การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลสำหรับพยากรณ์บัญชีเงินกู้เบิกเกินบัญชีและทบทวนวงเงิน

ปิยะมาศ กรัณย์ภควุฒิ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศสำหรับวิสาหกิจสมัยใหม่
คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2557

DPU

Applying Data Mining Technique for Bank Overdraft & Credit Limit Review Prediction

Piyamart Karanphakawuth

DPU

A Thematic Paper Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science

Department of Information Technology for Modern Enterprise

Faculty of Information Technology, Dhurakij Pundit University

2014

หัวข้อสารนิพนธ์ การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล
สำหรับพยากรณ์บัญชีเงินกู้เบิกเกินบัญชีและทบทวนวงเงิน
ชื่อผู้เขียน ปิยะมาศ กรัณย์ภควุฒิ
อาจารย์ที่ปรึกษา อาจารย์ ดร. นันทิกา ปริญญาพล
สาขาวิชา เทคโนโลยีสารสนเทศสำหรับวิสาหกิจสมัยใหม่
ปีการศึกษา 2556

บทคัดย่อ

งานวิจัยนี้ได้พัฒนาโมเดลของรูปแบบการตัดสินใจของโปรแกรม WEKA เวอร์ชั่น 3.6.2 โดยเรียนรู้รูปแบบการตัดสินใจตามเงื่อนไขของหลักการทบทวนวงเงิน หรือที่เรียกว่างาน Auto review ด้วยวิธีการจำแนกข้อมูล และวิเคราะห์เปรียบเทียบผลการประมวลผลข้อมูลระหว่างเทคนิคที่ใช้ในการจำแนกข้อมูล 2 เทคนิค คือ J48 และ ID3 เพื่อคัดเลือกรูปแบบโมเดลการตัดสินใจที่เหมาะสมที่สุด ผลการวิจัยพบว่า รูปแบบโมเดลการตัดสินใจที่ได้จากเทคนิค J48 มีข้อดีมากกว่าเทคนิค ID3 เพราะมีลำดับชั้นการตัดสินใจที่ไม่เยอะจนเกินไป และมีค่า Correctly Classified, Incorrectly Classified, True Positive, False Positive และค่า Precision ที่แสดงถึงความมีประสิทธิภาพในการจำแนกข้อมูลของเทคนิค J48 ซึ่งจากการทดลองกับข้อมูลสำหรับเรียนรู้ จำนวน 12,893 เร็คคอร์ด และข้อมูลชุดทดสอบ จำนวน 872 เร็คคอร์ด ได้ผลลัพธ์เป็นไปในแนวทางเดียวกัน

ผู้วิจัยได้พัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ที่มีลักษณะการทำงานแบบ Web Application โดยใช้ภาษา PHP และกำหนดให้เรียกใช้รูปแบบโมเดลการตัดสินใจโดยตรงจากโปรแกรม WEKA จากการประเมินประสิทธิภาพเชิงคุณภาพของโปรแกรมโอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน จากผู้เชี่ยวชาญด้านสินเชื่อ จำนวน 33 คน ได้ผลคะแนนเฉลี่ยในภาพรวมอยู่ในระดับปานกลาง หรือได้คะแนนเฉลี่ย 3.35 จากคะแนนเต็ม 5

Thematic Paper Title	Applying Data Mining Technique for Bank Overdraft & Credit Limit Review Prediction
Author	Piyamart Karanphakawuth
Thematic Paper Advisor	Dr. Nantika Prinyapol
Department	Information Technology for Modern Enterprise
Academic Year	2013

ABSTRACT

This research develop a model of decision rules of WEKA version 3.6.2 by learning decision tree pattern based on the credit limit review or called Auto review. The study applies J48 and ID3 for selecting a suitable decision model. The results found that the pattern of the decision model from J48 technique is better than ID3 technique because ID3 has too many level of decision hierarchy. The accuracy rates of J48 technique is better than ID3 Technique such as correctly classified rate, incorrectly classified rate, true positive rate or recall rate, false positive rate and precision rate.

We develop the web application for predicting a chance of Bank Overdraft Account that can be possible to review credit limit to extend the maturity date by using PHP and import decision tree models direct from WEKA application. There are 33 experts in credit analysis evaluated the effectiveness and quality of application, the overall result are in middle level and average is 3.35 out of 5.

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์ของ อาจารย์ ดร.นันทิกา ปริญญาพล ที่สละเวลาอันมีค่ามาให้คำปรึกษา แนะแนวทางการดำเนินการตามระเบียบวิธีวิจัย อีกทั้งยังช่วยตรวจทานความถูกต้องของเนื้อหาสาระก่อนการจัดพิมพ์รูปเล่ม ผู้วิจัยรู้สึกทราบซึ้งใจเป็นอย่างยิ่งและขอขอบพระคุณอาจารย์ที่ดูแลเอาใจใส่เป็นอย่างดี

ขอขอบพระคุณคณาจารย์ผู้สอนทุกท่านที่ได้สั่งสอนอบรมให้มีความรู้ทางวิชาการ

ขอขอบพระคุณ คุณยุกต์ ศุทธรัตน์ คุณอุทัย ลลิตธรรมศิริ คุณนภดล เอกเอื้อปฏิภาณ และเพื่อนพนักงานสายงานบริหารความเสี่ยงสินเชื่อธุรกิจ SME ธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) ที่อนุเคราะห์ข้อมูลเพื่อดำเนินการศึกษาวิจัย พร้อมทั้งให้ความรู้และคำแนะนำต่างๆ เกี่ยวกับแนวคิดและเงื่อนไขที่ใช้ประกอบการพิจารณาของการทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชี และให้ความร่วมมือในการตอบแบบประเมินประสิทธิภาพโปรแกรมพยากรณ์เป็นอย่างดี

ขอขอบพระคุณ หสม. ดาต้า คิวบ์ โดย ดร.เอกสิทธิ์ พัชรวงศ์ศักดา ที่อนุญาตให้ใช้เทมเพลทเพื่อการแสดงผลของโปรแกรมพยากรณ์

คุณงามความดีอันเกิดจากงานสารนิพนธ์นี้ ผู้วิจัยขอมอบให้แด่ บิดา มารดา ญาติ พี่น้อง และผู้มีอุปการะคุณทุกท่าน

ปิยะมาศ กรัณย์ภควุฒิ

สารบัญ

หน้า

บทคัดย่อภาษาไทย…… ฆ
บทคัดย่อภาษาอังกฤษ…… ง
กิตติกรรมประกาศ…… จ
สารบัญตาราง…… ซ
สารบัญภาพ…… ฌ
บทที่
1. บทนำ…… 1
1.1 ความเป็นมาและความสำคัญของปัญหา…… 1
1.2 คำถามงานวิจัย…… 3
1.3 วัตถุประสงค์ของการวิจัย…… 4
1.4 ขอบเขตของการวิจัย…… 4
1.5 ประโยชน์ที่คาดว่าจะได้รับ…… 5
1.6 นิยามศัพท์เฉพาะ…… 5
2. แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง…… 7
2.1 ประวัติธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) …… 7
2.2 ความหมายสินเชื่อในทางเศรษฐศาสตร์…… 8
2.3 ความสัมพันธ์ระหว่างสินเชื่อกับระบบเศรษฐกิจ…… 8
2.4 ความรู้ทั่วไปด้านสินเชื่อ…… 9
2.5 แนวทางการแสวงหาความรู้…… 11
2.6 การทำเหมืองข้อมูล…… 12
2.7 ตัวจำแนกข้อมูลแบบ C4.5 หรือ J48…… 14
2.8 งานวิจัยที่เกี่ยวข้อง…… 15
3. ระเบียบวิธีวิจัย…… 19
3.1 ขั้นตอนการดำเนินงาน…… 19
3.2 อุปกรณ์และเครื่องมือที่ใช้ในการวิจัย…… 20
3.3 ระยะเวลาในการดำเนินการวิจัย…… 20
3.4 การดำเนินการวิจัย…… 22

สารบัญ (ต่อ)

บทที่	หน้า
4. ผลการดำเนินการวิจัย	36
4.1 การวิเคราะห์และสรุปผลการดำเนินการวิจัย	36
4.2 การประยุกต์ใช้โมเดลการตัดสินใจ	39
4.3 ผลการเปรียบเทียบผลลัพธ์จากโปรแกรมพยากรณ์ กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง	41
5. สรุปผลการวิจัยและข้อเสนอแนะ	44
5.1 สรุปผลการวิจัย	44
5.2 ข้อเสนอแนะ	48
บรรณานุกรม	50
ภาคผนวก	54
ประวัติผู้เขียน	83

สารบัญตาราง

ตารางที่ **หน้า**

3.1 ระยะเวลาในการดำเนินการวิจัย..... 20
3.2 การแปลงข้อมูลของแต่ละแอตทริบิวท์ให้เป็น Nominal..... 24
4.1 เปรียบเทียบความสามารถในการจำแนกข้อมูล หรือการพยากรณ์ของโมเดล..... 36
4.2 สูตรคำนวณค่า TP, FP และ Precision..... 37
4.3 การเปรียบเทียบจำนวนข้อมูลจำแนกตามคลาสจริง
กับผลการพยากรณ์ของโมเดล..... 37
4.4 เปรียบเทียบค่าดัชนีชี้วัดประสิทธิภาพในการจำแนกข้อมูล
หรือการพยากรณ์ของโมเดล..... 38
4.5 เปรียบเทียบผลลัพธ์จากโปรแกรมพยากรณ์กับผลการพิจารณา
ทบทวนวงเงินจากข้อมูลจริง..... 42

DPU

สารบัญภาพ

ภาพที่ **หน้า**

1.1 ปัญหาระบบงานทบทวนวงเงินแบบเดิม…..……..……..……..... 3
1.2 ความต้องการแก้ไขปัญหาระบบงานทบทวนวงเงิน ……..……..... 3
3.1 กระบวนการมาตรฐานในการวิเคราะห์ข้อมูล ตามมาตรฐาน CRISP-DM …….. 22
3.2 การประมวลผลโดยโปรแกรม Ms. Access.…….…… 27
3.3 การบันทึกไฟล์ Excel เป็นไฟล์ .csv…….……..……..…….. 28
3.4 โปรแกรม WEKA เวอร์ชั่น 3.6.2…….……..…… 28
3.5 ตัวอย่างการนำเข้าข้อมูล .csv…….……..…… 29
3.6 ตัวอย่างการลบแอตทริบิวท์ที่ไม่เกี่ยวข้อง…..…….. 29
3.7 ตัวอย่างการบันทึกไฟล์เป็น .arff.……..……... 30
3.8 การเลือกวิธีการจำแนกข้อมูลและวิธีการใช้ข้อมูล…… 31
3.9 รูปแบบโมเดลการตัดสินใจจากการจำแนกข้อมูลโดยใช้เทคนิค J48…… 32
3.10 รูปแบบโมเดลการตัดสินใจจากการจำแนกข้อมูลโดยใช้เทคนิค ID3…… 33
3.11 ผลการประเมินประสิทธิภาพโมเดลการตัดสินใจของเทคนิค J48…….. 34
3.12 ผลการประเมินประสิทธิภาพโมเดลการตัดสินใจของเทคนิค ID3…… 35
4.1 หน้าจอแสดงผลที่ใช้สำหรับพยากรณ์……..…… 40
4.2 หน้าจอแสดงผล Decision Tree สำหรับผู้จัดทำข้อมูล……..…… 40
4.3 หน้าจอแสดงผล Decision Tree สำหรับผู้ใช้ทั่วไป……..…. 41
5.1 สัดส่วนร้อยละของจำนวนข้อมูลที่ใช้เป็นชุดเรียนรู้และชุดทดสอบ..……....…. 45
5.2 สัดส่วนร้อยละของความถูกต้องของผลการพยากรณ์ (ชุดทดสอบ)..……... 47

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันองค์กรภาครัฐและภาคเอกชนต่างให้ความสำคัญกับการสร้างองค์ความรู้ใหม่ๆ ให้เกิดขึ้นกับองค์กร เนื่องจากสามารถนำไปใช้ให้เกิดประโยชน์ต่อการดำเนินงานขององค์กร เพื่อการบรรลุเป้าหมายตามที่องค์กรกำหนดไว้ โดยเฉพาะอย่างยิ่งด้านการสร้างผลกำไรสูงสุด อย่างไรก็ตามการดำเนินธุรกิจย่อมต้องมีคู่แข่งทางการค้าอย่างหลีกเลี่ยงไม่ได้ จึงทำให้องค์กรต่างคิดค้นหากลยุทธ์เพื่อชิงความได้เปรียบต่อคู่แข่งทุกวิถีทาง ไม่ว่าจะเป็นการเพิ่มการลงทุนในด้านต่างๆ เช่น ด้านทรัพยากรบุคคล โดยการเพิ่มจำนวนพนักงานในฝ่ายงานที่ก่อให้เกิดรายได้ต่อองค์กร ด้านเทคโนโลยีสารสนเทศ โดยทำการพัฒนาระบบงานด้านสารสนเทศรวมถึงการจัดซื้อฮาร์ดแวร์และซอฟต์แวร์ที่ทันสมัย ด้านการพัฒนาผลิตภัณฑ์และบริการ ด้านโลจิสติกส์ และด้านการบริการหลังการขาย เป็นต้น โดยมุ่งหมายว่าจะเป็นการช่วยให้เกิดการขยายตลาดไปยังกลุ่มลูกค้าใหม่ๆ หรือเป็นการเพิ่มช่องทางในการเข้าถึงสินค้าและบริการให้แก่ลูกค้า และเพื่อต้องการให้ลูกค้าได้รับความสะดวกและมีความพึงพอใจต่อผลิตภัณฑ์และบริการยิ่งขึ้น ซึ่งการลงทุนดังกล่าวทำให้องค์กรมีค่าใช้จ่ายที่เป็นตัวเงินทั้งสิ้น อาจส่งผลให้ผลประกอบการขององค์กรมีผลกำไรน้อยลง หรือในกรณีร้ายแรงที่สุดอาจทำให้เกิดผลประกอบการขาดทุนได้

ด้วยเหตุนี้ องค์กรต่างๆ จึงเริ่มให้ความสำคัญกับการลดค่าใช้จ่ายที่ไม่จำเป็น และพยายามสร้างมูลค่าเพิ่มจากสิ่งที่มีอยู่ในองค์กรและสามารถจัดหาได้ เช่น การสร้างมูลค่าเพิ่มจากข้อมูลที่องค์กรมีอยู่ แหล่งข้อมูลสำคัญ คือ ข้อมูลจากแหล่งข้อมูลภายในที่ถูกเก็บไว้ฐานข้อมูลขององค์กร และข้อมูลจากแหล่งข้อมูลภายนอก เช่น เว็บไซต์ โซเชียลมีเดีย เป็นต้น ซึ่งข้อมูลเหล่านี้สามารถนำไปสร้างมูลค่าเพิ่มได้ โดยการแปลงข้อมูลดิบให้กลายเป็นสารสนเทศ หรือการทำให้เกิดเป็นองค์ความรู้ใหม่ๆ ที่สามารถนำไปใช้ให้เกิดประโยชน์ต่อองค์กรต่อไป

การทำเหมืองข้อมูล (Data Mining) หรือเรียกทับศัพท์ว่า “ดาต้าไมนิ่ง” เป็นกระบวนการ เทคนิค และวิธีการที่ใช้เพื่อการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (Big Data) หรือข้อมูลที่มีจำนวนมหาศาล การทำเหมืองข้อมูลถูกนำมาใช้เป็นเครื่องมือในการค้นหาความรู้ใหม่ที่เป็นประโยชน์จากฐานข้อมูลขนาดใหญ่ที่ได้รับความนิยมในกลุ่มนักวิเคราะห์ข้อมูล

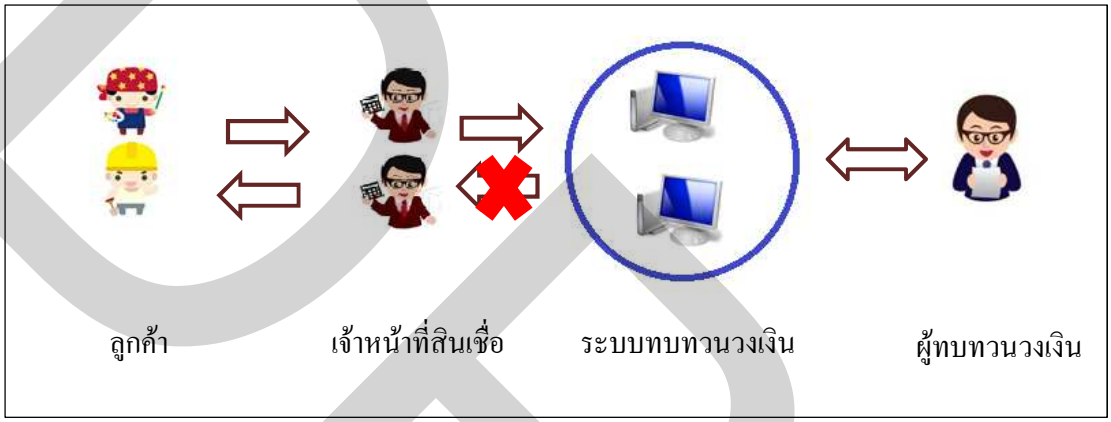
ประโยชน์ที่ได้รับจากการทำเหมืองข้อมูล คือ การค้นพบความรู้ใหม่ๆ ที่ยังไม่เคยถูกค้นพบมาก่อน และสามารถนำไปใช้ประโยชน์ได้ เช่น เพื่อการวางแผนกลยุทธ์ขององค์กรด้านต่าง ๆ เพื่อสร้างความได้เปรียบคู่แข่งทางธุรกิจ เพื่อสร้างความพึงพอใจในผลิตภัณฑ์และบริการของลูกค้า เพื่อการบรรลุเป้าหมายขององค์กรในการที่จะมีผลประกอบการที่ดี ซึ่งส่งผลต่อภาพลักษณ์และความมั่นคงขององค์กร ทำให้ผู้ถือหุ้นมีความเชื่อมั่นในศักยภาพขององค์กรยิ่งขึ้น นอกจากนี้ ประโยชน์จากการทำเหมืองข้อมูลยังสามารถนำรูปแบบองค์ความรู้ที่ได้ไปประยุกต์ใช้กับงานด้านอื่นๆ ได้ เช่น การนำไปใช้เพื่อการพัฒนาระบบสนับสนุนงานด้านปฏิบัติการต่างๆ ทำให้การปฏิบัติงานมีความคล่องตัวยิ่งขึ้น เป็นต้น

สถาบันการเงินอย่างธนาคารก็มีเป้าหมายในการประกอบธุรกิจเช่นเดียวกันกับองค์กรธุรกิจอื่นๆ นั่นคือ ความต้องการมีผลกำไรสูงสุด ดังนั้น การมีข้อมูลหรือความรู้ใหม่ๆ เพื่อสนับสนุนการดำเนินงานขององค์กรให้บรรลุเป้าหมายสูงสุดจึงเป็นสิ่งจำเป็นอย่างยิ่ง ซึ่งถือเป็นความรับผิดชอบที่สำคัญอย่างหนึ่งของหน่วยงานสนับสนุนข้อมูลทางธุรกิจของธนาคาร นอกจากนี้ การบริหารงานภายในของธนาคารยังมีการแบ่งหน้าที่รับผิดชอบอย่างชัดเจน เช่น งานด้านการบริหารความสัมพันธ์ลูกค้า กับงานด้านปฏิบัติการในสำนักงาน การแบ่งหน้าที่ความรับผิดชอบที่แยกขาดจากกันเช่นนี้ บางครั้งอาจทำให้เกิดปัญหาการขาดการเชื่อมโยงกันของงานที่เกี่ยวเนื่องกัน

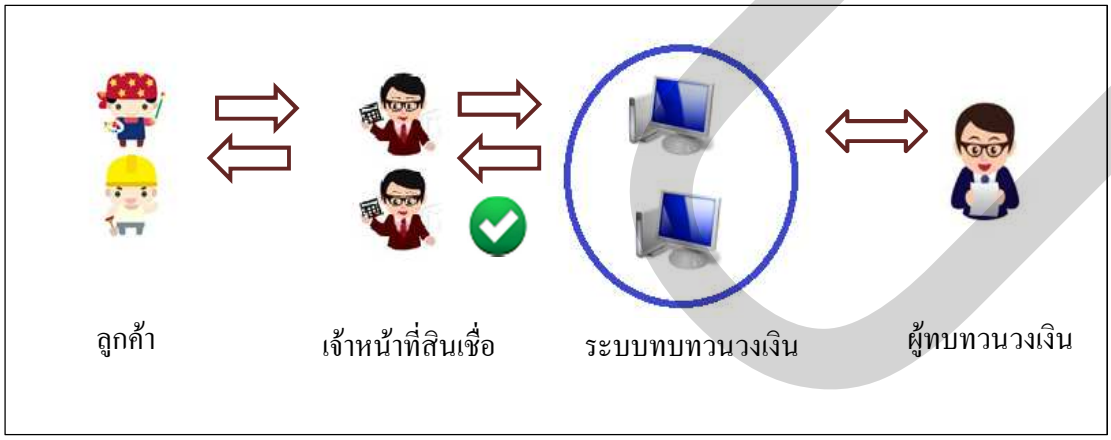
จากการพิจารณากระบวนการทำงานของงานทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชีวงเงิน ปัจจุบันพบว่าหน้าที่รับผิดชอบเกี่ยวกับงานด้านทบทวนวงเงิน ทุกขั้นตอนอยู่ในความควบคุมดูแลของเจ้าหน้าที่ทบทวนวงเงินเพียงหนึ่งคน ซึ่งถ้าหากเจ้าหน้าที่สินเชื่อต้องการสอบถามรายละเอียดต่างๆ ต้องไปสอบถามที่เจ้าหน้าที่ทบทวนวงเงินคนนั้นคนเดียวเท่านั้น และถ้าหากเจ้าหน้าที่ทบทวนวงเงินหยุดงาน ก็ต้องรอจนกว่าจะกลับมาปฏิบัติงาน ทำให้ไม่ได้รับความสะดวกในการสอบถามข้อมูลว่าลูกค้าที่อยู่ในความดูแลของตนเองนั้นอยู่ในเกณฑ์ หรือมีโอกาสที่จะได้รับการทบทวนวงเงินหรือไม่ มีปัจจัยที่จะนำมาใช้ประกอบการพิจารณาทบทวนวงเงินอะไรบ้าง และมีหลักเกณฑ์เงื่อนไขอย่างไร ทั้งนี้ เพื่อใช้เป็นข้อมูลสำหรับติดตามดูแลและให้บริการลูกค้าเพิ่มเติมต่อไป

ปัญหาดังกล่าวเกิดขึ้นเนื่องมาจากสาเหตุการขาดการเชื่อมโยงของงานในความรับผิดชอบระหว่างผู้ปฏิบัติงาน 2 ฝ่าย คือ เจ้าหน้าที่สินเชื่อซึ่งมีบทบาทหน้าที่ในปฏิบัติงานเป็นด่านหน้าของหน่วยธุรกิจของธนาคาร มีหน้าที่รับผิดชอบในการให้บริการและติดตามดูแลลูกค้าอย่างใกล้ชิด กับอีกฝ่ายหนึ่งคือ เจ้าหน้าที่ทบทวนวงเงิน ซึ่งมีบทบาทหน้าที่เป็นผู้ปฏิบัติงานเบื้องหลัง คือ เมื่อลูกค้าได้รับอนุมัติสินเชื่อเงินกู้เบิกเกินบัญชีเรียบร้อยแล้ว เจ้าหน้าที่ทบทวนวงเงินมีหน้าที่รับผิดชอบในการพิจารณาทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชีที่ใกล้ถึงวันครบ

กำหนดอายุสัญญา และถ้าหากพิจารณาแล้วเห็นว่าเป็นลูกค้าดีและอยู่ในเกณฑ์ที่กำหนด ก็จะทำการอนุมัติให้ขยายวันครบกำหนดอายุสัญญาออกไป ซึ่งลูกค้าจะสามารถใช้วงเงินนั้นๆ ต่อไปได้ โดยไม่จำเป็นต้องยื่นขอสินเชื่อใหม่อีกครั้ง ดังนั้น เพื่อเป็นการแก้ไขปัญหาให้กับเจ้าหน้าที่สินเชื่อ ผู้วิจัยจึงมีแนวคิดที่จะนำความรู้ด้านการทำเหมืองข้อมูลมาประยุกต์ใช้กับงานทบทวนวงเงินดังกล่าว และเป็นที่มาของงานวิจัยครั้งนี้



ภาพที่ 1.1 ปัญหาระบบงานทบทวนวงเงินแบบเดิม



ภาพที่ 1.2 ความต้องการแก้ไขปัญหาระบบงานทบทวนวงเงิน

1.2 คำถามงานวิจัย

1. ความรู้ที่ถูกค้นพบจากการทำเหมืองข้อมูลสามารถนำมาใช้กับการพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ได้จริงหรือไม่

2. ผลการพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน อันเกิดจากความรู้ที่ถูกค้นพบจากการทำเหมืองข้อมูล มีความถูกต้องและน่าเชื่อถือหรือไม่

3. ความรู้ที่ถูกค้นพบจากการทำเหมืองข้อมูล สามารถนำไปประยุกต์ใช้โดยพัฒนาเป็นโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ได้หรือไม่

4. โปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน สามารถแบ่งเบาภาระงานของเจ้าหน้าที่ผู้เกี่ยวข้อง ได้หรือไม่

1.3 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาเปรียบเทียบรูปแบบโมเดลการตัดสินใจที่ถูกค้นพบด้วยวิธีการจำแนกข้อมูลของการทำเหมืองข้อมูล กับหลักการจริงของงานทบทวนวงเงิน

2. เพื่อพัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

3. เพื่อประเมินทัศนคติของผู้ใช้งานที่มีต่อประสิทธิภาพของโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

