



การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสม
 เพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา

An Efficiency Comparison of Algorithms and Feature Selection
 Methods to Predict the Learning Achievement of Vocational Students

รัชพล กลัดชื่น^{1*} และ จริญญา แสนราช²

¹แผนกคอมพิวเตอร์ธุรกิจ วิทยาลัยอาชีวศึกษานครปฐม นครปฐม 73000

²ภาควิชาคอมพิวเตอร์ศึกษา มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร 10800

Corresponding Author E-mail :charun.s@fte.kmutnb.ac.th

ARTICLE INFO	ABSTRACT
Article history:	The objectives of this research were to compare an efficiency of
Received 30 April 2018	algorithm and feature selection methods to predict the learning
Accept 04 June 2018	achievement of vocational students. By studying the vocational certificate
Online 30 June 2018	student's profiles during the 2007-2016 academic year for 5,100 records 27
Keywords:	features. In this research includes two methods: Simple Classification and
Data Mining,	Hybrid Classification, Simple Classification used only classification
Decision Tree,	algorithms. The three candidate classification algorithms were used in this
Naïve Bayes,	experiment for both Simple and Hybrid Classification include: 1) Decision
Rule Induction,	Tree: J48graft 2) Naïve Bayes and 3) Rule Induction. Hybrid Classification
Feature Selection	used forward selection techniques to select attributes. The data was
	analyzed by using Rapid Miner Studio 8 with data mining technique on 10-
	fold cross validation. The comparative results showed that the accuracy of
	Decision Tree: J48graft by forwarding selection and all feature selection is
	83.08% and 81.71%, respectively. In conclusion, the prediction model using
	Decision Tree: J48graft is most appropriate for predicting the learning
	achievement of vocational students.

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายและคุณลักษณะที่มีต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา โดยทำการศึกษาข้อมูลนักศึกษาระดับประกาศนียบัตรวิชาชีพ จำนวน 5,100 ระเบียน ตั้งแต่ปีการศึกษา 2550 - 2559 9 สาขาวิชา 27 คุณลักษณะ โดยใช้เทคนิคการจำแนกข้อมูล 3 เทคนิค ได้แก่ Decision Tree : J48graft, Naïve Bayes และ Rule Induction ทำการเปรียบเทียบประสิทธิภาพตัวแบบการทำนายระหว่างการใช้คุณลักษณะทั้งหมดกับการเลือกคุณลักษณะแบบ forward select ทดสอบประสิทธิภาพตัวแบบทำนายด้วยวิธีการ 10-fold cross validation โดยใช้โปรแกรม Rapid Miner Studio 8 จากนั้นนำผลการทดสอบประสิทธิภาพที่มีค่าความถูกต้องที่สูงที่สุด 2 ค่า มาทำการเปรียบเทียบด้วยวิธี T-Test ผลการศึกษาพบว่าการใช้เทคนิค Decision Tree : J48graft ด้วยการเลือกคุณลักษณะแบบ Forward Selection และ การเลือกคุณลักษณะทั้งหมด มีค่าความถูกต้องเท่ากับ 83.08% และ 81.71% ตามลำดับ และทดสอบด้วยวิธี T-Test พบว่าการทดสอบทั้งสองแบบมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จากผลการเปรียบเทียบประสิทธิภาพในครั้งนี้ สามารถนำเทคนิค Decision Tree : J48graft ไปใช้ในการทำนายผลสัมฤทธิ์ทางการเรียน และเป็นแนวทางในการสอนเสริมหรือแนะแนวให้กับนักศึกษาต่อไป

คำสำคัญ: เหมืองข้อมูล, วิชิตันไม้ตัดสินใจ, วิธีแบบเบย์, วิธีการสร้างกฎ, การคัดเลือกคุณลักษณะ

คำนำ

จุดมุ่งหมายของหลักสูตรประกาศนียบัตรวิชาชีพชั้นสูง พุทธศักราช 2557 มุ่งเน้นให้ผู้เรียนมีความรู้พื้นฐาน

ในการดำรงชีวิต มีทักษะและสมรรถนะในงานอาชีพ สามารถบูรณาการความรู้ ทักษะจากศาสตร์ต่าง ๆ มาประยุกต์ใช้ในงานอาชีพ แต่จากรายงานผลการทดสอบทางการศึกษาระดับชาติด้านอาชีวศึกษา (V-Net) ซึ่งเป็นการทดสอบเพื่อวัดความรู้และความคิดของผู้เรียน ตั้งแต่ปีการศึกษา 2556-2560 พบว่านักศึกษาระดับประกาศนียบัตรวิชาชีพชั้นสูง มีคะแนนเฉลี่ยเท่ากับ 39.27, 41.37, 41.69, 37.35 และ 37.11 ตามลำดับระดับคุณภาพอยู่ในเกณฑ์ ต้องปรับปรุง แสดงให้เห็นถึงคุณภาพทางด้านวิชาการอยู่ในระดับต่ำกว่าเกณฑ์มาตรฐาน [1] ทั้งนี้เนื่องจากตัวผู้สอนใช้การบรรยายให้ผู้เรียนฟัง ขาดแรงจูงใจให้ผู้เรียนสนใจในการเรียน ประกอบกับผู้เรียนอาชีวศึกษาส่วนมากมีผลสัมฤทธิ์ทางการเรียนค่อนข้างต่ำ ไม่ค่อยสนใจในการเรียนเท่าที่ควร ขาดวินัย ขาดความขยันและขาดความอดทน

