แบบจำลองพยากรณ์ผลการเรียนของนักศึกษาจากพฤติกรรมการใช้งานอินเทอร์เน็ต โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยราชภัฏยะลา

A Forecast Model of Student's Learning Performance Based on Internet Using Behaviors by Data Mining Techniques Case Study Yala Rajabhat University

อับดุลเลาะ บากา[1], อรรถพล อดุลยศาสน์[1], อิสมาแอ ล่าเตะเกะ[1], สุลัยมาน เภอโส๊ะ[1], จีรวุธ มุนินทร์นพมาศ[1], อิมรอน แวมง[1], มูฮัมหมัด ปุ[1]
[1]สาขาวิชาคอมพิวเตอร์ คณะวิทยาศาสตร์เทคโนโลยีและการเกษตร มหาวิทยาลัยราชภัฏยะลา
E-mail: abdulloh.b@yru.ac.th

บทคัดย่อ

งานวิจัยนี้นำเสนอแบบจำลองการพยากรณ์ผลการเรียนของนักศึกษา จากข้อมูลพฤติกรรมการใช้งานอินเทอร์เน็ต โดยใช้เทคนิคการทำเหมืองข้อมูลด้วยต้นไม้ตัดสินใจ J48, Naïve Bays, โครงข่ายประสาทเทียม RBF และ MLP กรณีศึกษามหาวิทยาลัยราชภัฏยะลา โดยใช้ข้อมูลพฤติกรรมการใช้งานอินเทอร์เน็ตจากข้อมูลจราจรทางคอมพิวเตอร์ (log file) ช่วงเดือนสิงหาคมถึงเดือนพฤศจิกายน 2558 และข้อมูลผลการเรียนปีการศึกษา 1/2558 จำนวน 4,380 คน เป็นเป็นคลาสเป้าหมาย โดยใช้วิธีการทดลอง 4 รูปแบบเพื่อค้นหารูปแบบพฤติกรรมที่ดีที่สุด ผลลัพธ์ที่ได้แสดงให้เห็นว่าการใช้แบบจำลองการพยากรณ์ผลการเรียนของนักศึกษาโดยใช้เทคนิคการทำเหมืองข้อมูลด้วยต้นไม้ตัดสินใจ J48 ร่วมกับวิธีการกำจัดข้อมูลที่ผิดพลาดและการคัดเลือกคุณลักษณะของแอตทริบิวต์ให้ค่าความถูกต้องสูงที่สุด นอกจากนี้ยังแสดงรูปแบบพฤติกรรมการใช้งานอินเทอร์เน็ตของนักศึกษาที่มีผลการเรียนอยู่ในระดับสูง ปานกลาง และต่ำ ซึ่งสามารถนำไปใช้ประโยชน์ในเรื่องของการบริหารจัดการเครือข่ายอินเทอร์เน็ตภายในมหาวิทยาลัยได้อย่างมีประสิทธิภาพ

คำสำคัญ: การทำเหมืองข้อมูล, ต้นไม้ตัดสินใจ, เครือข่ายประสาทเทียม, แบบจำลองการพยากรณ์ผลการเรียน, พฤติกรรมการใช้งานอินเทอร์เน็ต

Abstract

This paper proposed a forecast model of student's learning performance, based on Internet using behaviors. Four techniques of data mining, J48 decision tree, Naïve bays, RBF and MLP, were studied and compared. Log files recorded Internet using behaviors of students in Yala Rajaphat University from August to November 2015 was used as a case study. In addition, academic outcome of 4380 students in 1/2015 semester was observed simultaneously. We examined and compared those of four techniques to find the best model. The results of this study indicated that the pre-processes prior to enter a data mining process in order to eliminate missing values and to select features were important. The J48 decision tree technique matched with such pre-processes gave highest accurate values, compared to other techniques. Furthermore, the J48 decision tree technique could present patterns, which relate between Internet using behaviour and learning abilities, i.e., high medium and low, of students. These patterns could help administrators better manage networks and Internet use according policies of university.

Key word: Data Mining, Decision Tree, Neural Network, Forecasting Model of Student's Learning Performance, Internet Using Behaviors.

