซ็นต์ยอดการผลิต (Production yield) ต้องมากกว่าหรือเท่ากับ 95% จึงจะถือว่าประสบผลสำเร็จ โดยรวมทั้งงานดีและงานเสียที่เกิดจากกระบวนการผลิต

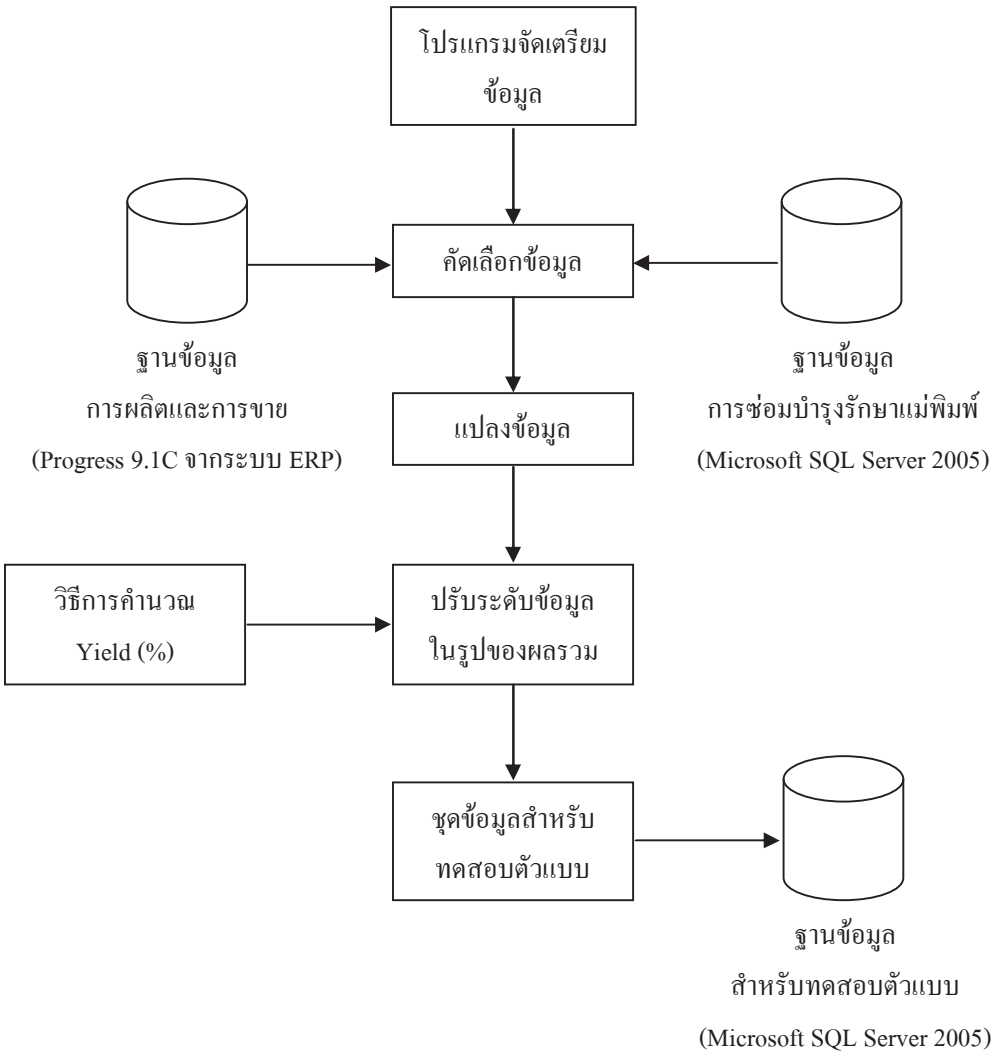
1.2.2 แผนกขาย ดัชนีชี้วัดหลัก คือ ยอดขาย ซึ่งก็คือเปอร์เซ็นต์ยอดงานดี (Final yield) ที่หักยอดของเสียออกแล้วต้องมากกว่าหรือเท่ากับ 95% จึงจะถือว่าประสบผลสำเร็จ (ยอดงานที่ผ่านการตรวจจากแผนก QC)

1.2.3 แผนกซ่อมบำรุงแม่พิมพ์ ดัชนีชี้วัดหลัก คือ การบำรุงรักษาแม่พิมพ์ตามจำนวนการผลิตที่กำหนด โดยจะกำหนดไว้เป็นช่วงจำนวนการผลิตขั้นต่ำ (Minimum Cycle Grinding) และจำนวนการผลิตขั้นสูง (Maximum Cycle Grinding) เช่น แม่พิมพ์ A มีอายุการผลิตชิ้นงานทั้งสิ้น 1 ล้านหน่วย และกำหนดจำนวนการผลิตขั้นต่ำและขั้นสูงในแต่ละรอบการผลิตที่

ต้องนำแม่พิมพ์มาบำรุงรักษาไว้ที่ 10,000–13,000 หน่วย ถ้าแม่พิมพ์ใดสามารถผลิตชิ้นงานจนถึงช่วงที่กำหนด หรือสามารถผลิตชิ้นงานได้มากกว่าช่วงที่กำหนดถือว่าประสบผลสำเร็จ

1.3 การแปลงข้อมูลและปรับระดับข้อมูล

เนื่องจากข้อมูลที่ใช้ในการทดสอบนำมาจาก 2 แหล่งข้อมูล ซึ่งมีโครงสร้างฐานข้อมูลที่แตกต่างกัน ผู้วิจัยได้พัฒนาโปรแกรมจัดเตรียมข้อมูลด้วย Microsoft Visual Basic 6 และพัฒนาระบบฐานข้อมูลด้วย Microsoft SQL Server 2005 เพื่อใช้จัดเก็บข้อมูลที่ถูกจัดเตรียมไว้สำหรับทดสอบอัลกอริทึม ข้อมูลที่ถูกคัดเลือกจะได้รับการปรับระดับข้อมูลให้อยู่ในรูปของผลรวมที่สอดคล้องกับลำดับที่ของประวัติการซ่อมบำรุงรักษาแม่พิมพ์ โดยข้อมูลที่ได้รับการปรับระดับแล้วมีจำนวนทั้งสิ้น 2,141 ระเบียน ขั้นตอนการแปลงข้อมูลและปรับระดับข้อมูลแสดงดังภาพที่ 16



ภาพที่ 16 ขั้นตอนการแปลงข้อมูลและปรับระดับข้อมูล

1.3.1 การคัดเลือกข้อมูล

- คัดเลือกข้อมูลประวัติการซ่อมบำรุงรักษาแม่พิมพ์ ระหว่างเดือน ม.ค. 2550 – ธ.ค. 2552 จากฐานข้อมูล Microsoft SQL Server 2005 ของระบบซ่อมบำรุงรักษาแม่พิมพ์

- คัดเลือกข้อมูลการผลิตและการขาย ระหว่างเดือน ม.ค. 2550 – ธ.ค. 2552 ที่สอดคล้องกับลำดับที่ในการซ่อมบำรุงรักษาแม่พิมพ์จากฐานข้อมูล Progress 9.1C ของระบบ ERP

1.3.2 การทำความสะอาดข้อมูลและการแปลงข้อมูล

- ลบข้อมูลแม่พิมพ์ที่ไม่มีประวัติการซ่อมบำรุงรักษาแม่พิมพ์

- แก้ไขลำดับที่ประวัติการซ่อมบำรุงรักษาแม่พิมพ์ที่จัดลำดับผิด

- แปลงรูปแบบข้อมูล (Data type) ที่ไม่เหมือนกันให้อยู่ในรูปแบบเดียวกัน

1.3.3 การปรับระดับข้อมูล

- คำนวณเปอร์เซ็นต์ดัชนีชี้วัดการผลิต (Production yield)

- คำนวณเปอร์เซ็นต์ดัชนีชี้วัดการขาย (Final yield)

- คำนวณค่าผลรวมเปอร์เซ็นต์ดัชนีชี้วัดการผลิต (Production yield) และเปอร์เซ็นต์ดัชนีชี้วัดการขาย (Final yield) ที่สอดคล้องกับลำดับที่ในการซ่อมบำรุงรักษาแม่พิมพ์ตามประวัติการซ่อมบำรุงรักษาแม่พิมพ์

- ตัวอย่างแอตทริบิวท์ที่ถูกคัดเลือกจากตารางข้อมูลยอดการผลิตและการซ่อมบำรุงแม่พิมพ์จำนวน 9 แอตทริบิวท์ที่ใช้ในการทดสอบอัลกอริทึม แสดงดังตารางที่ 8

ตารางที่ 8 ตัวอย่างแอตทริบิวท์ที่ใช้ในการทดสอบอัลกอริทึม

ลำดับที่	ชื่อรายการข้อมูล	ชนิดข้อมูล	ขนาด	รายละเอียด	ตัวอย่างข้อมูล
1	toolItem	ตัวอักษร	80	รหัสแม่พิมพ์ที่ใช้ในระบบการผลิต	7L1512DLS0304502-602
2	toolId	ตัวอักษร	40	รหัสแม่พิมพ์ที่ใช้ในระบบซ่อมบำรุงแม่พิมพ์	DLS0304502-602
3	toolOption	ตัวอักษร	40	รหัสแม่พิมพ์ย่อย	DLS0304502
4	minCG	ตัวเลข	8	จำนวนการผลิตขั้นต่ำที่ต้องบำรุงรักษาแม่พิมพ์	0.9
5	maxCG	ตัวเลข	8	จำนวนการผลิตขั้นสูงที่ต้องบำรุงรักษาแม่พิมพ์	1.4

ตารางที่ 8 (ต่อ)

ลำดับที่	ชื่อรายการข้อมูล	ชนิดข้อมูล	ขนาด	รายละเอียด	ตัวอย่างข้อมูล
6	productionOutput	ตัวเลข	8	ผลรวมของยอดการผลิตตามรอบการซ่อมบำรุงรักษาแม่พิมพ์	1.59
7	qcOutput	ตัวเลข	8	ผลรวมของยอดงานดีตามรอบการซ่อมบำรุงรักษาแม่พิมพ์	1.48
8	productionYield	ตัวเลข	8	อัตราผลผลิตทั้งหมดที่เกิดจากกระบวนการผลิต (เปอร์เซ็นต์)	92.55
9	finalYield	ตัวเลข	8	อัตราผลผลิตของยอดงานดีที่เกิดจากกระบวนการผลิต (เปอร์เซ็นต์)	86.31

2. การจัดกลุ่มแม่พิมพ์ตามดัชนีชี้วัดผลการปฏิบัติงาน (KPI)

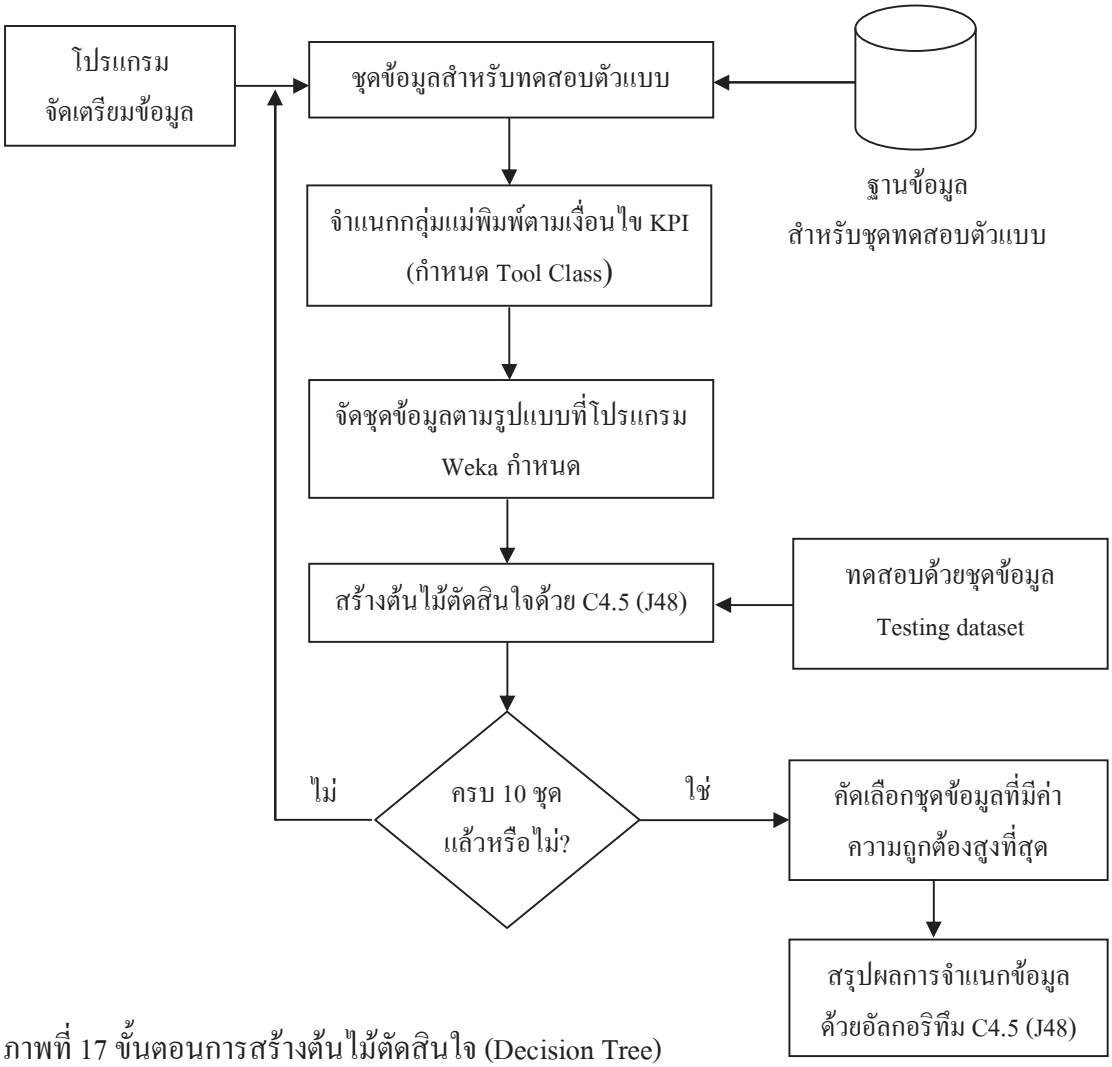
ผู้วิจัยได้กำหนดเงื่อนไขการจัดกลุ่มแม่พิมพ์ให้สอดคล้องกับดัชนีชี้วัดผลการปฏิบัติงาน (KPI) เพื่อใช้ในการทดสอบอัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยนำชุดข้อมูลที่จัดเตรียมไว้มาทดสอบตามเงื่อนไขเพื่อจำแนกกลุ่มแม่พิมพ์ออกเป็น 3 กลุ่ม (Class) คือ Good, Normal และ Bad ตัวอย่างเงื่อนไขการจัดกลุ่มแม่พิมพ์แสดงในตารางที่ 9

ตารางที่ 9 เงื่อนไขการจัดกลุ่มแม่พิมพ์ตามดัชนีชี้วัดผลการปฏิบัติงาน (KPI)

การดูแลรักษาแม่พิมพ์	ยอดการผลิต	ยอดงานดี
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)

3. การทดสอบการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ

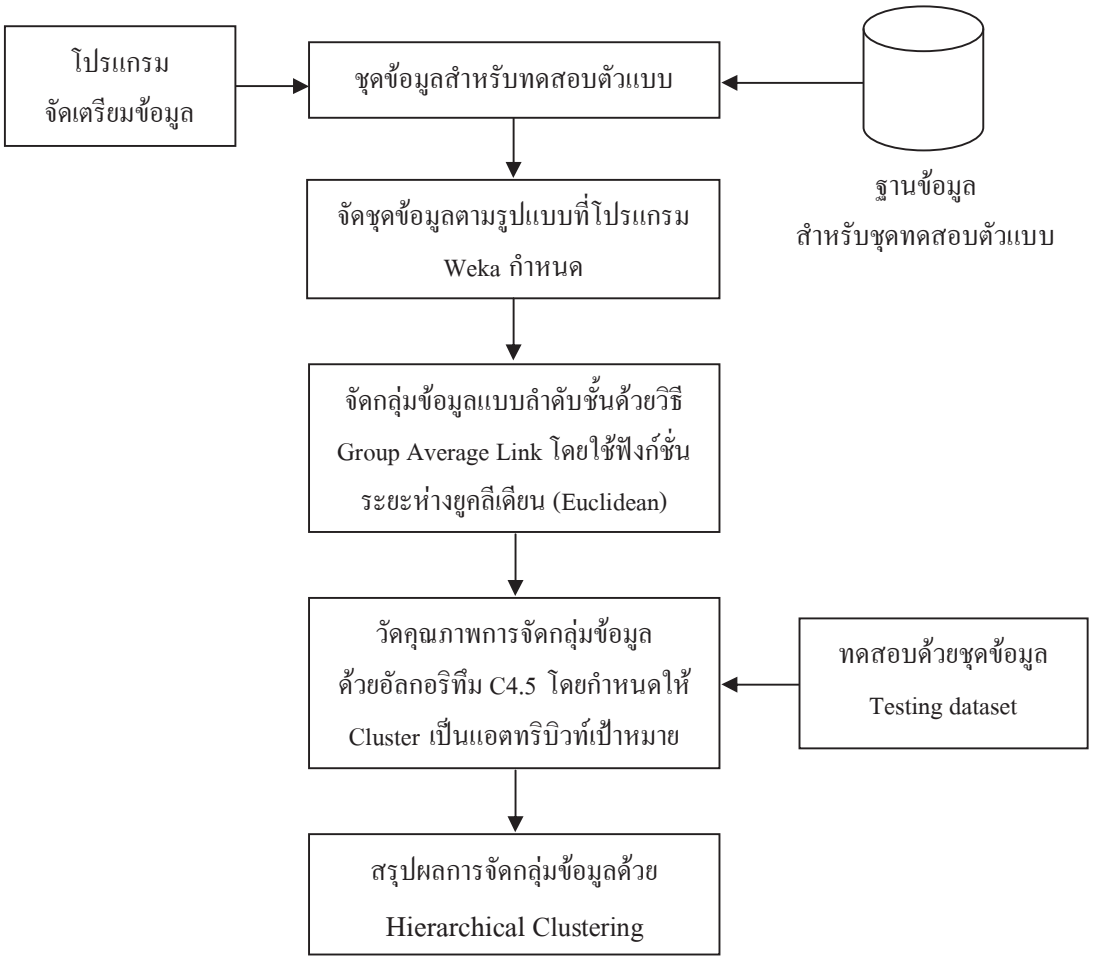
เพื่อให้การจำแนกข้อมูลมีค่าความถูกต้อง (Accuracy rate) สูงสุด และเพื่อให้ได้ต้นไม้ตัดสินใจที่ดี ผู้วิจัยนำชุดข้อมูลที่จัดเตรียมไว้มาจำแนกกลุ่มแม่พิมพ์ในเงื่อนไขต่างๆ จำนวน 10 รูปแบบ และนำมาสร้างเป็นชุดข้อมูลสำหรับทดสอบอัลกอริทึมจำนวน 10 ชุด จากนั้นนำข้อมูลแต่ละชุดมาทดสอบการจำแนกข้อมูลด้วยโปรแกรม Weka โดยใช้โมดูล J48 ซึ่งเป็นการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เพื่อคัดเลือกรูปแบบเงื่อนไขการจำแนกกลุ่มแม่พิมพ์ที่ให้ค่าความถูกต้องสูงที่สุด โดยกำหนดให้กลุ่มแม่พิมพ์ (Tool class) เป็นแอตทริบิวท์เป้าหมาย ในการแบ่งข้อมูลสำหรับเรียนรู้ (Training dataset) และข้อมูลสำหรับทดสอบ (Testing dataset) ผู้วิจัยใช้วิธีกำหนดความหลากหลายแบบสุ่ม (n-Fold Cross Validation) โดยกำหนดจำนวนครั้งที่สลับ (Fold) เท่ากับ 10 ซึ่งเป็นค่ามาตรฐานที่นิยมใช้โดยทั่วไป สำหรับขั้นตอนการสร้างต้นไม้ตัดสินใจเพื่อจำแนกกลุ่มแม่พิมพ์แสดงดังภาพที่ 17



ภาพที่ 17 ขั้นตอนการสร้างต้นไม้ตัดสินใจ (Decision Tree)

4. การทดสอบการจัดกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น

ผู้วิจัยทำการทดสอบการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) โดยใช้วิธีวัดค่าความเหมือนแบบเฉลี่ย (Group Average Link) ด้วยฟังก์ชันระยะห่างยูคลีเดียน (Euclidean function) จากโปรแกรม Weka โดยกำหนดจำนวนกลุ่มที่ต้องการจำนวน 3 กลุ่ม (Cluster) จากนั้นทำการวัดคุณภาพการจัดกลุ่มข้อมูล (Cluster Evaluation) ด้วยอัลกอริทึม C4.5 ซึ่งเป็นวิธีการสร้างต้นไม้ตัดสินใจ โดยกำหนดให้กลุ่มข้อมูล (Cluster) ที่ได้จากการจัดกลุ่มแบบลำดับชั้นเป็นแอตทริบิวท์ (Class) เป้าหมาย ในการวัดคุณภาพผู้วิจัยใช้วิธีกำหนดความหลากหลายแบบสุ่ม (n-Fold Cross Validation) ในการแบ่งข้อมูลสำหรับเรียนรู้ (Training dataset) และข้อมูลสำหรับทดสอบ (Testing dataset) โดยกำหนดจำนวนครั้งที่สลับ (Fold) เท่ากับ 10 ซึ่งเป็นค่ามาตรฐานที่นิยมใช้โดยทั่วไป ขั้นตอนการจัดกลุ่มแม่พิมพ์แบบลำดับชั้น (Hierarchical Clustering) แสดงดังภาพที่ 18