การทำเหมืองข้อมูล เป็นวิธีการในการค้นหา รูปแบบและแนวโน้มที่มีประโยชน์จากแหล่งข้อมูลขนาดใหญ่ [2] เช่น ฐานข้อมูล คลังข้อมูล เว็บ หรือแหล่งจัดเก็บข้อมูลอื่น ๆ โดยเฉพาะในด้านการศึกษาถือว่าเป็นแหล่งที่มีข้อมูลขนาดใหญ่ถูกเก็บรวบรวมอยู่จำนวนมาก เช่น ประวัติผู้เรียน ประวัติผลการเรียน ข้อมูลรายวิชาต่างๆ เป็นต้น

วิทยาลัยอาชีวศึกษาส่วนใหญ่มีการเก็บข้อมูลต่างๆ ของนักศึกษาไว้เป็นจำนวนมาก และไม่ได้นำมาใช้ให้เกิดประโยชน์หลังจากที่นักศึกษาเรียนจบไปแล้ว ปัจจุบันนักวิจัยได้เริ่มนำข้อมูลเหล่านั้นมาทำการศึกษาเพื่อทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษา โดยนำข้อมูลที่เก็บไว้จำนวนมากเหล่านั้นมาสร้างเป็นตัวแบบ โดยใช้เทคนิคเหมืองข้อมูล (Data Mining) [3] จากการทบทวนงานวิจัยพบว่าได้มีงานวิจัยเป็นจำนวนมากที่ได้ให้ความสำคัญและได้ทำการศึกษาถึงปัจจัยที่มีผลต่อผลสัมฤทธิ์ทางการเรียน รวมถึงการพัฒนาตัวแบบสำหรับการสร้างระบบเพื่อวิเคราะห์แนวทางการเรียนให้กับผู้เรียน ดังเช่น สุกิจ และคณะ [4] ได้เปรียบเทียบ

ประสิทธิภาพของการจำแนกข้อมูลด้วยเทคนิคเหมืองข้อมูล 3 เทคนิค คือ C4.5 (J48), K-Nearest Neighbors (K-NN) และ Naïve Bayes เพื่อวิเคราะห์ปัจจัยพฤติกรรมด้านวินัยในตนเองที่ส่งผลต่อระดับผลการเรียนของนักศึกษา ผลการเปรียบเทียบพบว่าการใช้ C4.5 (J48) ให้ประสิทธิภาพสูงสุดร้อยละ 78 และ เสกสรร [5] ได้ทำการสร้างตัวแบบสำหรับหาปัจจัยเพื่อพยากรณ์ผลการเรียนของนักเรียนโรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาลัยกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา โดยใช้ข้อมูลนักเรียนระดับมัธยมศึกษาปีที่ 4 ระหว่างปีการศึกษา 2553 – 2556 จำนวน 525 ระเบียบ ประกอบด้วย 16 คุณลักษณะ พบว่า ชุดข้อมูลแบบไม่จัดกลุ่มนำมาคัดเลือกคุณลักษณะด้วยวิธี Correlation-based Feature Selection (CFS) ร่วมกับเทคนิคโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอน ให้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 94.48 นอกจากนี้ สุภาวดี และคณะ [6] ได้นำเสนอรูปแบบการทำนายผลการทดสอบทางการศึกษาระดับชาตินี้พื้นฐาน (O-Net) โดยทำการเปรียบเทียบเทคนิคเหมืองข้อมูล 3 เทคนิค ได้แก่ Decision Tree, K-Nearest Neighbors (K-NN) และ Naïve Bayes พบว่าการใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) สร้างตัวแบบในการทำนายรายวิชาภาษาไทยมีค่าเฉลี่ยความถูกต้องสูงสุดร้อยละ 74.14 และวิชาวิทยาศาสตร์มีค่าเฉลี่ยความถูกต้องสูงสุดร้อยละ 81.46 อรณุช และคณะ [7] ได้เปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล 3 วิธี คือ วิธีแบบเบย์ วิธีต้นไม้ตัดสินใจ และวิธีฐานกฎ พบว่าการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ ด้วยอัลกอริทึม J48 มีประสิทธิภาพสูงสุดให้ค่าความถูกต้องเท่ากับ 82.78%

ปัญหาในการทำเหมืองข้อมูลที่สำคัญอย่างหนึ่งคือ คุณลักษณะของข้อมูลที่ทำนายมีจำนวนมากซึ่ง บางส่วนก็ไม่มีความสัมพันธ์กับการสร้างตัวแบบในการทำนาย ดังนั้นในการคัดเลือกคุณลักษณะจึงเป็นขั้นตอนที่

สำคัญในการทำเหมืองข้อมูล อย่างที่ ไกรุ่ง และคณะ [8] ได้ใช้เทคนิคเหมืองข้อมูล 2 เทคนิค คือ ต้นไม้ตัดสินใจ C4.5 และข่ายงานความเชื่อเบย์ (BayesNet) โดยแบ่งการทดลองออกเป็น 2 ลักษณะ คือ การทดลองกับชุดข้อมูลทั้งหมดทุกแอททริบิวต์ และการทดลองด้วยการเลือกทุกแอททริบิวต์ด้วยวิธีการ SNR (Signal to Noise Ratio) พบว่า วิธีการต้นไม้ตัดสินใจ C4.5 ให้ประสิทธิภาพในการทำงานดีกว่าเมื่อเปรียบเทียบกับวิธีการโครงข่ายเบย์

ดังนั้นงานวิจัยนี้จะนำเสนอเทคนิคเหมืองข้อมูลมาใช้ในการเรียนรู้การเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมในการจำแนกและหาปัจจัยที่มีผลต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา เพื่อเป็นแนวทางในการปรับปรุงระบบการเรียนการสอนต่อไป ด้วยวิธีการจำแนกข้อมูล 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree : J48graft) เทคนิคนาอิวเบย์ (Naïve Bayes) และ เทคนิคกฎการอุปนัย (Rule Induction)