คำนำ

เทคโนโลยีอินเทอร์เน็ตเป็นปัจจัยพื้นฐานหนึ่งในปัจจุบันที่จำเป็นในการดำรงชีวิต โดยเฉพาะด้านการศึกษาในยุคศตวรรษที่ 21 เพื่อใช้ในการเรียนการสอน และการค้นคว้าวิจัย ด้วยจำนวนผู้ใช้งานอินเทอร์เน็ตที่เพิ่มขึ้นอย่างรวดเร็ว ข้อมูลจากเว็บไซต์ wearesocial.com ได้มีการสำรวจผู้ใช้งานอินเทอร์เน็ตในปี 2559 พบว่าทั่วโลกมีผู้ใช้งานอินเทอร์เน็ตประมาณ 3,400 ล้านคน และประเทศไทยมีประมาณ 38 ล้านคน คิดเป็น 56% ของประชากรไทยทั้งหมด (Kemp, 2016) ทำให้นักวิจัยหลายคนสนใจประเด็นพฤติกรรมการใช้งานอินเทอร์เน็ต โดยใช้เฉพาะในสถาบันการศึกษาที่ทุกมหาวิทยาลัยมีบริการอินเทอร์เน็ตให้กับนักศึกษาได้ค้นคว้าข้อมูลด้านการศึกษาและอื่นๆ งานวิจัยของ ชเนตตี สยนานนท์ (2555) และ พัชราภรณ์ หงส์สิบสอง (2557) ได้ข้อสรุปที่คล้ายกันพบว่า ด้านการเรียนการสอนเป็นปัจจัยอันดับที่ 1 ที่มีผลต่อการใช้งานอินเทอร์เน็ตของนักศึกษา และนักศึกษาที่มีผลสัมฤทธิ์ทางการเรียนแตกต่างกันมีพฤติกรรมการใช้อินเทอร์เน็ตที่แตกต่างกันในด้านการศึกษา และความบันเทิง โดยใช้ข้อมูลจากแบบสอบถาม ซึ่งข้อเสียข้อหนึ่งของแบบสอบก็คือ บางครั้งผู้กรอกแบบสอบถามไม่ได้กรอกข้อมูลตามความเป็นจริงที่เกิดขึ้น ด้วยประเทศไทยมี พ.ร.บ. คอมพิวเตอร์ฉบับแรกปี 2550 ที่บังคับให้ผู้ให้บริการอินเทอร์เน็ตในประเทศไทยต้องมีการจัดเก็บข้อมูลจราจรทางคอมพิวเตอร์ (traffic log หรือ log file) อย่างน้อย 90 วัน เพื่อใช้เป็นเครื่องมือในการตรวจสอบการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ ซึ่งทำให้เกิดฐานข้อมูลขนาดใหญ่ แต่มีองค์กรส่วนน้อยเท่านั้นที่สามารถประยุกต์ใช้และนำข้อมูลเหล่านี้ไปใช้ประโยชน์ในด้านอื่นๆ เช่น เพื่อศึกษาพฤติกรรมการซื้อขายสินค้าออนไลน์ของลูกค้า เพื่อศึกษาพฤติกรรมในการโจมตีเครือข่ายอินเทอร์เน็ตของผู้ไม่หวังดี และเพื่อวิเคราะห์พฤติกรรมการใช้งานอินเทอร์เน็ตของนักศึกษาในสถาบันการศึกษา เป็นต้น อีกทั้งข้อมูล log file เป็นข้อมูลที่เกิดขึ้นจากการใช้งานจริงและมีลำดับของเหตุการณ์ที่ชัดเจน ไม่มีการแต่งเติมข้อมูล ซึ่งจะทำให้ผลการวิเคราะห์ข้อมูลมีความน่าเชื่อถืออย่างมาก (Greiff และคณะ, 2015) ซึ่งเหมืองข้อมูลเป็นวิธีการหนึ่งที่ได้รับความนิยมอย่างสูง เพราะสามารถใช้ประโยชน์จากฐานข้อมูลขนาดใหญ่นี้ เพื่อสร้างแบบจำลองการพยากรณ์พฤติกรรมจากข้อมูลต่างๆ ที่เกิดขึ้นในอดีตเพื่อทำนายเหตุการณ์ที่จะเกิดขึ้นในอนาคต (อับดุลเลาะ บากา และคณะ, 2555) ปัจจุบันได้มีการนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับด้านการศึกษามากยิ่งขึ้น โดยเฉพาะการวิเคราะห์ปัจจัยที่มีผลต่อผลการเรียนและประเมินประสิทธิภาพของนักศึกษา (Shahiri และคณะ, 2015) ดังนั้นผู้วิจัยจึงเห็นความสำคัญของฐานข้อมูลขนาดใหญ่ที่มีการจัดเก็บข้อมูล log file มาใช้ประโยชน์เพื่อวิเคราะห์ปัจจัย รูปแบบ และพฤติกรรมของนักศึกษาที่มีผลการเรียนอยู่ในระดับ สูง ปานกลาง และต่ำ เพื่อเป็นแนวทางให้สถาบันการศึกษาสามารถกำหนดนโยบายในการบริหารจัดการเครือข่ายอินเทอร์เน็ตให้เหมาะสมและคุ้มค่ากับงบประมาณที่เสียไป พร้อมทั้งสอดคล้องกับพันธกิจของสถาบันการศึกษาได้อย่างมีประสิทธิภาพ

