เทคนิคเหมืองข้อมูลสำหรับการพยากรณ์ผลสัมฤทธิ์ทางการศึกษาด้วยวิธีการจัดการเรียนการสอนแบบผสมผสาน

DATA MINING TECHNIQUES FOR PREDICTING ACHIEVEMENT OF STUDENTS BY BLENDED LEARNING INSTRUCTION

วีระยุทธ พิมพาภรณ์
หัวหน้าสาขาวิทยาการคอมพิวเตอร์คณะวิทยาศาสตร์ ศรีราชา
มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา
E-mail : werayut.scisrc@gmail.com

พยุง มีสัจ
คณบดี คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
E-mail : pym@kmutnb.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองสำหรับการพยากรณ์ผลสัมฤทธิ์ทางการศึกษาด้วยวิธีการจัดการเรียนการสอนแบบผสมผสานโดยใช้เทคนิคเหมืองข้อมูลทั้งนี้ข้อมูลที่ใช้สำหรับสร้างแบบจำลองได้จากชุดข้อมูลคะแนนการทดสอบของผู้เรียนในระหว่างการศึกษามีทั้งสิ้น 20 ตัวแปร การวิจัยได้แบ่งข้อมูลออกเป็น 2 ชุดข้อมูลโดยใช้วิธีการคัดเลือกคุณลักษณะ (Feature Selection) ด้วยวิธี Information Gain และGain Ratio จากนั้นลดมิติของข้อมูลเพื่อนำไปวัดค่าประสิทธิภาพของแบบจำลองการพยากรณ์โดยใช้วิธี 10-fold Cross Validation ผลการวิจัยพบว่าแบบจำลองสำหรับการพยากรณ์แบบเคเนียเรสเนเบอร์ (KNN) มีค่าความถูกต้องสูงสุดเท่ากับ 86.13% ต้นไม้ตัดสินใจ (Decision Tree) เท่ากับ 81.74% กฎพื้นฐาน (Rule-Base) เท่ากับ 81.67% และนาอีฟเบย์ (Naïve Bays) เท่ากับ 55.05%

คำสำคัญ : แบบจำลองการพยากรณ์ การเรียนแบบผสมผสาน เหมืองข้อมูล การจำแนกข้อมูล

ABSTRACT

The purpose of this research is to develop a data mining technique model for predicting learning achievement of students based on blended learning instruction. The data that was used to create the model was gathered from the learner's exercise scores throughout the semesters. There were 20 variables. The data was separated into 2 sets using the feature selection techniques of information gain (IG) and gain ratio (GR). After the feature selection process, the data was reduced in order to evaluate the prediction model's performance by using the 10-fold cross validation method. The research findings show that the K-Nearest Neighbor Model for prediction is the most accurate model with the accuracy rate of 86.13%, while the Decision Tree Model has the accuracy rate of 81.74%, the Rule Base Model has the accuracy rate of 81.67%, and the Naïve Bays Model has the accuracy rate of 55.05%.

KEYWORDS : Prediction Model, Blended Learning Instruction, Data Mining, Classification

1. บทนำ

ปัจจุบันสภาพแวดล้อมทางด้านเทคโนโลยีการศึกษาได้มีการปรับเปลี่ยนรูปแบบและมีการพัฒนาอย่างต่อเนื่อง เป็นการตอบสนองต่อเทคโนโลยีการสื่อสารและโทรคมนาคมที่มีความก้าวหน้าอย่างรวดเร็ว การนำเทคโนโลยีทางการศึกษาที่ทันสมัยมาประยุกต์ใช้กับการเรียนการสอนในองค์กรการศึกษาอย่างมีประสิทธิภาพนั้น ถือได้ว่าเป็นการสร้างความได้เปรียบทางการแข่งขันขององค์กรทางการศึกษา

การทำเหมืองข้อมูลถือได้ว่าเป็นเทคนิคในการวิเคราะห์ข้อมูลขั้นสูงอย่างหนึ่ง เป็นการนำข้อมูลในฐานข้อมูลขนาดใหญ่มาวิเคราะห์ สื่อค้นองค์ความรู้ และรวบรวมองค์ความรู้ให้อยู่ในรูปของ ฐานความรู้ (Knowledge base) ซึ่งถือเป็นรากฐานสำคัญในการประยุกต์ใช้ในงานด้านต่างๆ และที่สำคัญการทำเหมืองข้อมูลยังถูกนำมาประยุกต์ใช้ในงานด้านการศึกษา โดยเทคนิคที่ได้รับความนิยมและมีความน่าสนใจได้แก่ เทคนิคการจำแนกข้อมูล (Classification) ถือได้ว่าเป็นการแบ่งประเภทของข้อมูล โดยการแบ่งข้อมูลทำได้โดยการใช้ชุดฝึกฝน (Training Data Set) ซึ่งเป็นชุดข้อมูลที่ใช้สำหรับอธิบายและแบ่งประเภทข้อมูล ใช้สำหรับการสร้างต้นแบบ (Model) กระบวนการสร้างต้นแบบในเทคนิคเหมืองข้อมูล สามารถใช้อัลกอริทึมได้หลายแบบ เช่น ต้นไม้ตัดสินใจ (Decision Tree) กฎพื้นฐาน (Rule-Base) นาอีฟเบย์ (Naïve Bays) เคเนียเรสเนเบอร์ (KNN) เป็นต้น อย่างไรก็ตามการเลือกใช้เทคนิคเหมืองข้อมูล ในการพัฒนาแบบจำลองเพื่อการพยากรณ์ จำเป็นต้องคำนึงถึงประสิทธิภาพค่าความแม่นยำในการจำแนกข้อมูล ซึ่งขึ้นอยู่กับกระบวนการเตรียมข้อมูล (Data Preparation) ก่อนการสร้างแบบจำลอง โดยประกอบไปด้วย กระบวนการเลือกแอตทริบิวต์ (Feature Selection) ประเภทแอตทริบิวต์ ชนิดของข้อมูลที่มีความหลากหลาย การตรวจค่าความผิดปกติของข้อมูล (Mladenic & Grobelnik, 1999)