วัตถุประสงค์ของการวิจัย เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายและคุณลักษณะที่มีผลสัมฤทธิ์ผลทางการเรียนของนักศึกษาอาชีวศึกษา

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. การจำแนกประเภทข้อมูล (Data Classification) เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยเป็นการสร้างโมเดลจำแนกประเภทเพื่อทำนายกลุ่มของข้อมูลใหม่ ซึ่งคำตอบที่เป็นประเภทค่าต่าง ๆ เหล่านี้ในการวิเคราะห์ข้อมูลเรียกว่าคลาส (Class) หรือ ลาเบล (Label) การสร้างโมเดลจำแนกประเภทข้อมูลเกิดจากการหาความสัมพันธ์ของข้อมูล โดยข้อมูลทั้งหมดจะแบ่งเป็น 2 ส่วน คือ ส่วนแรกข้อมูลเรียนรู้ (Training data) เป็นชุดข้อมูลที่ใช้ในการสร้างโมเดลจำแนกประเภทข้อมูลขึ้นมาใหม่ เพื่อให้โมเดลที่สร้างได้เรียนรู้ข้อมูล และส่วนที่สองของข้อมูลที่ใช้ในการทดสอบโมเดลที่สร้างขึ้นมา (Testing Data) เป็นชุดข้อมูลประเมินความถูกต้องของโมเดลจำแนกประเภทข้อมูล [9] งานวิจัย

นี้คณะผู้วิจัยได้เลือกเทคนิคของการทำเหมืองข้อมูล 3 เทคนิค [10]

1.1 ต้นไม้ตัดสินใจ (Decision Tree) เป็นตัวแบบทางคณิตศาสตร์ที่ใช้จำแนกประเภทข้อมูลที่มีความแม่นยำสูง แต่ประสิทธิภาพของตัวแบบขึ้นอยู่กับคุณลักษณะของข้อมูลที่น่ามาใช้ ซึ่งค่าของคุณลักษณะเหล่านี้จะอยู่ในรูปของค่าที่ไม่ต่อเนื่อง (Discrete Value) เช่น เพศ สาขาวิชา เป็นต้น ตัวอย่างการนำต้นไม้ตัดสินใจไปใช้ เช่น วงการแพทย์ การวิเคราะห์ทางการเงิน ดาราศาสตร์ [11] ต้นไม้ตัดสินใจเป็นโครงสร้างข้อมูลชนิดเป็นลำดับชั้น (Hierarchy) ใช้สนับสนุนการตัดสินใจ โดยจะมีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดราก (Root Node) อยู่ด้านบนสุดและโหนดใบอยู่ล่างสุดของต้นไม้ ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดจะมีคุณลักษณะ (Attribute) เป็นตัวทดสอบ กิ่งของต้นไม้ (Branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ (Leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (Class) หรือผลลัพธ์ที่ได้จากการทำนาย โหนดที่อยู่บนสุดของต้นไม้เรียกโหนดราก (Root Node)

จากรูปที่ 1 จะแทนกฎการจำแนกประเภทดังนี้

IF <age>=youth and <student>=no THEN no

IF <age>=youth and <student>=yes THEN yes

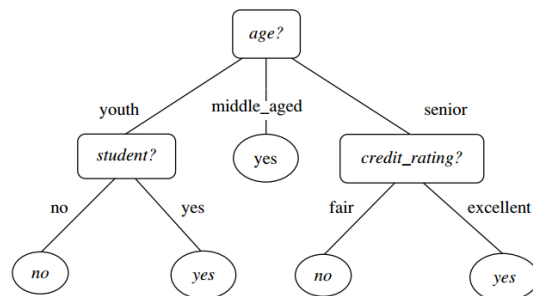
IF <age>=middle_aged THEN yes

IF <age>=senior and <credit_rating>=fair

THEN no

IF <age>=senior and <credit_rating>=excellent

THEN yes



รูปที่ 1 ตัวอย่างต้นไม้ตัดสินใจ [10]

1.2 อัลกอริทึมนาอิวเบย์ (Naïve Bayes) นาอิวเบย์เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น (Probability) ตามทฤษฎีของเบย์ (Bayes's theorem) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ หลักการของนาอิวเบย์ใช้หลักการของความน่าจะเป็น โดยมีสมมติฐานว่าปริมาณของความสนใจขึ้นอยู่กับกระจายความน่าจะเป็น (Probability Distribution) [10] เป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณลักษณะ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ

1.3 Rule Induction เป็นวิธีการสำหรับการดึงเอาชุดกฎเกณฑ์ต่าง ๆ เพื่อจัดแบ่งเงื่อนไขหรือกรณีโครงสร้างต้นไม้สามารถสร้างชุดของกฎต่าง ๆ และขณะที่บางครั้งเรียกวิธีการนี้ว่า การสร้างกฎใหม่จากตัวอย่าง แต่วิธีการนี้ยังมีความหมายที่แตกต่างกันเนื่องจากวิธีการใช้การอุปนัยจะสร้างชุดของกฎที่เป็นอิสระ ซึ่งไม่จำเป็นต้องอยู่ในรูปโครงสร้างของต้นไม้ เพราะตัวสร้างการอนุมานกฎ (Rule Induction) ไม่ได้บังคับการแตกข้อมูลเป็นแต่ละระดับ แต่อาจจะสามารถค้นหารูปแบบ (Pattern) ที่แตกต่างกันได้

