ปัจจัยที่มีอิทธิพลต่อความแม่นยำของการจำแนกข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล เปรียบเทียบระหว่างอัลกอริธึม Naïve Bayes Simples และ Id3 กรณีศึกษาธนาคารกรุงศรีอยุธยา จำกัด(มหาชน)

ปิยะมาศ กรัณย์ภควุฒิ[1] และ นันทิกา ปริญญาพล[2]

[1]คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์
[2]หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศสำหรับวิสาหกิจสมัยใหม่
คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์
Emails: karan4064@hotmail.com, nantika.pri@dpu.ac.th,

บทคัดย่อ

บทความนี้นำเสนอผลการศึกษา ทดลอง วิเคราะห์และสรุปผลเรื่องปัจจัยที่มีอิทธิพลต่อความแม่นยำของการจำแนกข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล โดยทำการวิเคราะห์และสรุปผลจากผลลัพธ์ที่ได้จากการประมวลผลโดยโปรแกรม Waikato Environment for Knowledge Analysis(WEKA) เวอร์ชัน 3.6.10 [1] ซึ่งผู้วิจัยได้ให้ความสำคัญกับการวิเคราะห์และเปรียบค่าความถูกต้องแม่นยำของโมเดลเพื่อการจำแนกข้อมูลที่สร้างขึ้นจากสองอัลกอริธึมคือ แบบ Naïve Bayes Simples และแบบ Id3 ทั้งสองอัลกอริธึมเป็นอัลกอริธึมที่อยู่ในกลุ่มการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลโดยวิธีจำแนกข้อมูล(Classify) ประเภทการเรียนรู้ของเครื่องจักรแบบมีครู(Supervised Leaning)

วัตถุประสงค์ของการทดลองนี้ เพื่อพิสูจน์สมมติฐานที่ว่า “การแบ่งช่วงข้อมูลที่เหมาะสม เพื่อจัดกลุ่มค่าข้อมูลในแอตทริบิวต์จากตัวเลข(Numeric) เป็นแบบนาม(Nominal) มีผลต่อค่าดัชนีที่ใช้ในการวัดผลความสามารถในการจำแนกข้อมูลหรือการพยากรณ์ข้อมูลของโมเดล” และเพื่อนำข้อสรุปจากการทดลองไปสร้างองค์ความรู้ที่เป็นประโยชน์สำหรับผู้เริ่มต้นทำการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล โดยสามารถนำไปใช้ประกอบการวางแผนในขั้นตอนการเตรียมข้อมูลเพื่อให้การวิเคราะห์ข้อมูลมีประสิทธิภาพและมีความถูกต้อง ซึ่งผู้วิจัยได้ทำการศึกษา ทดลอง วิเคราะห์และสรุปผลการทดลองในเบื้องต้น เกี่ยวกับปัจจัยสำคัญที่บ่งบอกถึงการวางแผนที่ดีในการจัดการข้อมูลที่เหมาะสมในขั้นตอนการเตรียมข้อมูล ดังได้แก่ ความถูกต้องแม่นยำ ความน่าเชื่อถือ และความมีประสิทธิภาพของโมเดลในการจำแนกหรือการพยากรณ์ข้อมูล

คำสำคัญ-- การทำเหมืองข้อมูล, Data Mining, การพยากรณ์, การจำแนกข้อมูล

1. บทนำ

ปัจจุบันองค์กรภาครัฐและภาคเอกชนต่างให้ความสำคัญกับการสร้างองค์ความรู้ใหม่ๆ เนื่องจากองค์ความรู้สามารถนำไปใช้ให้เกิดประโยชน์ต่อการดำเนินงานขององค์กร เพื่อการบรรลุวัตถุประสงค์ขององค์กร หรือเพื่อการมีผลกำไรสูงสุด แต่การดำเนินธุรกิจย่อมต้องมีคู่แข่งทางการค้าอย่างหลีกเลี่ยงไม่ได้ จึงทำให้เกิดการคิดค้นหากลยุทธ์ในการดำเนินธุรกิจเพื่อชิงความได้เปรียบต่อคู่แข่งทุกวิถีทาง ไม่ว่าจะเป็นการลงทุนด้านต่างๆ เช่น ด้านทรัพยากรบุคคล โดยการเพิ่มจำนวนพนักงานในฝ่ายงานที่ก่อให้เกิดรายได้ต่อองค์กร ด้านเทคโนโลยีสารสนเทศ โดยทำการพัฒนาระบบงานด้านสารสนเทศรวมถึงการจัดซื้อฮาร์ดแวร์และซอฟแวร์ที่ทันสมัย ด้านการพัฒนาผลิตภัณฑ์และบริการ ด้านโลจิสติกส์ และด้านการบริการหลังการขาย โดยคาดหวังว่าจะเป็นการช่วยให้เกิดการขยายตลาดไปยังกลุ่มลูกค้าใหม่ๆ และเป็นการเพิ่มช่องทางการเข้าถึงสินค้าและบริการของลูกค้าได้อย่างสะดวกรวดเร็ว ลูกค้ามีความประทับใจในผลิตภัณฑ์และบริการ ซึ่งการลงทุนด้านต่างๆ ที่กล่าวมาข้างต้นนั้น ทำให้องค์กรมีค่าใช้จ่ายที่เป็นตัวเงินทั้งสิ้น

ดังนั้น องค์กรธุรกิจต่างๆ จึงให้ความสำคัญกับการลดค่าใช้จ่ายที่ไม่จำเป็น และให้ความสำคัญกับการสร้าง

มูลค่าเพิ่มจากข้อมูลที่มีอยู่ไม่ว่าจะเป็นแหล่งข้อมูลภายในหรือภายนอกองค์กร เช่น เว็บไซต์หรือโซเชียลมีเดียต่างๆ ซึ่งเป็นช่องทางสำคัญที่จะทำให้เกิดองค์ความรู้ใหม่ๆ อย่างไรก็ตาม แหล่งข้อมูลสำคัญที่สุด คือ ฐานข้อมูลขององค์กร

การทำเหมืองข้อมูลหรือ “ดาต้าไมนิ่ง” จึงได้รับความนิยมในกลุ่มนักวิเคราะห์ข้อมูลและถูกนำมาใช้เป็นเครื่องมือในการค้นหาความรู้ใหม่จากฐานข้อมูลขนาดใหญ่ ประโยชน์ที่ได้รับจากการทำเหมืองข้อมูล คือ ความรู้ใหม่ที่ยังไม่เคยถูกค้นพบมาก่อนและสามารถนำความรู้ใหม่นี้ไปใช้ในการวางแผนกลยุทธ์ขององค์กร ทั้งนี้ เพื่อสร้างความได้เปรียบคู่แข่งทางธุรกิจ เพื่อสร้างความพึงพอใจในผลิตภัณฑ์และบริการของลูกค้า เพื่อการมีผลประกอบการที่ดี ตลอดจนเพื่อสร้างความเชื่อมั่นแก่ผู้ถือหุ้น เป็นต้น

สถาบันการเงิน อย่างธนาคารก็มีวัตถุประสงค์ในการประกอบธุรกิจเช่นเดียวกันกับองค์กรธุรกิจอื่นๆ ดังนั้น การมีข้อมูลหรือความรู้ใหม่ๆ จึงเป็นสิ่งจำเป็นต่อการดำเนินธุรกิจการที่จะได้มาซึ่งข้อมูลเหล่านั้น จึงเป็นหน้าที่ความรับผิดชอบสำคัญอย่างหนึ่งของหน่วยงานสนับสนุนข้อมูลทางธุรกิจของธนาคาร

ธนาคารเป็นองค์กรที่มีฐานข้อมูลขนาดใหญ่ และเป็นแหล่งข้อมูลที่เหมาะแก่การทำเหมืองข้อมูล โดยผลลัพธ์ที่ได้จะถูกนำไปใช้เพื่อสนับสนุนงานที่อยู่ในความรับผิดชอบของหน่วยที่เกี่ยวข้อง เช่น การนำไปใช้เพื่อการวางแผนการตลาดและกลยุทธ์ ตลอดจนการนำไปใช้เพื่อการพัฒนาระบบสนับสนุนงานด้านปฏิบัติการต่างๆ ซึ่งช่วยให้กระบวนการดำเนินงานต่างๆ ของธนาคารถูกขับเคลื่อนไปได้อย่างราบรื่น