1.4 ขอบเขตของการวิจัย

1. การวิจัยนี้ใช้ข้อมูลบัญชีเงินกู้เบิกเกินบัญชีที่เคยเข้าสู่กระบวนการทบทวนวงเงินล่วงหน้าระหว่างเดือนมกราคม 2556 ถึงเดือนกันยายน 2557 และเป็นข้อมูลของลูกค้าที่อยู่ในความดูแลของฝ่ายพิจารณาสินเชื่อธุรกิจ SME ของธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) จำนวนทั้งสิ้น 13,765 เร็คคอร์ด โดยทำการแบ่งข้อมูลสำหรับการดำเนินการวิจัยออกเป็น 2 ชุด โดยใช้เกณฑ์เดือนของข้อมูลเป็นตัวแบ่งชุดข้อมูล

2. ข้อมูลชุดแรกเป็นข้อมูลการทบทวนวงเงินตั้งแต่เดือนมกราคม 2556 ถึงเดือนสิงหาคม 2557 จำนวน 12,893 เร็คคอร์ด ชุดนี้ใช้สำหรับเรียนรู้และทดสอบโมเดลโดยโปรแกรม WEKA

3. ข้อมูลชุดที่สองเป็นข้อมูลการทบทวนวงเงิน ของเดือนกันยายน 2557 จำนวน 872 เร็คคอร์ด ชุดนี้ใช้สำหรับประเมินประสิทธิภาพโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

4. ผู้วิจัยทำการคัดเลือกแอตทริบิวท์ที่เป็นข้อมูลปัจจัยที่ใช้ในการพิจารณาตามหลักเกณฑ์การทบทวนวงเงินและข้อมูลที่เป็นผลการพิจารณา จำนวน 5 แอตทริบิวท์ ได้แก่ Utilization, LTV , Movement, Payment และ Class

5. ตัวแปรต้น ที่ใช้เป็นเงื่อนไขเพื่อการพยากรณ์ในงานวิจัยนี้ ได้แก่ แอตทริบิวท์ Utilization, LTV, Movement และ Payment ซึ่งเป็นข้อมูลเจ้าหน้าที่สินเชื่อสามารถหาได้จากข้อมูลบัญชีเงินกู้เบิกเกินบัญชีของลูกค้ารายนั้นๆ

6. แอตทริบิวท์ Class เป็นข้อมูลของผลการพิจารณาว่าบัญชีนั้นๆ ผ่านเกณฑ์การทบทวนวงเงินหรือไม่ หรือในงานวิจัยนี้เรียกว่า “ตัวแปรตาม” หรือสำหรับการทำเหมืองข้อมูลด้วยวิธีจำแนกข้อมูล เรียกตัวแปรตามนี้ว่า “คลาส” (Class)

7. งานวิจัยนี้ใช้โปรแกรม WEKA เวอร์ชั่น 3.6.2 สำหรับประมวลผลข้อมูลและสร้างโมเดลการตัดสินใจ เนื่องจากเป็นเวอร์ชั่นที่สนับสนุนการทำงานระหว่างฮาร์ดแวร์และซอฟต์แวร์ที่ใช้ในการพัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

8. การเรียนรู้รูปแบบการตัดสินใจของงานวิจัยนี้ ใช้วิธีการจำแนกข้อมูล (Classification) ของเทคนิคการทำเหมืองข้อมูล (Data Mining)

ดังนั้น เนื้อหาของงานวิจัยนี้ ผู้วิจัยจึงขอกล่าวถึงเฉพาะในส่วนที่เกี่ยวข้องกับขอบเขตของการวิจัย พร้อมกับอ้างอิงหลักการของการทำเหมืองข้อมูล กระบวนการสร้างแบบจำลองหรือโมเดลการตัดสินใจที่เกี่ยวข้องพอสังเขป

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. มีโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน เพื่อให้ผู้ใช้งานที่ต้องการคาดคะเนความน่าจะเป็น หรือประเมินโอกาสที่บัญชีเงินกู้เบิกเกินบัญชีของลูกค้าที่อยู่ในความดูแลของตนเองนั้น จะได้รับการทบทวนวงเงินหรือไม่ เพียงใด

2. ทำให้ทำการสอบถามข้อมูลหรือพยากรณ์โอกาสที่วงเงินเบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ของผู้ใช้งานทำได้ง่ายยิ่งขึ้น

3. ทำให้ผู้ใช้งานเห็นภาพรวมของหลักเกณฑ์และปัจจัยสำคัญๆ ที่ถูกนำมาใช้พิจารณาและมีผลต่อผลการพิจารณาทบทวนวงเงินในปัจจุบัน

4. ทำให้ทราบถึงทัศนคติของผู้ใช้งานที่มีต่อประสิทธิภาพของโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

1.6 นิยามศัพท์เฉพาะ

เงินกู้เบิกเกินบัญชี (Bank Overdraft) (ภิเษก ชัยนิรันดร์, 2553, น. 21-22) หมายถึง เงินกู้ที่ธนาคารอนุมัติให้บุคคลหรือกิจการที่มีความต้องการเพิ่มสภาพคล่องของเงินทุนหมุนเวียนในกิจการ โดยธนาคารจะพิจารณาตามสภาวะความต้องการใช้เงินหมุนเวียนของแต่ละกิจการ ซึ่งผู้กู้

จะสามารถทำรายการเบิกเงินเกินบัญชี จากบัญชีเงินฝากประเภทบัญชีกระแสรายวันได้ โดยที่ผู้กู้จะทำการออกเช็คเพื่อสั่งจ่ายค่าสินค้าและบริการได้ภายในวงเงินตามที่ได้ตกลงกันไว้ระหว่างธนาคารและผู้กู้ และภาระหนี้ที่เกิดขึ้นจากการเบิกเกินบัญชีนี้จะถูกคิดดอกเบี้ยตามอัตราที่ธนาคารกำหนด

การทบทวนวงเงิน (Credit Review) หมายถึง การที่สถาบันการเงินได้พิจารณาถึงความเสี่ยงที่อาจเกิดขึ้นกับวงเงินที่ได้อนุมัติไว้กับลูกค้ารายใดๆ และเมื่อระยะเวลาผ่านไปสภาพความเป็นจริงทางธุรกิจของลูกค้าอาจเปลี่ยนแปลงไป หรือมีความเสี่ยงที่อาจจะไม่สามารถชำระหนี้คืนได้ โดยเฉพาะอย่างยิ่งความเสี่ยงกับสินเชื่อประเภทเงินกู้เบิกเกินบัญชีเงิน (Bank Overdraft) ที่ลูกค้าสามารถใช้จ่ายเงินได้โดยง่ายและอาจใช้ผิดวัตถุประสงค์จากที่สถาบันการเงินอนุมัติไว้ ดังนั้น จึงต้องมีการกำหนดให้มีการทบทวนวงเงินของลูกค้าตามระยะที่กำหนด สำหรับข้อมูลที่ใช้เพื่อการทบทวนวงเงินนั้น โดยสถาบันการเงินจะใช้ข้อมูลจากประวัติการทำธุรกรรมทางการเงิน การติดต่อและภาระผูกพันธ์ต่างๆ ระหว่างสถาบันการเงินกับลูกค้ารายนั้นๆ มาประกอบการพิจารณา

แบบจำลอง (Model) คือ ตัวแบบที่ช่วยในการนำเสนอข้อมูลต่างๆ เพื่อจะนำข้อมูลเหล่านั้นไปใช้ในการแก้ไขปัญหา แบบจำลองมีหลายประเภท เช่น แบบจำลองเชิงบรรยาย (Description Model) แบบจำลองคงที่และแบบพลวัตร (Static And Dynamic Model) และแบบจำลองทางคณิตศาสตร์ (Mathematics Model) เป็นต้น (Jiawei Han, Jian Pei & Micheline Kamber, 2011) ประโยชน์ของแบบจำลอง คือ ทำให้ประหยัดเวลาและค่าใช้จ่ายในการวิเคราะห์ข้อมูล มีความน่าเชื่อถือ ผู้ใช้งานสามารถทำความเข้าใจและมองเห็นภาพได้ชัดเจนขึ้น แบบจำลองบางประเภทสามารถนำไปทดลองแทนมนุษย์ในสถานการณ์ที่อันตรายได้ เครื่องมือที่ใช้ในการสร้างแบบจำลอง ได้แก่ การเขียนโปรแกรมด้วยภาษาทางคอมพิวเตอร์ การใช้โปรแกรม Spread Sheet การใช้สูตรทางคณิตศาสตร์และสถิติ รวมถึงการใช้โปรแกรมวิเคราะห์ข้อมูลสำเร็จรูปอื่นๆ

Utilization คือ สัดส่วนร้อยละของจำนวนเงินที่มีการเบิกใช้เกินบัญชีของบัญชีเงินกู้เบิกเกินบัญชีต่อวงเงินที่ได้รับอนุมัติให้สามารถเบิกเกินบัญชีได้

Loan To Value หรือ LTV คือ สัดส่วนภาระหนี้สินทั้งสิ้นต่อมูลค่าหลักประกันของลูกค้า

Movement คือ สัดส่วนร้อยละของจำนวนเงินเข้า-ออก หรือเรียกว่า รายการเคลื่อนไหวทางบัญชี ต่อวงเงินที่ได้รับอนุมัติให้สามารถเบิกเกินบัญชีได้

Payment History คือ ประวัติการจ่ายชำระหนี้ของลูกค้า ซึ่งงานวิจัยนี้ใช้ประวัติการจ่ายชำระหนี้ย้อนหลัง 24 เดือน

Class คือ ผลการพิจารณาว่าบัญชีเงินกู้เบิกเกินบัญชีนั้นๆ ผ่านเกณฑ์การทบทวนวงเงินหรือไม่ โดยมีค่าข้อมูลเป็น Pass และ No

บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

2.1 ประวัติธนาคารกรุงศรีอยุธยา จำกัด (มหาชน)

ธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) (2556) ก่อตั้งขึ้นเมื่อวันที่ 27 มกราคม 2488 และได้นำหลักทรัพย์เข้าจดทะเบียนในตลาดหลักทรัพย์แห่งประเทศไทยในปี 2520 ปัจจุบันเป็นธนาคารพาณิชย์ที่มีเงินให้สินเชื่อและเงินฝากใหญ่เป็นอันดับ 5 ของประเทศ และเป็นธนาคารที่ให้บริการทางการเงินอย่างครบวงจรแก่ลูกค้าธุรกิจและลูกค้าบุคคล

การเปลี่ยนแปลงครั้งสำคัญของธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) มีดังนี้

1. 3 มกราคม 2550 จีอี มันนี่ ซึ่งเป็นสถาบันการเงินเพื่อรายย่อยชั้นนำของโลกได้บรรลุข้อตกลงการเป็นพันธมิตรทางธุรกิจกับธนาคารกรุงศรีอยุธยา

2. 14 กุมภาพันธ์ 2551 ธนาคารฯ เข้าซื้อกิจการบริษัท จีอี แคปปิตอล ออโต้ ลีส จำกัด (มหาชน) (GECAL) ทำให้พอร์ตสินเชื่อของธนาคารฯเติบโตขึ้นประมาณร้อยละ 17 และต่อมา GECAL เปลี่ยนชื่อเป็น บริษัท อยุธยา แคปปิตอล ออโต้ ลีส จำกัด (มหาชน) (AYCAL)

3. 8 เมษายน 2552 ธนาคารฯ เข้าซื้อหุ้นธนาคาร เอไอจี เพื่อรายย่อย จำกัด (มหาชน) (AIGRB) และ บริษัท เอไอจี คาร์ด (ประเทศไทย) จำกัด (AIGCC) ทำให้ธนาคารฯ มีสินทรัพย์เพิ่มขึ้นประมาณ 32,800 ล้านบาท มีฐานสินเชื่อเพิ่มขึ้น 21,900 ล้านบาท เงินฝากเพิ่มขึ้น 18,600 ล้านบาท และมีจำนวนบัตรเครดิตเพิ่มขึ้นประมาณ 222,000 บัตร

4. 9 กันยายน 2552 ธนาคารฯ เข้าซื้อกิจการบริษัท ซีเอฟจี เซอร์วิสเซส จำกัด (CFGS) ซึ่งเป็นบริษัทในเครือของบริษัทอเมริกัน อินเตอร์เนชั่นแนล กรุ๊ป อิงค์ จำกัด ทั้งนี้ CFGS เป็นหนึ่งในผู้ให้บริการไมโครไฟแนนซ์โดยเฉพาะสินเชื่อที่ใช้ทะเบียนรถเป็นหลักประกัน มีสาขากว่า 163 แห่งทั่วประเทศ และเป็นที่รู้จักในนาม "ศรีสวัสดิ์ เงินติดล้อ"

5. 5 พฤศจิกายน 2552 ธนาคารฯ เข้าซื้อธุรกิจการเงินเพื่อผู้บริโภคของจีอี แคปปิตอล ในประเทศไทย ทำให้ธุรกิจการเงินเพื่อลูกค้ารายย่อยของธนาคารฯ เติบโตเร็วขึ้น และทำให้สัดส่วนสินเชื่อรายย่อยของธนาคารเพิ่มขึ้นจาก 36% เป็น 42% ของสินเชื่อรวม การผนวกธุรกิจของจีอี มันนี่ ประเทศไทย เข้ากับธนาคารทำให้ธนาคารฯ กลายเป็นผู้ให้บริการบัตรเครดิตรายใหญ่ที่สุดใน

ประเทศไทย ด้วยจำนวนบัตรหมุนเวียนในระบบมากกว่า 3 ล้านใบ และให้บริการลูกค้ากว่า 8 ล้านราย

6. 25 มกราคม 2555 คณะกรรมการธนาคารมีมติอนุมัติเห็นชอบให้ธนาคารฯ ดำเนินการรับโอนธุรกิจลูกค้ารายย่อยในประเทศไทยของธนาคารฮ่องกงและเซี่ยงไฮ้แบงกิ้งคอร์ปอเรชั่น จำกัด (HSBC) ซึ่งประกอบธุรกิจบัตรเครดิต ธุรกิจสินเชื่อส่วนบุคคล ธุรกิจสินเชื่ออสังหาริมทรัพย์ ธุรกิจเงินฝากส่วนบุคคลและตั๋วแลกเงิน

7. 18 ธันวาคม 2556 มิตซูบิชิ ยูเอฟเจ ไฟแนนเชียล กรุ๊ป (MUFG) เข้ามาเป็นพันธมิตรทางธุรกิจกับธนาคารฯ แทนจีอี มันนี่

2.2 ความหมายสินเชื่อในทางเศรษฐศาสตร์

สำนักงานเศรษฐกิจการคลัง (2549) สินเชื่อ หมายถึง อำนาจในการเข้าถึงการใช้สินค้าและบริการ โดยสัญญาว่าจะชดใช้คืนในอนาคต โดยสินเชื่ออาจอยู่ในรูปของสินค้าและบริการ หรืออยู่ในรูปของเงินก็ได้

ในทางเศรษฐศาสตร์มองว่า หน่วยเศรษฐกิจในระบบเศรษฐกิจนั้นอาจมีลักษณะการใช้ทรัพยากรที่ไม่สมดุลกัน กล่าวคือ บางหน่วยเศรษฐกิจอาจมีความต้องการใช้สินค้าและบริการที่มากกว่าทรัพยากรที่ตนมีอยู่ในปัจจุบัน ในขณะที่บางหน่วยเศรษฐกิจอาจมีทรัพยากรเหลือใช้เกินความต้องการ ดังนั้น หากมีการโอนทรัพยากรส่วนที่เหลือใช้นั้นไปให้ผู้ที่มีความต้องการใช้ทรัพยากรและก่อให้เกิดประโยชน์ได้ในอนาคต โดยมีข้อตกลงว่าจะมีการชำระคืนแก่เจ้าของเดิมในอนาคต การจัดสรรการใช้ทรัพยากรต่างๆ ในระบบเศรษฐกิจก็จะมีประสิทธิภาพมากขึ้น

2.3 ความสัมพันธ์ระหว่างสินเชื่อกับระบบเศรษฐกิจ

สำนักงานเศรษฐกิจการคลัง (2549) ความสัมพันธ์ระหว่างสินเชื่อกับระบบเศรษฐกิจโดยมีภาคการเงินเป็นส่วนส่งเสริมให้เกิดการใช้ทรัพยากรในภาคเศรษฐกิจได้อย่างมีประสิทธิภาพ เริ่มจากการที่หน่วยเศรษฐกิจที่มีรายได้มากกว่าความต้องการใช้จ่าย ทำให้เกิดส่วนรั่วไหลจากระบบเศรษฐกิจกลายเป็นการออมเงินในรูปแบบต่างๆ เช่น การออมเงินผ่านระบบธนาคารพาณิชย์ เป็นต้น จากนั้นหน่วยเศรษฐกิจที่มีความต้องการใช้จ่ายมากกว่าทรัพยากรที่ตนมีอยู่ ก็จะไปขอกู้เงินหรือขอสินเชื่อจากหน่วยเศรษฐกิจที่รับออมเงิน โดยมีอัตราดอกเบี้ยเป็นราคาของทุนและใช้เป็นเครื่องมือช่วยให้การจัดสรรทรัพยากรจากผู้ออมไปสู่ผู้กู้ ให้เป็นไปอย่างมีประสิทธิภาพ โดยผู้กู้นั้นไม่ว่าจะเป็นกรณีการใช้นโยบายขาดดุลงบประมาณของภาครัฐ หรือภาคเอกชน เม็ดเงินจากการกู้ยืมเหล่านี้จะถูกนำไปใช้จ่าย ซึ่งเป็นการอัดฉีดเงินที่รั่วไหลกลับเข้าสู่ภาคเศรษฐกิจอีกครั้ง

2.4 ความรู้ทั่วไปด้านสินเชื่อ (ภิเษก ชัยนิรันดร์, 2553, น. 9 - 22)

1. ความหมายของสินเชื่อ Joseph French Johnson นักเศรษฐศาสตร์ชาวอเมริกัน ให้นิยามของสินเชื่อไว้ว่า “Credit is the power to obtain goods or service by given a promise to pay at a specified date in the future”

สินเชื่อ คือ ความสามารถที่จะได้มาซึ่งสินค้าหรือบริการ โดยการให้คำมั่นสัญญาว่าจะชำระคืนตามวันที่กำหนดไว้ในอนาคต

จากนิยามข้างต้นแสดงถึงองค์ประกอบของสินเชื่อ 3 ส่วน ได้แก่

1.1 ความสามารถ หมายถึง ความสามารถของผู้กู้ที่จะสร้างความน่าเชื่อถือให้ประจักษ์ต่อผู้ให้กู้เกิดความมั่นใจว่า หากให้สินเชื่อแก่ผู้กู้แล้ว ผู้กู้มีศักยภาพเพียงพอที่จะบริหารจัดการองค์กรได้อย่างมีประสิทธิภาพ นำพาธุรกิจขององค์กรเจริญก้าวหน้าได้อย่างยั่งยืน และมีรายได้เพียงพอกับการจ่ายชำระหนี้ในอนาคต

1.2 การได้มาซึ่งสินค้าและบริการ หมายถึง ผู้กู้ต้องนำเงินจากการกู้ยืมนั้นไปใช้ให้ตรงตามวัตถุประสงค์ของการอนุมัติสินเชื่อ

1.3 คำมั่นสัญญาว่าจะชำระคืนตามวันที่กำหนด หมายถึง พันธสัญญาร่วมกันระหว่างผู้กู้มีต่อผู้ให้กู้ที่จะหาเงินมาจ่ายชำระหนี้ ภายในวันหรือระยะเวลาที่กำหนดไว้

2. ความหมายของสินเชื่อ ในกรณีกู้ยืมเงินตามประมวลกฎหมายแพ่งและพาณิชย์

การกู้ยืมเงิน ตามประมวลกฎหมายแพ่งและพาณิชย์ถือเป็นสัญญายืมใช้สิ้นเปลือง ที่มีวัตถุแห่งสัญญาเป็นเงินตรา ตามมาตรา 650 บัญญัติไว้ ดังนี้

“อันว่ายืมใช้สิ้นเปลือง คือ สัญญาซึ่งผู้ให้ยืมโอนกรรมสิทธิ์ในทรัพย์สินชนิดใช้ไปสิ้นไปนั้น เป็นปริมาณมีกำหนดให้แก่ผู้ยืมและผู้ยืมตกลงว่าจะคืนทรัพย์สินประเภท ชนิด และปริมาณเช่นเดียวกันให้แทนทรัพย์สินซึ่งให้ยืมนั้น สัญญาย่อมบริบูรณ์ต่อเมื่อส่งมอบทรัพย์สินที่ยืม”

3. หลักการพิจารณาในการเลือกใช้สินเชื่อระยะสั้นและสินเชื่อระยะยาว

โดยทั่วไปแหล่งที่มาของเงินทุนที่ใช้เพื่อการดำเนินธุรกิจมาจาก 2 แหล่ง คือ จากส่วนของเจ้าของ หรือเรียกว่า “ทุน” และจากการกู้ยืม สำหรับหลักการในการพิจารณาว่าการที่กิจการจะลงทุนในสินทรัพย์ถาวรและสินทรัพย์หมุนเวียนนั้น ควรจะเลือกใช้เงินทุนจากแหล่งใด และถ้าหากเลือกใช้วิธีการกู้ยืมจะใช้สินเชื่อประเภทใดและมีระยะเวลาแบบใด มีหลักการพิจารณาอยู่ 3 รูปแบบ ดังนี้

3.1 พิจารณาตามหลักการจับคู่ตามระยะเวลา (Maturity Matching Approach) หลักการ คือ กรณีที่กิจการต้องการลงทุนในสินทรัพย์ถาวร เช่น เพื่อซื้อที่ดิน อาคาร และอุปกรณ์ ควรใช้สินเชื่อระยะยาว เนื่องจากสินทรัพย์ถาวรเหล่านี้ ต้องใช้ระยะเวลานานพอสมควร ในการที่จะสร้างรายได้

และสามารถนำเงินมาชำระหนี้ได้ ในกรณีกิจการที่ต้องการลงทุนในสินทรัพย์หมุนเวียน เช่น เพื่อซื้อสินค้าเข้าร้าน ควรใช้สินเชื่อระยะสั้น เช่น สินเชื่อเงินกู้เบิกเกินบัญชี (Bank Overdraft) หรือเรียกว่าวงเงินโอดี ทั้งนี้ เพื่อเป็นการเสริมสภาพคล่องของกิจการให้มีเงินทุนเพียงพอต่อการซื้อสินค้าหรือบริการ ในสถานการณ์ที่กิจการค้าขายดี กิจการก็มีจำเป็นที่ต้องสำรองสินค้าไว้เพื่อขายในปริมาณมากขึ้นจากเดิม และเมื่อมีรายได้ก็นำไปชำระหนี้เงินต้นพร้อมดอกเบี้ย จากนั้นวงเงินของบัญชีเงินกู้เบิกเกินบัญชีที่เคยอนุมัติไว้ ก็จะกลับมาคงดังเดิมและสามารถใช้เป็นเงินทุนหมุนเวียนในกิจการต่อไปได้อีกเรื่อยๆ จนกว่าจะสิ้นสุดระยะเวลาของสัญญากู้ยืม

3.2 พิจารณาตามหลักการที่ค่อนข้างเสี่ยง (Aggressive Approach) หลักการคือ ในสถานการณ์ที่อัตราดอกเบี้ยของสินเชื่อระยะสั้นต่ำกว่าอัตราดอกเบี้ยสินเชื่อระยะยาว กิจการอาจพิจารณาเลือกใช้สินเชื่อระยะสั้น เพื่อนำมาลงทุนในสินทรัพย์ถาวรของกิจการ หลักการนี้ข้อเสีย คือ สินทรัพย์ถาวรนั้นอาจสร้างรายได้ไม่ทันกับรอบการจ่ายชำระหนี้ อาจทำให้โอกาสที่จะไม่สามารถนำเงินมาจ่ายชำระหนี้ภายในระยะเวลาที่กำหนดมีค่อนข้างสูง

3.3 พิจารณาตามหลักความปลอดภัยไว้ก่อน (Conservative Approach) หลักการคือ กิจการอาจใช้จุดแข็งสินเชื่อระยะยาวที่มีกำหนดระยะเวลาการชำระคืนค่อนข้างนาน แต่อาจมีความเสี่ยงเรื่องความผันผวนของอัตราดอกเบี้ยลอยตัวในอนาคต ดังนั้น บางกิจการอาจนำเงินทุนที่เหลือใช้จากสินเชื่อระยะยาวไปลงทุนในสินทรัพย์หมุนเวียนที่มีสภาพคล่องสูงและสามารถขายคืนหรือแปลงเป็นเงินสดได้เร็ว

4. ประเภทของสินเชื่อ แบ่งตามลักษณะต่างๆ ได้ดังนี้

4.1 ประเภทสินเชื่อแบ่งตามลักษณะของการชำระคืน ได้แก่ สินเชื่อชำระครั้งเดียว (Single Payment Credit) และสินเชื่อผ่อนส่ง (Installment Credit)

4.2 ประเภทสินเชื่อแบ่งตามระยะเวลา ได้แก่ สินเชื่อระยะสั้น (Short Term Credit) เป็นสินเชื่อที่มีกำหนดระยะเวลาชำระคืนภายใน 1 ปี สินเชื่อระยะปานกลาง (Intermediate Term Credit) เป็นสินเชื่อที่มีกำหนดระยะเวลาชำระคืนตั้งแต่ 1 - 5 ปี และสินเชื่อระยะยาว (Long Term Credit) เป็นสินเชื่อที่มีกำหนดระยะเวลาชำระคืนตั้งแต่ 5 ปี ขึ้นไป ในทางปฏิบัติโดยทั่วไปนิยมเรียกสินเชื่อที่มีกำหนดระยะเวลาชำระคืนตั้งแต่ 1 ปีขึ้นไปว่า “สินเชื่อระยะยาว”

4.3 ประเภทสินเชื่อแบ่งตามหลักประกัน ได้แก่ สินเชื่อที่มีหลักประกัน (Secured Credit) สินเชื่อที่ไม่มีหลักประกัน (Unsecured Credit)

4.4 ประเภทสินเชื่อแบ่งตามลักษณะของผู้ให้สินเชื่อ ได้แก่ สินเชื่อรัฐบาลหรือสินเชื่อสาธารณะ (Public Credit) เช่น ตั๋วเงินคลัง พันธบัตรรัฐบาล การกู้ยืมทางตรงของรัฐบาลทั้งในและต่างประเทศ และสินเชื่อในภาคเอกชน (Private Credit) ได้แก่ สินเชื่อทุกประเภทของสถาบัน