2. การเลือกคุณลักษณะ (Feature Selection)

การวิเคราะห์ข้อมูลที่มีมิติจำนวนมากนั้น ถือเป็นเรื่องยาก และมีความซับซ้อนสูง ขั้นตอนวิธีสำหรับการเรียนรู้ที่จะนำมาใช้จัดการกับข้อมูลเหล่านี้จะต้องใช้ทรัพยากรทั้งในด้านการคำนวณและการใช้หน่วยความจำจำนวนมาก ในขณะที่ประสิทธิภาพของกระบวนการเรียนรู้อาจลดลงเนื่องจากอาจมีสัญญาณรบกวน (Noise) ในข้อมูลที่เกิดจากมิติที่ไม่เกี่ยวข้องกับการวิเคราะห์ข้อมูลนั้น ๆ ดังนั้นเพื่อแก้ปัญหาดังกล่าว มิติของข้อมูลเหล่านั้นจะต้องถูกลดจำนวนลงด้วยวิธีการเลือกคุณลักษณะที่สำคัญสำหรับการวิเคราะห์ข้อมูล [12]

การเลือกคุณลักษณะแบ่งได้ 3 ประเภท ได้แก่ Embedded, Filter และ Wrapper ซึ่งมีรายละเอียดดังต่อไปนี้

1) การเลือกคุณลักษณะแบบฝังตัว (Embedded) จะเกิดขึ้นโดยขั้นตอนวิธีสำหรับการเรียนรู้เอง ซึ่งในขั้นตอนวิธีสำหรับการเรียนรู้เอง จะมีการเลือกคุณลักษณะที่เหมาะสมสำหรับการสร้างแบบจำลองในการแก้ปัญหาต่าง ๆ โดยไม่ต้องเพิ่มกระบวนการวิธีในการคัดเลือกคุณลักษณะที่เหมาะสมอื่นๆ เข้ามาช่วย

2) การเลือกคุณลักษณะแบบกรอง (Filter) จะเป็นกระบวนการประเมินประสิทธิภาพของคุณลักษณะของข้อมูลแต่ละตัวว่ามีความเหมาะสมกับการวิเคราะห์ข้อมูลมากน้อยเพียงใด โดยไม่ขึ้นกับขั้นตอนวิธีของการเรียนรู้แบบใดแบบหนึ่ง การเลือกคุณลักษณะแบบนี้จะทำการจัดลำดับ (Ranking) ตามความสำคัญของคุณลักษณะแต่ละตัว และเลือกคุณลักษณะที่มีระดับความสำคัญสูงสุดตามจำนวนที่ผู้ใช้ระบุ หรืออาจจะระบุเป็นค่าขีดแบ่ง (Threshold) ของคุณลักษณะที่จะเลือกก็ได้ ข้อดีของการเลือกคุณลักษณะแบบนี้คือ การประมวลผลที่รวดเร็วและไม่ขึ้นกับขั้นตอนวิธีการเรียนรู้แบบใด ๆ แต่ข้อเสียที่สำคัญคือ จำนวนคุณลักษณะหรือค่าขีดแบ่งที่เหมาะสมนั้น ไม่สามารถทราบได้

3. การเลือกคุณลักษณะแบบควรรวม (Wrapper)

เป็นขั้นตอนกระบวนการหนึ่งสำหรับคัดเลือกเซตย่อยจากคุณลักษณะทั้งหมดของข้อมูล โดยจะเน้นที่กระบวนการค้นหาเซตย่อยของคุณลักษณะที่เหมาะสมกับขั้นตอนวิธีการเรียนรู้วิธีใดวิธีหนึ่งโดยเฉพาะ ดังนั้นวิธีการเลือกคุณลักษณะแบบนี้ ถือได้ว่าเป็นการเพิ่มประสิทธิภาพของขั้นตอนการเรียนรู้ได้ดีที่สุด แต่ข้อเสียที่สำคัญคือ จะต้องใช้เวลาในการเรียนรู้มาก และเซตย่อยของคุณลักษณะที่เลือกได้ จะเหมาะกับวิธีการเรียนรู้แบบหนึ่ง ซึ่งอาจไม่เหมาะกับวิธีการเรียนรู้แบบอื่น ๆ และปัญหาที่สำคัญอีกประการหนึ่งคือ อาจเกิดปัญหา Overfitting ได้ ตัวอย่างของวิธีการเลือกคุณลักษณะแบบ Wrapper เช่น Forward Selection, Backward Elimination และ Evolutionary Selection

3. การวิเคราะห์ความแม่นยำตรงของโมเดล K-fold cross validation เป็นวิธีการตรวจสอบค่าความผิดพลาดในการคาดการณ์ของโมเดล โดยการแบ่งข้อมูลออกเป็นกลุ่มจำนวน K กลุ่ม (K-fold) ในตอนแรกเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอน นำข้อมูลไปจัดหมวดหมู่ จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่น ๆ ที่เหลือเป็นชุดทดสอบ สลับอย่างนี้ไปเรื่อย ๆ จนครบ K กลุ่ม ในขั้นตอนสุดท้ายจะหาค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของค่าความถูกต้องในแต่ละกลุ่ม วิธีการนี้ข้อมูลทุกตัวจะได้เป็นทั้งชุดทดสอบและชุดสอน [8]

4. หลักการทำงานของ Forward Selection เป็นการใช้หลักการสร้างชุดของคุณลักษณะโดยการใช้การเรียงลำดับ โดยมีขั้นตอนดังนี้