บทความนี้ผู้วิจัยได้นำตัวอย่างการนำความรู้ด้านเทคโนโลยีสารสนเทศมาช่วยลดขั้นตอนการทำงานของธนาคารเกี่ยวกับงานด้านการขยายวันครบกำหนดอายุสัญญาของบัญชีที่มีวงเงินเบิกเกินบัญชี(โอดี)แบบอัตโนมัติ ภายใต้เงื่อนไขของธนาคาร ซึ่งระบบงานเดิมมีการควบคุมและตรวจสอบโดยเจ้าหน้าที่ธนาคารผู้เกี่ยวข้องอีกชั้นหนึ่ง เนื่องจากไม่มีเครื่องมือมาตรฐานมาช่วยสนับสนุนระบบงานดังกล่าวให้มีความน่าเชื่อถือและเป็นที่ยอมรับอย่างเพียงพอ ดังนั้น ผู้วิจัยจึงมีแนวคิดที่จะนำเทคนิคการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับงานดังกล่าว เพื่อหารูปแบบโมเดลที่เหมาะสม มีความถูกต้องและเชื่อถือได้

ปัจจุบันธนาคารมีจำนวนบัญชีที่มีวงเงินเบิกเกินบัญชีมากกว่า 200,000 บัญชี และในแต่ละเดือนมีบัญชีที่ครบกำหนดอายุสัญญากว่า 1,000 บัญชี การทดลองนี้ผู้วิจัยจึงได้นำข้อมูลตัวอย่างที่ผ่านการคัดกรองและเข้าข่ายการพิจารณาให้ขยายอายุสัญญา ประจำเดือนตุลาคม 2556 จำนวน 584 บัญชี มาทำการวิเคราะห์และวางแผนสำหรับการจัดเตรียมข้อมูลที่เหมาะสม ก่อนที่จะเข้าสู่กระบวนการทำเหมืองข้อมูลเต็มรูปแบบต่อไป

2. งานวิจัยหรือทฤษฎีที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data Mining)

เป็นเทคนิคที่ใช้สำหรับการวิเคราะห์ข้อมูลเพื่อค้นหาความรู้ที่มีประโยชน์จากฐานข้อมูลขนาดใหญ่(Knowledge Discovery in Databases - KDD) โดยอาศัยหลักสถิติ การรู้จำหรือการเรียนรู้ของเครื่องจักร และหลักคณิตศาสตร์ มากระทำกับข้อมูลจำนวนมหาศาลโดยอัตโนมัติ ความรู้ที่ได้จะแสดงออกมาในลักษณะของรูปแบบ(Pattern) กฎ(Rule) และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น

แนวปฏิบัติสำหรับการทำให้เครื่องจักรเรียนรู้[1] มี 2 แนวทาง คือ

1) การวิเคราะห์ข้อมูลโดยการเรียนรู้ข้อมูลในอดีต (Supervised Learning) จะมีชุดคำตอบเตรียมไว้ล่วงหน้าเพื่อใช้เป็นแบบอย่างในการวิเคราะห์ข้อมูล หรือสำหรับสอนคอมพิวเตอร์

2) การวิเคราะห์ข้อมูลโดยไม่มีการเตรียมชุดคำตอบไว้ล่วงหน้า(Unsupervised Learning) เป็นการให้คอมพิวเตอร์ทำการวิเคราะห์หาความเหมือนหรือความคล้ายคลึงกัน หรือความสัมพันธ์กันของข้อมูล

การทำเหมืองข้อมูล มีรูปแบบการวิเคราะห์ข้อมูล 3 วิธี คือ

1) วิธีจำแนกข้อมูล(Classification) หรือการพยากรณ์หรือการทำนายผลข้อมูล(Prediction) วิธีนี้จะทำการจำแนกหรือพยากรณ์ข้อมูลตามโมเดลต้นแบบที่ได้จากการเรียนรู้ข้อมูลในอดีต

2) วิธีหาความสัมพันธ์ของข้อมูล(Association) วิธีนี้ใช้สำหรับหาความสัมพันธ์ของข้อมูลที่เกิดขึ้นซ้ำกันบ่อยๆ โดยไม่มีการเรียนรู้ข้อมูลในอดีต

3) วิธีจัดกลุ่มข้อมูล(Clustering) วิธีนี้จัดกลุ่มข้อมูลที่มีลักษณะคล้ายกันไว้ด้วยกัน โดยไม่มีการเรียนรู้ข้อมูลในอดีต

ในปี 2539 บริษัท DaimlerChrysler บริษัท SPSS และบริษัท NCR [1] ได้ร่วมกันกำหนดมาตรฐานในการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล เรียกว่า “Cross-Industry

Standard Process For Data Mining: CRISP-DM" ประกอบด้วย 6 ขั้นตอน ดังนี้

1) ขั้นตอนการทำความเข้าใจปัญหาทางธุรกิจ และวางแผนแนวทางดำเนินการเบื้องต้น(Business Understanding)

2) ขั้นตอนการเก็บรวบรวมข้อมูล การตรวจสอบข้อมูล และวิเคราะห์ข้อมูลที่จำเป็นต้องใช้ในการวิเคราะห์ข้อมูล(Data Understanding)

3) ขั้นตอนการเตรียมข้อมูล(Data Preparation) คือ การปรับปรุงคุณภาพข้อมูลโดยรวมให้อยู่ในสภาพที่สามารถนำไปใช้ในการทำเหมืองข้อมูลได้ วิธีการที่ใช้สำหรับการเตรียมข้อมูล [2][4] มี 4 วิธี ดังนี้

1) การทำความสะอาดข้อมูล(Data Cleansing) คือ การจัดการข้อมูลในกรณีข้อมูลจริงไม่สมบูรณ์หรือมีค่าที่ขาดหายไป(Missing Value)

2) การผสานข้อมูล(Data Integration) คือ การผสานข้อมูลที่รวบรวมมาจากแหล่งข้อมูลที่ต่างกัน เพื่อลดความซ้ำซ้อนของข้อมูล(Data Redundancies) ซึ่งจะช่วยลดปัญหาความไม่สอดคล้องกันข้องข้อมูล(Data Inconsistencies)

3) การแปลงข้อมูล(Data Transformation) คือ การทำนอร์มอลไลซ์(Normalization) ด้วยการแปลงค่าข้อมูลให้อยู่ในช่วงสั้นๆ เช่น (0,1) หรือ (-1,0)

4) การลดรูปข้อมูล(Data Reduction) คือ การจัดการข้อมูลให้อยู่ในรูปแบบกะทัดรัด ประหยัดพื้นที่ในการเก็บข้อมูล

4) ขั้นตอนการวิเคราะห์และออกแบบ(Modeling) เพื่อให้ได้รูปแบบหรือโมเดล

5) ขั้นตอนการวัดประสิทธิภาพของรูปแบบหรือโมเดล (Evaluation) ว่ามีความถูกต้องและน่าเชื่อถือตามวัตถุประสงค์ที่ตั้งไว้หรือไม่

6) ขั้นตอนการนำผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลไปใช้ให้เกิดประโยชน์ต่อองค์กร(Deployment)

2.2 แบบจำลอง (Model)

แบบจำลอง คือ ตัวแบบที่ช่วยในการนำเสนอข้อมูลต่างๆ เพื่อจะนำข้อมูลเหล่านั้นไปใช้ในการแก้ไขปัญหา แบบจำลองมีหลายประเภท เช่น แบบจำลองเชิงบรรยาย(Description Model) แบบจำลองคงที่และแบบพลวัตร(Static And Dynamic Model) และแบบจำลองทางคณิตศาสตร์(Mathematics Model) [4]

ประโยชน์ของแบบจำลอง คือ ประหยัดเวลาและค่าใช้จ่ายในการวิเคราะห์ข้อมูล มีความน่าเชื่อถือ สามารถทำความเข้าใจและมองเห็นภาพได้ชัดเจนขึ้น รวมถึงสามารถนำไปใช้ทดลองแทนมนุษย์ในสถานการณ์ที่อันตราย เครื่องมือที่ใช้ในการสร้างและพัฒนาแบบจำลองมีหลายรูปแบบ เช่น การเขียนโปรแกรมด้วยภาษาต่างๆ การใช้โปรแกรม Spread sheet การคำนวณด้วยสูตรทางคณิตศาสตร์และสถิติ รวมถึงการใช้โปรแกรมสำเร็จรูปต่างๆ

2.3 ตัวจำแนกข้อมูลแบบ ID3

ID3 เป็นอัลกอริธึมสำหรับการจำแนกข้อมูลในกลุ่มเทคนิคแผนภาพต้นไม้ตัดสินใจ (Decision Tree) เป็นแบบจำลองทางคณิตศาสตร์ เพื่อหาทางเลือกที่ดีที่สุด เหมาะสำหรับแก้ปัญหาที่ง่ายและมีทางเลือกน้อย ไม่ซับซ้อน คุณลักษณะเด่น คือ ช่วยให้เห็นภาพข้อมูลสำหรับตัดสินใจง่ายขึ้น