ภาพที่ 18 ขั้นตอนการจัดกลุ่มแม่พิมพ์แบบลำดับชั้น (Hierarchical Clustering)

5. การประเมินผลอัลกอริทึม

ในโครงงานนี้ต้นไม้การตัดสินใจถูกสร้างขึ้นเพื่อศึกษารูปแบบค่าของแอตทริบิวท์ในกลุ่มเป้าหมายที่ถูกจำแนกเป็น 3 กลุ่ม (Class) โดยผู้วิจัยใช้ Success rate เป็นค่าชี้วัดคุณภาพการจำแนกโดยรวม และใช้ True Positive rate (TP), False Positive rate (FP) และ Precision เป็นค่าชี้วัดคุณภาพการจำแนกคลาสในแต่ละคลาส ค่าเหล่านี้คำนวณได้จาก Confusion Matrix ดังภาพที่ 19

$$success\ \ rate = \left(\sum_{i=1}^{n} f_{ii} \right) / total$$

$$TP(a) = f_{aa} / actual(a)$$

$$FP(a) = (predict(a) - f_{aa}) / (total - actual(a))$$

$$precision(a) = f_{aa} / predict(a)$$

$TP(a)$ หรือ True Positive rate บอกความถูกต้องในการจำแนกอินสแตนซ์มายังคลาส a ในขณะที่ $FP(a)$ หรือ False Positive rate บอกความเอนเอียงในการจำแนกอินสแตนซ์มายังคลาส a ซึ่งทำให้เกิดสัญญาณหลอก (False alarm) ส่วน $precision(a)$ เป็นค่าความเที่ยงในการจำแนกคลาส a โดยคุณสมบัติของคลาสที่เราต้องการคือ มีค่า $TP(a)$ และ $precision(a)$ สูง แต่มีค่า $FP(a)$ ต่ำ

คลาสที่แท้จริง (Actual Classes) \ คลาสที่จำแนกโดยคลาสซิฟายเออร์ (Predicted Classes)	a	b	c	
a	f_{aa}	f_{ab}	f_{ac}	$\sum_{k=1}^{n} f_{ak}$
b	f_{ba}	f_{bb}	f_{bc}	
c	f_{ca}	f_{cb}	f_{cc}	
	$\sum_{i=1}^{n} f_{ia}$			$total$

ภาพที่ 19 Confusion Matrix ในการจำแนกคลาส a, b และ c

f_{ac} เป็นจำนวนอินสแตนซ์ที่มีคลาสที่แท้จริง คือ คลาส a และถูกจำแนกให้อยู่ในคลาส c

$total$ เป็นจำนวนอินสแตนซ์ทั้งหมด

$\sum_{k=1}^{n} f_{ak}$ หรือ $actual(a)$ เป็นจำนวนอินสแตนซ์ทั้งหมดที่มีคลาสที่แท้จริง คือ คลาส a

$\sum_{i=1}^{n} f_{ia}$ หรือ $predict(a)$ เป็นจำนวนอินสแตนซ์ทั้งหมดที่ถูกจำแนกให้อยู่ใน คลาส a

การวิเคราะห์และประเมินผลเพื่อคัดเลือกอัลกอริทึมที่เหมาะสมในการคัดเลือกกลุ่มแม่พิมพ์ ระหว่างอัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ผู้วิจัยใช้วิธีวัดค่าความคลาดเคลื่อน (Error Measurement) ในการประเมินประสิทธิภาพของอัลกอริทึม ซึ่งอัลกอริทึมที่เหมาะสม คือ อัลกอริทึมที่มีค่าความคลาดเคลื่อนต่ำที่สุด โดยพิจารณาเปรียบเทียบจากค่าสถิติต่างๆ จำนวน 3 วิธี คือ

5.1 Root Mean Squared Error (RMSE)

มีสูตรคำนวณดังนี้ คือ

$$RMSE = \sqrt{\sum \frac{(p-a)^2}{n}}$$

5.2 Mean Absolute Error (MAE)

มีสูตรคำนวณดังนี้ คือ

$$MAE = \frac{\sum |p-a|}{n}$$

5.3 Relative Absolute Error (RAE)

มีสูตรคำนวณดังนี้ คือ

$$RAE = \frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \overline{a}| + \ldots + |a_n - \overline{a}|}$$

โดยที่ p = ค่าที่ทำนายได้

a = ค่าที่วัดได้

n = จำนวนข้อมูล

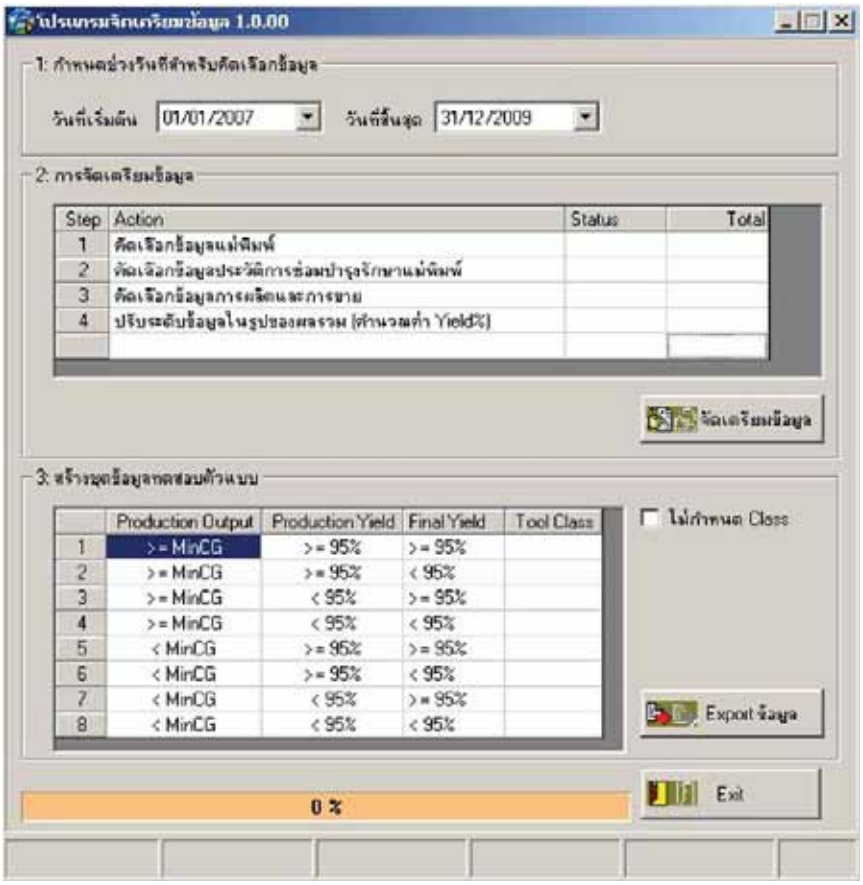
6. การสรุปและรายงานผล

ผู้วิจัยจะสรุปผลการทดสอบอัลกอริทึมในรูปแบบของข้อความ รูปภาพ ตาราง และกราฟ โดยข้อมูลกราฟจะสร้างจากโปรแกรม Microsoft Excel ซึ่งเป็นเครื่องมือที่สามารถทำความเข้าใจและใช้งานได้ง่าย โดยจะเปรียบเทียบให้เห็นถึงความถูกต้องของค่าต่างๆ ที่ได้จากการดำเนินงานวิจัย

บทที่ 4
ผลการดำเนินงาน

ในการดำเนินงานวิจัยเพื่อคัดเลือกอัลกอริทึมที่เหมาะสมสำหรับการจัดกลุ่มแม่พิมพ์ ผู้วิจัยได้พัฒนาโปรแกรมจัดเตรียมข้อมูล เพื่อใช้ในการจัดเตรียมชุดข้อมูลให้อยู่ในรูปแบบตามที่โปรแกรม Weka กำหนด โดยพัฒนาโปรแกรมด้วย Microsoft Visual Basic 6 และใช้ฐานข้อมูล Microsoft SQL Server 2005 ในการจัดเก็บข้อมูล ผลการดำเนินงานแบ่งออกเป็น 4 ส่วน คือ ผลการพัฒนาโปรแกรมจัดเตรียมข้อมูล ผลการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) ผลการจัดกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) และผลการประเมินและคัดเลือกอัลกอริทึมที่เหมาะสม รายละเอียดของผลการดำเนินงานมีดังนี้

1. ผลการพัฒนาโปรแกรมจัดเตรียมข้อมูล



ภาพที่ 20 หน้าจอโปรแกรมจัดเตรียมข้อมูล

จากภาพที่ 20 หน้าจอการทำงานของโปรแกรมจัดเตรียมข้อมูลแบ่งออกเป็น 3 ส่วน คือ ส่วนแรกสำหรับกำหนดช่วงวันที่ที่ต้องการคัดเลือกข้อมูล ส่วนที่สอง คือ กระบวนการคัดเลือกและปรับระดับข้อมูล ส่วนที่สาม คือ กระบวนการสร้างชุดข้อมูลสำหรับทดสอบอัลกอริทึมตามรูปแบบที่โปรแกรม Weka กำหนด

ในกระบวนการคัดเลือกและปรับระดับข้อมูล จะประกอบด้วย 4 กระบวนการย่อย คือ กระบวนการคัดเลือกข้อมูลแม่พิมพ์ กระบวนการคัดเลือกข้อมูลประวัติการซ่อมบำรุงรักษาแม่พิมพ์ กระบวนการคัดเลือกข้อมูลการผลิตและการขาย และกระบวนการสุดท้าย คือ กระบวนการคำนวณเปอร์เซ็นต์ยอดการผลิต (Production yield) และเปอร์เซ็นต์ยอดงานดี (Final yield) เพื่อปรับระดับข้อมูลให้อยู่รูปของผลรวม โดยผู้ใช้งานสามารถทราบผลการทำงานของแต่ละกระบวนการได้จากคอลัมภ์ Status และสามารถทราบจำนวนข้อมูลที่ถูกคัดเลือกทั้งหมดได้จากคอลัมภ์ Total ซึ่งโปรแกรมจะแจ้งสถานะการทำงานใน 3 ลักษณะ คือ Processing (กำลังประมวลผล) Success (กระบวนการทำงานเสร็จสมบูรณ์) และ Failed (กระบวนการทำงานล้มเหลว)

ในกระบวนการสร้างชุดข้อมูลสำหรับทดสอบ จะแบ่งการทำงานออกเป็น 2 ลักษณะ คือ การสร้างชุดข้อมูลสำหรับทดสอบอัลกอริทึมการจำแนกข้อมูลด้วยโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และการสร้างชุดข้อมูลสำหรับทดสอบอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ในการสร้างชุดข้อมูลสำหรับทดสอบอัลกอริทึมโครงสร้างต้นไม้ตัดสินใจ ผู้ใช้งานจะต้องกำหนดกลุ่มแม่พิมพ์ (Tool Class) ในตารางตามเงื่อนไขดัชนีชี้วัดผลการปฏิบัติงาน (KPI) โดยไม่ต้องคลิกตัวเลือก “ไม่กำหนด Class” ในทางตรงกันข้ามถ้าต้องการสร้างชุดข้อมูลสำหรับทดสอบอัลกอริทึมการจัดกลุ่มแบบลำดับชั้น ผู้ใช้งานต้องคลิกตัวเลือก “ไม่กำหนด Class” และไม่ต้องกำหนดกลุ่มแม่พิมพ์ในตาราง จากนั้นกดปุ่ม “Export ข้อมูล” โปรแกรมจะบันทึกข้อมูลในรูปแบบของ CSV (Comma-Separated Value) ซึ่งเป็นรูปแบบข้อมูลรูปแบบหนึ่งที่สามารถใช้งานกับโปรแกรม Weka ได้ โดยรูปแบบที่นิยมใช้ทั่วไป คือ ARFF (Attribute-Relation File Format) สำหรับโครงสร้างตารางข้อมูลแสดงดังตารางที่ 10 ตัวอย่างชุดข้อมูลแสดงดังภาพที่ 21

ตารางที่ 10 สรุปโครงสร้างตารางข้อมูลโปรแกรมจัดเตรียมข้อมูล

ชื่อโครงสร้างตารางข้อมูล	**รายละเอียด**
ข้อมูลยอดการผลิตและการซ่อมบำรุงแม่พิมพ์	ภาคผนวก ก หน้าที่ 73 ตารางที่ 15
ข้อมูลการผลิตและการขายแยกตามคำสั่งผลิต	ภาคผนวก ก หน้าที่ 75 ตารางที่ 16

A	B	C	D	E	F	G	H	I	J
toolId	toolItem	toolOption	minCG	maxCG	productionOutput	qcOutput	productionYield	finalYield	toolClass
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	0.88	0.87	94.54	93.8	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	0.89	0.89	86.82	86.82	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.69	1.48	92.55	88.31	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.42	1.42	90.11	90.11	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.29	1.29	95.59	95.59	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.37	1.37	93.94	93.94	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.09	1.09	93.52	93.52	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.07	1.07	93.93	93.93	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.25	1.25	96.15	96.15	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.47	1.47	96.02	96.02	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.6	1.5	94.05	94.06	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	95.51	95.51	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	0.97	0.95	88.97	86.68	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.01	1.01	95.82	95.82	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.41	1.41	94.31	94.31	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.14	1.14	91.67	91.67	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.09	1.08	94.37	93.33	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.38	1.38	95.59	95.59	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	91.87	91.87	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.76	1.76	95.22	95.22	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	111.7	111.7	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.81	1.81	95.82	95.82	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	18.66	18.73	94.12	93.46	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.41	1.41	94.31	94.31	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.14	1.14	91.67	91.67	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.09	1.08	94.37	93.33	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.38	1.38	95.59	95.59	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	91.87	91.87	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.76	1.76	95.22	95.22	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	111.7	111.7	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.81	1.81	95.82	95.82	Good
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.89	1.81	96.25	91.01	Normal
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.95	1.95	94.66	94.66	Bad
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.84	1.84	95.78	95.78	Good
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.93	1.93	96.24	96.24	Good
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.62	1.5	95.54	88.72	Normal
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.56	1.56	88.35	88.35	Bad
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	2.06	1.95	96.72	91.26	Normal
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	2.18	2.06	86.61	81.65	Bad
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.8	1.8	95.89	95.89	Good
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.77	1.77	97.09	97.09	Good
DLS0404102	7L2091DLS0404102	DLS0404102	0.8	1.2	1.91	1.91	90.51	90.51	Bad

ภาพที่ 21 ตัวอย่างชุดข้อมูลสำหรับทดสอบอัลกอริทึมโครงสร้างต้นไม้ตัดสินใจ

2. ผลการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ

จากการนำชุดข้อมูลมาจำแนกกลุ่มแม่พิมพ์แบ่งเป็น 3 กลุ่ม (Class) คือ Good, Normal และ Bad ตามดัชนีชี้วัดผลการปฏิบัติงาน (KPI) ในเงื่อนไขต่างๆ จำนวน 10 ชุด จากนั้นทดสอบการจำแนกข้อมูลแต่ละชุดด้วยโปรแกรม Weka โดยใช้โมดูล J48 ซึ่งเป็นการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เพื่อคัดเลือกรูปแบบเงื่อนไขการจำแนกกลุ่มแม่พิมพ์ที่ให้ค่าความถูกต้องสูงที่สุด โดยกำหนดให้กลุ่มแม่พิมพ์ (Tool Class) เป็นแอตทริบิวท์เป้าหมาย ในการจำแนกผู้วิจัยใช้วิธีกำหนดความหลากหลายแบบสุ่ม (n-Fold Cross Validation) ในการแบ่งข้อมูลสำหรับเรียนรู้ (Training dataset) และข้อมูลสำหรับทดสอบ (Testing dataset) โดยกำหนดจำนวนครั้งที่สลับ (Fold) เท่ากับ 10 ซึ่งเป็นค่ามาตรฐานที่นิยมใช้โดยทั่วไป รายละเอียดการจำแนกกลุ่มแม่พิมพ์ตามดัชนีชี้วัดผลการปฏิบัติงาน (KPI) และค่าความถูกต้อง (Success rate) แสดงดังตารางที่ 11

ตารางที่ 11 รายละเอียดการจำแนกกลุ่มแม่พิมพ์และค่าความถูกต้อง (Success rate)

การดูแลรักษา (ยอดผลิต)	ยอดการผลิต	ยอดงานดี	รูปแบบการจำแนกกลุ่มข้อมูล (Class)									
			1	2	3	4	5	6	7	8	9	10
ง่าย (>= Min)	มาก (>= 95%)	มาก (>= 95%)	G	G	G	G	G	G	G	G	G	G
ง่าย (>= Min)	มาก (>= 95%)	น้อย (< 95%)	G	G	G	G	G	G	G	G	G	N
ง่าย (>= Min)	น้อย (< 95%)	มาก (>= 95%)	G	N	N	N	G	N	N	N	N	G
ง่าย (>= Min)	น้อย (< 95%)	น้อย (< 95%)	G	N	N	N	N	N	N	B	B	B
ยาก (< Min)	มาก (>= 95%)	มาก (>= 95%)	N	N	N	B	N	G	G	N	G	G
ยาก (< Min)	มาก (>= 95%)	น้อย (< 95%)	N	N	B	B	N	G	G	N	N	N
ยาก (< Min)	น้อย (< 95%)	มาก (>= 95%)	N	B	B	B	N	B	N	N	N	G
ยาก (< Min)	น้อย (< 95%)	น้อย (< 95%)	B	B	B	B	B	B	B	B	B	B
ค่าความถูกต้อง			94.02	93.13	92.39	92.62	93.13	96.73	96.73	95.98	99.49	**99.95**

หมายเหตุ

G หมายถึง Good, N หมายถึง Normal และ B หมายถึง Bad

เมื่อทดลองจำแนกกลุ่มแม่พิมพ์ด้วยเงื่อนไขต่างๆ ตามตารางที่ 11 ค่าความถูกต้อง (Success rate) ที่ได้จากข้อมูลชุดที่ 1 – 8 จะมีค่าอยู่ระหว่าง 92.39 - 96.73 เปอร์เซ็นต์ และชุดข้อมูลที่ 9 - 10 จะให้ค่าความถูกต้องที่ 99.49 และ 99.95 เปอร์เซ็นต์ตามลำดับ ซึ่งจะเห็นได้ว่าข้อมูลชุดที่ 10 ให้ค่าความถูกต้องสูงที่สุด คือ 99.95 เปอร์เซ็นต์ ดังนั้นผู้วิจัยจึงเลือกข้อมูลชุดที่ 10 เพื่อใช้สรุปและอธิบายผลลัพธ์จากการสร้างต้นไม้ตัดสินใจ และจากการทดลองจะสังเกตได้ว่าข้อมูลชุดที่ 2 กับข้อมูลชุดที่ 5 และข้อมูลชุดที่ 6 กับข้อมูลชุดที่ 7 มีค่าความถูกต้องเท่ากัน เนื่องจากผลการจำแนกกลุ่มแม่พิมพ์ในเงื่อนไขที่ต่างกันให้จำนวนตัวอย่างแม่พิมพ์ในแต่ละคลาสเท่ากัน โดยผลการจำแนกกลุ่มแม่พิมพ์ด้วยวิธีโครงสร้างต้นไม้ตัดสินใจข้อมูลชุดที่ 1- 9 แสดงรายละเอียดในตารางที่ 12

ตารางที่ 12 ผลสรุปการจำแนกกลุ่มแม่พิมพ์ด้วยอัลกอริทึม C4.5 ข้อมูลชุดที่ 1-9

ชุดข้อมูลที่ใช้ในการทดสอบ	รายละเอียดผลลัพธ์จากการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5
ข้อมูลชุดที่ 1	ภาคผนวก ข หน้าที่ 78
ข้อมูลชุดที่ 2	ภาคผนวก ข หน้าที่ 81
ข้อมูลชุดที่ 3	ภาคผนวก ข หน้าที่ 84
ข้อมูลชุดที่ 4	ภาคผนวก ข หน้าที่ 87
ข้อมูลชุดที่ 5	ภาคผนวก ข หน้าที่ 90
ข้อมูลชุดที่ 6	ภาคผนวก ข หน้าที่ 93
ข้อมูลชุดที่ 7	ภาคผนวก ข หน้าที่ 96
ข้อมูลชุดที่ 8	ภาคผนวก ข หน้าที่ 99
ข้อมูลชุดที่ 9	ภาคผนวก ข หน้าที่ 102

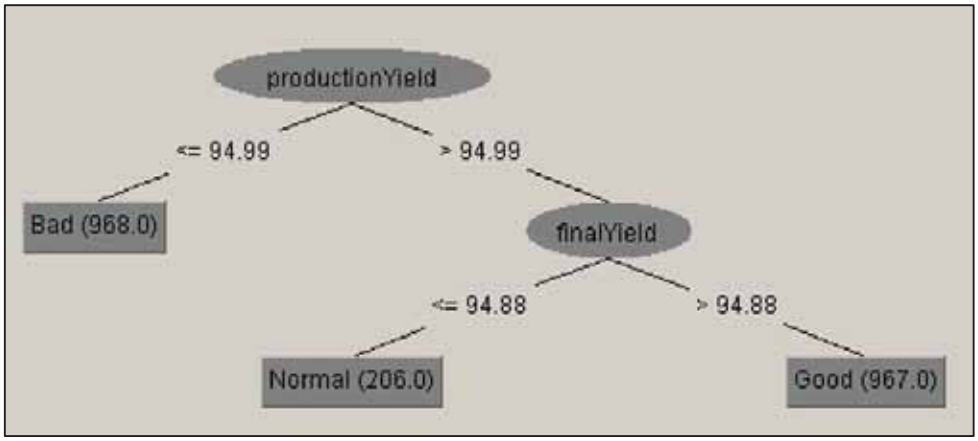
ผลการสร้างต้นไม้ตัดสินใจข้อมูลชุดที่ 10 ด้วยอัลกอริทึม C4.5 (โมดูล J48) จากโปรแกรม Weka แสดงรายละเอียดได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R10
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
```