4.1 เริ่มต้นจากเซตว่าง แล้ววนรอบการทำงาน โดยเริ่มต้นจากรอบที่ 1 จะทำการหาค่าคุณลักษณะทั้งหมดมา 1 คุณลักษณะ เพื่อคำนวณหา Gain ของทุกคุณลักษณะ ซึ่งเป็นค่าที่บอกถึงอำนาจในการจำแนก หากคุณลักษณะใดคำนวณได้ค่า Gain สูง ก็จะเลือกเอาคุณลักษณะนั้นเป็นตัวตั้งสำหรับรอบต่อไป

4.2 ทำการจับคู่คุณลักษณะนั้นกับคุณลักษณะอื่น ๆ เพิ่มครั้งละ 1 คุณลักษณะ แล้วก็คำนวณค่า Gain ถ้าค่าใดได้ค่า Gain สูงก็จะถูกเลือกเป็นตัวตั้งสำหรับรอบถัดไป

4.3 วนรอบเพื่อเพิ่มคุณลักษณะขึ้นครั้งละ 1 คุณลักษณะ จนกระทั่งสุดท้ายได้ชุดของคุณลักษณะที่ถูกเลือกออกมา

วิธีการทดลอง

วิธีการวิจัยครั้งนี้แบ่งออกเป็น 3 ขั้นตอน ประกอบด้วย 1) การจัดเตรียมข้อมูล 2) การสร้างตัวแบบทำนาย และ 3) การวัดประสิทธิภาพของตัวแบบทำนาย

1. การจัดเตรียมข้อมูล ข้อมูลที่นำมาใช้ในการวิจัยครั้งนี้ เป็นข้อมูลของนักศึกษาระดับประกาศนียบัตรวิชาชีพ ของสถานศึกษารัฐบาลแห่งหนึ่งในประเทศไทย จำนวน 5,100 ระเบียบ ที่เก็บรวบรวมไว้ตั้งแต่ปีการศึกษา 2550 ถึง 2559 ประกอบด้วย 9 สาขาวิชา ได้แก่ สาขาการบัญชี, สาขาคอมพิวเตอร์ธุรกิจ, สาขาการขาย, สาขาเลขานุการ, สาขาการโรงแรม, สาขาธุรกิจค้าปลีก, สาขากหกรรมการผลิต, สาขาการตลาด, สาขาธุรกิจค้าปลีกทั่วไป ประกอบด้วยคุณลักษณะทั้งสิ้น 27 คุณลักษณะ ได้แก่ กรู๊ปเลือด, ส่วนสูง, น้ำหนัก, จังหวัดของสถานศึกษาเดิม, รายได้บิดา, รายได้มารดา, อาชีพบิดา, อาชีพมารดา, รายได้ผู้ปกครอง, ความสัมพันธ์กับผู้ปกครอง, สถานะบิดา, สถานะมารดา, ประเภทการเข้าศึกษา, รอบการเข้าเรียน, รูปแบบการศึกษา, สถานะการศึกษา, เกรดเฉลี่ยเดิม, สาขาวิชาที่เรียน, ผลการเรียนรู้เฉลี่ยกลุ่มภาษาไทย, ผลการเรียนรู้เฉลี่ยกลุ่มภาษาต่างประเทศ, ผลการเรียนรู้เฉลี่ยกลุ่มสังคม, ผลการเรียนรู้เฉลี่ยกลุ่มวิทยาศาสตร์, ผลการเรียนรู้เฉลี่ยกลุ่มคณิตศาสตร์, ผลการเรียนรู้เฉลี่ยกลุ่มพลศึกษา, ผลการเรียนรู้เฉลี่ยกลุ่มศิลปะ, ผลการเรียนรู้เฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี, ผลการเรียนรู้เฉลี่ยเมื่อจบการศึกษาตามสาขาที่เลือกเรียน

หลังจากเก็บรวบรวมข้อมูลของนักศึกษาแต่ละคนแล้ว ก็จะนำข้อมูลที่ได้มาทำการตรวจสอบความสมบูรณ์ ซึ่งอยู่ในรูปแบบของตาราง Excel และทำการ Clean ข้อมูลเพื่อตัดข้อมูลที่ผิดปกติออก รวมทั้งข้อมูลที่ไม่มีค่าว่าง (Missing Values) จากนั้นทำการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ข้อมูลตามอัลกอริทึมที่เลือกใช้ ดังนี้

1) จัดกลุ่มจังหวัดของสถานศึกษาเดิมแยกตามภูมิภาคเป็น 6 กลุ่ม คือ

- N ภาคเหนือ
- NE ภาคตะวันออกเฉียงเหนือ
- C ภาคกลาง
- E ภาคตะวันออก
- W ภาคตะวันตก
- S ภาคใต้

2) จัดกลุ่มของคะแนนสอบแต่ละกลุ่มสาขา และผลการเรียนของโรงเรียนเดิมเป็น 3 กลุ่ม คือ

- คะแนนเฉลี่ย 3.00 – 4.00 ช่วงผลการเรียนเป็น High (สูง)
- คะแนนเฉลี่ย 2.00 – 2.99 ช่วงผลการเรียนเป็น Medium (ปานกลาง)
- คะแนนเฉลี่ย 1.00 – 1.99 ช่วงผลการเรียนเป็น Low (ต่ำ)