การเงิน (Banking Credit) และสินเชื่อการค้า (Trade Credit) หรือการให้เครดิตวันชำระเงินค่าสินค้าระหว่างกิจการคู่ค้าด้วยกันเอง รวมถึงการออกตราสารหนี้เพื่อระดมทุนของกิจการ เป็นต้น

2.5 แนวทางการแสวงหาความรู้

ณรงค์ โพธิ์พฤกษานันท์ (2556, น. 16 - 29) ในอดีตมนุษย์ค้นหาความรู้ต่างๆ จากสภาพแวดล้อมทางธรรมชาติรอบตัวและวิธีการแสวงหาความรู้ดังกล่าวยังไม่เป็นระบบระเบียบเท่าที่ควร แนวทางการแสวงหาความรู้ มีดังนี้

1. การลองผิดลองถูก เป็นความรู้ที่ได้จากการปฏิบัติแล้วจดจำไปใช้ในคราวต่อไป
2. ความบังเอิญ เป็นความรู้ที่ได้มาโดยไม่คาดคิดมาก่อน
3. ความเชื่อหรือขนบธรรมเนียมประเพณีที่สืบทอดมาแต่โบราณ
4. ความรู้จากผู้รู้ หรือนักปราชญ์ หรือผู้เชี่ยวชาญ
5. ความรู้จากประสบการณ์จริง หรือประสาทสัมผัส
6. ความรู้จากการหยั่งรู้ หรือการรู้แจ้ง
7. ความรู้จากการหาเหตุผล มี 2 วิธี คือ

7.1 การหาเหตุผลแบบนิรนัย (Deductive Method) หรือวิธีอนุมาน เป็นการค้นหาความรู้จากการเก็บรวบรวมข้อเท็จจริงใหญ่ไปหาข้อเท็จจริงย่อย แล้วสรุปผล การหาความรู้รูปแบบนี้ อริสโตเติล (Aristotle) เป็นผู้ใช้เป็นคนแรก เรียกว่า “Syllogistic Method” หรือ “Aristotle Deduction”

7.2 การหาเหตุผลแบบอุปนัย (Inductive Method) หรือวิธีอุปมาน เป็นการค้นหาความรู้จากการเก็บรวบรวมข้อเท็จจริงย่อยไปหาข้อเท็จจริงใหญ่ แล้วสรุปผล เป็นทฤษฎีการหาเหตุผลในแบบของ ฟรานซิส เบคอน (Francis Bacon) ซึ่งมีความคิดเห็นขัดแย้งและต้องการแก้ไขจุดอ่อนของทฤษฎีการค้นหาความรู้แบบนิรนัยของอริสโตเติล (Aristotle) รูปแบบการหาความรู้แบบอุปนัย แบ่งออกเป็น 2 แบบ คือ

7.2.1 อุปนัยแบบสมบูรณ์ (Perfect Induction) เป็นการหาความรู้โดยการเก็บรวบรวมข้อมูลทุกหน่วยของประชากร (Population) ที่จะศึกษา

7.2.2 อุปนัยแบบไม่สมบูรณ์ (Imperfect Induction) เป็นการหาความรู้โดยการเก็บรวบรวมข้อมูลจากบางหน่วยของประชากร (Population) ที่จะศึกษา แล้วนำผลที่ได้ไปอ้างอิงว่าเป็นการศึกษาของประชากรทั้งหมด

8. วิธีทางวิทยาศาสตร์ จากแนวคิดของ ชาร์ล ดาร์วิน (Charles Darwin) ที่ได้นำรูปแบบการค้นหาความรู้ทั้งแบบนิรนัยและอุปนัยมาใช้ในกระบวนการหาความรู้ทางวิทยาศาสตร์ธรรมชาติ

โดยวิธีการค้นหาความรู้แบบนี้ เรียกว่า วิธีการอนุมาน-อุปมาน (Deductive-Inductive Method) คือ การใช้วิธีการอนุมานคำตอบหรือตั้งสมมติฐานก่อน แล้วใช้วิธีอุปมานเก็บรวบรวมข้อมูลเพื่อทดสอบสมมติฐาน ต่อมา จอห์น ดิวอี้ (John Dewey) ได้ปรับปรุงแนวคิดดังกล่าวที่เรียกว่า “Reflective Thinking” โดยแบ่งขั้นตอนการแก้ปัญหาออกเป็น 5 ขั้นตอน และในเวลาต่อมาเรียกว่า “วิธีการทางวิทยาศาสตร์” ประกอบด้วย

8.1 กำหนดปัญหา (Problem)

8.2 ตั้งสมมติฐาน (Hypothesis)

8.3 เก็บรวมรวมข้อมูล (Collection of Data)

8.4 วิเคราะห์ข้อมูล (Analysis of Data)

8.5 สรุปผล (Conclusion)

2.6 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลเป็นเทคนิคที่ใช้สำหรับการวิเคราะห์ข้อมูล เพื่อค้นหาความรู้ที่มีประโยชน์จากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in Databases - KDD) โดยอาศัยหลักสถิติ หลักคณิตศาสตร์ และการเรียนรู้ของเครื่องจักร โดยเครื่องจักรในที่นี้หมายถึงโปรแกรมวิเคราะห์ข้อมูลที่ติดตั้งไว้ในเครื่องคอมพิวเตอร์ มากระทำกับข้อมูลจำนวนมหาศาลโดยอัตโนมัติ ความรู้ที่ได้จะแสดงออกมาในลักษณะที่เป็นรูปแบบ (Pattern) กฎ (Rule) และความสัมพันธ์ที่ซ่อนอยู่ในข้อมูลเหล่านั้น

วิธีการที่จะทำให้โปรแกรมวิเคราะห์ข้อมูลที่ติดตั้งในคอมพิวเตอร์เรียนรู้ข้อมูล (เอกสิทธิ์ พัชรวงศ์ศักดา, 2556, น. 42) มี 2 แนวทาง คือ

1. การวิเคราะห์ข้อมูลโดยการเรียนรู้ข้อมูลในอดีตที่มีชุดคำตอบเตรียมไว้ล่วงหน้า (Supervised Learning) เพื่อใช้เป็นแบบแผนในการวิเคราะห์ข้อมูล หรือเรียกว่าเป็นการสอนคอมพิวเตอร์จากข้อมูลที่ถูกจัดเตรียมแนวคำตอบไว้เรียบร้อยแล้ว

2. การวิเคราะห์ข้อมูลโดยไม่มีการเตรียมชุดคำตอบไว้ล่วงหน้า (Unsupervised Learning) เป็นการให้โปรแกรมคอมพิวเตอร์ทำการวิเคราะห์ข้อมูลโดยการหาความเหมือนหรือความคล้ายคลึงกัน หรือที่เรียกว่า รูปแบบความสัมพันธ์กันของข้อมูลโดยอัตโนมัติ

2.6.1 รูปแบบวิธีการวิเคราะห์ข้อมูลของการทำเหมืองข้อมูล มีดังนี้

1. วิธีจำแนกข้อมูล (Classification) เพื่อใช้ในการพยากรณ์ (Prediction) หรือเพื่อการทำนายผล หรือเพื่อการคาดคะเนคำตอบ วิธีนี้จะพยากรณ์ข้อมูลตามโมเดลต้นแบบที่ได้จากการเรียนรู้ข้อมูลในอดีต

2. วิธีหาความสัมพันธ์ของข้อมูล (Association) วิธีนี้ใช้สำหรับค้นหาความสัมพันธ์ของข้อมูลรายการที่เกิดขึ้นซ้ำกันบ่อยๆ

3. วิธีจัดกลุ่มข้อมูล (Clustering) วิธีนี้ใช้เพื่อจัดกลุ่มข้อมูลที่มีลักษณะคล้ายกันไว้ด้วยกัน

2.6.2 มาตรฐานในการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล

ในปี 2539 บริษัท Daimler Chrysler บริษัท SPSS และบริษัท NCR (เอกสิทธิ์ พัชรวงศ์ศักดา, 2556, น. 10-12) ได้ร่วมกันกำหนดมาตรฐานของการวิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูล เรียกว่า “Cross-Industry Standard Process for Data Mining: CRISP-DM” ประกอบด้วย 6 ขั้นตอน ดังนี้

1. ขั้นตอนการทำความเข้าใจปัญหาทางธุรกิจ และวางแผนแนวทางดำเนินการเบื้องต้น (Business Understanding)

2. ขั้นตอนการเก็บรวบรวม การตรวจสอบ และวิเคราะห์ข้อมูลที่จำเป็นต้องใช้ในการวิเคราะห์ข้อมูล (Data Understanding)

3. ขั้นตอนการเตรียมข้อมูล (Data Preparation) คือ การปรับปรุงคุณภาพข้อมูลให้อยู่ในสภาพที่สามารถนำไปใช้ในการทำเหมืองข้อมูลได้ วิธีการที่ใช้สำหรับการเตรียมข้อมูล (ฉวีวรรณ เพ็ชรศิริ, 2556; Jiawei Han, Jian Pei & Micheline Kamber, 2011) มี 4 วิธี ดังนี้

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

3.2 การผสานข้อมูล (Data Integration) คือ การผสานข้อมูลที่รวบรวมมาจากแหล่งข้อมูลที่ต่างกัน เพื่อลดความซ้ำซ้อนของข้อมูล (Data Redundancies) และช่วยลดปัญหาความไม่สอดคล้องกันของข้อมูล (Data Inconsistencies)

3.3 การแปลงข้อมูล (Data Transformation) คือ การทำนอร์มอลไลซ์ (Normalization) ด้วยการแปลงค่าข้อมูลให้อยู่ในช่วงสั้นๆ เช่น (0,1)

3.4 การลดรูปข้อมูล (Data Reduction) คือ การจัดการข้อมูลให้อยู่ในรูปแบบกะทัดรัด ประหยัดพื้นที่ในการเก็บข้อมูล

4. ขั้นตอนการวิเคราะห์และออกแบบ (Modeling) เพื่อให้ได้รูปแบบหรือโมเดล

5. ขั้นตอนการวัดประสิทธิภาพของรูปแบบหรือโมเดล (Evaluation) ว่ามีความถูกต้องตรงตามวัตถุประสงค์ที่ตั้งไว้หรือไม่

6. ขั้นตอนการนำผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลไปใช้ให้เกิดประโยชน์ต่อองค์กร (Deployment)

2.7 ตัวจำแนกข้อมูลแบบ C4.5 หรือ J48

ตัวจำแนกข้อมูลแบบ C4.5 หรือในโปรแกรม WEKA เรียกว่าเทคนิค “J48” (เสถียร วิชาเรือง, 2553, น. 52) เป็นอัลกอริธึมสำหรับการจำแนกข้อมูลในกลุ่มเทคนิคแผนภาพต้นไม้ตัดสินใจ (Decision Tree) เป็นแบบจำลองหรือโมเดลการตัดสินใจทางคณิตศาสตร์สถิติ ใช้สำหรับหาทางเลือกที่ดีที่สุด เหมาะสำหรับใช้เพื่อแก้ปัญหาที่ไม่ซับซ้อน และมีทางเลือกเพื่อการตัดสินใจน้อย จุดเด่นของแผนภาพต้นไม้ตัดสินใจ คือ ช่วยให้ผู้ใช้งานมองเห็นภาพรวมของรูปแบบการตัดสินใจชัดเจนขึ้น

$$\text{อัตราส่วนความน่าจะเป็น} = \frac{\text{ความน่าจะเป็นที่จะเกิดเหตุการณ์}}{\text{ความน่าจะเป็นที่จะไม่เกิดเหตุการณ์}}$$

แผนภาพต้นไม้ตัดสินใจเป็นแบบจำลองที่ได้จากการเรียนรู้รูปแบบองค์ประกอบที่ใช้ในการตัดสินใจโดยโปรแกรมคอมพิวเตอร์ เพื่อพยากรณ์หรือคาดคะเนคำตอบที่ควรจะเป็นตามหลักการความน่าจะเป็น จากข้อมูลชุดที่ใช้เพื่อการเรียนรู้นั้น การเรียนรู้แบบนี้เรียกว่า “การเรียนรู้แบบมีครู”(Supervised Learning) (สิทธิโชค มุกดาสกุลภิบาล, 2551; Jiawei Han, Jian Pei &Micheline Kamber, 2011; N. Vandana Sawant, Ketan Shah & Vinayak Ashok Bharadi, 2011) 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

Berry, Michael J.A. & Linoff, Gordon S. (2004, pp. 166 - 167) ได้เปรียบเทียบลักษณะของการจำแนกข้อมูลไว้ว่า การจำแนกข้อมูลมีลักษณะคล้ายกับการเล่นเกมส์ 20 คำถาม โดยที่ผู้เล่นคนหนึ่งๆ พยายามค้นหาคำตอบโดยจะถามคำถามเชิงกว้างแล้วค่อยๆ แคบลงไปเรื่อยๆ จนเกิดความมั่นใจและสามารถระบุคำตอบได้ว่าสิ่งนั้นคืออะไร

Ohlhorst, Frank. (2013, p.4) ให้นิยามของการทำเหมืองข้อมูล (Data Mining) ไว้ว่า “Data Mining is process in which data are analyzed from different perspectives and then turned into summary data that are deemed useful. Data Mining is normally used with data at rest or with archival data. Data Mining techniques focus on modeling and knowledge discovery for predictive,

rather than purely descriptive, purposed—an ideal process for uncovering new patterns from data sets."

กระบวนการสร้างแบบจำลอง (Model) หรือการสร้างโมเดลตามทฤษฎีของการทำเหมืองข้อมูล ต้องมีการแบ่งข้อมูลออกเป็น 2 ชุด ข้อมูลชุดแรก เรียกว่า Training Data เป็นข้อมูลที่ใช้สำหรับสอนโปรแกรมคอมพิวเตอร์ให้ทำการเรียนรู้รูปแบบปัญหาและหาข้อสรุป แล้วนำไปสร้างเป็นแบบจำลองหรือที่เรียกว่าโมเดล และข้อมูลชุดที่สอง เรียกว่า Testing Data เป็นข้อมูลที่ใช้สำหรับทดสอบประสิทธิภาพของโมเดลที่โปรแกรมวิเคราะห์ข้อมูลนั้นค้นพบจากข้อมูลชุด Training Data ว่ามีความถูกต้องน่าเชื่อถือมากน้อยเพียงใด ทั้งนี้ ขนาดของข้อมูลและรูปแบบการเลือกใช้ข้อมูลเพื่อการ Training และ Testing นั้นขึ้นอยู่กับความต้องการและความเชี่ยวชาญของผู้วิเคราะห์ข้อมูล

งานวิจัยนี้ได้ทำการประมวลผลและวิเคราะห์ข้อมูลโดยใช้โปรแกรม WEKA เวอร์ชั่น 3.6.2 ซึ่งมีความสามารถพิเศษที่ผู้ใช้สามารถเลือกได้ว่าจะทำการแบ่งข้อมูลเพื่อการเรียนรู้ (Training) และเพื่อการทดสอบ (Testing) โมเดลแบบใดได้ในคราวเดียวกัน

2.8 งานวิจัยที่เกี่ยวข้อง

ปิยะมาศ กรัณย์ภควุฒิ (2556) ได้ทำการวิจัยเกี่ยวกับปัจจัยที่มีอิทธิพลต่อค่าความแม่นยำของการจำแนกข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล เปรียบเทียบระหว่างอัลกอริธึม Naïve Bayes Simples และ ID3 แนวทางการวิจัยได้เน้นให้เห็นถึงความสำคัญของขั้นตอนการเตรียมข้อมูลก่อนการสร้างโมเดล โดยทำการแบ่งข้อมูลชุดเรียนรู้รูปแบบโมเดลออกเป็น 2 ชุด เรียกว่า ชุด A และชุด B โดยมีความแตกต่างของข้อมูลตรงที่การจัดกลุ่มค่าข้อมูลในแต่ละแอตทริบิวท์ ซึ่งค่าของข้อมูลชุด B มีความละเอียดและมีความเป็นเอกลักษณ์ที่ชัดเจนกว่าชุด A จากนั้นทำการวิเคราะห์เปรียบเทียบผลการทดลองจากทั้ง 2 อัลกอริธึม พบว่า การจัดกลุ่มค่าข้อมูลที่ความละเอียดและมีความเป็นเอกลักษณ์ที่ชัดเจนกว่าจะทำให้การจำแนกข้อมูลมีความถูกต้องยิ่งขึ้น และพบว่าอัลกอริธึม Naïve Bayes Simple มีความสามารถในการจำแนกข้อมูลที่ดีกว่า ID3

อย่างไรก็ตาม ผู้วิจัยมีประเด็นที่พิจารณาแล้วเห็นว่า แม้ว่าอัลกอริธึม Naïve Bayes Simple จะมีประสิทธิภาพในการจำแนกหรือพยากรณ์ข้อมูลได้ดีกว่า ID3 จริง ในแง่ของการนำไปอธิบายต่อผู้ใช้งานคนอื่นๆ มีความยุ่งยากที่จะอธิบายให้บุคคลเหล่านั้นมองเห็นภาพรวมของรูปแบบการตัดสินใจที่ชัดเจนได้ ในขณะเดียวกันอัลกอริธึม ID3 เป็นอัลกอริธึมในกลุ่มแผนภาพต้นไม้ตัดสินใจ (Decision Tree) ซึ่งมีจุดเด่นด้านการอธิบายรูปแบบการตัดสินใจเป็นแบบแผนภาพที่เข้าใจง่าย แต่มีข้อเสีย คือ ในกรณีที่ข้อมูลชุดเรียนรู้มีรูปแบบการตัดสินใจแบบเดียวกัน และมี

คำตอบทั้งเห็นด้วยและปฏิเสธ อัลกอริธึม ID3 จะไม่พยากรณ์ความน่าจะเป็นของคำตอบของรูปแบบการตัดสินใจดังกล่าว

การไม่แสดงผลการพยากรณ์ของอัลกอริธึม ID3 นี้ ผู้วิจัยมีความเห็นว่าเป็นทั้งจุดเด่นและจุดด้อย ขึ้นอยู่กับการไปใช้กับงานแต่ละประเภท งานมีความละเอียดมากน้อยเพียงใด เช่น ถ้านำไปใช้กับงานวินิจฉัยโรค ก็จะถือว่าเป็นจุดเด่น เพราะบางครั้งต้องใช้ความรู้ความเชี่ยวชาญเฉพาะของแพทย์นั้นๆ มาประกอบการตัดสินใจ แต่สำหรับบางงานอาจไม่จำเป็นต้องใช้ความละเอียดในการพิจารณาที่ยุ่งยากซับซ้อนมาก หรือเพียงแค่ต้องการคำตอบในภาพรวม หรือเพื่อวัตถุประสงค์ที่จะนำไปเป็นแนวทางในการทำงานที่เกี่ยวข้องต่อไปเท่านั้น ซึ่งกรณีนี้อาจกล่าวได้ว่าเป็นจุดด้อยของอัลกอริธึม ID3

จากเหตุผลข้างต้น ผู้วิจัยจึงได้นำข้อดีและข้อเสียในแต่ละประเด็นที่จากการวิจัยครั้งก่อนมาพิจารณาปรับปรุงให้ดีขึ้น พร้อมกับทดลองหารูปแบบการตัดสินใจที่เหมาะสมที่สุดจากอัลกอริธึมอื่นๆ ดังที่จะแสดงไว้ในงานวิจัยครั้งนี้

ณัฐพร อาทรธรรมคุณ (2552, น. 43 - 62) ได้ทำการประยุกต์ใช้วิธีการจำแนกข้อมูล (Classification) ของเทคนิคการทำเหมืองข้อมูลโดยใช้เทคนิค Classifiers.Trees.J48 ของโปรแกรม WEKA ทำการประมวลผลวิเคราะห์หารูปแบบโมเดลจากข้อมูลผลการเรียนของนักเรียนเตรียมทหาร โรงเรียนเตรียมทหาร สถาบันวิชาการป้องกันประเทศ กองบัญชาการกองทัพไทย ที่จบการศึกษาตั้งแต่ปี พ.ศ. 2547-2549 จำนวน 4 รายวิชา ได้แก่ คณิตศาสตร์ ภาษาอังกฤษ ฟิสิกส์ และเคมี เพื่อนำผลลัพธ์จากโมเดลมาเป็นแนวทางสำหรับพยากรณ์ผลการเรียนของนักเรียนในชั้นปีปัจจุบัน

จากการศึกษาวิจัยนี้ ทำให้ผู้วิจัยมีความเชื่อมั่นว่าเทคนิค Classifiers.Trees.J48 น่าจะสามารถนำมาประยุกต์ใช้กับงานทบทวนวงเงินได้เช่นเดียวกัน

เสถียร วิชาเรือง (2553, น. 38 - 73) ได้ทำการประยุกต์ใช้วิธีการจำแนกข้อมูล (Classification) ของเทคนิคการทำเหมืองข้อมูลในกระบวนการอนุมัติสินเชื่อ โดย ได้ทำการวิจัยเชิงเปรียบเทียบประสิทธิภาพของอัลกอริธึมที่มีในโปรแกรม WEKA จำนวน 3 อัลกอริธึม ได้แก่ 1) เทคนิคการจำแนกกลุ่มแบบ Dicision Rules : Part หรือในโปรแกรม WEKA เรียกว่า Classifiers.rules.PART 2) เทคนิคการจำแนกกลุ่มแบบ Dicision Tree : C4.5 หรือในโปรแกรม WEKA เรียกว่า เทคนิค Classifiers.Trees.J48 และ 3) เทคนิคการจำแนกกลุ่มแบบ Dicision Table หรือในโปรแกรม WEKA เรียกว่าเทคนิค Classifiers.Decision Table โดยทำการประมวลผลข้อมูลเพื่อวิเคราะห์หารูปแบบโมเดลจากข้อมูลที่ใช้ประกอบการพิจารณาอนุมัติสินเชื่อของบริษัทเอกนครหลวงลิสซิ่ง จำกัด ตั้งแต่ปี พ.ศ. 2545 - 2550 จำนวน 17,964 สัญญา จากนั้นวิเคราะห์เปรียบเทียบ

ความถูกต้องของแต่ละโมเดลและเลือกรูปแบบที่เหมาะสมตรงกับความเป็นจริงมากที่สุด โดยผลวิจัยได้คัดเลือกโมเดลที่ได้จากเทคนิค Classifiers.rules.PART เนื่องมีความถูกต้องและมีเหตุผลสอดคล้องในทางธุรกิจมากที่สุด และนำรูปแบบเงื่อนไขดังกล่าวไปสร้างเป็นโปรแกรมจำแนกลูกค้าว่าเป็นลูกค้ากลุ่มดีหรือไม่ดี

จากการศึกษาวิจัยข้างต้น ผู้วิจัยมีความเห็นว่า รูปแบบการตัดสินใจที่ได้จากเทคนิค Classifiers.rules.PART เป็นลักษณะของการใช้คำเชื่อม ซึ่งเหมาะกับการนำไปเขียนเป็นโปรแกรม แต่ไม่เหมาะกับการที่จะนำมาใช้อธิบายแก่ผู้ใช้งาน เพราะอาจก่อให้เกิดความสับสนและเข้าใจผิดพลาดได้

สิทธิโชค มุกดาสกุลภิบาล (2551, น. 48 - 92) ได้ทำการวัดประสิทธิภาพของขั้นตอนวิธีตัวจำแนก C4.5, ADTree และ Naïve Bayes ในการจำแนกข้อมูลการซุกซ่อนสิ่งเสพติดสำหรับไปรษณีย์ระหว่างประเทศ ได้ผลสรุปว่า ตัวจำแนกแต่ละตัวมีข้อดีข้อเสียแตกต่างกัน ในแง่ความถูกต้องตัวจำแนก ADTree ให้ค่าความถูกต้องได้ดี อยู่ที่ช่วงร้อยละ 94 - 97 แต่ค่อนข้างใช้เวลาในการประมวลผลนานกว่าตัวจำแนกอื่นๆ ขณะที่ตัวจำแนก C4.5 และ Naïve Bayes ใช้เวลาในการประมวลผลใกล้เคียงกัน ตัวจำแนก C4.5 สามารถทำนายข้อมูลที่ไม่มีสิ่งเสพติดซุกซ่อนได้ดี อยู่ที่ช่วงร้อยละ 99.0 - 99.9 และตัวจำแนก Naïve Bayes สามารถทำนายข้อมูลที่มีสิ่งเสพติดซุกซ่อนได้ดี อยู่ที่ช่วงร้อยละ 21.0 - 84.4 และจะมากกว่าตัวจำแนกอื่นเมื่อเพิ่มขนาดของข้อมูลการเรียนรู้ที่สูงขึ้น

จากการศึกษาวิจัยนี้ ทำให้ผู้วิจัยมีความเชื่อมั่นว่าเทคนิคตัวจำแนก C4.5 หรือที่โปรแกรม WEKA เรียกว่า “Classifiers.Trees.J48” น่าจะมีประสิทธิภาพในการจำแนกข้อมูลที่ดีและน่าจะสามารถนำมาประยุกต์ใช้กับงานทบทวนวงเงินได้

ชุติมา อุตมะมุณีย์ และ ประสงค์ ปราณีตพลกรัง (2553) ได้ทำการวิจัยเชิงเปรียบเทียบประสิทธิภาพของอัลกอริธึม C4.5 และ Bayesian Networks พบว่า Bayesian Networks มีประสิทธิภาพในการจำแนกข้อมูลได้ดีกว่าอัลกอริธึม C4.5 จึงได้นำโมเดลการจำแนกข้อมูลของ Bayesian Networks ไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจแบบอัตโนมัติออนไลน์สำหรับการเลือกสาขาวิชาเรียนของนักศึกษาระดับอุดมศึกษา

จากการศึกษาวิจัยนี้ ทำให้ผู้วิจัยได้ทราบถึงจุดด้อยของการจำแนกข้อมูลด้วยอัลกอริธึม C4.5 ที่ไม่เหมาะกับงานบางประเภท