แผนภาพต้นไม้ตัดสินใจเป็นแบบจำลองที่มีการเรียนรู้แบบมีครู(Supervised Learning)[3][4][5] โดยชุดข้อมูลที่เรียกว่า Training Data จะถูกนำมาสอนให้โปรแกรมคอมพิวเตอร์เกิดการเรียนรู้รูปแบบปัญหาและข้อสรุป แล้วนำมาสร้างเป็นแบบจำลอง จากนั้นนำไปพยากรณ์ข้อมูลในชุดข้อมูลทดสอบ(Testing Data) เพื่อพิสูจน์ความถูกต้องและน่าเชื่อถือของแบบจำลองหรือโมเดลนั้นๆ จนเป็นที่ยอมรับและนำออกใช้งานจริง องค์ประกอบของแผนภาพต้นไม้ตัดสินใจ มี 4 ส่วน ดังนี้

1) Root node คือ node แรกตัวบนสุด

2) Branch คือ กิ่งก้านสาขาของ node มีทั้งกิ่งด้านซ้าย(Left Branch) และกิ่งด้านขวา(Right Branch)

3) Child คือ ลูกๆ ของ Branch

4) Leaf node คือ node ที่อยู่ลำดับสุดท้าย

2.4 ตัวจำแนกข้อมูลแบบ Naïve Bayes Simple

Naïve Bayes Simples เป็นอัลกอริธึมสำหรับการจำแนกข้อมูลอย่างง่ายในกลุ่มเทคนิคเบย์ (Bayes) ทำการวิเคราะห์ข้อมูลโดยการเรียนรู้รูปแบบของปัญหาและข้อสรุป แล้วนำมาสร้างเป็นโมเดลการจำแนกข้อมูลด้วยทฤษฎีความน่าจะเป็นแบบมีเงื่อนไข [3][6] หลักการทฤษฎีความน่าจะเป็นแบบมีเงื่อนไข คือ ถ้าเหตุการณ์ A และ B เป็นเหตุการณ์ใดๆ ที่ไม่เป็นอิสระต่อกัน สามารถหาความน่าจะเป็นของเหตุการณ์ A เมื่อทราบว่าเหตุการณ์ B เกิดขึ้นแล้ว [3] ดังนี้

$$P(A|B) = P(A,B)/P(B)$$

โดย P(A|B) คือ ความน่าจะเป็นแบบมีเงื่อนไขของเหตุการณ์ A และ B

P(A,B) คือ ความน่าจะเป็นที่เหตุการณ์ A และ B จะเกิดขึ้นพร้อมกัน

P(B) คือ ความน่าจะเป็นของเหตุการณ์ B และมีค่ามากกว่า 0

3. การทดลอง

การทดลองนี้ใช้ข้อมูลบัญชีที่มีวงเงินเบิกเกินบัญชีที่ผ่านการคัดกรองและเข้าข่ายการพิจารณาให้ขยายอายุสัญญา ประจำเดือนตุลาคม 2556 จำนวน 584 บัญชี จากเก็บรวบรวมข้อมูลเงื่อนไขสำหรับการจำแนกข้อมูลยังไม่ชัดเจนเท่าที่ควร ผู้วิจัยจึงจำลองรูปแบบการจัดการกับข้อมูลออกเป็น 2 รูปแบบ ก่อนที่จะนำข้อมูลดังกล่าวไปสอนโปรแกรม WEKA ให้เรียนรู้รูปแบบของปัญหา ภายใต้สมมติฐานที่ว่า “การแบ่งช่วงข้อมูลที่เหมาะสมเพื่อจัดกลุ่มค่าข้อมูลในแอตทริบิวต์จากตัวเลข(Numeric) เป็นแบบนาม(Nominal) มีผลต่อค่าดัชนีที่ใช้ในการวัดผลความสามารถในการจำแนกข้อมูลหรือการพยากรณ์ข้อมูลของโมเดล”

ดังนั้น ขอบเขตการทดลองในบทความนี้จึงจำกัดอยู่เฉพาะในส่วนของขั้นตอนการเตรียมข้อมูล(Data Preparation) ของเทคนิคการทำเหมืองข้อมูล เพื่อที่จะนำไปสู่การได้มาซึ่งรูปแบบโมเดลการจำแนกข้อมูลที่เหมาะสม มีความถูกต้องและน่าเชื่อถือที่สุดเท่านั้น

3.1 การเตรียมข้อมูล(Data Preparation)

การทดลองนี้ได้จำลองข้อมูลออกเป็น 2 ชุด เรียกชุดข้อมูลว่า ชุด A และ ชุด B โดยมีความแตกต่างระหว่าง 2 ชุด คือ การแบ่งช่วงข้อมูลเพื่อจัดกลุ่มค่าข้อมูลในแอตทริบิวต์จากตัวเลข(Numeric) เป็นแบบนาม(Nominal) มีความละเอียดแตกต่างกัน เช่น ถ้า x หมายถึง ค่าข้อมูล , x =0 ค่าของข้อมูลชุด A จะถูกแทนด้วย Low แต่ในชุด B จะถูกแทน ด้วย None นั่นหมายความว่าในกรณีที่มีการแบ่งข้อมูลออกเป็นช่วงๆ ค่าของข้อมูลในชุด B จะมีความเป็นเอกลักษณ์ หรือมีความละเอียดมากกว่า ชุด A

ข้อมูลที่ใช้ทดลองประกอบด้วย 7 แอตทริบิวท์ ได้แก่

1) Utilization คือ สัดส่วนจำนวนเงินที่มีการเบิกใช้เกินบัญชีต่อวงเงิน

2) Movement คือ สัดส่วนจำนวนเงินที่มีการเคลื่อนไหวทางบัญชีต่อวงเงิน

3) Loan To Value (LTV) คือ สัดส่วนภาระหนี้สินต่อมูลค่าหลักประกัน

4) Return Cheque Debit (Return_Cheque_Dr) คือ สัดส่วนจำนวนเงินของเช็คคืนด้านสั่งจ่ายต่อวงเงิน

5) Return Cheque Credit (Return_Cheque_Cr) คือ สัดส่วนจำนวนเงินของเช็คคืนด้านฝากเข้าต่อวงเงิน

6) Debt Behavior History (History_24M) คือ ประวัติการจ่ายชำระหนี้ย้อนหลัง 24 เดือน มีค่าเป็น Good และ Bad

7) Auto Review (Auto_Review_Class) คือ ผลการพิจารณาให้ขยายอายุสัญญาหรือไม่ ซึ่งใช้เป็นตัวจัดคลาสในการทดลองครั้งนี้

3.2 การแปลงข้อมูล(Data Transformation)

การทำเหมืองข้อมูลด้วยวิธีการจำแนกข้อมูลหรือการพยากรณ์ข้อมูล ต้องทำการแปลงค่าของข้อมูลในแอตทริบิวท์ที่เก็บข้อมูลเป็นตัวเลข ให้อยู่รูปแบบนาม(Nominal) โดยแบ่งค่าข้อมูลออกเป็นกลุ่ม(Cluster) เพื่อลดการกระจายข้อมูลที่มากเกินไปในการทดลองนี้ ทำการแปลงค่าข้อมูลให้อยู่ในรูปของกลุ่มที่สื่อถึงระดับความเสี่ยง 3 ระดับ ได้แก่ สูง(High) ปานกลาง(Medium) ต่ำ(Low) และไม่มีข้อมูล(None) และเพื่อความเข้าใจง่ายในการอ่านผลการทดลอง ผู้วิจัยจึงได้กำหนดตัวอักษรภาษาอังกฤษที่เป็นชื่อย่อของแต่ละแอตทริบิวท์ ตามด้วย “_” นำหน้าชื่อกลุ่มข้อมูลที่ถูกแปลงค่าเป็นกลุ่มนั้นๆ

3.3 เกณฑ์การแบ่งช่วงข้อมูล เพื่อการจัดกลุ่มข้อมูล

Algorithm: Utilization แบบที่ 1

Case when Utilization <0.30 **Then** ‘U_Low’
when Utilization>=0.30 and Utilization<=0.70 **Then** 'U_Medium’
Else 'U_High'
End Utilization

Algorithm: Utilization แบบที่ 2

Case when Utilization=0 **Then** ‘U_None’
when Utilization>0 and Utilization <0.30 **Then** ‘U_Low’
when Utilization>=0.30 and Utilization<=0.70 **Then** 'U_Medium’
Else 'U_High'
End Utilization