```
Test mode:    10-fold cross-validation

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

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

productionYield <= 94.99: Bad (968.0)
productionYield > 94.99
|   finalYield <= 94.88: Normal (206.0)
|   finalYield > 94.88: Good (967.0)
Number of Leaves: 3
Size of the tree: 5
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2140               99.9533 %
Incorrectly Classified Instances         1                0.0467 %
Kappa statistic                          0.9992
Mean absolute error                      0.0003
Root mean squared error                  0.0176
Relative absolute error                  0.0802 %
Root relative squared error              4.0052 %
Coverage of cases (0.95 level)          99.9533 %
Mean rel. region size (0.95 level)      33.3333 %
Total Number of Instances             2141

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision  Recall  F-Measure   ROC Area  Class
1         0          1         1         1          1        Good
0.999     0          1         0.999     0.999      0.999    Bad
1         0.001      0.995     1         0.998      1
Normal

=== Confusion Matrix ===

   a   b   c   <-- classified as
 967   0   0 |   a = Good
   0 967   1 |   b = Bad
   0   0 206 |   c = Normal
```

ผลลัพธ์ที่ได้มีระเบียนที่จำแนกถูกต้องจำนวน 2,140 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 99.9533 % โดยจำแนกอยู่ในคลาส Good จำนวน 967 ระเบียน อยู่ในคลาส Normal จำนวน 206 ระเบียน และในคลาส Bad จำนวน 967 ระเบียน

โดยมีระเบียนที่จำแนกผิดจำนวน 1 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 0.0467 % สำหรับข้อมูล 1 ระเบียน ที่จำแนกผิดนั้น คือ มี 1 ระเบียน ที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Good



ภาพที่ 22 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ในข้อมูลชุดที่ 10

ภาพที่ 22 แสดงผลลัพธ์การจำแนกกลุ่มแม่พิมพ์ข้อมูลชุดที่ 10 ด้วยอัลกอริทึม C4.5 ในรูปแบบโครงสร้างต้นไม้ตัดสินใจ โดยต้นไม้ที่ได้มีโหนดใบทั้งสิ้น 3 โหนด ในจำนวนนี้มี 2 โหนดที่ใช้ยอดงานดี (Final yield) เป็นกฎการตัดสินใจ และอีก 1 โหนดใช้ยอดการผลิต (Production yield) เป็นกฎการตัดสินใจ ซึ่งสามารถสรุปได้ดังนี้

คลาส Good: กลุ่มแม่พิมพ์ที่ให้ยอดการผลิต (Production yield) มากกว่าเกณฑ์ปกติ

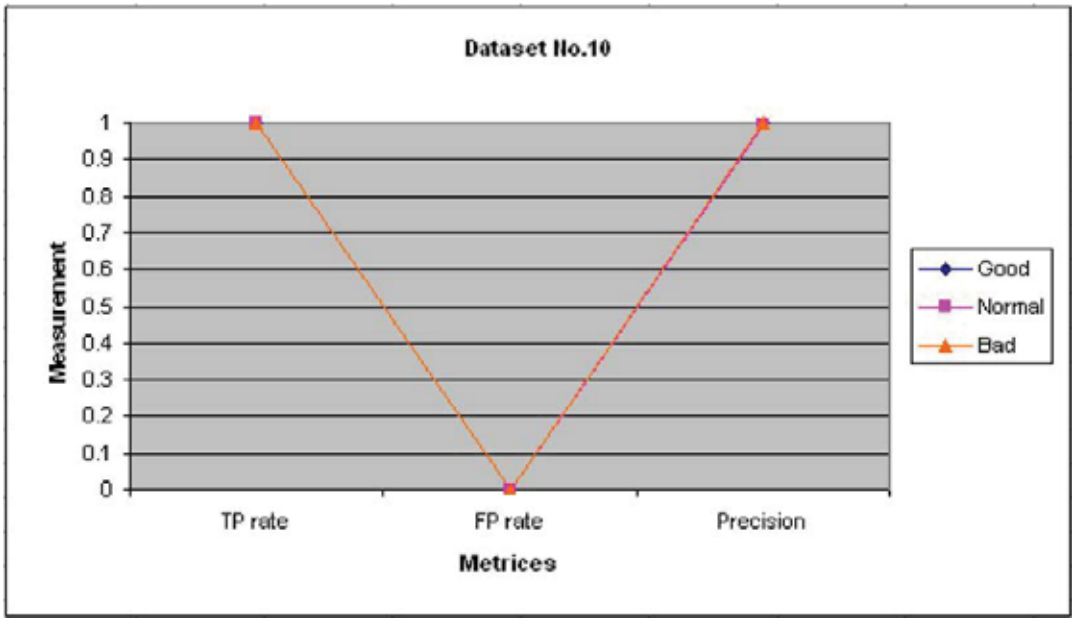
- เป็นแม่พิมพ์ที่ให้ยอดการผลิต (Production yield) มากกว่า 94.99%
- เป็นแม่พิมพ์ที่ให้ยอดงานดี (Final yield) มากกว่า 94.88%
- มีแม่พิมพ์เป็นสมาชิกในคลาส Good จำนวน 967 ระเบียน ครอบคลุม 45% ของข้อมูลทั้งหมด -> อัตราการจำแนกผิด 0.0

คลาส Normal: กลุ่มแม่พิมพ์ที่ให้ยอดการผลิต (Production yield) ในเกณฑ์ปกติ

- เป็นแม่พิมพ์ที่ให้ยอดการผลิต (Production yield) มากกว่า 94.99%
- เป็นแม่พิมพ์ที่ให้ยอดงานดี (Final yield) น้อยกว่าหรือเท่ากับ 94.88%
- มีแม่พิมพ์เป็นสมาชิกในคลาส Normal จำนวน 206 ระเบียน ครอบคลุม 10% ของข้อมูลทั้งหมด -> อัตราการจำแนกผิด 0.001

คลาส Bad: กลุ่มแม่พิมพ์ที่ให้ยอดการผลิต (Production yield) ต่ำกว่าเกณฑ์ปกติ

- เป็นแม่พิมพ์ที่ให้ยอดการผลิต (Production yield) น้อยกว่าหรือเท่ากับ 94.99%
- มีแม่พิมพ์เป็นสมาชิกในคลาส Bad จำนวน 968 ระเบียน ครอบคลุม 45% ของข้อมูลทั้งหมด -> อัตราการจำแนกผิด 0.0



ภาพที่ 23 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 10

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 23 เมื่อพิจารณาค่าดังกล่าว พบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Normal (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.001 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad โดยมีอัตราการจำแนกผิดเท่ากับ 0.0 และคลาส Good มีอัตราการจำแนกผิดเท่ากับ 0.0 เช่นกัน นอกจากนี้ยังพบว่าการจำแนกข้อมูลในแต่ละกลุ่มมีค่า TP rate และ Precision ที่สูง ดังนั้นจึงสรุปได้ว่าเงื่อนไขการจำแนกกลุ่มแม่พิมพ์ตามดัชนีชี้วัดผลการปฏิบัติงาน (KPI) ที่เลือกมานี้มีความสัมพันธ์กัน อีกทั้งตัวแบบที่ได้มีความน่าเชื่อถือและสามารถจำแนกกลุ่มแม่พิมพ์ได้ผลเป็นอย่างดี โดยให้ค่าความถูกต้องที่ 99.95%

3. ผลการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น

ผลการทดสอบอัลกอริทึมการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) โดยใช้วิธีวัดค่าความเหมือนแบบเฉลี่ย (Group Average Link) ด้วยฟังก์ชันระยะห่างยูคลีเดียน (Euclidean function) และกำหนดจำนวนกลุ่มที่ต้องการจำนวน 3 กลุ่ม (Cluster) ข้อมูลที่ใช้ทดสอบมีจำนวนทั้งสิ้น 2,141 ระเบียน สามารถแสดงผลการจัดกลุ่มแบบลำดับชั้นได้ดังนี้

```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 3-L AVERAGE
              -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     ClusteringR10
Instances:    2141
Attributes:   9
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield

Test mode:    evaluate on training data

=== Model and evaluation on training set ===

Cluster 0
((((((((((((((((((109.91:0.00764,(111.79:0.00458,123.49:0.00458)
:0.00306):0.00145,((97.28:0.00081,95.98:0.00081):0.0014,...

Cluster 1
(((((((96.0:0.01205,(92.45:0.00721,112.73:0.00721):0.00484):0.005
77,129.23:0.01783):0.01014,...

Cluster 2
((106.9:0.14083,((((124.82:0.00737,107.91:0.00737):0.00866,(151.78
:0.00222,148.47:0.00222):0.01381):0.00166,...