นโรดม อารีกุล และ ชุลีรัตน์ จรัสกุลชัย (2554) ได้ทำการวิจัยเชิงเปรียบเทียบประสิทธิภาพของอัลกอริธึม Decision trees (J48) และอัลกอริธึม Bayes (BayesNet) จากการประมวลผลข้อมูลเพื่อวิเคราะห์หารูปแบบโมเดลของโปรแกรม WEKA พบว่า อัลกอริธึม Decision

trees (J48) มีความสามารถในการจำแนกข้อมูลค้นหาผู้หลบหลีกภาษีได้ดีกว่า อัลกอริธึม Bayes และมีประสิทธิภาพในการพยากรณ์ถูกต้องเฉลี่ยร้อยละ 71.97

จากการศึกษาวิจัยนี้ ทำให้ผู้วิจัยมีความเชื่อมั่นว่าการวิเคราะห์รูปแบบโมเดลด้วยโปรแกรม WEKA โดยเลือกใช้เทคนิค Classifiers.Trees.J48 น่าจะเป็นเทคนิคที่เหมาะสมและควรนำมาประยุกต์ใช้กับงานทบทวนวงเงิน

DPU

บทที่ 3
ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นการประยุกต์ใช้โมเดลที่เกิดการเรียนรู้รูปแบบการตัดสินใจด้วยวิธีการจำแนกข้อมูล ของโปรแกรม WEKA เวอร์ชั่น 3.6.2 โดยทำการเรียนรู้รูปแบบการตัดสินใจของข้อมูลบัญชีเงินกู้เบิกเกินบัญชีที่เคยเข้าสู่กระบวนการทบทวนวงเงินล่วงหน้า ระหว่างเดือนมกราคม 2556 ถึงเดือนกันยายน 2557 จากนั้นได้พัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ที่มีลักษณะการทำงานแบบ Web Application โดยใช้ภาษา PHP และเรียกใช้รูปแบบโมเดลการตัดสินใจจากโปรแกรม WEKA โดยตรง และมีการประเมินผลว่า ผลการพยากรณ์ที่ได้นั้น มีความถูกต้องน่าเชื่อถือ และสอดคล้องกับความเป็นจริงทางธุรกิจหรือไม่ มากน้อยเพียงใด รายละเอียดเกี่ยวกับระเบียบวิธีวิจัย มีดังต่อไปนี้

3.1 ขั้นตอนการดำเนินงาน

1. กำหนดประเด็นปัญหาในการทำวิจัย
2. เก็บรวบรวมข้อมูลที่เกี่ยวข้อง
3. ศึกษาและทำความเข้าใจกับข้อมูล
4. ศึกษารูปแบบและแนวทางการทำวิจัย
5. ศึกษาหาความรู้เพิ่มเติมเกี่ยวกับเทคนิคการทำเหมืองข้อมูล
6. จัดเตรียมข้อมูลให้อยู่รูปแบบที่พร้อมใช้งาน
7. ประมวลผลข้อมูลและคัดเลือกรูปแบบโมเดลที่เหมาะสม
8. พัฒนาโปรแกรมพยากรณ์
9. ทดสอบโปรแกรมพยากรณ์กับการปฏิบัติงานจริง
10. ประเมินผลประสิทธิภาพของโปรแกรมพยากรณ์
11. สรุปผลการวิจัย
12. เก็บรวบรวมเนื้อหา นำเสนอและจัดพิมพ์เล่มงานวิจัย

3.2 อุปกรณ์และเครื่องมือที่ใช้ในการวิจัย

1. ฮาร์ดแวร์ ได้แก่ ชุดคอมพิวเตอร์ Notebook
   - 1.1 OS : Windows 8 (64-bit)
   - 1.2 Processor : Intel Core i5 CPU 1.80 GHz
   - 1.3 RAM : 4.00 GB
   - 1.4 Hard Disk : 500 GB
2. ซอฟต์แวร์ ได้แก่
   - 2.1 ชุดโปรแกรม Microsoft Office 2010 (Access, Excel, Word, Power Point)
   - 2.2 โปรแกรม WEKA เวอร์ชั่น 3.6.2
   - 2.3 ชุดโปรแกรม Appserv (Apache2.2, MySQL, PHP5)
   - 2.4 โปรแกรมอรรถประโยชน์อื่นๆ ได้แก่ EditPlus3, Sublime Text และ SnagIt

3.3 ระยะเวลาในการดำเนินการวิจัย

ตารางที่ 3.1 ระยะเวลาในการดำเนินการวิจัย ตั้งแต่เดือนตุลาคม 2556 – สิงหาคม 2557

ช่วงเวลา (เดือน) / ขั้นตอน	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.
1. กำหนดประเด็นปัญหาในการทำวิจัย	⟷										
2. เก็บรวบรวมข้อมูล		⟷	⟷	⟷	⟷	⟷	⟷	⟷	⟷		
3. ศึกษาและทำความเข้าใจกับข้อมูล		⟷	⟷	⟷	⟷	⟷	⟷	⟷			

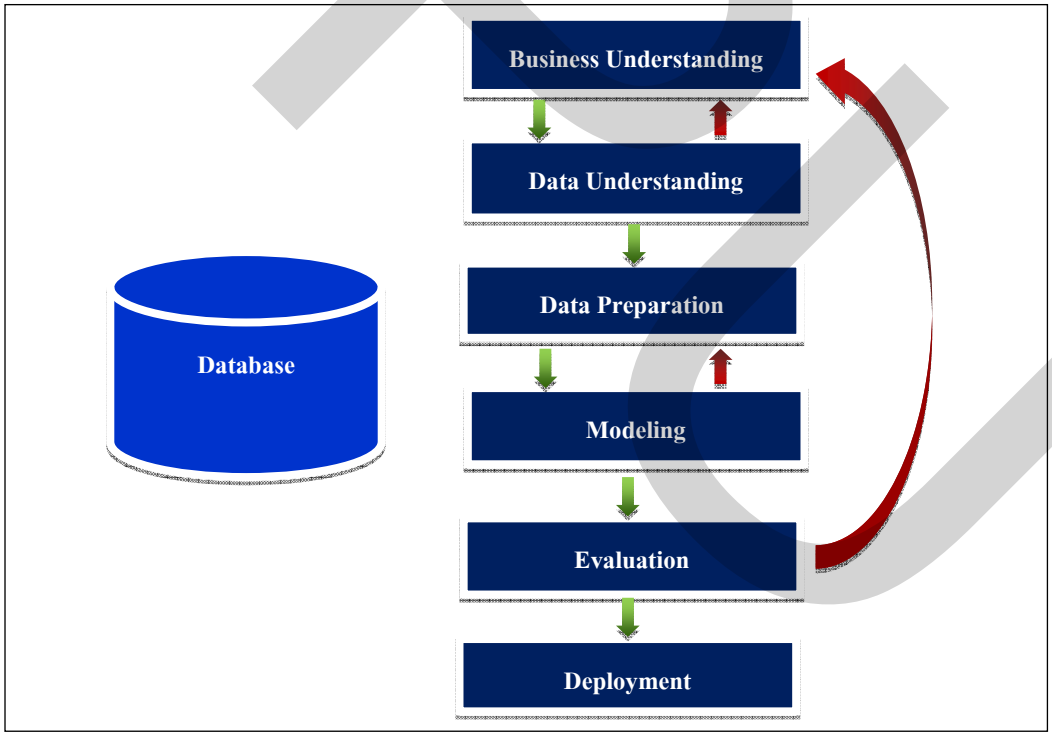
ตารางที่ 3.1 (ต่อ)

ช่วงเวลา (เดือน) / ขั้นตอน	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.
4. ศึกษารูปแบบและแนวทางการทำวิจัย		⟷	⟷	⟷	⟷	⟷	⟷	⟷	⟷		
5. ศึกษาหาความรู้เพิ่มเติม			⟷	⟷	⟷	⟷	⟷	⟷	⟷		
6. จัดเตรียมข้อมูล							⟷	⟷	⟷		
7. ประมวลผลและคัดเลือกโมเดล									⟷	⟷	
8. พัฒนาโปรแกรมพยากรณ์									⟷	⟷	
9. ทดสอบโปรแกรมพยากรณ์										⟷	⟷
10. ประเมินผลโปรแกรมพยากรณ์										⟷	⟷
11. สรุปผลการวิจัย											⟷

ตารางที่ 3.1 (ต่อ)

ช่วงเวลา (เดือน) / ขั้นตอน	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.
12. เก็บรวบรวมเนื้อหา นำเสนอ และจัดพิมพ์		←									→

3.4 การดำเนินการวิจัย



ภาพที่ 3.1 กระบวนการมาตรฐานในการวิเคราะห์ข้อมูล ตามมาตรฐาน CRISP-DM

งานวิจัยนี้ดำเนินการวิจัยตามขั้นตอนมาตรฐาน CRISP-DM ประกอบด้วย 6 ขั้นตอน ดังนี้

1. ขั้นตอนการทำความเข้าใจปัญหาทางธุรกิจ (Business Understanding)

เนื่องจากกระบวนการทำงานเกี่ยวกับงานทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชีเดิมอยู่ในความควบคุมดูแลของเจ้าหน้าที่ธนาคารเพียงหนึ่งคน ซึ่งทำให้เจ้าหน้าที่สินเชื่อท่านอื่นไม่ได้รับความสะดวกในการสอบถามข้อมูลว่าลูกค้าที่อยู่ในความดูแลนั้นอยู่ในเกณฑ์หรือมีโอกาสที่จะได้รับการทบทวนวงเงินหรือไม่ ดังนั้น เพื่ออำนวยความสะดวกแก่เจ้าหน้าที่สินเชื่อให้สามารถทำการสอบถามข้อมูลด้วยตนเองว่า บัญชีนั้นๆ จะได้รับการทบทวนวงเงินหรือไม่และมีความเป็นไปได้มากน้อยเพียงไร มีปัจจัยอะไรบ้างที่จะนำมาใช้ประกอบการพิจารณาทบทวนวงเงิน และมีหลักเกณฑ์เงื่อนไขอย่างไร ผู้วิจัยจึงมีแนวคิดที่จะนำความรู้ใหม่ที่ได้จากการทำเหมืองข้อมูลมาประยุกต์ใช้กับงานทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชีดังกล่าว และพัฒนาเป็นโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

2. ขั้นตอนการเก็บรวบรวม การตรวจสอบ และวิเคราะห์ข้อมูล (Data Understanding)

จากการทำความเข้าใจปัญหาทางธุรกิจข้างต้น ผู้วิจัยจึงได้ทำการศึกษาและเก็บรวบรวมข้อมูลบัญชีเงินกู้เบิกเกินบัญชีของลูกค้าที่อยู่ในความดูแลของฝ่ายพิจารณาสินเชื่อธุรกิจ SME ของธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) และเคยเข้าสู่กระบวนการทบทวนวงเงินล่วงหน้า ตั้งแต่เดือนมกราคม 2556 ถึงเดือนกันยายน 2557 จำนวนทั้งสิ้น 13,765 เร็คคอร์ด มาทำการวิเคราะห์และคัดเลือกแอตทริบิวท์ที่เป็นข้อมูลปัจจัยสำคัญๆ ที่ใช้ในการพิจารณาตามหลักเกณฑ์การทบทวนวงเงิน จำนวน 4 แอตทริบิวท์ และข้อมูลที่เป็นผลการพิจารณาว่าบัญชีนั้นๆ ผ่านเกณฑ์การพิจารณาให้ทบทวนวงเงินหรือไม่ จำนวน 1 แอตทริบิวท์ รวมเป็น 5 แอตทริบิวท์ ได้แก่ Utilization, LTV, Movement, Payment และ Class

3. ขั้นตอนการเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูล ผู้วิจัยได้ทำการแบ่งข้อมูลบัญชีเงินกู้เบิกเกินบัญชี จำนวน 13,765 เร็คคอร์ด ออกเป็น 2 ชุด โดยใช้เดือนของข้อมูลเป็นเกณฑ์ในการแบ่ง ดังนี้

3.1 ข้อมูลชุดแรกเป็นข้อมูลการทบทวนวงเงิน ตั้งแต่เดือนมกราคม 2556 ถึงเดือนสิงหาคม 2557 จำนวน 12,893 เร็คคอร์ด ชุดนี้ใช้สำหรับเรียนรู้และทดสอบโมเดลด้วยโปรแกรม WEKA

3.2 ข้อมูลชุดที่สองเป็นข้อมูลการทบทวนวงเงิน ของเดือนกันยายน 2557 จำนวน 872 เร็คคอร์ด ชุดนี้ใช้สำหรับประเมินประสิทธิภาพโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

การแปลงข้อมูล (Data Transformation) ตามทฤษฎีการทำเหมืองข้อมูลด้วยวิธีการจำแนกข้อมูลนั้น ต้องทำการแปลงค่าข้อมูลในที่เก็บอยู่ในรูปแบบตัวเลข ให้เป็นรูปแบบนาม (Nominal) ทั้งนี้ เพื่อลดการกระจายข้อมูลที่มากเกินไป วิธีการ คือ ให้แบ่งค่าข้อมูลออกเป็นช่วงหรือเป็นกลุ่ม (Cluster) การแบ่งช่วงข้อมูลที่เหมาะสม ถูกต้องและสอดคล้องกับความเป็นจริงของงาน จะมีผลต่อค่าความถูกต้องของการพยากรณ์ ซึ่งค่าความถูกต้องที่สูงหรือต่ำนี้ จะเป็นตัวบ่งบอกว่าผลลัพธ์จากการพยากรณ์นั้น มีความน่าเชื่อถือหรือไม่ ซึ่งจะเห็นผลชัดเจนมากถ้าใช้วิธีการสร้างรูปแบบโมเดลจากชุดข้อมูลเรียนรู้แบบมีครู (Supervised Learning)

แอตทริบิวท์ที่ไม่ต้องทำการแปลงข้อมูลให้อยู่ในรูป Nominal คือ แอตทริบิวท์ Class เนื่องจากข้อมูลถูกเก็บอยู่ในรูปแบบ Nominal เรียบร้อยแล้ว โดยมีค่าข้อมูลเป็น No และ Pass ส่วนที่ต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบ Nominal มีจำนวน 4 แอตทริบิวท์ ได้แก่ Utilization, LTV, Movement และ PaymentHistory รายละเอียดการแปลงข้อมูล ในตารางที่ 3.2

ตารางที่ 3.2 การแปลงข้อมูลของแต่ละแอตทริบิวท์ให้เป็น Nominal

Cluster No. / Attribute	1	2	3	4	5	6	7
LTV	<=75	>75-80	>80-100	>100	Deposit	CleanLoan	NoData
PaymentHistory	Bad	Good	NoData				
Utilization	<=80	>80-90	>90-95	>95-100	>100	NoData	
Movement	<10	10-14.99	15-19.99	>=20	NoData		

เงื่อนไขการแปลงข้อมูลของแอตทริบิวท์ LTV

```
--Query1
UPDATE IS_Src SET IS_Src.LTV_txt = ">100" WHERE (((IS_Src.LTV)>1));
--Query2
UPDATE IS_Src SET IS_Src.LTV_txt = ">80-100" WHERE (((IS_Src.LTV)>0.8 And
(IS_Src.LTV)<=1));
--Query3
UPDATE IS_Src SET IS_Src.LTV_txt = ">75-80" WHERE (((IS_Src.LTV)>0.75 And
(IS_Src.LTV)<=0.8));
```

--Query4

UPDATE IS_Src SET IS_Src.LTV_txt = "<=75" WHERE (((IS_Src.LTV)<=0.75));

--Query5

UPDATE IS_Src SET IS_Src.LTV_txt = "CleanLoan" WHERE (((IS_Src.LTV2)="Clean"));

--Query6

UPDATE IS_Src SET IS_Src.LTV_txt = "Deposit" WHERE (((IS_Src.LTV1)="เงินฝาก"));

--Query7

UPDATE IS_Src SET IS_Src.LTV_txt = "NoData" WHERE (((IS_Src.LTV2)="#N/A" Or (IS_Src.LTV2) Is Null));

เงื่อนไขการแปลงข้อมูลของแอตทริบิวท์ PaymentHistory

--Query8

UPDATE IS_Src SET IS_Src.History_24M_txt = "Good" WHERE (((IS_Src.History_24M)="Yes"));

--Query9

UPDATE IS_Src SET IS_Src.History_24M_txt = "Bad" WHERE (((IS_Src.History_24M)="No"));

--Query10

UPDATE IS_Src SET IS_Src.History_24M_txt = "NoData" WHERE (((IS_Src.History_24M) Is Null Or (IS_Src.History_24M) ="#N/A"));

เงื่อนไขการแปลงข้อมูลของแอตทริบิวท์ Utilization

--Query11

UPDATE IS_Src SET IS_Src.Utilization_txt = ">100" WHERE (((IS_Src.Utilization)>1));

--Query12

UPDATE IS_Src SET IS_Src.Utilization_txt = ">95-100" WHERE (((IS_Src.Utilization)>0.95 And (IS_Src.Utilization)<=1));

--Query13

UPDATE IS_Src SET IS_Src.Utilization_txt = ">90-95" WHERE (((IS_Src.Utilization)>0.9 And (IS_Src.Utilization)<=0.95));

```
--Query14
UPDATE IS_Src SET IS_Src.Utilization_txt = ">80-90" WHERE (((IS_Src.Utilization)>0.8 And
(IS_Src.Utilization)<=0.9));
--Query15
UPDATE IS_Src SET IS_Src.Utilization_txt = "<=80" WHERE (((IS_Src.Utilization)<=0.8));
--Query16
UPDATE IS_Src SET IS_Src.Utilization_txt = "NoData" WHERE (((IS_Src.Utilization) Is
Null));
```

เงื่อนไขการแปลงข้อมูลของแอตทริบิวท์ Movement

```
--Query17
UPDATE IS_Src SET IS_Src.Movement_txt = "<10" WHERE (((IS_Src.Movement)<0.1));
--Query18
UPDATE IS_Src SET IS_Src.Movement_txt = "10-14.99" WHERE (((IS_Src.Movement)>=0.1
And (IS_Src.Movement)<0.15));
--Query19
UPDATE IS_Src SET IS_Src.Movement_txt = "15-19.99" WHERE (((IS_Src.Movement)>=0.15
And (IS_Src.Movement)<0.2));
--Query20
UPDATE IS_Src SET IS_Src.Movement_txt = ">=20" WHERE (((IS_Src.Movement)>=0.2));
--Query21
UPDATE IS_Src SET IS_Src.Movement_txt = "NoData" WHERE (((IS_Src.Movement) Is
Null));
```

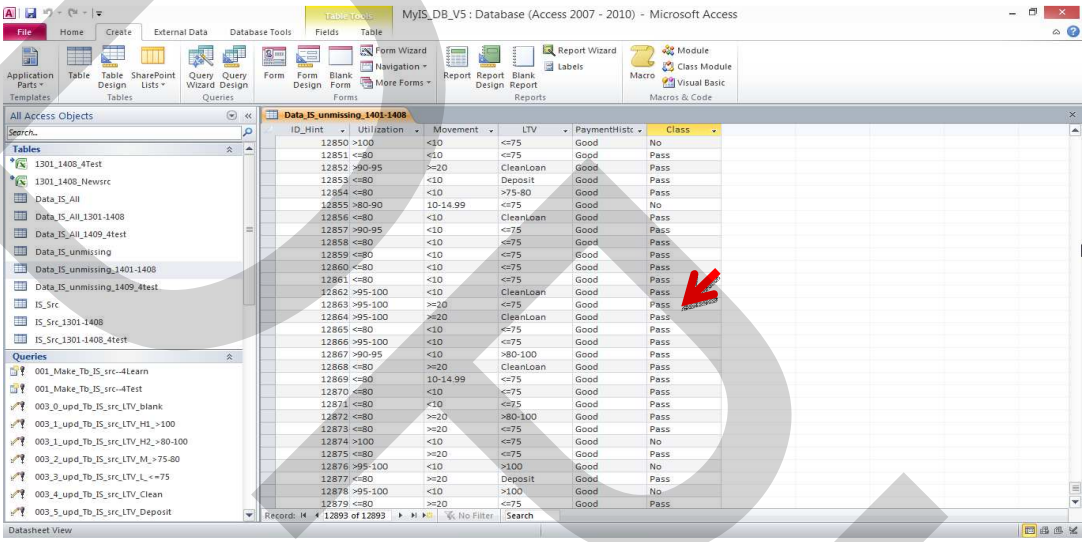
เงื่อนไขการเลือกข้อมูลเพื่อนำไปประมวลผลกับโปรแกรม WEKA

```
--Query22
SELECT IS_Src.ID AS ID_Hint, IS_Src.Utilization_txt AS Utilization, IS_Src.LTV_txt AS
LTV, IS_Src.Movement_txt AS Movement, IS_Src.History_24M_txt AS
PaymentHistory,IS_Src.Auto_Review_Class AS Class
FROM IS_Src
WHERE (((IS_Src.Utilization_txt)<>"NoData") AND ((IS_Src.LTV_txt)<>"NoData") AND
((IS_Src.Movement_txt)<>"NoData") AND ((IS_Src.History_24M_txt)<>"NoData"))
```

ORDER BY IS_Src.ID;

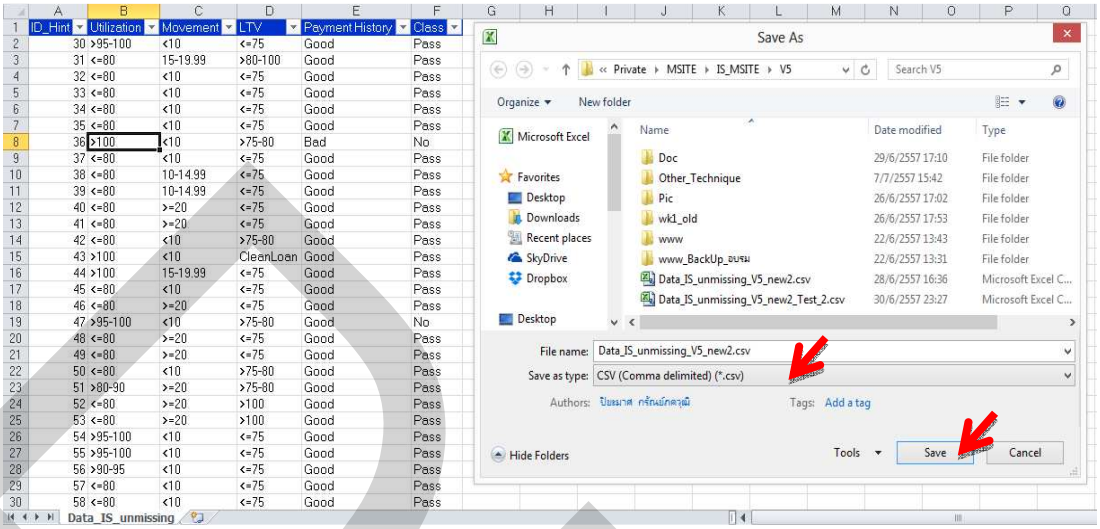
4. ขั้นตอนการวิเคราะห์และออกแบบ (Modeling)

4.1 นำข้อมูลที่ได้เก็บรวบรวมไว้ในขั้นตอนการจัดเตรียมข้อมูล มาเข้าสู่กระบวนการแปลงข้อมูลให้อยู่ในรูป Nominal โดยโปรแกรม Ms. Access ตามเงื่อนไขที่กำหนดไว้ในเงื่อนไขการแปลงข้อมูล ดังภาพที่ 3.2



ภาพที่ 3.2 การประมวลผลโดยโปรแกรม Ms. Access

4.2 Export ข้อมูลที่ได้ทำการแปลงข้อมูลจากโปรแกรม Ms. Access (.accdb) เรียบร้อยแล้ว ไปเป็นไฟล์ Excel (.xlsx) จากนั้นเปิดโปรแกรม Ms. Excel เปิดไฟล์ที่ Export มาจาก Ms. Access นั้น แล้ว Save as เป็นไฟล์ CSV (.csv) ในขั้นตอนนี้ผู้วิจัยเลือกเฉพาะเร็คคอร์ดที่มีค่าข้อมูลครบทุกแอตทริบิวท์เท่านั้น ทั้งนี้ เพื่อให้ค่าความถูกต้องของโมเดลการตัดสินใจมีความถูกต้องและมีความน่าเชื่อถือ ดังภาพที่ 3.3



ภาพที่ 3.3 การบันทึกไฟล์ Excel เป็นไฟล์ .csv

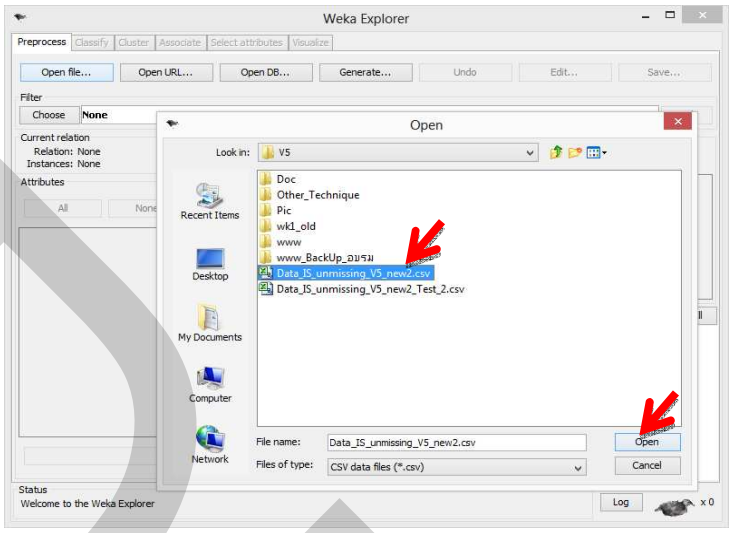
4.3 ทำการประมวลผลข้อมูลเพื่อหารูปแบบโมเดลการตัดสินใจตามกระบวนของการทำเหมืองข้อมูล โดยใช้โปรแกรม WEKA และมีลำดับขั้นตอนดังนี้

1. เปิดโปรแกรม WEKA > เลือก explorer > เลือก open file > เลือกไฟล์ .arff หรือ .csv ที่ต้องการใช้เป็นชุดข้อมูลเรียนรู้ ดังภาพที่ 3.4 และภาพที่ 3.5