รูปที่ 1. อัลกอริธึมเงื่อนไขสำหรับการจัดกลุ่มข้อมูลของแอตทริบิวท์ Utilization

จากรูปที่ 1. เงื่อนไขการแบ่งช่วงข้อมูล เพื่อจัดกลุ่มข้อมูลของแอตทริบิวท์ Utilization อธิบายได้ดังนี้

แบบที่ 1 ถ้า Utilization มีค่าน้อยกว่า 0.30 ให้จัด

อยู่ในกลุ่ม U_Low หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.30 และน้อยกว่าหรือเท่ากับ 0.70 ให้จัดอยู่ในกลุ่ม U_Medium หรือ ถ้าไม่เข้าเกณฑ์ข้างต้น กล่าวคือ มีค่ามากกว่า 0.70 ให้จัดอยู่ในกลุ่ม U_High

แบบที่ 2 ถ้า Utilization มีค่าเท่ากับ 0 ให้จัดอยู่ในกลุ่ม U_None หรือ ถ้ามีค่าน้อยกว่า 0.30 ให้จัดอยู่ในกลุ่ม U_Low หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.30 และน้อยกว่าหรือเท่ากับ 0.70 ให้จัดอยู่ในกลุ่ม U_Medium หรือ ถ้าไม่เข้าเกณฑ์ข้างต้น กล่าวคือ มีค่ามากกว่า 0.70 ให้จัดอยู่ในกลุ่ม U_High

```
Algorithm: Movement แบบที่ 1
Case when Movement<0.30 Then 'M_High'
     when Movement>=0.30 and Movement<=0.70 Then
        'M_Medium'
    Else 'M_Low'
End Movement
Algorithm: Movement แบบที่ 2
Case when Movement=0 Then 'M_None'
     when Movement>0 and Movement<0.30 Then
        'M_High'
     when Movement>=0.30 and Movement<=0.70 Then
        'M_Medium'
    Else 'M_Low'
End Movement
```

รูปที่ 2. อัลกอริธึมเงื่อนไขสำหรับการจัดกลุ่มข้อมูลของแอตทริบิวท์ Movement

จากรูปที่ 2. เงื่อนไขการแบ่งช่วงข้อมูล เพื่อจัดกลุ่มข้อมูลของแอตทริบิวท์ Movement อธิบายได้ดังนี้

แบบที่ 1 ถ้า Movement มีค่าน้อยกว่า 0.30 ให้จัดอยู่ในกลุ่ม M_High หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.30 และน้อยกว่าหรือเท่ากับ 0.70 ให้จัดอยู่ในกลุ่ม M_Medium หรือถ้าไม่เข้าเกณฑ์ข้างต้น กล่าวคือ ถ้ามีค่ามากกว่า 0.70 ให้จัดอยู่ในกลุ่ม M_Low

แบบที่ 2 ถ้า Movement มีค่าเท่ากับ 0 ให้จัดอยู่ในกลุ่ม M_None หรือ ถ้ามีค่าน้อยกว่า 0.30 ให้จัดอยู่ในกลุ่ม M_High หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.30 และน้อยกว่าหรือเท่ากับ 0.70 ให้จัดอยู่ในกลุ่ม M_Medium หรือ ถ้าไม่เข้าเกณฑ์ข้างต้น กล่าวคือ ถ้ามีค่ามากกว่า 0.70 ให้จัดอยู่ในกลุ่ม M_Low

```
Algorithm: LTV แบบที่ 1
Case when LTV<0.50 Then 'L_Low'
     when LTV>=0.50 and LTV<=0.80 Then 'L_Medium'
    Else 'L_High'
End LTV
Algorithm: LTV แบบที่ 2
Case when LTV=0 Then 'L_None'
     when LTV>0 and LTV<0.50 Then 'L_Low'
     when LTV>=0.50 and LTV<=0.80 Then 'L_Medium'
    Else 'L_High'
End LTV
```

รูปที่ 3. อัลกอริธึมเงื่อนไขสำหรับการจัดกลุ่มข้อมูลของแอตทริบิวท์ LTV

จากรูปที่ 3. เงื่อนไขการแบ่งช่วงข้อมูล เพื่อจัดกลุ่มข้อมูลของแอตทริบิวท์ LTV อธิบายได้ดังนี้

แบบที่ 1 ถ้า LTV มีค่าน้อยกว่า 0.50 ให้จัดอยู่ในกลุ่ม L_Low หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.50 และน้อยกว่าหรือเท่ากับ 0.80 ให้จัดอยู่ในกลุ่ม L_Medium หรือ ถ้ามีค่ามากกว่า 0.80 ให้จัดอยู่ในกลุ่ม L_High

แบบที่ 2 ถ้า LTV ค่าเท่ากับ 0 ให้จัดอยู่ในกลุ่ม L_None หรือ ถ้ามีค่าน้อยกว่า 0.50 ให้จัดอยู่ในกลุ่ม L_Low หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.50 และน้อยกว่าหรือเท่ากับ 0.80 ให้จัดอยู่ในกลุ่ม L_Medium หรือ ถ้าไม่เข้าเกณฑ์ข้างต้น กล่าวคือ ถ้ามีค่ามากกว่า 0.80 ให้จัดอยู่ในกลุ่ม L_High

```
Algorithm: Return_Cheque_Dr แบบที่ 1
Case when Return_Cheque_Dr>0 Then 'Yes'
    Else 'No'
End Return_Cheque_Dr
Algorithm: Return_Cheque_Dr แบบที่ 2
Case when Return_Cheque_Dr=0 Then 'CDr_None'
     when Return_Cheque_Dr>0 and
        Return_Cheque_Dr<0.03 Then 'CDr_Low'
     when Return_Cheque_Dr>=0.03 and
        Return_Cheque_Dr<=0.05 Then 'CDr_ Medium'
    Else 'CDr_High'
End Return_Cheque_Dr
```

รูปที่ 4. อัลกอริธึมเงื่อนไขสำหรับการจัดกลุ่มข้อมูลของแอตทริบิวท์ Return_Cheque_Dr

จากรูปที่ 4. เงื่อนไขการแบ่งช่วงข้อมูลเพื่อจัดกลุ่มข้อมูลของแอตทริบิวท์ Return_Cheque_Dr อธิบายได้ดังนี้

แบบที่ 1 ถ้า Return_Cheque_Dr มีค่ามากกว่า 0 ให้จัดอยู่ในกลุ่ม Yes หรือ ถ้ามีค่าเท่ากับ 0 ให้จัดอยู่ในกลุ่ม No

แบบที่ 2 ถ้า Return_Cheque_Dr มีค่าเท่ากับ 0 ให้

จัดอยู่ในกลุ่ม CDr_None หรือ ถ้ามีค่ามากกว่า 0 และน้อยกว่า 0.03 ให้จัดอยู่ในกลุ่ม CDr_Low หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.03 และน้อยกว่าหรือเท่ากับ 0.05 ให้จัดอยู่ในกลุ่ม CDr_Medium หรือ ถ้าไม่เข้าเกณฑ์ข้างต้น กล่าวคือ ถ้ามีค่ามากกว่า 0.05 ให้จัดอยู่ในกลุ่ม CDr_High

```
Algorithm: Return Cheque Cr แบบที่ 1
Case when Return_Cheque_Cr>0 Then 'Yes'
     Else 'No'
End Return_Cheque_Cr
Algorithm: Return Cheque Cr แบบที่ 2
Case when Return_Cheque_Cr=0 Then 'CCr_None'
     when Return_Cheque_Cr>0 and
          Return_Cheque_Cr<0.03 Then 'CCr_Low'
     when Return_Cheque_Cr>=0.03 and
          Return_Cheque_Cr<=0.05 Then 'CCr_Medium'
     Else 'CCr_High'
End Return_Cheque_Cr
```

รูปที่ 5. อัลกอริธึมเงื่อนไขสำหรับการจัดกลุ่มข้อมูลของแอตทริบิวท์ Return_Cheque_Cr

จากรูปที่ 5. อธิบายเงื่อนไขการแบ่งช่วงข้อมูลเพื่อจัดกลุ่มข้อมูลของแอตทริบิวท์ Return_Cheque_Cr ดังนี้

แบบที่ 1 ถ้า Return_Cheque_Cr มีค่ามากกว่า 0 ให้จัดอยู่ในกลุ่ม Yes หรือ ถ้ามีค่าเท่ากับ 0 ให้จัดอยู่ในกลุ่ม No