Clustered Instances

0      2105 ( 98%)
1        28 (  1%)
2         8 (  0%)
```

ผลลัพธ์จากการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) แสดงให้เห็นว่าสมาชิกในแต่ละกลุ่มมีลักษณะโดยรวมของข้อมูลที่แตกต่างกันมาก โดย Cluster 0 มีสมาชิกในกลุ่มมากที่สุด คือ 2,105 ระเบียน คิดเป็น 98% ของข้อมูลทั้งหมด Cluster 1 มีสมาชิกในกลุ่มจำนวน 28 ระเบียน คิดเป็น 1% ของข้อมูลทั้งหมด และ Cluster 2 มีสมาชิกในกลุ่มจำนวน 8 ระเบียน คิดเป็น 0% ของข้อมูลทั้งหมด หรือกล่าวในอีกทางหนึ่งก็คือ ข้อมูลที่นำมาทดสอบนี้มีลักษณะของข้อมูลที่คล้ายคลึงกันเป็นจำนวนมาก ส่งผลให้ Cluster 0 มีจำนวนสมาชิกมากที่สุด และเมื่อนำชุดข้อมูลที่ได้จากการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ของทุกระเบียน มาทำการวิเคราะห์คุณลักษณะของสมาชิกในแต่ละกลุ่ม สามารถแสดงผลได้ดังภาพที่ 24

toolId	toolItem	toolOption	minCG	maxCG	productionOutput	qcOutput	productionYield	finalYield	Cluster
DLS0304502-602	7L1512DLS0304502-602	DLS0304602	0.90	1.40	1.81	1.81	95.82	95.82	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.89	1.81	95.25	91.01	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.95	1.95	94.66	94.66	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.84	1.84	95.78	95.78	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.83	1.83	96.24	96.24	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.62	1.50	95.54	88.72	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.56	1.56	88.35	88.35	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	2.06	1.95	96.72	91.26	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	2.18	2.06	86.61	81.65	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.80	1.80	95.89	95.89	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.77	1.77	97.09	97.09	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.91	1.91	90.51	90.51	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	2.06	1.93	95.60	89.33	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.89	1.77	83.56	78.16	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.97	1.83	96.14	89.29	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.25	1.25	97.32	97.32	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.45	1.45	105.78	105.78	cluster0
DLS0404102	7L2091DLS0404102	DLS0404102	0.80	1.20	1.20	1.20	91.60	91.60	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.52	1.52	89.33	89.33	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.17	1.10	114.13	107.30	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.21	1.10	84.58	76.83	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.33	1.30	90.65	88.56	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	0.95	0.95	88.79	88.79	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.35	1.35	95.90	95.90	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.37	1.37	95.21	95.21	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.59	1.48	93.77	87.10	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.44	1.44	95.87	95.87	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.24	1.24	91.63	91.63	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.43	1.43	96.02	96.02	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.46	1.46	92.42	92.42	cluster0
DLS0404202	7L2111DLS0404202	DLS0404202	1.00	1.50	1.45	1.44	96.61	95.79	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	1.72	1.54	92.58	82.79	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	1.42	1.42	136.42	136.42	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	1.36	1.32	95.54	92.73	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	1.34	1.34	96.91	96.91	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	0.98	0.98	94.17	94.17	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	1.25	1.12	86.13	77.45	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	1.50	1.50	95.96	95.96	cluster0
DLS0407002	7L2241DLS0407002	DLS0407002	1.00	1.30	1.14	1.14	85.37	85.37	cluster0

ภาพที่ 24 ตัวอย่างชุดข้อมูลที่ได้จากการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

ผลลัพธ์การจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) สามารถสรุปได้ดังนี้

Cluster 0:

- แม่พิมพ์ที่เป็นสมาชิกของกลุ่มนี้มีรอบการบำรุงรักษาตามจำนวนการผลิตขั้นต่ำ (Minimum Cycle Grinding) และขั้นสูง (Maximum Cycle Grinding) อยู่ในช่วง 0.4 – 4.0 KKStroke (เค-เค-สโตรค)

- สมาชิกจำนวน 965 ระเบียนในกลุ่มนี้มียอดการผลิต (Production yield) และยอดงานดี (Final yield) ต่ำกว่า 95% คลอบคลุม 46% ของจำนวนสมาชิกทั้งหมด

- สมาชิกจำนวน 934 ระเบียนในกลุ่มนี้มียอดการผลิต (Production yield) และยอดงานดี (Final yield) มากกว่าหรือเท่ากับ 95% คลอบคลุม 44% ของจำนวนสมาชิกทั้งหมด

- สมาชิกจำนวน 206 ระเบียนในกลุ่มนี้มียอดการผลิต (Production yield) มากกว่าหรือเท่ากับ 95% แต่ให้ยอดงานดี (Final yield) น้อยกว่า 95% คลอบคลุม 10% ของจำนวนสมาชิกทั้งหมด

Cluster 1:

- แม่พิมพ์ที่เป็นสมาชิกของกลุ่มนี้มีรอบการบำรุงรักษาตามจำนวนการผลิตขั้นต่ำ (Minimum Cycle Grinding) และขั้นสูง (Maximum Cycle Grinding) อยู่ใน 3 ช่วง คือ 4.0 – 6.0, 5.0 – 8.0 และ 6.0 – 10.0 KKStroke

- สมาชิกจำนวน 25 ระเบียนในกลุ่มนี้มียอดการผลิต (Production yield) และยอดงานดี (Final yield) มากกว่าหรือเท่ากับ 95% คลอบคลุม 89% ของจำนวนสมาชิกทั้งหมด

- สมาชิกที่เหลือจำนวน 3 ระเบียนในกลุ่มนี้มียอดการผลิต (Production yield) และยอดงานดี (Final yield) ต่ำกว่า 95% คลอบคลุม 11% ของจำนวนสมาชิกทั้งหมด

Cluster 2:

- แม่พิมพ์ที่เป็นสมาชิกของกลุ่มนี้มีรอบการบำรุงรักษาตามจำนวนการผลิตขั้นต่ำ (Minimum Cycle Grinding) และขั้นสูง (Maximum Cycle Grinding) อยู่ในช่วง 3.0 – 3.5 KKStroke

- สมาชิกทั้งหมดจำนวน 8 ระเบียนในกลุ่มนี้เป็นแม่พิมพ์ที่มียอดการผลิต (Production yield) และยอดงานดี (Final yield) มากกว่า 95%

3.1 ผลการวัดคุณภาพการจัดกลุ่มแบบลำดับชั้นด้วยอัลกอริทึม C4.5

การวัดคุณภาพการจัดกลุ่มข้อมูล (Cluster Evaluation) ที่ได้จากการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ผู้วิจัยใช้วิธีวัดคุณภาพด้วยอัลกอริทึม C4.5 ซึ่งเป็นการสร้างต้นไม้ตัดสินใจ โดยกำหนดให้กลุ่มข้อมูล (Cluster) ที่ได้จากการจัดกลุ่มแบบลำดับชั้นเป็นแอตทริบิวท์ (Class) เป้าหมาย และใช้วิธีกำหนดความหลากหลายแบบสุ่ม (n-Fold Cross Validation) ในการแบ่งข้อมูลสำหรับเรียนรู้ (Training dataset) และข้อมูลสำหรับทดสอบ (Testing dataset) โดยกำหนดจำนวนครั้งที่สลับ (Fold) เท่ากับ 10 ซึ่งเป็นค่ามาตรฐานที่นิยมใช้โดยทั่วไป

ผลการวัดคุณภาพการจัดกลุ่มข้อมูล (Cluster Evaluation) ด้วยอัลกอริทึม C4.5 โครงสร้างต้นไม้ตัดสินใจที่ได้มีโหนดใบทั้งสิ้น 2 โหนด โดยทั้ง 2 โหนดใช้จำนวนการผลิตขั้นสูง (Maximum Cycle Grinding) ของรอบการบำรุงรักษาแม่พิมพ์เป็นกฎการตัดสินใจ ซึ่งนำมาสรุปได้ดังตารางที่ 13

ตารางที่ 13 ผลลัพธ์เบื้องต้นจากการวัดคุณภาพการจัดกลุ่มแบบลำดับชั้นด้วยอัลกอริทึม C4.5

สถิติรวม	**Class 1**	**Class 2**	**Class 3**
Leaf nodes = 2 Success rate = 99.67	TP rate = 1 FP rate = 0.194 Precision = 0.997	TP rate = 1 FP rate = 0 Precision = 1	TP rate = 0.125 FP rate = 0 Precision = 1

เมื่อวิเคราะห์ลักษณะการจัดกลุ่มข้อมูลแบบลำดับชั้น (Hierarchical Clustering) จากโครงสร้างต้นไม้ตัดสินใจที่ได้ สามารถสรุปได้ดังนี้

Class 1: Cluster 0

- แม่พิมพ์ที่เป็นสมาชิกของกลุ่มนี้มีรอบการบำรุงรักษาตามจำนวนการผลิตขั้นสูง (Maximum Cycle Grinding) น้อยกว่าหรือเท่ากับ 4.0 KKStroke

Class 2: Cluster 1

- แม่พิมพ์ที่เป็นสมาชิกของกลุ่มนี้มีรอบการบำรุงรักษาตามจำนวนการผลิตขั้นสูง (Maximum Cycle Grinding) มากกว่า 4.0 KKStroke

Class 3: Cluster 2

- แม่พิมพ์ที่เป็นสมาชิกของกลุ่มนี้มีรอบการบำรุงรักษาตามจำนวนการผลิตขั้นสูง (Maximum Cycle Grinding) น้อยกว่าหรือเท่ากับ 4.0 KKStroke โดยสมาชิกในกลุ่มนี้จะถูกนำไปรวมเป็นโหนดเดียวกับ Class 1: Cluster 0 เนื่องจากใช้กฎการตัดสินใจเดียวกัน

เมื่อทำการเปรียบเทียบผลลัพธ์การวัดคุณภาพการจัดกลุ่ม (Cluster Evaluation) ด้วยอัลกอริทึม C4.5 กับผลลัพธ์ที่ได้จากการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ที่อธิบายผลไว้ก่อนหน้านี้พบว่าข้อมูลส่วนใหญ่นั้นตรงกัน กล่าวคือ รอบการบำรุงรักษาตามจำนวนการผลิตขั้นต่ำ (Minimum Cycle Grinding: MinCG) และขั้นสูง (Maximum Cycle Grinding: MaxCG) ของแต่ละกลุ่มอยู่ในช่วงเดียวกัน เช่น รอบการบำรุงรักษาของ Cluster 0 อยู่ในช่วง 0.4-4.0 KKStroke และรอบบำรุงรักษาของ Class 1: Cluster 0 ตามจำนวนการผลิตขั้นสูง (MaxCG) มีค่าน้อยกว่าหรือเท่ากับ 4.0 KKStroke เป็นต้น

ผลจากการวัดคุณภาพยังแสดงให้เห็นว่าการจำแนกข้อมูลในแต่ละกลุ่มมีค่า TP rate และ Precision ที่สูง มีค่า FP rate ที่ต่ำ ต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทาง Class 1: Cluster 0 (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.194 เมื่อกลับไปพิจารณาที่คลาสดังกล่าว พบว่าสัญญาณหลอกเกิดขึ้นที่ Class 3: Cluster 2 ส่วนคลาสที่เหลือไม่มีอัตราการจำแนกผิด ดังนั้นจึงสรุปได้ว่าต้นไม้ตัดสินใจที่ได้มีความน่าเชื่อถือ สามารถนำไปทำนายข้อมูลเพื่อจัดกลุ่มแม่พิมพ์แบบลำดับชั้น (Hierarchical Clustering) ในเบื้องต้นได้ สำหรับผลการวัดคุณภาพการจัดกลุ่มแสดงรายละเอียดในภาคผนวก ค หน้าที่ 106

4. ผลการประเมินและคัดเลือกอัลกอริทึมที่เหมาะสม

การประเมินประสิทธิภาพการจัดกลุ่มแม่พิมพ์ เพื่อคัดเลือกอัลกอริทึมที่เหมาะสม ระหว่างอัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ผู้วิจัยใช้วิธีวัดค่าความคลาดเคลื่อน (Error Measurement) ในการประเมินประสิทธิภาพของอัลกอริทึม ซึ่งอัลกอริทึมที่เหมาะสม คือ อัลกอริทึมที่มีค่าความคลาดเคลื่อนต่ำที่สุด โดยพิจารณาเปรียบเทียบจากค่าสถิติต่างๆ จำนวน 3 วิธี ผลลัพธ์ที่ได้สรุปดังตารางที่ 14

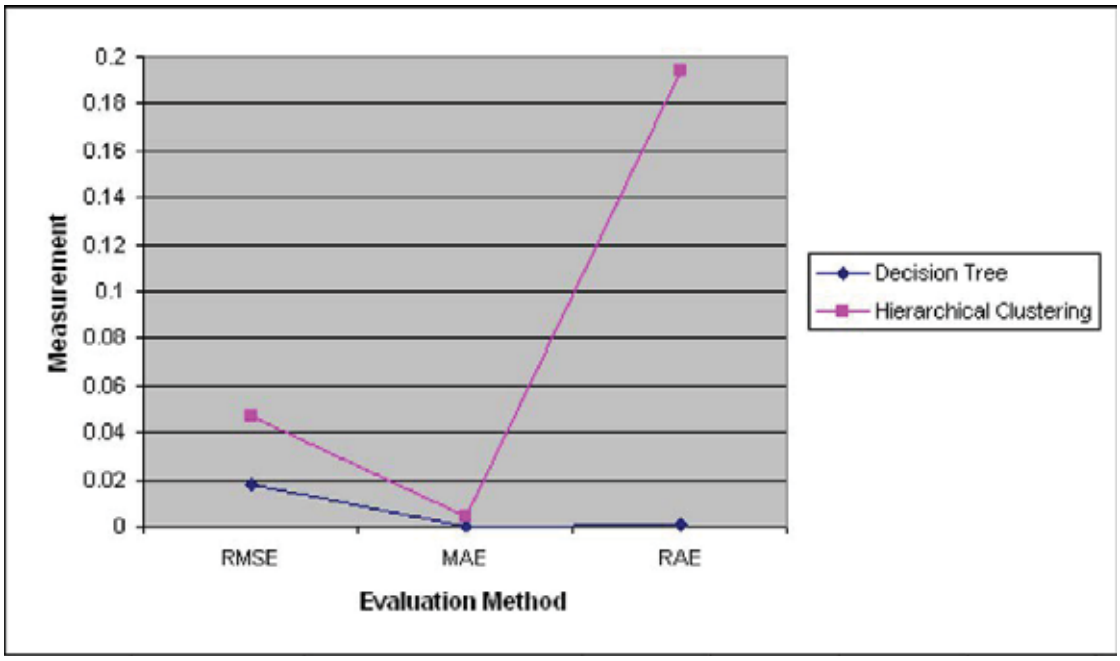
ตารางที่ 14 ตารางเปรียบเทียบค่าความคลาดเคลื่อน (Error rate)

ค่าความคลาดเคลื่อน (Error rate)	อัลกอริทึม	
	ต้นไม้ตัดสินใจ (Decision Tree)	การจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering)
RMSE (Root Mean Squared Error)	0.0176	0.0466
MAE (Mean Absolute Error)	0.0003	0.0044
RAE (Relative Absolute Error)	0.0802%	19.384%

เมื่อพิจารณาค่าความคลาดเคลื่อนจากตารางที่ 14 พบว่าค่าความคลาดเคลื่อนที่ได้จากอัลกอริทึมการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) มีค่าสูงกว่าอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) อย่างเห็นได้ชัดทั้ง 3 วิธี โดยเฉพาะค่าความคลาดเคลื่อนสัมพัทธ์ (Relative Absolute Error) ซึ่งมีค่าสูงกว่ามาก โดยอัลกอริทึมการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) มีค่าความคลาดเคลื่อนสัมพัทธ์ (Relative Absolute Error) ที่ 19.384% และอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) มีค่าความคลาดเคลื่อนสัมพัทธ์ (Relative Absolute Error) ที่ 0.0802% สำหรับค่าความคลาดเคลื่อนตัวอื่นๆ มีค่าแตกต่างกันเล็กน้อย

ข้อมูลจากการทดลองนี้สรุปได้ว่า อัลกอริทึมที่เหมาะสมในการคัดเลือกกลุ่มแม่พิมพ์คืออัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) เนื่องจากมีค่าความคลาดเคลื่อนต่ำกว่าอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) อย่างชัดเจน ดังนั้นอาจกล่าวได้ว่าอัลกอริทึมการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ยังไม่เหมาะสมกับการนำมาสร้างตัวแบบจากข้อมูลชุดนี้

จากตารางที่ 14 เมื่อปรับฐานค่าความคลาดเคลื่อนสัมพัทธ์ (Relative Absolute Error) จากค่าเปอร์เซ็นต์ให้อยู่ในรูปเลขจำนวนเต็ม สามารถแสดงผลข้อมูลค่าความคลาดเคลื่อนทั้งหมดในรูปแบบกราฟได้ดังภาพที่ 25



ภาพที่ 25 กราฟแสดงผลการเปรียบเทียบค่าความคลาดเคลื่อน

บทที่ 5
สรุปผลการวิจัย

การดำเนินงานวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและคัดเลือกเทคนิคการทำเหมืองข้อมูลที่เหมาะสมกับการคัดเลือกกลุ่มแม่พิมพ์โลหะแบบ Progressive Die ซึ่งเป็นแม่พิมพ์ที่มีราคาสูงและออกแบบยาก อัลกอริทึมในการทำเหมืองข้อมูลนั้นมีอยู่หลายวิธี ผู้วิจัยได้คัดเลือกอัลกอริทึมการทำเหมืองข้อมูลชนิดการเรียนรู้แบบมีการควบคุม (Supervised Learning) และอัลกอริทึมการเรียนรู้แบบไม่มีการควบคุม (Unsupervised Learning) มาอย่างละหนึ่งอัลกอริทึม ซึ่งอัลกอริทึมที่ใช้ในการทดสอบ คือ อัลกอริทึมการจำแนกข้อมูล (Classification) ด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการรวมกลุ่มข้อมูล (Clustering) ด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ในการพิจารณาอัลกอริทึมการทำเหมืองข้อมูลที่เหมาะสม ผู้วิจัยใช้วิธีวัดค่าความคลาดเคลื่อน (Error Measurement) โดยพิจารณาเปรียบเทียบจากค่าสถิติต่างๆ จำนวน 3 วิธี คือ Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) และ Relative Absolute Error (RAE) โดยอัลกอริทึมที่เหมาะสม คือ อัลกอริทึมที่มีค่าความคลาดเคลื่อนต่ำที่สุด

ผลการวิจัยเบื้องต้นจากโปรแกรม Weka พบว่าอัลกอริทึมที่เหมาะสม คือ อัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยมีค่าความคลาดเคลื่อนต่ำกว่าอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) อย่างชัดเจน และมีค่าต่ำกว่าทั้ง 3 วิธี โดยเฉพาะค่าความคลาดเคลื่อนสัมพัทธ์ (RAE) ที่มีค่าน้อยกว่ามาก กล่าวคือ โครงสร้างต้นไม้ตัดสินใจมีค่าความคลาดเคลื่อนสัมพัทธ์ที่ 0.0802% และการจัดกลุ่มแบบลำดับชั้นมีค่าความคลาดเคลื่อนสัมพัทธ์ที่ 19.384% สำหรับค่าความคลาดเคลื่อนที่เหลือมีความแตกต่างกันเล็กน้อย

ข้อมูลที่ใช้ในการศึกษานำมาจากฐานข้อมูลของบริษัท อาพิค ยามาดะ (ประเทศไทย) จำกัด ซึ่งเป็นข้อมูลการดำเนินงานระหว่างปี 2550-2552 นอกจากนี้ผู้วิจัยยังได้คัดเลือกดัชนีชี้วัดผลการปฏิบัติงาน (KPI: Key Performance Indicator) จากหน่วยงานที่เกี่ยวข้อง คือ แผนกผลิต แผนกขาย และแผนกซ่อมบำรุงแม่พิมพ์ เพื่อใช้เป็นเกณฑ์ในการจัดกลุ่มและอธิบายผลที่ได้จากการทดสอบอัลกอริทึม

ในการทดสอบอัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) ด้วยอัลกอริทึม C4.5 จากโปรแกรม Weka ในโมดูล J48 ผู้วิจัยได้ทำการทดลองจำแนกกลุ่มแม่พิมพ์ออกเป็น 3 กลุ่ม คือ Good, Normal และ Bad ตามดัชนีชี้วัดผลการปฏิบัติงาน (KPI) ใน

เงื่อนไขต่างๆ จำนวน 10 รูปแบบ เพื่อคัดเลือกรูปแบบการจำแนกกลุ่มที่ให้ค่าความถูกต้อง (Success rate) สูงสุด ซึ่งจะให้ต้นไม้ตัดสินใจที่ดีด้วยเช่นกัน ในการจำแนกผู้วิจัยใช้วิธีกำหนดความหลากหลายแบบสุ่ม (n-Fold Cross Validation) ในการแบ่งข้อมูลสำหรับเรียนรู้ (Training dataset) และข้อมูลสำหรับทดสอบ (Testing dataset) โดยกำหนดจำนวนครั้งที่สลับ (Fold) เท่ากับ 10 ซึ่งเป็นค่ามาตรฐานที่นิยมใช้โดยทั่วไป การวัดคุณภาพการจำแนกในแต่ละคลาส ผู้วิจัยพิจารณาจากค่า TP rate, FP rate และ Precision ซึ่งคำนวณได้จากตาราง Confusion Matrix ซึ่งต้นไม้ตัดสินใจที่ได้มีค่า TP rate และ Precision ที่สูง และมีค่า FP rate ที่ต่ำ และให้ค่าความถูกต้องโดยรวมที่ 99.95%

สำหรับการทดสอบอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) โดยใช้วิธีวัดค่าความเหมือนแบบเฉลี่ย (Group Average Link) ด้วยฟังก์ชันระยะห่างยูคลีเดียน (Euclidean function) จากโปรแกรม Weka โดยกำหนดจำนวนกลุ่มที่ต้องการเท่ากับ 3 กลุ่ม (Cluster) ผลลัพธ์ที่ได้จากการจัดกลุ่มแบบลำดับชั้นพบว่า ลักษณะโดยรวมของข้อมูลมีความคล้ายคลึงกันมาก ส่งผลให้จำนวนสมาชิกใน Cluster 0 มีจำนวนสมาชิกมากที่สุด จำนวน 2,105 ระเบียน จากจำนวนข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 98% ของข้อมูลทั้งหมดในการวัดคุณภาพการจัดกลุ่ม (Cluster Evaluation) ที่ได้จากการจัดกลุ่มแบบลำดับชั้น ผู้วิจัยเลือกใช้วิธีวัดคุณภาพการจัดกลุ่มด้วยอัลกอริทึม C4.5 ซึ่งเป็นการสร้างต้นไม้ตัดสินใจ โดยกำหนดให้กลุ่มข้อมูล (Cluster) ที่ได้จากการจัดกลุ่มแบบลำดับชั้นเป็นแอตทริบิวท์ (Class) เป้าหมาย จากนั้นพิจารณาคุณภาพการจัดกลุ่ม (Cluster) จากค่า TP rate, FP rate และ Precision การวัดคุณภาพด้วยวิธีนี้มีข้อดีอย่างน้อย 2 ประการ คือ ประการแรกทำให้สามารถเห็นรูปแบบการจัดกลุ่ม (Clustering) ของอัลกอริทึมจากโครงสร้างต้นไม้ตัดสินใจ ประการที่สองผลลัพธ์ที่ได้จากโครงสร้างต้นไม้ตัดสินใจสามารถนำไปวิเคราะห์คุณภาพรูปแบบการจัดกลุ่มเพิ่มเติมได้ (Roiger and Geatz 2003: 59) และเมื่อเปรียบเทียบผลลัพธ์การวัดคุณภาพการจัดกลุ่ม (Cluster Evaluation) ด้วยอัลกอริทึม C4.5 กับผลลัพธ์ที่ได้จากการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) พบว่าสมาชิกที่อยู่ในแต่ละกลุ่มมีคุณลักษณะของข้อมูลส่วนใหญ่ตรงกัน

โดยสรุปการศึกษาวิจัยในครั้งนี้สามารถบรรลุวัตถุประสงค์ของการศึกษาทั้ง 2 ประการ คือ ประการแรก สามารถวิเคราะห์ข้อมูลแม่พิมพ์ด้วยเทคนิคเหมืองข้อมูล โดยศึกษาและเปรียบเทียบระหว่างอัลกอริทึมการจำแนกข้อมูลด้วยวิธีโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการรวมกลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ได้เป็นผลสำเร็จ ประการที่สอง สามารถใช้ผลลัพธ์ที่ได้จากการทดสอบอัลกอริทึมเป็นแนวทางในการคัดเลือกและจัดกลุ่มแม่พิมพ์ได้ ซึ่งค่าความถูกต้องของการจัดกลุ่มแม่พิมพ์อยู่ในเกณฑ์ที่สูง

ข้อจำกัดของการศึกษา

1. การจัดเตรียมข้อมูลใช้เวลานาน เนื่องจากข้อมูลที่ใช้ทดสอบอัลกอริทึมนำมาจากฐานข้อมูลที่มีโครงสร้างแตกต่างกันและมีข้อมูลในปริมาณมาก

2. การคำนวณเปอร์เซ็นต์ยอดการผลิต (Production yield) และเปอร์เซ็นต์ยอดงานดี (Final yield) เป็นการคำนวณย้อนหลัง ซึ่งอาจทำให้ข้อมูลที่ได้คลาดเคลื่อนจากความเป็นจริง เนื่องจากระบบ ERP แยกข้อมูลการผลิตตามช่วงเวลาในการผลิตและซ่อมบำรุงที่เกิดขึ้นวันเดียวกันไม่ได้ เช่น หากแม่พิมพ์ทำการผลิตในช่วงเช้าและถูกนำมาบำรุงรักษาในรอบบ่าย จากนั้นนำกลับไปผลิตต่อในช่วงกลางคืนของวันเดียวกัน กรณีนี้การคำนวณเปอร์เซ็นต์ยอดการผลิต (Production yield) และเปอร์เซ็นต์ยอดงานดี (Final yield) ของรอบการบำรุงรักษาแม่พิมพ์ที่ติดกันอาจคลาดเคลื่อนได้

3. ข้อมูลเกาะกลุ่มอยู่กลุ่มเดียวในปริมาณมาก เนื่องจากแอตทริบิวท์อิสระที่นำมาใช้ทดสอบอัลกอริทึมมีความคล้ายคลึงกันมาก ทำให้กลุ่มข้อมูลที่ได้จากการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) เกาะกลุ่มอยู่กลุ่มเดียวในปริมาณมาก

ข้อเสนอแนะ

1. จากการที่ผู้วิจัยได้นำเสนออัลกอริทึมการจำแนกข้อมูลด้วยโครงสร้างต้นไม้ตัดสินใจเป็นอัลกอริทึมที่เหมาะสมในการคัดเลือกกลุ่มแม่พิมพ์นั้น สามารถที่จะศึกษาอัลกอริทึมดังกล่าวเพิ่มเติมได้ด้วยการเพิ่มแอตทริบิวท์ประเภทอื่นๆ ที่สนใจ เช่น ข้อมูลลูกค้า ข้อมูลวัตถุดิบ เพื่อพิจารณาความสัมพันธ์ในการจำแนกกลุ่มแม่พิมพ์

2. การเก็บข้อมูลเปอร์เซ็นต์ยอดการผลิต (Production yield) และเปอร์เซ็นต์ยอดงานดี (Final yield) ควรทำการคำนวณและเก็บข้อมูลทันทีที่ทำการบำรุงรักษาแม่พิมพ์ เพื่อให้ค่าที่ได้ตรงกับความเป็นจริงมากที่สุด

3. การจำแนกกลุ่มแม่พิมพ์ อาจทำการทดสอบจากอัลกอริทึมอื่นๆ แล้วนำผลที่ได้เปรียบเทียบกัน เช่น อัลกอริทึม NaiveBayes อัลกอริทึม OneR เป็นต้น

4. การจัดเตรียมข้อมูลที่เป็นตัวเลข (Numeric) เช่น ข้อมูลการผลิต (Production output) อาจทำการแปลงข้อมูลตัวเลขเป็นข้อมูลชนิดมีค่าไม่ต่อเนื่อง (Nominal) เพื่อช่วยในการวิเคราะห์การจำแนกกลุ่มหรือวิเคราะห์การเกาะกลุ่ม ตัวอย่างการแปลงข้อมูล (Discretization) ชนิดตัวเลขเป็นข้อมูลชนิดมีค่าไม่ต่อเนื่อง แสดงรายละเอียดในภาคผนวก ง หน้าที่ 110

บรรณานุกรม

ภาษาไทย

เกรียงไกร พิพิธหิรัญการ. “การเปรียบเทียบประสิทธิภาพในการทำนายพฤติกรรมผู้บริโภคโดยใช้เทคนิคการทำเหมืองข้อมูลระหว่างกฎความสัมพันธ์สำหรับจำแนกและต้นไม้ตัดสินใจ กรณีศึกษา: ศูนย์หนังสือมหาวิทยาลัยเกษตรศาสตร์.” สารนิพนธ์ปริญญามหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2550.

เสาวคนธ์ วิทวัสโอฬาร. หลุมพรางกลยุทธ์ KPI & BSC [ออนไลน์]. เข้าถึงเมื่อ 25 ธันวาคม 2551. เข้าถึงได้จาก http://www.bangkokbiznews.com/2007/10/16/news_24849366.php

ก้องศักดิ์ จงเกษมวงศ์. “การตัดเล็มอย่างอ่อนสำหรับต้นไม้ตัดสินใจโดยการใช้แป๊กพรอพาเกชันนิวรอลเน็ตเวิร์ก.” วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2543.

กัลยา วินิชย์บัญชา. การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows. กรุงเทพฯ : ธรรมสาร จำกัด, 2546.

คมกริช อุดมมณีธนกิจ. “เทคนิคการแบ่งกลุ่มกระแสข้อมูล เพื่อเพิ่มคุณภาพของกลุ่มข้อมูลผลลัพธ์.” โครงการวิจัยวิทยานิพนธ์ สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์, 2548.

ณัฐมณฑ์ สิริวัฒนานันท์. “การตรวจสอบความเหมาะสมในการขนส่งสินค้าโดยวิธีต้นไม้ตัดสินใจ.” สารนิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2551.

ดนัย เทียนพุฒ. ดัชนีวัดผลสำเร็จธุรกิจ = KPIs: Key Performance Indicators. กรุงเทพฯ : ดี เอ็น ที คอนซัลแตนท์, 2544.

บุญเสริม กิจศิริกุล. “อัลกอริทึมการทำเหมืองข้อมูล.” รายงานวิจัยฉบับสมบูรณ์ โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ 2545 ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย, 2546.

บุญทัน คำพาศัย. “การจัดกลุ่มแขวงต่างๆ ในประเทศสาธารณรัฐประชาธิปไตยประชาชนลาวด้วยเทคนิคการวิเคราะห์จัดกลุ่ม 3 วิธี คือ Two-steps, K-Means และ Hierarchy.” สารนิพนธ์ปริญญามหาบัณฑิต สาขาสถิติประยุกต์ มหาวิทยาลัยขอนแก่น, 2549.

มลธิดา ฤทธิ์สมบูรณ์ และสุชา สมานชาติ. “การพัฒนาระบบสนับสนุนการพิจารณาอนุมัติให้สินเชื่อเพื่อการเช่าซื้อสินค้าโดยใช้เทคนิคต้นไม้ตัดสินใจ.” วารสารเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ 4, 7 (มกราคม – มิถุนายน 2551) : 8-14.

สถาบันไทย-เยอรมัน. ทฤษฎีแม่พิมพ์โลหะ. กรุงเทพฯ : สถาบันไทย-เยอรมัน, 2550.

สิทธา ดอร์จิ. “ระบบอัจฉริยะแบบผสมผสานสำหรับการทำเหมืองข้อมูลการใช้งานโทรศัพท์ทางไกลของลูกค้าสำหรับการแบ่งกลุ่มลูกค้า.” วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ (นานาชาติ) บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2552.

ห้องสมุดกรมส่งเสริมอุตสาหกรรม. การประเมินผลการปฏิบัติงาน (Evaluation) [ออนไลน์]. เข้าถึงเมื่อ 25 ธันวาคม 2551. เข้าถึงได้จาก http://library.dip.go.th/multim5/ebook/I%20%E0%B8%81%E0%B8%AA%E0%B8%AD15%20T19H6.pdf

ห้องสมุดกรมส่งเสริมอุตสาหกรรม. เทคโนโลยีแม่พิมพ์ [ออนไลน์]. เข้าถึงเมื่อ 24 กันยายน 2551. เข้าถึงได้จาก http://library.dip.go.th/multim4/eb/EB%20122.2%20ม47.doc

อรนุช ชัยหมื่น. “การศึกษาเปรียบเทียบการแบ่งกลุ่มข้อมูลลูกค้าสินค้าหัตถกรรมไทยโดยใช้วิธีการ 2 ขั้นตอนของ SOM กับ K-Means Algorithm และ Hierarchical Cluster กับ K-Means Algorithm.” วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์, 2548.

อัญชลิสา แต้ตระกูล. “การสร้างโมเดลสำหรับวิเคราะห์ความผิดพลาดในการผลิตฮาร์ดดิสก์ด้วยเทคนิคการขุดเหมืองข้อมูล.” วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, 2552.

ภาษาต่างประเทศ

Berry, Michael J.A., and Gordon S. Linnoff. Data Mining Techniques For Marketing, Sale and Customer Relationship Management. New York : Wiley Publishing, 2004.

Han, Jiawei, and Micheline Kamber. Data Mining Concepts and Techniques. U.S.A. : Morgan Kaufman, 2001.

Kardi Teknomo’s Tutorials [Online]. Accessed 9 August 2010. Available from http://people.revoledu.com/kardi/tutorial/index.html

Larose, Daniel T. Discovering Knowledge in Data an Introduction to Data Mining. New Jersey : John Wiley & Sons Inc., 2005.

Roiger, Richard J., and Michael W. Geatz. Data Mining A TUTORIAL-BASED PRIMER. U.S.A. : Pearson Addison Wesley, 2003.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. U.S.A. : Pearson Addison Wesley, 2003.

Witten, Ian H., and Eibe Frank. Data Mining Practical Machine Learning Tools and Techniques Second Edition. U.S.A. : Morgan Kaufman, 2005.

ภาคผนวก

ภาคผนวก ก
โครงสร้างข้อมูล

โครงสร้างของตารางข้อมูลที่ใช้ในงานวิจัย มีดังนี้

ตารางที่ 15 โครงสร้างตารางข้อมูลยอดการผลิตและการซ่อมบำรุงแม่พิมพ์

ลำดับที่	ชื่อรายการข้อมูล	ชนิดข้อมูล	ขนาด	รายละเอียด	ตัวอย่างข้อมูล
1	toolItem	ตัวอักษร	80	รหัสแม่พิมพ์ที่ใช้ในระบบการผลิต	7L1512DLS0304502-602
2	toolId	ตัวอักษร	40	รหัสแม่พิมพ์ที่ใช้ในระบบซ่อมบำรุงแม่พิมพ์	DLS0304502-602
3	toolOption	ตัวอักษร	40	รหัสแม่พิมพ์ย่อย	DLS0304502
4	cgCount	ตัวเลข	1	ลำดับที่ในการบำรุงรักษาแม่พิมพ์	2
5	toolDesc	ตัวอักษร	255	รายละเอียดแม่พิมพ์	4044490/503PDIP16L(80X100)/(90X120)
6	minCG	ตัวเลข	8	จำนวนการผลิตขั้นต่ำที่ต้องบำรุงรักษาแม่พิมพ์	0.9
7	maxCG	ตัวเลข	8	จำนวนการผลิตขั้นสูงที่ต้องบำรุงรักษาแม่พิมพ์	1.4
8	targetPunch	ตัวเลข	8	ระยะมาตรฐานที่ต้อง grinding ในแต่ละครั้งของ Punch (มิลลิเมตร)	0.1
9	targetDie	ตัวเลข	8	ระยะมาตรฐานที่ต้อง grinding ในแต่ละครั้งของ Die (มิลลิเมตร)	0.08
10	matlType	ตัวอักษร	50	ชนิดของ Material	A194
11	matlThickness	ตัวเลข	8	ความหนาของ Material	0.254
12	cgDate	ตัวเลข	4	วันที่บำรุงรักษาแม่พิมพ์	ปีเดือนวัน : 20091025
13	regrindPunch	ตัวเลข	8	ระยะจริงที่ทำการ grinding Punch (มิลลิเมตร)	0.13
14	regrindDie	ตัวเลข	8	ระยะจริงที่ทำการ grinding Die (มิลลิเมตร)	0.11
15	beforePunch	ตัวเลข	8	ระยะก่อน grinding Punch	57.26

ตารางที่ 15 (ต่อ)

ลำดับที่	ชื่อรายการข้อมูล	ชนิดข้อมูล	ขนาด	รายละเอียด	ตัวอย่างข้อมูล
16	afterPunch	ตัวเลข	8	ระยะที่เหลือหลังจาก grinding Punch	57.13
17	beforeDie	ตัวเลข	8	ระยะก่อน grinding Die	11.64
18	afterDie	ตัวเลข	8	ระยะที่เหลือหลังจาก grinding Die	11.53
19	useablePunch	ตัวเลข	8	ระยะความสูงของ Punch ที่เหลืออยู่	3.5
20	productionOutput	ตัวเลข	8	ผลรวมของยอดการผลิตของรอบการซ่อมบำรุงในแต่ละครั้ง	1.59
21	qcOutput	ตัวเลข	8	ผลรวมของยอดงานดีของรอบการซ่อมบำรุงในแต่ละครั้ง	1.48
22	productionYield	ตัวเลข	8	อัตราผลผลิตทั้งหมดที่เกิดจากกระบวนการผลิต (เปอร์เซ็นต์)	92.55
23	finalYield	ตัวเลข	8	อัตราผลผลิตของยอดงานดีที่เกิดจากกระบวนการผลิต (เปอร์เซ็นต์)	86.31
24	toolClass	ตัวอักษร	10	กลุ่มแม่พิมพ์ที่กำหนดสำหรับอัลกอริทึม C4.5	Good, Normal และ Bad

ตารางที่ 16 โครงสร้างตารางข้อมูลการผลิตและการขายแยกตามคำสั่งผลิต

ลำดับที่	ชื่อรายการข้อมูล	ชนิดข้อมูล	ขนาด	รายละเอียด	ตัวอย่างข้อมูล
1	toolItem	ตัวอักษร	80	รหัสแม่พิมพ์ที่ใช้ในระบบการผลิต	7L1512DLS0304502-602
2	toolId	ตัวอักษร	40	รหัสแม่พิมพ์ที่ใช้ในระบบซ่อมบำรุงแม่พิมพ์	DLS0304502-602
3	toolOption	ตัวอักษร	40	รหัสแม่พิมพ์ย่อย	DLS0304502
4	cgCount	ตัวเลข	1	ลำดับที่ในการบำรุงรักษาแม่พิมพ์	2
5	jobNo	ตัวอักษร	20	รหัสงานที่สั่งผลิต	LN02827
6	suffixNo	ตัวเลข	2	รหัสงานย่อยที่สั่งผลิต	0, 100
7	productLot	ตัวอักษร	50	รหัส Product Lot No	K67267-712
8	item	ตัวอักษร	40	รหัสสินค้า	124151010
9	itemDesc	ตัวอักษร	100	รายละเอียดสินค้า	4044490 IDLF16L(80x100)mtx
10	itemUps	ตัวเลข	8	จำนวนชิ้นงานที่ได้ในหนึ่งครั้งที่ทำการปั๊มชิ้นงาน	4
11	grossWeight	ตัวเลข	8	ค่าที่ใช้แปลงหน่วยวัตถุดิบจากกิโลกรัมเป็นจำนวนชิ้นงาน	0.7995
12	itemMatl	ตัวอักษร	50	รหัสวัตถุดิบ	310144010
13	subLot	ตัวอักษร	50	รหัสย่อย Product Lot No.	K67267
14	qtyIssuedKg	ตัวเลข	8	จำนวนวัตถุดิบที่เบิกเข้ากระบวนการผลิต (หน่วยเป็นกิโลกรัม)	242
15	qtyIssuedKp	ตัวเลข	8	จำนวนวัตถุดิบที่เบิกเข้ากระบวนการผลิต (หน่วยเป็น Kilo pieces)	0
16	prodCompleted	ตัวเลข	8	จำนวนที่ผลิตจากแผนกผลิตได้ทั้งงานดีและงานเสีย	110

ตารางที่ 16 (ต่อ)

ลำดับที่	ชื่อรายการข้อมูล	ชนิดข้อมูล	ขนาด	รายละเอียด	ตัวอย่างข้อมูล
17	qcCompleted	ตัวเลข	8	จำนวนยอดการผลิตที่ผ่านการตรวจจากแผนก QC (ยอดพร้อมขาย)	98
18	locNo	ตัวอักษร	8	รหัส Location ที่เก็บสินค้า	1FNGD
19	mcNo	ตัวอักษร	10	เครื่องจักรที่ผลิต	K80-08
20	lineNo	ตัวอักษร	10	ไลน์การผลิต	L1
21	transDate	ตัวเลข	4	วันที่บันทึกข้อมูลการผลิต	ปีเดือนวัน : 20091025
22	eduId	ตัวอักษร	5	รหัสลูกค้าหลัก	02
23	eduName	ตัวอักษร	50	รายละเอียดลูกค้าหลัก	SONY
24	eduGroup	ตัวอักษร	10	รหัสกลุ่มลูกค้า	JAP
25	lfGroup	ตัวเลข	1	รหัสกลุ่มการผลิต	0 = Leadframe 1 = Non-Leadframe

ภาคผนวก ข

ผลการจำแนกข้อมูลด้วยอัลกอริทึม C4.5

ผลการจำแนกข้อมูลด้วยอัลกอริทึม C4.5
ตามเงื่อนไขดัชนีวัดผลการปฏิบัติงาน (KPI) ชุดข้อมูลที่ 1-9

ตารางที่ 17 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 1

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Good
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 1 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R01
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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

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

productionOutput <= 0.79
|   productionYield <= 95.05: Bad (83.0/10.0)
|   productionYield > 95.05
|   |   minCG <= 0.7: Good (17.0/1.0)
|   |   minCG > 0.7: Normal (39.0)
productionOutput > 0.79: Good (2002.0/114.0)

Number of Leaves: 4
Size of the tree: 7
```

