การใช้เหมืองข้อมูลสำหรับพิจารณาการให้สินเชื่อสำหรับธนาคาร

วิทยานิพนธ์

ของ

ขวัญนภา พิมพ์ชารีย์

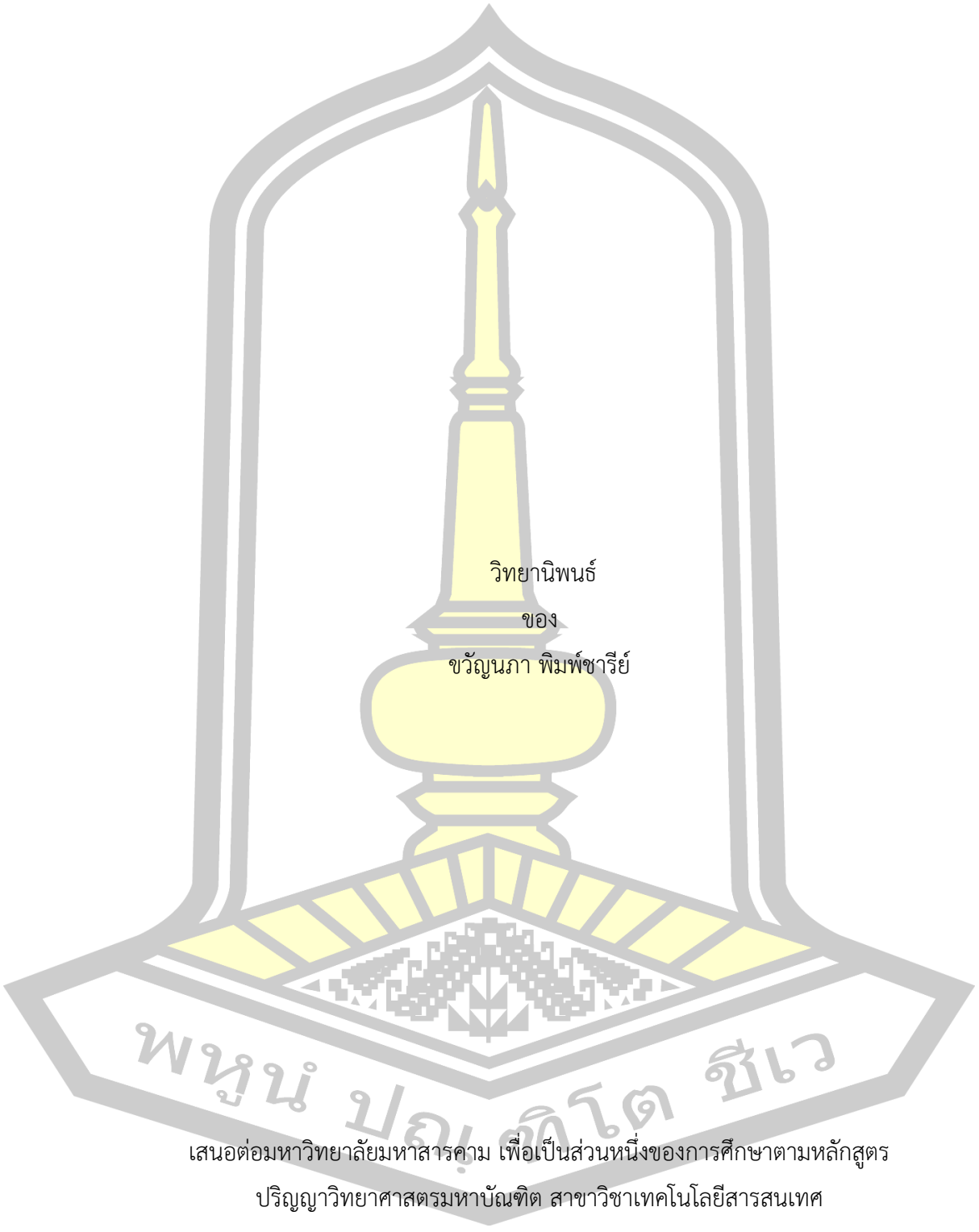
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

กรกฎาคม 2564

ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

การใช้เหมืองข้อมูลสำหรับพิจารณาการให้สินเชื่อสำหรับธนาคาร



วิทยานิพนธ์

ของ

ขวัญนภา พิมพ์ชารีย์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

กรกฎาคม 2564

ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

Data Mining Approaches for Loan Decision of a Bank

Kuannabha Pimcharee

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

July 2021

Copyright of Mahasarakham University

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนางสาวขวัญนภา พิมพ์ชารีย์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รศ. ดร. สิทธิชัย บุษหมั่น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผศ. ดร. โอฬาริก สุรินต๊ะ)

.....กรรมการ

(ผศ. ดร. ฉัตรเกล้า เจริญผล)

.....กรรมการ

(ผศ. ดร. จิรัฏฐา ภูบุญอบ)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....

(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....

(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

พหุน ปญฺ ฑิโต ชีวเ

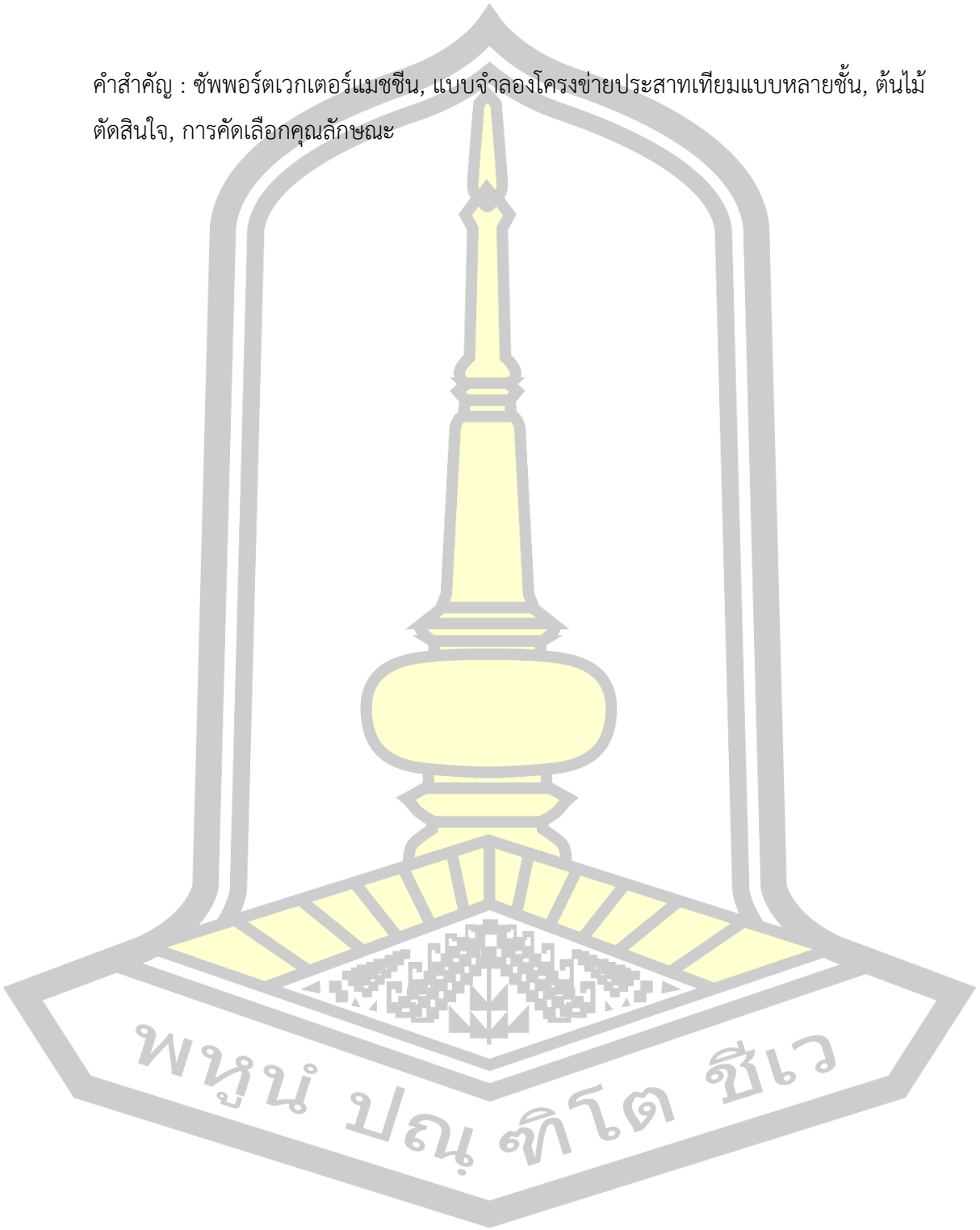
ชื่อเรื่อง การใช้เหมืองข้อมูลสำหรับพิจารณาการให้สินเชื่อสำหรับธนาคาร
ผู้วิจัย ขวัญนภา พิมพ์ชารีย์
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. โอฬาริก สุรินต๊ะ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2564

บทคัดย่อ

การพิจารณาเพื่ออนุมัติสินเชื่อของธนาคารให้แก่บุคคลต้องอาศัยปัจจัยประกอบการพิจารณา ได้แก่ นโยบายการให้สินเชื่อของธนาคาร วงเงินที่ขอกู้ วัตถุประสงค์ในการขอกู้ ความสามารถในการชำระหนี้ เงินทุนสินทรัพย์ และเงินฝากของผู้ขอสินเชื่อ เป็นต้น ทั้งนี้ การให้สินเชื่อทุกประเภทย่อมมีความเสี่ยงจากการที่ไม่ได้รับชำระเงินคืน และเกิดหนี้ที่ไม่ก่อรายได้ (Non-Performing Load: NPLs) ซึ่งส่งผลกระทบต่อสภาพคล่องในการดำเนินงานของธนาคาร ดังนั้น การบริหารสินเชื่อที่ดีมีส่วนช่วยลดความเสี่ยงในการเกิดหนี้ค้างชำระ งานวิจัยฉบับนี้ได้ศึกษาการนำเทคนิคเหมืองข้อมูลมาใช้เพื่อระบุปัจจัยสำคัญในการพิจารณาสินเชื่อ โดยพิจารณาจากข้อมูลผู้ที่ยื่นสินเชื่อโครงการธนาคารประชาชนจำนวน 14 ปัจจัย จากข้อมูลผู้ยื่นสินเชื่อทั้งหมด 1,000 ตัวอย่าง โดยมีวัตถุประสงค์เพื่อเปรียบเทียบกระบวนการอนุมัติสินเชื่อของธนาคารโดยไม่ใช้วิธีการคัดเลือกคุณลักษณะ และใช้วิธีการคัดเลือกคุณลักษณะ ซึ่งประกอบด้วยวิธี Chi-Square และวิธี Information Gain จากนั้นนำข้อมูลที่ได้ไปสร้างโมเดลด้วยวิธีซัพพอร์ทเว็กเตอร์แมชชีน (Support Vector Machine: SVM) โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron: MLP) และต้นไม้ตัดสินใจ (Decision Tree) เพื่อใช้สำหรับพิจารณาการได้สินเชื่อของธนาคาร ในการทดสอบได้เลือกใช้วิธี K-Fold Cross-Validation โดยกำหนดให้ K มีค่าเท่ากับ 10 จากการทดลองโดยไม่ใช้วิธีการคัดเลือกคุณลักษณะ พบว่าวิธีซัพพอร์ทเว็กเตอร์แมชชีนมีความถูกต้องสูงที่สุดที่ 97.12% จากนั้นทดสอบด้วยวิธีการคัดเลือกคุณลักษณะพิเศษ พบว่าวิธี Chi-Square คัดเลือกคุณลักษณะพิเศษจำนวนทั้งสิ้น 10 คุณลักษณะ และวิธี Information Gain คัดเลือกคุณลักษณะพิเศษจำนวนเพียง 3 คุณลักษณะ จากนั้นได้นำข้อมูลไปสร้างโมเดลด้วยวิธีซัพพอร์ทเว็กเตอร์แมชชีน โครงข่ายประสาทเทียมแบบหลายชั้น และต้นไม้ตัดสินใจ จากการทดลองพบกว่าวิธี Chi-Square ร่วมกับวิธีโครงข่ายประสาทเทียมแบบหลายชั้นมีความถูกต้อง 90.40% และวิธี Information Gain ร่วมกับวิธีโครงข่ายประสาทเทียมแบบหลายชั้นมีความถูกต้อง 91.70% จากการทดลองจึงสรุปได้ว่าการพิจารณาเพื่ออนุมัติสินเชื่อของธนาคารด้วยวิธีซัพพอร์ทเว็กเตอร์แมชชีนมีความถูกต้องสูงที่สุด

และไม่จำเป็นต้องใช้กระบวนการคัดเลือกคุณลักษณะพิเศษ

คำสำคัญ : ซัพพอร์ตเวกเตอร์แมชชีน, แบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น, ต้นไม้ตัดสินใจ, การคัดเลือกคุณลักษณะ



TITLE Data Mining Approaches for Loan Decision of a Bank

AUTHOR Kuannabha Pimcharee

ADVISORS Assistant Professor Olarik Surinta , Ph.D.

DEGREE Master of Science **MAJOR** Information Technology

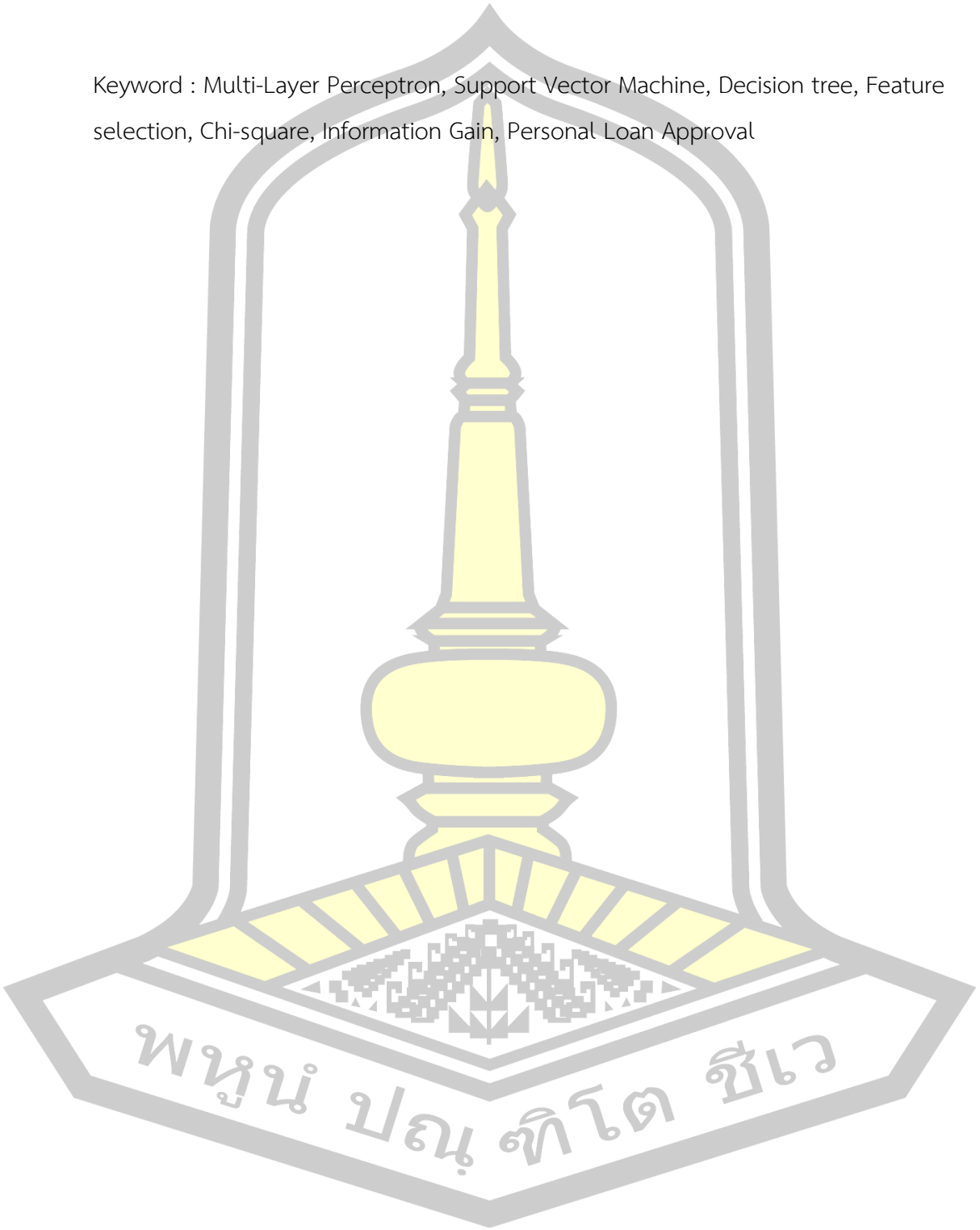
UNIVERSITY Mahasarakham University **YEAR** 2021

ABSTRACT

The approval of a bank's credit to an individual requires several factors to be considered, such as bank credit policy, loan amount, the purpose of the loan, ability to pay a debt, asset capital loan applicant's deposit. However, every type of credit is subject to the risk of non-repayment and non-performing loans (NLPs), which affect the liquidity of the banking operation. Therefore, good credit management aids decrease the risk of incurring debt. In this research, we studied the application of data mining techniques to identify key factors for the loan decision of a bank. We considered the 14 factors from 1,000 load applicants collected from the people who submitted the loan application to the bank. The main objective is to compare the data mining process of personal loan approval with and without feature selection techniques. The feature selection techniques including, Chi-square and information gain. In the evaluation process, we performed three models with K-fold cross-validation using k=10. For the experiments, we first created the data mining models using three methods, including support vector machine (SVM), multi-layer perceptron (MLP), and decision tree. The results showed that the SVM method outperformed other data mining models. The accuracy of the SVM method was obtained at 97.12%. Second, we experimented with feature selection techniques consisting of Chi-square and information gain. The Chi-square considered the ten factors, while information gain selected the best three factors. The experimental result showed that the Chi-square and information gain combined with the MLP method obtained an accuracy of 90.40% and 91.70%, respectively. Therefore, we concluded that the SVM classifier without combining the feature selection method was the best method

to approve the personal loan.