ในปัจจุบันสถาบันการศึกษาโดยเฉพาะในระดับอุดมศึกษา มีการพัฒนาระบบการเรียนการสอนออนไลน์ เพื่อเพิ่มผลสัมฤทธิ์ทางการศึกษาของผู้เรียน (Achievement of students) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) ให้สูงขึ้น อย่างไรก็ตามในช่วงการเปลี่ยนผ่านจากระบบการสอนในชั้นเรียน (face to face) ไปสู่ระบบการเรียนการสอนแบบออนไลน์ (Online Learning) ในลักษณะต่างๆ รูปแบบการเรียนการสอนที่ได้รับความนิยมในปัจจุบันคือ การเรียนการสอนแบบผสมผสาน (Blended Learning) และจากคุณลักษณะของผลสัมฤทธิ์ทางการเรียนที่ว่า ความรู้ ความเข้าใจ และความสามารถของผู้เรียนที่เกิดจากการเรียนการสอนที่วัดผลได้จากการทดสอบ โดยนำผลไป พิจารณาเทียบกับเกณฑ์ที่ได้กำหนดไว้ ซึ่งทำให้สามารถแยกผู้เรียนออกเป็นกลุ่มๆ ตามระดับความรู้ การวิจัยครั้งนี้ผู้วิจัยกำหนดกลุ่มระดับความรู้ตามมาตรฐานเกณฑ์คะแนน 8 ระดับ คือ A (Excellent), B+ (Very Good), B (Good), C (Average) จนถึงระดับ F ซึ่งถือเป็นคลาสของชุดข้อมูลแบบหลายค่า (Multi-Class) ดังนั้นการหาปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ของผู้เรียน ใช้วิธีการวัดผลสัมฤทธิ์ทางการเรียนด้วยการทดสอบ ในการวิจัยครั้งนี้แบ่งการทดสอบออกเป็น 3 ประเภทได้แก่ 1) การทดสอบก่อนเรียน เป็นการทดสอบเพื่อสำรวจความพร้อมของผู้เรียน และวัดความรู้พื้นฐานข้อมูลผู้เรียน 2) การทดสอบประจำบทเรียน โดยทดสอบตามวัตถุประสงค์ของการเรียนรู้ เพื่อสำรวจความรู้ความสามารถของผู้เรียน 3) แบบทดสอบหลังเรียน เป็นการทดสอบหลังจากสิ้นสุดการเรียนการสอน เพื่อสรุปผลการเรียน ข้อมูลที่ได้จากการทดสอบจะนำมาใช้เป็นชุดข้อมูล (Data Set) สำหรับการวิเคราะห์และพัฒนาแบบจำลองการพยากรณ์ (Prediction Model) โดยใช้เทคนิคเหมืองข้อมูล (Data Mining)

งานวิจัยครั้งนี้ เป็นการนำเสนอผลการพัฒนาแบบจำลองการพยากรณ์ (Prediction Model) ผลสัมฤทธิ์ทางการศึกษาผ่านการจัดการเรียนการสอนแบบผสมผสาน (Blended Learning Instructional) ด้วยเทคนิคเหมืองข้อมูล (Data Mining) โดยการวิเคราะห์ข้อมูลการจัดการศึกษาในระดับอุดมศึกษา เพื่อสืบค้นความรู้หรือการคาดคะเนผลสัมฤทธิ์ทางการศึกษาของผู้เรียนเมื่อสิ้นสุดการศึกษาในรายวิชา ซึ่งถือได้ว่าเป็นผลลัพธ์ที่สำคัญในกระบวนการจัดการศึกษาในระดับอุดมศึกษา

2. วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาเทคนิคการทำเหมืองข้อมูลและนำมาประยุกต์ใช้ในงานด้านการศึกษา

2. เพื่อเปรียบเทียบประสิทธิภาพและเลือกเทคนิคที่เหมาะสมในการจำแนกข้อมูลในการจัดการเรียนการสอนด้วย

การทำเหมืองโดยเทคนิคต้นไม้ตัดสินใจกฎพื้นฐานนาอีฟเบย์ เคเนียเรสเนเบอร์

3. เพื่อพัฒนาแบบจำลองในการพยากรณ์ผลสัมฤทธิ์ของผู้เรียนผ่านระบบการจัดการเรียนออนไลน์

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 เหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล ถือเป็นกระบวนการของการกลั่นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อใช้ในการทำนายแนวโน้ม แลพฤติกรรม โดยอาศัยข้อมูลในอดีตเพื่อค้นหารูปแบบความสัมพันธ์ และองค์ความรู้ใหม่จากข้อมูล (Witten, Frank, & Hall, 2011) โดยมีขั้นตอนดังนี้

1. ทำความเข้าใจปัญหา โดยการเลือกข้อมูลให้มีความเหมาะสมกับอัลกอริทึมที่ใช้งาน จำนวนที่ต้องการ และค่าเป้าหมายเพื่อให้ได้ผลลัพธ์ที่ต้องการ