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

Decision Tree J48

ต้นไม้ตัดสินใจเป็นวิธีการหนึ่งที่นิยมใช้ในการทำเหมืองข้อมูลเพื่อการพยากรณ์พฤติกรรมข้อมูลต่างๆ ที่จะเกิดขึ้นในอนาคต เพราะง่ายต่อการทำความเข้าใจและมีความรวดเร็วในการประมวลผลข้อมูล (Corani และ Guariso, 2005) มีลักษณะโครงสร้างเป็นต้นไม้ แต่ละโหนดแสดงคุณลักษณะที่ใช้ในการทดสอบ แต่ละกิ่งแสดงผลในการทดสอบ และโหนดใบแสดงคำตอบที่กำหนดไว้ แสดงดังสมการที่ (1) (Perveen และคณะ, 2016)

$$\text{Gain\_Ratio}(D, A) = \frac{\text{Entropy}(D)\sum_{j=1}^{i}(P_j * \text{Entropy}(P_j))}{\text{Splitting\_info}} \quad (1)$$

Naïve Bays

นาอีฟเบย์เป็นเทคนิคการจำแนกข้อมูล โดยมีการตั้งสมมติฐานเพื่อกำหนดให้การเกิดของเหตุการณ์ต่างๆ ที่ใช้ในการจัดกลุ่มนั้นเป็นอิสระต่อกัน ซึ่งจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นของแต่ละความสัมพันธ์ (Baitharu และ Pani, 2016) จุดประสงค์เพื่อต้องการสร้างแบบจำลองที่อยู่ในรูปของความน่าจะเป็น เพื่อหาว่าสมมติฐานใดถูกต้องมากที่สุด ข้อดีของวิธีการแบบเบย์คือสามารถใช้ข้อมูลและความรู้ก่อนหน้านี้ เพื่อใช้ในการเรียนรู้ นอกจากนี้ยังเหมาะสมกับชุดข้อมูลที่มีขนาดใหญ่ ทฤษฎีเบย์ (Bayes' Theorem) แสดงดังสมการที่ (2) (Farid และคณะ,2014)

$$P(H|E) = [P(E|H) \times P(H)]/P(E) \quad (2)$$

กำหนดให้ $P(H)$ คือ ความน่าจะเป็นที่จะเกิดสมมติฐาน H
$P(E)$ คือ ความน่าจะเป็นของชุดข้อมูล E
$P(H|E)$ คือ ความน่าจะเป็นของ H เมื่อทราบ E
$P(E|H)$ คือ ความน่าจะเป็นของ E เมื่อทราบ H

Artificial Neural Networks

โครงข่ายประสาทเทียม มีพื้นฐานการทำงานเลียนแบบมาจากสมองมนุษย์ เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) เพื่อให้คอมพิวเตอร์เรียนรู้สามารถจดจำรูปแบบที่เกิดขึ้นในอดีต แล้วทำนายสิ่งที่เกิดขึ้นในอนาคต จุดเด่นของโครงข่ายประสาทเทียมคือ มีความถูกต้องสูง และรองรับข้อมูลที่ไม่สมบูรณ์หรือมีสิ่งรบกวนได้ (Renno และคณะ, 2016) การทำงานของโครงข่ายประสาทเทียมประกอบด้วย หน่วยประมวลผลย่อยหรือเพอร์เซฟตรอน (Perceptron) หลายๆ หน่วยเชื่อมต่อกันเป็นโครงข่าย โดยโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นประกอบด้วย ชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) แต่ละหน่วยเพอร์เซฟตรอนจะมีการคำนวณฟังก์ชันผลรวม (Summation Function) แสดงดังสมการที่ (3) ส่วนฟังก์ชันกระตุ้น (Activation Function) จะทำหน้าที่แปลงผลลัพธ์จากฟังก์ชันผลรวมให้อยู่ในช่วงที่ผู้ใช้ต้องการ โดยโครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP) จะใช้ฟังก์ชันซิกมอยด์ (Sigmoid Function) แสดงดังสมการที่ (4) ส่วนโครงข่ายประสาทเทียมแบบ Raidial Basis Function (RBF) จะใช้ฟังก์ชันเกาส์เซียน (Gaussian Function) แสดงดังสมการที่ (5) (Roiger และคณะ, 2003)