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

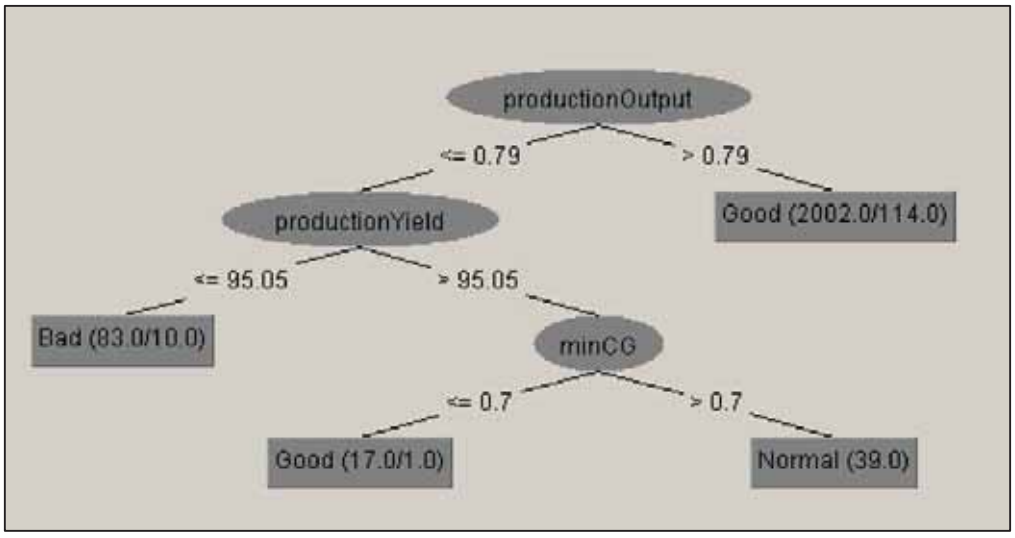
Correctly Classified Instances	2013	94.0215 %
Incorrectly Classified Instances	128	5.9785 %
Kappa statistic	0.6122	
Mean absolute error	0.0742	
Root mean squared error	0.1942	
Relative absolute error	56.8645 %	
Root relative squared error	76.1513 %	
Coverage of cases (0.95 level)	97.1042 %	
Mean rel. region size (0.95 level)	65.7325 %	
Total Number of Instances	2141	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.994	0.511	0.943	0.994	0.967	0.723	Good
0.545	0.005	0.878	0.545	0.673	0.748	Bad
0.411	0.001	0.951	0.411	0.574	0.681	Normal

=== Confusion Matrix ===

a	b	c	<-- classified as
1902	10	2	a = Good
60	72	0	b = Bad
56	0	39	c = Normal



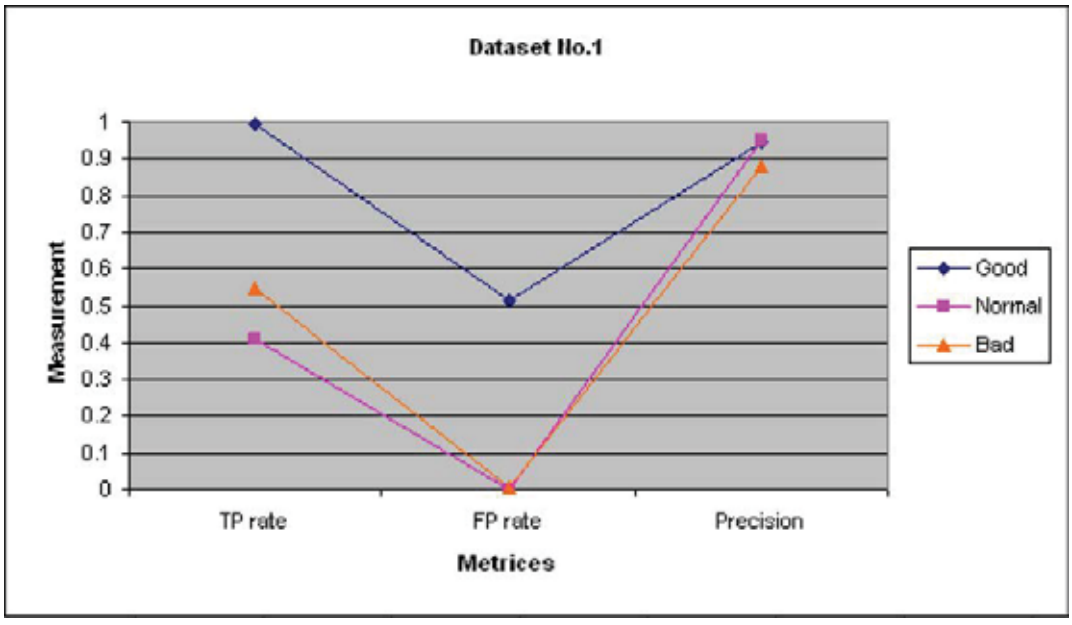
ภาพที่ 26 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 1

สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 1

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 2,013 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 94.0215 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,902 ระเบียน อยู่ในคลาส Normal จำนวน 39 ระเบียน และในคลาส Bad จำนวน 72 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 128 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 5.9785 % โดยข้อมูลทั้ง 128 ระเบียนที่จำแนกผิดนั้น คือ

- มี 12 ระเบียนที่อยู่ในคลาส Good แต่โปรแกรมจำแนกให้อยู่ในคลาส Bad จำนวน 10 ระเบียน และในคลาส Normal จำนวน 2 ระเบียน
- มี 56 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good ทั้ง 56 ระเบียน
- มี 60 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Good ทั้ง 60 ระเบียน



ภาพที่ 27 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 1

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 27 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Good (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.511 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad และ Normal โดยมีอัตราการจำแนกผิดเท่ากับ 0.005 และ 0.001 ตามลำดับ

ตารางที่ 18 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 2

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Normal
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Bad
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 2 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R02
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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

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

productionYield <= 94.99
|   productionOutput <= 0.79: Bad (83.0/10.0)
|   productionOutput > 0.79: Normal (885.0/59.0)
productionYield > 94.99
|   qcOutput <= 0.79: Normal (66.0/23.0)
|   qcOutput > 0.79: Good (1107.0/52.0)

Number of Leaves:        4
Size of the tree:        7
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1994               93.134  %
Incorrectly Classified Instances       147                6.866  %
```

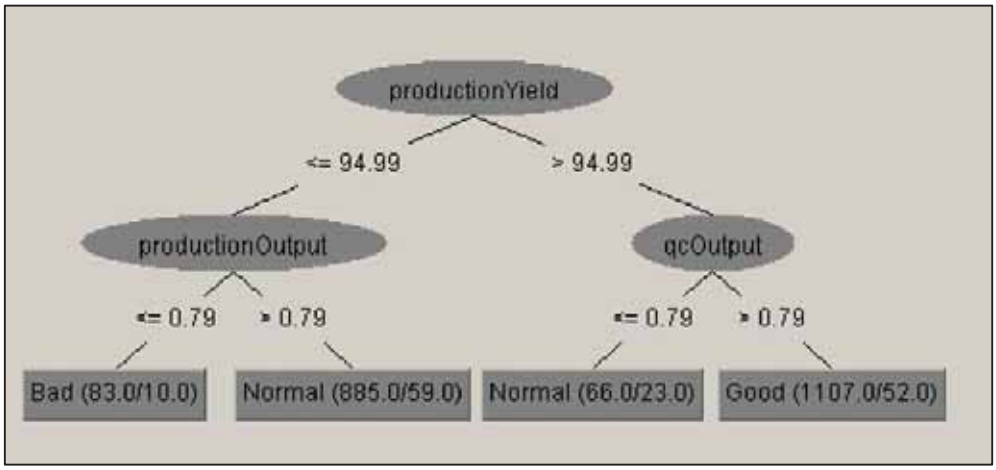
```
Kappa statistic                          0.8735
Mean absolute error                      0.0837
Root mean squared error                  0.206
Relative absolute error                 22.6781 %
Root relative squared error             47.9576 %
Coverage of cases (0.95 level)          98.9724 %
Mean rel. region size (0.95 level)      61.0462 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.995	0.069	0.936	0.995	0.965	0.968	Good
0.553	0.005	0.88	0.553	0.679	0.883	Bad
0.911	0.053	0.93	0.911	0.92	0.931	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1073    0    5 |    a = Good
    0   73   59 |    b = Bad
   73   10  848 |    c = Normal
```



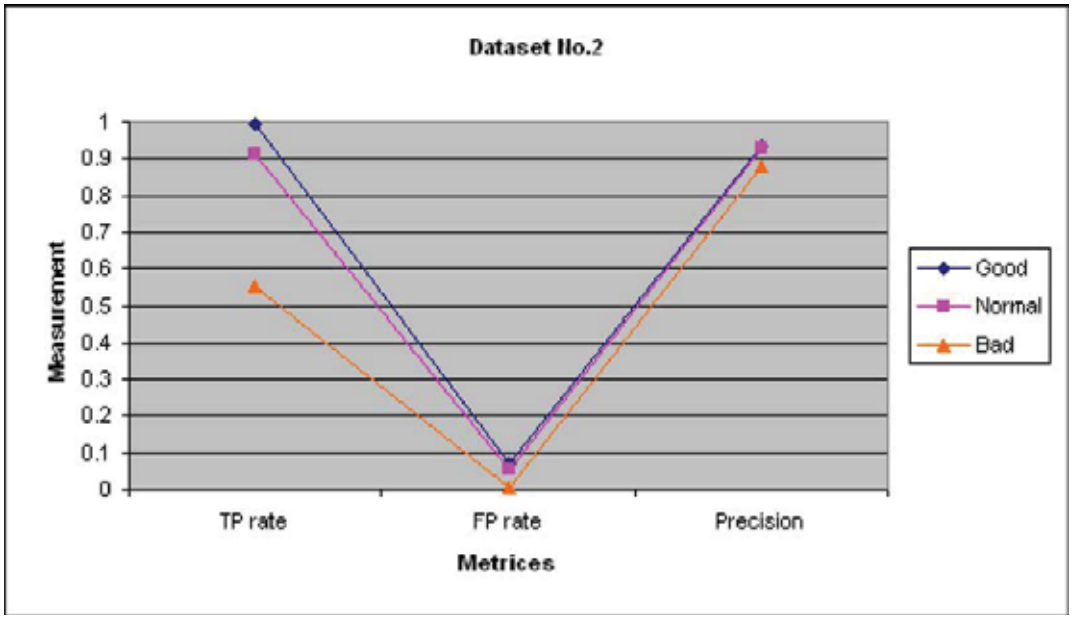
ภาพที่ 28 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 2

สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 2

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 1,994 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 93.134 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,073 ระเบียน อยู่ในคลาส Normal จำนวน 848 ระเบียน และในคลาส Bad จำนวน 73 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 147 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 6.866 % โดยข้อมูลทั้ง 147 ระเบียนที่จำแนกผิดนั้น คือ

- มี 5 ระเบียนที่อยู่ในคลาส Good แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal จำนวน 5 ระเบียน
- มี 83 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good จำนวน 73 ระเบียน และในคลาส Bad จำนวน 10 ระเบียน
- มี 59 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal ทั้ง 59 ระเบียน



ภาพที่ 29 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 2

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 29 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Good (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.069 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Normal โดยมีอัตราการจำแนกผิดเท่ากับ 0.053 และคลาส Bad มีอัตราการจำแนกผิดที่ 0.005

ตารางที่ 19 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 3

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Normal
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Bad
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Bad
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 3 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R03
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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

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

productionYield <= 94.99
|   productionOutput <= 0.79: Bad (83.0/10.0)
|   productionOutput > 0.79: Normal (885.0/59.0)
productionYield > 94.99: Good (1173.0/95.0)

Number of Leaves:        3
Size of the tree:        5
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1978               92.3867 %
Incorrectly Classified Instances       163                7.6133 %
Kappa statistic                          0.86
```

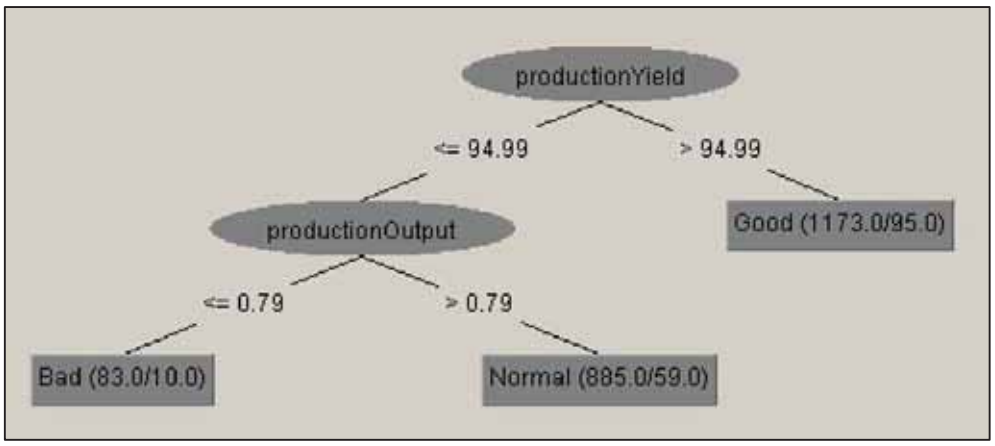
```
Mean absolute error                      0.0932
Root mean squared error                  0.2165
Relative absolute error                 25.0802 %
Root relative squared error             50.2497 %
Coverage of cases (0.95 level)          99.0659 %
Mean rel. region size (0.95 level)      64.8762 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.998	0.086	0.922	0.998	0.959	0.951	Good
0.514	0.005	0.88	0.514	0.649	0.845	Bad
0.9	0.051	0.93	0.9	0.915	0.918	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1076    0    2 |    a = Good
    9   73   60 |    b = Bad
   82   10  829 |    c = Normal
```