2. ทำความเข้าใจข้อมูล โดยการรวบรวมตรวจสอบความถูกต้องและกำหนดคุณสมบัติที่ต้องการให้กับข้อมูล

3. เตรียมข้อมูล โดยการคัดเลือกข้อมูลเพื่อทำการแปลให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ข้อมูลด้วยเทคนิคต่างๆ

4. สร้างแบบจำลอง โดยแบ่งเป็น 2 ประเภทคือ 1) การสร้างแบบจำลองเพื่อการทำนาย (Predictive Data Mining) เป็นการคาดคะแนนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น โดยใช้ข้อมูลในอดีต 2) การสร้างแบบจำลองเพื่อใช้บรรยาย (Descriptive Data Mining) เพื่อหาแบบจำลองมาอธิบายลักษณะบางอย่างของข้อมูล

3.2 การจำแนกประเภทและการทำนาย (Classification and Prediction)

การจำแนกประเภทเป็นกระบวนการในการหารูปแบบของชุดข้อมูลที่มีความใกล้เคียงกันที่สุดเพื่อใช้สำหรับทำนายชุดข้อมูลว่าอยู่ในประเภทใดของชุดข้อมูลที่ได้มีการแบ่งตามคลาส (Class) แล้วซึ่งชุดข้อมูลที่ถูกแบ่งแล้วได้จากการเรียนรู้ข้อมูลจากข้อมูลฝึกฝน (Training Data Set) โดยแบบจำลองสามารถแสดงได้ในหลายรูปแบบเช่นต้นไม้ตัดสินใจ (Decision Tree) กฎพื้นฐาน (Rule-Base) นาอีฟเบย์ (Naïve Bays) เคเนียเรสเนเบอร์ (KNN) โครงข่ายประสาทเทียม (Artificial Neural Network) ในกระบวนการสร้างแบบจำลอง (Model) สำหรับการจำแนกข้อมูลมีกระบวนการแบ่งออกเป็น 3 ขั้นตอนได้แก่ 1) การสร้างแบบจำลอง (Model Construction) โดยการเรียนรู้ข้อมูลที่ได้กำหนดคลาสไว้แล้วหรือข้อมูลฝึกฝน (Training Data) 2) การประเมินแบบจำลอง (Model Evaluation) โดยอาศัยข้อมูลทดสอบ (Testing Data) ซึ่งเป็นคลาสของข้อมูลนี้จะใช้ในการทดสอบ ด้วยวิธีการเปรียบเทียบกับคลาสที่แบบจำลองหามาเพื่อทดสอบความถูกต้อง 3) การนำแบบจำลองไปใช้งาน (Model Usage) เป็นการนำแบบจำลองที่มีประสิทธิภาพไปใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen Data) โดยแบบจำลองจะให้ผลลัพธ์เป็นการพยากรค่าของคลาสตามที่แบบจำลองกำหนด

3.3 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจเป็นแบบจำลองเพื่อการทำนาย (Predictive model) เป็นแบบจำลองที่มีลักษณะคล้ายกับแบบจำลองที่มีลำดับขั้นของการตัดสินใจ (Rokach & Maimon, 2008) เป็นวิธีการที่ได้รับความนิยมเนื่องจากมีความซับซ้อนน้อยเมื่อเทียบกับอัลกอริทึมอื่นๆ ซึ่งต้นไม้การตัดสินใจ เป็นการนำข้อมูลฝึกฝน (Training Data) มาสร้างแบบจำลองเพื่อพยากรณ์มีการทำงานแบบ การเรียนรู้แบบมีผู้สอน (Supervised Learning) คือสามารถสร้างแบบจำลองได้จากกลุ่มตัวอย่างของข้อมูลได้อัตโนมัติ และสามารถพยากรณ์กลุ่มตัวอย่างของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ หรือ ข้อมูลทดสอบ (Testing Data) การแสดงรูปแบบของ ต้นไม้ตัดสินใจ ประกอบไปด้วย โหลด (Node) แรกสุดเรียกว่า โหนดราก (Root Node) และแตกออกเป็นโหนดย่อยจนโหลดสุดท้ายเรียกว่า โหนดปลาย (Leaf Node) การวิจัยครั้งนี้ผู้วิจัยเลือกใช้วิธีการสร้างต้นไม้ตัดสินใจ ด้วยขั้นตอนวิธีแบบ C4.5 โดยกำหนดให้(Quinlan, 1986, 1990)

$$Entropy(s) = -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

โดยที่ p_i จำนวนความถี่ของ คลาส (Class) i ในโหนด (Node) s เพื่อใช้สำหรับคำนวณค่าความหน้าจะเป็นซึ่งจะเป็นหนึ่งคลาสโดยใช้ค่า Entropy จะมีค่าเป็น 0 และมีค่าเป็น 1 นั้นหมายถึงทุกคลาส ค่าความหน้าจะเป็นที่เท่ากันซึ่งมีโอกาสเกิดขึ้นได้โดยนิยาม

$$p = (k_i | N) \qquad (2)$$

ที่ N เท่ากับค่าทั้งหมดของรูปแบบกลุ่มของคลาส โดย K_i เท่ากับเหตุการณ์ที่เกิดขึ้นใน N การคำนวณหาค่า Information Gain ในการแบ่งกลุ่ม p ในกลุ่ม k ด้วยการวัดผลโดยนำค่า Gain ของ p ที่มีค่าน้อยที่สุดในการรวมกันจากกลุ่มย่อย k แล้วนำไปลบออกจากค่าของ $Entropy(p)$ โดยค่าคุณสมบัติ (Attributes) จะใช้สำหรับการเลือกโหนด (Node) ในการแบ่งกลุ่ม โดยเลือกค่า Gain ที่มีค่ามากที่สุดของ k โดยมีนิยาม (Tan, Steinbach, & Kumar, 2006) ดังนี้

$$GAIN_{split} = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)) \quad (3)$$