$$y = \sum_{i=1}^{n} w_i x_i + \beta \quad (3)$$

$$z = \frac{1}{1 + e^{-y}} \quad (4)$$

$$z = e^{-y^2} \quad (5)$$

กำหนดให้ y คือ ผลลัพธ์ของฟังก์ชันผลรวม
x_i คือ ค่าข้อมูลเข้าตัวที่ i
n คือ จำนวนข้อมูลเข้าทั้งหมด
w_i คือ ค่าน้ำหนักของข้อมูลเข้าตัวที่ i
β คือ ค่าโน้มถ่วงเอียง

วิจัยที่เกี่ยวข้อง

ปัจจุบันได้มีการพัฒนางานวิจัยด้านการทำเหมืองข้อมูลเพื่อสร้างแบบจำลองในการพยากรณ์พฤติกรรมหรือเหตุการณ์ต่างๆ มากมาย โดยเฉพาะด้านการศึกษา Kavuk และคณะ (2011) ได้นำเสนอรูปของการตรวจสอบพฤติกรรมที่ผิดจรรยาบรรณของนักเรียนจากข้อมูล log file และ ชเนตตี สยนานนท์ (2555) ได้ศึกษาพฤติกรรมการใช้อินเทอร์เน็ตของนักศึกษาปริญญาตรี จากแบบสอบถามจำนวน 400 คน พบว่าพฤติกรรมการใช้งานอินเทอร์เน็ตของนักศึกษาที่ต่างกัน จะมีผลการเรียนที่แตกต่างเช่นเดียวกัน และ Shahiri และคณะ(2015) ใช้วิธีการในการทำเหมืองข้อมูลเพื่อทำนายปัจจัยสำคัญที่มีผลต่อประสิทธิภาพในการออกแบบและวิธีการดำเนินการทดลองของนักศึกษาในมาเลเซีย ทำให้สามารถยกระดับผลการเรียนได้ดีขึ้น ปี 2015 Greiff และคณะ สร้างรูปแบบวิธีการประเมินผลและวัดทักษะการแก้ปัญหาของนักเรียนที่เรียนสายวิทยาศาสตร์โดยใช้วิธีการทำเหมืองข้อมูล จากข้อมูล log file ของนักเรียนที่เข้าร่วมโครงการ PISA 2012 ต่อมาในปี 2016 Greiff และคณะ มีการนำข้อมูล log file ของนักเรียนในประเทศฟินแลนด์ เพื่อศึกษาและวิเคราะห์รูปแบบของพฤติกรรมการแก้ปัญหาที่ซับซ้อนของนักเรียน จากทั้ง 2 งานวิจัย พบว่ารูปแบบที่ได้มีความน่าเชื่อถือ และสามารถลำดับเหตุการณ์ที่เกิดขึ้นจริง เพราะข้อมูลจาก log file เป็นข้อมูลที่เกิดขึ้นจริงจากพฤติกรรมของผู้ใช้ ไม่มีการปลอมแปลงข้อมูล Crockett และคณะ (2016) นำเสนอวิธีการสร้างแบบจำลองโดยใช้ต้นไม้ตัดสินใจ (decision tree) ในการพยากรณ์รูปแบบพฤติกรรมการเรียนรู้ของผู้เรียน พบว่าแบบจำลองการเรียนรู้ที่ได้สามารถค้นหาความสัมพันธ์ของแอตทริบิวต์ที่เกิดขึ้น ที่มีผลต่อการเรียนรู้ของผู้เรียน และได้ค่าความถูกต้องที่สูงขึ้นด้วย Marbouti และคณะ (2016) นำเสนอแบบจำลองสำหรับการพยากรณ์นักศึกษาที่มีความเสี่ยงที่จะพ้นสภาพการเป็นนักศึกษาในหลักสูตร โดยใช้ผลการเรียนเป็นคลาสเป้าหมาย (target class) ทำให้อาจารย์ผู้สอนและนักศึกษาสามารถเตรียมการกับเหตุการณ์ที่อาจจะเกิดขึ้นในอนาคตได้ โดยใช้โครงข่ายประสาทเทียม (ANN: Artificial Neural Network) ต้นไม้ตัดสินใจ (Decision Tree) และนาอีฟเบย์ (Naïve Bays) ในการสร้างแบบจำลอง และมีการคัดเลือกคุณลักษณะของแอตทริบิวต์ (Feature Selection) เพื่อลดขนาดของข้อมูล และเพิ่มค่าความถูกต้องให้กับแบบจำลองที่ได้