แบบที่ 2 ถ้า Return_Cheque_Cr มีค่าเท่ากับ 0 ให้จัดอยู่ในกลุ่ม CCr_None หรือ ถ้ามีค่ามากกว่า 0 และน้อยกว่า 0.03 ให้จัดอยู่ในกลุ่ม CCr_Low หรือ ถ้ามีค่ามากกว่าหรือเท่ากับ 0.03 และน้อยกว่าหรือเท่ากับ 0.05 ให้จัดอยู่ในกลุ่ม CCr_Medium หรือ ถ้าไม่เข้าเกณฑ์ข้างต้น กล่าวคือ มีค่ามากกว่า 0.05 ให้จัดอยู่ในกลุ่ม CCr_High

4. ผลการทดลอง

4.1 ผลการทดลอง แบบอัลกอริธึม Naïve Bayes Simple

ตาราง 1. เปรียบเทียบค่าความน่าจะเป็น(Probability) ของข้อมูลชุด A และชุด B

Class / Probability	Cluster	Set A		Set B	
		Pass	Not Pass	Pass	Not Pass
Probability of Class		**0.920**	**0.080**	**0.920**	**0.080**
Utilization	High	0.401	0.878	0.400	0.860
	Medium	0.216	0.061	0.216	0.060
	Low	0.383	0.061	0.236	0.060
	None	-	-	0.148	0.02
Movement	High	0.619	0.735	0.424	0.400
	Medium	0.192	0.143	0.192	0.140
	Low	0.189	0.122	0.188	0.120
	None	-	-	0.196	0.340
LTV	High	0.213	0.490	0.213	0.490
	Medium	0.488	0.245	0.488	0.245
	Low	0.299	0.265	0.299	0.265
	None	-	-	-	-
Return_Cheque_Dr	Yes/High	0.039	0.292	0.009	0.200
	No/Medium	0.961	0.708	0.013	0.060
	Low	-	-	0.020	0.060
	None	-	-	0.958	0.680
Return_Cheque_Cr	Yes/High	0.144	0.125	0.090	0.060
	No/Medium	0.856	0.875	0.026	0.080
	Low	-	-	0.031	0.020
	None	-	-	0.852	0.840
History_24M	Good	0.996	0.833	0.996	0.833
	Bad	0.004	0.167	0.004	0.167

ตารางที่ 1. แสดงถึงการเปรียบเทียบค่าความน่าจะเป็น (Probability) ระหว่างผลการทดลอง 2 ชุด ทั้งในระดับคลาสและระดับกลุ่มข้อมูล(Cluster) ในแต่ละแอตทริบิวต์ ผลการทดลองพบว่าค่าความน่าจะเป็นในระดับคลาสของข้อมูลชุด A และชุด B มีค่าเท่ากัน คือ 0.920 และ 0.080 ตามลำดับ

เมื่อพิจารณาค่าความน่าจะเป็นในระดับคลัสเตอร์ของแต่ละแอตทริบิวต์พบว่า เมื่อเปลี่ยนแปลงเงื่อนไขการแบ่งช่วงข้อมูลเพื่อจัดกลุ่มหรือคลัสเตอร์ข้อมูล ตามข้อ 3.3 มีผลทำให้ค่าความน่าจะเป็นของข้อมูลชุด A และชุด B มีความแตกต่างค่อนข้างชัดเจน โดยเฉพาะอย่างยิ่งในคลัสเตอร์ Low และคลัสเตอร์ High ของแอตทริบิวท์ Utilization, Movement และ LTV และคลัสเตอร์ Yes และคลัสเตอร์ No ของแอตทริบิวท์ Return_Cheque_Dr และ Return_Cheque_Cr

ค่าความน่าจะเป็นของคลัสเตอร์กลุ่มต่างๆ ของข้อมูลชุด A ถูกกระจายไปยังคลัสเตอร์ต่างๆ ของชุด B คือที่คลัสเตอร์ None

ตาราง 2. เปรียบเทียบความสามารถในการจำแนกข้อมูลของโมเดล

Measure	Set A	Set B
Correctly Classified (No. Instances /(%))	543 (92.98%)	548 (93.84%)
Incorrectly Classified (No. Instances /(%))	41 (7.02%)	36 (6.16%)
Root Mean Squared Error (%)	0.235	0.230

ข้อมูลจากตารางที่ 2. แสดงให้เห็นถึงความสามารถในการจำแนกข้อมูลของโมเดลที่ได้จากการวิคราะห์ข้อมูล พบว่าค่าความถูกต้องในการจำแนกหรือการพยากรณ์(Correctly Classified) ที่ตรงกับค่าจริง ของชุด A และ B จำนวน 543 เรคคอร์ด และ 548 เรคคอร์ด คิดเป็น 92.98% และ 93.84% ตามลำดับ ค่าความผิดพลาดในการจำแนกหรือการพยากรณ์(Incorrectly Classified) ที่ไม่ที่ตรงกับค่าจริง ของชุด A และ B จำนวน 41 เรคคอร์ด และ 36 เรคคอร์ด คิดเป็น 7.02% และ 6.16% ตามลำดับ และค่าดัชนีความคลาดเคลื่อนระหว่างค่าจริงและค่าพยากรณ์ (Root Mean Squared Error) ของชุด A และ B เท่ากับ 0.235 และ 0.230 ตามลำดับ

จากจำนวนข้อมูลทดลองทั้งหมด 584 เรคคอร์ด โปรแกรม WEKA สามารถนำมาทำการจำแนกข้อมูล ของชุด A และ B ได้ทั้งหมด คือเท่ากับ 584 เรคคอร์ด

ตาราง 3. เปรียบเทียบค่าความจริงและค่าที่จะพยากรณ์

Classified as / Fact Value	Set A		Set B	
	Pass	Not Pass	Pass	Not Pass
Pass	531 (a)	7 (b)	534 (a)	4 (b)
Not Pass	34 (c)	12 (d)	32 (c)	14 (b)

ตาราง 4.สูตรคำนวณค่า Correctly และ Incorrectly

Correctly Classified	Incorrectly Classified
(a+d)/(a+b+c+d)	(b+c)/(a+b+c+d)

ตาราง 5. สูตรคำนวณค่า TP, FP และ Precision

True Positive Rate(TP)		False Positive Rate(FP)		Precision	
Pass	Not Pass	Pass	Not Pass	Pass	Not Pass
a/ (a+b)	d/ (c+d)	c/ (c+d)	b/ (a+d)	a/ (a+c)	d/ (b+d)

ตาราง 6. เปรียบเทียบค่าข้อมูลจริงกับผลการพยากรณ์ของโมเดล

	Set A			Set B		
	Pass	Not Pass	Avg.	Pass	Not Pass	Avg.
TP Rate	0.987	0.261	0.930	0.993	0.304	0.938
FP Rate	0.739	0.013	0.682	0.696	0.007	0.641
Pre-cision	0.940	0.632	0.916	0.943	0.778	0.930

จากค่าในตารางที่ 3. สามารถนำไปคำนวณหาค่าต่างๆ ได้โดยใช้สูตรคำนวณดังในตารางที่ 4. และ 5. ในส่วนตารางที่ 6. เป็นการแสดงผลจากการคำนวณโดยโปรแกรม WEKA ซึ่งมาจากสูตรคำนวณในตารางที่ 5. เมื่อพิจารณาผลการทดลองในตารางที่ 6. ในส่วนที่เป็นค่าเฉลี่ย(Average) พบว่า ค่า True Postitive Rate หรือค่าที่ใช้อธิบายผลการทำนายด้วยรูปแบบหรือโมเดลที่ถูกค้นพบ แล้วให้คำตอบเป็นจริงของชุด A และ B เท่ากับ 0.930 และ 0.938 ตามลำดับ ค่า False Positive Rate คือค่าที่ใช้อธิบายผลการทำนายด้วยรูปแบบหรือโมเดลที่ถูกค้นพบ แล้วให้คำตอบเป็นเท็จ กล่าวคือ ข้อมูลที่ไม่ได้อยู่ในคลาสนั้นแต่คำตอบจากโมเดลบอกว่าอยู่ในคลาสนั้นของชุด A และ B เท่ากับ 0.682 และ 0.641 ตามลำดับ และค่า Precision คือค่าวัดความเชื่อมั่นของชุด A และ B เท่ากับ 0.916 และ 0.930 ตามลำดับ