จากนั้นจึงทำการแจกแจงข้อมูลในแต่ละกลุ่ม p ตามโหนด ที่ได้ทำการแบ่งไว้แล้วข้างต้นตัวจำแนกข้อมูล C4.5 ได้ทำการขยายส่วนของการจำแนกข้อมูลที่เป็นตัวเลข ด้วยการแบ่งข้อมูลให้เป็นช่วงเพื่อใช้ในการสร้างต้นไม้ตัดสินใจ โดยการแบ่งค่าต่อเนื่องออกเป็นช่วง (Discretization)

3.4 การจำแนกอีฟเบเซียน (Naïve Bayesian Classifier)

เทคนิคการแบ่งกลุ่มแบบเบเซียนอาศัยพื้นฐานจากกฎของเบย์ (Bayes' Rules) ในการเรียนรู้ในการวิจัยครั้งนี้ ผู้วิจัยเลือกใช้อัลกอริธึมประเภทนาอีฟเบเซียน ในการวิเคราะห์ข้อมูลที่ได้จากการจัดการเรียนการสอนแบบผสมผสาน เนื่องจากเป็นอัลกอริธึมที่สามารถเรียนรู้และง่ายต่อความเข้าใจ โดยนาอีฟเบเซียนเป็นการคำนวณหาความหน้าจะเป็นของแต่ละกลุ่มข้อมูล ในรูปแบบของคลาส (Class) เมื่อมีการกำหนดแอททริบิว (Attribute) และค่าข้อมูลในแอททริบิวมาให้ การทำนายจะคำนวณหาค่าความน่าจะเป็นของทุกๆ คลาส (Class Membership Probabilities) มาเปรียบเทียบกัน แล้วเลือกค่าความน่าจะเป็นที่สูงที่สุดของคลาสใดๆ มาเป็นผลของการทำนายเพียงค่าเดียว โดยถือว่าค่าคุณสมบัติในแต่ละตัวไม่ขึ้นต่อกันกับค่าคุณสมบัติอื่น (Class conditional independence) ซึ่งสามารถเขียน $p(a_1, a_2, ...a_n)$ ด้วยผลคูณของความน่าจะเป็นดังสมการ

$$p(a_1,..a_n) = \prod_{i=1}^{n} p(a_i|v_j) \quad (4)$$

3.5 เคเนียเรสเนเบอร์ (K-Nearest Neighbor : KNN)

เทคนิคจำแนกข้อมูลแบบ Non-Parametric และ Parametric มีประสิทธิภาพสำหรับข้อมูลการสอนที่มากในการวิจัยครั้งนี้จึงพิจารณาว่า KNN เป็นตัวเลือกที่จะใช้ในการวิเคราะห์ที่ดีโดยอัลกอริทึมนี้จะจำแนกประเภทข้อมูล โดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัวจากข้อมูลบนชุดข้อมูลตัวอย่างทำงานโดยขึ้นกับระยะทางน้อยที่สุดด้วยวิธีแบบ Euclidean กับชุดข้อมูลทดสอบและจะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัวจากนั้นจะรวมสมาชิดที่ใกล้ที่สุด K ตัวแล้วเลือกคลาสที่สมาชิดส่วนใหญ่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่ให้กับสมาชิกใหม่

3.6 การวัดประสิทธิภาพแบบจำลอง

วิธีวัดประสิทธิภาพแบบจำลอง เป็นการเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมที่ใช้ในการพัฒนาแบบจำลองการพยากรณ์ สามารถทำได้โดยการวัดประสิทธิภาพการจำแนกข้อมูลตามแนวคิดด้านการค้นคืนสารสนเทศ (Information Retrieval) ซึ่งเป็นการวัดค่าต่างๆ ดังนี้ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measurement) (Han & Kamber, 2006)

3.7 งานวิจัยที่เกี่ยวข้อง

การวิจัยเพื่อพัฒนาประสิทธิภาพการเรียน (Yadav & Pal, 2012) ด้วยใช้เทคนิคเหมืองข้อมูล โดยทดสอบข้อมูลต่างๆ ของผู้เรียน เช่น ผลการเรียนในระดับมัธยมศึกษาตอนปลาย วิธีการรับเข้าศึกษา เป็นต้น ด้วยเทคนิคต้นไม้ตัดสินใน ID3 C4.5 และ CARTจากนั้นหาประสิทธิภาพของแบบจำลองโดยวิธี 10-fold cross validation ผลวิจัยพบว่า C4.5 มีค่าความถูกต้องสูงที่สุด (67.7778%) และ ID3 CART ตามลำดับ

การวิจัยเพื่อการลดอัตราการออก (Dropout Rate) ของผู้เรียน(Pal, 2012)ด้วยการใช้การจำแนกนาอีฟเบเซียน (Naïve Bayesian Classifier) โดยใช้โปรแกรม Wake สร้างแบบจำลองในการทำนาย ผลการวิจัยพบว่าแบบจำลองที่สร้างขึ้นให้ค่าการความแม่นยำ (Precision) และค่าการระลึก (Recall) ในการทำนายถูกต้อง (Dropout = Yes) ที่ระดับ 0.917 และ 0.924 ตามลำดับ