Keyword : Multi-Layer Perceptron, Support Vector Machine, Decision tree, Feature selection, Chi-square, Information Gain, Personal Loan Approval



กิตติกรรมประกาศ

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.โอฬาริก สุรินต๊ะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่เสียสละเวลาช่วยเหลือให้คำปรึกษา คำแนะนำ แนวคิด ชี้แนะข้อบกพร่องและร่วมแก้ไขปัญหาติดตามความก้าวหน้าของงานวิจัย รวมทั้งฝึกฝนให้ผู้วิจัยมีทักษะทางด้านการคิด การอ่าน การเขียนและการนำเสนอผลงานทางวิชาการ ซึ่งเป็นประโยชน์อย่างมากในการพัฒนาตนเอง อีกทั้งให้ความเอาใจใส่ทำให้ผู้วิจัยสามารถดำเนินงานวิจัยจนประสบผลสำเร็จลุล่วงไปด้วยดี สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณบิดามารดา ครอบครัวของข้าพเจ้า ที่ได้ให้ชีวิตและโอกาสทางการศึกษา คอยเป็นกำลังใจและให้ความห่วงใยเสมอมา รวมทั้งเพื่อนร่วมงาน คุณครูและอาจารย์ทุกท่าน ที่กรุณาประสิทธิ์ประสาทวิชาความรู้อันเป็นประโยชน์แก่ผู้วิจัย คุณค่าและประโยชน์อันพึงมาจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

ขวัญนภา พิมพ์ชารีย์

พหูน ปณฺฑิโต ชีเว

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... ฉ

กิตติกรรมประกาศ ..... ซ

สารบัญ ..... ฌ

สารบัญรูปภาพ ..... ฎ

สารบัญตาราง ..... ฎ

บทที่ 1 บทนำ 1

1.1 หลักการและเหตุผล 1

1.2 วัตถุประสงค์ของการวิจัย 2

1.3 ความสำคัญของการวิจัย 2

1.4 ขอบเขตงานวิจัย 2

1.4.1 เก็บรวบรวมข้อมูล 2

1.4.2 เตรียมข้อมูลก่อนประมวลผล 3

1.5 นิยามคำศัพท์เฉพาะ 3

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 4

2.1 ทฤษฎีที่เกี่ยวข้อง 4

2.1.1 สินเชื่อโครงการธนาคารประชาชน 4

2.1.2 วิธีการคำนวณการให้กู้ยืมเงิน 6

2.1.2.1 กระบวนการคำนวณของธนาคารในการพิจารณาสินเชื่อ 6

2.1.2.2 กระบวนการของธนาคารในการพิจารณาสินเชื่อ 6

2.1.2.3 กระบวนการพิจารณาด้วยวิธีการทางด้านเหมืองข้อมูล 7

2.1.3 เหมืองข้อมูล 7
2.1.4 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) 10
2.1.5 โครงข่ายประสาทเทียม (Neural Network) 12
2.1.5.1 โครงสร้างโครงข่ายประสาทเทียม 12
2.1.5.2 หลักการของโครงข่ายประสาทเทียม 13
2.1.5.3 หลักการทำงาน 13
2.1.5.4 Back Propagation Algorithm 14
2.1.5.5 ข้อจำกัดของ Neural Network 15
2.1.6 เทคนิคต้นไม้ตัดสินใจ (Decision tree) 16
2.1.6.1 ส่วนประกอบของต้นไม้ตัดสินใจ 16
2.1.6.2 การสร้างต้นไม้ตัดสินใจ 17
2.1.6.3 การคำนวณค่า Information Gain ต้นไม้ตัดสินใจ 18
2.1.7 การคัดเลือกคุณลักษณะ (Feature Selection) 19
2.1.7.1 การทดสอบไคสแควร์ (Chi-Square) 19
2.1.7.2 การจัดลำดับคุณสมบัติแบบ Information Gain 19
2.1.8 การประเมินประสิทธิภาพของแบบจำลอง (Model Evaluation) 21
2.2 งานวิจัยที่เกี่ยวข้อง 21
บทที่ 3 วิธีการดำเนินการวิจัย 26
3.1 ขั้นตอนการดำเนินการวิจัย 26
3.2 การเก็บรวบรวมข้อมูล 27
3.3 ขั้นตอนก่อนประมวลผล 28
3.4 การสร้างแบบจำลอง 29
3.5 การประเมินประสิทธิภาพของแบบจำลอง 29
บทที่ 4 ผลการศึกษา 31

4.1 ข้อมูลที่ใช้ในการทดลอง 31
4.2 ผลการวิเคราะห์ชุดข้อมูล 32
4.3 ผลการวิเคราะห์คุณลักษณะหรือปัจจัยที่มีผลต่อการพิจารณาอนุมัติ 33
4.4 ผลการทดลองประสิทธิภาพของเทคนิค 36
บทที่ 5 สรุปผลการวิจัย ข้อเสนอแนะและงานวิจัยในอนาคต 38
5.1 สรุปผลการวิจัย 38
5.2 ข้อเสนอแนะและงานวิจัยในอนาคต 39
บรรณานุกรม 40
ประวัติผู้เขียน 45

พหุน ปณ ฑิโต ชีเว

สารบัญรูปภาพ

ภาพประกอบที่ 2. 1หลักการทำงานของ SVM [13] 11
ภาพประกอบที่ 2. 2 แสดงโมเดลของ Neural Network [17] 13
ภาพประกอบที่ 2. 3 โครงสร้าง Neural Network [17] 14
ภาพประกอบที่ 2. 4 รูปแบบ Back Propagation neural network [17] 15
ภาพประกอบที่ 2. 5 โครงข่ายประสาทเทียม Multilayer Perceptron แบบ 1 hidden layer [19] 16
ภาพประกอบที่ 2. 6 ตัวอย่างของต้นไม้ตัดสินใจ [22] 17
ภาพประกอบที่ 3.1 ขั้นตอนการดำเนินงานวิจัย 27
ภาพประกอบที่ 3.2 ตัวอย่างการเก็บข้อมูลลูกค้าที่ยื่นขอสินเชื่อ 28

พหูน ปณ ฑิโต ชีเว

สารบัญตาราง

ตารางที่ 4. 1 ค่าพารามิเตอร์ที่ดีที่สุดของ Multi-Layer Perceptron (MLP)..... 32
ตารางที่ 4.2 ค่าพารามิเตอร์ที่ดีที่สุดของซัพพอร์ทเวกเตอร์แมชชีน (SVM)..... 32
ตารางที่ 4. 3 ผลการเปรียบเทียบแบบจำลอง 3 เทคนิคด้วยวิธี Cross-Validation จาก 14 คุณลักษณะ..... 33
ตารางที่ 4. 4 น้ำหนักของแต่ละคุณลักษณะเมื่อใช้วิธีการ Chi Square..... 34
ตารางที่ 4.5 น้ำหนักของแต่ละคุณลักษณะที่มีผลต่อการพิจารณาสินเชื่อเมื่อใช้วิธีการ Chi Square34
ตารางที่ 4.6 ผลการทดสอบตามแบบจำลอง 3 เทคนิค โดยใช้วิธีการ Chi Square..... 35
ตารางที่ 4.7 น้ำหนักของแต่ละคุณลักษณะที่มีผลต่อการพิจารณาสินเชื่อเมื่อใช้วิธีการ Information Gain..... 35
ตารางที่ 4.8 ผลการทดสอบตามแบบจำลอง 3 เทคนิค โดยใช้วิธีการ Information Gain 36

พหุน ปณ ฑิโต ชีเว

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

สินเชื่อโครงการธนาคารประชาชนเป็นโครงการที่จัดตั้งขึ้นตามนโยบายของรัฐบาลเพื่อกระจายโอกาสในการเข้าถึงแหล่งเงินทุนของลูกค้ารายย่อย ซึ่งมีวัตถุประสงค์เพื่อตอบสนอง นโยบาย ของรัฐบาลและสอดคล้องกับวิสัยทัศน์ของธนาคาร คือเป็นสถาบันการเงินที่มั่นคงเพื่อการออมการ พัฒนาเศรษฐกิจและสังคมของประเทศโดยเฉพาะเศรษฐกิจฐานราก เพื่อพื้นฐานและความมั่นคง แก่ผู้ประกอบการรายย่อยและประชาชนทั่วไป [1]

ธนาคารได้ตระหนักถึงความสำคัญที่จะทำหน้าที่ส่งเสริมคุณภาพชีวิตของประชาชนและส่งเสริมให้ประชาชนผู้ด้อยโอกาสสามารถเข้าถึงแหล่งเงินทุนของสถาบันการเงินเพื่อทดแทนการ กู้เงินนอกระบบซึ่งถือเป็นรากฐานที่สำคัญของเศรษฐกิจขั้นพื้นฐานของประเทศ ด้วยเหตุนี้ธนาคาร จึงได้เปิดดำเนินการ โครงการธนาคารประชาชนขึ้นเมื่อวันที่ 25 มิถุนายน 2544 จนถึงปัจจุบัน [2] สาขาทั่วประเทศของธนาคารได้ดำเนินการปล่อยสินเชื่อโครงการธนาคารประชาชนตามนโยบายรัฐ

ในการพิจารณาเพื่อขออนุมัติสินเชื่อแต่ละประเภทนั้น ธนาคารได้อาศัยนโยบายและ หลักเกณฑ์ของผู้ใช้สินเชื่อแต่ละรายโดยทั่วไปแล้วจะมีปัจจัยหลักที่ใช้ประกอบการพิจารณา คือ นโยบายการให้สินเชื่อของสินเชื่อแต่ละประเภท เช่น ผู้ให้สินเชื่อบางรายอาจะกำหนดว่าผู้ยื่น ขอสินเชื่อต้องประกอบธุรกิจหรือมีอายุงานตั้งแต่ 1 ปีขึ้นไป วงเงินที่ขอกู้มีความจำเป็น สอดคล้องกับวัตถุประสงค์ในการขอสินเชื่อ เช่น เพื่อซื้อที่ดินพร้อมสิ่งปลูกสร้างเพื่ออยู่อาศัย คุณลักษณะและความสามารถในการชำระหนี้ของผู้ขอสินเชื่อ สถาบันการเงินส่วนใหญ่จะใช้หลัก 5Cs ซึ่งประกอบไปด้วย คุณลักษณะและความน่าเชื่อถือของผู้ขอสินเชื่อ ความสามารถ ในการชำระหนี้คืนได้ตามระยะเวลาที่กำหนด เงินทุน สินทรัพย์ หรือเงินฝากของผู้ขอสินเชื่อ ผู้ค้ำประกันหรือหลักประกันที่ผู้ขอสินเชื่อนำมาจำนำหรือจำนองเพื่อให้สถาบันการเงินมีความ มั่นใจและลดความเสี่ยง และปัจจัยภายนอกที่มีผลกระทบต่อรายได้ของผู้ขอสินเชื่อ [3] ซึ่งจากปัจจัยที่กล่าวมาข้างต้น ธนาคารได้มีการเก็บข้อมูลต่างที่เกี่ยวข้องกับผู้ขอสินเชื่อ เพื่อ รวบรวมข้อมูลประกอบการพิจารณาความน่าเชื่อถือและปัจจัยอื่นๆ ที่เกี่ยวข้องกับธุรกิจ และรายได้ของผู้ขอสินเชื่อเพื่อสอดคล้องกับวัตถุประสงค์และหลักการดังกล่าว

ซึ่งการให้สินเชื่อทุกประเภทนั้นย่อมมีความเสี่ยงจากการที่ไม่ได้รับชำระเงินคืนหรือ เกิดหนี้ที่ไม่ก่อรายได้ขึ้น (Non-Performing Loan: NPLs) [4]

ซึ่งขณะนี้ทุกธนาคารต้องกันเงินสำรองเพื่อตั้งหนี้สงสัยจะสูญตามเกณฑ์ของธนาคารแห่งประเทศไทย ซึ่งส่งผลกระทบต่อสภาพคล่องในการดำเนินงานของธนาคาร จากข้อมูล สินเชื่อคงเหลือประเภทสินเชื่อตามนโยบายรัฐธนาคาร ณ 31 ธันวาคม 2561 จำนวน 3,136 บัญชี ยอดเงิน 206.10

ล้านบาท เป็นหนี้ที่ไม่ก่อให้เกิดรายได้ 13.76 ล้านบาท คิดเป็น 6.68% [5] อย่างไรก็ตามการบริหารสินเชื่อที่ดีมีส่วนช่วยลดความเสี่ยงในการเกิดหนี้ค้างชำระโดยเริ่มตั้งแต่ การวิเคราะห์เพื่ออนุมัติเงินกู้ที่มีประสิทธิภาพ และดูแนวโน้มความสามารถในการชำระหนี้ของ ลูกค้าในอนาคตและครอบคลุมไปจนถึงการวิเคราะห์เพื่อพิจารณาประนอมหนี้ให้แก่ลูกค้า การบริหารสินเชื่อที่ดีต้องอาศัยข้อมูลทางการบัญชีเป็นเครื่องมือที่สำคัญในการวิเคราะห์ฐานะ การเงิน ผลการดำเนินงาน สภาพคล่องของกิจการ ประเมินผลการดำเนินทั้งในอดีตและปัจจุบัน รวมทั้งการวางแผนในอนาคต เพื่อใช้ประกอบการวิเคราะห์ อนุมัติสินเชื่อ ติดตามฐานะของลูกหนี้ เพื่อดูแนวโน้มความสามารถในการชำระหนี้ว่าอาจจะเกิดการค้างชำระได้หรือไม่ และช่วย ในการตัดสินใจพิจารณาอนุมัติการแก้ไขหนี้ค้างชำระให้เหมาะสมมีประสิทธิภาพ

จากที่กล่าวมาข้างต้น ผู้วิจัยได้ตระหนักถึงปัจจัยต่าง ๆ ที่มีผลต่อการพิจารณาอนุมัติสินเชื่อเพื่อลดขั้นตอนและลดระยะเวลาในการปฏิบัติงาน จึงได้นำกระบวนการเหมืองข้อมูล (Data Mining) มาช่วยในการพิจารณาการให้สินเชื่อโครงการธนาคารประชาชน

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้ในการศึกษาปัจจัยที่ส่งผลต่อการพิจารณาอนุมัติสินเชื่อโครงการธนาคารประชาชน
2. เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในการให้สินเชื่อด้วยกระบวนการเหมืองข้อมูล

1.3 ความสำคัญของการวิจัย

เพื่อศึกษาปัจจัยที่มีผลต่อการอนุมัติสินเชื่อ และเปรียบเทียบประสิทธิภาพของแบบจำลองด้านเหมืองข้อมูล

1.4 ขอบเขตงานวิจัย

ขอบเขตของการวิจัยมีดังต่อไปนี้

1.4.1 เก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลของกลุ่มลูกค้าที่ขอสินเชื่อโครงการธนาคารประชาชน ที่กรอกข้อมูลผ่านระบบพิจารณาสินเชื่อของธนาคาร โดยข้อมูลจะถูกนำมาแบ่งออกเป็น 2 กลุ่ม ประกอบด้วย กลุ่มที่มีความสามารถในการชำระหนี้ (กลุ่มที่ผ่านการพิจารณาอนุมัติ) และกลุ่มที่ไม่มีความสามารถในการชำระหนี้ (กลุ่มที่ไม่ผ่านการพิจารณาอนุมัติ) โดยใช้ข้อมูลจำนวนทั้งสิ้น 1,000 รายการ

1.4.2 เตรียมข้อมูลก่อนประมวลผล

1) การแบ่งข้อมูล

งานวิจัยฉบับนี้ ใช้วิธี Cross Validation เพื่อแบ่งข้อมูลออกเป็นกลุ่ม (k) โดยกำหนดให้ K=10

2) การสร้างแบบจำลองด้วยวิธีการเหมืองข้อมูล

งานวิจัยฉบับนี้ได้เลือกข้อมูลลูกค้าของธนาคารที่ยื่นขอสินเชื่อจากธนาคารมาสร้างแบบจำลอง ซึ่งประกอบด้วย

อายุ
เพศ
อาชีพ
สถานภาพ
รายได้หลัก
รายได้เสริมและรายได้อื่น ๆ
ค่าใช้จ่าย
รายได้สุทธิ
ค่าใช้จ่ายอุปโภคและบริโภค
ภาระหนี้
หนี้ครั้งนี้, เงินงวด
วงเงินกู้
จำนวนงวดที่จะผ่อนชำระ
จำนวนหนี้ในการยื่นกู้ครั้งนี้
ผลการอนุมัติ

ข้อมูลที่นำมาศึกษาเป็นข้อมูล ณ วันที่ 1 มกราคม พ.ศ. 2561 – 31 ธันวาคม พ.ศ. 2561 ซึ่งข้อมูลที่นำมาทำวิจัยนี้ไม่มีข้อมูล ชื่อ-นามสกุล, ที่อยู่, เลขที่บัตรประชาชน, หมายเลขโทรศัพท์ จึงไม่สามารถอ้างอิงไปยังตัวบุคคลได้

1.5 นิยามคำศัพท์เฉพาะ

1.5.1 ลูกค้า หมายถึง ลูกค้าทั่วไปหรือบุคคลที่ยื่นขอสินเชื่อโครงการธนาคารประชาชน

1.5.2 ข้อมูลส่วนบุคคล หมายถึง ข้อมูลทั่วไปเบื้องต้นของลูกค้าที่มายื่นเอกสารเพื่อขอสินเชื่อ

1.5.3 สินเชื่อ หมายถึง สินเชื่อโครงการธนาคารประชาชน

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 สินเชื่อโครงการธนาคารประชาชน

ธนาคารในฐานะเป็นสถาบันการเงินเฉพาะกิจของรัฐเป็นเครื่องมือสำคัญในการดำเนินนโยบายการคลังเพื่อกระตุ้นเศรษฐกิจให้กับรัฐบาลทุกยุคทุกสมัย ธนาคารได้ขยายบทบาท ในการส่งเสริมเศรษฐกิจในระดับฐานราก โดยการกำหนดวิสัยทัศน์เพื่อมุ่งสู่ “การเป็นสถาบัน การเงินที่มั่นคง เพื่อการออม การพัฒนาเศรษฐกิจและสังคมของประเทศไทย โดยเฉพาะเศรษฐกิจ” ตามรายงานประจำปี 2551 [6]

ทำให้ธนาคารต้องตระหนักถึงความสำคัญที่จะทำหน้าที่ส่งเสริมคุณภาพชีวิตของประชาชน และส่งเสริมให้ประชาชนผู้ด้อยโอกาสสามารถเข้าถึงแหล่งเงินทุนของสถาบันการเงินเพื่อทดแทน การกู้เงินนอกระบบ ซึ่งถือเป็นรากฐานที่สำคัญของเศรษฐกิจขั้นพื้นฐานของประเทศ ด้วยเหตุนี้ ธนาคารจึงได้เปิดดำเนินการโครงการธนาคารประชาชนขึ้น เมื่อวันที่ 25 มิถุนายน 2544 จนถึงปัจจุบัน

เงื่อนไขการรับบริการสินเชื่อประชาชน

1. คุณสมบัติผู้กู้

1.1 เป็นผู้ประกอบการรายย่อย เช่น ผู้ประกอบอาชีพค้าขาย บริการ และผลิตสินค้าอุปโภคบริโภค เพื่อจำหน่าย

1.2 ผู้มีรายได้ประจำ เช่น ข้าราชการ ทหาร ตำรวจ ลูกจ้างของหน่วยงานราชการ รัฐวิสาหกิจ หรือบริษัทเอกชน ที่มีเงินเดือนประจำ

1.3 มีอายุครบ 20 ปีบริบูรณ์ขึ้นไป และเมื่อรวมอายุผู้กู้ กับระยะเวลาที่ชำระเงินต้อง ไม่เกิน 65 ปีบริบูรณ์ ในกรณีที่ไม่เป็นไปตามหลักเกณฑ์ เงื่อนไขดังกล่าวให้ อยู่ในดุลยพินิจของคณะกรรมการสินเชื่อสาขา โดยความเห็นชอบจากผู้จัดการเขต ส่วนกรณีที่ผู้กู้อายุไม่ครบ 20 ปีบริบูรณ์ ต้องให้บิดา-มารดา หรือผู้ปกครองเป็นผู้แทน

1.4 เป็นผู้มีถิ่นที่อยู่แน่นอน สามารถติดต่อได้

1.5 มีสถานที่ประกอบอาชีพแน่นอน

1.6 เปิดบัญชีเงินฝากประเภทเผื่อเรียก ณ สาขาที่ยื่นขอกู้เงิน

2. วัตถุประสงค์การกู้

2.1 เพื่อเป็นเงินทุน

2.2 เป็นเงินทุนหมุนเวียนในการประกอบอาชีพ

2.3 เป็นค่าใช้จ่ายที่จำเป็นต่อการดำรงชีพ และ

2.4 ชำระหนี้สินอื่น ๆ

3. จำนวนเงินให้กู้ ให้กู้ตามความจำเป็นและตามความสามารถในการชำระคืน ทั้งนี้ ไม่เกินรายละ 200,000 บาท

4. ระยะเวลาชำระคืนเงินกู้ ระยะเวลาชำระคืนเงินกู้และดอกเบี้ยไม่เกิน 8 ปี (96 งวด)

5. วิธีการคำนวณการชำระคืนเงินกู้

5.1 ชำระเงินต้นและดอกเบี้ยเป็นงวดรายเดือน และ

5.2 การคำนวณชำระหนี้แบบอัตราดอกเบี้ยคงที่ (Flat rate)

6. อัตราดอกเบี้ยเงินกู้ อัตราดอกเบี้ยและการคิดอัตราดอกเบี้ยกรณีผิดนัด ไม่ชำระหนี้ ให้เป็นไปตามประกาศธนาคาร

7. หลักประกันการกู้เงิน สามารถใช้หลักประกันประเภทหนึ่งประเภทใด ดังนี้

7.1 กรณีใช้บุคคลค้ำประกัน ผู้ค้ำประกันต้องมีคุณสมบัติ ดังนี้

7.1.1 มีอายุครบ 20 ปีบริบูรณ์ขึ้นไป และเมื่อรวมอายุผู้ค้ำประกันกับระยะเวลาที่ ต้องชำระเงินกู้ต้องไม่เกิน 65 ปี