4.2 ผลการทดลอง แบบอัลกอริธึม ID3

```
TrainingSetA
Id3

Utilization = U_High
|  History_24M = Good
|  |  LTV = L_Low
|  |  |  Return_Cheque_Dr = No
|  |  |  |  Movement = M_High
|  |  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  |  Return_Cheque_Cr = No: Pass
|  |  |  |  Movement = M_Low: Pass
|  |  |  |  Movement = M_Medium: Pass
|  |  |  Return_Cheque_Dr = Yes
|  |  |  |  Movement = M_High: Pass
|  |  |  |  Movement = M_Low: Pass
|  |  |  |  Movement = M_Medium: Not pass
|  |  LTV = L_Medium
|  |  |  Return_Cheque_Dr = No
|  |  |  |  Movement = M_High
|  |  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  |  Return_Cheque_Cr = No: Pass
|  |  |  |  Movement = M_Low: Pass
|  |  |  |  Movement = M_Medium: Pass
|  |  |  Return_Cheque_Dr = Yes
|  |  |  |  Movement = M_High: Pass
|  |  |  |  Movement = M_Low
|  |  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  |  Return_Cheque_Cr = No: Pass
|  |  |  |  Movement = M_Medium
|  |  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  |  Return_Cheque_Cr = No: Pass
|  |  LTV = L_High
|  |  |  Movement = M_High
|  |  |  |  Return_Cheque_Dr = No
|  |  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  |  Return_Cheque_Cr = No: Pass
|  |  |  |  Return_Cheque_Dr = Yes: Not pass
|  |  |  Movement = M_Low
|  |  |  |  Return_Cheque_Dr = No: Pass
|  |  |  |  Return_Cheque_Dr = Yes
|  |  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  |  Return_Cheque_Cr = No: Not pass
|  |  |  Movement = M_Medium
|  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  Return_Cheque_Cr = No: Pass
|  History_24M = Bad
|  |  LTV = L_Low: Not pass
|  |  LTV = L_Medium: Not pass
|  |  LTV = L_High: Pass
Utilization = U_Medium
|  Return_Cheque_Dr = No
|  |  LTV = L_Low: Pass
|  |  LTV = L_Medium: Pass
|  |  LTV = L_High
|  |  |  Movement = M_High
|  |  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  |  Return_Cheque_Cr = No: Pass
|  |  |  Movement = M_Low: Pass
|  |  |  Movement = M_Medium: Pass
|  Return_Cheque_Dr = Yes
|  |  LTV = L_Low: Not pass
|  |  LTV = L_Medium: Pass
|  |  LTV = L_High: null
Utilization = U_Low
|  Movement = M_High: Pass
|  Movement = M_Low: Pass
|  Movement = M_Medium
|  |  LTV = L_Low: Pass
|  |  LTV = L_Medium: Pass
|  |  LTV = L_High
|  |  |  Return_Cheque_Cr = Yes: Pass
|  |  |  Return_Cheque_Cr = No: Pass
```

รูปที่ 6. โมเดลต้นไม้ตัดสินใจ ข้อมูลชุด A

```
TrainingSetB
Id3

Utilization = U_High
|  History_24M = Good
|  |  LTV = L_Low
|  |  |  Return_Cheque_Dr = CDr_None
|  |  |  |  Movement = M_High
|  |  |  |  |  Return_Cheque_Cr = CCr_Medium: Pass
|  |  |  |  |  Return_Cheque_Cr = CCr_None: Pass
|  |  |  |  |  Return_Cheque_Cr = CCr_High: Pass
|  |  |  |  |  Return_Cheque_Cr = CCr_Low: null
|  |  |  |  Movement = M_None: Pass
|  |  |  |  Movement = M_Low: Pass
|  |  |  |  Movement = M_Medium: Pass
|  |  |  Return_Cheque_Dr = CDr_Medium: Pass
|  |  |  Return_Cheque_Dr = CDr_High: Not pass
|  |  |  Return_Cheque_Dr = CDr_Low: Pass
|  |  LTV = L_Medium
|  |  |  Return_Cheque_Dr = CDr_None
|  |  |  |  Movement = M_High
|  |  |  |  |  Return_Cheque_Cr = CCr_Medium: null
|  |  |  |  |  Return_Cheque_Cr = CCr_None: Pass
|  |  |  |  |  Return_Cheque_Cr = CCr_High: Pass
|  |  |  |  |  Return_Cheque_Cr = CCr_Low: Pass
|  |  |  |  Movement = M_None: Pass
|  |  |  |  Movement = M_Low: Pass
|  |  |  |  Movement = M_Medium: Pass
|  |  |  Return_Cheque_Dr = CDr_Medium
|  |  |  |  Movement = M_High: Pass
|  |  |  |  Movement = M_None: null
|  |  |  |  Movement = M_Low: Pass
|  |  |  |  Movement = M_Medium: Pass
|  |  |  Return_Cheque_Dr = CDr_High: Not pass
|  |  |  Return_Cheque_Dr = CDr_Low
|  |  |  |  Movement = M_High: Pass
|  |  |  |  Movement = M_None: null
|  |  |  |  Movement = M_Low: Pass
|  |  |  |  Movement = M_Medium: Pass
|  |  LTV = L_High
|  |  |  Movement = M_High
|  |  |  |  Return_Cheque_Dr = CDr_None
|  |  |  |  |  Return_Cheque_Cr = CCr_Medium: null
|  |  |  |  |  Return_Cheque_Cr = CCr_None: Pass
|  |  |  |  |  Return_Cheque_Cr = CCr_High: Not pass
|  |  |  |  |  Return_Cheque_Cr = CCr_Low: Pass
|  |  |  |  Return_Cheque_Dr = CDr_Medium: null
|  |  |  |  Return_Cheque_Dr = CDr_High: Pass
|  |  |  |  Return_Cheque_Dr = CDr_Low: Not pass
|  |  |  Movement = M_None
|  |  |  |  Return_Cheque_Dr = CDr_None: Pass
|  |  |  |  Return_Cheque_Dr = CDr_Medium: null
|  |  |  |  Return_Cheque_Dr = CDr_High: Not pass
|  |  |  |  Return_Cheque_Dr = CDr_Low: null
|  |  |  Movement = M_Low
|  |  |  |  Return_Cheque_Dr = CDr_None: Pass
|  |  |  |  Return_Cheque_Dr = CDr_Medium: null
|  |  |  |  Return_Cheque_Dr = CDr_High
|  |  |  |  |  Return_Cheque_Cr = CCr_Medium: null
|  |  |  |  |  Return_Cheque_Cr = CCr_None: Not pass
|  |  |  |  |  Return_Cheque_Cr = CCr_High: Pass
|  |  |  |  |  Return_Cheque_Cr = CCr_Low: null
|  |  |  |  Return_Cheque_Dr = CDr_Low: null
|  |  |  Movement = M_Medium
|  |  |  |  Return_Cheque_Cr = CCr_Medium: Pass
|  |  |  |  Return_Cheque_Cr = CCr_None: Pass
|  |  |  |  Return_Cheque_Cr = CCr_High: null
|  |  |  |  Return_Cheque_Cr = CCr_Low: Pass
|  History_24M = Bad
|  |  LTV = L_Low: Not pass
|  |  LTV = L_Medium: Not pass
|  |  LTV = L_High: Pass
Utilization = U_Medium
|  Return_Cheque_Dr = CDr_None
|  |  LTV = L_Low: Pass
|  |  LTV = L_Medium: Pass
|  |  LTV = L_High
|  |  |  Movement = M_High
|  |  |  |  Return_Cheque_Cr = CCr_Medium: null
|  |  |  |  Return_Cheque_Cr = CCr_None: Pass
|  |  |  |  Return_Cheque_Cr = CCr_High: Pass
|  |  |  |  Return_Cheque_Cr = CCr_Low: Pass
|  |  |  Movement = M_None: Pass
|  |  |  Movement = M_Low: Pass
|  |  |  Movement = M_Medium: Pass
|  Return_Cheque_Dr = CDr_Medium: null
|  Return_Cheque_Dr = CDr_High
|  |  LTV = L_Low: Not pass
|  |  LTV = L_Medium: Pass
|  |  LTV = L_High: null
|  Return_Cheque_Dr = CDr_Low: null
Utilization = U_Low
|  Movement = M_High: Pass
|  Movement = M_None: Pass
|  Movement = M_Low: Pass
|  Movement = M_Medium
|  |  LTV = L_Low: Pass
|  |  LTV = L_Medium: Pass
|  |  LTV = L_High
|  |  |  Return_Cheque_Cr = CCr_Medium: Pass
|  |  |  Return_Cheque_Cr = CCr_None: Pass
|  |  |  Return_Cheque_Cr = CCr_High: null
|  |  |  Return_Cheque_Cr = CCr_Low: null
Utilization = U_None: Pass
```