4. วิธีดำเนินการวิจัย

4.1 ศึกษาปัญหาและวิเคราะห์ข้อมูล

การวิจัยครั้งนี้เน้นการวิเคราะห์ข้อมูลที่ได้จากการทดสอบเพื่อวัดความรู้ความสามารถด้านพุทธิพิสัย (Cognitive Domain) ที่ส่งผลต่อผลสัมฤทธิ์ทางการเรียน ดังนั้นผู้วิจัยจึงใช้แบบทดสอบวัดความรู้ ความเข้าใจ ที่สอดคล้องกับวัตถุประสงค์และเป้าหมายของรายวิชา โดยแบ่งการทดสอบออกเป็น 3 ประเภทดังนี้

1. แบบทดสอบก่อนเรียน (Pre-test) เป็นแบบทดสอบที่ใช้ประเมินผู้เรียนก่อนดำเนินการเรียนการสอน เพื่อสำรวจความพร้อมของผู้เรียนและวัดความรู้พื้นฐานเดิมของผู้เรียน

2. แบบทดสอบประจำบทเรียน (Formative test) เป็นการทดสอบตามวัตถุประสงค์การเรียนรู้ของแต่ละหน่วยการเรียน เพื่อสำรวจความรู้ ความเข้าใจ ที่ผู้เรียนได้จากการเรียนผ่านระบบการเรียนแบบผสมผสาน

3. แบบทดสอบหลังเรียน (Post-test) เป็นแบบทดสอบที่ใช้ในการประเมินผลการเรียนของผู้เรียนเมื่อสิ้นสุดการเรียนทั้งรายวิชา

โดยชุดข้อมูลมีการจัดเก็บข้อมูลผลการทดสอบของผู้เรียนจำนวน 3 ภาคการศึกษา ได้จำนวนข้อมูล 864 ข้อมูล และมีแอตทริบิวต์ จำนวน 20 แอตทริบิวต์ โดยมีรายละเอียดข้อมูลดังตารางที่ 1

ตารางที่ 1 ลักษณะข้อมูลที่ใช้ในงานวิจัย

รายละเอียด	ข้อมูล	หมายเหตุ
จำนวนข้อมูล (รวม)	864	
จำนวนข้อมูล (2/2553)	173	
จำนวนข้อมูล (1/2554)	332	
จำนวนข้อมูล (2/2554)	359	
จำนวนแอทริบิวต์	20	
ชื่อคลาสที่ใช้จำแนก	Grade	8 ระดับ (A,B+,B,C+,C,D+,D,F)

4.2 การเตรียมข้อมูล (Data Preparation)

งานวิจัยนี้มุ่งเน้นการสร้างแบบจำลองเพื่อใช้สำหรับพยากรณ์ผลสัมฤทธิ์ทางการเรียนของผู้เรียน โดยใช้เทคนิคเหมืองข้อมูล 4 เทคนิค ดังนี้ ต้นไม้ตัดสินใจ (Decision Tree) กฎพื้นฐาน (Rule-Base) นาอีฟเบย์ (Naïve Bays) เคเนียเรสเนเบอร์ (KNN) โดยการนำผลลัพธ์ค่าประสิทธิภาพที่ได้จากการสร้างแบบจำลองด้วยเทคนิคต่างๆ มาเปรียบเทียบประสิทธิภาพ โดยข้อมูลที่ได้มาจะถูกแบ่งออกเป็น 2 ลักษณะคือ 1) ข้อมูลที่ได้จากการทดสอบจะอยู่ในรูปแบบ จำนวนร้อยละ (Percentage) โดยมีการกำหนดให้ผู้ที่สอบผ่านต้องมีคะแนนคิดเป็นร้อยละ 60 ขึ้นไป 2) คะแนนสรุปผลการสอบ จะเป็นการคำนวณค่าเฉลี่ย ($\overline{X}$) ของผลการทดสอบ และจำนวนครั้งที่ผ่านเรียนทดสอบผ่าน 3) คะแนนทั่วไปสำหรับคำนวณเกรด ได้แก่ ค่าคะแนนสอบกลางภาค ปลายภาค เป็นต้น

ผู้วิจัยทำการ Normalization ในส่วนแอทริบิวต์ทุกตัวที่มีลักษณะต่อเนื่อง ให้มีค่าใกล้เคียงกันเพื่อให้ลักษณะเป็นข้อมูลเชิงกลุ่ม สำหรับการแปรผู้วิจัยเลือกใช้วิธีการ min-max normalization (Kotsiantis, Tzelepis, Koumanakos, & Tampakas, 2006) ปรับเสกลของค่าตัวแปรทุกตัวอย่างอยู่ในช่วง 0 - 1 จากนั้นทำการวิเคราะห์หาค่า GAIN ของข้อมูลในแต่ละชุดเพื่อใช้ในการเลือกแอทริบิวต์สำหรับสร้างแบบจำลอง โดยการค่าของ GAIN ในงานวิจัยครั้งนี้ใช้วิธี Information Gain และ Gain ratio ด้วยโปรแกรม Weka รุ่น 3.6.7 และเมื่อพิจารณาจากค่า GAIN ที่คำนวณด้วยวิธี Information Gain แอทริบิวต์ที่มีค่าสูงสุด 8 อันดับ ได้แก่ Score.FinalScore.MitScore. eLearningAvg.AllTest Avg.1-8Test NumOfPass Post-test Score.Labตามลำดับ และเมื่อพิจารณาจากค่า GAIN ที่คำนวณด้วยวิธี Gain ratio แอทริบิวต์ที่มีค่าสูงสุด 8 ลำดับ ได้แก่ Score.FinalScore.MitAvg.AllTest Post-test NumOfPass Score.Lab Avg.1-8Test และ Score.eLearning ตามลำดับ จากความแตกต่างของผลการเลือกแอทริบิวต์ด้วย วิธีการทั้ง 2 แบบ ผู้วิจัยจึงจำแนกชุดข้อมูลฝึกฝน (Training Data Set) ออกเป็น 2 ชุด เพื่อทดสอบประสิทธิภาพในการสร้างแบบจำลองเพื่อการทำนาย รวมทั้งผู้วิจัยจะทำการลดมิติของข้อมูลให้เหลือแบบจำลองละ 5 แอทริบิวต์ โดยรายละเอียดดังแสดงในตารางที่ 2