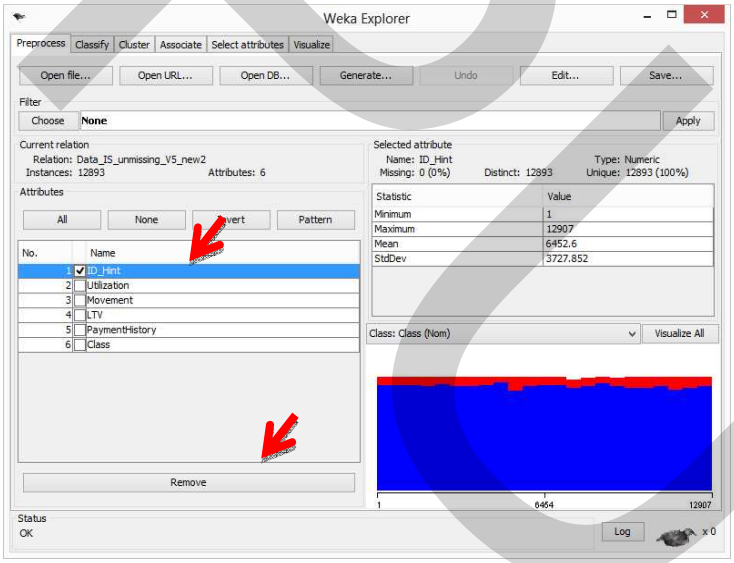
2. ลบแอตทริบิวท์ที่ไม่เกี่ยวข้องออก > คลิกเลือก > Remove ดังภาพที่ 3.6

3. เลือกแอตทริบิวท์ที่ต้องการ รวมถึงแอตทริบิวท์ที่จะใช้เป็นตัวจัดคลาสด้วย (กรณีใช้วิธีจำแนกข้อมูล) กรณีนำเข้าข้อมูลที่เป็นไฟล์ .csv ให้ save เป็นไฟล์ .arff ดังภาพที่ 3.7

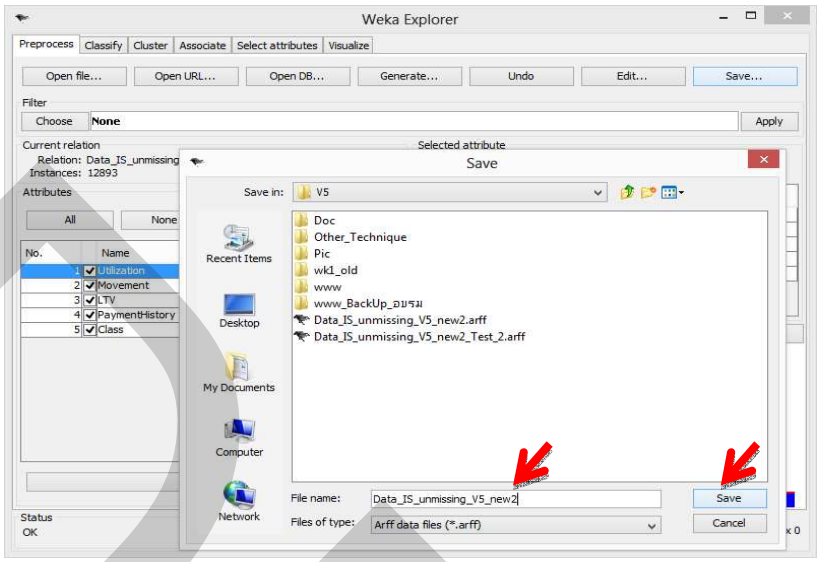
ภาพที่ 3.4 โปรแกรม WEKA เวอร์ชั่น 3.6.2



ภาพที่ 3.5 ตัวอย่างการนำเข้าข้อมูล .csv



ภาพที่ 3.6 ตัวอย่างการลบแอตทริบิวท์ที่ไม่เกี่ยวข้อง



ภาพที่ 3.7 ตัวอย่างการบันทึกไฟล์เป็น .arff

4.4 เลือกวิธีวิเคราะห์ข้อมูลด้วยวิธีจำแนกข้อมูล (Classify) งานวิจัยนี้ใช้เทคนิค J48 และ ID3 และเลือกใช้วิธีทดสอบประสิทธิภาพโมเดลกับทั้ง 2 เทคนิค ดังนี้

4.4.1 คลิก Tab classify > เลือกเทคนิคที่ต้องการใช้จำแนกข้อมูล ที่ Classifier ดังภาพที่ 3.8

4.4.2 เลือกวิธีการวัดประสิทธิภาพโมเดลที่กำลังจะถูกสร้างขึ้นที่ Test Options > เลือกวิธีวัดประสิทธิภาพโมเดลแบบ Cross-validation 10 Folds

ลักษณะการทำงานของ Test Options แต่ละแบบของโปรแกรม WEKA มี 4 รูปแบบ ดังนี้

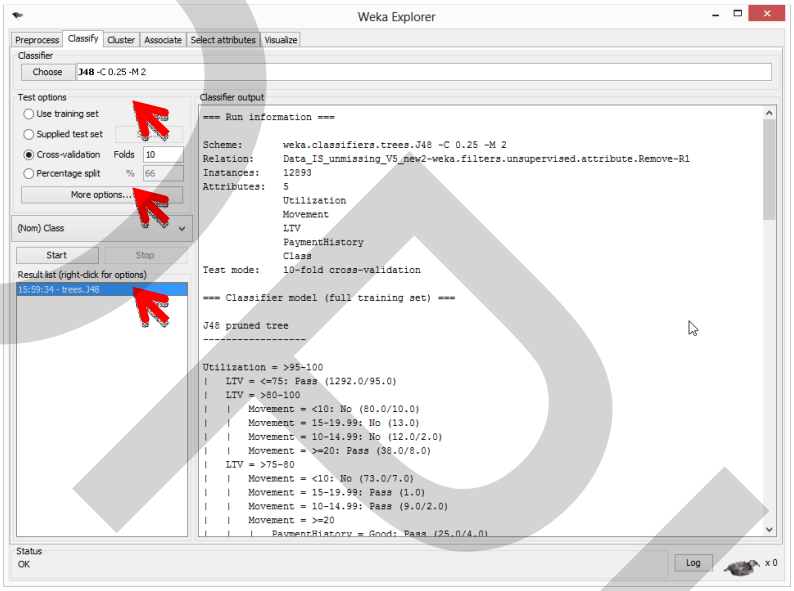
1. Use training set คือ เป็นการกำหนดให้โปรแกรม WEKA ใช้ชุดข้อมูลเรียนรู้นั้นเป็นตัวทดสอบโมเดลที่สร้างขึ้น

2. Supplied test set คือ เป็นการใช้ชุดข้อมูลทดสอบคนละชุดกับข้อมูลเรียนรู้ ซึ่งถูกจัดเตรียมไว้เรียบร้อยแล้ว และต้องเป็นไฟล์ .arff ที่มีโครงสร้างและคุณสมบัติของข้อมูลเหมือนกับของชุดข้อมูลเรียนรู้

3. Cross-validation คือ เป็นการใช้ความสามารถพิเศษที่มีอยู่ในโปรแกรม WEKA ให้ทำการแบ่งชุดข้อมูลเรียนรู้ออกเป็นชุดๆ ละเท่าๆ กัน ตามจำนวน Folds ที่กำหนดโดยทุกๆ รอบของการเรียนรู้รูปแบบข้อมูลเพื่อที่จะสร้างโมเดลนั้น โปรแกรม WEKA จะเหลือข้อมูลไว้ 1 ชุด เพื่อ

ใช้สำหรับทดสอบและวัดประสิทธิภาพของโมเดล และจะทำเช่นนี้ไปเรื่อยๆ จนกว่าจะครบทุกชุด ดังนั้น ค่าความน่าจะเป็นที่ได้จึงเป็นค่าความน่าจะเป็นเฉลี่ย

4. Percentage split คือ เป็นการใช้ความสามารถพิเศษที่มีอยู่ในโปรแกรม WEKA ให้จัดการแบ่งชุดข้อมูลเรียนรู้ออกเท่ากับ % split ที่ถูกกำหนดและส่วนที่เหลือนำไปใช้สำหรับทดสอบและวัดประสิทธิภาพของโมเดล



ภาพที่ 3.8 การเลือกวิธีการจำแนกข้อมูลและวิธีการใช้ข้อมูลทดสอบ

4.5 หากต้องการ Set Option เพิ่มเติม ให้เลือกที่ More Option

4.6 คลิก Start เพื่อรันโปรแกรม

4.7 ผลจากการประมวลผลข้อมูลจะได้รูปแบบโมเดลการตัดสินใจของเทคนิค J48 และ ID3 ดังที่แสดงผลรูปแบบการตัดสินใจอย่างย่อ ในภาพที่ 3.9 และภาพที่ 3.10

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Data_IS_unmissing_V5_new2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    12893
Attributes:   5
              Utilization
              Movement
              LTV
              PaymentHistory
              Class
Test mode:    10-fold cross-validation

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

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

Utilization = >95-100
|   LTV = <=75: Pass (1292.0/95.0)
|   LTV = >80-100
|   |   Movement = <10: No (80.0/10.0)
.
.
Utilization = <=80: Pass (8202.0/58.0)
Utilization = >100
|   LTV = <=75
|   |   Movement = <10: No (293.0/4.0)
|   |   Movement = 15-19.99: Pass (2.0/1.0)
.
.
Utilization = >80-90
|   LTV = <=75: Pass (858.0/16.0)
|   LTV = >80-100
|   |   Movement = <10: No (42.0/3.0)
.
.
Utilization = >90-95
|   LTV = <=75: Pass (656.0/33.0)
|   LTV = >80-100
|   |   Movement = <10: No (25.0/4.0)
.
.
|   LTV = Deposit: Pass (50.0/2.0)

Number of Leaves  :     55

Size of the tree :      72


Time taken to build model: 0.06 seconds
```

ภาพที่ 3.9 รูปแบบโมเดลการตัดสินใจจากการจำแนกข้อมูลโดยใช้เทคนิค J48

```
Scheme:       weka.classifiers.trees.Id3
Relation:     Data_IS_unmissing_V5_new2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    12893
Attributes:   5
              Utilization
              Movement
              LTV
              PaymentHistory
              Class
Test mode:    10-fold cross-validation

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

Id3

Utilization = >95-100
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
Utilization = <=80
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
Utilization = >100
| LTV = <=75
| | Movement = <10
| | | PaymentHistory = Good: No
| | | PaymentHistory = Bad: No
.
.
Utilization = >80-90
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
Utilization = >90-95
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
.
.
| | Movement = >=20: Pass

Time taken to build model: 0.03 seconds
```

ภาพที่ 3.10 รูปแบบโมเดลการตัดสินใจจากการจำแนกข้อมูลโดยใช้เทคนิค ID3

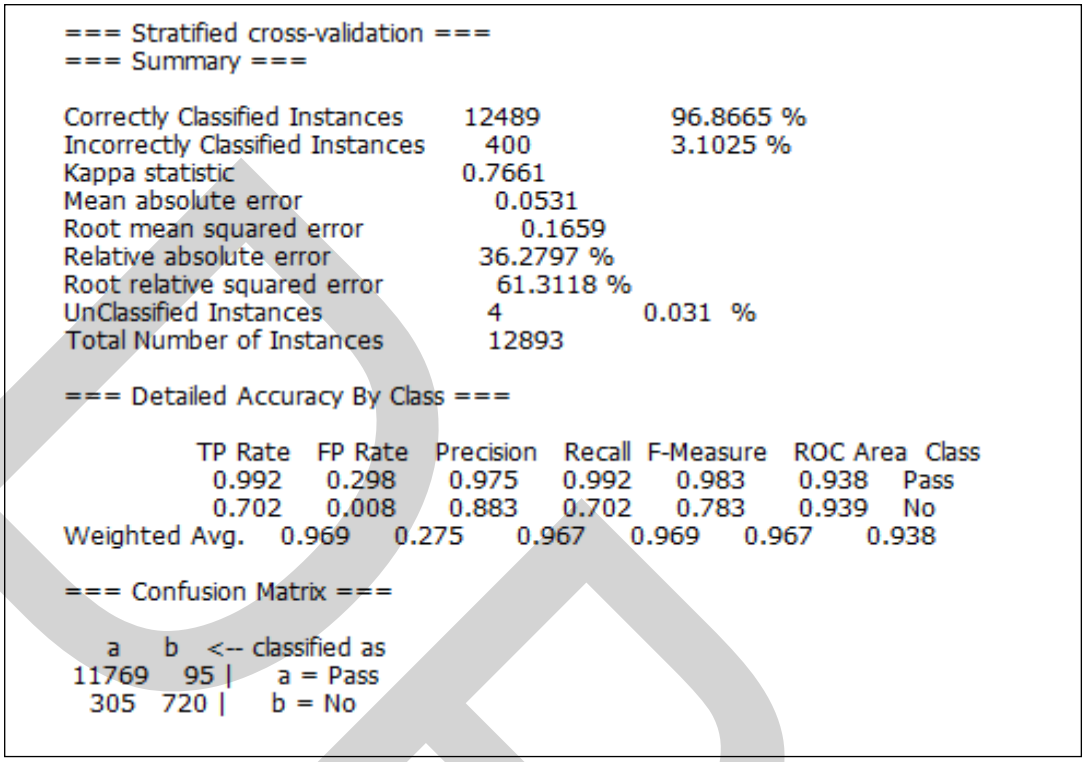
จากการวิเคราะห์ผลการประมวลผลข้อมูลของโปรแกรม WEKA ในขั้นตอนนี้พบว่าเทคนิค J48 มีลำดับชั้นของการจำแนกข้อมูล 72 ชั้น (layers) และได้รูปแบบการตัดสินใจ 55 รูปแบบ และพบว่าเทคนิค ID3 มีลำดับชั้นของการจำแนกข้อมูล 181 ชั้น (layers) และได้รูปแบบการตัดสินใจ 132 รูปแบบ นอกจากนี้ ยังพบว่าเทคนิค ID3 มีรูปแบบการตัดสินใจที่ไม่สามารถพยากรณ์ข้อมูลได้ จำนวน 4 รูปแบบ คือแสดงผลการพยากรณ์เป็น “null”

5. ขั้นตอนการวัดประสิทธิภาพของรูปแบบหรือโมเดล (Evaluation)

จากการวัดประสิทธิภาพของโมเดลการตัดสินใจที่ได้จากเทคนิค J48 และ ID3 ของโปรแกรม WEKA ได้ผลลัพธ์ ดังนี้

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

Correctly Classified Instances       12495               96.9131 %
Incorrectly Classified Instances       398                3.0869 %
Kappa statistic                          0.7673
Mean absolute error                      0.0545
Root mean squared error                  0.1671
Relative absolute error                 37.2147 %
Root relative squared error             61.7539 %
Total Number of Instances            12893

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.992    0.297      0.975     0.992     0.983      0.92     Pass
                 0.703    0.008      0.885     0.703     0.784      0.92     No
Weighted Avg.    0.969    0.274      0.968     0.969     0.968      0.92

=== Confusion Matrix ===

     a     b   <-- classified as
 11774    94 |     a = Pass
   304   721 |     b = No
```

ภาพที่ 3.11 ผลการประเมินประสิทธิภาพโมเดลการตัดสินใจของเทคนิค J48



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

Correctly Classified Instances       12489               96.8665 %
Incorrectly Classified Instances       400                3.1025 %
Kappa statistic                          0.7661
Mean absolute error                      0.0531
Root mean squared error                  0.1659
Relative absolute error                 36.2797 %
Root relative squared error             61.3118 %
UnClassified Instances                   4                0.031  %
Total Number of Instances            12893

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.992    0.298      0.975     0.992     0.983      0.938    Pass
                 0.702    0.008      0.883     0.702     0.783      0.939    No
Weighted Avg.    0.969    0.275      0.967     0.969     0.967      0.938