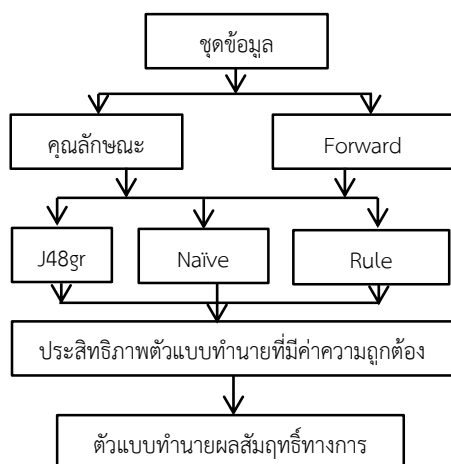
3) จัดกลุ่มผลการเรียนเฉลี่ยตอนจบการศึกษาเป็น 4 กลุ่ม คือ

- คะแนนเฉลี่ย 3.50 – 4.00 ช่วงผลการเรียนเป็น Excellent (ดีเยี่ยม)
- คะแนนเฉลี่ย 3.00 – 3.49 ช่วงผลการเรียนเป็น Good (ดี)
- คะแนนเฉลี่ย 2.50 – 2.99 ช่วงผลการเรียนเป็น Averagely (ปานกลาง)
- คะแนนเฉลี่ย 2.00 – 2.49 ช่วงผลการเรียนเป็น Fairly (พอใช้)

2. การสร้างตัวแบบทำนาย เป็นขั้นตอนการออกแบบและสร้างตัวแบบทำนายทั้ง 3 ตัวแบบ ได้แก่ Decision Tree : J48 graft, Naïve Bayes และ Rule Induction ในการวิจัยจะเปรียบเทียบการทำงาน 2 แบบ คือ

2.1 Simple Classification คือ ไม่มีการเลือกคุณลักษณะของข้อมูลก่อน แต่ใช้อัลกอริทึมในการจำแนกประเภทข้อมูลต่าง ๆ เพื่อทำการประมวลผลเพื่อหาคำตอบอย่างเดียว

2.2 Hybrid Classification คือ การคัดเลือกคุณลักษณะของข้อมูลก่อนแล้วจึงนำไปประมวลผลต่อในอัลกอริทึมการจำแนกประเภทข้อมูลต่าง ๆ โดยในการวิจัยนี้ใช้เทคนิคการค้นหาคุณลักษณะแบบ Forward Selection



รูปที่ 2 รูปแบบการสร้างตัวแบบทำนาย

จากรูปภาพที่ 2 เป็นรูปแบบการสร้างตัวแบบทำนาย โดยขั้นตอนแรกเป็นการนำข้อมูลของนักศึกษาส่งเข้าไปยังตัวแบบการทำนายทั้ง 3 ตัวแบบ ขั้นแรกใช้คุณลักษณะทั้งหมด และทำการวัดประสิทธิภาพตัวแบบการทำนาย ขั้นที่สองใช้การเลือกคุณลักษณะแบบ Forward Select กับตัวแบบทำนายทั้ง 3 ตัวแบบ เช่นกัน และวัดประสิทธิภาพ เพื่อเปรียบเทียบหาค่าความถูกต้องที่สูงที่สุดมาใช้ในการทำนายผลสัมฤทธิ์ทางการเรียน

3. การหาประสิทธิภาพของตัวแบบทำนาย การวิจัยนี้ผู้วิจัยได้ทำการทดสอบประสิทธิภาพตัวแบบทำนายด้วยวิธีการ 10-fold cross validation ซึ่งจะแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่า ๆ กัน คือ 510 ระเบียบ หลังจากนั้นข้อมูลหนึ่งส่วนมาเป็นกลุ่มชุดข้อมูลทดสอบ (Testing Data) ส่วนที่เหลือ 9 ส่วนนำมาใช้เป็นกลุ่มข้อมูลชุดสอน (Training Data) แล้วทำการวนไปจนครบจำนวน 10 รอบ โดยเปลี่ยนกลุ่มชุดข้อมูลทดสอบไปเรื่อย ๆ จนครบ จากนั้นนำผลการทดสอบประสิทธิภาพที่มีค่าความถูกต้อง (Accuracy) ที่สูงที่สุด 2 ค่า มาทำการเปรียบเทียบกับวิธี T-Test เพื่อตรวจสอบว่าค่าความถูกต้องทั้งสองนั้นมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติหรือไม่ และสุดท้ายได้นำผลของตัวแบบที่ให้ผลดีที่สุดมาใช้สำหรับเป็นต้นแบบในการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษาต่อไป

ผลการทดลอง

การวิจัยนี้ใช้ข้อมูลนักศึกษาระดับประกาศนียบัตรวิชาชีพ จำนวน 5,100 ระเบียบ 27 คุณลักษณะ ไปทดสอบกับตัวแบบทำนาย 3 ตัวแบบ ที่ใช้เทคนิคค้นไม่ตัดสินใจ (Decision Tree : J48graft) เทคนิคนาอิวเบย์ (Naïve Bayes) และ เทคนิคกฎการอุปนัย (Rule Induction) เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบทำนาย นอกจากนั้นยังได้ทำการเปรียบเทียบระหว่างการไม่เลือกคุณลักษณะของข้อมูลก่อนและเมื่อเลือกคุณลักษณะโดยการใช้หลักการการทำงานของ Forward Selection ผลการเปรียบเทียบดังตารางที่ 1

จากตารางที่ 1 พบว่า การใช้เทคนิค Decision Tree: J48graft ด้วยการเลือกคุณลักษณะแบบ Forward Selection ค่าความถูกต้อง (Accuracy) ในการจำแนกประเภทข้อมูลได้ดีที่สุด โดยให้ค่าความถูกต้องที่ 83.08% สามารถสรุปเป็นกฎการตัดสินใจได้ 88 ข้อ แสดงดังรูปที่ 3