7.1.2 เป็นผู้มีถิ่นที่อยู่แน่นอน สามารถติดต่อได้

7.1.3 ค้ำประกันผู้กู้ได้ไม่เกิน 2 คน

7.1.4 หลักเกณฑ์และเงื่อนไข

(1) วงเงินกู้ไม่เกิน 50,000 บาท เป็นบุคคลที่ธนาคารให้ความเชื่อถือ หรือเป็นผู้มีอาชีพและรายได้แน่นอน ตั้งแต่เดือนละ 7,000 บาทขึ้นไป ไม่น้อยกว่า 1 คน

(2) วงเงินกู้ตั้งแต่ 50,001-200,000 บาท

(2.1) เป็นบุคคลที่ธนาคารให้ความเชื่อถือ หรือเป็นผู้มีอาชีพ และรายได้แน่นอนตั้งแต่เดือนละ 7,000 บาทขึ้นไป ไม่น้อยกว่า 2 คน

(2.2) เป็นผู้มีอาชีพและรายได้แน่นอน ตั้งแต่เดือนละ 15,000 บาท ขึ้นไปไม่น้อยกว่า 1 คน *การให้บุคคลค้ำประกันผู้กู้เกินกว่า 2 คน ให้อยู่ในดุลยพินิจของ คณะ กรรมการ สินเชื่อสาขา โดยพิจารณาจากรายได้และความน่าเชื่อถือ

7.2 หลักประกันประเภทอื่น สามารถใช้หลักประกันที่เป็นของผู้กู้หรือของบุคคลอื่น ที่ยินยอมให้ใช้ค้ำประกันได้ ดังนี้

7.2.1 สมุดเงินฝากและหรือสลากพิเศษ ให้กู้ได้ไม่เกินร้อยละ 95 ของจำนวน เงินฝากคงเหลือในสมุดฝากเงิน หรือให้กู้ได้ไม่เกินร้อยละ 95 ของมูลค่าสลากพิเศษ

7.2.2 อสังหาริมทรัพย์ซึ่งตั้งอยู่ในแหล่งชุมชนที่มีความเจริญ มีไฟฟ้า สาธารณูปโภคอื่น ๆ ตามความจำเป็น และมีทางสาธารณประโยชน์ซึ่งสามารถ เข้า-ออกได้สะดวก โดยให้กู้ได้ ดังนี้

(1) ไม่เกินร้อยละ 80 ของราคาประเมินที่ดินพร้อมสิ่งปลูกสร้าง

(2) ไม่เกินร้อยละ 70 ของราคาประเมินที่ดินว่างเปล่าหรือห้องชุด

(3) ไม่เกินร้อยละ 60 ของราคาประเมินที่ดิน ที่เป็นที่สวน ที่ไร่ ที่นา

8. เงื่อนไขอื่น ๆ กรณีหน่วยงานรวบรวมนำส่งชำระหนี้เป็นงวดรายเดือน ธนาคารจะจ่าย ค่าตอบแทนให้แก่หน่วยงานภาครัฐ/รัฐวิสาหกิจและเอกชน ร้อยละ 1.00ของเงินต้นที่นำส่งธนาคารในการพิจารณาการให้สินเชื่อแก่ผู้กู้คือจะมีคณะกรรมการสินเชื่อสาขาในการพิจารณา อนุมัติ 3 คน ประกอบด้วย

1. ผู้จัดการสาขา
2. ผู้ช่วยผู้จัดการสาขา และ
3. พนักงานสินเชื่อสาขา

2.1.2 วิธีการคำนวณการให้กู้ยืมเงิน

2.1.2.1 กระบวนการคำนวณของธนาคารในการพิจารณาสินเชื่อ

ธนาคารจะนำข้อมูลของลูกค้ามาคำนวณ ดังต่อไปนี้

รายได้สุทธิ

รายได้สุทธิ = รายได้หลัก + รายได้เสริมและรายได้อื่น ๆ (1)

ภาระหนี้รวม

รวมภาระหนี้ = ภาระหนี้ + จำนวนหนี้ในการยื่นกู้ครั้งนี้ (2)

3. จำนวนหนี้ในการยื่นกู้ครั้งนี้ เนื่องจากสินเชื่อโครงการธนาคารประชาชนเป็นสินเชื่อที่ใช้อัตราดอกเบี้ยเงินกู้แบบเงินต้นคงที่ (Flat Rate) ซึ่งจะคำนวณดอกเบี้ยที่ลูกค้า ต้องชำระจากเงินต้นทั้งหมดที่คงที่ตลอดอายุสัญญา

ดอกเบี้ยที่ต้องชำระทั้งหมด = เงินต้น x อัตราดอกเบี้ยต่อปี x ระยะเวลา (ปี)

จำนวนหนี้ในการยื่นกู้ครั้งนี้ = $\frac{\text{เงินต้น + ดอกเบี้ยที่ต้องชำระทั้งหมด}}{\text{จำนวนงวดที่ต้องผ่อนชำระทั้งหมด}}$ (3)

2.1.2.2 กระบวนการของธนาคารในการพิจารณาสินเชื่อ

เกณฑ์ที่สถาบันการเงินใช้พิจารณาสินเชื่อ [3] เพื่อพิจารณาอนุมัติสินเชื่อขึ้นอยู่กับนโยบายและหลักเกณฑ์ของผู้ใช้สินเชื่อแต่ละรายโดยทั่วไป มีปัจจัยหลักที่ใช้ประกอบการพิจารณาคือ

นโยบายสินเชื่อของผู้ใช้สินเชื่อ

วัตถุประสงค์ในการขอกู้

คุณลักษณะและความสามารถในการชำระหนี้ของผู้ขอสินเชื่อ ซึ่งสถาบันการเงินส่วนใหญ่จะใช้หลัก 5 Cs ประกอบด้วย

1. Character คือ คุณลักษณะและความน่าเชื่อถือของผู้ขอสินเชื่อ วินัยในการใช้และการชำระสินเชื่อ ในอดีตซึ่งบอกถึงความสามารถในการใช้หนี้และการบริหารจัดการสินเชื่อ

2. Capacity คือความสามารถในการจ่ายชำระหนี้คืนได้ตามระยะเวลาที่กำหนด ไว้ รวมถึงความมั่นคงของรายได้ที่จะนำมาชำระหนี้ในอนาคต

3. Capital คือ เงินทุน สินทรัพย์ หรือเงินฝากของผู้ขอสินเชื่อ เพื่อเป็นหลักประกัน

4. Collateral คือ ผู้ค้ำประกันหรือหลักประกันที่ผู้ขอสินเชื่อนำมาจำนำ หรือจำนองเพื่อให้สถาบันการเงินมีความมั่นใจและลดความเสี่ยง

5. Conditions คือ ปัจจัยภายนอกที่มีผลกระทบต่อรายได้ของผู้ขอสินเชื่อ

จากการคำนวณสมการที่ (4) และ (5) สามารถแบ่งข้อมูลออกเป็น 2 กลุ่ม ประกอบด้วย

1. กลุ่มที่มีความสามารถในการชำระหนี้ (กลุ่มที่ผ่านการพิจารณาอนุมัติ)

2. กลุ่มที่ไม่มีความสามารถในการชำระหนี้ (กลุ่มที่ไม่ผ่านการพิจารณาอนุมัติ)

2.1.2.3 กระบวนการพิจารณาด้วยวิธีการทางด้านเหมืองข้อมูล

1) การแบ่งข้อมูล งานวิจัยฉบับนี้ ใช้วิธี Cross Validation เพื่อแบ่งข้อมูลออกเป็นกลุ่ม (k) โดยกำหนดให้ k=10 ดังนั้น ข้อมูลที่ใช้ใช้ในการเรียนรู้ (Training Data) และข้อมูล ที่ใช้ในการทดสอบ (Test Data) จึงถูกแบ่งออกเป็น 10 ส่วน และทำการทดสอบจำนวน 10 ครั้ง เพื่อหาค่าเฉลี่ยของความถูกต้อง และค่าเบี่ยงเบนมาตรฐาน [7]

2) การสร้างแบบจำลองด้วยวิธีการเหมืองข้อมูล งานวิจัยฉบับนี้ เลือกข้อมูลลูกค้าของธนาคารที่ยื่นขอสินเชื่อจากธนาคารมาสร้างแบบจำลองด้วยวิธีการเหมือง ข้อมูล ได้แก่ วิธีซัพพอร์ทเว็กเตอร์แมชชีน (Support Vector Machine: SVM) โครงข่ายประสาทเทียม แบบหลายชั้น (Multi-Layer Perceptron: MLP) และต้นไม้ตัดสินใจ (Decision Tree)

3) ข้อมูลลูกค้าธนาคารไปผ่านกระบวนการวิเคราะห์หาปัจจัยที่มีผลต่อการพิจารณาอนุมัติสินเชื่อโดยใช้วิธีChi-Square ซึ่งเป็นวิธีการทางสถิติและวิธี Information Gain เพื่อหาปัจจัย ที่มีผลต่อการพิจารณาได้สินเชื่อ ก่อนที่จะนำไปสร้างแบบจำลอง เพื่อใช้เปรียบเทียบหาวิธีการ ที่เหมาะสมเพื่อใช้ยื่นขอสินเชื่อจากธนาคาร

2.1.3 เหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น [8] โดยมีผู้ให้นิยามของการทำเหมือง ข้อมูลไว้หลายท่านเช่นการทำเหมืองข้อมูล คือ การวิเคราะห์ข้อมูลเพื่อหาความสัมพันธ์และสรุปผล จาก

ข้อมูลจำนวนมากเพื่อทำความเข้าใจและให้เกิดประโยชน์ต่อเจ้าของธุรกิจ หรือการทำเหมือง ข้อมูล คือ กระบวนการคัดเลือกและสำรวจข้อมูลตลอดจนเป็นการสร้างแบบจำลองข้อมูลเพื่อค้นหา รูปแบบและค้นหาความสัมพันธ์จากข้อมูลจำนวนมากเพื่อให้ได้ผลลัพธ์ที่เป็นประโยชน์ สรุปได้ว่า การทำเหมืองข้อมูลเป็นการนำเอาข้อมูลมาวิเคราะห์ เพื่อให้ได้ความรู้ใหม่ออกมา เพื่อนำมาช่วยในการตัดสินใจ

1. กระบวนการทำเหมืองข้อมูล

ในปัจจุบันความก้าวหน้าทางเทคโนโลยีสารสนเทศ ทำให้มีการจัดเก็บข้อมูลเป็นจำนวนมาก เพื่อช่วยในการตัดสินใจและดำเนินงานในหน่วยงาน อย่างไรก็ดี การใช้ข้อมูลโดย ส่วนใหญ่ ยังอยู่ในลักษณะที่ดึงข้อมูลจากฐานข้อมูลมาใช้งาน การวิเคราะห์ข้อมูล เพื่อให้ได้ความรู้ ที่เกี่ยวกับข้อมูล ได้แก่ ตัวแบบ (Pattern) ที่แสดงความสัมพันธ์ระหว่างข้อมูล ยังมีการทำอยู่ในวงจำกัด ความรู้ เหล่านี้สามารถนำมาใช้ประโยชน์ในการดำเนินงานและการตัดสินใจขององค์กร การทำเหมืองข้อมูลที่สำคัญ ได้แก่ การคัดเลือกข้อมูล (Selection) การเตรียมข้อมูล (Preprocessing) การแปลงข้อมูล (Transformation) การวิเคราะห์และค้นหารูปแบบข้อมูล (Data Mining) และ การแปล/ประเมินผล การวิเคราะห์ข้อมูล (Interpretation/Evaluation)

การทำเหมืองข้อมูล

เป็นกระบวนการที่มีขั้นตอนในการปฏิบัติ ทั้งนี้เพื่อให้ได้ผลลัพธ์ คือ ความรู้ที่เกี่ยวกับข้อมูลที่ถูกต้อง สามารถนำไปใช้ในการตัดสินใจและดำเนินงานได้โดยไม่เกิดความผิดพลาด หรือ สร้างความเสียหายจากการนำไปใช้งาน ขบวนการทำเหมืองข้อมูล ประกอบด้วย ขั้นตอนต่าง ๆ คล้ายกับการพัฒนาซอฟท์แวร์ หรือการดำเนินโครงการทางเทคโนโลยีสารสนเทศ โดยได้มีการ กำหนดขั้นตอนการทำเหมืองข้อมูลให้เป็นมาตรฐานเพื่อให้สามารถนำไปปฏิบัติ ทำให้การทำเหมืองข้อมูล มีประสิทธิภาพและได้ผลลัพธ์ที่มีความถูกต้อง มาตรฐานนี้เรียกว่า Cross-Industry Standard Process for Data Mining (CRISP-DM) [9]

ประเภทการทำเหมืองข้อมูล

ประเภทของการทำเหมืองข้อมูล สามารถแบ่งออกได้เป็น 2 ประเภท [10] ดังนี้

1. การทำเหมืองข้อมูลแบบทำนาย (Predictive Mining) คือ การนำข้อมูลที่มีอยู่ มาใช้ใน การทำนายผลข้อมูลในอนาคตที่ไม่ทราบมาก่อน ซึ่งการสร้างแบบจำลองรูปแบบนี้ จะเน้น การแบ่งข้อมูลออกเป็นกลุ่มตามคุณสมบัติของข้อมูล ในกรณีที่ข้อมูลไม่ต่อเนื่องจะใช้เทคนิคการ จำแนกประเภทข้อมูล (Classification) และในกรณีที่ข้อมูลมีความต่อเนื่องจะใช้เทคนิคการถดถอย (Regression)

2. การทำเหมืองข้อมูลแบบพรรณนา (Descriptive Mining) คือ การนำข้อมูลที่มี อยู่มาศึกษาหาคำอธิบายหรือคุณสมบัติทั่ว ๆ ไปของข้อมูล เพื่อใช้เป็นแนวทางในการตัดสินใจ เช่น เทคนิคการหาความสัมพันธ์ (Association) หรือ เทคนิคการจัดกลุ่ม (Clustering)

3) เทคนิคการทำเหมืองข้อมูล

การแก้ปัญหาของงานชนิดต่าง ๆ โดยวิธีการ Data Mining ในแต่ละงานจะมี เทคนิคที่นำมาใช้ได้อย่างเหมาะสม โดยเทคนิค Data Mining นั้นมีหลายหลาก ส่วนใหญ่มาจาก ศาสตร์ AI (Artificial Intelligence) หรือศาสตร์อื่น ๆ แบ่งเป็น 3 เทคนิค ในการทำเหมืองข้อมูล [11] ดังนี้

1) Cluster Analytic คือการจัดกลุ่มข้อมูลซึ่งมีลักษณะคล้ายกับการแบ่งประเภท แต่จะไม่เหมือนกันโดยการแบ่งประเภทจะวิเคราะห์ข้อมูลตามต้นแบบ แต่สำหรับการแบ่งกลุ่มเป็น การวิเคราะห์โดยไม่พิจารณาจัดกลุ่มตามประเภทที่มีหรือที่รู้จัก แต่จะใช้ขั้นตอนวิธีการจัดกลุ่ม เพื่อค้นหากลุ่มที่สามารถยอมรับได้เพื่อจัดเข้ากลุ่ม กล่าวคือ กลุ่มของวัตถุมีการสร้างขึ้นโดย เปรียบเทียบวัตถุที่มีความเหมือนกันจัดเข้ากลุ่มเดียวกัน

2) Association Rule เป็นการค้นหากฎความสัมพันธ์ของข้อมูล โดยค้นหาความสัมพันธ์ของข้อมูลทั้งสองชุดหรือมากกว่าสองชุดขึ้นไปไว้ด้วยกันความสำคัญของกฎทำการวัด โดยใช้ข้อมูลสองตัวด้วยกันคือค่าสนับสนุน (Support) ซึ่งเป็นเปอร์เซ็นต์ของการดำเนินการที่กฎ สามารถนำไปใช้ หรือเป็นเปอร์เซ็นต์ของการดำเนินการที่กฎที่ใช้มีความถูกต้อง และข้อมูลตัวที่ สองที่นำมาใช้วัดคือค่าความมั่นใจ (Confidence) ซึ่งเป็นจำนวนของกรณีที่กฎถูกต้องโดยสัมพันธ์ กับจำนวนของกรณีที่กฎสามารถนำไปใช้ได้ ในการหากฎความสัมพันธ์นั้นจะมีขั้นตอน วิธีการหาหลายวิธีด้วยกัน แต่ขั้นตอนวิธีที่เป็นที่รู้จักและใช้อย่างแพร่หลายคือ อัลกอริทึม Apriori

3) Classification Analytic เป็นการจัดแบ่งประเภทของข้อมูล โดยหาชุดต้นแบบ หรือชุดของการทำงานที่อธิบายและแบ่งประเภทข้อมูล วัตถุประสงค์เพื่อให้สามารถใช้เป็น ต้นแบบทำนายประเภทของวัตถุหรือข้อมูลที่ไม่มีการระบุประเภทหรือชนิดของข้อมูล ซึ่งต้นแบบ สร้างจากการวิเคราะห์ชุดของข้อมูลฝึกสอน (Training Data) โดยอาจจะเป็นกลุ่มข้อมูลที่มีการระบุ ประเภทหรือกลุ่มเรียบร้อยแล้ว รูปแบบของต้นแบบแสดงได้หลายแบบเช่น Classification Rules, Decision Trees และ Neural Networks เป็นต้น

4) ประโยชน์ของการทำเหมืองข้อมูล

การทำเหมืองข้อมูลถูกนำมาใช้สนับสนุนการตัดสินใจ โดยการสร้างมูลค่า เพิ่มกับข้อมูลที่มีอยู่ โดยประโยชน์ที่แต่ละองค์กรได้รับจากการกระทำเหมืองข้อมูลสรุปได้ดังนี้

1. การแข่งขันทางการค้า ลูกค้าที่ดีมักจะเป็นที่ชื่นชอบของบริษัทคู่แข่งเช่นกัน บริษัทเหล่านั้นจะพยายามแย่งส่วนแบ่งการตลาดที่สามารถสร้างผลกำไรให้กับบริษัทได้ และ พยายามที่จะ

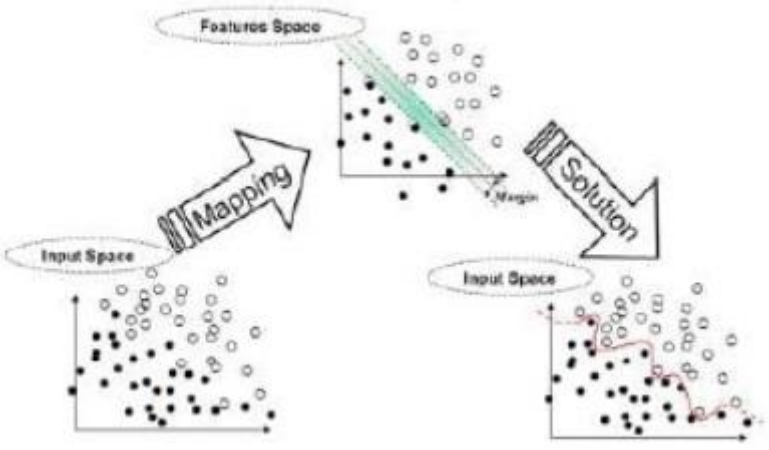
แย่งชิงส่วนแบ่งการตลาดนั้นมา การทำเหมืองข้อมูลสามารถนำมาใช้ประโยชน์ ได้ทั้งการแย่งส่วนแบ่งตลาด และในแง่การป้องกันมิให้เกิดการเสียส่วนแบ่งตลาด

2. ทำให้เกิดความรู้ที่สามารถนำมาใช้ หรือประกอบการตัดสินใจได้ เนื่องจาก การทำเหมืองข้อมูลจะใช้เทคนิคที่ซับซ้อน และมีลักษณะเป็นปัญญาประดิษฐ์ (Artificial Intelligence) ในการสร้างโมเดลที่อิงกับข้อมูล ซึ่งรวบรวมจากแหล่งต่าง ๆ เช่น รายการทางธุรกิจ ข้อมูลประวัติลูกค้า และข้อมูลอื่นๆ จากแหล่งภายนอก ความรู้ที่ได้จึงช่วยให้องค์กร สามารถ คาดการณ์อนาคต และสามารถเจาะกลุ่มตลาดได้ถูกต้องมากขึ้น เช่น ใช้ในการหาข้อผิดพลาด ของการปฏิบัติงานหรือการให้บริการได้ (Fraud Detection) โดยการทำให้ขั้นตอนการทำงาน มีประสิทธิภาพมากขึ้น ทำให้ประหยัดค่าใช้จ่าย และช่วยในการกำหนดเป้าหมายกลุ่มเป้าหมาย ได้อย่างมีประสิทธิภาพมากขึ้น ทำให้ยอดขายเพิ่มขึ้นจากจำนวนลูกค้า และลดโอกาสของ ความเสี่ยงต่าง ๆ