=== Confusion Matrix ===

     a     b   <-- classified as
 11769    95 |     a = Pass
   305   720 |     b = No
```

ภาพที่ 3.12 ผลการประเมินประสิทธิภาพโมเดลการตัดสินใจของเทคนิค ID3

6. ขั้นตอนการนำผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลไปใช้ให้เกิดประโยชน์ต่อองค์กร (Deployment)

จากการวิเคราะห์เปรียบเทียบรูปแบบโมเดลและค่าดัชนีที่ใช้วัดประสิทธิภาพของรูปแบบโมเดลการตัดสินใจที่ได้จากทั้งเทคนิค J48 และ ID3 ผลปรากฏว่า เทคนิค J48 มีประสิทธิภาพในการจำแนกหรือพยากรณ์ข้อมูลได้ดีกว่าเทคนิค ID3 ผู้วิจัยจึงได้เลือกใช้โมเดลการตัดสินใจที่ได้จากเทคนิค J48 ไปใช้กับโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ที่มีลักษณะการใช้งานแบบ Web Application โดยใช้ภาษา PHP และเรียกใช้รูปแบบโมเดลการตัดสินใจจากโปรแกรม WEKA เวอร์ชั่น 3.6.2 โดยตรง ดังรายละเอียดเพิ่มเติมในบทที่ 4

บทที่ 4
ผลการดำเนินการวิจัย

4.1 การวิเคราะห์และสรุปผลการดำเนินการวิจัย

จากการประมวลผลข้อมูลด้วยโปรแกรม WEKA เวอร์ชั่น 3.6.2 สามารถสรุปผลการประเมินประสิทธิภาพโมเดลการตัดสินใจของทั้ง 2 เทคนิค ได้แก่ เทคนิค J48 และเทคนิค ID3 ได้ดังรายละเอียดในตารางต่อไปนี้

ตารางที่ 4.1 เปรียบเทียบความสามารถในการจำแนกข้อมูล หรือการพยากรณ์ของโมเดล

Technique / Measurement	J48	ID3
Correctly Classified (No. Instances, (%))	12,495 (96.9131%)	12,489 (96.8665%)
Incorrectly Classified (No. Instances, (%))	398 (3.0869%)	400 (3.1025)
Unclassified Instances	-	4
Total Instances	**12,893**	**12,893**
Root Mean Squared Error	0.1671	0.1659

หมายเหตุ. คำว่า “อินสแตนท์ (Instance)” ในโปรแกรม WEKA หมายถึง เร็คคอร์ด (Record) หรือ แถว (Row) และคำว่า “แอตทริบิวท์ (Attribute)” ในโปรแกรม WEKA หมายถึง คอลัมน์ (Column) หรือฟิลด์ (Field)

จากตารางที่ 4.1 ผลการเปรียบเทียบความสามารถในการจำแนกข้อมูลของโมเดลที่ได้จากเทคนิค J48 และ ID3 สรุปผลได้ดังนี้

1. ค่าความถูกต้องแม่นยำในการจำแนกข้อมูล (Correctly Classified) ของเทคนิค J48 และ ID3 มีค่าเท่ากับ 96.91% และ 96.86% ตามลำดับ

2. ค่าความผิดพลาดในการจำแนกข้อมูล (Incorrectly Classified) ของเทคนิค J48 และ ID3 มีค่าเท่ากับ 3.09% และ 3.10% ตามลำดับ

3. จำนวนเร็คคอร์ด ที่ไม่สามารถจำแนกข้อมูลได้ (Unclassified Instances) ของเทคนิค J48 สามารถจำแนกข้อมูลได้ทุกเร็คคอร์ด และ ID3 มีจำนวน 4 เร็คคอร์ด

4. จำนวนเร็คคอร์ด ที่ใช้หรับการเรียนรู้และสร้างรูปแบบโมเดลนี้มีจำนวน 12,893 เร็คคอร์ด

5. ค่าดัชนีความคลาดเคลื่อนในการจำแนกข้อมูลระหว่างค่าจริงและค่าพยากรณ์ (Root Mean Squared Error) ของเทคนิค J48 และ ID3 มีค่าเท่ากับ 0.1671 และ 0.1659 ตามลำดับ

ตารางที่ 4.2 สูตรคำนวณค่า TP, FP และ Precision

True Positive Rate (TP) or Recall Rate		**False Positive Rate (FP)**		**Precision**	
Pass	No	Pass	No	Pass	No
a/(a+b)	d/(c+d)	c/(c+d)	b/(a+d)	a/(a+c)	d/(b+d)

ตารางที่ 4.3 การเปรียบเทียบจำนวนข้อมูลจำแนกตามคลาสจริง กับผลการพยากรณ์ของโมเดล

Technique & Classified as / Fact Value	**J48**		**ID3**	
	Pass	**No**	**Pass**	**No**
Pass (11,868)	11,774 (a)	94 (b)	11,769 (a)	95 (b)
No (1,025)	304 (c)	721 (d)	305 (c)	720 (b)

จากตารางที่ 4.3 แสดงการเปรียบเทียบจำนวนข้อมูลจำแนกตามคลาสจริง กับผลการ

พยากรณ์ของโมเดลที่ได้จากการใช้เทคนิค J48 และ ID3 สรุปผลได้ดังนี้

1. จำนวนเร็คคอร์ดที่มีคลาสจริงมีค่าเป็น Pass จำนวน 11,868 นั้น พบว่าผลการพยากรณ์ของโมเดลที่ได้จากการใช้เทคนิค J48 ที่พยากรณ์เป็น Pass จำนวน 11,774 เร็คคอร์ด และพยากรณ์เป็น No จำนวน 94 เร็คคอร์ด และพบว่าผลการพยากรณ์ของเทคนิค ID3 ที่พยากรณ์เป็น Pass จำนวน 11,769 เร็คคอร์ด และพยากรณ์เป็น No จำนวน 95 เร็คคอร์ด แสดงว่าโมเดลที่ได้จากเทคนิค ID3 ไม่สามารถพยากรณ์ข้อมูลได้ จำนวน 4 เร็คคอร์ด

2. จำนวนเร็คคอร์ดที่มีคลาสจริงมีค่าเป็น No จำนวน 1,025 เร็คคอร์ด พบว่ามีผลการพยากรณ์ของโมเดลที่ได้จากการใช้เทคนิค J48 ที่พยากรณ์เป็น Pass มีจำนวน 304 เร็คคอร์ด และพยากรณ์เป็น No มีจำนวน 721 เร็คคอร์ด และพบว่าผลการพยากรณ์ของเทคนิค ID3 ที่พยากรณ์เป็น Pass จำนวน 305 เร็คคอร์ด และพยากรณ์เป็น No จำนวน 720 เร็คคอร์ด

ตารางที่ 4.4 เปรียบเทียบค่าดัชนีชี้วัดประสิทธิภาพในการจำแนกข้อมูล หรือการพยากรณ์ของโมเดล

Class / Measure	J48			ID3		
	Pass	No	Average	Pass	No	Average
True Positive Rate	0.992	0.703	0.969	0.992	0.702	0.969
False Positive Rate	0.297	0.008	0.274	0.298	0.008	0.275
Precision	0.975	0.885	0.968	0.975	0.883	0.967

จากตารางที่ 4.4 แสดงการเปรียบเทียบค่าดัชนีชี้วัดประสิทธิภาพในการจำแนกข้อมูล หรือการพยากรณ์ของโมเดล ซึ่งสามารถคำนวณได้จากการนำข้อมูลในตารางที่ 4.3 ซึ่งผู้วิจัยนำข้อมูลมาจากค่าที่แสดงอยู่ในส่วนของ Confusion Matrix ของโปรแกรม WEKA จากนั้นนำมาเข้าสูตรคำนวณตามสูตรในตารางที่ 4.2 และเมื่อพิจารณาค่าดัชนีชี้วัดประสิทธิภาพโมเดลจากค่าเฉลี่ย (Average) สามารถสรุปผลได้ดังนี้

1. ค่า True Positive Rate (Average) หรือค่าที่ใช้วัดผลการพยากรณ์ตามรูปแบบโมเดลที่ถูกค้นพบ แล้วให้คำตอบตรงกับข้อมูลจริง หรือเรียกว่าค่า “Recall” พบว่าทั้งเทคนิค J48 และ ID3 มีค่าเท่ากันคือ 0.969

2. ค่า False Positive Rate (Average) หรือค่าที่ใช้วัดผลการพยากรณ์ตามรูปแบบโมเดลที่ถูกค้นพบ แล้วให้คำตอบเป็นเท็จ หมายถึง การที่ข้อมูลจริงไม่ได้อยู่ในคลาสนั้นแต่คำพยากรณ์จากโมเดลบอกว่าอยู่ในคลาสนั้น พบว่าของเทคนิค J48 และ ID3 มีค่าเท่ากับ 0.274 และ 0.275 ตามลำดับ

3. ค่า Precision (Average) หรือค่าที่ใช้วัดความเชื่อมั่นของผลการพยากรณ์ พบว่าของเทคนิค J48 และ ID3 มีค่าเท่ากับ 0.968 และ 0.967 ตามลำดับ

จากการการเปรียบเทียบข้อดีข้อเสียของทั้ง 2 เทคนิค พบว่ารูปแบบโมเดลการตัดสินใจที่ได้จากเทคนิค J48 มีข้อดีมากกว่าเทคนิค ID3 ทั้งในแง่ของลำดับชั้นการตัดสินใจที่ไม่เยอะจนเกินไป ดังที่แสดงไว้ในบทที่ 3 และมีค่าดัชนีที่ใช้วัดประสิทธิภาพของรูปแบบโมเดลการตัดสินใจ ที่แสดงถึงความมีประสิทธิภาพในการจำแนกข้อมูลหรือสามารถพยากรณ์ข้อมูลของเทคนิค J48 นั้น ดีกว่าเทคนิค ID3

4.2 การประยุกต์ใช้โมเดลการตัดสินใจ

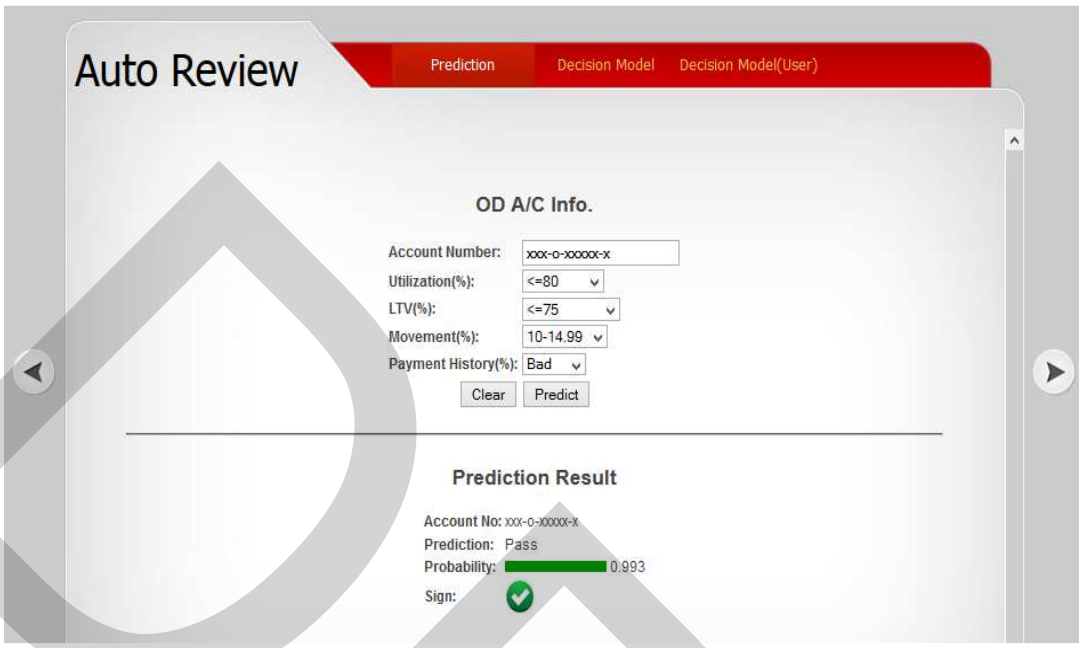
จากการวิเคราะห์และสรุปผลการประเมินประสิทธิภาพโมเดลการตัดสินใจข้างต้น ผู้วิจัยจึงได้พัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ที่มีลักษะการใช้งานแบบ Web Application โดยใช้ภาษา PHP และกำหนดให้การพยากรณ์ข้อมูลนั้นเรียกใช้รูปแบบโมเดลการตัดสินใจที่ได้จากเทคนิค J48 จากโปรแกรม WEKA โดยตรง

ส่วนประกอบของโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ประกอบด้วย 3 ส่วน ดังนี้

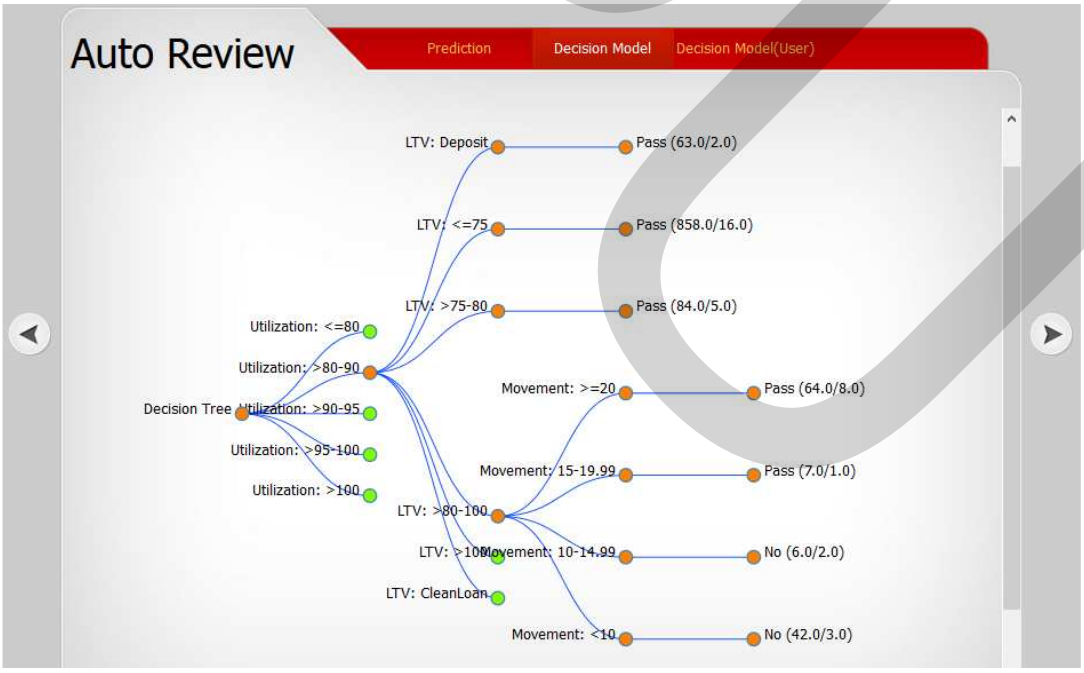
1. หน้าจอหลัก ใช้สำหรับการพยากรณ์

2. หน้าจอแสดง Decision Tree ใช้สำหรับผู้จัดทำข้อมูล เพื่อทำความเข้าใจรูปแบบการตัดสินใจของตัวโปรแกรมพยากรณ์ โดยตัวเลขที่อยู่ในวงเล็บใช้สำหรับอธิบายข้อมูล หรือเป็นการแจกแจงให้ทราบเกี่ยวกับข้อมูลทางเทคนิคสำหรับผู้วิเคราะห์ข้อมูล ตัวเลขด้านหน้า คือ จำนวนเร็คคอร์ดของข้อมูลชุดเรียนรู้ที่มีเงื่อนรูปแบบการตัดสินใจแบบนั้นๆ ว่ามีกี่เร็คคอร์ด และตัวเลขด้านหลัง คือ จำนวนเร็คคอร์ดที่ข้อมูล Class ของผลการการพิจารณาทบทวนวงเงินจริงขัดแย้งกับผลการพยากรณ์ของโมเดล

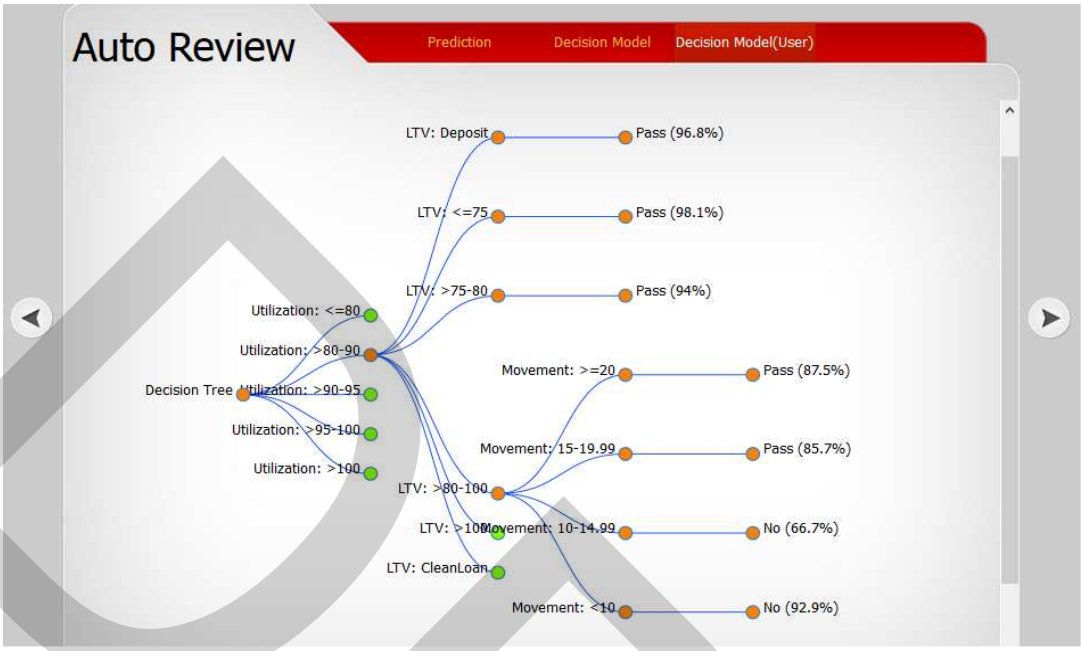
3. หน้าจอแสดง Decision Tree ใช้สำหรับผู้ใช้ทั่วไป เพื่อทำความเข้าใจรูปแบบการตัดสินใจของตัวโปรแกรมพยากรณ์ ทั้งนี้ เพื่อความเข้าใจง่าย ผู้วิจัยจึงได้แปลงข้อมูลทางเทคนิคในข้อ 2. ให้แสดงผลเป็นเปอร์เซนต์ (%) ของความน่าจะเป็นตามผลการพยากรณ์



ภาพที่ 4.1 หน้าจอแสดงผลที่ใช้สำหรับพยากรณ์



ภาพที่ 4.2 หน้าจอแสดงผล Decision Tree สำหรับผู้จัดทำข้อมูล



ภาพที่ 4.3 หน้าจอแสดงผล Decision Tree สำหรับผู้ใช้ทั่วไป

4.3 ผลการเปรียบเทียบผลลัพธ์จากโปรแกรมพยากรณ์ กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง

จากการนำข้อมูลทดสอบชุดที่ 2 จำนวน 872 เร็คคอร์ด มาทำการทดสอบเพื่อประเมินประสิทธิภาพของโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงินแล้วเปรียบเทียบผลลัพธ์จากโปรแกรมพยากรณ์กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง โดยสามารถแสดงผลการเปรียบเทียบ จำแนกตามรูปแบบโมเดลการตัดสินใจของเทคนิค J48 ทั้ง 55 รูปแบบ ดังนี้

ตารางที่ 4.5 เปรียบเทียบผลลัพธ์จากโปรแกรมพยากรณ์กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง

Seq.	Utilization	LTV	Movement	Payment	Predict. Result	Prob.	% Prob.	No. Test Case	TRUE	FALSE	% False/Total Test Case
1	<=80				Pass	0.993	99.3	520	517	3	1%
2	>80-90	Deposit			Pass	0.968	96.8	6	6	0	0%
3	>80-90	<=75			Pass	0.981	98.1	49	48	1	2%
4	>80-90	>75-80			Pass	0.94	94	2	2	0	0%
5	>80-90	>80-100	>=20		Pass	0.875	87.5	6	6	0	0%
6	>80-90	>80-100	15-19.99		Pass	0.857	85.7	1	1	0	0%
7	>80-90	>80-100	10-14.99		No	0.667	66.7	1	0	1	100%
8	>80-90	>80-100	<10		No	0.929	92.9	1	1	0	0%
9	>80-90	>100	>=20		Pass	0.871	87.1	6	6	0	0%
10	>80-90	>100	15-19.99		Pass	1	100				
11	>80-90	>100	10-14.99		Pass	0.889	88.9				
12	>80-90	>100	<10		No	0.682	68.2	1	0	1	100%
13	>80-90	CleanLoan			Pass	0.895	89.5	10	8	2	20%
14	>90-95	Deposit			Pass	0.96	96	2	2	0	0%
15	>90-95	<=75			Pass	0.95	95	43	43	0	0%
16	>90-95	>75-80	>=20		Pass	0.946	94.6	1	1	0	0%
17	>90-95	>75-80	15-19.99		Pass	1	100	1	1	0	0%
18	>90-95	>75-80	10-14.99		Pass	1	100	1	1	0	0%
19	>90-95	>75-80	<10		No	0.963	96.3	1	1	0	0%
20	>90-95	>80-100	>=20		Pass	0.8	80	2	1	1	50%
21	>90-95	>80-100	15-19.99		Pass	0.8	80				
22	>90-95	>80-100	10-14.99		No	1	100	1	1	0	0%
23	>90-95	>80-100	<10		No	0.84	84	2	2	0	0%
24	>90-95	>100	>=20		Pass	0.905	90.5	3	3	0	0%
25	>90-95	>100	15-19.99		No	0.531	53.1				
26	>90-95	>100	10-14.99		No	1	100				
27	>90-95	>100	<10		No	0.833	83.3	3	3	0	0%
28	>90-95	CleanLoan			Pass	0.938	93.8	7	7	0	0%
29	>95-100	Deposit			Pass	0.984	98.4	11	11	0	0%
30	>95-100	<=75			Pass	0.926	92.6	77	71	6	8%
31	>95-100	>75-80	>=20	Good	Pass	0.84	84				
32	>95-100	>75-80	>=20	Bad	No	1	100				
33	>95-100	>75-80	15-19.99		Pass	1	100				
34	>95-100	>75-80	10-14.99		Pass	0.778	77.8	1	0	1	100%
35	>95-100	>75-80	<10		No	0.904	90.4	7	7	0	0%
36	>95-100	>80-100	>=20		Pass	0.789	78.9				
37	>95-100	>80-100	15-19.99		No	1	100				
38	>95-100	>80-100	10-14.99		No	0.833	83.3	3	3	0	0%
39	>95-100	>80-100	<10		No	0.875	87.5	10	8	2	20%

ตารางที่ 4.5 (ต่อ)

Seq.	Utilization	LTV	Movement	Payment	Predict. Result	Prob.	% Prob.	No. Test Case	TRUE	FALSE	% False/Total Test Case
40	>95-100	>100	>=20		Pass	0.867	86.7	1	0	1	100%
41	>95-100	>100	15-19.99		No	0.857	85.7				
42	>95-100	>100	10-14.99		No	0.8	80				
43	>95-100	>100	<10		No	0.663	66.3	9	7	2	22%
44	>95-100	CleanLoan			Pass	0.91	91	38	37	1	3%
45	>100	Deposit			Pass	0.955	95.5	2	2	0	0%
46	>100	<=75	>=20	Good	Pass	0.75	75	1	0	1	100%
47	>100	<=75	>=20	Bad	No	1	100	1	1	0	0%
48	>100	<=75	15-19.99		Pass	0.5	50				
49	>100	<=75	10-14.99	Good	Pass	0.667	66.7	1	1	0	0%
50	>100	<=75	10-14.99	Bad	No	1	100				
51	>100	<=75	<10		No	0.986	98.6	23	22	1	4%
52	>100	>75-80			No	0.917	91.7	6	5	1	17%
53	>100	>80-100			No	0.9	90	5	3	2	40%
54	>100	>100			No	0.682	68.2	2	1	1	50%
55	>100	CleanLoan			Pass	0.971	97.1	4	4	0	0%
Total								**872**	**844**	**28**	**3.21%**

จากตารางที่ 4.5 ผลการทดสอบโปรแกรมพยากรณ์กับข้อมูลจริง จำนวน 872 เร็คคอร์ด สามารถสรุปผลได้ดังนี้

1. ผลลัพธ์จากโปรแกรมพยากรณ์ กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริงมีความถูกต้องตรงกัน (Correctly Classified) จำนวน 844 เร็คคอร์ด คิดเป็นร้อยละ 96.79 หรือเรียกว่าค่า True Positive Rate หรือค่า Recall มีค่าเท่ากับ 0.9679

2. ผลลัพธ์จากโปรแกรมพยากรณ์ไม่ตรง กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง (Incorrectly Classified) จำนวน 28 เร็คคอร์ด มีค่าเท่ากับ 0.0321 หรือคิดเป็นร้อยละ 3.21

3. รูปแบบโมเดลการติดสินใจที่ไม่ได้ทดสอบ เนื่องจากข้อมูลในชุดสอบไม่มีรูปแบบการตัดสินใจที่ตรงกับรูปแบบการตัดสินใจของโมเดล จำนวน 13 รูปแบบ ได้แก่ รูปแบบลำดับที่ 10, 11, 21, 25, 26, 31, 32, 33, 36, 37, 41, 42, 48 และ 50

4. รูปแบบโมเดลการติดสินใจที่ผลการพยากรณ์ผิดพลาด จากผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง ตั้งแต่ร้อยละ 20 ขึ้นไป มีจำนวน 11 รูปแบบ ได้แก่ รูปแบบลำดับที่ 7, 12, 13, 20, 34, 39, 40, 43, 46, 53 และ 54

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

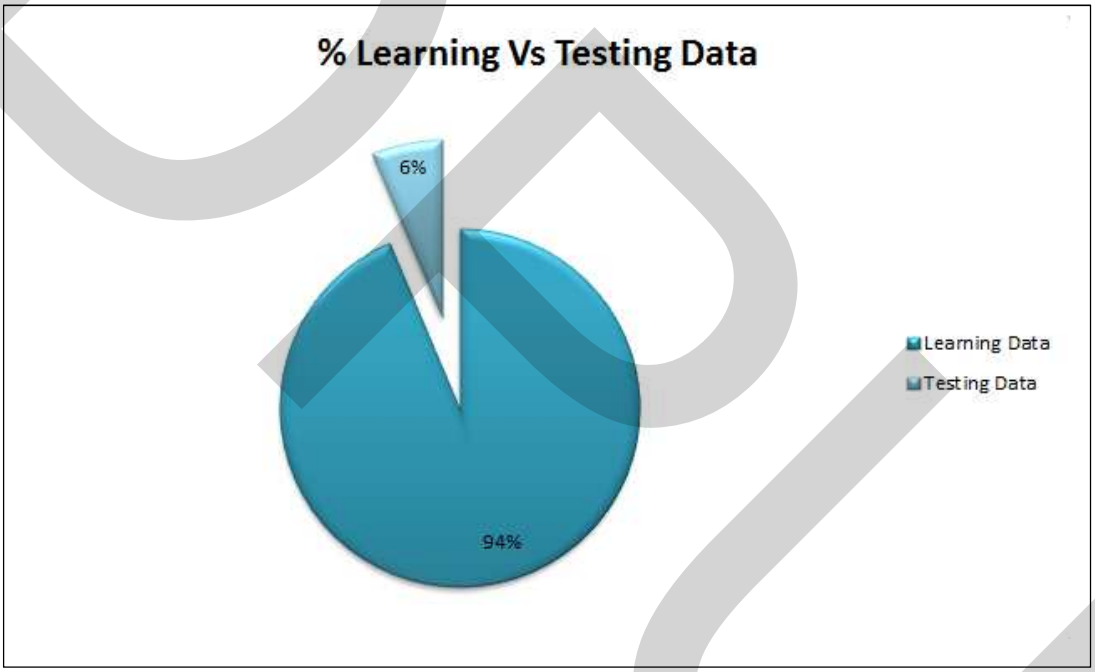
5.1 สรุปผลการวิจัย

จากปัญหาการขาดการเชื่อมโยงของการปฏิบัติงานเกี่ยวกับงานทบทวนวงเงิน ระหว่างผู้ปฏิบัติงาน 2 ฝ่าย คือ เจ้าหน้าที่สินเชื่อ ซึ่งมีบทบาทหน้าที่ในปฏิบัติงานเป็นด่านหน้าของหน่วยธุรกิจของธนาคาร คือ มีหน้าที่รับผิดชอบในการให้บริการและติดตามดูแลลูกค้าอย่างใกล้ชิด กับอีกฝ่ายหนึ่ง คือ เจ้าหน้าที่ทบทวนวงเงิน ซึ่งมีบทบาทหน้าที่เป็นผู้ปฏิบัติงานเบื้องหลัง คือ เมื่อลูกค้าได้รับอนุมัติสินเชื่อเงินกู้เบิกเกินบัญชีเรียบร้อยแล้ว เจ้าหน้าที่ทบทวนวงเงิน มีหน้าที่รับผิดชอบในการพิจารณาทบทวนวงเงินของบัญชีเงินกู้เบิกเกินบัญชีที่ใกล้ถึงวันครบกำหนดอายุสัญญา หากพิจารณาแล้วเห็นว่าเป็นลูกค้าดีและอยู่ในเกณฑ์ที่กำหนด ก็จะทำการอนุมัติให้ขยายวันครบกำหนดอายุสัญญาออกไป ซึ่งลูกค้าจะสามารถใช้วงเงินนั้นต่อไปได้ โดยไม่จำเป็นต้องยื่นขอสินเชื่อใหม่อีกครั้ง

การแบ่งหน้าที่ความรับผิดชอบที่แยกขาดจากกันเช่นนี้ ทำให้เกิดอุปสรรคในการทำงานระหว่างเจ้าหน้าที่สินเชื่อและเจ้าหน้าที่ทบทวนวงเงิน ดังนั้น เพื่อเป็นการแก้ไขปัญหาที่เกิดขึ้น ผู้วิจัยจึงแนวคิดที่จะนำความรู้ด้านการทำเหมืองข้อมูลมาประยุกต์ใช้กับงานทบทวนวงเงิน พร้อมกับทำการพัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ที่มีลักษณะการทำงานแบบ Web Application โดยใช้ภาษา PHP และกำหนดให้เรียกใช้รูปแบบโมเดลการตัดสินใจโดยตรงจากโปรแกรม WEKA พร้อมกับมีการประเมินผลเชิงปริมาณจากการทดสอบข้อมูลผลการพิจารณาทบทวนวงเงินจริง มาเปรียบเทียบกับผลการพยากรณ์ที่ได้จากโปรแกรมพยากรณ์ฯ ว่ามีความถูกต้องน่าเชื่อถือ และสอดคล้องกับความเป็นจริงทางธุรกิจหรือไม่ เพียงใด และมีการประเมินผลเชิงคุณภาพจากความคิดเห็นของผู้ใช้โปรแกรมฯ ซึ่งเป็นผู้เชี่ยวชาญด้านสินเชื่อ

งานวิจัยนี้เป็นการประยุกต์ใช้โมเดลที่เกิดการเรียนรู้รูปแบบการตัดสินใจของโปรแกรม WEKA เวอร์ชั่น 3.6.2 โดยทำการเรียนรู้รูปแบบการตัดสินใจตามเงื่อนไขของหลักการของงานทบทวนวงเงิน หรือที่เรียกว่างาน Auto review ด้วยวิธีการจำแนกข้อมูล จากข้อมูลบัญชีเงินกู้เบิกเกินบัญชีที่เคยเข้าสู่กระบวนการทบทวนวงเงินล่วงหน้า ระหว่างเดือนมกราคม 2556 ถึงเดือน

กันยายน 2557 และเป็นข้อมูลของลูกค้าที่อยู่ในความดูแลของฝ่ายพิจารณาสินเชื่อธุรกิจ SME ของธนาคารกรุงศรีอยุธยา จำกัด (มหาชน) จำนวนทั้งสิ้น 13,765 เร็คคอร์ด สำหรับการดำเนินการวิจัยได้ทำการแบ่งข้อมูลออกเป็น 2 ชุด โดยใช้เกณฑ์เดือนของข้อมูลเป็นตัวแบ่งชุดข้อมูล ชุดแรกเป็นข้อมูลระหว่างเดือนมกราคม 2556 ถึงเดือนสิงหาคม 2557 จำนวน 12,893 เร็คคอร์ด ชุดนี้ใช้เพื่อการเรียนรู้และทดสอบโมเดลโดยโปรแกรม WEKA และชุดที่สองเป็นของเดือนกันยายน 2557 จำนวน 872 เร็คคอร์ด ชุดนี้ใช้เพื่อการประเมินประสิทธิภาพโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน



ภาพที่ 5.1 สัดส่วนร้อยละของจำนวนข้อมูลที่ใช้เป็นชุดเรียนรู้และชุดทดสอบ

ผลการทดลองจากการประมวลผลข้อมูลด้วยเทคนิค J48 และเทคนิค ID3 พบว่า

1. เทคนิค J48 มีลำดับชั้นของการจำแนกข้อมูล 72 ชั้น (layers) และได้รูปแบบการตัดสินใจ 55 รูปแบบ

2. เทคนิค ID3 มีลำดับชั้นของการจำแนกข้อมูล 181 ชั้น (layers) และได้รูปแบบการตัดสินใจ 132 รูปแบบ นอกจากนี้ ยังพบว่าเทคนิค ID3 มีรูปแบบการตัดสินใจที่ไม่สามารถพยากรณ์ข้อมูลได้ จำนวน 4 รูปแบบ คือ แสดงผลการพยากรณ์เป็น “null”

3. ค่า True Positive Rate (Average) หรือค่า Recall (Average) หรือค่าที่ใช้วัดผลการพยากรณ์ตามรูปแบบโมเดลที่ถูกค้นพบ แล้วให้คำตอบตรงกับข้อมูลจริง ซึ่งทั้งของเทคนิค J48 และ ID3 มีค่าเท่ากันคือ 0.969

4. ค่า False Positive Rate (Average) หรือค่าที่ใช้วัดผลการพยากรณ์ตามรูปแบบโมเดลที่ถูกค้นพบ แล้วให้คำตอบเป็นเท็จ หมายถึง การที่ข้อมูลจริงไม่ได้อยู่ในคลาสนั้นแต่คำตอบจากโมเดลบอกว่าอยู่ในคลาสนั้น ของเทคนิค J48 และ ID3 มีค่าเท่ากับ 0.274 และ 0.275 ตามลำดับ

5. ค่า Precision (Average) หรือค่าที่ใช้วัดความเชื่อมั่นของผลการพยากรณ์ ของเทคนิค J48 และ ID3 มีค่าเท่ากับ 0.968 และ 0.967 ตามลำดับ

จากการวิเคราะห์และเปรียบเทียบข้อดีข้อเสียของเทคนิค J48 และเทคนิค ID3 พบว่ารูปแบบโมเดลการตัดสินใจที่ได้จากเทคนิค J48 มีข้อดีมากกว่าเทคนิค ID3 ทั้งในแง่ของลำดับชั้นการตัดสินใจที่ไม่เยอะจนเกินไป ดังที่แสดงไว้ในบทที่ 3 และเทคนิค J48 มีค่าดัชนีที่ใช้วัดประสิทธิภาพของรูปแบบโมเดลการตัดสินใจ ที่แสดงถึงความมีประสิทธิภาพในการจำแนกข้อมูลหรือมีความสามารถในการพยากรณ์ข้อมูลได้ดีกว่าเทคนิค ID3

ส่วนประกอบของโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน มี 3 ส่วน ดังนี้

1. หน้าจอหลัก ใช้สำหรับการพยากรณ์

2. หน้าจอแสดง Decision Tree สำหรับผู้จัดทำข้อมูลใช้เพื่อทำความเข้าใจรูปแบบการตัดสินใจของตัวโปรแกรมพยากรณ์

3. หน้าจอแสดง Decision Tree สำหรับผู้ใช้ทั่วไปใช้เพื่อทำความเข้าใจรูปแบบการตัดสินใจของตัวโปรแกรมพยากรณ์ เพื่อความเข้าใจง่าย ผู้วิจัยจึงได้แปลงข้อมูลทางเทคนิค ให้แสดงผลเป็นเปอร์เซนต์ (%) ของค่าความน่าจะเป็น

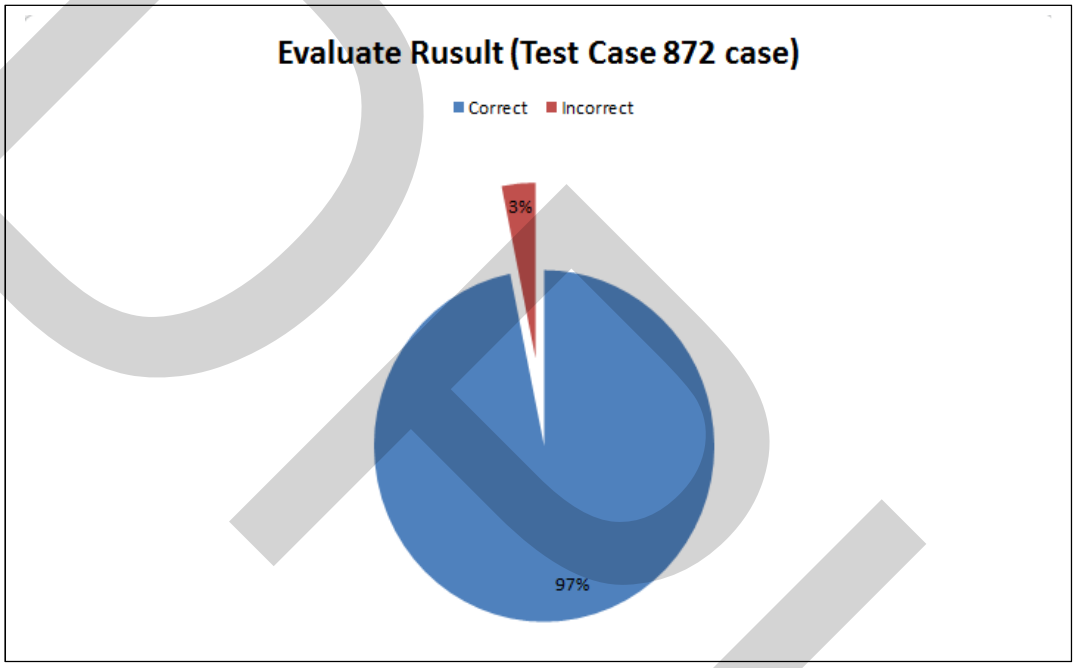
จากการนำข้อมูลชุดที่ 2 จำนวน 872 เร็คคอร์ด มาทำการทดสอบกับโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน เพื่อประเมินประสิทธิภาพเชิงปริมาณของโปรแกรมฯ สรุปผลการทดสอบได้ดังนี้

1. ผลลัพธ์จากโปรแกรมพยากรณ์ กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริงมีความถูกต้องตรงกัน (Correctly Classified) จำนวน 844 เร็คคอร์ด คิดเป็นร้อยละ 96.79 หรือเรียกว่าค่า True Positive Rate หรือค่า Recall มีค่าเท่ากับ 0.9679

2. ผลลัพธ์จากโปรแกรมพยากรณ์ไม่ตรง กับผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง (Incorrectly Classified) จำนวน 28 เร็คคอร์ด มีค่าเท่ากับ 0.0321 หรือคิดเป็นร้อยละ 3.21

3. รูปแบบโมเดลการติดสินใจที่ไม่ได้ทดสอบ เนื่องจากข้อมูลในชุดสอบไม่มีรูปแบบการตัดสินใจที่ตรงกับรูปแบบการตัดสินใจของโมเดล มีจำนวน 13 รูปแบบ

4. รูปแบบโมเดลการติดสินใจที่ผลการพยากรณ์ผิดพลาด จากผลการพิจารณาทบทวนวงเงินจากข้อมูลจริง ตั้งแต่ร้อยละ 20 ขึ้นไป มีจำนวน 11 รูปแบบ



ภาพที่ 5.2 สัดส่วนร้อยละของความถูกต้องของผลการพยากรณ์ (ชุดทดสอบ)

นอกจากนี้ ผู้วิจัยได้จัดทำแบบสอบถาม เพื่อประเมินประสิทธิภาพเชิงคุณภาพของโปรแกรมฯ จากผู้เชี่ยวชาญด้านสินเชื่อ จำนวน 33 คน แบ่งเป็นเพศชาย 18 คน และเพศหญิง 15 คน ส่วนใหญ่เป็นผู้มีตำแหน่งตั้งแต่ระดับ 7 หรือรองผู้จัดการฝ่ายขึ้นไป คิดเป็นร้อยละ 64 รองลงมาเป็นผู้มีตำแหน่งตั้งแต่ระดับ 5 - 6 หรือระดับหัวหน้างาน คิดเป็นร้อยละ 21 และตำแหน่งระดับ 1 – 4 หรือระดับเจ้าหน้าที่สินเชื่อ คิดเป็นร้อยละ 15 สามารถแปลผลการประเมินได้ดังนี้

1. เป็นผู้ที่ทราบเกี่ยวกับหลักเกณฑ์ของงานทบทวนวงเงิน ในระดับปานกลาง (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 3.27)

2. มีลูกค้าที่อยู่ในความดูแลที่เข้าร่วมโครงการของงานทบทวนวงเงิน ในระดับปานกลาง (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 2.70)

3. เห็นด้วยกับการนำ Utilization, LTV, Movement, Payment มาร่วมพิจารณาสำหรับการทบทวนวงเงินโอดี ในระดับมาก (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 3.88)

4. มีความคิดเห็นว่าการอธิบายรูปแบบการตัดสินใจด้วย Decision Tree ทำให้เข้าใจหลักเกณฑ์ของการทบทวนวงเงิน ในระดับมาก (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 3.52)

5. มีความคิดเห็นว่าการนำ Decision Tree ที่สร้างขึ้นจากโมเดลที่ได้จากการประมวลผลโดยโปรแกรม WEKA มีความสอดคล้องกับหลักความเป็นจริงทางธุรกิจ ในระดับปานกลาง (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 3.30)

6. มีความคิดเห็นว่าในส่วนผลลัพธ์ของการพยากรณ์ (Prediction) มีความน่าเชื่อถือและสอดคล้องกับความคาดหวัง ในระดับปานกลาง (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 3.24)

7. มีความคิดเห็นว่าโปรแกรมพยากรณ์ฯ ทำให้การสอบถาม หรือพยากรณ์โอกาสที่วงเงินเงินกู้เบิกเกินบัญชี (OD) จะได้รับการทบทวนวงเงิน ทำได้ง่ายขึ้น ในระดับมาก (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 3.55)

8. สรุปผลคะแนนเฉลี่ยในภาพรวมของการประเมินประสิทธิภาพเชิงคุณภาพของโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน ในระดับปานกลาง (จากคะแนนเต็ม 5 ได้คะแนนเฉลี่ย 3.35)

5.2 ข้อเสนอแนะ

1. หากผู้สนใจต้องการนำแนวทางการพัฒนาโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงินนี้ไปพัฒนาเพิ่มเติม อาจต้องทำการปรับเปลี่ยนเงื่อนไขสำหรับการจำแนกข้อมูลของแต่ละแอตทริบิวท์ ให้สอดคล้องกับนโยบายการทบทวนวงเงินของแต่ละองค์กร เนื่องจากแต่ละองค์กรอาจให้ความสำคัญกับแต่ละปัจจัยและมีมุมมองการจัดระดับความเสี่ยงที่แตกต่างกัน

2. LTV เป็นข้อมูลที่ว่าด้วยเรื่องมูลค่า หรือประเภทของหลักประกัน ที่ค้ำประกันต่อภาระหนี้ ซึ่งงานวิจัยนี้มีข้อมูล LTV ที่ระบุว่าเป็น CleanLoan หมายถึง ไม่มีมูลค่าหลักประกันที่ค้ำประกันวงเงินนั้นๆ ซึ่งในทางธุรกิจจะถือว่าวงเงินนั้นมีความเสี่ยงสูง และอาจจะพิจารณาไม่อนุมัติให้มีการทบทวนวงเงินในอนาคต แต่ผลการพยากรณ์ของโปรแกรมฯ ออกมาเป็น Pass ประเด็นนี้ผู้วิจัยได้ทำการตรวจสอบแล้ว พบว่า วงเงินเหล่านั้นมีการ Hold บัญชีเงินฝากออมทรัพย์ เพื่อหักชำระหนี้ แต่ในระบบข้อมูลหลักประกันไม่ได้บันทึกข้อมูลดังกล่าวไว้ และเมื่อข้อมูลดังกล่าวถูกตรวจพบโดยเจ้าหน้าที่ผู้เกี่ยวข้อง จึงทำการปรับเปลี่ยนผลการพิจารณาทบทวนวงเงินจาก No เป็น Pass เนื่องจากถือเสมือนว่าบัญชีเงินกู้เบิกเกินบัญชีนั้นๆ มีหลักประกันเป็นเงินฝาก

ดังนั้น เมื่อนำข้อมูลดังกล่าวมาเข้าสู่กระบวนการของการทำเหมืองข้อมูล จึงทำให้ผลการพยากรณ์ออกมาขัดแย้งกับความเป็นจริงในทางธุรกิจ แม้ว่าวงเงิน หรือข้อผิดพลาดเหล่านี้จะมีจำนวนน้อยมากก็ตาม แต่ถ้าหากนำโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนออกใช้งานจริงในอนาคต อาจส่งผลต่อความน่าเชื่อถือของตัวโปรแกรมฯ ได้ ในการวิจัยครั้งต่อไป ต้องเพิ่มความระมัดระวังในการใช้คำที่อาจสื่อความหมายคลาดเคลื่อน

3. โปรแกรม WEKA ถูกพัฒนาโดยภาษา Java ผู้ใช้อาจพบปัญหาการใช้งาน อันเนื่องมาจากเวอร์ชั่นของ Java ที่ติดตั้งไว้ในฮาร์ดแวร์ของผู้ใช้อาจไม่สนับสนุนการทำงานของโปรแกรม ซึ่งผู้ใช้อาจต้องทำศึกษาค้นคว้าเพิ่มเติมเกี่ยวกับการกำหนดค่า (Configure) และเลือกใช้เวอร์ชั่นของ Java ให้เหมาะสม

4. เกร็ดความรู้เพิ่มเติมที่ผู้วิจัยค้นพบระหว่างการดำเนินการวิจัย และเพื่อเป็นวิทยาทานแก่ผู้ที่กำลังสนใจศึกษาเกี่ยวกับการทำเหมืองข้อมูลโดยใช้โปรแกรม WEKA ช่วยในการวิเคราะห์ข้อมูลมี 2 ประเด็น ดังนี้

4.1 หลักการตั้งชื่อแอตทริบิวท์ ห้ามมีช่องว่าง คั่นภายในชื่อของแอตทริบิวท์

4.2 ค่าข้อมูลในแต่ละแอตทริบิวท์ รวมถึงค่าข้อมูลที่เกิดจากการแปลงค่าข้อมูลจากตัวเลข เป็นรูปแบบนาม (Nominal) ห้ามมีช่องว่างคั่นระหว่างค่าข้อมูล หรือ Nominal นั้นๆ เพราะจะทำให้ผลการพยากรณ์ผิดพลาด โดยเฉพาะอย่างยิ่งในกรณีที่กำหนดให้มีการเรียกใช้โมเดลจากโปรแกรม WEKA โดยตรง

บรรณานุกรม

DPU

บรรณานุกรม

ภาษาไทย

หนังสือ

ฉวีวรรณ เพ็ชรศิริ. (2556). *ระบบความชาญฉลาดทางธุรกิจ เพื่อสนับสนุนการตัดสินใจ* (พิมพ์ครั้งที่ 2). กรุงเทพฯ : มหาวิทยาลัยธุรกิจบัณฑิตย์.

ณรงค์ โพธิ์พฤกษานันท์. (2556). *ระเบียบวิธีวิจัยหลักการและแนวคิด เทคนิคการเขียนรายงานวิจัย* (พิมพ์ครั้งที่ 8). กรุงเทพมหานคร : แอคทีฟ พริ้นท์.

ณัฐพร อาทรธรรมคุณ. (2552). *การประยุกต์ค่าต้าไมนิ่งสำหรับการทำนายผลการศึกษาของนักเรียน กรณีศึกษา : โรงเรียนเตรียมทหาร สถาบันวิชาการป้องกันประเทศ* (สารนิพนธ์ปริญญามหาบัณฑิต). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์.

ภิเษก ชัยนิรันดร์. (2553). *การบริหารสินเชื่อ* (พิมพ์ครั้งที่ 1). กรุงเทพมหานคร : สายธาร.

เสถียร วิชาเรือง. (2553). *การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในกระบวนการอนุมัติ* (สารนิพนธ์ปริญญามหาบัณฑิต). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์.

สิทธิโชค มุกดาสกุลภิบาล. (2551). *การวัดประสิทธิภาพของขั้นตอนวิธีตัวจำแนก C4.5, ADTree และ Naive Bayes ในการจำแนกข้อมูลการซุกซ่อนสิ่งเสพติดสำหรับไปรษณีย์ระหว่างประเทศ* (สารนิพนธ์ปริญญามหาบัณฑิต). กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2556). *คู่มือการใช้งาน WEKA Explorer เบื้องต้น* (พิมพ์ครั้งที่ 1). กรุงเทพฯ : เอเชีย ดิจิตอลการพิมพ์.

บทความ

ชุติมา อุตมะมุณีย์ และ ประสงค์ ปราณีตพลกรัง. (2553). *การพัฒนาตัวแบบระบบสนับสนุนการตัดสินใจแบบอัตโนมัติออนไลน์สำหรับการเลือกสาขาวิชาเรียนของนักศึกษาระดับอุดมศึกษา* (รายงานผลการวิจัย).หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม

นโรดม อารีกุล และ ชุลีรัตน์ จรัสกุลชัย. (2554*). การใช้เทคนิคการทำเหมืองข้อมูลในการค้นหาผู้หลบหลีกภาษี* (รายงานผลการวิจัย). ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

ปิยะมาศ กรัณย์ภควุฒิ. (2556). *ปัจจัยที่มีอิทธิพลต่อค่าความแม่นยำของการจำแนกข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลเปรียบเทียบระหว่างอัลกอริธึม Naive Bayes Simples และ ID3* (รายงานผลการวิจัย). หลักสูตรวิทยาศาสตรมหาบัณฑิต เทคโนโลยีสารสนเทศสำหรับวิสาหกิจสมัยใหม่ มหาวิทยาลัยธุรกิจบัณฑิตย์.

สารสนเทศจากสื่ออิเล็กทรอนิกส์

ธนาคารกรุงศรีอยุธยาจำกัด (มหาชน). (2556). *เกี่ยวกับเรา.* สืบค้นเมื่อ 30 ธันวาคม 2556, จาก http://www.krungsri.com

สำนักงานเศรษฐกิจการคลัง. (2549). *ความหมายและลักษณะของประเภทสินเชื่อ.* สืบค้นเมื่อ 7 กรกฎาคม 2557, จาก http://www.fpo.go.th/S-I/Source/ECO/ECO31.htm

ภาษาต่างประเทศ

BOOKS

Berry, Michael J.A. & Linoff, Gordon S. (2004). *Data Mining Techniques : for marketing, sales and customer relationship management (2nd ed.).* United States of America : Wiley Publishing, Inc.

Ohlhorst, Frank. (2013). *Big Data Analytics.* United States of America : John Wiley & Sons, Inc.

DPU

ภาคผนวก

DPU

ภาคผนวก ก

รูปแบบโมเดลการตัดสินใจของเทคนิค J48

DPU

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Data_IS_unmissing_V5_new2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    12893
Attributes:   5
              Utilization
              Movement
              LTV
              PaymentHistory
              Class
Test mode:    10-fold cross-validation

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

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

Utilization = >95-100
|   LTV = <=75: Pass (1292.0/95.0)
|   LTV = >80-100
|   |   Movement = <10: No (80.0/10.0)
|   |   Movement = 15-19.99: No (13.0)
|   |   Movement = 10-14.99: No (12.0/2.0)
|   |   Movement = >=20: Pass (38.0/8.0)
|   LTV = >75-80
|   |   Movement = <10: No (73.0/7.0)
|   |   Movement = 15-19.99: Pass (1.0)
|   |   Movement = 10-14.99: Pass (9.0/2.0)
|   |   Movement = >=20
|   |   |   PaymentHistory = Good: Pass (25.0/4.0)
|   |   |   PaymentHistory = Bad: No (2.0)
|   LTV = CleanLoan: Pass (189.0/17.0)
|   LTV = >100
|   |   Movement = <10: No (92.0/31.0)
|   |   Movement = 15-19.99: No (7.0/1.0)
|   |   Movement = 10-14.99: No (10.0/2.0)
|   |   Movement = >=20: Pass (30.0/4.0)
|   LTV = Deposit: Pass (129.0/2.0)
Utilization = <=80: Pass (8202.0/58.0)
Utilization = >100
|   LTV = <=75
|   |   Movement = <10: No (293.0/4.0)
|   |   Movement = 15-19.99: Pass (2.0/1.0)
|   |   Movement = 10-14.99
|   |   |   PaymentHistory = Good: Pass (3.0/1.0)
|   |   |   PaymentHistory = Bad: No (2.0)
|   |   Movement = >=20
|   |   |   PaymentHistory = Good: Pass (16.0/4.0)
|   |   |   PaymentHistory = Bad: No (3.0)
|   LTV = >80-100: No (30.0/3.0)
|   LTV = >75-80: No (24.0/2.0)
|   LTV = CleanLoan: Pass (35.0/1.0)
|   LTV = >100: No (22.0/7.0)
|   LTV = Deposit: Pass (22.0/1.0)
Utilization = >80-90
|   LTV = <=75: Pass (858.0/16.0)
|   LTV = >80-100
|   |   Movement = <10: No (42.0/3.0)
|   |   Movement = 15-19.99: Pass (7.0/1.0)
|   |   Movement = 10-14.99: No (6.0/2.0)
|   |   Movement = >=20: Pass (64.0/8.0)
|   LTV = >75-80: Pass (84.0/5.0)
|   LTV = CleanLoan: Pass (76.0/8.0)
```