วิธีการดำเนินการวิจัย

ขั้นตอนในการสร้างแบบจำลองการพยากรณ์ผลการเรียนของนักศึกษา จากข้อมูลพฤติกรรมการใช้งานอินเทอร์เน็ต โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษา มหาวิทยาลัยราชภัฏยะลามีทั้งหมด 5 ขั้นตอนแสดงดังภาพที่ 1 โดยมีรายละเอียดดังนี้

ขั้นตอนที่ 1 การรวบรวมข้อมูลที่ใช้ในการวิจัยจะเป็นการรวบรวมข้อมูลจาก 2 แหล่งข้อมูลดังนี้

1.1) ข้อมูลจากศูนย์คอมพิวเตอร์มหาวิทยาลัยราชภัฏยะลาเป็นข้อมูล log file การใช้งานอินเทอร์เน็ตของนักศึกษาทุกชั้นปีในช่วงเดือนสิงหาคม ถึงเดือนพฤศจิกายน พ.ศ. 2558 ซึ่งมีตัวอย่างข้อมูล (Example) 117,907,560 แถว และมีจำนวนแอตทริบิวต์ 34 แอตทริบิวต์ แสดงตัวอย่างข้อมูลแต่ละแถวดังภาพที่ 2

1.2) ข้อมูลจากกองบริการนักศึกษามหาวิทยาลัยราชภัฏยะลาเป็นข้อมูลเกรดของนักศึกษาที่ลงทะเบียนเรียนในปีการศึกษา 1/2558 จำนวน 4,380 คน มีจำนวนแถวข้อมูล 43,546 แถว และมี 7 แอตทริบิวต์ แสดงตัวอย่างข้อมูลดังตารางที่ 1

ขั้นตอนที่ 2 การเตรียมข้อมูลประกอบด้วยขั้นตอนย่อย 3 ขั้นตอนดังนี้

2.1) การคัดเลือกข้อมูล (Data Selection) แบ่งออกเป็น 2 ชุดคือ ชุดที่ 1 ข้อมูล log file ที่มีทั้งหมด 34 แอตทริบิวต์ จะเลือกเฉพาะแอตทริบิวต์ที่มีนัยสำคัญต่อการสร้างแบบจำลองการพยากรณ์ผลการเรียนโดยใช้ข้อมูลพฤติกรรมการใช้งานอินเทอร์เน็ตซึ่งพิจารณาแล้วมีเพียงแค่ 5 แอตทริบิวต์ คือ data, time, catid, sessionid, user ส่วนข้อมูลชุดที่ 2 เป็นข้อมูลผลการเรียนที่มีทั้งหมด 6 แอตทริบิวต์ พิจารณาแล้วเลือกแค่ 2 แอตทริบิวต์ คือ Student_id

และ Grade ที่มีผลต่อการสร้างแบบจำลอง นอกจากนี้ผู้วิจัยได้ทำการคำนวณเกรดเฉลี่ยของนักศึกษาแต่ละคนเฉพาะปีการศึกษา 1/2558 แล้วรวม (join) ชุดข้อมูลทั้งสองโดยใช้รหัสนักศึกษา เป็นคีย์หลักแสดงตัวอย่างดังตารางที่ 2

2.2) การกำจัดข้อมูลที่ผิดพลาด (Missing Value) เช่น ในบางแถวข้อมูล รหัสประเภทของเว็บไซต์ไม่ปรากฏในชุดข้อมูล log file ผู้วิจัยก็จะทำการลบแถวข้อมูลที่ไม่สมบูรณ์ออก

ภาพที่ 1 ขั้นตอนการสร้างแบบจำลอง

2.3) แปลงค่าข้อมูล (Data Transformation) ให้อยู่ในรูปแบบที่เหมาะสมประกอบด้วย 2 ขั้นตอนย่อยคือ ขั้นตอนที่ 1 ผู้วิจัยทำการจัดรูปแบบข้อมูลใหม่เพื่อให้สามารถสร้างแบบจำลองการพยากรณ์ด้วยเทคนิคการทำเหมืองข้อมูล โดยการแปลงข้อมูลแอตทริบิวต์รหัสประเภทของเว็บไซต์ (catid) แต่ละประเภทให้เป็นหัวคอลัมน์ (1 ถึง 56)