ตารางที่ 2 ชุดข้อมูลฝึกฝน(Training Data Set)

No.	ชื่อชุดข้อมูล	Attribute	หมายเหตุ
IG01	Info.Gain 01	18	เรียงลำดับตาม Information Gain
IG02	Info.Gain 02	10	
IG03	Info.Gain 03	8	
IG04	Info.Gain 04	6	
IG05	Info.Gain 05	5	
GR01	Gain Ratio 01	18	เรียงลำดับตาม Gain Ratio
GR02	Gain Ratio 02	10	
GR03	Gain Ratio 03	8	
GR04	Gain Ratio 04	6	
GR05	Gain Ratio 05	5	

และเมื่อพิจารณาข้อมูลในชุดข้อมูลจะพบว่าข้อมูลเกรดของผู้เรียนที่ใช้ในการแบ่งคลาสของข้อมูล มีลักษณะของความไม่สมดุลกันของชุดข้อมูล (Imbalanced Data Set) ลักษณะดังกล่าวจะทำให้จะทำให้ประสิทธิภาพในการจำแนกข้อมูลของตัวแบบต่ำลง ผู้วิจัยจึงเลือกใช้วิธีแก้ปัญหาด้วยวิธีการใช้อัลกอริทึม SMOT (Synthetic Minority Oversampling Technique) ซึ่งเป็นการทำให้จำนวนคลาสของข้อมูลที่น้อยเพิ่มขึ้น (Over Sampling) (Chawla et al., 2002) โดยในการวิจัยครั้งนี้ ผู้วิจัยจะปรับค่าคลาส D (ร้อยละ 3 ของข้อมูลทั้งหมด) ให้มีค่าใกล้เคียงกับคลาส B+ (ร้อยละ 22 ของข้อมูลทั้งหมด) ซึ่งเป็นคลาสที่มีค่าร้อยละต่อข้อมูลทั้งหมดมากที่สุด

4.3 การเตรียมข้อมูล (Data Preparation)

ผู้วิจัยเลือกใช้โปรแกรม WEKA version 3.6.7 เพื่อทำการวิเคราะห์ข้อมูลและสร้างแบบจำลอง โดยเทคนิค ต้นไม้ตัดสินใจ (Decision Tree) กฎพื้นฐาน (Rule-Base) นาอีฟเบย์ (Naïve Bays) เคเนียเรสเนเบอร์ (KNN) วิธีการทั้งหมดจะให้ผลลัพธ์ในรูปแบบของกฎหรือโมเดลที่สามารถตัดสินใจ ซึ่งถือเป็นลักษณะของการ การแทนความรู้ (Knowledge Representation) แบบหนึ่ง การทดสอบจะใช้วิธีการเรียนแบบ 10 Cross-validation Folds โดยการแบ่งข้อมูลออกเป็น 10 ส่วน และเลือกข้อมูล 9 ส่วน สำหรับฝึกแบบจำลอง ในอีก 1 ส่วนที่เหลือนำมาทดสอบแบบจำลอง จากนั้นทำขั้นตอนดังกล่าวซ้ำอีก 10 ครั้ง โดยเลือกข้อมูลทดสอบ 9 ส่วน และข้อมูลทดสอบ 1 ส่วน เปลี่ยนจนครบทั้ง 10 ชุด จากนั้นนำข้อมูลวัดได้จากการสร้างแบบจำลองมากเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง

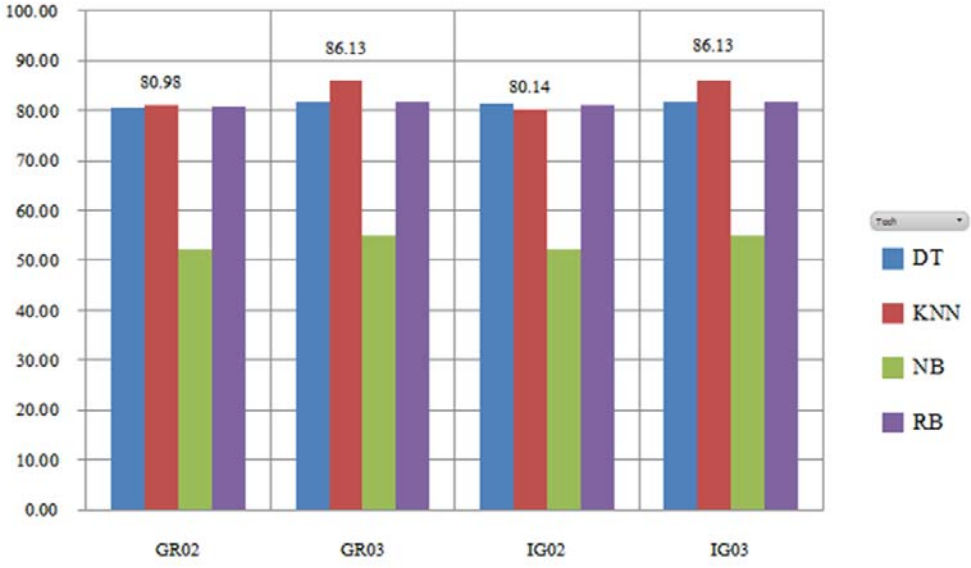
5. ผลการวิจัย

จากผลการทดลองได้ทำการวัดประสิทธิภาพด้วยโปรแกรม Wake ตามวิธีการดำเนินงานวิจัยและคำนวณค่าความถูกต้อง (Accuracy) ในการจำแนกกลุ่มผลการทดสอบและวัดประสิทธิภาพของอัลกอริทึมทั้ง 4 อัลกอริทึมโดยชุดข้อมูลฝึกฝน (Training Data Set) ออกเป็น 2 ชุดและลดมิติของข้อมูลตามขั้นตอนวิธีวิจัยผลการทดสอบแสดงในตารางที่ 3

ผลการทดสอบอัลกอริทึมโดยใช้ชุดข้อมูลทั้ง 10 แบบที่มีความแตกต่างกันตามเงื่อนไขที่กำหนดด้วยวิธี 10 fold-cross validation พบว่าค่าร้อยละของการทำนายผลที่ถูกต้องจากข้อมูลทั้งหมด (Correctly Classified Instances) ที่คำนวณได้จาก Confusion Matrix ของอัลกอริทึมเคเนียเรสเนเบอร์ (KNN) มีประสิทธิภาพมากที่สุดโดยมีค่าประสิทธิภาพเท่ากับ 86.13% ทั้ง IG03 และ GR03 และค่าความแม่นยำ (Precision) เท่ากับ 0.861ค่าความระลึก (Recall) 0.861และค่าความถ่วงดุล (F-Measurement) 0.86รองลงมาคืออัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) มีค่าประสิทธิภาพเท่ากับ 81.74% ทั้ง IG03 และ RG03 และค่าความแม่นยำ (Precision) เท่ากับ 0.818 ค่าความระลึก

ตารางที่ 3 แสดงประสิทธิภาพความถูกต้อง

No.	DT(J48)	KNN(IBk)	NB(Naivebays)	RB(PART)
IG02	81.32	80.14	52.06	80.98
IG03	81.74	86.13	55.05	81.67
GR02	80.56	80.98	52.20	80.84
GR03	81.74	86.13	55.05	81.67



ภาพประกอบที่ 1 การเปรียบเทียบผลการทดสอบความถูกต้องของอัลกอริทึม

(Recall) 0.817 และค่าความถ่วงดุล (F-Measurement) 0.817 เมื่อนำข้อมูลในตารางที่ 4 มาแสดงในรูปของกราฟเพื่อเปรียบเทียบกับอัลกอริทึมอื่นที่ได้มีการสร้างแบบจำลองจะได้กราฟดังภาพประกอบที่ 1

จากผลการวิเคราะห์หาประสิทธิภาพของการสร้างแบบจำลองในแต่ละรูปแบบของชุดข้อมูลพบว่า การลดจำนวนแอทริบิวต์โดยพิจารณาจากค่า Information Gain และGain Ratio ที่เหมาะสม ส่งผลให้ประสิทธิภาพของแบบจำลองดีขึ้น ไม่ว่าจะเลือกจากค่าใดและสามารถสรุปได้ว่าแอทริบิวต์ที่ส่งผลต่อการทำนายผลสัมฤทธิ์ทางการศึกษาของผู้เรียนด้วยวิธีการจัดการเรียนการสอนแบบผสมผสานด้วยเทคนิคเหมืองข้อมูลได้ 8 แอททริบิวตร์และจากแอทริบิวต์ที่ส่งผลต่อการทำนายผลการเรียนจะเห็นได้ว่ามีส่วนประกอบในการทำนายผลสัมฤทธิ์ทางการศึกษาที่มีประสิทธิภาพสูงสุดที่ได้จากระบบการจัดการเรียนออนไลน์แบบผสมผสานจำนวน 5 ส่วนซึ่งหมายถึงระบบการจัดการเรียนออนไลน์สามารถทำให้ผู้เรียนและผู้สอนคาดเดาผลสัมฤทธิ์ของผู้เรียนได้อย่างมีประสิทธิภาพสูงถึง 86.13% ด้วยเทคนิคการทำเหมืองข้อมูลแบบ เคเนียเรสเนเบอร์ (KNN)ทั้งนี้ส่งผลมาจากผู้เรียนมีการเรียนกันแบบกลุ่มทำให้อัลกอริทึมแบบ เคเนียเรสเนเบอร์ ซึ่งใช้วิธีการพิจารณาข้อมูลที่มีลักษณะใกล้เคียงกันหรือระยะห่างข้อข้อมูลมีน้อย ส่งผลให้ค่าประสิทธิภาพของอัลกอริทึมดังกล่าวสูงสุด