ตารางที่ 1 เปรียบเทียบประสิทธิภาพของการทำงานของ
ตัวแบบ 10-fold cross validation

Algorithm	Accuracy	
	Simple Classification	Hybrid Classification
Decision tree (J48graft)	81.71*	83.08*
Naïve Bayes	78.25	81.69
Rule Induction	81.25	81.22

J48graft pruned tree

Tech = Low

| Lang = Low: Fairly (781.0/87.0)

| Lang = Medium

| | Thai = Low: Fairly (51.0/11.0)

| | Thai = Medium

| | | Phy = Low: Fairly (3.0)

| | | Phy = Medium: Averagely (24.0/9.0)

| | | Phy = High: Averagely (45.0/17.0)

| | Thai = High

| | | Phy = Low: Fairly (3.0)

| | | Phy = Medium: Averagely (17.0/6.0)

| | | Phy = High: Averagely (29.0/7.0)

| Lang = High

| | Sci = Low: Fairly (16.0/2.0)

| | Sci = Medium

รูปที่ 3 ตัวอย่างกฎการตัดสินใจที่ได้จากต้นไม้ตัดสินใจ
โดยใช้อัลกอริทึม J48graft

จากรูปที่ 3 กฎการตัดสินใจที่ได้ เช่น IF
Tech=Low AND Lang=Low THEN Fairly หมายความว่า
ถ้าผลการเรียนเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี
ต่ำ และผลการเรียนเฉลี่ยกลุ่มภาษาต่างประเทศต่ำ ดังนั้น
ผลการเรียนตอนจบการศึกษาจะอยู่ในระดับพอใช้

จากนั้นเราจะทำการเทียบค่าความถูกต้องการใช้
เทคนิค Decision Tree: J48 graft แบบ Simple
Classification และ แบบ Hybrid Classification ว่ามี
ความแตกต่างกันอย่างมีนัยสำคัญทางสถิติหรือไม่ โดยใช้
T-Test พบว่า การทดสอบทั้งสองแบบมีค่าเท่ากับ 0.018
ซึ่งมีค่าน้อยกว่า 0.05 แสดงว่าค่าความถูกต้องทั้งสองค่ามี
ความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05 ดังตาราง
ที่ 2

ตารางที่ 2 การทดสอบค่าความถูกต้องด้วย T-Test

A	B	C
	0.817+/-0.014	0.831+/-0.010
0.817+/-0.014		0.018*
0.831+/-0.010		

สรุปผล

จากผลการวิจัยสามารถใช้เทคนิคเหมืองข้อมูลใน
การทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับ
อาชีวศึกษา ได้ทั้ง 3 ตัวแบบ ประกอบไปด้วย Decision
Tree : J48 graft, Naïve Bayes และ Rule Induction
แต่ทำการเปรียบเทียบเพื่อให้ได้วิธีการที่ได้ค่าความถูกต้อง
ที่สูงที่สุด จากนั้นจึงทำการเปรียบเทียบหาประสิทธิภาพ
ความถูกต้องในการทำนาย เพื่อหาตัวแบบทำนายที่ให้
ความถูกต้องสูงสุดที่เหมาะสมกับการนำไปใช้ในการ
ทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับ
อาชีวศึกษา ผลการวิจัยพบว่าตัวแบบการทำนายที่ใช้
เทคนิค Decision Tree : J48 graft แบบ Simple
Classification ให้ค่าความถูกต้องร้อยละ 81.71 และ
แบบ Hybrid Classification ร่วมกับการใช้หลักการ
ทำงานของ Forward Selection ให้ค่าความถูกต้องร้อยละ
83.08 ซึ่งมีค่าสูงสุด และผลการทดสอบค่าความ
ถูกต้องทั้งสองโดยใช้ T-Test พบว่าค่าความถูกต้องทั้งสอง
ค่ามีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ .05
แสดงให้เห็นว่าตัวแบบการทำนายที่ใช้เทคนิค Decision

Tree : J48graft แบบ Hybrid Classification ร่วมกับการใช้หลักการทำงานของ Forward Selection เหมาะสมในการนำไปใช้ในการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับประกาศนียบัตรวิชาชีพ ซึ่งสอดคล้องกับงานวิจัยสุกิจ คูชัยสิทธิ์ [4] สุภาวดี จุฑะสุวรรณ [6] อรุณ พันโท [7] และ ไกรรุ่ง เสงพะพรหม [8] ที่ใช้เทคนิคต้นไม้โดยใช้อัลกอริทึม J48graft ให้ประสิทธิภาพได้ดีที่สุด

ส่วนปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ทางการเรียนมีทั้งหมด 9 คุณลักษณะ ได้แก่ ผลการเรียนรู้เฉลี่ยกลุ่มการเรียนอาชีวและเทคโนโลยี ภาษาไทย คณิตศาสตร์ ภาษาต่างประเทศ วิทยาศาสตร์ สังคม สาขาวิชาที่เรียน รอบการเข้าเรียน(เข้า-บ่าย) และรูปแบบการศึกษา(ทวิภาคี-ปกติ) และจากกฎการตัดสินใจที่ได้จากต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48graft ในรูปภาพที่ 3 เราสามารถบอกได้ว่าผลการเรียนเฉลี่ยของกลุ่มสาระใดมีผลต่อผลสัมฤทธิ์ทางการเรียน ซึ่งช่วยให้อาจารย์ที่ปรึกษาและอาจารย์ผู้สอนร่วมกันวางแผนการเรียน และสอนเสริมหรือแนะแนวให้กับนักศึกษาที่ได้คะแนนในแต่ละวิชาในกลุ่มสาระนั้น ๆ