แล้วกำหนดค่า True กรณีที่นักศึกษามีการเข้าไปใช้งานเว็บไซต์ และ False กรณีนักศึกษาไม่มีการใช้งานเว็บไซต์ กำหนดต่อหนึ่งครั้งของการเชื่อมต่ออินเทอร์เน็ตเป็นหนึ่งแถวข้อมูล แสดงตัวอย่างการแปลงข้อมูลดังตารางที่ 3 จะทำให้แถวข้อมูลทั้งหมดเหลือเพียง 458,365 แถว ซึ่งจะทำให้อัลกอริทึมทำงานได้เร็วยิ่งขึ้น และขั้นตอนที่ 2 ผู้วิจัยทำการแบ่งช่วงข้อมูล (Data Discretization) ของแอตทริบิวต์ grade จากข้อมูลเชิงปริมาณ (Quantitative data) ให้เป็นเชิงคุณภาพ (Qualitative data) แสดงดังตารางที่ 4

2.4) การคัดเลือกคุณลักษณะ (Feature Selection) ที่ใช้ในการทำเหมืองข้อมูลจะใช้ Information Gain ในประเมินประสิทธิภาพของแต่ละแอตทริบิวต์

ขั้นตอนที่ 3 การแบ่งชุดข้อมูลทดสอบในการสร้างแบบจำลอง ใช้วิธีการแบ่งชุดข้อมูลออกเป็น 10 ชุด ใช้ข้อมูล 9 ชุดในการสอน และข้อมูล 1 ชุดในการทดสอบ โดยชุดข้อมูลแต่ละชุดจะถูกสอน และทดสอบสลับกันไป เรียกวิธีการนี้ว่า 10 Fold Cross Validation

ขั้นตอนที่ 4 การสร้างแบบจำลองการพยากรณ์พฤติกรรมการใช้งานอินเทอร์เน็ตจะใช้ 4 เทคนิคของการทำเหมืองข้อมูล คือ J48, Naïve Bays, RBF และ MLP โดยใช้โปรแกรม WEKA เวอร์ชั่น 3.6

ขั้นตอนที่ 5 การเปรียบเทียบผลการทดลอง โดยใช้วิธีการประเมินค่าความถูกต้อง (Accuracy) ในการพยากรณ์ดังแสดงในสมการที่ (6)

$$\text{ค่าความถูกต้อง} = \frac{\text{จำนวนข้อมูลที่ทำนายถูก}}{\text{จำนวนข้อมูลทั้งหมดในคลาส}} * 100 \quad (6)$$

itime=1439225940 date=2015-08-10 time=23:46:25 devid=FGT1KB3909601132 vd=root type=utm subtype=webfilter action=blocked catid=26 catdesc="Malicious Websites" crlevel=high crscore=60 direction=outgoing dstip=70.186.131.117 dstport=80 eventtype=ftgd_blk group=Authen_Student(LDAP_STD02) hostname=api.browsestudio.com level=warning logid=13056 logver=52 method=domain msg="URL belongs to a denied category in policy" profile=YRU_WebFilterL2 proto=6 rcvdbyte=460 reqtype=direct sentbyte=488 service=HTTP sessionid=1894045339 srcip=10.108.15.116 srcport=57622 url=VQLAchs9BIQnIV8z5UVNGhjOcUVCBJ0MBnFLmVVc7g2VEgakTAIVlC1bGk5A4FmC0Ea user=105568009

ภาพที่ 2 ตัวอย่างข้อมูลจราจรทางคอมพิวเตอร์

ตารางที่ 1 ตัวอย่างข้อมูลเกรดของนักศึกษา

Year	Semester	Student_id	Course_id	Course_name	Credit	Grade
2558	1	105568xxx	3162323	การพาณิชย์อิเล็กทรอนิกส์	3	A
2558	1	105568xxx	3156102	การจัดการทรัพยากรมนุษย์	3	B+

ตารางที่ 2 ตัวอย่างข้อมูลจราจรคอมพิวเตอร์รวมกับข้อมูลเกรด

ชื่อแอตทริบิวต์	คำอธิบาย	รูปแบบข้อมูล
date	วันเดือนปีที่ใช้งานอินเทอร์เน็ต	2015-08-10
time	ช่วงเวลาที่มีการใช้งานอินเทอร์เน็ต	23:46:25
catid	รหัสประเภทของเว็บไซต์	26
sessionid	รหัส session	1894045339
user	รหัสนักศึกษา	105568009
grade	เกรดเฉลี่ย	3.51