2.1.4 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) [12] เป็นตัวแบบที่ใช้ในการแยกแยะข้อมูล โดย SVM จะทำการแบ่งชั้นของข้อมูลด้วยระนาบหลายมิติ จากข้อมูล 2 กลุ่ม ชุดข้อมูล โดยตัวแบบของ SVM ที่เกี่ยวข้องกับโครงข่ายประสาทเทียม ตัวแบบของ SVM ใช้ Sigmoid kernel Function ซึ่งมีค่าเท่ากันทั้ง 2 ชั้น (Layer) ตัวแบบของ SVM มีความคล้ายคลึง กันกับเพอร์เซฟตรอน (Perception) ซึ่งเป็นข่ายงานประสาทเทียมแบบง่ายมีหน่วยเดียวที่จำลองลักษณะของเซลล์ประสาทด้วยการใช้ Kernel function ซึ่ง SVM จะใช้ข้อมูลของคุณสมบัติและ ตัวแปรที่เปลี่ยนแปลงใช้ในการกำหนดระนาบหลายมิติเรียกว่า โครงสร้าง (Feature) ส่วนการเลือก ที่มีความเหมาะสมที่สุด เรียกว่า โครงสร้างในการคัดเลือก (Feature selection) จำนวนเซตของโครงสร้าง ที่ใช้อธิบายในกรณีหนึ่ง (เช่น แถวของการค่าคาดการณ์) เรียกว่า เวกเตอร์ (Vector) ดังนั้นจุดมุ่งหมายของตัวแบบ SVM คือประโยชน์สูงสุดจากระนาบหลายมิติที่ แบ่งแยกกลุ่มของเวกเตอร์ที่ถูกเรียกว่า การหาเส้นแบ่ง Hyper-planes ซึ่งใช้แบ่งข้อมูลสองคลาสเพื่อให้ได้ผลลัพธ์ที่ดี โดยพิจารณาจากสมการเส้นตรง Hyper-planes และ SVMs จะทำการค้นหาเวกเตอร์ที่อยู่ใกล้ เส้นแบ่ง Hyper planes เรียกว่า ซัพพอร์ตเวกเตอร์ (Support Vectors) ดังภาพประกอบที่ 2.1

พหุน ปณฺฑิโต ชีเว



The SVM algorithm

ภาพประกอบที่ 2. 1 หลักการทำงานของ SVM[13]

ฟังก์ชันเคอร์เนล (Kernel Function) ในกรณีการจัดแบ่งกลุ่มข้อมูลโดย ใช้ระนาบแบบไม่เป็นเส้นตรง ซัพพอร์ตเวคเตอร์แมชชีนจะอาศัยหลักการของการแปลงข้อมูลจากพื้นที่ ข้อมูลนำเข้า (Input Space) ให้เป็นพื้นที่คุณลักษณะ (Feature Space) ที่มีมิติสูงขึ้น แสดงให้เห็นถึงแนวคิดของซัพพอร์ตเวคเตอร์แมชชีนซึ่งทำการแปลงข้อมูลที่ใช้ในการเรียนรู้แบบไม่เชิงเส้นไปเป็นขนาดพื้นที่คุณลักษณะที่ใหญ่ขึ้นผ่านฟังก์ชันเคอร์เนล (Kernel Function : Φ) และสร้างระนาบ ซึ่งแบ่งข้อมูลสองกลุ่มได้ดีที่สุด ทำให้เกิดเป็นขอบเขตการตัดสินใจ (Decision Surface) แบบไม่ เชิงเส้นในพื้นที่ข้อมูลนำเข้า ในขณะที่ซัพพอร์ตเวคเตอร์แมชชีนแบบเชิงเส้นจะสร้างระนาบใน พื้นที่คุณลักษณะที่ใหญ่ขึ้นภายใต้ทฤษฎีของ Mercer [14] ต้องการการคำนวณที่สิ้นเปลือง ในส่วน ของตัวอย่างเพื่อให้ได้ขนาดพื้นที่คุณลักษณะที่ใหญ่ขึ้น ปัญหาดังกล่าวสามารถแก้ไขได้โดยการ ใช้ฟังก์ชันเคอร์เนลเพื่อให้ได้ผลลัพธ์ที่น่าพึงพอใจเช่นเดียวกัน ซึ่งการใช้ฟังก์ชันเคอร์เนลจะทำให้สามารถคำนวณระนาบได้โดยไม่ต้องอาศัยการแปลงไปเป็นพื้นที่คุณลักษณะ

ฟังก์ชันเคอร์เนล K(xi , xj) เป็นฟังก์ชันที่แก้ปัญหาภายใต้เงื่อนไขของ Mercer’s ซึ่งมีค่าเท่ากับการคูณกันของสองเวคเตอร์ xi , xj ในพื้นที่คุณลักษณะ Φ(xi) และ Φ(xj)

$$K(xi , xj) = \Phi(xi) * \Phi(xj) \qquad (4)$$

โดยที่ Φ คือฟังก์ชันการแปลงแบบไม่เป็นเชิงเส้น (Nonlinear projection function) ซึ่งฟังก์ชันเคอร์เนล หลายตัวได้ถูกนำมาใช้กับซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เป็นเชิงเส้น อย่างประสบความสำเร็จ การฟังก์ชันเคอร์เนลที่แตกต่างกันของซัพพอร์ตเวกเตอร์แมชชีน สามารถนำมาซึ่งวิธีการเรียนรู้ที่หลากหลาย ซึ่งตัวอย่างของฟังก์ชันเคอร์เนล มีดังนี้ [15]

1. โพลิโนเมียลเคอร์เนล (Polynomial kernel)

K(xi , xj) = [(xi , xj) + 1]d (5)

2. เกาเซียเรเดียลเบสิสฟังก์ชัน (Gaussian Radial Basis Function)

K(xi , xj) = exp - $\left|\left|\frac{x_i - x_j}{2\sigma^2}\right|^2\right|$ (6)

3. แทนเจนต์เคอร์เนล (Tangent Kernel)

K(xi , xj) = tanh (v (xi * xj) + c) (7)

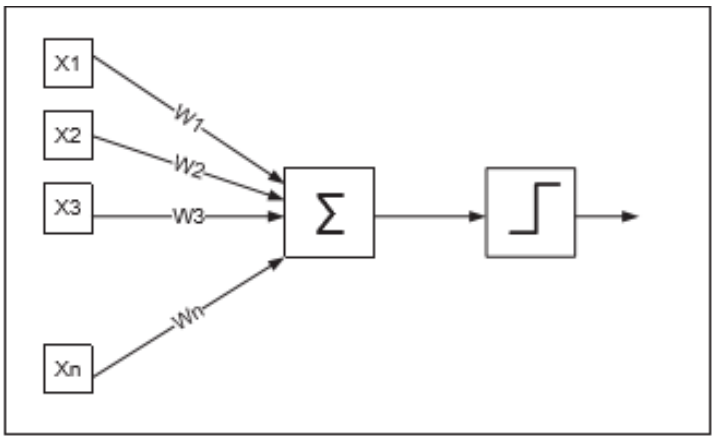
2.1.5 โครงข่ายประสาทเทียม (Neural Network)

การจำลองการทำงานของระบบประสาทเทียม ซึ่งเป็นการทำงานของระบบประสาท ในสมองของมนุษย์ การทำงานของ Neural Network แต่ละ Process จะรับ Input เข้าไปคำนวณ และสร้าง Output ออกมาในลักษณะที่ไม่ใช่เป็นการทำงานแบบเชิงเส้นตรง เพราะว่า Input แต่ละ ตัวจะถูกให้ลำดับความสำคัญของค่าไม่เท่ากัน ค่าของ Output ที่ได้จากการเชื่อมโยงกันนี้ จะถูก นำมาเปรียบเทียบกับ Output ที่ได้ตั้งเอาไว้ ถ้าค่าที่ออกมาเกิดความคลาดเคลื่อน ก็นำไปสู่การปรับค่าหรือน้ำหนักของค่าที่ใส่ไว้ให้แต่ละ Input [16]

Neural Network เป็นการสร้างแบบจำลองที่เลียนแบบการทำงานของสมองมนุษย์ มีโครงสร้างเป็นกลุ่มของ Node ที่เชื่อมโยงถึงกันในแต่ละ Layer คือ Input Layer, Hidden Layer และ Output Layer [16]

2.1.5.1 โครงสร้างโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมมีโครงสร้างแตกต่างจากข่ายงานของสมอง แต่ก็ยังเหมือนสมอง ในแง่ที่ว่าข่ายงานประสาทเทียม คือการรวมกลุ่มแบบขนานของหน่วย ประมวลผล ย่อย ๆ และการเชื่อมต่อนี้เป็นส่วนสำคัญที่ทำให้เกิดสติปัญญาของข่ายงาน เมื่อพิจารณาขนาดแล้ว สมองมีขนาดใหญ่กว่าข่ายงานประสาทเทียมอย่างมาก รวมทั้งเซลล์ประสาทยังมีความซับซ้อนกว่า หน่วยย่อยของ 2 ข่ายงาน อย่างไรก็ดีหน้าที่สำคัญของสมอง เช่น การเรียนรู้ยังคงสามารถถูกจำลอง ขึ้นอย่างง่ายด้วย โครงข่ายประสาทนี้



ภาพประกอบที่ 2. 2 แสดงโมเดลของ Neural Network [17]

2.1.5.2 หลักการของโครงข่ายประสาทเทียม

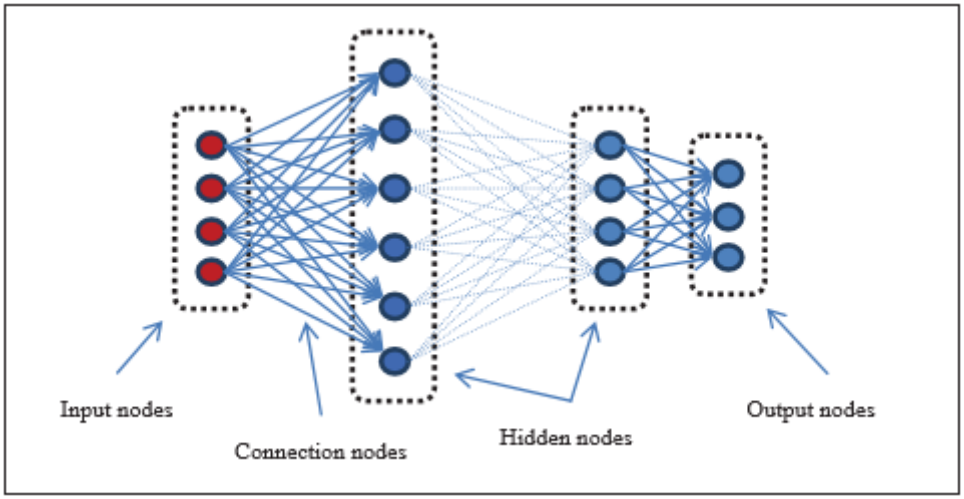
Neural Networks ประกอบด้วย input และ output เหมือนกัน โดยจำลองให้ input แต่ละอันมี weight เป็นตัวกำหนดน้ำหนักของ input โดย neuron แต่ละหน่วยจะมีค่า threshold เป็นตัวกำหนดว่าค่าน้ำหนักรวมของ input ต้องมากขนาดไหนจึงจะสามารถส่ง output ไปยังโหนดอื่นได้

2.1.5.3 หลักการทำงาน

การทำงานของ Neural networks คือเมื่อมี input เข้ามายัง network ก็เอา input มาคำนวณกับ weight ของแต่ละขา ผลที่ได้จาก input ทุก ๆ ขาของ neuron จะเอามารวมกัน แล้วก็เอามาเทียบกับ threshold ที่กำหนดไว้ ถ้าผลรวมมีค่ามากกว่า threshold แล้ว neuron ก็จะส่ง output ออกไป output นี้ ก็จะถูกส่งไปยัง input ของ neuron อื่น ๆ ที่เชื่อมกันใน network ถ้าค่า น้อยกว่า threshold ก็จะไม่เกิด output เขียนออกมาได้ดังนี้

If (sum (input * weight) > threshold) then output

สิ่งสำคัญคือต้องทราบค่า Weight และ threshold สำหรับสิ่งที่ต้องการ เพื่อให้คอมพิวเตอร์รู้จำ ซึ่งเป็นค่าที่ไม่แน่นอน แต่สามารถกำหนดให้คอมพิวเตอร์ปรับค่า เหล่านั้น ได้โดยการสอนให้มันรู้จัก Pattern ของสิ่งที่เราต้องการให้มันรู้จำ เรียกว่า “back propagation” ซึ่งเป็นกระบวนการย้อนกลับของการรู้จำในการฝึก feed-forward neural networks จะมีการใช้อัลกอริทึมแบบ back- propagation เพื่อใช้ในการปรับปรุงน้ำหนักคะแนนของเครือข่าย (network weight) หลังจากใส่รูปแบบข้อมูลสำหรับฝึกให้แก่เครือข่ายในแต่ละครั้งแล้ว ค่าที่ได้รับ (Output) จากเครือข่ายจะถูกนำไปเปรียบเทียบกับผลที่คาดหวังแล้ว ทำการคำนวณหาค่าความผิดพลาด ซึ่งค่าความผิดพลาดนี้จะถูกส่งกลับเข้าสู่เครือข่ายเพื่อใช้แก้ไขค่าน้ำหนักคะแนนต่อไป



ภาพประกอบที่ 2. 3 โครงสร้าง Neural Network [17]

Output ของแต่ละ Node

$$y_i = f\,(W_i^1 X_1 + W_i^2 X_2 + W_i^3 X_3 + \ldots\ldots + W_i^m X_m)$$

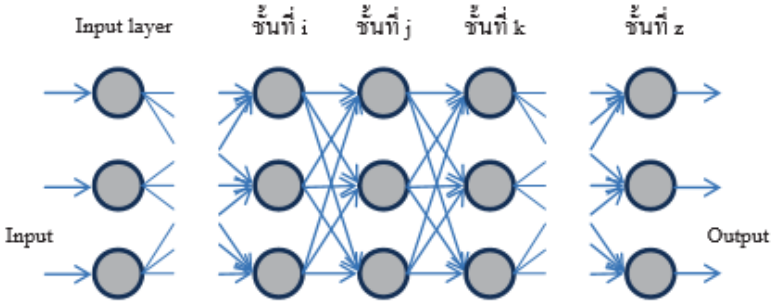
$$= f\,(\sum W_i^j X_j)$$

เมื่อX_i = input จากโหนดอื่นๆ

W_{ij} = น้ำหนัก (weight) ของแต่ละแขน (connection)

2.1.5.4 Back Propagation Algorithm

Back Propagation เป็นอัลกอริทึมที่ใช้ในการเรียนรู้ของเครือข่ายใยประสาทวิธีหนึ่ง ที่นิยมใช้ใน multilayer perception เพื่อปรับค่าน้ำหนักในเส้นเชื่อมต่อระหว่างโหนด ให้เหมาะสม โดยการปรับค่านี้จะขึ้นกับความแตกต่างของค่า Output ที่คำนวณได้กับค่า Output ที่ต้องการพิจารณา ดังภาพประกอบที่ 2.4



ภาพประกอบที่ 2. 4 รูปแบบ Back Propagation neural network [17]

ขั้นตอนของ Back Propagation Algorithm มีดังนี้

1. กำหนดค่าอัตราในการเรียนรู้ (rate parameter: r)

2. สำหรับแต่ละตัวอย่าง Input ให้ทำตามขั้นตอนต่อไปนี้จนกว่าได้ระดับ performance ที่ต้องการ

3. คำนวณหาค่า Output โดยใช้ค่าน้ำหนักเริ่มต้น ซึ่งอาจได้จากการสุ่ม

4. เพิ่มค่าน้ำหนักที่เปลี่ยนแปลง สำหรับตัวอย่าง Input ทั้งหมด และเปลี่ยนค่าน้ำหนัก Neural Network Taxonomy

2.1.5.5 ข้อจำกัดของ Neural Network

1. Neural Network รับข้อมูลได้เฉพาะข้อมูลตัวเลขที่อยู่ช่วง 0 – 1 เท่านั้น กรณีข้อมูลนำเข้ามีค่ามากกว่านี้ต้องทำการปรับลดค่าลง หรือในกรณีที่เป็นข้อมูลอื่น ที่ไม่ใช่ตัวเลขต้องทำการแปลงค่าก่อน

2. การสร้างแบบจำลองด้วย Neural Network นั้นจะไม่สามารถอธิบาย ได้ว่าผลลัพธ์ที่ได้นั้นมาจากไหน

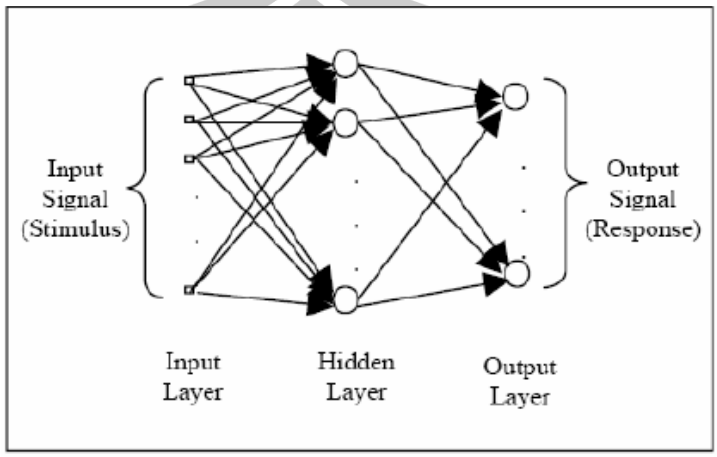
3. เนื่องจากการที่ไม่สามารถอธิบายผลลัพธ์ที่ได้มาได้ ดังนั้นการสร้างแบบจำลองด้วย Neural Network จะไม่สามารถรับรองได้เลยว่าเป็นแบบจำลองที่ดีหรือไม่ จนกว่าจะได้ทำการทดสอบกับข้อมูลทดสอบก่อนจนแน่ใจก่อน

โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron: MLP) [18] เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบชั้นใช้สำหรับงานที่มีความ ซับซ้อน ได้ผลเป็นอย่างดี โดยกระบวนการฝึกฝนเป็นแบบ Supervise และใช้ขั้นตอนการส่งค่า ย้อนกลับ (Backpropagation) สำหรับการฝึกฝนระบวนการส่งค่าย้อนกลับ ประกอบด้วย 2 ส่วนย่อยคือ

1. การส่งผ่านไปข้างหน้า (Forward Pass) เป็นการที่ข้อมูลจะส่งผ่านเข้าโครงข่าย ประสาทเทียมที่ชั้นของข้อมูลเข้า และจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่ง จนกระทั่งถึงชั้น ข้อมูลออก

2. การส่งผ่านย้อนกลับ (Backward Pass) เป็นการส่งผ่านข้อมูลย้อนกลับค่าน้ำหนัก การเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือ ผลต่าง ของผลตอบแทนที่แท้จริง (Actual Response) กับผลตอบแทนเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทาง

ตรงกันข้ามกับการเชื่อมต่อ ค่าน้ำหนักการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบแทนที่แท้จริง เข้าใกล้ผลตอบ เป้าหมาย ดังภาพประกอบที่ 2.5



ภาพประกอบที่ 2. 5 โครงข่ายประสาทเทียม Multilayer Perceptron แบบ 1 hidden layer [19]

2.1.6 เทคนิคต้นไม้ตัดสินใจ (Decision tree)

เทคนิคต้นไม้ตัดสินใจ คือ แบบจำลองทางคณิตศาสตร์เพื่อการหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ ข้อมูลแบบมีผู้สอน (Supervised Learning) สามารถสร้างแบบจำลองการจัดหมวดหมู่ ได้จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ล่วงหน้า (Training Set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วย [20]

โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” เช่น [20]