ภาพที่ 30 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 3

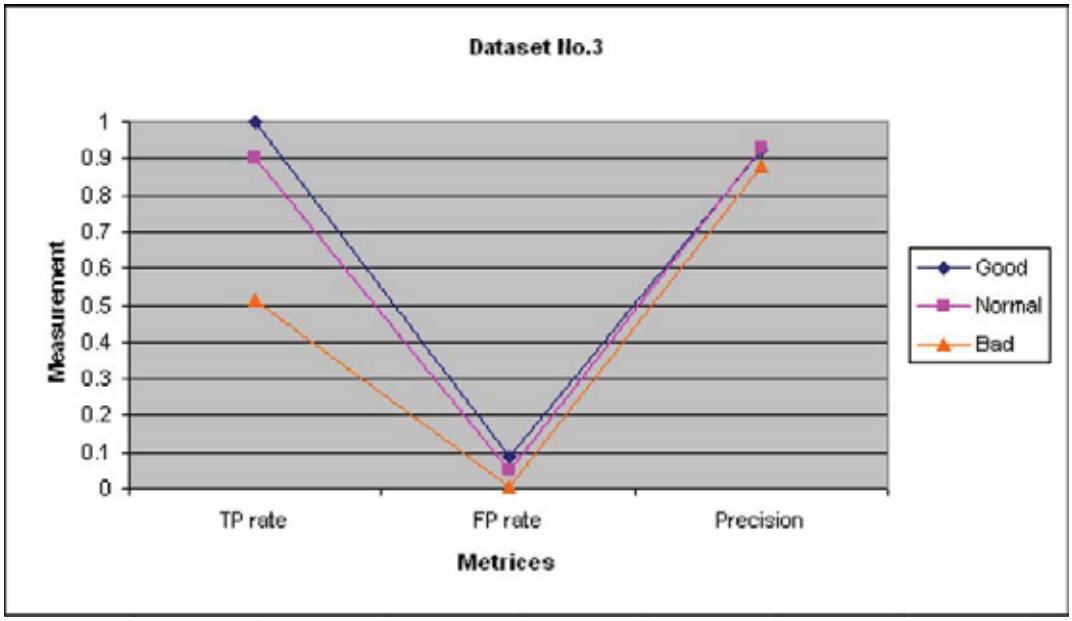
สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 3

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 1,978 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 92.3867 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,076 ระเบียน อยู่ในคลาส Normal จำนวน 829 ระเบียน และในคลาส Bad จำนวน 73 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 163 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 7.6133 % โดยข้อมูลทั้ง 163 ระเบียนที่จำแนกผิดนั้น คือ

- มี 2 ระเบียนที่อยู่ในคลาส Good แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal จำนวน 2 ระเบียน

- มี 92 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good จำนวน 82 ระเบียน และในคลาส Bad จำนวน 10 ระเบียน
- มี 69 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal ทั้ง 60 ระเบียน และในคลาส Good จำนวน 9 ระเบียน



ภาพที่ 31 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 3

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 31 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Good (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.086 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad และ Normal โดยมีอัตราการจำแนกผิดเท่ากับ 0.005 และ 0.051 ตามลำดับ

ตารางที่ 20 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 4

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Normal
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Bad
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Bad
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Bad
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 4 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R04
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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


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

productionYield <= 94.99
|   productionOutput <= 0.79: Bad (83.0/10.0)
|   productionOutput > 0.79: Normal (885.0/59.0)
productionYield > 94.99
|   qcOutput <= 0.79: Bad (66.0/23.0)
|   qcOutput > 0.79: Good (1107.0/52.0)

Number of Leaves:        4
Size of the tree:        7
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===
```

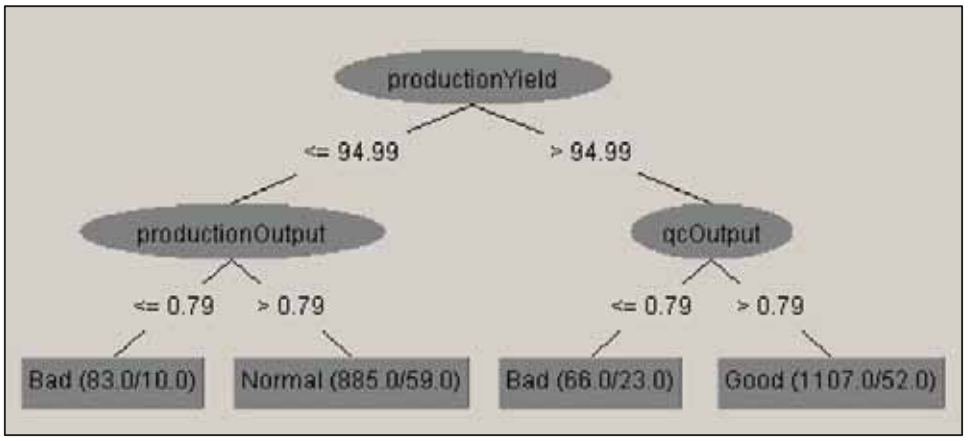
```
Correctly Classified Instances        1983               92.6203 %
Incorrectly Classified Instances       158                7.3797 %
Kappa statistic                          0.8682
Mean absolute error                      0.09
Root mean squared error                  0.2132
Relative absolute error                 23.1556 %
Root relative squared error             48.3711 %
Coverage of cases (0.95 level)          99.4395 %
Mean rel. region size (0.95 level)      64.5648 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.084	0.924	1	0.96	0.956	Good
0.352	0.005	0.889	0.352	0.505	0.668	Bad
0.987	0.045	0.933	0.987	0.959	0.972	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1078    0    0 |    a = Good
   88   80   59 |    b = Bad
    1   10  825 |    c = Normal
```



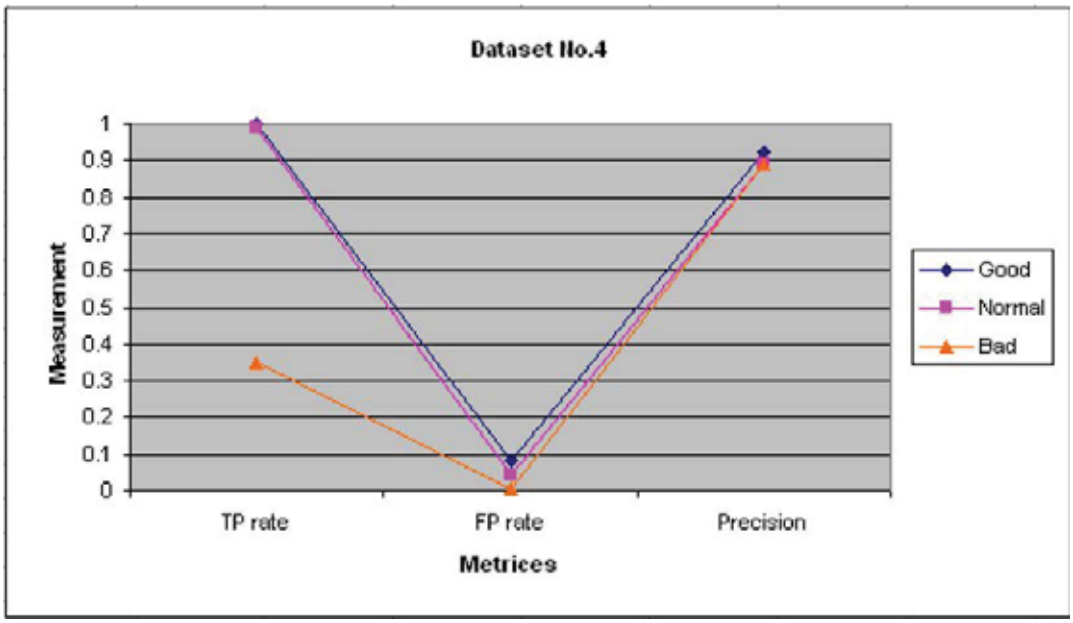
ภาพที่ 32 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 4

สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 4

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 1,983 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 92.6203 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,078 ระเบียน อยู่ในคลาส Normal จำนวน 825 ระเบียน และในคลาส Bad จำนวน 80 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 158 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 7.3797 % โดยข้อมูลทั้ง 158 ระเบียนที่จำแนกผิดนั้น คือ

- มี 11 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good จำนวน 1 ระเบียน และในคลาส Bad จำนวน 10 ระเบียน
- มี 147 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal จำนวน 59 ระเบียน และอยู่ในคลาส Good จำนวน 88 ระเบียน



ภาพที่ 33 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 4

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 33 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Good (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.084 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad และ Normal โดยมีอัตราการจำแนกผิดเท่ากับ 0.005 และ 0.045 ตามลำดับ

ตารางที่ 21 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 5

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 5 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R05
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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


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

productionYield <= 94.99
|   productionOutput <= 0.79: Bad (83.0/10.0)
|   productionOutput > 0.79: Normal (885.0/59.0)
productionYield > 94.99
|   qcOutput <= 0.79: Normal (66.0/23.0)
|   qcOutput > 0.79: Good (1107.0/52.0)

Number of Leaves: 4
Size of the tree: 7
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===
```

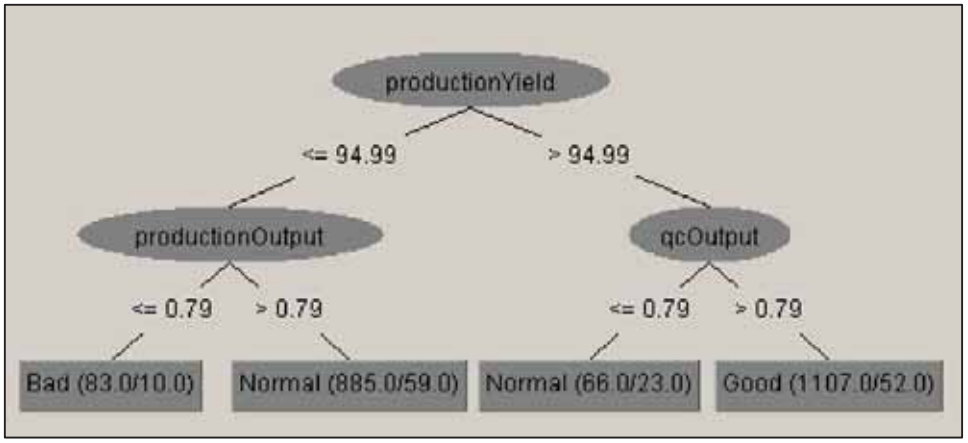
```
Correctly Classified Instances        1994               93.134  %
Incorrectly Classified Instances       147                6.866  %
Kappa statistic                          0.8735
Mean absolute error                      0.0837
Root mean squared error                  0.206
Relative absolute error                 22.6781 %
Root relative squared error             47.9576 %
Coverage of cases (0.95 level)          98.9724 %
Mean rel. region size (0.95 level)      61.0462 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.995	0.069	0.936	0.995	0.965	0.968	Good
0.553	0.005	0.88	0.553	0.679	0.883	Bad
0.911	0.053	0.93	0.911	0.92	0.931	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1073    0    5 |    a = Good
    0   73   59 |    b = Bad
   73   10  848 |    c = Normal
```



ภาพที่ 34 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 5

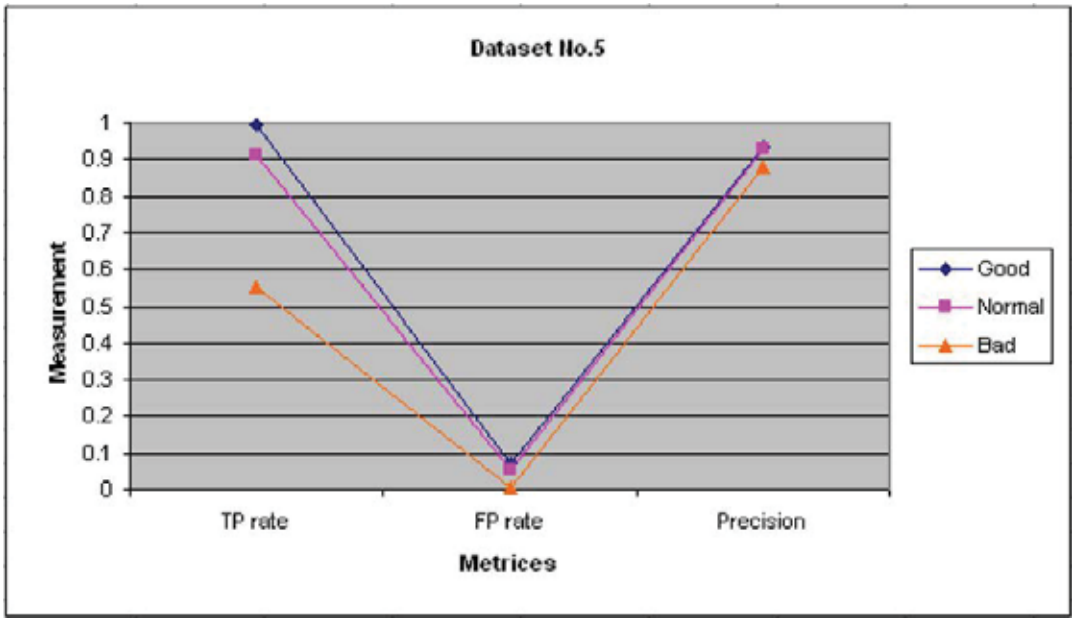
สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 5

ผลลัพธ์ที่ได้จากการทดสอบชุดข้อมูลที่ 5 นี้ ให้ผลลัพธ์เท่ากับการทดสอบการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 2 เนื่องจากสามารถจำแนกข้อมูลในแต่ละคลาสได้ในจำนวนที่เท่ากัน แม้จะกำหนดเงื่อนไขที่ต่างกัน กล่าวคือ

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 1,994 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 93.134 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,073 ระเบียน อยู่ในคลาส Normal จำนวน 848 ระเบียน และในคลาส Bad จำนวน 73 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 147 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 6.866 % โดยข้อมูลทั้ง 147 ระเบียนที่จำแนกผิดนั้น คือ

- มี 5 ระเบียนที่อยู่ในคลาส Good แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal จำนวน 5 ระเบียน
- มี 83 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good จำนวน 73 ระเบียน และในคลาส Bad จำนวน 10 ระเบียน
- มี 59 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal ทั้ง 59 ระเบียน



ภาพที่ 35 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 5

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 35 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Good (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.069 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Normal โดยมีอัตราการจำแนกผิดเท่ากับ 0.053 และคลาส Bad มีอัตราการจำแนกผิดที่ 0.005

ตารางที่ 22 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 6

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Normal
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Good
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Good
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Bad
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 6 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R06
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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


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

productionYield <= 94.99
|   productionOutput <= 0.79: Bad (83.0/10.0)
|   productionOutput > 0.79: Normal (885.0/59.0)
productionYield > 94.99: Good (1173.0)

Number of Leaves: 3
Size of the tree: 5
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2071               96.7305 %
Incorrectly Classified Instances        70                3.2695 %
```

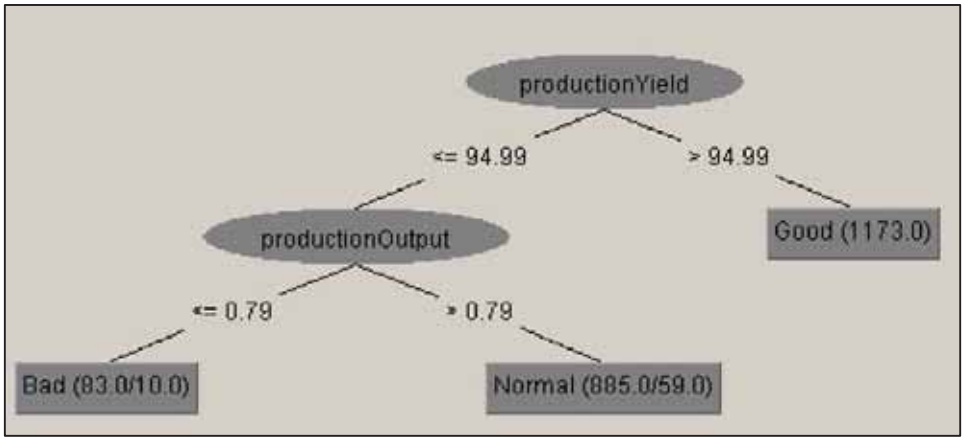
```
Kappa statistic                          0.939
Mean absolute error                      0.0398
Root mean squared error                  0.142
Relative absolute error                 10.9759 %
Root relative squared error             33.3628 %
Coverage of cases (0.95 level)          99.9533 %
Mean rel. region size (0.95 level)      48.2641 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.001	0.999	1	1	0.999	Good
0.545	0.004	0.889	0.545	0.676	0.89	Bad
0.988	0.046	0.932	0.988	0.959	0.973	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1173    0    0 |    a = Good
    0   72   60 |    b = Bad
    1    9  826 |    c = Normal
```



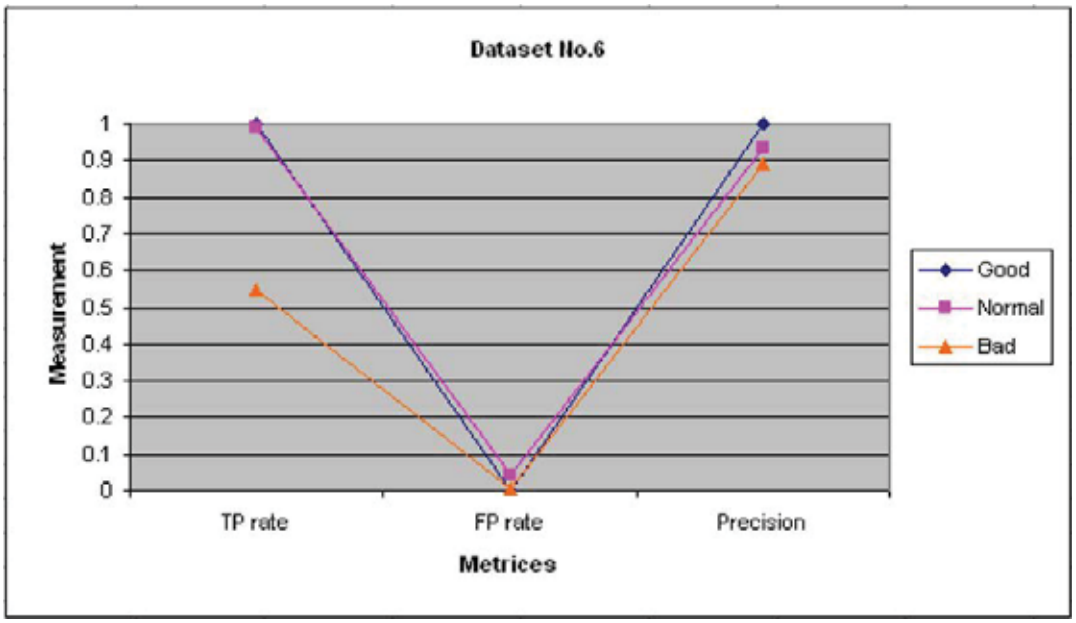
ภาพที่ 36 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 6

สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 6

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 2,071 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 96.7305 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,173 ระเบียน อยู่ในคลาส Normal จำนวน 826 ระเบียน และในคลาส Bad จำนวน 72 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 70 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 3.2695 % โดยข้อมูลทั้ง 70 ระเบียนที่จำแนกผิดนั้น คือ

- มี 10 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good จำนวน 1 ระเบียน และในคลาส Bad จำนวน 9 ระเบียน
- มี 60 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal ทั้ง 60 ระเบียน



ภาพที่ 37 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 6

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 37 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Normal (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.046 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad โดยมีอัตราการจำแนกผิดเท่ากับ 0.004 และคลาส Good มีอัตราการจำแนกผิดเท่ากับ 0.001

ตารางที่ 23 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 7

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Normal
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Good
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Good
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 7 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R07
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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


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


productionYield <= 94.99
|   productionOutput <= 0.79: Bad (83.0/10.0)
|   productionOutput > 0.79: Normal (885.0/59.0)
productionYield > 94.99: Good (1173.0)

Number of Leaves: 3
Size of the tree: 5
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2071               96.7305 %
```

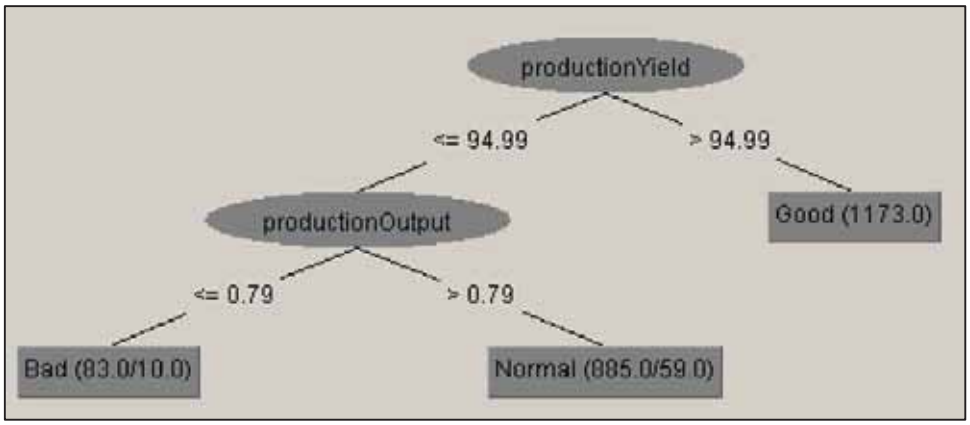
```
Incorrectly Classified Instances        70               3.2695 %
Kappa statistic                          0.939
Mean absolute error                      0.0398
Root mean squared error                  0.142
Relative absolute error                 10.9759 %
Root relative squared error             33.3628 %
Coverage of cases (0.95 level)          99.9533 %
Mean rel. region size (0.95 level)      48.2641 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.001	0.999	1	1	0.999	Good
0.545	0.004	0.889	0.545	0.676	0.89	Bad
0.988	0.046	0.932	0.988	0.959	0.973	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1173    0    0 |    a = Good
    0   72   60 |    b = Bad
    1    9  826 |    c = Normal
```