ตารางที่ 3 ตัวอย่างข้อมูลที่ผ่านการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมทำเหมืองข้อมูล

date	time	1	2	3	…	56	sessionid	Student_id	Class
2015-08-1	23:46:25	True	False	True	…	False	1894045339	105568xxx	High
2015-08-1	23:46:25	False	False	True	…	False	1894045125	405568xxx	Medium
2015-08-1	23:47:04	True	False	False	…	True	1894045986	405668xxx	Low

ตารางที่ 4 ตัวอย่างการแบ่งข้อมูลระดับของเกรดเฉลี่ย

ช่วงเกรดเฉลี่ย	ระดับของเกรดเฉลี่ย
0.00-2.00	Low
2.01-3.00	Medium
3.01-4.00	High

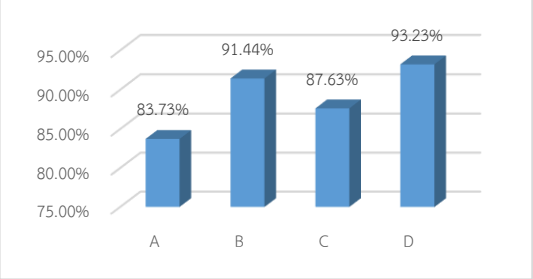
ตารางที่ 5 รูปแบบการทดลองที่ใช้ในการทำเหมืองข้อมูล

ชุดการทดลอง	กำจัด Missing Value	FS	J48	NB	RBF	MLP
A	X	X	✓	✓	✓	✓
B	X	✓	✓	✓	✓	✓
C	✓	X	✓	✓	✓	✓
D	✓	✓	✓	✓	✓	✓

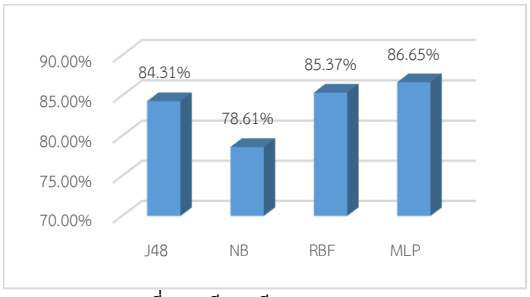
สำหรับชุดรูปแบบที่ใช้ในการทดลองได้แบ่งออกเป็น 4 ชุดการทดลองคือ ชุดการทดลอง A เป็นชุดการทดลองที่ไม่ผ่านกระบวนการกำจัดข้อมูลที่ผิดพลาด (Missing Value) และไม่ผ่านการคัดเลือกคุณลักษณะของข้อมูล (Feature Selection) ชุดการทดลอง B เป็นชุดการทดลองที่ไม่ผ่านกระบวนการกำจัดข้อมูลที่ผิดพลาด แต่ผ่านกระบวนคัดเลือกคุณลักษณะ ชุดการทดลอง C เป็นชุดการทดลองที่ตรงกันข้ามกับชุดการทดลอง B และสุดท้ายชุดการทดลอง D เป็นชุดการทดลองที่ผ่านกระบวนการทั้งการกำจัดข้อมูลที่ผิดพลาด และผ่านการคัดเลือกคุณลักษณะของข้อมูล แสดงรายละเอียดดังตารางที่ 5 โดยทั้ง 4 ชุดการทดลองจะใช้เทคนิคการทำเหมืองข้อมูล 4 แบบคือ J48, NB, RBF และ MLP โดยใช้โปรแกรม WEKA 3.6.13 ในการทดลอง