“If Income = High and Married” = No THEN Risk = Poor”

“If Income = High and Married” = Yes THEN Risk = Good”

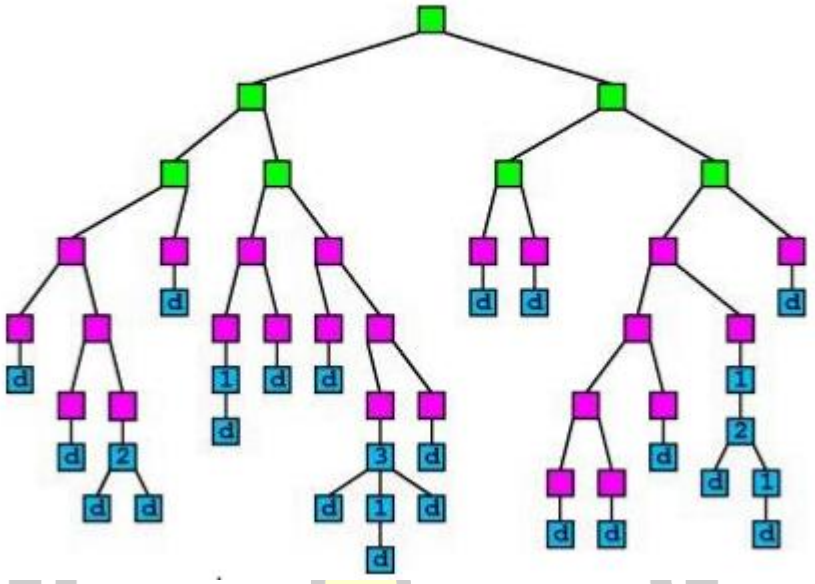
2.1.6.1 ส่วนประกอบของต้นไม้ตัดสินใจ

ประกอบด้วย [21]

1) โหนด (Node) คือคุณสมบัติต่างๆ เป็นจุดที่แยกข้อมูลว่าจะไปในทิศทาง ใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (Root Node)

2) กิ่ง (Branch) คือ คุณสมบัติของคุณสมบัติในโหนดที่แตกออกมา โดย จำนวนของกิ่งจะเท่ากับคุณสมบัติของโหนด

3) ใบ (Leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล โดยสามารถแสดงส่วนประกอบของต้นไม้ตัดสินใจ และแสดงตัวอย่างของต้นไม้ตัดสินใจได้ ดังภาพประกอบที่ 2.6



ภาพประกอบที่ 2. 6 ตัวอย่างของต้นไม้ตัดสินใจ [22]

2.1.6.2 การสร้างต้นไม้ตัดสินใจ

หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ เป็นการสร้างในลักษณะจากบนลงล่าง (Top-Down) คือเริ่มจากการสร้างราก ของต้นไม้ก่อน แล้วจึง แตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจดังนี้ [23]

1) ต้นไม้เริ่มต้นมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึก

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อ แยกตามกลุ่มของข้อมูลนั้น

3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าเกน (Gain) ของแต่ละแอททริบิวต์เพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือกแอททริบิวต์ ที่มีความสามารถในการ แบ่งแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุด โดยแอททริบิวต์ที่มีค่าเกนมากที่สุด จะถูกเลือก ให้เป็นตัวทดสอบ หรือแอททริบิวต์ใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ ถูกสร้างขึ้นจากค่าต่าง ๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

5) ทำการวนซ้ำหรือหาแอททริบิวต์ที่มีค่าเกนมากที่สุด สำหรับข้อมูลที่ถูก แบ่งออกมาในแต่ละกิ่ง เพื่อนำแอททริบิวต์นี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่แอททริบิวต์ ที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีก สำหรับโหนดในระดับต่อ ๆ ไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อย ๆ โดยการ วนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้เป็นจริง

2.1.6.3 การคำนวณค่า Information Gain ต้นไม้ตัดสินใจ

[24] เป็นโครงสร้างที่ใช้ แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล โดยต้นไม้ตัดสินใจจะมีลักษณะคล้าย โครงสร้างต้นไม้ ที่แต่ละโหนดแสดงแอททริบิวต์ ในการสร้างต้นไม้ตัดสินใจ ปัญหาสำคัญที่ต้อง พิจารณาคือ ควรจะตัดสินใจเลือกแอททริบิวต์ใดมาทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนของ การสร้างต้นไม้และต้นไม่ย่อย (Subtree) ของต้นไม้ตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือก แอททริบิวต์คือการคำนวณค่ามาตรฐาน (Gain Criterion) ซึ่งเป็นค่าที่บ่งบอกว่าแอททริบิวต์ที่ เป็นไปได้ จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าแอททริบิวต์ใดมีค่าเกนสูงที่สุด แสดงว่าแอททริบิวต์นั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่า Information Gain จะช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูล อีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจ ที่ได้ไม่มีความซับซ้อนมากจนเกินไป

ซึ่งค่า Information Gain นั้นสามารถคำนวณได้จากสมการ ดังสมการที่ (8)

$$I(S_1, S_2, \ldots, S_n) = -\sum \frac{s_1}{s} \log_2 \frac{s_1}{s} \qquad (8)$$

เมื่อ

S เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s เรคคอร์ด

n เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

C_i แทนกลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง n

S_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i

S_{ij} แทนจำนวนข้อมูลที่เป็นสมาชิกของ S ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของแอททริบิวต์ A

J ค่าระหว่าง 1 ถึง v

ค่าเอ็นโทรปีของแอททริบิวต์ A ซึ่งมีค่าของแอททริบิวต์เป็น $((a_1, a_2, a_3, \ldots, a_v))$ หาได้ในสมการที่ 9

$$E(A) = \sum \frac{s_{1j} + \ldots + s_{nj}}{s} I(s_{1j}, s_{2j}, \ldots, s_{nj}) \qquad (9)$$

ดังนั้นจะสามารถพิจารณาค่ามาตรฐานเกนได้ดังสมการที่ 10

$$\text{GAIN}(A) = I(S_{1j}, S_{2j}, \ldots\ldots, S_{\text{nj}}) - E(A) \quad (10)$$

2.1.7 การคัดเลือกคุณลักษณะ (Feature Selection)

การลดขนาดข้อมูลเป็นขั้นตอนหนึ่งที่จะต้องทำก่อนการเรียนรู้ด้วยเครื่องจักรการเรียนรู้ (Machine Learning) ซึ่งการลดคุณลักษณะมีหลากหลายวิธี ซึ่งอยู่บนพื้นฐานของการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรกับกลุ่มเป้าหมาย (Class) รวมถึงการสร้างคุณลักษณะใหม่ จากคุณลักษณะเดิม อาจจะนำคุณลักษณะพื้นฐานเหล่านี้มารวมกันเพื่อให้เป็นคุณลักษณะใหม่ [26]

2.1.7.1 การทดสอบไคสแควร์ (Chi-Square)

การประเมินค่าของคุณลักษณะ(Attribute) โดยการคำนวณค่า Chi-Square ทางสถิติ เพื่อแจกแจงความถี่หรือน้ำหนักของคุณลักษณะที่ได้ จากการศึกษาเพื่อประเมินค่าของคุณลักษณะ ดังแสดงในสมการ 11 [27]

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (11)$$

โดยที่O1,O2,..,Onป็นความถี่หรือน้ำหนักของคุณลักษณะที่ได้จากการศึกษา

E1, E2,.....,En เป็นความถี่น้ำหนักของคุณลักษณะที่คาดหวัง (หรือความถี่ที่ควรจะเป็น)

2.1.7.2 การจัดลำดับคุณสมบัติแบบ Information Gain

การจัดลำดับคุณสมบัติแบบ Information Gain มีหลักการพื้นฐานจากการ สุ่มตัวอย่าง (Entropy) โดยค่าที่ได้จาก Information Gain คือค่าของความต่างระหว่างตัวแปร x ที่เป็นตัวแปรเป้าหมาย (Target Variable) กับตัวแปรอิสระ A (Independent Variable) ซึ่งลักษณะ ของ Information Gain จะทำการลด Entropy ของตัวแปรเป้าหมาย x โดยการเรียนรู้จากสถานะของ ตัวแปรอิสระ A

Information Gain มีแนวโน้มที่จะเลือกคุณสมบัติ ที่มีจำนวนค่าที่แตกต่าง กันสูง มากกว่าคุณสมบัติที่มีจำนวนค่าที่แตกต่างกันต่ำ ถึงแม้ว่าคุณสมบัติที่มีจำนวนค่าที่แตกต่าง กันต่ำจะให้ข้อมูลที่เป็นประโยชน์

การคำนวณของเทคนิค Information Gain นั้นจะพิจารณาระหว่าง คุณสมบัติ X ซึ่งก็คือตัวแปรเป้าหมายกับคุณสมบัติที่เป็นคลาสของข้อมูล Y หรือตัวแปรอิสระ จากนั้นทำการดู

จาก ความน่าจะเป็นที่เกิดขึ้นระหว่างค่าของคุณสมบัติ X กับค่าของคุณสมบัติ Y ว่ามีมากน้อยเพียงใด หากมีความน่าจะเป็นเกิดขึ้นน้อย ก็จะทำการให้ คะแนนของคุณสมบัติ X น้อยตามไปด้วย โดยสามารถ เขียนเป็นสมการได้ดังนี้

$$\text{Gain}(Y;X) = H(Y) - H(Y \vee X) \quad (1)$$

โดย $H(Y)$ คือค่าความน่าจะป็นจากการสุ่มตัวอย่างของ Y

$H(Y \vee X)$ คือค่าความน่าจะป็นจากการสุ่มจัวอย่างของ Yเทียบกับ X

$\text{Gain}(Y;X)$ คือค่าของคะแนนจากการสุ่มตัวอย่างที่คำนวณได้โดยจะมีตั้งแต่ 0 ถึง 1

ค่าของ Y เป็นค่าคุณสมบัติที่เป็นคลาสของข้อมูลซึ่งคำนวณตั้งแต่ { $Y_1, Y_2, \ldots\ldots, Y_n$ }

ค่าของ X เป็นค่าคุณสมบัติอื่นๆ ที่ไม่ใช่คลาส ซึ่งคำนวณตั้งแต่ { $X_1, X_2, \ldots\ldots, X_n$ }

ค่าของ $H(Y)$และ $H(Y \vee X)$คำนวณมาจาก

$$H(Y) = -\sum P(Y = y_i) \log_2 P(Y = y_i) \quad (2)$$

$$H(Y \vee X) = -\sum P(X = x_i) \log_2 H(Y \vee X = x_i) \quad (3)$$

โดย $P(Y = y_i)$ คือ ค่าความน่าจะเป็นตั้งแต่ y_1 จนถึง y_k

$P(X = x_i)$ คือ ค่าความน่าจะเป็นตั้งแต่ x_1 จนถึง x_k

2.1.8 การประเมินประสิทธิภาพของแบบจำลอง (Model Evaluation)

เป็นการวัดประสิทธิภาพการทำนายหมวดหมู่ของลูกค้าของแบบจำลองด้วยเกณฑ์ทางด้านการค้นคืนสารสนเทศ คือ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) นอกจากนี้ ยังประเมินในส่วนของเวลาที่ใช้ในการสร้างแบบจำลองของ แต่ละเทคนิคด้วย สำหรับการคำนวณค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึก สามารถแสดงได้ดังสมการ ดังนี้

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

โดย ค่าของผลลัพธ์ที่ได้จากการจำแนกคือ

TP คือ จำนวนข้อมูลที่จำแนกลูกค้าที่มีความสามารถในการชำระหนี้ได้ถูกต้อง (True Positive)

TN คือ จำนวนข้อมูลที่จำแนกลูกค้าที่ไม่มีความสามารถในการชำระหนี้ได้ถูกต้อง (True Negative)

FP คือ จำนวนข้อมูลที่จำแนกลูกค้าที่มีความสามารถในการชำระหนี้ไม่ถูกต้อง (False Positive)

FN คือ จำนวนข้อมูลที่จำแนกลูกค้าที่ไม่มีความสามารถในการชำระหนี้ไม่ถูกต้อง (False Negative)

2.2 งานวิจัยที่เกี่ยวข้อง

ประเวช สุพรรณพยัคฆ์ [29] ได้ศึกษาเรื่อง การวิเคราะห์พฤติกรรมลูกค้าที่ก่อให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้ในธุรกิจสินเชื่อรถยนต์ โดยการทำเหมืองข้อมูล กรณีศึกษา บริษัท ลิสซิ่ง แห่งหนึ่ง ได้นำเทคนิคการทำเหมืองข้อมูลมาวิเคราะห์ (Data Mining) ด้วยวิธีการเกาะกลุ่มแบบ Simple K-mean ในการแบ่งกลุ่มหาลักษณะเด่นของพฤติกรรมลูกค้าและทำการตรวจสอบผลที่ได้จากการแบ่งกลุ่มโดยการสอบถามผู้เชี่ยวชาญด้านสินเชื่อรถยนต์ จากนั้นนำผลที่ได้มาหาความสัมพันธ์เพื่อหา

ลักษณะพฤตติกรรมที่ส่งผลให้เกิดหนี้ไม่ก่อให้เกิดรายได้ พบว่า กลุ่มของข้อมูลจะแบ่งออกเป็น 5 กลุ่ม ที่มีความแตกต่างกัน จากนั้นนำข้อมูลแต่ละกลุ่มมาทดสอบหาค่าเป้าหมายของกฎความสัมพันธ์ใน 2 ประเภทคือค่าเป้าหมายการถูกดำเนินคดีทางกฎหมายและค่าเป้าหมายหนี้ตามระดับยอดเงินค้างชำระ จากการศึกษาพบว่า ค่าเป้าหมายถูกดำเนินคดีทางกฎหมายและหนี้ตามระดับยอดเงินค้างชำระ ของกลุ่มข้อมูลที่ถูกแบ่งทั้ง 5 กลุ่ม จะมีลักษณะเด่นของกลุ่มข้อมูลและมีค่าความเชื่อมั่นดังนี้ กลุ่มที่ 1 อาชีพประกอบธุรกิจส่วนตัวและมีระดับการศึกษาต่ำ กลุ่มที่ 2 ค่างวดการผ่อนชำระต่ำและมียอดเงินการขอสินเชื่อต่ำ กลุ่มที่ 3 อาชีพพนักงานบริษัทเอกชนและมีอัตราเงินดาวนสูง กลุ่มที่ 4 อาชีพเกษตรกรและใช้ระยะเวลาผ่อนชำระนาน และกลุ่มที่ 5 รายได้สุทธิสูงและใช้ระยะเวลาผ่อนชำระสั้น โดยได้นำผลที่ได้เสนอเป็นแนวทางการนำไปใช้เพื่อการลดหนี้ไม่ก่อให้เกิดรายได้และลดค่าใช้จ่ายในขั้นตอนควบคุมสินเชื่อที่จะเกิดขึ้นลงอีกด้วย

เบญญา จันทพันธ์ [30] ได้ศึกษาการหาพฤติกรรมการยกเลิกการใช้ธุรกรรมทางการเงินผ่านโทรศัพท์เคลื่อนที่จากรายการบันทึกการทำธุรกรรมโดยการทำเหมืองข้อมูล โดยมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่สามารถสนับสนุนกิจกรรมทางการตลาด โดยใช้เทคนิคการทำเหมืองข้อมูล เพื่อหาพฤติกรรมของกลุ่มที่มีแนวโน้มยกเลิกการใช้บริการ ซึ่งศึกษาจากข้อมูลรายการธุรกรรมจากผู้ใช้บริการโทรศัพท์เคลื่อนที่แห่งหนึ่ง โดยขั้นตอนแรก นำข้อมูลธุรกรรมทั้งหมดมาทำการแบ่งหรือจัดกลุ่มออกเป็น 5 กลุ่มข้อมูล โดยใช้ระยะเวลาการใช้บริการเป็นตัวแบ่งกลุ่มข้อมูล เพื่อให้ง่ายต่อการนำผลลัพธ์ไปใช้งาน แบ่งกลุ่มข้อมูลออกเป็น 2 ส่วนคือ ข้อมูลเรียนรู้ และข้อมูลทดสอบ เพื่อใช้วัดประสิทธิภาพของแบบจำลองและใช้ขั้นตอนวิธีต้นไม้ตัดสินใจ เพื่อการจำแนกข้อมูล และสร้างแบบจำลองพฤติกรรมตามกรอบการทำเหมืองข้อมูลแบบ CRISP-DM พบว่า ความสัมพันธ์ของความถี่ในการใช้บริการ ค่าบริการ และสถานะแพ็คเกจบริการเสริม จากชุดข้อมูลเรียนรู้ และข้อมูลทดสอบ มีผลต่อแนวโน้มการยกเลิกการใช้บริการ และผลจากการวัดประสิทธิภาพของตัวแบบ เมื่อดำเนินการโดยใช้กลุ่มประชากรทั้งหมด พบว่า แบบจำลองมีประสิทธิภาพในการวิเคราะห์ข้อมูลถูกต้องถึงร้อยละ 86.68 แสดงให้เห็นว่าแบบจำลองมีความแม่นยำในการจำแนกข้อมูล และสามารถนำไปประยุกต์ใช้งานเพื่อสนับสนุนกิจกรรมทางการตลาดได้อย่างมีประสิทธิภาพ อาทิ การสร้างโปรโมชั่นที่เหมาะสมกับลูกค้าแต่ละกลุ่ม เพื่อให้เกิดความพึงพอใจสูงสุดและเกิดความภักดีกับองค์กร รวมถึงการกระตุ้นการซื้อของลูกค้า

ธรณินทร์ สัจวิริยทรัพย์ [31] ได้ศึกษาเรื่องเทคนิคการพยากรณ์แบบผสมของ ซัพพอร์ตเวกเตอร์แมชชีนและขั้นตอนวิธีเชิงพันธุกรรมสำหรับการพยากรณ์ปริมาณความต้องการพลังงานไฟฟ้าสูงสุดรายปีของประเทศไทย มีวัตถุประสงค์เพื่อพยากรณ์ความต้องการพลังงานไฟฟ้าสูงสุด โดยแรงจูงใจของการพัฒนาตัวแบบผสมนี้ เกิดจากการใช้ความสามารถในการสร้างตัวแบบที่ซับซ้อนของซัพพอร์ตเวกเตอร์แมชชีนเพื่อทำนายค่าในอนาคต แต่เนื่องจากประสิทธิภาพของซัพพอร์ตเวกเตอร์ แมชชีน

ขึ้นอยู่กับการเลือกค่าพารามิเตอร์ที่เหมาะสม ขั้นตอนวิธีการเชิงพันธุกรรมจึงถูกนำมาใช้ประโยชน์ในการหาค่าที่เหมาะสมของซัพพอร์ตเวกเตอร์แมชชีน นอกจากนี้ตัวแบบผสมที่นำเสนอได้ถูกเปรียบเทียบประสิทธิภาพการพยากรณ์กับตัวแบบดั้งเดิม ได้แก่ ตัวแบบอารีมา และซัพพอร์ตเวกเตอร์แมชชีนโดยปราศจากการหาค่าพารามิเตอร์ที่เหมาะสมที่สุดด้วยการประเมินประสิทธิภาพ 6 วิธี จากผลการศึกษาพบว่า ตัวแบบผสมที่นำเสนอมีประสิทธิภาพการพยากรณ์ที่ถูกต้องและแม่นยำกว่าตัวแบบดั้งเดิม โดยสามารถลดค่าความคลาดเคลื่อนของตัวแบบอารีมาและซัพพอร์ตเวกเตอร์แมชชีนได้อย่างน้อยร้อยละ 7.78 และ 5.72 ตามลำดับ