ภาพที่ 38 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 7

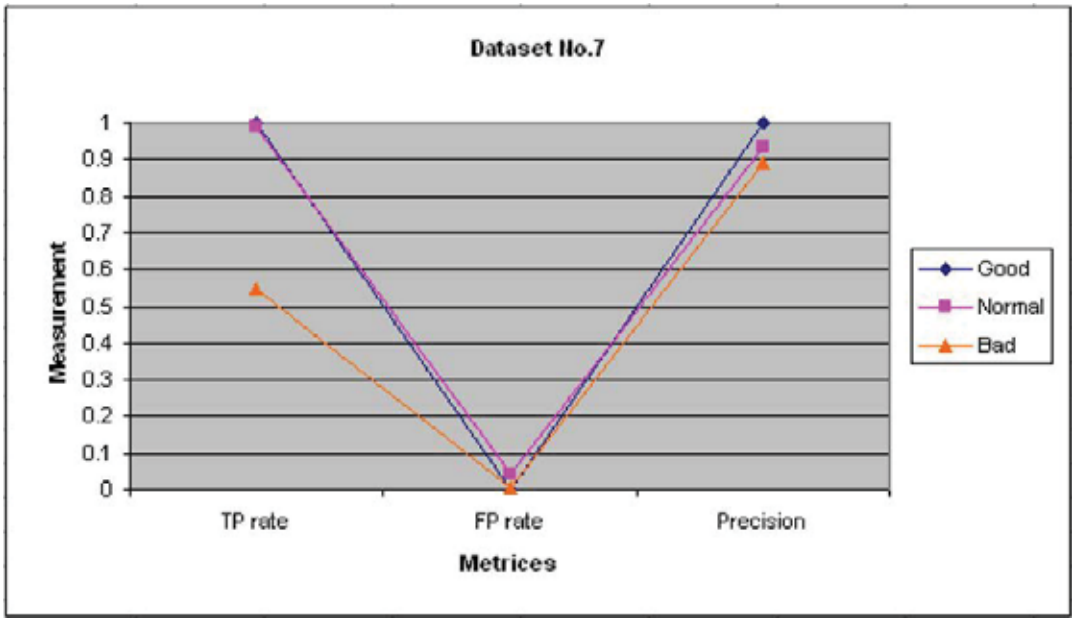
สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 7

ผลลัพธ์ที่ได้จากการทดสอบชุดข้อมูลที่ 7 นี้ ให้ผลลัพธ์เท่ากับการทดสอบการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 6 เนื่องจากสามารถจำแนกข้อมูลในแต่ละคลาสได้ในจำนวนที่เท่ากัน แม้จะกำหนดเงื่อนไขที่ต่างกัน กล่าวคือ

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 2,071 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 96.7305 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,173 ระเบียน อยู่ในคลาส Normal จำนวน 826 ระเบียน และในคลาส Bad จำนวน 72 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 70 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 3.2695 % โดยข้อมูลทั้ง 70 ระเบียนที่จำแนกผิดนั้น คือ

- มี 10 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good จำนวน 1 ระเบียน และในคลาส Bad จำนวน 9 ระเบียน
- มี 60 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal ทั้ง 60 ระเบียน



ภาพที่ 39 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 7

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 39 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Normal (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.046 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad โดยมีอัตราการจำแนกผิดเท่ากับ 0.004 และคลาส Good มีอัตราการจำแนกผิดเท่ากับ 0.001

ตารางที่ 24 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 8

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Normal
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 8 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R08
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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


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

productionYield <= 94.99: Bad (968.0)
productionYield > 94.99
|   qcOutput <= 0.79: Normal (66.0/23.0)
|   qcOutput > 0.79: Good (1107.0/52.0)

Number of Leaves: 3
Size of the tree: 5
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2055               95.9832 %
```

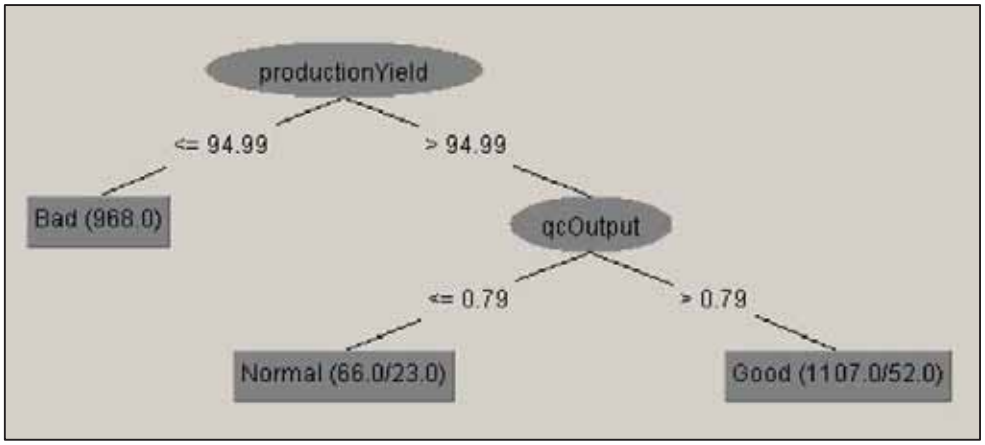
```
Incorrectly Classified Instances        86               4.0168 %
Kappa statistic                          0.9237
Mean absolute error                      0.0469
Root mean squared error                  0.1556
Relative absolute error                 13.0169 %
Root relative squared error             36.6689 %
Coverage of cases (0.95 level)          98.3185 %
Mean rel. region size (0.95 level)      41.3358 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.99	0.071	0.934	0.99	0.961	0.968	Good
0.999	0	1	0.999	0.999	0.999	Bad
0.221	0.005	0.656	0.221	0.331	0.815	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1067    0   11 |    a = Good
    1  967    0 |    b = Bad
   74    0   21 |    c = Normal
```



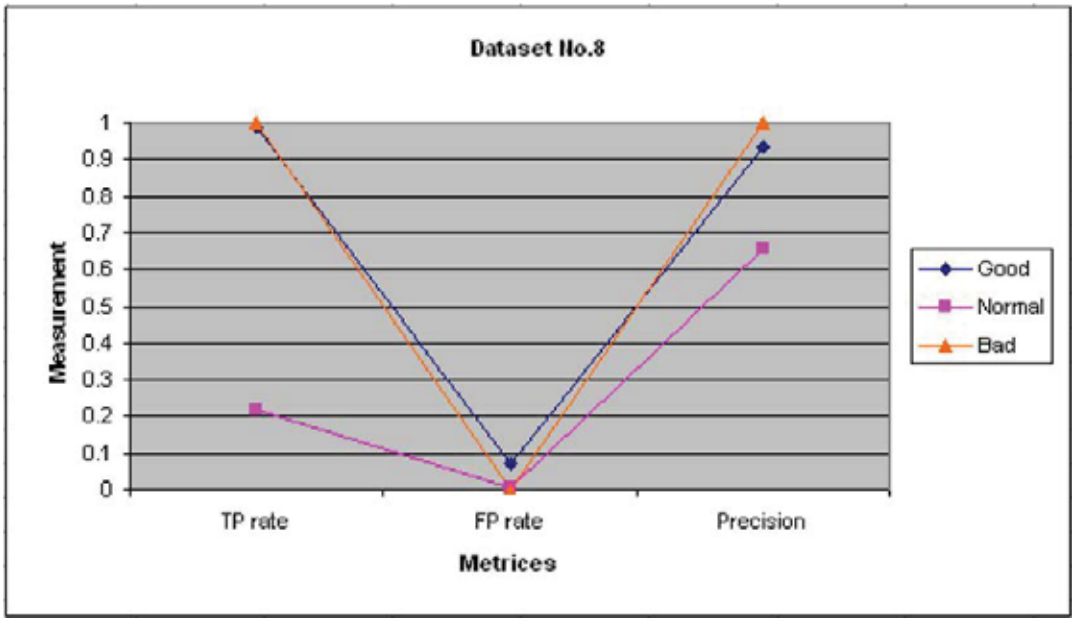
ภาพที่ 40 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 8

สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 8

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 2,055 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 95.9832 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,067 ระเบียน อยู่ในคลาส Normal จำนวน 21 ระเบียน และในคลาส Bad จำนวน 967 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 86 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 4.0168 % โดยข้อมูลทั้ง 86 ระเบียนที่จำแนกผิดนั้น คือ

- มี 11 ระเบียนที่อยู่ในคลาส Good แต่โปรแกรมจำแนกให้อยู่ในคลาส Normal จำนวน 11 ระเบียน
- มี 74 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good ทั้ง 74 ระเบียน
- มี 1 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Good ทั้ง 1 ระเบียน



ภาพที่ 41 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 8

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 41 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Good (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.071 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad และ Normal โดยมีอัตราการจำแนกผิดเท่ากับ 0.0 และ 0.005 ตามลำดับ

ตารางที่ 25 เงื่อนไขการจำแนกกลุ่มแม่พิมพ์ชุดข้อมูลที่ 9

การดูแลรักษาแม่พิมพ์ (ยอดผลิต)	ยอดการผลิต (Production Yield)	ยอดงานดี (Final Yield)	กลุ่มแม่พิมพ์ (Tool Class)
ง่าย (ยอดผลิต >= Min)	มาก (Production yield >=95%)	มาก (Final yield >= 95%)	Good
ง่าย (ยอดผลิต >=Min)	มาก (Production yield >=95%)	น้อย (Final yield < 95%)	Good
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	มาก (Final yield >=95%)	Normal
ง่าย (ยอดผลิต >= Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	มาก (Final yield >= 95%)	Good
ยาก (ยอดผลิต < Min)	มาก (Production yield >= 95%)	น้อย (Final yield < 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	มาก (Final yield >= 95%)	Normal
ยาก (ยอดผลิต < Min)	น้อย (Production yield < 95%)	น้อย (Final yield < 95%)	Bad

ผลลัพธ์จากโปรแกรม Weka เมื่อจำแนกชุดข้อมูลที่ 9 ด้วยอัลกอริทึม C4.5 (โมดูล J48) แสดงได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     R09
Instances:    2141
Attributes:   10
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, toolClass
Test mode:    10-fold cross-validation

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


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

productionYield <= 94.99: Bad (968.0)
productionYield > 94.99: Good (1173.0/10.0)

Number of Leaves: 2
Size of the tree: 3
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2130               99.4862 %
Incorrectly Classified Instances        11                0.5138 %
Kappa statistic                          0.9897
Mean absolute error                      0.0065
```

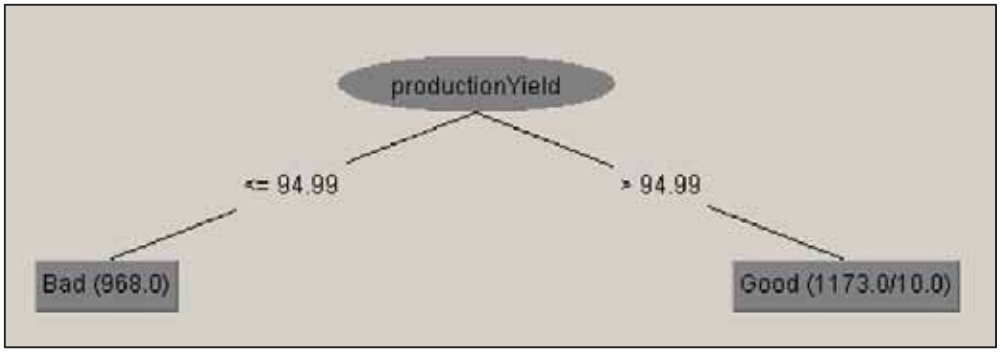
```
Root mean squared error                  0.0583
Relative absolute error                  1.9429 %
Root relative squared error             14.2675 %
Coverage of cases (0.95 level)          99.4862 %
Mean rel. region size (0.95 level)      33.3333 %
Total Number of Instances             2141
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.011	0.991	1	0.995	0.994	Good
0.999	0	1	0.999	0.999	0.999	Bad
0	0	0	0	0	0.726	Normal

=== Confusion Matrix ===

```
    a    b    c   <-- classified as
 1163    0    0 |    a = Good
    1  967    0 |    b = Bad
   10    0    0 |    c = Normal
```



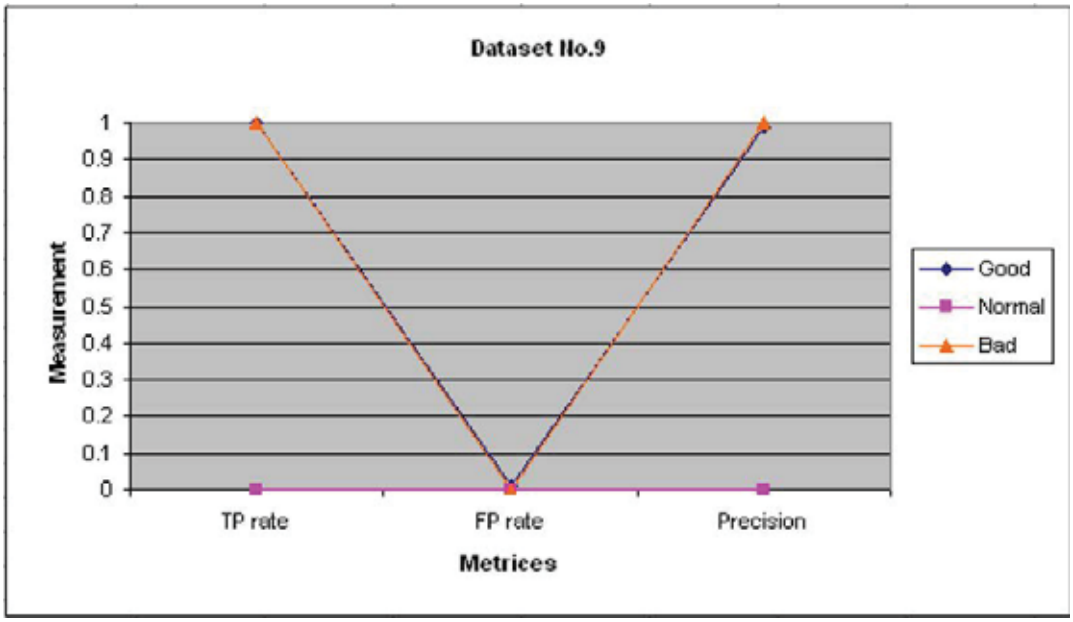
ภาพที่ 42 ผลการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 กับชุดข้อมูลที่ 9

สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจในชุดข้อมูลที่ 9

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 2,130 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 99.4862 % โดยจำแนกอยู่ในคลาส Good จำนวน 1,163 ระเบียน อยู่ในคลาส Normal จำนวน 0 ระเบียน และในคลาส Bad จำนวน 967 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 11 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 0.5138 % โดยข้อมูลทั้ง 11 ระเบียนที่จำแนกผิดนั้น คือ

- มี 10 ระเบียนที่อยู่ในคลาส Normal แต่โปรแกรมจำแนกให้อยู่ในคลาส Good ทั้ง 10 ระเบียน
- มี 1 ระเบียนที่อยู่ในคลาส Bad แต่โปรแกรมจำแนกให้อยู่ในคลาส Good ทั้ง 1 ระเบียน



ภาพที่ 43 กราฟแสดงค่าการวัดคุณภาพการจำแนกข้อมูลชุดที่ 9

จากค่า TP rate, FP rate และ Precision จะได้กราฟดังภาพที่ 43 เมื่อพิจารณาค่าดังกล่าวจะพบว่าต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทางคลาส Good (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.011 และเมื่อกลับไปพิจารณาที่คลาสดังกล่าวพบว่าสัญญาณหลอกเกิดขึ้นที่คลาส Bad และ Normal โดยคลาส Bad มีอัตราการจำแนกผิดเท่ากับ 0.0 และคลาส Normal เป็นการจำแนกผิดทั้งจำนวน

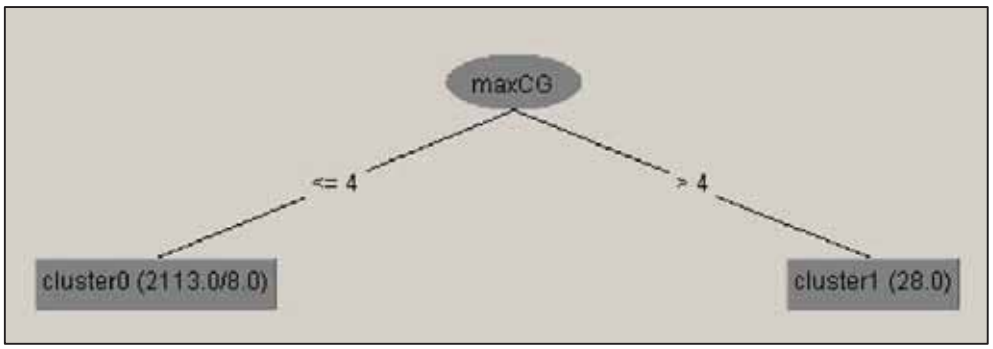
ภาคผนวก ค
ผลการวัดคุณภาพการจัดกลุ่มแบบลำดับชั้น

ผลการวัดคุณภาพการจัดกลุ่มแบบลำดับชั้น
ด้วยอัลกอริทึม C4.5

ผลการวัดคุณภาพการจัดกลุ่มแบบลำดับ (Hierarchical Clustering) ด้วยอัลกอริทึม C4.5 โดยกำหนดให้ Cluster เป็นแอตทริบิวท์เป้าหมาย แสดงผลได้ดังนี้

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     ClusteringR10
Instances:    2141
Attribute name: toolId, toolItem, toolOption, minCG, maxCG,
                productionOutput, qcOutput, productionYield,
                finalYield, cluster
Test mode:    10-fold cross-validation

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

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

maxCG <= 4: cluster1 (2113.0/8.0)
maxCG > 4: cluster2 (28.0)

Number of Leaves: 2
Size of the tree: 3
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2134               99.673  %
Incorrectly Classified Instances         7                0.327  %
Kappa statistic                          0.8909
Mean absolute error                      0.0044
Root mean squared error                  0.0466
Relative absolute error                 19.384  %
Root relative squared error             44.3266 %
Coverage of cases (0.95 level)          99.673  %
Mean rel. region size (0.95 level)      33.3333 %
Total Number of Instances             2141

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
1         0.194      0.997     1         0.998      0.893    cluster0
1         0          1         1         1          1        cluster1
0.125     0          1         0.125     0.222      0.524    cluster2


=== Confusion Matrix ===
```

```
    a     b     c   <-- classified as
 2105     0     0 |    a = cluster0
    0    28     0 |    b = cluster1
    7     0     1 |    c = cluster2
```



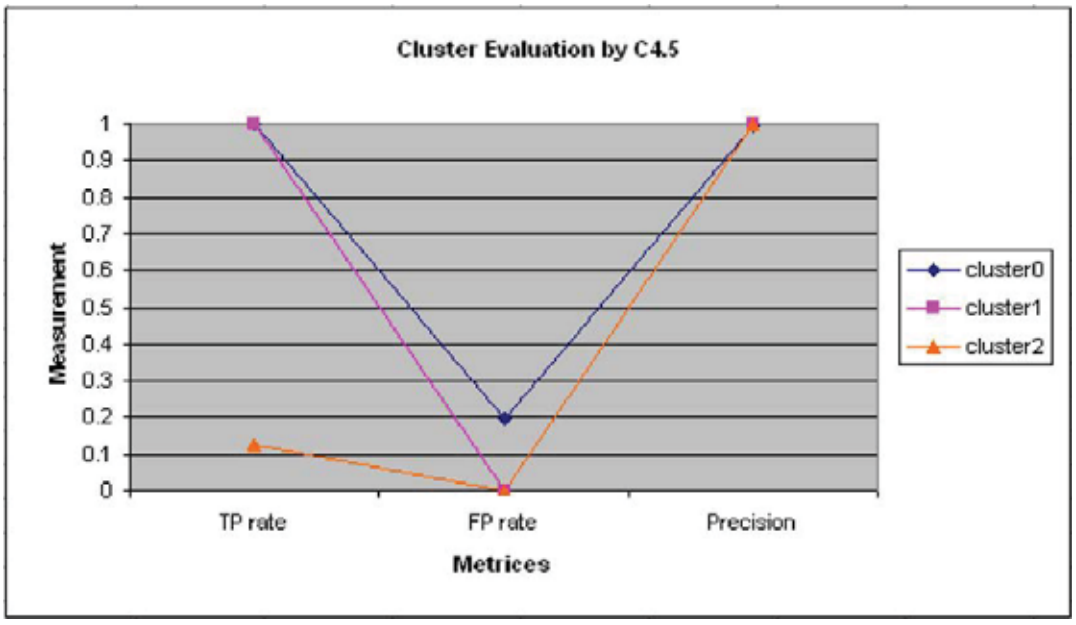
ภาพที่ 44 ผลการวัดคุณภาพการจัดกลุ่มแบบลำดับชั้นด้วยอัลกอริทึม C4.5

สรุปผลที่ได้จากการสร้างต้นไม้ตัดสินใจเมื่อกำหนดให้ Cluster เป็นแอตทริบิวท์เป้าหมาย

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกถูกต้อง 2,134 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน และมีเปอร์เซ็นต์ความถูกต้องที่ 99.673 % โดยจำแนกอยู่ใน cluster0 จำนวน 2,105 ระเบียน อยู่ใน cluster1 จำนวน 28 ระเบียน และใน cluster2 จำนวน 1 ระเบียน

ผลลัพธ์ที่ได้นั้น มีระเบียนที่จำแนกผิดจำนวน 7 ระเบียน จากข้อมูลทั้งหมด 2,141 ระเบียน คิดเป็น 0.327 % โดยข้อมูล 1 ระเบียนที่จำแนกผิดนั้น คือ