รูปที่ 7. โมเดลต้นไม้ตัดสินใจ ข้อมูลชุด B

รูปที่ 6. แสดงถึงผลการจำแนกข้อมูลในรูปแบบต้นไม้ตัดสินใจ โดยพบว่าโมเดลต้นไม้ตัดสินใจของข้อมูลชุด A ส่วนใหญ่สามารถจำแนกข้อมูลได้ว่าข้อมูลแบบใดบ้างจะถูกจัดอยู่ในคลาส Pass หรือ Not Pass และมีเพียงหนึ่งรูปแบบเท่านั้นที่ให้คำตอบเป็น Null คือไม่สามารถจำแนกได้ ผู้วิจัยจึงได้ทำการตรวจสอบจากข้อมูลต้นแบบสำหรับสอนโปรแกรมพบว่ารูปแบบเงื่อนไขเพื่อการตัดสินใจตามที่โมเดลสร้างขึ้นนั้น มีคำตอบทั้ง 2 แบบ คือ Pass และ Not Pass ทำให้โมเดลไม่สามารถจำแนกข้อมูลได้ จึงแสดงผลลัพธ์เป็น Null

รูปที่ 7. แสดงผลการทดลองของข้อมูลชุด B ซึ่งชุด B เป็นชุดที่มีการแบ่งกลุ่มหรือคลัสเตอร์ข้อมูลในแต่ละแอตทริบิวท์ละเอียดมากกว่าชุด A พบว่า โมเดลต้นไม้ตัดสินใจของข้อมูลชุด B สามารถจำแนกข้อมูลได้ว่าข้อมูลแบบใดบ้างจะถูกจัดอยู่ในคลาส Pass หรือ Not Pass เช่นกัน แต่ไม่ดีเท่าที่ควรและเป็นที่น่าสังเกตว่าโมเดลของชุด B แสดงคำตอบเป็น Null ค่อนข้างมาก

ผู้วิจัยจึงได้ทำการตรวจสอบจากข้อมูลจริงพบว่า มีสาเหตุมาจากรูปแบบข้อมูลสำหรับตัดสินใจตามที่โมเดลสร้างขึ้นนั้น ไม่ได้มีอยู่จริงในข้อมูลชุดทดลอง ดังนั้น จึงไม่มีคำตอบให้กับรูปแบบของปัญหาดังกล่าว ซึ่งผู้วิจัยตั้งข้อสันนิษฐานว่าอาจมีความเป็นไปได้ที่การจัดกลุ่มข้อมูลในแอตทริบิวท์ใดๆ ที่ละเอียดมากเกินไป อาจไม่สัมพันธ์ตามหลักความเป็นจริงของเงื่อนไขการพิจารณาขยายอายุสัญญาของระบบงานเดิม

จากเหตุการณ์นี้ ผู้วิจัยให้ความสนใจกับแอตทริบิวท์เกี่ยวเช็คคืนเป็นพิเศษ คือ Return_Cheque_Dr และ Return_Cheque_Cr เนื่องจากในขั้นตอนการเตรียมข้อมูล มีความแตกต่างในเรื่องการแปลงค่าข้อมูล คือ ชุด A มีค่าข้อมูลเป็น Yes และ No ขณะที่ค่าข้อมูลชุด B มีค่าเป็น Low, Medium และ High

ตาราง 7. เปรียบเทียบความสามารถในการจำแนกข้อมูลของโมเดล

Measure	Set A	Set B
Correctly Classified (No. Instances /(%))	540 (92.47%)	537 (91.95%)
Incorrectly Classified (No. Instances /(%))	44 (7.53%)	40 (6.85%)
Root Mean Squared Error (%)	0.256	0.246

จากตารางที่ 7. แสดงให้เห็นถึงความสามารถในการจำแนกข้อมูลของโมเดลของอัลกอริธึม ID3 พบว่าค่าความถูกต้องในการจำแนกหรือการพยากรณ์(Correctly Classified) ที่ตรงกับค่าจริง ของชุด A และ B จำนวน 540 เรคคอร์ด และ 537 เรคคอร์ด หรือคิดเป็น 92.47% และ 91.95% ตามลำดับ ซึ่งเป็นที่น่าสังเกตว่า ขณะที่มีการจัดกลุ่มมูลที่ละเอียดมากขึ้นแต่ทำไมความถูกต้องจึงลดต่ำลง ค่าความผิดพลาดในการจำแนกหรือการพยากรณ์(Incorrectly Classified) ที่ไม่ที่ตรงกับค่าจริง ของชุด A และ B จำนวน 44 เรคคอร์ด และ 40 เรคคอร์ด หรือคิดเป็น 7.53% และ 6.85% ตามลำดับ และค่าดัชนีความคลาดเคลื่อนระหว่างค่าจริงและค่าพยากรณ์(Root Mean Squared Error) ของชุด A และ B เท่ากับ 0.256 และ 0.246 ตามลำดับ

จากจำนวนข้อมูลทั้งหมด 584 เรคคอร์ด โปรแกรม WEKAสามารถทำการจำแนกข้อมูล ของชุด A เท่ากับ 584 เรคคอร์ด และชุด B เท่ากับ 577 เรคคอร์ด ซึ่งน้อยกว่าจำนวนข้อมูลทั้งหมด

ตาราง 8. เปรียบเทียบค่าความจริงและค่าที่จำแนกหรือพยากรณ์

Classified as / Fact Value	Set A		Set B	
	Pass	Not Pass	Pass	Not Pass
Pass	531 (a)	7 (b)	526 (a)	7 (b)
Not Pass	37 (c)	9 (d)	33 (c)	11 (b)

ตาราง 9 เปรียบเทียบค่าข้อมูลจริงกับผลการพยากรณ์ของโมเดล

	Set A			Set B		
	Pass	Not Pass	Avg.	Pass	Not Pass	Avg.
TP Rate	0.987	0.196	0.925	0.987	0.250	0.931
FP Rate	0.804	0.013	0.742	0.750	0.013	0.694
Precision	0.935	0.563	0.906	0.941	0.611	0.916

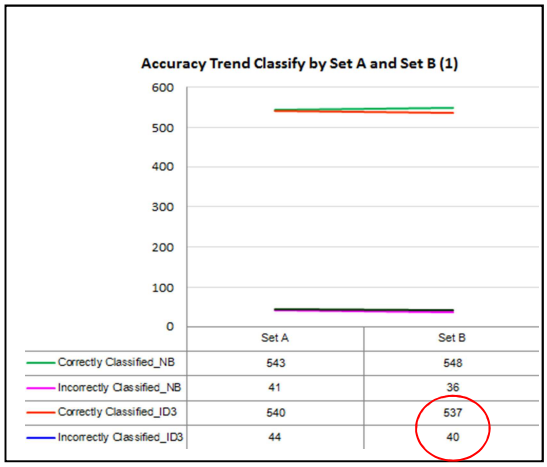
จากค่าในตารางที่ 8. สามารถนำไปคำนวณหาค่าต่างๆได้โดยใช้สูตรคำนวณดังในตารางที่ 4. และ 5. ในส่วนตารางที่ 9. เป็นการแสดงผลจากการคำนวณโดยโปรแกรม WEKA ซึ่งมาจากสูตรคำนวณในตารางที่ 5. เมื่อพิจารณาผลการทดลองในตารางที่ 9. ในส่วนที่เป็นค่าเฉลี่ย(Average) พบว่า ค่า True Postitive Rate หรือค่าที่ใช้อธิบายผลการทำนายด้วยรูปแบบหรือโมเดลที่ถูกค้นพบ แล้วให้คำตอบเป็นจริงของชุด A และ B เท่ากับ 0.925 และ 0.931 ตามลำดับ ค่า False Positive Rate คือค่าที่ใช้อธิบายผลการทำนายด้วยรูปแบบหรือโมเดลที่ถูกค้นพบ แล้วให้

คำตอบเป็นเท็จ กล่าวคือ ข้อมูลที่ไม่ได้อยู่ในคลาสนั้นแต่คำตอบจากโมเดลบอกว่าอยู่ในคลาสนั้นของชุด A และ B เท่ากับ 0.742 และ 0.694 ตามลำดับ และค่า Precision คือค่าวัดความเชื่อมั่นของชุด A และ B เท่ากับ 0.906 และ 0.916 ตามลำดับ