ไพจิต สุขสมบูรณ์ และณิชา นภาพร จงกะสิกิจ [32] ศึกษาเรื่องผลของตัวแบบที่ส่งผลต่อความสามารถในการชำระสินเชื่อของสมาชิกสหกรณ์เครดิตยูเนียน ซึ่งได้ศึกษาเกี่ยวกับการนำเทคนิคทางด้านเหมืองข้อมูลมาประยุกต์ใช้ในการระบุดัชนีที่เกี่ยวข้องกับความสามารถในการชำระสินเชื่อของสมาชิกสหกรณ์เครดิตยูเนียน เพื่อทำการศึกษาค้นหาว่าควรจะต้องพิจารณาอะไรบ้างในการอนุมัติสินเชื่อ จึงจะทำให้อัตราการเกิดหนี้สูญที่ลดลง ซึ่งรวบรวมข้อมูลจากสหกรณ์เครดิตยูเนียนในเขตอำเภอเมืองจังหวัดลำปาง จำนวน 5 แห่ง พบว่าสหกรณ์แต่ละแห่งจะมีหลักเกณฑ์ในการให้อนุมัติสินเชื่อที่คล้ายคลึงกัน โดยสามารถกำหนดเป็นปัจจัยที่เกี่ยวข้องในการพิจารณาสินเชื่อจำนวน 19 ปัจจัย และเมื่อนำมาวิเคราะห์โดยวิธีการ Evolutionary Selection พบว่ามีปัจจัยที่มีผลกระทบโดยตรงกับการชำระสินเชื่อของสมาชิกเพียง 12 ปัจจัย จากนั้นจึงนำปัจจัยที่ได้มาทำการสร้างเป็นตัวแบบด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคเบย์เซียนแบบง่าย (Naïve Bayes) และเทคนิคการค้นหาเพื่อนบ้านที่ใกล้ที่สุด k ตัว (K-Nearest Neighbors) และนำตัวแบบที่สร้างขึ้นด้วยเทคนิคข้างต้นมาทำการทดสอบด้วยวิธีการทดสอบแบบไขว้ (Cross Validation) แล้วทำการวัดค่าความถูกต้อง (Correct) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) โดยในการทดสอบเพื่อหาค่าตัวแบบที่เหมาะสมจะนำไปใช้ได้ จากข้อมูลตัวอย่างจำนวน 200 ตัวอย่าง พบว่าตัวแปรที่สร้างขึ้นด้วยวิธีการต้นไม้ตัดสินใจ ให้ผลลัพธ์ในการทดสอบที่ดีที่สุด โดยมีค่าความถูกต้องที่ร้อยละ 95.96 มีค่าความแม่นยำที่ร้อยละ 95.71 มีค่าความระลึกที่ร้อยละ 91.67 และมีค่าความถ่วงดุลที่ร้อยละ 91.94 ทั้งนี้เมื่อนำไปใช้ในการทดสอบกับข้อมูลสมาชิกที่ขอสินเชื่อและติดตามข้อมูลการชำระสินเชื่อพบว่า มีสมาชิกเพียง 2 ราย จาก 5 ราย ที่ได้ประเมินว่ามีโอกาสเกิดหนี้สูญหรือมียอดสินเชื่อค้างชำระ

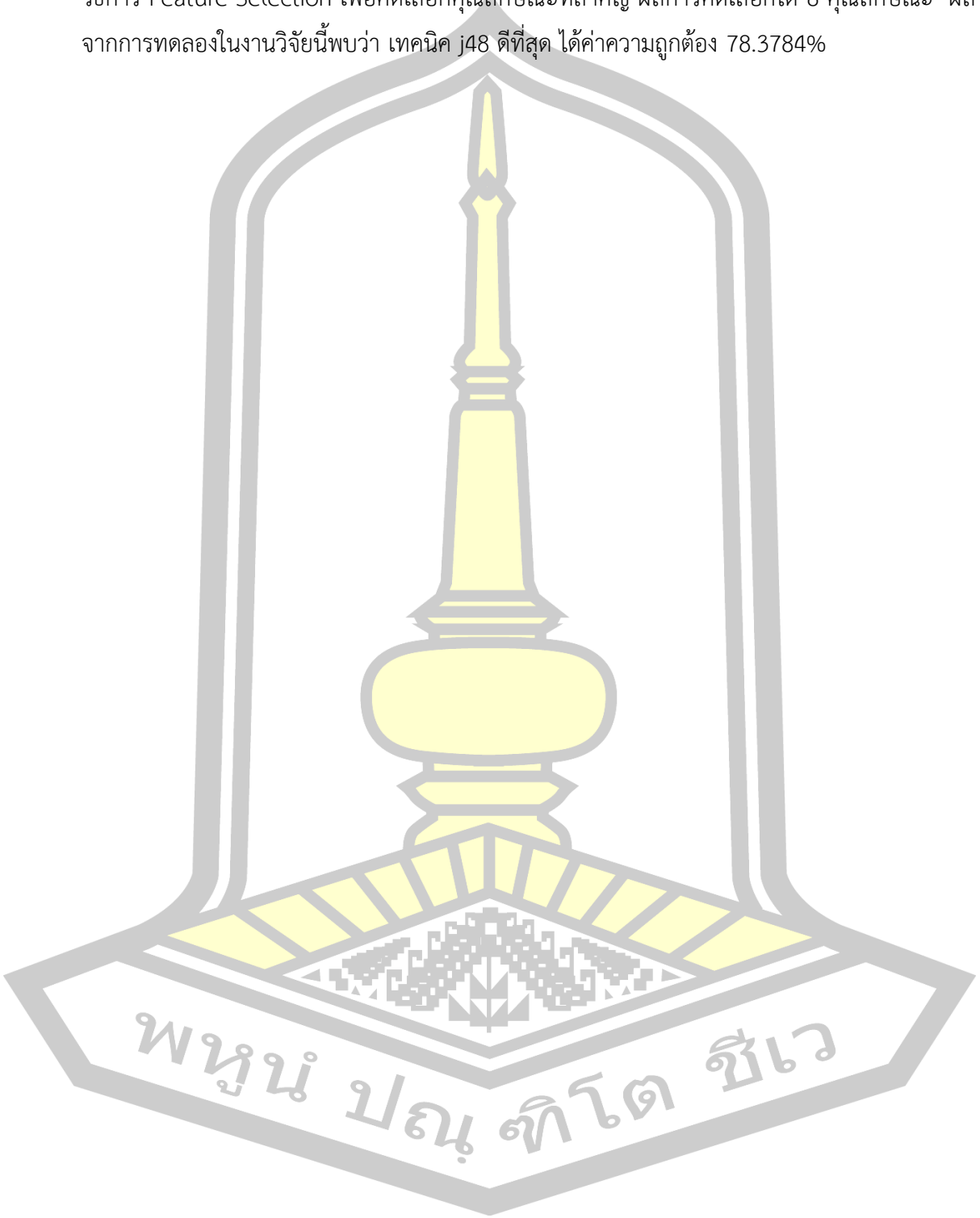
อัยนาอ์ บุญญาลัย [33] ศึกษาเกี่ยวกับการจัดกลุ่มลูกค้าที่ก่อให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้ NPLs เฉพาะสินเชื่อรายย่อย โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษาสถาบันการเงินแห่งหนึ่ง วัตถุประสงค์เพื่อนำไปใช้เป็นแนวทางในการปรับปรุงการพิจารณาอนุมัติและวิเคราะห์สินเชื่อรายย่อยช่วยในการจัดการความเสี่ยงด้านเครดิต ด้วยวิธีการวิเคราะห์การจัดกลุ่ม (Cluster Analysis) หลังจากที่ได้ผลการจัดกลุ่มแล้วทำการตรวจสอบผลโดยทำแบบสอบถามเพื่อสัมภาษณ์ผู้เชี่ยวชาญ

ด้านการวิเคราะห์ และนำผลสัมภาษณ์การเปรียบเทียบการ credit scoring เพื่อสะท้อนความมีประสิทธิภาพ โดยผลการวิจัยพบว่า ผลการพิจารณาของผู้เชี่ยวชาญไม่ผ่านเกณฑ์การพิจารณาสินเชื่อทั้ง 5 กลุ่ม และนำมาวิเคราะห์ความเสี่ยงภาคตัดขวาง ได้ดังนี้ ลูกค้ากลุ่มที่ 3 มีความเสี่ยงมากที่สุด รองลงมาคือลูกค้ากลุ่มที่ 4 ลูกค้ากลุ่มที่ 1 ลูกค้ากลุ่มที่ 5 สุดท้ายคือลูกค้ากลุ่มที่ 2 และสามารถเรียงลำดับความเสี่ยงตามหลักการวิเคราะห์สินเชื่อ ดังนี้ การวิเคราะห์ความสามารถในการชำระหนี้ มีความมากที่สุด รองลงมาคือการวิเคราะห์คุณลักษณะของผู้ขอสินเชื่อ ด้านสภาพผลกระทบจากเศรษฐกิจโดยทั่วไป การวิเคราะห์ส่วนทุน และสุดท้าย คือการวิเคราะห์หลักประกันสินเชื่อ โดยนำผลที่ได้ไปจัดแนวทางการประยุกต์ เพื่อลดการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ จัดการความเสี่ยงด้านเครดิต และลดค่าใช้จ่ายในขั้นตอนการควบคุมสินเชื่อ

ญาณินี สิทธิสาร [34] ได้ศึกษาโมเดลทำนายความสามารถในการชำระหนี้เงินกู้ กรณีศึกษาสำนักงานส่งเสริมสวัสดิการและสวัสดิการครูและบุคลากรทางการศึกษา จังหวัดอุดรธานี โดยได้ศึกษาและรวบรวมปัจจัยที่คาดว่าจะมีผลต่อความสามารถในการชำระหนี้เงินกู้ จำนวน 8 ปัจจัย ทำการวิเคราะห์ความสัมพันธ์ของปัจจัยโดยใช้ การวิเคราะห์การถดถอยพหุคูณ ผลการศึกษาพบว่าปัจจัยที่มีความสัมพันธ์กับความสามารถในการชำระเงินกู้ได้แก่ จำนวนบุตร เพศ อายุ รายได้ต่อเดือน จำนวนเงินกู้ ภาระหนี้ การทำประกันเงินกู้ จำนวนเงินงวด ข้อมูลได้จากการรวบรวมประวัติการกู้เงินระหว่างปี พ.ศ.2547 – 2557 จำนวนทั้งสิ้น 11,732 รายการ การสร้างโมเดลเพื่อทำนายความสามารถในการชำระหนี้เงินกู้ โดยใช้ต้นไม้ตัดสินใจ เครือข่ายความเชื่อเบย์ และโครงข่ายประสาทเทียม ผลการศึกษาพบว่าประสิทธิภาพการทำนายของโมเดลโครงข่ายประสาทเทียมมีค่าความแม่นยำสูงสุดที่ร้อยละ 98.77 โมเดลเครือข่ายความเชื่อเบย์ มีค่าความแม่นยำร้อยละ 98.76 และโมเดลต้นไม้การตัดสินใจ มีค่าความแม่นยำร้อยละ 98.74 ตามลำดับ เมื่อเปรียบเทียบประสิทธิภาพการทำนายแล้วพบว่าโมเดลโครงข่ายประสาทเทียมมีประสิทธิภาพการทำนายสูงที่สุด

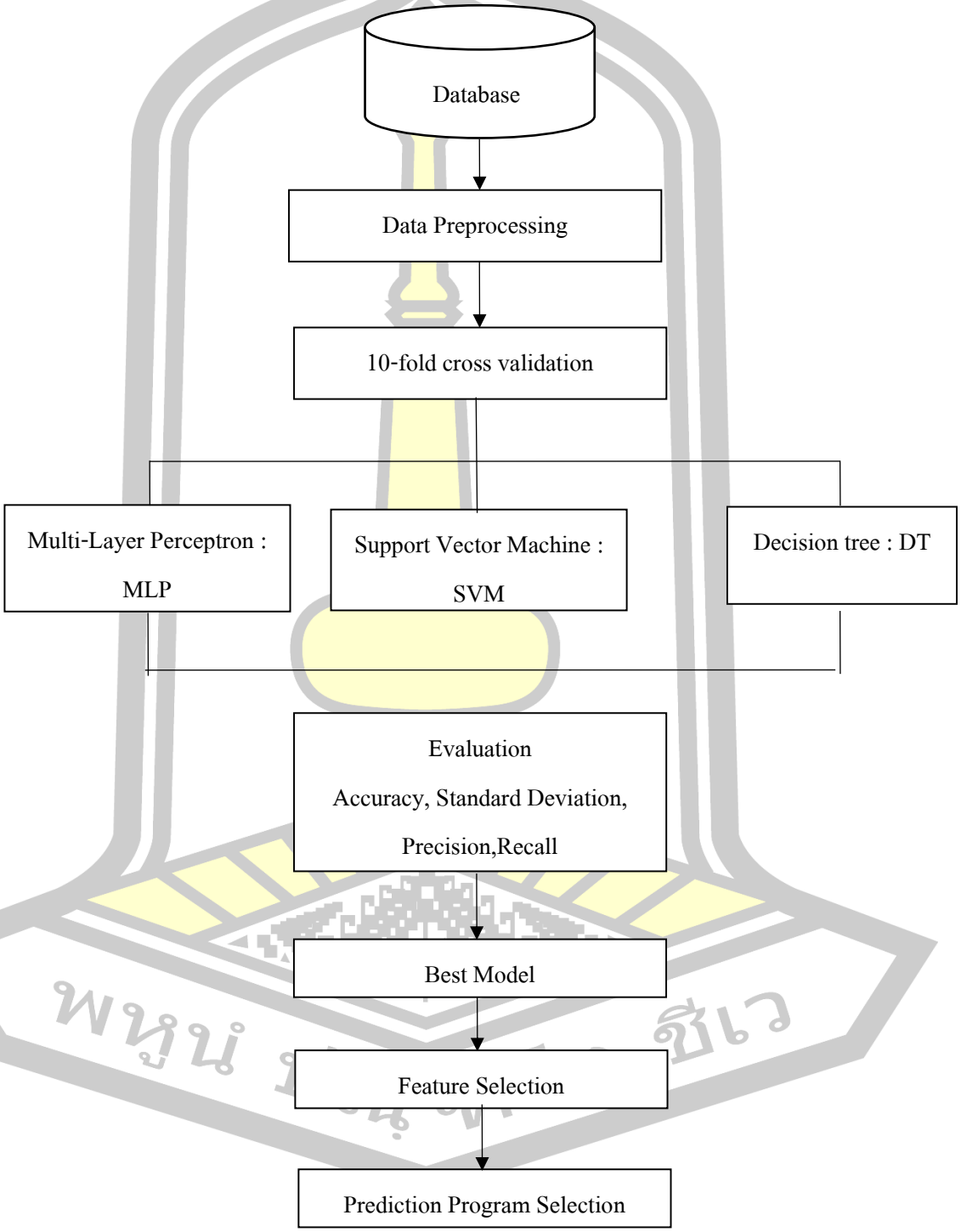
Jafar Hamid & Ahmed [35] ได้ศึกษาเรื่อง การพัฒนารูปแบบการทำนายความเสี่ยงด้านเงินกู้ในธนาคารโดยใช้เหมืองข้อมูล จากผลการศึกษาพบว่า ปัจจุบันมีความเสี่ยงมากมายที่เกี่ยวข้องกับการกู้ยืมเงินสำหรับธนาคารและสำหรับผู้ที่ได้รับเงินกู้ การวิเคราะห์ความเสี่ยงในการกู้ยืมเงินของธนาคารจำเป็นต้องเข้าใจว่าความหมายของความเสี่ยงคืออะไร นอกจากนี้จำนวนธุรกรรมในภาคธนาคารยังเติบโตอย่างรวดเร็วและมีข้อมูลปริมาณมากซึ่งแสดงถึงพฤติกรรมของลูกค้าและความเสี่ยงจากการกู้ยืมที่เพิ่มขึ้น เหมืองข้อมูลเป็นหนึ่งในงานวิจัยที่มีแรงจูงใจและมีความสำคัญมากที่สุด โดยมีจุดประสงค์เพื่อดึงข้อมูลจากเหมืองข้อมูลที่สะสมจำนวนมหาศาล จากการศึกษานี้จำแนกความเสี่ยงด้านเงินกู้ในภาคธนาคารโดยใช้เหมืองข้อมูล แบบจำลองนี้ถูกสร้างขึ้นโดยใช้รูปแบบข้อมูลภาคธนาคารเพื่อทำนายสถานะของสินเชื่อ โดยใช้อัลกอริทึม 3 แบบเพื่อสร้างโมเดลเพื่อจำแนกความเสี่ยงที่จะเกิดขึ้น ได้แก่ j48, bayesNet และ NaiveBayes ด้วยการใช้โปรแกรม Weka Model ชุดข้อมูล

ที่ใช้งานวิจัยนี้ อยู่ในรูป ARFF ซึ่งประกอบไปด้วยชื่อของ คุณลักษณะ ชนิดของคุณลักษณะ และใช้วิธีการ Feature Selection เพื่อคัดเลือกคุณลักษณะที่สำคัญ ผลการคัดเลือกได้ 8 คุณลักษณะ ผลจากการทดลองในงานวิจัยนี้พบว่า เทคนิค j48 ดีที่สุด ได้ค่าความถูกต้อง 78.3784%



บทที่ 3
วิธีการดำเนินการวิจัย

3.1 ขั้นตอนการดำเนินการวิจัย



ภาพประกอบที่ 3.1 ขั้นตอนการดำเนินงานวิจัย

3.2 การเก็บรวบรวมข้อมูล

ในการวิจัยครั้งนี้กำหนดให้กลุ่มประชากรคือ กลุ่มลูกค้าสินเชื่อโครงการธนาคารประชาชนของธนาคาร ที่กรอกข้อมูลผ่านระบบพิจารณาสินเชื่อ โดยข้อมูลจะถูกนำมาแบ่งออกเป็น 2 กลุ่ม ประกอบด้วย กลุ่มที่มีความสามารถในการชำระหนี้ (กลุ่มที่ผ่านการพิจารณาอนุมัติ) จำนวน 500 รายการ และกลุ่มที่ไม่มีความสามารถในการชำระหนี้ (กลุ่มที่ไม่ผ่านการพิจารณาอนุมัติ) จำนวน 500 รายการ รวมใช้ข้อมูลจำนวนทั้งสิ้น 1,000 รายการ โดยรายละเอียดของข้อมูลที่เก็บรวบรวมประกอบด้วย

ลำดับ	คุณลักษณะ	type
	ผลการทดลอง	label
1	อายุ	integer
2	เพศ	binominal
3	อาชีพ	polynominal
4	สถานะ	polynominal
5	รายได้หลัก	integer
6	รายได้เสริม	integer
7	ค่าใช้จ่าย	integer
8	รายได้สุทธิ	integer
9	อุปโภค-บริโภค	integer
10	ภาระหนี้	integer
11	หนี้ครั้งนี้	integer
12	รวมภาระหนี้	integer
13	วงเงินกู้	integer
14	จำนวนงวด	integer

โดยพิจารณาข้อมูลที่จะนำมาวิเคราะห์ ซึ่งเป็นข้อมูลลูกค้าที่ยื่นกู้ขอสินเชื่อกับธนาคารโดยกรอกผ่านระบบพิจารณาสินเชื่อของธนาคาร ดังภาพประกอบที่ 3.2