6. สรุปผลการทดลอง

งานวิจัยนี้ได้ทำการทดสอบประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างแบบจำลองเพื่อพยากรณ์ผลสัมฤทธิ์ทางการศึกษาของผู้เรียนด้วยวิธีการจัดการเรียนการสอนออนไลน์แบบผสมผสานโดยใช้เทคนิคเหมืองข้อมูลสำหรับการจำแนกประเภทและการทำนาย (Classification and Prediction) ผู้วิจัยเลือกใช้อัลกอริทึมที่ได้รับความนิยมในการทำเหมืองข้อมูลจำนวน 4 อัลกอริทึมได้แก่เช่น ต้นไม้ตัดสินใจ (Decision Tree) กฎพื้นฐาน (Rule-Base) นาอีฟเบย์ (Naïve Bays) เคเนียเรสเนเบอร์ (KNN) กับชุดข้อมูลผู้เรียน 864 ข้อมูลและทำการคำนวณหาค่าGain เพื่อเลือกแอทริบิวต์ที่ส่งผลต่อคลาสที่ได้กำหนดไว้ในขั้นตอนการทดสอบหาประสิทธิภาพของแต่ละอัลกอริทึมผู้วิจัยใช้วิธี 10 fold-cross validation เพื่อใช้สำหรับวัดค่าประสิทธิภาพของแบบจำลองผลการทดสอบพบว่าอัลกอริทึม KNN มีประสิทธิภาพในการจำแนกข้อมูลได้ดีที่สุดเมื่อลดมิติของข้อมูลให้เหลือ 8 มิติจาก 16 มิติโดยค่าค่าความถูกต้องวัดได้ 86.13% รองลงมาคือต้นไม้ตัดสินใจ (Decision Tree) มีค่าประสิทธิภาพเท่ากับ 81.74% และในการวิจัยยังพบอีกว่าการจัดการเรียนการสอนออนไลน์แบบผสมผสารมีผลต่อปัจจัยในการพยากรณ์ผลสัมฤทธิ์ทางการศึกษามากถึง 5 ใน 8 ส่วนของประสิทธิ์ภาพในการพยากรณ์ผลสัมฤทธิ์ทางการเรียน

ผลจากการใช้เทคนิคเหมืองข้อมูลเพื่อการพัฒนาแบบจำลองการพยากรณ์จะเป็นประโยชน์ในการประยุกต์ใช้ในกระบวนการให้คำแนะนำผู้เรียนในการวางแผนการเรียนผ่านระบบผู้เชี่ยวชาญหรือระบบการพยากรณ์ผลสัมฤทธิ์ของผู้เรียนต่อไป

7. เอกสารอ้างอิง

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. "SMOTE: synthetic minority over-sampling technique." **J. Artif. Int. Res.,** 16(1), 321-357.

Han, J., & Kamber, M. 2006. Data Mining: **Concepts and Techniques.** Elsevier.

Kotsiantis, S., Tzelepis, D., Koumanakos, E., & Tampakas, V. 2006. "Forecasting Fraudulent Financial Statements Using Data Mining," **International Journal of Computer Science,** 1(2), 99-107.

Mladenic, D., & Grobelnik, M. 1999. **Feature Selection for Unbalanced Class Distribution and Naive Bayes.** Paper presented at the Proceedings of the Sixteenth International Conference on Machine Learning.

Pal, S. 2012. Mining Educational Data Using Classification to Decrease Dropout Rate of Students. **CoRR, abs/1206.3078.**

Quinlan, J. R. 1986. "Induction of Decision Trees," **Mach. Learn.,** 1(1), 81-106. doi: 10.1023/a:1022643204877

Quinlan, J. R. 1990. "Decision trees and decision-making," **Systems, Man and Cybernetics, IEEE Transactions on,** 20(2), 339-346. doi: 10.1109/21.52545

Rokach, L., & Maimon, O. Z. 2008. **Data mining with decision trees : theroy and applications.** Singapore ; Hackensack, NJ: World Scientific.

Tan, P. N., Steinbach, M., & Kumar, V. 2006. **Introduction to Data Mining:** Pearson Addison Wesley.

Witten, I. H., Frank, E., & Hall, M. A. 2011. **Data Mining: Practical Machine Learning Tools and Techniques.** : Elsevier Science.

Yadav, S. K., & Pal, S. 2012. "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," **CoRR, abs/1203.3832 %U** http://dblp.uni-trier.de/rec/bibtex/journals/corr/abs-1203-3832.

>> วีระยุทธ พิมพาภรณ์

สำเร็จการศึกษา ปริญญาโท วิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาเทคโนโลยีสารสนเทศ จากมหาวิทยาลัยเทคโนโลยีมหานคร ปี พ.ศ. 2552 วิศวกรรมศาสตรบัณฑิต (วศ.บ.) สาขาวิศวกรรมไฟฟ้า มหาวิทยาลัยเทคโนโลยีมหานคร ปี พ.ศ. 2548

ปัจจุบันทำงานในตำแหน่ง หัวหน้าสาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ ศรีราชา มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา

>> พยุง มีสัจ

จบการศึกษาหลักสูตรครุศาสตร์อุตสาหกรรมบัณฑิต สาขาวิศวกรรมไฟฟ้า จากสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือปี พ.ศ. 2537 จบการศึกษา MS และ Ph.D. in Electrical Engineering จาก Oklahoma State University ประเทศสหรัฐอเมริกา ปี พ.ศ. 2541 และ 2545 ตามลำดับ

ปัจจุบันดำรงตำแหน่งคณบดีคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