- มี 7 ระเบียนที่อยู่ใน cluster2 แต่โปรแกรมจำแนกให้อยู่ cluster0 ทั้ง 7 ระเบียน



ภาพที่ 45 กราฟแสดงค่าการวัดคุณภาพการจัดกลุ่มแบบลำดับชั้นด้วยอัลกอริทึม C4.5

เมื่อพิจารณาค่า TP rate, FP rate และ Precision จะพบว่าทั้ง 3 cluster ค่อนข้างมีความน่าเชื่อถือเนื่องจากมีค่า TP rate และ Precision ที่สูง และมีค่า FP rate ที่ต่ำ โดยต้นไม้ตัดสินใจค่อนข้างเอนเอียงไปทาง cluster0 (พิจารณาจากค่า FP rate) โดยมีอัตราการจำแนกผิดเท่ากับ 0.194 เมื่อกลับไปพิจารณาที่คลาสดังกล่าว พบว่าสัญญาณหลอกเกิดขึ้นที่ cluster 2 ส่วนคลาสที่เหลือไม่มีอัตราการจำแนกผิด

ภาคผนวก ง
ตัวอย่างการแปลงข้อมูล (Discretization)

ตัวอย่างการแปลงข้อมูล (Discretization)

ชนิดตัวเลข (Numeric) เป็นข้อมูลชนิดไม่ต่อเนื่อง (Nominal)

ตารางที่ 26 ตัวอย่างการแปลงข้อมูลชนิดตัวเลขเป็นข้อมูลชนิดไม่ต่อเนื่อง

แอตทริบิวท์	รายละเอียด	ช่วงข้อมูล	จำนวนระเบียน
minCG	จำนวนการผลิตขั้นต่ำที่ต้องบำรุงรักษาแม่พิมพ์ (หน่วย: เค-เค-สโตรค)	**M1** จำนวนการผลิตขั้นต่ำน้อยกว่า 1.99	2,022
		M2 จำนวนการผลิตขั้นต่ำอยู่ในช่วง 2.00 – 3.99	91
		M3 จำนวนการผลิตขั้นต่ำมากกว่า 4.00	28
maxCG	จำนวนการผลิตขั้นสูงที่ต้องบำรุงรักษาแม่พิมพ์ (หน่วย: เค-เค-สโตรค)	**X1** จำนวนการผลิตขั้นสูงน้อยกว่า 2.99	2,043
		X2 จำนวนการผลิตขั้นสูงอยู่ในช่วง 3.00 – 5.99	70
		X3 จำนวนการผลิตขั้นสูงมากกว่า 6.00	28
productionOutput	ผลรวมของยอดการผลิตของรอบการซ่อมบำรุงในแต่ละครั้ง (หน่วย: เค-เค-สโตรค)	**L1** ยอดการผลิต < จำนวนการผลิตขั้นต่ำ (minCG)	227
		L2 ยอดการผลิต >= จำนวนการผลิตขั้นต่ำ (minCG) และ ยอดการผลิต < จำนวนการผลิตขั้นสูง (maxCG)	846
		L3 ยอดการผลิต >= จำนวนการผลิตขั้นสูง (maxCG)	1,068
qcOutput	ผลรวมของยอดงานดีของรอบการซ่อมบำรุงในแต่ละครั้ง (หน่วย: เค-เค-สโตรค)	**Hard** ยอดงานดี < จำนวนการผลิตขั้นต่ำ (minCG)	295
		Easy ยอดงานดี >= จำนวนการผลิตขั้นสูง (maxCG)	1,846

ตารางที่ 26 (ต่อ)

แอตทริบิวท์	รายละเอียด	ช่วงข้อมูล	จำนวนระเบียน
productionYield	อัตราผลผลิตทั้งหมดที่เกิดจากกระบวนการผลิต (เปอร์เซ็นต์)	**Low** อัตราผลผลิต < 90%	364
		Medium อัตราผลผลิตอยู่ในช่วง >= 90% และ < 95%	604
		High อัตราผลผลิต >= 95%	1,173
finalYield	อัตราผลผลิตของยอดงานดีที่เกิดจากกระบวนการผลิต (เปอร์เซ็นต์)	**Loss** อัตราผลผลิตของยอดงานดี < 95%	1,174
		Success อัตราผลผลิตของยอดงานดี >= 95%	967

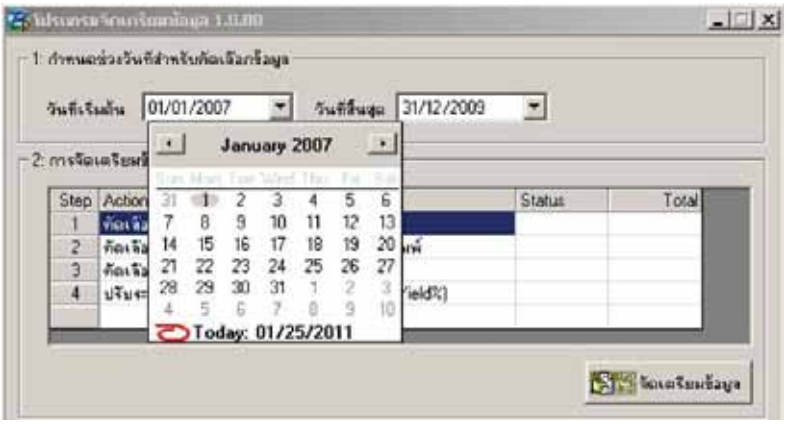
ภาคผนวก จ
คู่มือการใช้งานโปรแกรมจัดเตรียมข้อมูล

คู่มือการใช้งาน
โปรแกรมจัดเตรียมข้อมูล

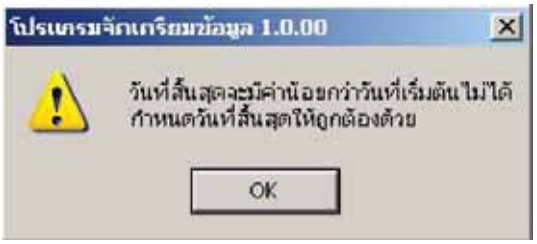
คู่มือการใช้งานโปรแกรมจัดเตรียมข้อมูลจัดทำขึ้นเพื่อให้ผู้ที่ใช้งานเข้าใจการทำงานของโปรแกรมจัดเตรียมข้อมูลได้อย่างถูกต้อง โดยมีวิธีการใช้งานดังนี้

1. การกำหนดช่วงข้อมูลและจัดเตรียมข้อมูล

1.1 กำหนดช่วงวันที่ที่ต้องการคัดเลือกข้อมูล ในช่องวันที่เริ่มต้น และวันที่สิ้นสุด โดยกำหนดเป็น วัน/เดือน/ปี ดังภาพที่ 1 และวันที่สิ้นสุดห้ามมีค่าวันที่น้อยกว่าวันที่เริ่มต้น หากกำหนดผิดโปรแกรมจะฟ้องข้อความแจ้งเตือนดังภาพที่ 2



ภาพที่ 1 การกำหนดช่วงวันที่เริ่มต้นและสิ้นสุด



ภาพที่ 2 ข้อความแจ้งเตือนเมื่อกำหนดวันที่สิ้นสุดไม่ถูกต้อง

1.2 เมื่อกำหนดช่วงวันที่เสร็จแล้ว ให้กดปุ่ม “จัดเตรียมข้อมูล” โดยโปรแกรมจะยุติการใช้งานปุ่ม “จัดเตรียมข้อมูล” ชั่วคราว เพื่อป้องกันการกดซ้ำ จากนั้นจะเริ่มกระบวนการจัดเตรียมข้อมูล โดยแบ่งออกเป็น 4 กระบวนการย่อย คือ

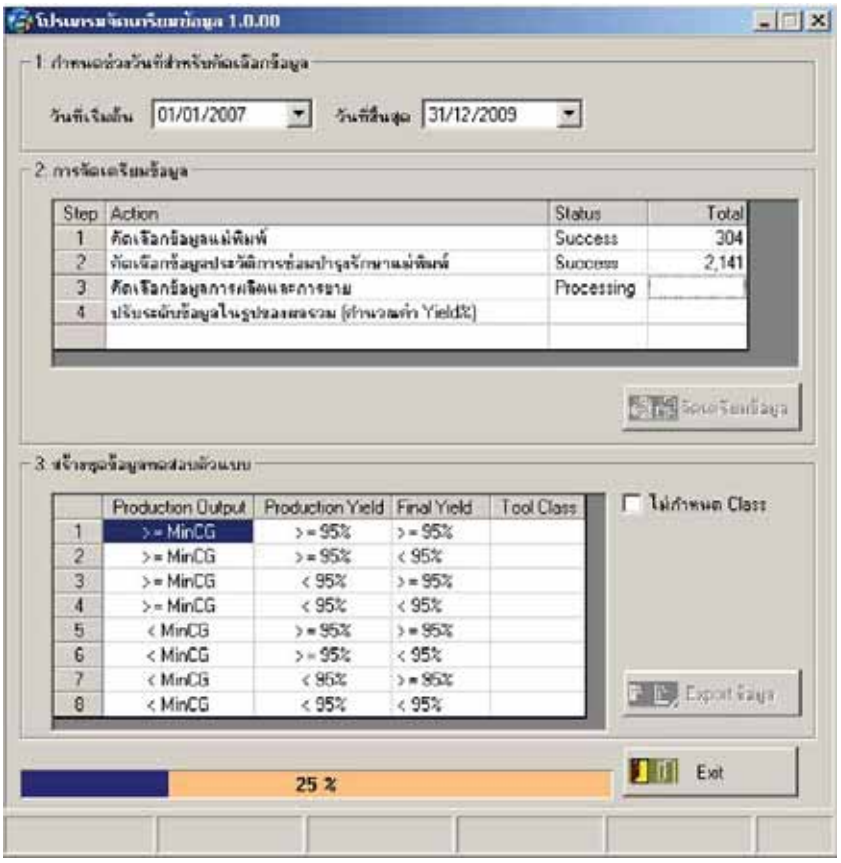
1.2.1 กระบวนการคัดเลือกข้อมูลแม่พิมพ์ คือ กระบวนการคัดเลือกข้อมูลแม่พิมพ์จากฐานข้อมูล Progress 9.1C ของระบบ ERP ที่มีสถานะใช้งานในปัจจุบัน (Active)

1.2.2 กระบวนการคัดเลือกข้อมูลประวัติการซ่อมบำรุงรักษาแม่พิมพ์ คือ กระบวนการคัดเลือกข้อมูลประวัติซ่อมบำรุงรักษาแม่พิมพ์จากฐานข้อมูล Microsoft SQL Server 2005 ของระบบซ่อมบำรุงรักษาแม่พิมพ์ โดยคัดเลือกประวัติการซ่อมบำรุงรักษาแม่พิมพ์ที่สอดคล้องกับข้อมูลแม่พิมพ์ที่ได้จากกระบวนการทำงานที่ 1.2.1

1.2.3 กระบวนการคัดเลือกข้อมูลการผลิตและการขาย คือ กระบวนการคัดเลือกข้อมูลการผลิตและการขายจากฐานข้อมูล Progress 9.1C ของระบบ ERP ที่สอดคล้องกับรอบการบำรุงรักษาแม่พิมพ์ที่ได้จากกระบวนการทำงานที่ 1.2.2

1.2.4 กระบวนการปรับระดับข้อมูลในรูปของผลรวม คือ การคำนวณหาค่ายอดการผลิต (Production yield) และยอดงานดี (Final yield) ซึ่งก็คือยอดขาย ที่สอดคล้องกับข้อมูลการผลิตและการขายที่ได้จากกระบวนการทำงานที่ 1.2.3

ตัวอย่างผลการของทำงานในกระบวนการต่างๆ แสดงได้ดังภาพที่ 3

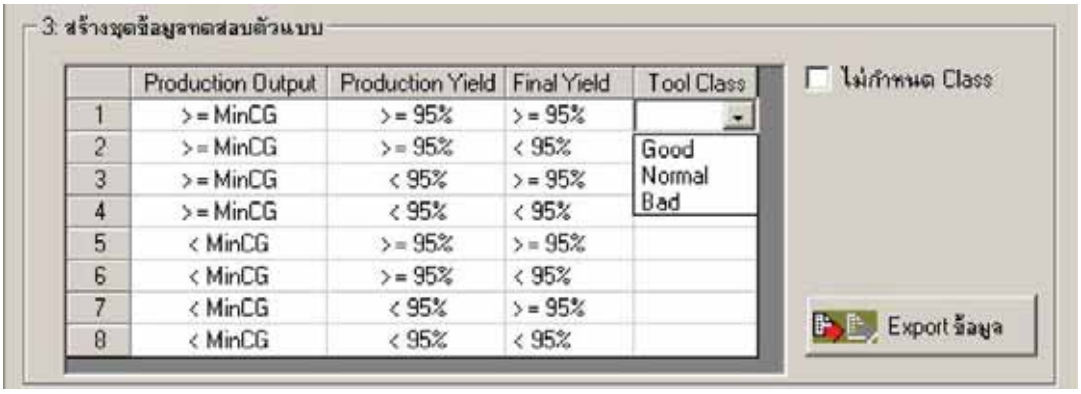


ภาพที่ 3 แสดงการทำงานของกระบวนการจัดเตรียมข้อมูล

2. การสร้างชุดข้อมูล

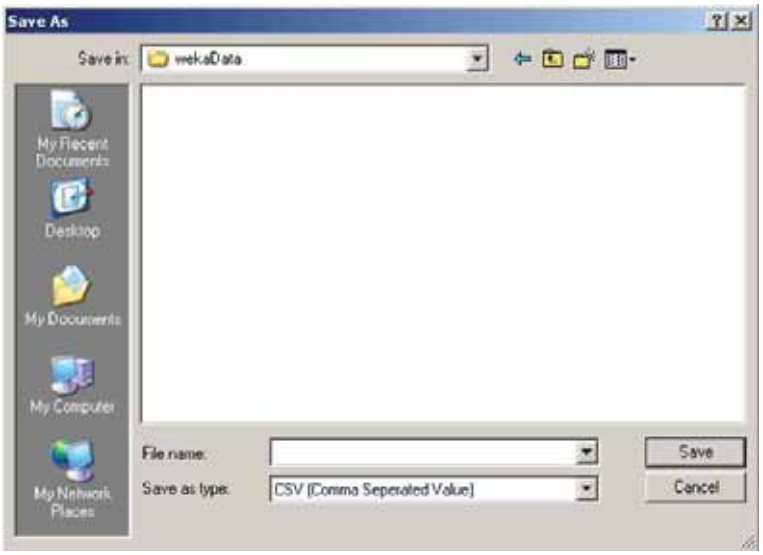
2.1 การสร้างชุดข้อมูลสำหรับการจำแนกข้อมูลด้วยโครงสร้างต้นไม้ตัดสินใจ (Decision Tree)

2.1.1 กำหนดกลุ่มแม่พิมพ์ให้ครบทั้ง 8 เงื่อนไข ด้วยการกดเลือกกลุ่มแม่พิมพ์ในช่อง Tool Class โดยไม่ต้องคลิกเลือกที่ช่อง “ไม่กำหนด Class” จากนั้นให้กดปุ่ม “Export ข้อมูล” ดังภาพที่ 4



ภาพที่ 4 การสร้างชุดข้อมูลพร้อมกับการกำหนดกลุ่มแม่พิมพ์

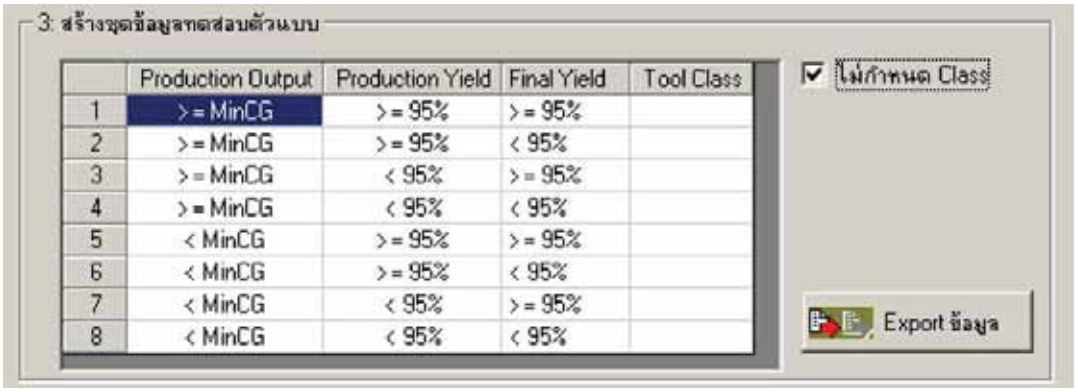
2.1.2 โปรแกรมจะแสดงหน้าต่าง “Save As” ดังภาพที่ 5 ให้ทำการกำหนดชื่อไฟล์และเลือกสถานที่จัดเก็บไฟล์ จากนั้นให้กดปุ่ม “Save”



ภาพที่ 5 การบันทึกชุดข้อมูลสำหรับทดสอบตัวแบบ

2.2 การสร้างชุดข้อมูลสำหรับการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

2.2.1 คลิกเลือกที่ช่อง “ไม่กำหนด Class” จากนั้นให้กดปุ่ม “Export ข้อมูล” ดังภาพที่ 6 โดยที่ไม่ต้องทำการกำหนดกลุ่มแม่พิมพ์ในช่อง Tool Class



ภาพที่ 6 การสร้างชุดข้อมูลสำหรับการจัดกลุ่มแบบลำดับชั้น

2.2.2 โปรแกรมจะแสดงหน้าต่าง “Save As” ดังภาพที่ 5 ให้ทำการกำหนดชื่อไฟล์และเลือกสถานที่จัดเก็บไฟล์ จากนั้นให้กดปุ่ม “Save”

แสดงตัวอย่างผลการสร้างชุดข้อมูลได้ดังภาพที่ 7

A	B	C	D	E	F	G	H	I	J
toolId	toolItem	toolOption	minCG	maxCG	productionOutput	qcOutput	productionYield	finalYield	toolClass
DLS0303502	7L1281DLS0303602	DLS0303602	1.28	1.6	1.22	1.22	96.52	96.52	Good
DLS0303502	7L1281DLS0303602	DLS0303602	1.28	1.6	1.9	1.9	96.18	96.18	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.45	1.45	97.9	97.9	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	0.89	0.87	94.54	93.9	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	0.89	0.89	86.82	86.82	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.59	1.49	92.56	86.31	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.42	1.42	90.11	90.11	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.29	1.29	95.59	95.59	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.37	1.37	93.94	93.94	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.09	1.09	93.52	93.52	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.07	1.07	93.83	93.83	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.25	1.25	96.15	96.15	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.47	1.47	96.02	96.02	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.6	1.6	94.06	94.06	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	95.51	95.51	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	0.97	0.95	88.97	86.68	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.01	1.01	95.02	95.02	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.41	1.41	94.31	94.31	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.14	1.14	91.67	91.67	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.09	1.08	94.37	93.33	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.39	1.36	95.59	95.59	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	91.97	91.97	Bad
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.76	1.76	95.22	95.22	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.49	1.49	111.7	111.7	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	1.01	1.01	95.02	95.02	Good
DLS0304502-602	7L1512DLS0304502-602	DLS0304502	0.9	1.4	18.85	18.73	94.12	93.46	Bad

ภาพที่ 7 ตัวอย่างผลการสร้างชุดข้อมูลจากโปรแกรมจัดเตรียมข้อมูล

ภาคผนวก ฉ
หนังสือขอความอนุเคราะห์ขอข้อมูลเพื่อใช้ในการศึกษาวิจัย

ที่ ศธ 0520.107 / 1715

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
22 ถนนบรมราชชนนี เขตตลิ่งชัน กรุงเทพฯ 10170

๒๙ มีนาคม 2554

เรื่อง ขอความอนุเคราะห์ขอข้อมูล

เรียน ผู้จัดการฝ่าย Business บริษัท ยามาดา (ประเทศไทย) จำกัด

ด้วย นายแสนศักดิ์ ชาวปากน้ำ นักศึกษาระดับบัณฑิตศึกษา สาขาวิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร กำลังทำการค้นคว้าอิสระ เรื่อง "การเปรียบเทียบเทคนิคการจำแนกข้อมูลและการรวมกลุ่มข้อมูลในการคัดเลือกแม่พิมพ์โลหะแบบ Progressive Die" มีความประสงค์จะขอข้อมูลการผลิตตั้งแต่ปีพ.ศ. 2550 – 2552 เพื่อใช้เป็นข้อมูลประกอบการทำการค้นคว้าอิสระ ในการนี้บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร จึงขอความอนุเคราะห์จากท่านโปรดอนุญาต ตามความละเอียดแจ้งแล้วนั้น ให้แก่นักศึกษาดังกล่าวด้วย

จึงเรียนมาเพื่อโปรดให้ความอนุเคราะห์ จักขอบพระคุณยิ่ง

ขอแสดงความนับถือ

(ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์)
คณบดีบัณฑิตวิทยาลัย

สำนักงานบัณฑิตวิทยาลัย ตลิ่งชัน
โทร. / โทรสาร 0-2849-7503

ประวัติผู้วิจัย

ชื่อ-สกุล	นายแสนศักดิ์ ชาวปากน้ำ
ที่อยู่	46/18 ม.1 ต.ท่าข้าม อ.สามพราน จ.นครปฐม 73110

ประวัติการศึกษา

พ.ศ. 2542	สำเร็จการศึกษาปริญญาบัณฑิต สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ มหาวิทยาลัยเอเชียอาคเนย์
พ.ศ. 2549	ศึกษาต่อระดับปริญญามหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ประวัติการทำงาน

พ.ศ. 2542 – 2543	เจ้าหน้าที่อบรมหลักสูตรคอมพิวเตอร์ บริษัท เวบ เน็ต เทค จำกัด
พ.ศ. 2543 – 2547	โปรแกรมเมอร์และนักวิเคราะห์ระบบ บริษัท ซาบีน่า ฟาร์อีสท์ จำกัด
พ.ศ. 2547 – 2548	โปรแกรมเมอร์และนักวิเคราะห์ระบบ บริษัท จีเนียส คอมมิวนิเคชั่น ซิสเต็ม จำกัด
พ.ศ. 2548-ปัจจุบัน	โปรแกรมเมอร์และนักวิเคราะห์ระบบ บริษัท อาพิค ยามาดะ (ประเทศไทย) จำกัด