ภาพที่ ก.1 รูปแบบโมเดลการตัดสินใจของเทคนิค J48

```
|  LTV = >100
|  |  Movement = <10: No (22.0/7.0)
|  |  Movement = 15-19.99: Pass (2.0)
|  |  Movement = 10-14.99: Pass (9.0/1.0)
|  |  Movement = >=20: Pass (31.0/4.0)
|  LTV = Deposit: Pass (63.0/2.0)
Utilization = >90-95
|  LTV = <=75: Pass (656.0/33.0)
|  LTV = >80-100
|  |  Movement = <10: No (25.0/4.0)
|  |  Movement = 15-19.99: Pass (5.0/1.0)
|  |  Movement = 10-14.99: No (6.0)
|  |  Movement = >=20: Pass (45.0/9.0)
|  LTV = >75-80
|  |  Movement = <10: No (27.0/1.0)
|  |  Movement = 15-19.99: Pass (3.0)
|  |  Movement = 10-14.99: Pass (6.0)
|  |  Movement = >=20: Pass (37.0/2.0)
|  LTV = CleanLoan: Pass (64.0/4.0)
|  LTV = >100
|  |  Movement = <10: No (24.0/4.0)
|  |  Movement = 15-19.99: No (0.0)
|  |  Movement = 10-14.99: No (4.0)
|  |  Movement = >=20: Pass (21.0/2.0)
|  LTV = Deposit: Pass (50.0/2.0)

Number of Leaves  :     55

Size of the tree :      72


Time taken to build model: 0.06 seconds

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

Correctly Classified Instances       12495               96.9131 %
Incorrectly Classified Instances       398                3.0869 %
Kappa statistic                          0.7673
Mean absolute error                      0.0545
Root mean squared error                  0.1671
Relative absolute error                 37.2147 %
Root relative squared error             61.7539 %
Total Number of Instances            12893

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.992    0.297      0.975     0.992     0.983      0.92     Pass
                 0.703    0.008      0.885     0.703     0.784      0.92     No
Weighted Avg.    0.969    0.274      0.968     0.969     0.968      0.92

=== Confusion Matrix ===

     a     b   <-- classified as
 11774    94 |     a = Pass
   304   721 |     b = No
```

ภาพที่ ก.1 (ต่อ)

ภาคผนวก ข

รูปแบบโมเดลการตัดสินใจของเทคนิค ID3

DPU

```
=== Run information ===

Scheme:       weka.classifiers.trees.Id3
Relation:     Data_IS_unmissing_V5_new2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    12893
Attributes:   5
              Utilization
              Movement
              LTV
              PaymentHistory
              Class
Test mode:    10-fold cross-validation

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

Id3


Utilization = >95-100
| LTV = <=75
| | Movement = <10: Pass
| | Movement = 15-19.99: Pass
| | Movement = 10-14.99: Pass
| | Movement = >=20: Pass
| LTV = >80-100
| | Movement = <10
| | | PaymentHistory = Good: No
| | | PaymentHistory = Bad: No
| | Movement = 15-19.99: No
| | Movement = 10-14.99: No
| | Movement = >=20: Pass
| LTV = >75-80
| | Movement = <10
| | | PaymentHistory = Good: No
| | | PaymentHistory = Bad: No
| | Movement = 15-19.99: Pass
| | Movement = 10-14.99
| | | PaymentHistory = Good: Pass
| | | PaymentHistory = Bad: No
| | Movement = >=20
| | | PaymentHistory = Good: Pass
| | | PaymentHistory = Bad: No
| LTV = CleanLoan
| | Movement = <10
| | | PaymentHistory = Good: Pass
| | | PaymentHistory = Bad: Pass
| | Movement = 15-19.99: Pass
| | Movement = 10-14.99: Pass
| | Movement = >=20: Pass
| LTV = >100
| | Movement = <10
| | | PaymentHistory = Good: No
| | | PaymentHistory = Bad: No
| | Movement = 15-19.99: No
| | Movement = 10-14.99
| | | PaymentHistory = Good: No
| | | PaymentHistory = Bad: No
| | Movement = >=20
| | | PaymentHistory = Good: Pass
| | | PaymentHistory = Bad: No
| LTV = Deposit
| | Movement = <10
| | | PaymentHistory = Good: Pass
| | | PaymentHistory = Bad: Pass
| | Movement = 15-19.99: Pass
```

ภาพที่ ข.1 รูปแบบโมเดลการตัดสินใจของเทคนิค J48

```
66  | | Movement = 10-14.99: Pass
67  | | Movement = >=20: Pass
68  Utilization = <=80
69  | LTV = <=75
70  | | Movement = <10: Pass
71  | | Movement = 15-19.99: Pass
72  | | Movement = 10-14.99: Pass
73  | | Movement = >=20: Pass
74  | LTV = >80-100
75  | | Movement = <10: Pass
76  | | Movement = 15-19.99: Pass
77  | | Movement = 10-14.99: Pass
78  | | Movement = >=20: Pass
79  | LTV = >75-80
80  | | Movement = <10: Pass
81  | | Movement = 15-19.99: Pass
82  | | Movement = 10-14.99: Pass
83  | | Movement = >=20: Pass
84  | LTV = CleanLoan: Pass
85  | LTV = >100
86  | | Movement = <10: Pass
87  | | Movement = 15-19.99: Pass
88  | | Movement = 10-14.99: Pass
89  | | Movement = >=20: Pass
90  | LTV = Deposit: Pass
91  Utilization = >100
92  | LTV = <=75
93  | | Movement = <10
94  | | | PaymentHistory = Good: No
95  | | | PaymentHistory = Bad: No
96  | | Movement = 15-19.99
97  | | | PaymentHistory = Good: Pass
98  | | | PaymentHistory = Bad: No
99  | | Movement = 10-14.99
100 | | | PaymentHistory = Good: Pass
101 | | | PaymentHistory = Bad: No
102 | | Movement = >=20
103 | | | PaymentHistory = Good: Pass
104 | | | PaymentHistory = Bad: No
105 | LTV = >80-100
106 | | PaymentHistory = Good: No
107 | | PaymentHistory = Bad: No
108 | LTV = >75-80
109 | | PaymentHistory = Good: No
110 | | PaymentHistory = Bad: No
111 | LTV = CleanLoan
112 | | PaymentHistory = Good
113 | | | Movement = <10: Pass
114 | | | Movement = 15-19.99: null
115 | | | Movement = 10-14.99: Pass
116 | | | Movement = >=20: Pass
117 | | PaymentHistory = Bad: Pass
118 | LTV = >100
119 | | Movement = <10
120 | | | PaymentHistory = Good: No
121 | | | PaymentHistory = Bad: No
122 | | Movement = 15-19.99: null
123 | | Movement = 10-14.99: null
124 | | Movement = >=20: Pass
125 | LTV = Deposit
126 | | PaymentHistory = Good: Pass
127 | | PaymentHistory = Bad: Pass
128 Utilization = >80-90
129 | LTV = <=75
130 | | Movement = <10: Pass
```