Label	อายุ	เพศ	อาชีพ	สถานภาพ	รายได้หลัก	รายได้เสริม/รายได้อื่น ๆ	ค่าใช้จ่ายในกิจการ/รายการหักกลบ	รายได้สุทธิ	ค่าใช้จ่ายอุปโภค-บริโภค	ภาระหนี้	หนี้ครั้งนี้	รวมภาระหนี้	วงเงินกู้	จำนวนงวด
1 Pass	41	1 ชาย	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	63,000.00	-	48,840.00	14,160.00	9,400.00	-	1,417.00	1,417.00	50,000.00	48.00
1 Pass	31	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	1 โสด	63,000.00	-	49,600.00	13,400.00	7,100.00	-	1,417.00	1,417.00	50,000.00	48.00
1 Pass	54	1 ชาย	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	1 โสด	75,000.00	15,000.00	60,500.00	29,500.00	7,500.00	15,122.00	1,889.00	17,011.00	50,000.00	36.00
1 Pass	49	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	122,595.00	-	88,569.00	34,326.00	8,600.00	14,326.00	1,889.00	16,215.00	50,000.00	36.00
1 Pass	44	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	74,500.00	-	62,150.00	12,350.00	6,800.00	-	1,417.00	1,417.00	50,000.00	48.00
1 Pass	34	1 ชาย	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	74,500.00	-	49,925.00	24,575.00	6,150.00	11,844.00	2,667.00	14,511.00	100,000.00	60.00
1 Pass	43	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	1 โสด	14,700.00	3,851.00	10,849.00	7,702.00	1,300.00	-	1,209.00	1,209.00	100,000.00	60.00
1 Pass	35	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	48,850.00	10,690.00	38,161.00	21,379.00	3,210.00	2,480.00	1,209.00	3,689.00	100,000.00	60.00
0 Failed	48	2 หญิง	2 ข้าราชการและลูกจ้างในหน่วยงานของรัฐ	1 โสด	13,750.00	-	3,981.00	9,769.00	2,940.00	6,640.00	2,910.00	9,550.00	100,000.00	24.00
1 Pass	26	1 ชาย	3 พนักงานและลูกจ้างในบริษัท	1 โสด	9,830.00	3,330.00	492.00	12,668.00	4,000.00	-	1,764.00	1,764.00	50,000.00	30.00
1 Pass	44	2 หญิง	2 ข้าราชการและลูกจ้างในหน่วยงานของรัฐ	3 สมรสไม่จดทะเบียน	8,500.00	-	-	8,500.00	2,600.00	-	1,209.00	1,209.00	50,000.00	60.00
1 Pass	29	1 ชาย	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	1 โสด	48,300.00	-	39,161.00	9,139.00	2,800.00	-	1,867.00	1,867.00	70,000.00	60.00
1 Pass	52	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	68,285.00	-	56,079.00	12,206.00	3,670.00	2,650.00	2,667.00	5,317.00	100,000.00	60.00
1 Pass	52	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	62,467.00	-	41,853.00	20,614.00	5,200.00	7,400.00	2,667.00	10,067.00	100,000.00	60.00
1 Pass	52	2 หญิง	3 พนักงานและลูกจ้างในบริษัท	2 สมรสจดทะเบียน	8,000.00	2,986.00	434.00	10,552.00	3,167.00	3,294.00	1,867.00	5,161.00	70,000.00	60.00
1 Pass	42	2 หญิง	1 ผู้ประกอบธุรกิจ / เจ้าของกิจการ	2 สมรสจดทะเบียน	260,000.00	-	213,200.00	46,800.00	11,700.00	24,596.00	2,667.00	27,263.00	100,000.00	60.00
0 Failed	54	1 ชาย	2 ข้าราชการและลูกจ้างในหน่วยงานของรัฐ	3 สมรสไม่จดทะเบียน	26,250.00	-	21,263.00	4,987.00	3,100.00	450.00	2,834.00	3,284.00	100,000.00	48.00
1 Pass	41	1 ชาย	2 ข้าราชการและลูกจ้างในหน่วยงานของรัฐ	3 สมรสไม่จดทะเบียน	15,840.00	-	-	15,840.00	4,000.00	-	2,417.00	2,417.00	100,000.00	60.00

ภาพประกอบที่ 3.2 ตัวอย่างการเก็บข้อมูลลูกค้าที่ยื่นขอสินเชื่อ

3.3 ขั้นตอนก่อนประมวลผล

ขั้นตอนก่อนการประมวลผล (Preprocessing) เป็นการเตรียมข้อมูล ซึ่งในงานวิจัยนี้ได้มีขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล ดังนี้

ผู้วิจัยได้ทำการกำหนดแบ่งข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มที่มีความสามารถในการชำระหนี้ (กลุ่มที่ผ่านการพิจารณาอนุมัติ) แทนค่าด้วย 1 Pass และกลุ่มที่ไม่มีความสามารถในการชำระหนี้ (กลุ่มที่ไม่ผ่านการพิจารณาอนุมัติ) แทนค่าด้วย 0 Failed และทำการแทนค่า คุณลักษณะอื่น ๆ ดังนี้

1. เพศ

1. ชาย

2. หญิง

2.อาชีพ

1. ผู้ประกอบธุรกิจ/เจ้าของกิจการ

2. ข้าราชการและลูกจ้างในหน่วยงานของรัฐ

3. พนักงานและลูกจ้างในบริษัท

3. สถานภาพทางครอบครัว

1. โสด

2. สมรสจดทะเบียน

3. สมรสไม่จดทะเบียน

4. หย่า

5. หม้าย

4. รายได้เสริม

0. ไม่มีรายได้เสริม

1. มีรายได้เสริม

3.4 การสร้างแบบจำลอง

นำชุดข้อมูลลูกค้าของธนาคารที่ยื่นขอสินเชื่อมาสร้างแบบจำลอง และแบ่งชุดข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลการเรียนรู้ (Training set) และข้อมูลที่ใช้ในการทดสอบ (Test set) โดยนำชุดข้อมูลได้ทำการเตรียมไว้เรียบร้อยแล้ว มาสร้างแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุดในการทำเหมืองข้อมูลเพื่อการตัดสินใจให้สินเชื่อของธนาคารออมสิน ซึ่งผู้วิจัยได้เลือกเทคนิคในการใช้เหมืองข้อมูลสำหรับธนาคารออมสินเพื่อการตัดสินใจให้สินเชื่อ ซึ่งมีทั้งหมด 3 เทคนิค ได้แก่

1) โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron: MLP)

2) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM)

3) เทคนิคต้นไม้ตัดสินใจ (Decision tree)

3.4 ขั้นตอนการคัดเลือกคุณลักษณะ

1) การแบ่งข้อมูล

งานวิจัยฉบับนี้ กำหนดให้แบ่งข้อมูลออกเป็น 2 ส่วน ประกอบด้วย ข้อมูลที่ใช้ในการเรียนรู้ (Training set) และข้อมูลที่ใช้ในการทดสอบ (Test set) โดยแบ่งเป็นอัตราส่วน 90:10

2) คัดเลือกคุณลักษณะด้วยวิธี

1) Information Gain

2) Chi-Square

3) การสร้างแบบจำลองด้วยวิธีการเหมืองข้อมูล

1) วิธีการ Information Gain โดยใช้ operator

Weight by Information Gain

Select by Weights

กำหนด Weight relation เป็น greater equals และ weight 0.1 หาน้ำหนักจากการ Weight by Information Gain

2) แบบจำลองโดยนำคุณลักษณะที่ผ่านการคัดเลือกด้วยวิธี Chi-Square โดยกำหนดค่ากำหนดค่าความเชื่อมั่น df = 0.01 ทำให้ได้น้ำหนักมากกว่าหรือเท่ากับ 27.7 และนำไปสร้างแบบจำลองด้วยวิธี SVM, MLP, Decision Tree

3.5 การประเมินประสิทธิภาพของแบบจำลอง

ข้อมูลมี 1000 ระเบียง ดังนั้นใช้ Cross-Validation ในการทดสอบประสิทธิภาพโดยกำหนดให้มีจำนวนกลุ่มเป็น 10 กลุ่ม (K=10) และทดสอบ 10 รอบเพื่อหาค่าเฉลี่ยของ ความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าเบี่ยงเบนมาตรฐาน ในการทดสอบเพื่อวัดประสิทธิภาพของแบบจำลอง

นั้น จะทำการทดสอบด้วยวิธี Cross-Validation โดยกำหนดให้มีการทดสอบจำนวน 10 รอบเพื่อหาค่าความถูกต้อง (Accuracy) เฉลี่ย และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ในการประเมินผลของแบบจำลองในการทดสอบเพื่อวัดประสิทธิภาพของแบบจำลองนั้น จะทำการทดสอบด้วยวิธี 10-fold cross validation โดยแบ่งกลุ่มข้อมูลออกเป็น 10 ส่วนเท่าๆ กัน จากนั้นใช้ข้อมูล 1 ส่วนในแต่ละรอบเป็นตัวทดสอบโมเดลประสิทธิภาพของโมเดล ซึ่งผู้วิจัยได้จะทำการทดสอบแบบจำลอง โดยใช้แบบจำลอง 1 แบบจำลอง นำข้อมูลทั้งหมดที่ได้จากการเก็บข้อมูลซึ่งผ่านกระบวนการเตรียมข้อมูลทั้งหมด 1,000 ข้อมูล

พหูน ปณ ฑิโต ชีเว

บทที่ 4
ผลการศึกษา

ในงานวิจัยฉบับนี้ แบ่งการทดลองออกเป็น 2 กระบวนการ 1) กระบวนการสร้างแบบจำลองโดยใช้วิธีการทางเหมืองข้อมูล ได้แก่ โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron: MLP) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) และ เทคนิคต้นไม้ตัดสินใจ (Decision tree) เพื่อหาแบบจำลองที่มีประสิทธิภาพสูงสุด 2) กระบวนการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งได้ใช้วิธี Chi-Square และ information gain จากนั้นจึงนำไปสร้างแบบจำลองเพื่อทดสอบประสิทธิภาพ

ผลการดำเนินการวิจัยในงานนี้ได้ใช้เทคนิค โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron: MLP) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) และ เทคนิคต้นไม้ตัดสินใจ (Decision tree) จากนั้นใช้วิธีการคัดเลือกคุณลักษณะ (Feature Selection) ในการหาค่าที่เหมาะสมและทำการลดมิติข้อมูล เข้ามาทดสอบร่วมในการสร้างแบบจำลองเพื่อจำแนกข้อมูล การวัดประสิทธิภาพแบบจำลองได้ทดลองแบ่งชุดข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วย Cross Validation และวัดประสิทธิภาพการจำแนกข้อมูลของแบบจำลองที่สร้างจากแต่ละเทคนิคโดยวัดค่าความถูกต้อง (Accuracy) และแสดงผลตัวอย่างการแปลผลการจำแนกข้อมูล ซึ่งในบทนี้จะแสดงรายละเอียดของแต่ละส่วนในงานวิจัย ดังนี้

4.1 ข้อมูลที่ใช้ในการทดลอง

4.2 ผลการวิเคราะห์ชุดข้อมูล

4.3 ผลการวิเคราะห์คุณลักษณะและลดมิติข้อมูล

4.4 ผลการทดลองประสิทธิภาพของเทคนิค

4.1 ข้อมูลที่ใช้ในการทดลอง

จากข้อมูลที่ได้รับทั้งหมด 1,000 ชุด แบ่งเป็นกลุ่มที่ได้รับการอนุมัติ 500 ราย และ ไม่อนุมัติ 500 ราย มาทำการเตรียมข้อมูลเพื่อเข้าสู่กระบวนการเหมืองข้อมูลเรียบร้อยแล้วนั้น ผู้วิจัยได้ทำการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ด้วยวิธี Cross Validation กำหนดค่า K=10-fold จากข้อมูล 1,000 ชุด เพื่อสร้างโมเดล (Training)

4.2 ผลการวิเคราะห์ชุดข้อมูล

การสร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพ โดยการนำชุดข้อมูลลูกค้าของธนาคารที่ยื่นขอสินเชื่อมาสร้างแบบจำลอง โดยแบ่งชุดข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลการเรียนรู้ (Training set) และข้อมูลที่ใช้ในการทดสอบ (Test set) โดยนำชุดข้อมูลที่ได้ทำการเตรียมไว้เรียบร้อยแล้ว มาสร้างแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุดในการทำเหมืองข้อมูลเพื่อการตัดสินใจให้สินเชื่อของธนาคาร จากคุณลักษณะทั้งหมด 14 คุณลักษณะ ซึ่งผู้วิจัยได้เลือกเทคนิคในการใช้เหมืองข้อมูลสำหรับพิจารณาการตัดสินใจให้สินเชื่อสำหรับธนาคาร ซึ่งมีทั้งหมด 3 เทคนิค ได้แก่ โครงข่ายประสาทเทียมแบบหลายชั้น ซัพพอร์ทเวกเตอร์แมชชีน และ เทคนิคต้นไม้ตัดสินใจ โดยทำการ Optimize Parameter จนได้ค่าที่ดีที่สุด ทั้งหมด 2 เทคนิค ได้แก่ โครงข่ายประสาทเทียมแบบหลายชั้น ตามตารางที่ 2 และ ซัพพอร์ทเวกเตอร์แมชชีน ตามตารางที่ 4.1 โดยเลือก Cross validation เป็น 10-fold เพื่อหาพารามิเตอร์ที่ดีที่สุดในการสร้างโมเดล

ตารางที่ 4. 1 ค่าพารามิเตอร์ที่ดีที่สุดของ Multi-Layer Perceptron (MLP)

ลำดับ	ค่าพารามิเตอร์	จำนวน
1	Input Node	14
2	Hidden Node	21
3	Output Node	2

2. Support Vector Machine: SVM โดยผู้วิจัยได้เลือกใช้ Kernel type แบบ dot และ Optimize ค่า Kernal cache, C โดยได้ค่าพารามิเตอร์ตามตารางที่ 4.3

ตารางที่ 4.2 ค่าพารามิเตอร์ที่ดีที่สุดของซัพพอร์ทเวกเตอร์แมชชีน (SVM)

ลำดับ	ค่าพารามิเตอร์	จำนวน/ ค่าพารามิเตอร์
1	จำนวน Input	14
2	จำนวน Output	2
3	C	59.5999

จากนั้นทำการประเมินประสิทธิภาพของแบบจำลอง ทำการทดสอบเพื่อวัดประสิทธิภาพของแบบจำลอง โดยทดสอบด้วยวิธี Cross-Validation กำหนดให้มีการทดสอบจำนวน 10 รอบเพื่อหาค่าความถูกต้อง (Accuracy) เฉลี่ย และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) ค่าความ

แม่นยำ (Precision) และค่าความระลึก (Recall) ในการประเมินผลของแบบจำลองในการทดสอบเพื่อวัดประสิทธิภาพของแบบจำลองนั้น จะทำการทดสอบด้วยวิธี 10-fold cross validation

จากการสร้างแบบจำลองการใช้เหมืองข้อมูลในการพิจารณาการให้สินเชื่อสำหรับธนาคาร พบว่าแบบจำลองซัพพอร์ทเวกเตอร์แมชชีน มีค่าความถูกต้อง 97.12% สูงกว่าวิธีอื่น ตามด้วยแบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้อง 88.74% และแบบจำลองเทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง 60.06% ตามลำดับ ตามตารางที่ 4.3

ตารางที่ 4. 3 ผลการเปรียบเทียบแบบจำลอง 3 เทคนิคด้วยวิธี Cross-Validation จาก 14 คุณลักษณะ

เทคนิค	10-fold cross validation		
	ค่าความถูกต้อง (Accuracy)	ค่าความแม่นยำ (Precision)	ค่าความระลึก (Recall)
MLP	83.99 ± 4.08	83.07 ± 5.98	85.98 ± 7.04
SVM	97.12 ± 2.83	96.42 ± 3.97	97.99 ± 1.97
Decision tree	60.06 ± 4.36	59.14 ± 3.50	64.30 ± 11.57

4.3 ผลการวิเคราะห์คุณลักษณะหรือปัจจัยที่มีผลต่อการพิจารณาอนุมัติ

การทดสอบคุณลักษณะที่ส่งผลต่อการพิจารณาอนุมัติ เราสามารถใช้เทคนิคการคัดเลือกคุณลักษณะ (Feature Selection) เพื่อหาปัจจัยที่ส่งผลต่อการพิจารณาอนุมัติและลดมิติของข้อมูล ในงานวิจัยนี้ได้ใช้เทคนิคจำนวนทั้งสิ้น 2 วิธี คือ Chi Square และ Information Gain

Chi Square เพื่อทำการคัดเลือกคุณลักษณะ ด้วยวิธีการทางสถิติจากคุณลักษณะทั้งหมด 14 คุณลักษณะ จากนั้นทำการทดลองโดยใช้เทคนิค ทั้ง 3 เทคนิค คือ 1) Multi-Layer Perceptron: MLP 2) Support Vector Machine: SVM และ 3) Decision tree ในการวัดประสิทธิภาพการจำแนกข้อมูลของการพิจารณาการให้สินเชื่อ จากคุณลักษณะทั้งหมด 14 คุณลักษณะ ได้น้ำหนัก ของแต่ละคุณลักษณะ ตามตารางที่ 4.4

ตารางที่ 4. 4 น้ำหนักของแต่ละคุณลักษณะเมื่อใช้วิธีการ Chi Square

ลำดับที่	คุณลักษณะ	น้ำหนัก
1	รวมภาระหนี้	131.679
2	ภาระหนี้	128.632
3	รายได้สุทธิ	61.284
4	จำนวนงวด	51.351
5	ค่าใช้จ่าย	50.511
6	รายได้หลัก	49.884
7	อาชีพ	46.822
8	หนี้ครั้งนี้, เงินงวด	42.466
9	วงเงินกู้	36.923
10	เพศ	33.188
11	ค่าใช้จ่ายอุปโภค – บริโภค	23.154
12	สถานภาพ	22.785
13	รายได้เสริม	4.946
14	เพศ	1.210

เมื่อใช้วิธีการคัดเลือกคุณลักษณะด้วยวิธีการ Chi Square โดยกำหนดค่าความเชื่อมั่น df = 0.01 ทำให้ได้น้ำหนักมากกว่าหรือเท่ากับ 27.7 ได้คุณลักษณะที่มีผล 10 คุณลักษณะ ตามตารางที่ 4.5

ตารางที่ 4.5 น้ำหนักของแต่ละคุณลักษณะที่มีผลต่อการพิจารณาสินเชื่อเมื่อใช้วิธีการ Chi Square

ลำดับที่	คุณลักษณะ	ค่าน้ำหนัก
1	รวมภาระหนี้	131.679
2	ภาระหนี้	128.632
3	รายได้สุทธิ	61.284
4	จำนวนงวด	51.351
5	ค่าใช้จ่าย	50.511
6	รายได้หลัก	49.884
7	อาชีพ	46.822

ลำดับที่	คุณลักษณะ	ค่าน้ำหนัก
8	หนี้ครั้งนี้, เงินงวด	42.466
9	วงเงินกู้	36.923
10	อายุ	33.188

เมื่อนำข้อมูลชุดนี้ทำการทดสอบแบบจำลองการใช้เหมืองข้อมูลในการพิจารณาการให้สินเชื่อสำหรับธนาคาร พบว่า Multi-Layer Perceptron มีค่าความถูกต้อง 90.40% สูงกว่าวิธีอื่น ตามด้วย Support Vector Machine มีค่าความถูกต้อง 89.40% และ Decision tree มีค่าความถูกต้อง 62.10% ตามลำดับ ตาม ตารางที่ 4.6

ตารางที่ 4.6 ผลการทดสอบตามแบบจำลอง 3 เทคนิค โดยใช้วิธีการ Chi Square

การวัดประสิทธิภาพ	เทคนิค		
	MLP	SVM	Decision tree
ความถูกต้อง (Accuracy)	90.40 ± 2.67	89.40 ± 6.64	62.10 ± 3.73
ค่าความแม่นยำ (Precision)	90.40	90.87	64.44
ค่าความระลึก (Recall)	90.40	87.60	54.00

Information Gain โดยกำหนด Weight relation เป็น greater equals และ weight 0.1 Weight จากการ Weight by Information Gain ที่ถูกเลือกหลังการทำ Feature selection ด้วย Weight by Information Gain ตามตารางที่ 4.7

ตารางที่ 4.7 น้ำหนักของแต่ละคุณลักษณะที่มีผลต่อการพิจารณาสินเชื่อเมื่อใช้วิธีการ Information Gain

Name	Type	Missing	Statistics	Fliter	Average
ผลการอนุมัติ	Binominal	0	Pass (500)	Failed (500)	Failed (500), Pass (500)
รายได้สุทธิ	Integer	0	-60000	100885	17188.568
ภาระหนี้	Integer	0	0	115000	7978.752
รวมภาระหนี้	Integer	0	656	117417	10916.618

จากชุดข้อมูลข้างต้น แสดงให้เห็นว่าคุณลักษณะ ที่ได้รับการเลือกเหลือ 3 คุณลักษณะ จากทั้งหมด 14 คุณลักษณะ จากนั้น นำชุดข้อมูลที่ได้มาสร้าง Model โดยใช้ 3 เทคนิค คือ SVM, MLP และ Decision Tree เพื่อทดสอบค่าความถูกต้อง ได้ผลการทดลอง ตามตารางที่ 4.8 ดังนี้

ตารางที่ 4.8 ผลการทดสอบตามแบบจำลอง 3 เทคนิค โดยใช้วิธีการ Information Gain

การวัดประสิทธิภาพ	เทคนิค		
	MLP	SVM	Decision tree
ความถูกต้อง (Accuracy)	91.70 ± 3.59	91.20 ± 1.69	51.80 ± 2.10
ค่าความแม่นยำ (Precision)	90.49	88.72	51.22
ค่าความระลึก (Recall)	93.20	94.40	75.80