5. สรุป

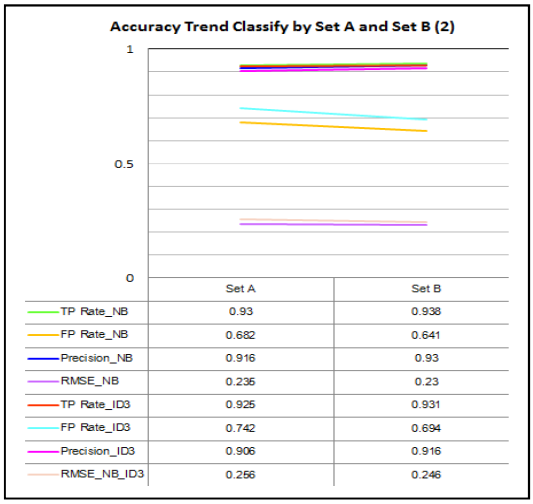
จากผลการทดลอง เปรียบเทียบการแบ่งกลุ่มหรือการคลัสเตอร์ค่าของข้อมูลในแอตทริบิวท์หนึ่งๆ ด้วยการแบ่งกลุ่มที่แตกต่างกันในชุดข้อมูลทดลอง 2 ชุด คือ ชุด A และ ชุด B ด้วยวิธีการกำหนดเงื่อนไขการแบ่งช่วงในลักษณะที่แตกต่างกัน เพื่อพิสูจน์ว่ารูปแบบการจัดการข้อมูลในลักษณะดังกล่าวจะมีผลต่อความแม่นยำในการจำแนกข้อมูลหรือไม่ โดยทำการทดลองและเปรียบเทียบผลการทดลองด้วย 2 อัลกอริธึม ได้แก่ Naïve Bayes Simple และ ID3 ผลปรากฏว่า

ในการกรณีที่มีการแบ่งกลุ่มข้อมูลหรือคลัสเตอร์ข้อมูลออกเป็นหลายกลุ่มหรือหลายคลัสเตอร์ ในที่นี้หมายถึงชุดทดลอง “ชุด B” จะช่วยเพิ่มความสามารถในการจำแนกข้อมูลของอัลกอริธึม Naïve Bayes Simple แต่รูปแบบการแบ่งกลุ่มหรือคลัสเตอร์ในลักษณะดังกล่าว ไม่เหมาะสำหรับนำไปใช้กับการหารูปแบบเพื่อสร้างโมเดลด้วยอัลกอริธึม ID3 เนื่องจาก ID3 มีข้อจำกัดด้านความสามารถในการจำแนกข้อมูลที่มีความซับซ้อน ทำให้ไม่สามารถจำแนกข้อมูลได้ดีเท่าที่ควร



รูปที่ 8. เปรียบเทียบแนวโน้มค่าดัชนีวัดความแม่นยำในการจำแนกระหว่างโมเดลที่ได้จากอัลกอริธึม Naïve Bayes Simple และ ID3 ของชุดข้อมูล A และ B

จากกราฟในรูปที่ 8. พบว่าอัลกอริธึม Naïve Bayes Simple สามารถจำแนกข้อมูลทั้ง ชุด A และ ชุด B ได้ทั้งหมด คือ 584 เรคคอร์ด แม้ว่าจำนวนการจำแนกจะแตกต่างกันก็ตาม แต่สำหรับอัลกอริธึม ID3 สามารถจำแนกข้อมูลได้ทั้งหมดเฉพาะข้อมูลชุด A ในขณะที่ ชุด B จำแนกได้เพียง คือ 577 เรคคอร์ดเท่านั้น



รูปที่ 9. เปรียบเทียบแนวโน้มค่าดัชนีวัดความสามารถในการจำแนกคลาสระหว่างโมเดลที่ได้จากอัลกอริธึม Naïve Bayes Simple และ ID3 ของชุดข้อมูล A และ B

จากกราฟในรูปที่ 9. แสดงให้เห็นว่าการแบ่งกลุ่มข้อมูลหรือคลัสเตอร์ข้อมูลออกเป็นหลายกลุ่มหรือหลายคลัสเตอร์ จะทำให้ค่าความสามารถในการจำแนกคลาสเป็นไปในทางที่ดีขึ้น ทั้งในส่วนของอัลกอริธึม Naïve Bayes Simple และ ID3 ดังจะสังเกตได้จากค่าความเชื่อมั่น(Precision) ค่าการทำนายถูกต้องตรงกับข้อมูลจริง(TP) ค่าการทำนายผิดพลาดจากข้อมูลจริง(FP) ค่าดัชนีความคลาดเคลื่อนระหว่างค่าจริงกับค่าที่จะจำแนกหรือพยากรณ์(RMSE) แต่อย่างไรก็ตาม การคำนวณค่าดังกล่าวอยู่บนพื้นฐานการคำนวณเฉพาะข้อมูลในส่วนที่นำมาจำแนกได้ในแต่ละอัลกอริธึมเท่านั้น

ในบทความนี้พิสูจน์ให้เห็นว่าการวางแผนในการจัดเตรียมข้อมูลที่เหมาะสมมีความสำคัญต่อความสามารถในการจำแนกข้อมูลหรือการพยากรณ์ข้อมูล และต้องมีจัดเตรียมข้อมูลในรูปแบบที่เหมาะสมกับการที่จะนำไปใช้งานกับอัลกอริธึมแต่ละประเภท และสอดคล้องกับเงื่อนไขการใช้งานตามหลักความเป็นจริงของระบบงานนั้นๆ

ดังจะเห็นได้จากความสามารถในการจำแนกข้อมูลของอัลกอริธึม ID3 ในส่วนของผลการทดลองข้อมูลชุด B ที่จำแนกข้อมูลได้เพียง 577 เรคคอร์ดเท่านั้น ผู้วิจัยได้ตั้งข้อสังเกตว่าข้อผิดพลาดดังกล่าวอาจมีสาเหตุมาจากความแตกต่างของเงื่อนไขการแปลงค่าข้อมูลของแอตทริบิวท์เกี่ยวกับเช็คคืน คือ

Return_Cheque_Dr และReturn_Cheque_Cr ผู้วิจัยจึงได้ทำการสอบถามเงื่อนไขที่ใช้พิจารณาจากเจ้าหน้าธนาคารผู้เกี่ยวข้อง และได้รับคำตอบว่าเกณฑ์ที่ใช้พิจารณาที่ถูกต้องคือตรวจสอบเฉพาะว่ามีเช็คคืนหรือไม่ ฉะนั้น ในขั้นตอนการเตรียมข้อมูลต้องแปลงค่าเพื่อจัดกลุ่มข้อมูลเป็น Yes และ No เท่านั้น

ข้อเสนอแนะเพิ่มเติม การแบ่งกลุ่มข้อมูลหรือคลัสเตอร์ข้อมูลออกเป็นหลายกลุ่มหรือหลายคลัสเตอร์ แม้ว่าจะไม่เหมาะกับการนำไปใช้สร้างโมเดลด้วยอัลกอริธึม ID3 แต่ก็สามารถนำอัลกอริธึม ID3 มาช่วยในการตรวจสอบเงื่อนไขหรือหาจุดบกพร่องของชุดข้อมูลทดลองในส่วนที่ยังไม่มีคำตอบสำหรับสอนโปรแกรมคอมพิวเตอร์เพื่อให้เรียนรู้ได้ดังที่แสดงในผลการทดลองในส่วนคลัสเตอร์ Null ของแผนภาพต้นไม้ตัดสินใจ

เอกสารอ้างอิง

[1] เอกสิทธิ์ พัชรวงศ์ศักดา. คู่มือการใช้งาน WEKA Explorer เบื้องต้น. พิมพ์ครั้งที่ 1. กรุงเทพฯ : สำนักพิมพ์ เอเชีย ดิจิตอล การพิมพ์, 2556.

[2] ณวีวรรณ เพ็ชรศิริ. ระบบความชาญฉลาดทางธุรกิจ เพื่อสนับสนุนการตัดสินใจ. พิมพ์ครั้งที่ 2. กรุงเทพฯ : สำนักพิมพ์มหาวิทยาลัยธุรกิจบัณฑิตย์, 2556.

[3] สิทธิโชค มุกดาสกุลภิบาล. การวัดประสิทธิภาพของขั้นตอนวิธีตัวจำแนก C4.5, ADTree และ Naïve Bayes ในการจำแนกข้อมูลการซุกซ่อนสิ่งเสพติดสำหรับไปรษณีย์ระหว่างประเทศ. กรุงเทพฯ :บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์, 2551.

[4] Jiawei Han, Jian Pei and Micheline Kamber. Data Mining: Concepts and Techniques: Concepts and Techniques (3rd Edition), 2011.

[5] N. Vandana Sawant, Ketan Shah and Vinayak Ashok Bharadi, “Survey on Data Mining Classification Techniques,” *Proceedings of the International Conference & Workshop on Emerging Trends in Technology (ICWET’ 11)*, p 1380, ACM New York, 2011

[6] John Galloway and Simeon J. Simoff, “Network Data Mining: Methods and Techniques for Discovering Deep Linkage Between Attributes,” *Proceedings of the 3rd Asia-Pacific Conference on Conceptual Modelling (APCCM’06)*, Vol. 53, pp 21-32, Australian Computer Society, Inc, 2006.