ภาพที่ ข.1 (ต่อ)

```
131  | |  Movement = 15-19.99: Pass
132  | |  Movement = 10-14.99: Pass
133  | |  Movement = >=20: Pass
134  | LTV = >80-100
135  | |  Movement = <10
136  | | |  PaymentHistory = Good: No
137  | | |  PaymentHistory = Bad: No
138  | |  Movement = 15-19.99: Pass
139  | |  Movement = 10-14.99: No
140  | |  Movement = >=20: Pass
141  | LTV = >75-80
142  | |  Movement = <10: Pass
143  | |  Movement = 15-19.99: Pass
144  | |  Movement = 10-14.99: Pass
145  | |  Movement = >=20: Pass
146  | LTV = CleanLoan
147  | |  Movement = <10: Pass
148  | |  Movement = 15-19.99: Pass
149  | |  Movement = 10-14.99: Pass
150  | |  Movement = >=20: Pass
151  | LTV = >100
152  | |  Movement = <10: No
153  | |  Movement = 15-19.99: Pass
154  | |  Movement = 10-14.99: Pass
155  | |  Movement = >=20: Pass
156  | LTV = Deposit
157  | |  Movement = <10: Pass
158  | |  Movement = 15-19.99: Pass
159  | |  Movement = 10-14.99: Pass
160  | |  Movement = >=20: Pass
161  Utilization = >90-95
162  | LTV = <=75
163  | |  Movement = <10: Pass
164  | |  Movement = 15-19.99: Pass
165  | |  Movement = 10-14.99: Pass
166  | |  Movement = >=20: Pass
167  | LTV = >80-100
168  | |  Movement = <10: No
169  | |  Movement = 15-19.99: Pass
170  | |  Movement = 10-14.99: No
171  | |  Movement = >=20: Pass
172  | LTV = >75-80
173  | |  Movement = <10
174  | | |  PaymentHistory = Good: No
175  | | |  PaymentHistory = Bad: No
176  | |  Movement = 15-19.99: Pass
177  | |  Movement = 10-14.99: Pass
178  | |  Movement = >=20: Pass
179  | LTV = CleanLoan
180  | |  Movement = <10: Pass
181  | |  Movement = 15-19.99: Pass
182  | |  Movement = 10-14.99: Pass
183  | |  Movement = >=20: Pass
184  | LTV = >100
185  | |  Movement = <10
186  | | |  PaymentHistory = Good: No
187  | | |  PaymentHistory = Bad: Pass
188  | |  Movement = 15-19.99: null
189  | |  Movement = 10-14.99: No
190  | |  Movement = >=20
191  | | |  PaymentHistory = Good: Pass
192  | | |  PaymentHistory = Bad: Pass
193  | LTV = Deposit
194  | |  Movement = <10
195  | | |  PaymentHistory = Good: Pass
196  | | |  PaymentHistory = Bad: Pass
197  | |  Movement = 15-19.99: Pass
198  | |  Movement = 10-14.99: Pass
199  | |  Movement = >=20: Pass
200
201  Time taken to build model: 0.03 seconds
202
203  === Stratified cross-validation ===
204  === Summary ===
205
206  Correctly Classified Instances       12489               96.8665 %
207  Incorrectly Classified Instances       400                3.1025 %
208  Kappa statistic                          0.7661
209  Mean absolute error                      0.0531
210  Root mean squared error                  0.1659
211  Relative absolute error                 36.2797 %
212  Root relative squared error             61.3118 %
213  UnClassified Instances                   4                0.031  %
214  Total Number of Instances            12893
215
216  === Detailed Accuracy By Class ===
217
218                TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
219                  0.992    0.298      0.975     0.992     0.983      0.938    Pass
220                  0.702    0.008      0.883     0.702     0.783      0.939    No
221  Weighted Avg.    0.969    0.275      0.967     0.969     0.967      0.938
222
223  === Confusion Matrix ===
224
225     a     b   <-- classified as
226  11769    95 |     a = Pass
227    305   720 |     b = No
```

ภาพที่ ข.1 (ต่อ)

ภาคผนวก ค

แบบประเมินประสิทธิภาพโปรแกรมพยากรณ์และการแปลผล

DPU

แบบประเมินประสิทธิภาพ

โปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

คำชี้แจง โปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน เป็นการประยุกต์ใช้โมเดลจากการเรียนรู้รูปแบบการตัดสินใจ ของซอร์ฟแวร์ WEKA โดยเรียนรู้รูปแบบการตัดสินใจตามเงื่อนไขของหลักเกณฑ์งาน Auto review ด้วยวิธีการจำแนกข้อมูลของเทคนิคการทำเหมืองข้อมูล (Data Mining) โปรดทำเครื่องหมาย / ลงในช่องว่างที่กำหนด

ส่วนที่ 1 ข้อมูลทั่วไป

1. เพศ ( ) ชาย ( ) หญิง
2. ระดับตำแหน่ง ( ) ระดับ 1 - 4 ( ) ระดับ 5 – 6 ( ) ระดับ 7 ขึ้นไป

ส่วนที่ 2 ความคิดเห็นของผู้เชี่ยวชาญด้านสินเชื่อ โดยแบ่งน้ำหนักระดับคะแนนความคิดเห็นออกเป็น 5 ระดับ ดังนี้

1 = น้อยที่สุด, 2 = น้อย, 3 =ปานกลาง, 4 = มาก, 5 = มากที่สุด

ประเด็นคำถาม	1	2	3	4	5
1. ท่านทราบเกี่ยวกับหลักเกณฑ์งาน Auto review ของวงเงินโอดี ในระดับใด					
2. ท่านมีลูกค้าที่เข้าร่วมโครงการ Auto review ของวงเงินโอดี ระดับใด					
3. ท่านเห็นด้วยกับการให้ Auto review ของวงเงินโอดี ควรพิจารณาจาก Utilization, LTV, Movement, Payment ตามลำดับ ในระดับใด					
4. Decision Tree (เอกสารแนบ) ทำให้ท่านมีความเข้าใจในหลักเกณฑ์ของงาน Auto review ในระดับใด					
5. Decision Tree (เอกสารแนบ) ที่พัฒนาขึ้นจากโมเดลที่ได้จากการประมวลผลโดยซอฟแวร์ WEKA มีความสอดคล้องกับหลักความเป็นจริงทางธุรกิจ ในระดับใด					

ประเด็นคำถาม	1	2	3	4	5
6. การทดสอบการใช้งานในส่วน Prediction ผลลัพธ์ที่ได้มีความน่าเชื่อถือและสอดคล้องกับความคาดหวังของท่านในระดับใด					
7. โปรแกรมนี้สามารถทำให้การสอบถาม หรือพยากรณ์โอกาสที่วงเงินเบิกเกินบัญชีจะได้รับการต่ออายุสัญญาโดยอัตโนมัติ ทำได้ง่ายขึ้นหรือไม่ ในระดับใด					

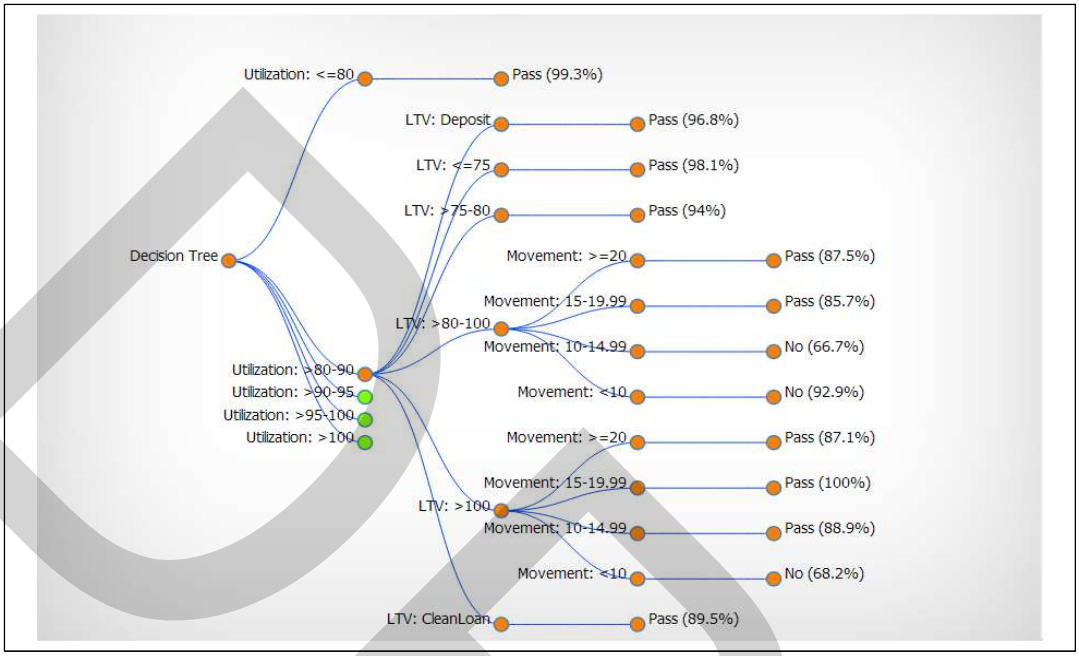
ส่วนที่ 3 ข้อเสนอแนะ

.....

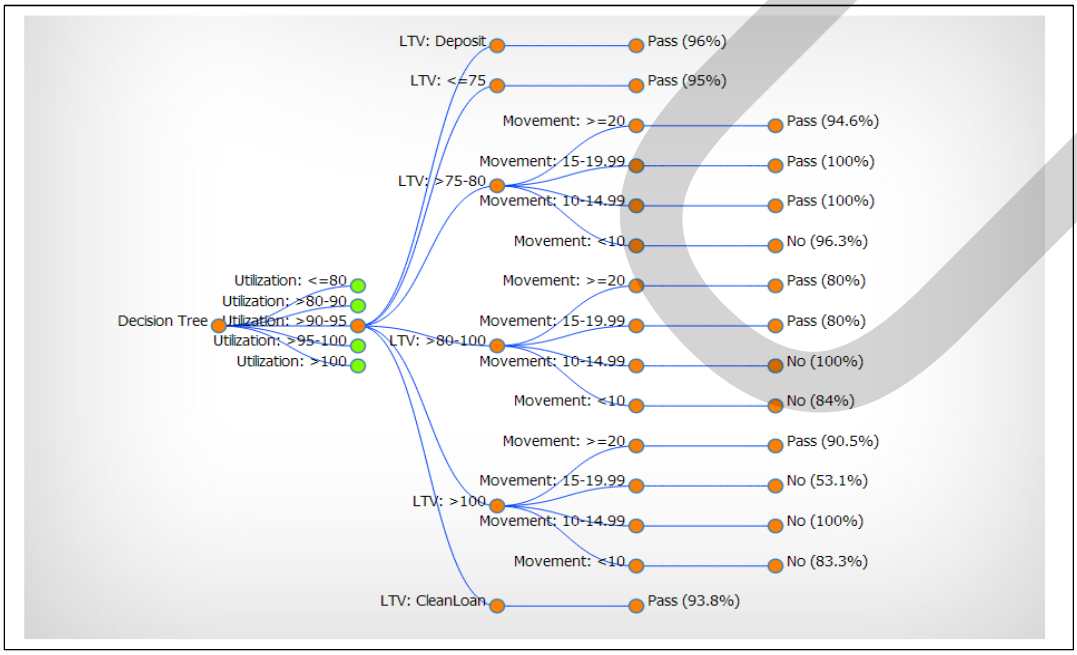
.....

DPU

(เอกสารแนบ 1/2)

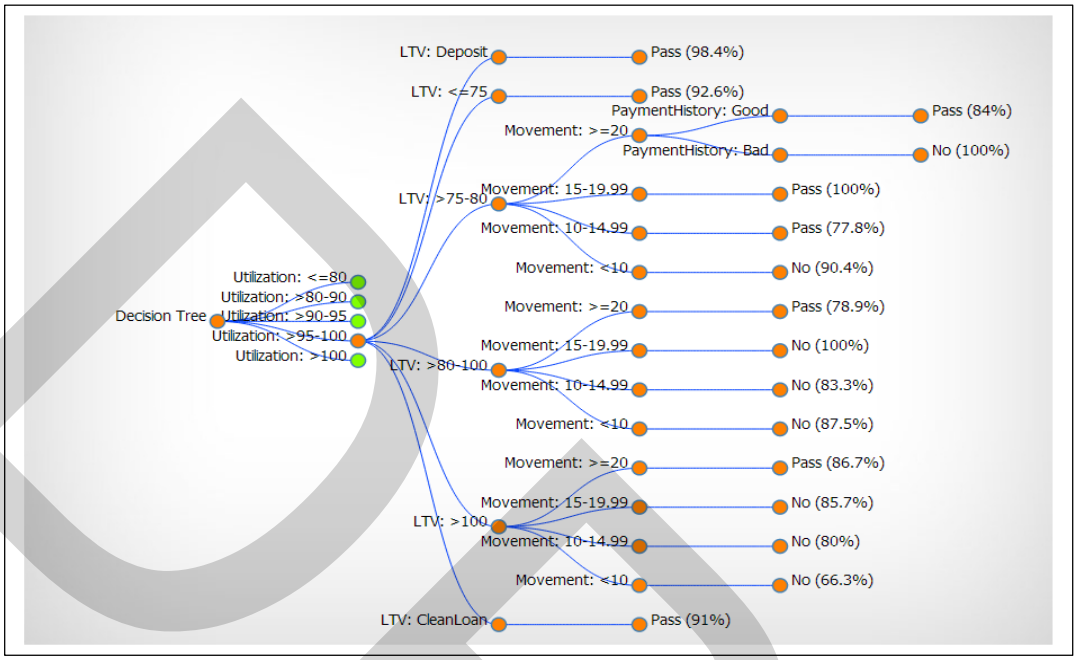


ภาพที่ ค.1 รูปแบบการตัดสินใจของโมเดล กรณี Utilize <= 80 และ Utilize > 80-90

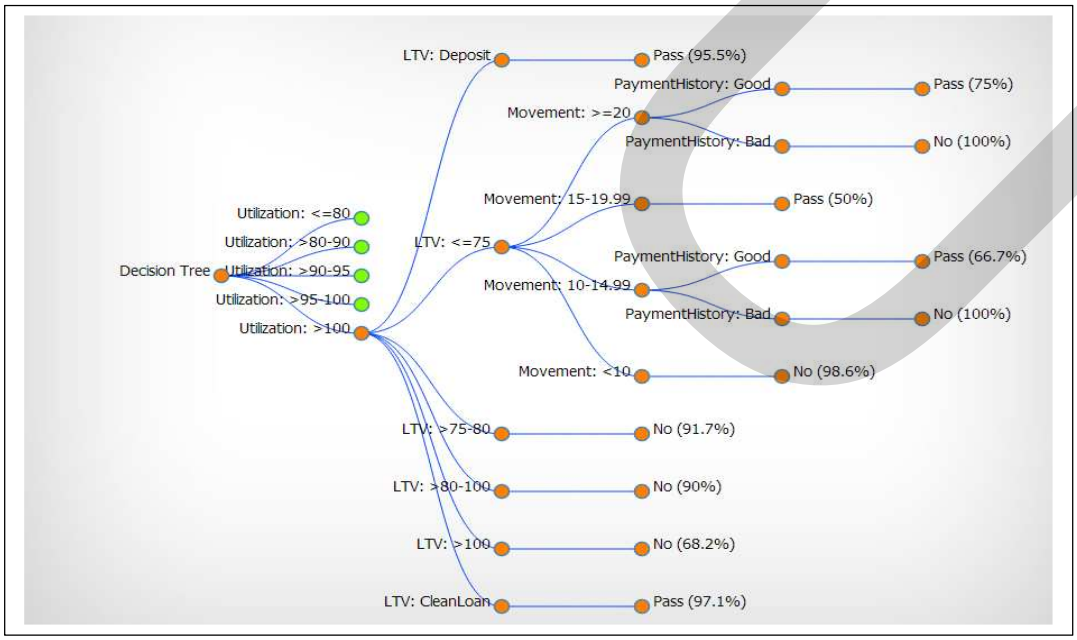


ภาพที่ ค.2 รูปแบบการตัดสินใจของโมเดล กรณี Utilize > 90-95

(เอกสารแนบ 2/2)



ภาพที่ ค.3 รูปแบบการตัดสินใจของโมเดล กรณี Utilize > 95-100



ภาพที่ ค.4 รูปแบบการตัดสินใจของโมเดล กรณี Utilize > 100

สรุปผลการประเมินประสิทธิภาพ
โปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน

งานวิจัยนี้ได้จัดให้มีการประเมินประสิทธิภาพของโปรแกรมพยากรณ์โอกาสที่บัญชีเงินกู้เบิกเกินบัญชีจะได้รับการทบทวนวงเงิน จากผู้เชี่ยวชาญด้านสินเชื่อ จำนวน 33 คน โดยสามารถสรุปผลจากแบบประเมิน ได้ดังนี้

ส่วนที่ 1 ข้อมูลทั่วไป

1. เพศของผู้ตอบแบบประเมิน

เพศ	จำนวน	ร้อยละ
ชาย	18	55%
หญิง	15	45%
รวม	**33**	**100%**

2. ระดับตำแหน่งของผู้ตอบแบบประเมิน

ระดับตำแหน่ง	จำนวน	ร้อยละ
1 - 4	5	15%
5 - 6	7	21%
7 ขึ้นไป	21	64%
รวม	**33**	**100%**

ส่วนที่ 2 ประเมินผลประสิทธิภาพของโปรแกรมจากความคิดเห็นของผู้เชี่ยวชาญด้านสินเชื่อ โดยใช้เกณฑ์คะแนนเฉลี่ยแบบลิเคิร์ท (Likert Scale) โดยแบ่งน้ำหนักของคะแนนความคิดเห็นออกเป็น 5 ระดับ ได้แก่ 1 = น้อยที่สุด, 2 = น้อย, 3 =ปานกลาง, 4 = มาก, 5 = มากที่สุด

ช่วงคะแนนของแต่ละระดับ $= \frac{\text{ค่าสูงสุด} - \text{ค่าต่ำสุด}}{\text{จำนวนระดับ}} = \frac{5-1}{5} = 0.80$

ดังนั้น คะแนนเฉลี่ยแต่ละช่วงมีเกณฑ์ ดังนี้

ช่วงคะแนนเฉลี่ย	ความหมาย
1 – 1.80	น้อยที่สุด
>1.80 - 2.60	น้อย
>2.60 - 3.40	ปานกลาง
>3.40 - 4.20	มาก
>4.20 - 5.00	มากที่สุด

จากแบบประเมินสามารถหาค่าคะแนนเฉลี่ยได้ดังนี้

1. ค่าคะแนนเฉลี่ย $(\overline{X}) = \left(\frac{\sum(\text{จำนวนคน} * \text{คะแนน})}{\text{จำนวนคนรวม}}\right)$

2. คะแนนเฉลี่ยภาพรวม $= \frac{\sum(\overline{X})}{\text{จำนวนข้อคำถาม}}$

ประเด็นคำถาม	จำนวนผู้ตอบแบบประเมิน จำแนกตามระดับคะแนน					ค่าคะแนนเฉลี่ย ($\bar{X}$)	แปลผล (ระดับ)
	1	2	3	4	5		
1. ท่านทราบเกี่ยวกับหลักเกณฑ์งาน Auto review ของวงเงินโอดี ในระดับใด		3	19	8	3	$\frac{6+57+30+15}{33} = 3.27$	ปานกลาง
2. ท่านมีลูกค้าที่เข้าร่วมโครงการ Auto review ของวงเงินโอดี ระดับใด	2	13	13	3	2	$\frac{2+26+39+12+10}{33} = 2.70$	ปานกลาง
3. ท่านเห็นด้วยกับการให้ Auto review ของวงเงินโอดี ควรพิจารณาจาก Utilization, LTV, Movement, Payment ตามลำดับ ในระดับใด		3	8	12	10	$\frac{6+24+48+50}{33} = 3.88$	มาก
4. Decision Tree (เอกสารแนบ) ทำให้ท่านมีความเข้าใจในหลักเกณฑ์ของงาน Auto review ในระดับใด		3	13	14	3	$\frac{6+39+56+15}{33} = 3.52$	มาก
5. Decision Tree (เอกสารแนบ) ที่พัฒนาขึ้นจากโมเดลที่ได้จากการประมวลผลโดยซอฟแวร์ WEKA มีความสอดคล้องกับหลักความเป็นจริงทางธุรกิจ ในระดับใด		2	19	12		$\frac{4+57+48}{33} = 3.30$	ปานกลาง
6. การทดสอบการใช้งานในส่วน Prediction ผลลัพธ์ที่ได้มีความน่าเชื่อถือและสอดคล้องกับความคาดหวังของท่าน ในระดับใด		2	23	8		$\frac{6+69+32}{33} = 3.24$	ปานกลาง
7. โปรแกรมนี้สามารถทำให้การสอบถาม หรือพยากรณ์โอกาสที่วงเงินเบิกเกินบัญชีจะได้รับการต่ออายุสัญญาโดยอัตโนมัติ ทำได้ง่ายขึ้นหรือไม่ ในระดับใด		1	15	15	2	$\frac{2+45+60+10}{33} = 3.55$	มาก
คะแนนเฉลี่ยภาพรวม						$\frac{23.46}{7} = 3.35$	**ปานกลาง**

สรุปผล

จากการประเมินประสิทธิภาพโปรแกรมพยากรณ์โอกาสที่วงเงินเบิกเกินบัญชีจะได้รับการทบทวนวงเงิน โดยผู้เชี่ยวชาญด้านสินเชื่อ จำนวน 33 คน เป็นเพศชาย 18 คน และเพศหญิง 15 คน ส่วนใหญ่เป็นผู้มีตำแหน่งตั้งแต่ระดับ 7 หรือรองผู้จัดการฝ่ายขึ้นไป คิดเป็นร้อยละ 64 รองลงมาเป็นผู้มีตำแหน่งตั้งแต่ระดับ 5 - 6 หรือระดับหัวหน้างาน คิดเป็นร้อยละ 21 และตำแหน่งระดับ 1 – 4 หรือระดับเจ้าหน้าที่สินเชื่อ คิดเป็นร้อยละ 15 และสามารถแปลผลการประเมินได้ดังนี้

เป็นผู้ที่ทราบเกี่ยวกับหลักเกณฑ์การ Auto review ของวงเงินโอดีและมีลูกค้าที่เข้าร่วมโครงการ Auto review ในระดับปานกลาง และเห็นด้วยกับการนำ Utilization, LTV, Movement, Payment มาร่วมพิจารณาสำหรับการทบทวนวงเงินโอดี ในระดับมาก โดยเห็นว่าการนำ Decision Tree ที่สร้างขึ้นจากโมเดลที่ได้จากการประมวลผลโดยซอฟแวร์ WEKA มีความสอดคล้องกับหลักความเป็นจริงทางธุรกิจ ในระดับปานกลาง ในส่วนผลลัพธ์ของการพยากรณ์(Prediction) มีความน่าเชื่อถือและสอดคล้องกับความคาดหวัง ในระดับปานกลาง และมีความเห็นว่าโปรแกรมนี้สามารถทำให้การสอบถาม หรือพยากรณ์โอกาสที่วงเงินเบิกเกินบัญชี(OD)จะได้รับการทบทวนวงเงินหรือไม่ ทำได้ง่ายขึ้น ในระดับมาก

คะแนนเฉลี่ยในภาพรวมของโปรแกรมพยากรณ์โอกาสที่วงเงินเบิกเกินบัญชี(OD)จะได้รับการทบทวนวงเงิน ในระดับปานกลาง

ภาคผนวก ง
บทความนำเสนอในงานประชุมวิชาการระดับประเทศ
ด้านเทคโนโลยีสารสนเทศ
ครั้งที่ 6

(National Conference on Information Technology: NCIT)

DPU

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 จึงเป็นสิ่งจำเป็นต่อการดำเนินธุรกิจ การที่จะได้มาซึ่งข้อมูลเหล่านั้น 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 และมีค่ามากว่า 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

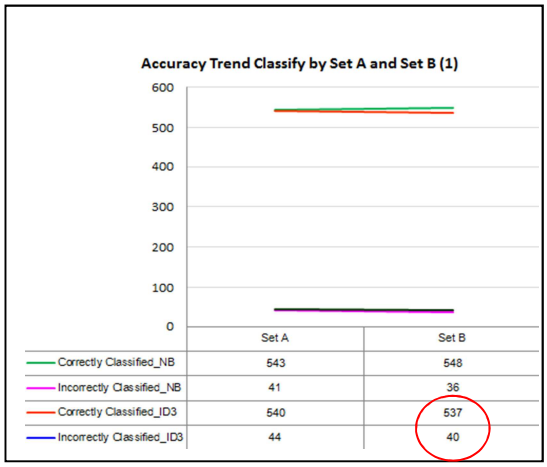
จากค่าในตารางที่ 8. สามารถนำไปคำนวณหาค่าต่างๆได้โดยใช้สูตรคำนวณดังในตารางที่ 4. และ 5. ในส่วนตารางที่ 9. เป็นการแสดงผลจากการคำนวณโดยโปรแกรม WEKA ซึ่งมาจากสูตรคำนวณในตารางที่ 5. เมื่อพิจารณาผลการทดลองในตารางที่ 9. ในส่วนที่เป็นค่าเฉลี่ย(Average) พบว่า ค่า True Postitive Rate หรือค่าที่ใช้อธิบายผลการทำนายด้วยรูปแบบหรือโมเดลที่ถูกค้นพบ 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

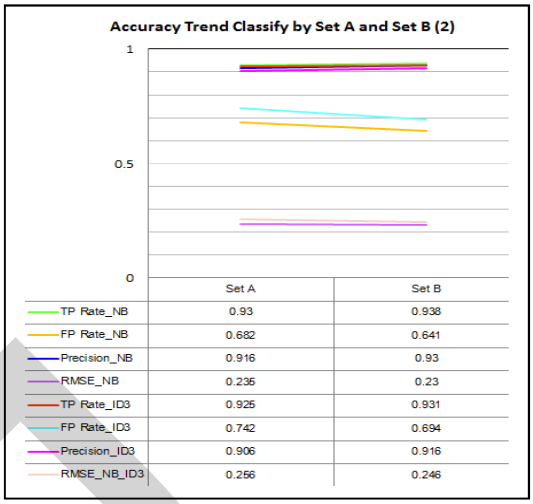
จากผลการทดลอง เปรียบเทียบการแบ่งกลุ่มหรือการคลัสเตอร์ค่าของข้อมูลในแอตทริบิวท์หนึ่งๆ ด้วยการแบ่งกลุ่มที่แตกต่างกันในชุดข้อมูลทดลอง 2 ชุด คือ ชุด A และ ชุด B ด้วยวิธีการกำหนดเงื่อนไขการแบ่งช่วงในลักษณะที่แตกต่างกัน เพื่อพิสูจน์ว่ารูปแบบการจัดการข้อมูลในลักษณะดังกล่าวจะมีผลค่าต่อความแม่นยำในการจำแนกข้อมูลหรือไม่ 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] ฉวีวรรณ 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.

ประวัติผู้เขียน

ชื่อ – นามสกุล	น.ส.ปิยะมาศ กรัณย์ภควุฒิ
ประวัติการศึกษา	ปีการศึกษา 2545 ศิลปศาสตรบัณฑิต (รัฐศาสตร์) มหาวิทยาลัยรามคำแหง
ตำแหน่งและสถานที่ทำงานปัจจุบัน	เจ้าหน้าที่วิเคราะห์ข้อมูล ฝ่ายบริหารความเสี่ยง Portfolio สินเชื่อธุรกิจ SME กลุ่มสนับสนุนและบริหารข้อมูลสารสนเทศสินเชื่อธุรกิจ SME ธนาคารกรุงศรีอยุธยา จำกัด (มหาชน)

DPU