จากตารางที่ 4.8 เมื่อนำข้อมูลชุดนี้ทำการทดสอบแบบจำลองการใช้เหมืองข้อมูลในการพิจารณาการให้สินเชื่อสำหรับธนาคาร พบว่า Multi-Layer Perceptron มีค่าความถูกต้อง 91.70% สูงกว่าวิธีอื่น ตามด้วย Support Vector Machine มีค่าความถูกต้อง 91.20% และ Decision tree มีค่าความถูกต้อง 51.80% ตามลำดับ

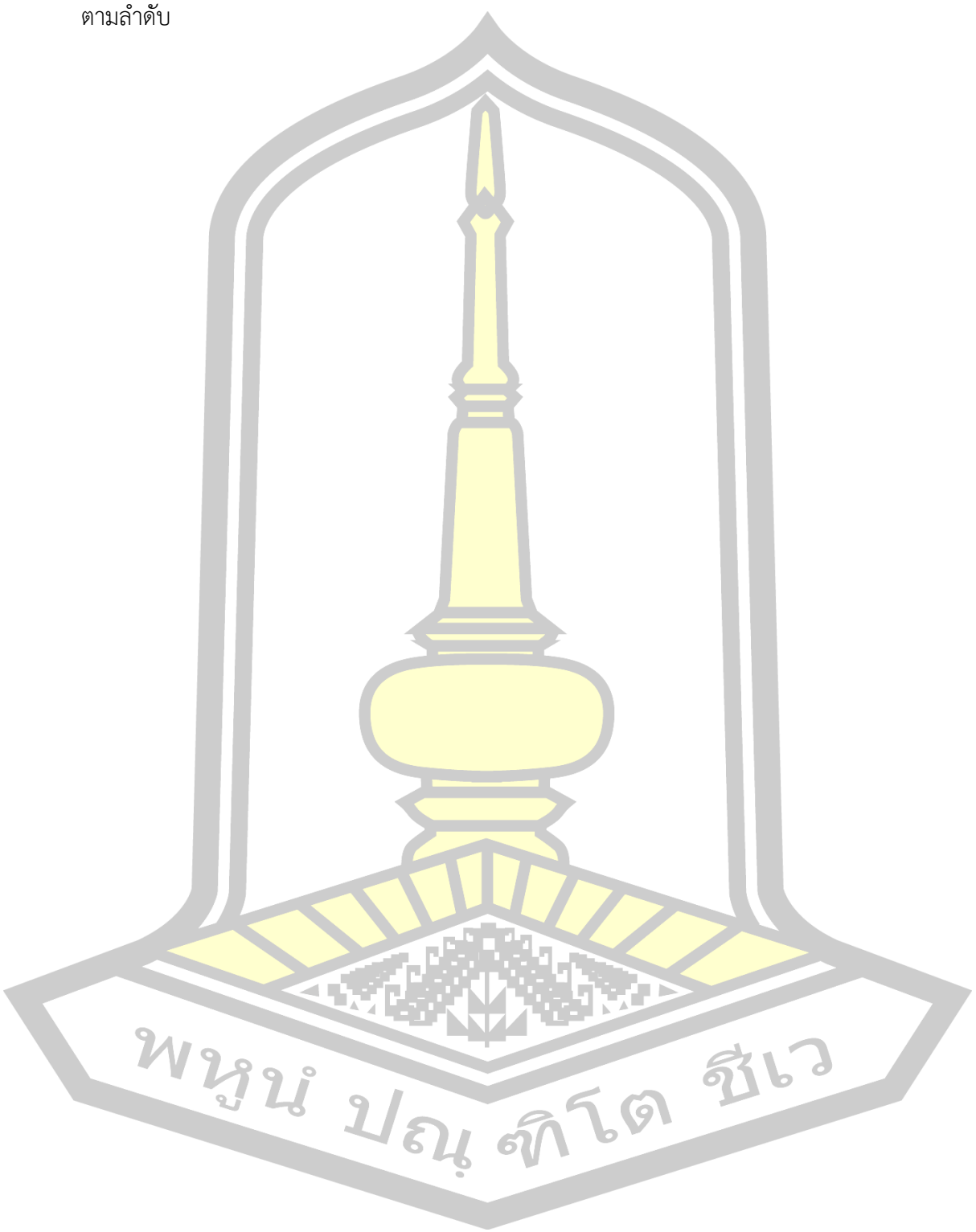
4.4 ผลการทดลองประสิทธิภาพของเทคนิค

จากการประเมินประสิทธิภาพของแบบจำลอง ข้อมูลผู้กู้จำนวน 1,000 ราย ด้วยวิธี Cross-Validation กำหนด k=10 พบว่า ปัจจัยที่มีผลต่อการพิจารณาสินเชื่อ ทั้งหมด 14 คุณลักษณะโดยใช้แบบจำลองซัพพอร์ทเวกเตอร์แมชชีน มีค่าความถูกต้อง 97.12% ตามด้วยแบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้อง 88.74% และแบบจำลองเทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง 60.06% ตามลำดับ

เมื่อใช้วิธีการทางสถิติ Chi-Square จากคุณลักษณะทั้งหมด 14 คุณลักษณะ ได้คุณลักษณะที่มีผล 10 คุณลักษณะ และเมื่อนำข้อมูลชุดนี้ทำการทดสอบแบบจำลองการใช้เหมืองข้อมูลในการพิจารณาการให้สินเชื่อสำหรับธนาคาร พบว่า Multi-Layer Perceptron มีค่าความถูกต้อง 90.40% สูงกว่าวิธีอื่น ตามด้วย Support Vector Machine มีค่าความถูกต้อง 89.40% และ Decision tree มีค่าความถูกต้อง 62.10% ตามลำดับ

และเมื่อใช้วิธี Information Gain เพื่อค้นหาคุณลักษณะที่ดีที่สุด จากการทดลองได้คุณลักษณะเพียง 3 คุณลักษณะ จากทั้งหมด 14 คุณลักษณะ จากนั้นนำชุดข้อมูลที่ได้จากวิธี Information Gain ทำการทดสอบแบบจำลองการใช้เหมืองข้อมูลในการพิจารณาการให้สินเชื่อสำหรับธนาคาร พบว่า พบว่า Multi-Layer Perceptron มีค่าความถูกต้อง 91.70% สูงกว่าวิธีอื่น ตามด้วย Support

Vector Machine มีค่าความถูกต้อง 91.20% และ Decision tree มีค่าความถูกต้อง 51.80% ตามลำดับ



บทที่ 5
สรุปผลการวิจัย ข้อเสนอแนะและงานวิจัยในอนาคต

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อนำเสนอกระบวนการวิเคราะห์ด้วยการเรียนรู้ของ ปัจจัยภายใต้แนวคิดของการจำแนกข้อมูล สามารถสรุปผลการศึกษา อภิปรายผล รวมทั้ง ข้อเสนอแนะแนวทางในการวิจัยได้ดังต่อไปนี้

5.1 สรุปผลการวิจัย

5.2 ข้อเสนอแนะและงานวิจัยในอนาคต

5.1 สรุปผลการวิจัย

เมื่อทำการทดลองโดยใช้วิธี Cross Validation เพื่อแบ่งข้อมูลออกเป็นกลุ่ม (k) โดยกำหนดให้ k=10 จากข้อมูลผู้กู้จำนวน 1,000 ราย กับแบบจำลองซัพพอร์ทเวกเตอร์แมชชีน แบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น และ แบบจำลองเทคนิคต้นไม้ตัดสินใจ พบว่า แบบจำลองซัพพอร์ทเวกเตอร์แมชชีน มีค่าความถูกต้อง 97.12% ตามด้วยแบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้อง 88.74% และแบบจำลองเทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง 60.06% ตามลำดับ ดังนั้น แบบจำลองซัพพอร์ทเวกเตอร์แมชชีน มีความเหมาะสมที่จะนำมาใช้ในการตัดสินใจให้สินเชื่อ

การคัดเลือกคุณลักษณะข้อมูลโดยใช้วิธีทางสถิติ Chi-square และวิธี Information Gain จากการทดลองเมื่อใช้วิธีการทางสถิติ Chi-Square เพื่อหาคุณลักษณะที่มีผลต่อการพิจารณาสินเชื่อ พบว่าคุณลักษณะที่มีผลมี 10 คุณลักษณะ จากนั้นนำทั้ง 10 คุณลักษณะมาสร้างแบบจำลองทั้ง 3 แบบด้วยวิธีแบบจำลองซัพพอร์ทเวกเตอร์แมชชีน, แบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น และ แบบจำลองเทคนิคต้นไม้ตัดสินใจ พบว่า เมื่อใช้วิธี Chi-square ร่วมกับวิธี แบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้อง 90.40% และเมื่อใช้วิธี Information Gain เพื่อค้นหาคุณลักษณะที่ดีที่สุด จากการทดลองได้คุณลักษณะเพียง 3 คุณลักษณะ จากทั้งหมด 14 คุณลักษณะ จากนั้นทำการทดสอบแบบจำลองการใช้เหมืองข้อมูลในการพิจารณาการให้สินเชื่อสำหรับธนาคาร พบว่า แบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้อง 91.70%

ดังนั้น จากงานวิจัย พบว่าการคัดเลือกคุณลักษณะ (Feature Selection) ด้วยวิธี Information Gain ได้คุณลักษณะที่ดีที่สุด 3 คุณลักษณะคือ รายได้สุทธิ, ภาระหนี้ และรวมภาระหนี้ และเมื่อใช้ร่วมกับแบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้องสูงสุดถึง 91.70%

ดังนั้น จากงานวิจัย พบว่าการคัดเลือกคุณลักษณะ (Feature Selection) ด้วยวิธี Information Gain ได้คุณลักษณะที่ดีที่สุด 3 คุณลักษณะคือ รายได้สุทธิ, ภาระหนี้ และรวมภาระหนี้ และเมื่อใช้ร่วมกับ แบบจำลองโครงข่ายประสาทเทียมแบบหลายชั้น มีค่าความถูกต้องสูงสุดถึง 91.70%

Feature Selection	**คุณลักษณะที่ดีที่สุด**
Chi-square	รวมภาระหนี้, ภาระหนี้, รายได้สุทธิ
Information Gain	รายได้สุทธิ, ภาระหนี้ และรวมภาระหนี้

5.2 ข้อเสนอแนะและงานวิจัยในอนาคต

งานวิจัยในอนาคต ควรพิจารณาเรื่องของ Instance Selection เพื่อลดจำนวนของข้อมูล ที่นำมาเรียนรู้ และใช้วิธี Ensemble Learning เพื่อสร้างโมเดลที่เหมาะสมกับการพิจารณาสินเชื่อ

พหุน ปณ ฑิโต ชีเว

บรรณานุกรม

[1] ธีรวัฒน์ พสุสิริโรจน์ และธีระวัฒน์ จันทึก, “การประเมินความสำเร็จโครงการสินเชื่อธนาคารประชาชนของธนาคารออมสิน”

[2] ณัฐญา เกิดเพชร, “การประเมินผลการสื่อสารโครงการธนาคารประชาชนของธนาคารออมสิน”, วารสารศาสตรมหาบัณฑิต สาขาวิชาการจัดการสื่อสารองค์กร คณะวารสารศาสตร์และสื่อสารมวลชน มหาวิทยาลัยธรรมศาสตร์ 2552.

[3] เกณฑ์ที่สถาบันการเงินใช้พิจารณาสินเชื่อ. https://www.1213.or.th/th/serviceunderbot/loans/Pages/beforeloan.aspx

[4] ธนาคารแห่งประเทศไทย, “คำอธิบายข้อมูล”, http://www2.bot.or.th/statistics MetaData/FI_NP_005_TH.PDF

[5] ธนาคารออมสิน, ผลการดำเนินงานสาขาบัวใหญ่ เดือน ธันวาคม 2561.

[6] ธนาคารออมสิน, สินเชื่อโครงการธนาคารประชาชนhttps://www.gsb.or.th/personal/products/loan/government/bank_people.aspx

[7] เอกสิทธิ์ พัชรวงศ์ศักดา, การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น, กรุงเทพฯ: หสม.ดาต้า คิวบ์, 2557.

[8] เสกสรรค์ วิลัยลักษณ์ วิภา เจริญภัณฑารักษ์และดวงดาว วิชาดากุล, “การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียน โรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา”, วารสารวิชาการสาขาวิทยาศาสตร์และเทคโนโลยี, 2(2), ก.ค. - ธ.ค. 2558.

[9] Pete Chapman, Julian Clinton, et al. (2000). CRISP-DM 1.0: Step-by-step datamining guide , SPSS inc 78.

[10] สายชล สินสมบูรณ์ทอง. (2558). การทำเหมืองข้อมูล. กรุงเทพมหานคร: จามจุรีโปรดักท์.

[11] Na-Wichien, S. (2017). Data Mining. Chulalongkorn University.

[12] Wang, L., Jia, X., 2009. Integration of soft and hard classifications using extended support vector machines. IEEE Geoscience and Remote Sensing Letters 6 (3), 543–547.

[13] R.O Duda and P.E.Hart. Pattern Classification and Scene Analysis. New York : Wiley, 1973.

[14] Courant, R. and Hilbert, D. “Methods of Mathematical Physics” Vol.1, Interscience Publishers, New York, 1953.

[15] วิทยา พรพัชรพงศ์, “โครงข่ายประสาทเทียม (Artificial Neural Networks-ANN). Access 29 November. Available from http://www.gotoknow.org/posts/163433.”2556)

[16] Zhang GP, Patuwo BA. Cross validation analysis of neural network out-of-sampleperformance in exchange rate forecasting. Int J Decision Science. 1998 ; 30(1) : 197-216.

[17] โครงข่ายประสาทเทียม. (2013), Access 29 November 2013, Available from http://www.scribd.com/doc/45905452/Report-Neural-Network

[18] Frederick M. Chester. Internal Structure and Weakening Mechanisms of the San Andreas Fault 1993. Journal of geophysical research, January 10 Vol.98 : 771-786.

[19] เอกรินทร์ แซ่เฮ้ง, “โครงข่ายประสาทเทียมกับการประยุกต์ใช้งาน (ตอนที่ 1 รู้จักกับโครงข่ายประสาทเทียม)”, แผนกสารสนเทศ สำนักวิชาการ วิทยาลัยนอร์ท กรุงเทพฯ.2553

[20] พยูน พาณิชย์กุล, “การพัฒนาระบบดาต้าไมนิ่งโดยใช้ Decision Tree.” โครงการพัฒนาระบบงานปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ แขนงวิทยาการสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง, 2548.

[21] จิตตภู พูลวัน. ระบบวิเคราะห์โรคทั่วไปโดยใช้การจำแนกข้อมูลบนอินเทอร์เน็ต ภาควิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2550.

[22] ศุภชัย ประคองศิลป์, “การออกแบบและพัฒนาระบบสนับสนุนการตัดสินใจในการอนุมัติลูกบ้านเข้าโครงการโดยใช้เทคนิคต้นไม่ตัดสินใจ กรณีศึกษา มูลนิธิที่อยู่อาศัยเพื่อมนุษยชาติ”. ปัญหาพิเศษปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2551.

[23] Ham, J. and M. Kamber. “Data Mining : Concepts and Techniques”. CA: Morgan Kaufmann, San Francisco, 2001.

[24] ขจรศักดิ์ ศรีอ่อน, “การทำนายสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้อง โดยใช้เทคนิคการทำเหมืองข้อมูลในระบบจำหน่ายของการไฟฟ้าส่วนภูมิภาค เขต 1 ภาคกลาง” วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมไฟฟ้า) สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์, 2552.

[25] Michell Tom M, “Machine Learning” McGraw Hill, Newjersy, 1997.

[26] Jiawei Han, Michelin Kameber and Jain Pei, “Data Mining : Concepts and Techniques”, Morgan Kaufmann plubliher, San Francisco, CA, 2006.

[27] M. A. Hall, “Correlation-based Feature Selection for Machine Learning,” no. April, 1999.

[28] รศ.ดร. กัลยา วานิชย์บัญชา, การวิเคราะห์ข้อมูลหลายตัวแปร, ภาควิชาสถิติ คณะพาณิชศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย; 2552.

[29] ประเวช สุพรรณพยัคฆ์. “การวิเคราะห์พฤติกรรมลูกค้าที่ก่อให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้ในธุรกิจสินเชื่อรถยนต์ โดยการทำเหมืองข้อมูล กรณีศึกษา บริษัท ลิสซิ่ง แห่งหนึ่ง”, ค้นคว้าอิสระปริญญามหาบัณฑิต, สาขาวิชาการบริหารเทคโนโลยีวิทยาลัยนวัตกรรม มหาวิทยาลัยธรรมศาสตร์, 2553.

[30] เบญญา จันทพันธ์. “การศึกษาการหาพฤติกรรมการยกเลิกการใช้บริการธุรกรรมทางการเงินผ่านโทรศัพท์เคลื่อนที่จากรายการการบันทึกการทำธุรกรรมโดยการทำเหมืองข้อมูล”, ค้นคว้าอิสระปริญญามหาบัณฑิต, สาขาวิชาบริหารเทคโนโลยี วิทยาลัยนวัตกรรม มหาวิทยาลัยธรรมศาสตร์, 2554.

[31] ธรณินทร์ สัจวิริยทรัพย์. “เทคนิคการพยากรณ์แบบผสมของ ซัพพอร์ตเวกเตอร์ แมชชีน และขั้นตอนวิธีเชิงพันธุกรรมสำหรับการพยากรณ์ปริมาณความต้องการพลังงานไฟฟ้าสูงสุดรายปีของประเทศไทย”,วารสารวิชาการพระจอมเกล้าพระนครเหนือ, 27(3), 2560.

[32] ไพจิต สุขสมบูรณ์ และณิชา นภาพร จงกะสิกิจ. “ผลของตัวแบบที่ส่งผลต่อความสามารถในการชำระสินเชื่อของสมาชิกสหกรณ์เครดิตยูเนี่ยน”, วารสารวิชาการคณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏลำปาง, 10(1), มกราคม – มิถุนายน 2560.

[33] อัยนาอ์ บุญญาลัย, “การจัดกลุ่มลูกค้าที่ก่อให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้ NPL เฉพาะสินเชื่อรายย่อย โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษาสถาบันการเงินแห่งหนึ่ง”, ค้นคว้าอิสระปริญญามหาบัณฑิต, มหาวิทยาลัยธรรมศาสตร์.

[34] ญาณินี สิทธิสาร, “โมเดลทำนายความสามารถในการชำระหนี้เงินกู้ กรณีศึกษาสำนักงานส่งเสริมสวัสดิการและสวัสดิการครูและบุคลากรทางการศึกษา จังหวัดอุดรธานี” หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น, 2561.

[35] Jafar Hamid, A., & Ahmed, T. M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining. *Machine Learning and Applications: An International Journal*, *3*(1), 1–9.

พหูน ปณฺฑิโต ชีเว

ประวัติผู้เขียน

ชื่อ	นางสาวขวัญนภา พิมพ์ชารีย์
วันเกิด	วันที่ 28 ตุลาคม พ.ศ. 2530
สถานที่เกิด	โรงพยาบาลวังน้ำเย็น อำเภอ วังน้ำเย็น สระแก้ว
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 108/1 ถนนเทศบาล 2 ตำบลบัวใหญ่ อำเภอบัวใหญ่ จังหวัดนครราชสีมา รหัสไปรษณีย์ 30120
ตำแหน่งหน้าที่การงาน	พนักงานธุรกิจสาขา ธนาคารออมสิน
สถานที่ทำงานปัจจุบัน	บ้านเลขที่ 153 ถนนนิเวศรัตน์ ตำบลบัวใหญ่ อำเภอบัวใหญ่ จังหวัดนครราชสีมา รหัสไปรษณีย์ 30120
ประวัติการศึกษา	พ.ศ. 2547 มัธยมศึกษาตอนต้น โรงเรียนบัวใหญ่ อำเภอบัวใหญ่ จังหวัดนครราชสีมา พ.ศ. 2549 มัธยมศึกษาตอนปลาย โรงเรียนบัวใหญ่ อำเภอบัวใหญ่ จังหวัดนครราชสีมา พ.ศ. 2553 วิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม พ.ศ. 2564 วิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม

พหูน ปณฺ ฑิโต ชีเว