ข้อเสนอแนะ

จากผลการดำเนินงานวิจัย มีข้อเสนอแนะเพื่อดำเนินการวิจัยครั้งต่อไปหรือการนำผลการวิจัยไปประยุกต์ใช้ในประเด็นต่างๆ ดังนี้

1. ในการจัดเก็บข้อมูลไว้ในฐานข้อมูลผู้ที่มีส่วนเกี่ยวข้องควรมีการพิจารณาคุณลักษณะข้อมูลของนักศึกษาที่จำเป็นต้องจัดเก็บ เพื่อให้ได้ข้อมูลที่เป็นประโยชน์มากที่สุด และในการทำงานของโปรแกรมควรมีการกำหนดให้ผู้กรอกข้อมูลต้องกรอกให้ครบทุกช่อง
2. อาจารย์ที่ปรึกษาและผู้ที่เกี่ยวข้องควรนำผลการทำนายผลสัมฤทธิ์ทางการเรียนไปพิจารณาและหาแนวทางแก้ไข

3. การเตรียมข้อมูลก่อนทำการวิเคราะห์นั้นถือว่ามีความสำคัญเป็นอย่างมาก เพราะถ้าทำ Data Cleaning โดยใช้วิธีการตัดข้อมูลที่มีค่าว่าง ซึ่งบางครั้งข้อมูลที่ตัดออกไปนั้นอาจมีผลต่อตัวแบบการทำนาย ในการดำเนินการจึงควรพิจารณาให้ละเอียดรอบคอบ

4. ในการนำเทคนิคเหมืองข้อมูลมาใช้นั้น ควรนำเทคนิคหลาย ๆ เทคนิคมาใช้ เพื่อเปรียบเทียบค่าความถูกต้อง ซึ่งจะทำให้ตัวแบบการทำนายเกิดประสิทธิภาพมากที่สุด

เอกสารอ้างอิง

- [1] สถาบันทดสอบทางการศึกษาแห่งชาติ. (2559). ผลการทดสอบทางการศึกษาระดับชาติด้านอาชีวศึกษา Vocational National Educational Test (V-Net). สืบค้นจาก http://www.niets.or.th/uploads/content_pdf/pdf_1503649895.pdf
- [2] Daniel T. L. and Chantal D. L. (2015). Data mining and predictive analytics. 2nd ed. USA : John Wiley and Sons incorporate.
- [3] พรณิภา บุตรเอก และสุรเดช บุญลือ.(2557). การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาโดยใช้ซอฟต์แวร์เวกเตอร์แมชชีน. Veridian E-Journal Science and Technology Silpakorn University, 1(6): 40-49.
- [4] กิจ คูชัยสิทธิ์ และเพ็ญศิริ มโนมัยสุพรรณ. “การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยพฤติกรรมด้านวินัยในตนเองที่ส่งผลต่อระดับผลการเรียนของนักศึกษาด้วยเทคนิคเหมืองข้อมูล” ใน การประชุมวิชาการ

- ระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 8, วัสดุ.วารสารวิชาการราชภัฏตะวันตก, 9(1): 71-80.
- โรงแรมดิวาน่า พล่าซ่า จังหวัดกระบี่, 2016.
- [5] เสกสรร วิลัยลักษณ์, วิภา เจริญภัณฑารักษ์ และ ดวงดาว วิชาตากุล.(2558).การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียน โรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา. *Veridian E-Journal Science and Technology Silpakorn University*, 2(2): 1-17.
- [6] สุภาวดี จุฑะสุวรรณ และ สมบูรณ์ อเนกฤทธิ์มิ่งคล. รูปแบบการพยากรณ์ผลการทดสอบทางการศึกษาระดับชาติด้านพื้นฐาน (O-Net) โดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษานักเรียนชั้นประถมศึกษาปีที่ 6 โรงเรียนอนุบาลสิงห์บุรี. ใน *ICMSIT 2017: 4 th International Conference on Management Science, Innovation and Technology*, Suan Sunandha Rajabhat University, Bangkok, Thailand. 2560.
- [7] อรุณ พันโท และมนต์ชัย เทียนทอง. (2014). เปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล. *วารสารเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏอุบลราชธานี*, 4(1): 1-11.
- [8] ไกรรุ่ง เสงพะพรหม, สุพจน์ เสงพะพรหม และ สุวิมล มรรควิบุลย์ชัย.(2557). การค้นหาความรู้คุณลักษณะสำคัญของนักศึกษาที่มีต่อผลสัมฤทธิ์ทางการเรียนในกลุ่มโปรแกรมคอมพิวเตอร์ มหาวิทยาลัยราชภัฏนครปฐม ด้วยเทคนิคเหมือง
- [9] เอกสิทธิ์ พัทธวงศ์ศักดิ์.(2557). *การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิ่ง เบื้องต้น*. พิมพ์ครั้งที่ 2. กรุงเทพฯ: เอเชีย ดิจิตอลการพิมพ์.
- [10] Han, J, Kamber Mi. and Pei J. (2012). *Data Mining Concepts and Techniques*. 3rd ed. USA: Morgan Kaufmann.
- [11] Thanaruk Theeramunkong.(2012). *Introduction to Concepts and Techniques in Data Mining and Application to Text Mining*. 2nd ed. Thai: Thammasat Printing House (Tha Prachan Campus).
- [12] เอกสิทธิ์ พัทธวงศ์ศักดิ์. *การคัดเลือก feature (feature selection) ด้วยวิธี Information Gain*. สืบค้นจาก: <http://dataminingtrend.com/2014/data-mining-techniques/feature-selection-information-gain>