ผลการทดลอง

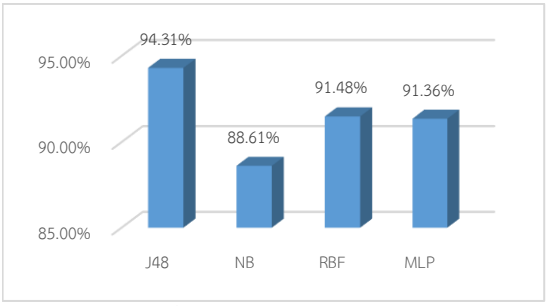
ผลการทดลองเพื่อประเมินประสิทธิภาพของแบบจำลองการพยากรณ์ผลการเรียนของนักศึกษาจากข้อมูลพฤติกรรมการใช้งานอินเทอร์เน็ต โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษา มหาวิทยาลัยราชภัฏยะลา สรุปได้ว่าชุดการทดลอง D ที่ผ่านกระบวนการกำจัด Missing Value และการทำ Feature Selection ได้ค่าความถูกต้องเฉลี่ยสูงสุดเมื่อเทียบกับชุดการทดลองอื่น ส่วนรองลงมาเป็นชุดการทดลอง B, C และ A ตามลำดับ แสดงดังภาพที่ 3 ผลการทดลอง A สรุปได้ว่า MLP ได้ค่าความถูกต้องสูงสุดคือ 93.23% รองลงคือ RBF J48 และ NB คือ 85.37% 84.31% และ 78.61% ตามลำดับ แสดงดังภาพที่ 4 ผลการทดลอง B สรุปได้ว่า J48 ได้ค่าความถูกต้องสูงสุดคือ 94.31% รองลงคือ RBF MLP และ NB คือ 91.48% 91.36% และ 88.61% ตามลำดับ แสดงดังภาพที่ 5 ผลการทดลอง C สรุปได้ว่า J48 ได้ค่าความถูกต้องสูงสุดคือ 89.00% รองลงคือ MLP RBF และ NB คือ 88.96% 88.01% และ 84.56% ตามลำดับ แสดงดังภาพที่ 6 ผลการทดลอง D สรุปได้ว่า J48 ได้ค่าความถูกต้องสูงสุดคือ 96.92% รองลงคือ MLP RBF และ NB คือ 92.71% 92.46% และ 90.85% ตามลำดับ แสดงดังภาพที่ 7



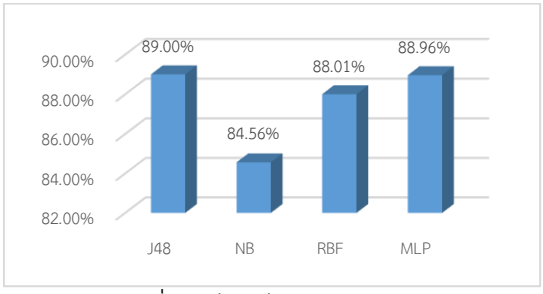
ภาพที่ 3 เปรียบเทียบค่าความถูกต้องเฉลี่ย 4 ชุดการทดลอง



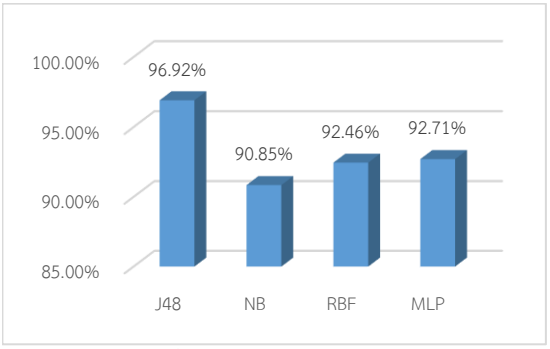
ภาพที่ 4 เปรียบเทียบผลการทดลอง A
โดยใช้ 4 เทคนิคทำเหมืองข้อมล



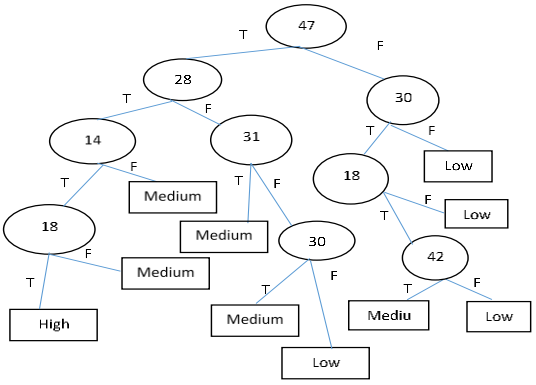
ภาพที่ 5 เปรียบเทียบผลการทดลอง B
โดยใช้ 4 เทคนิคการทำเหมืองข้อมูล



ภาพที่ 6 เปรียบเทียบผลการทดลอง C
โดยใช้ 4 เทคนิคการทำเหมืองข้อมูล



ภาพที่ 7 เปรียบเทียบผลการทดลอง D
โดยใช้ 4 เทคนิคการทำเหมืองข้อมูล



ภาพที่ 8 รูปแบบความสัมพันธ์ที่ได้จากต้นไม้ตัดสินใจ J48

จากภาพที่ 7 ผลการทดลอง D ต้นไม้ตัดสินใจ J48 ที่ได้ค่าความถูกต้องสูงสุดสามารถแปลงเป็นกฎความสัมพันธ์ (Rule) ได้ 10 กฎแสดงดังภาพที่ 8 ซึ่งแต่ละโหนดจะแสดงรหัสประเภทเว็บไซต์ (catid) กิ่งจะแสดงพฤติกรรมการเข้าใช้งานเว็บไซต์หรือไม่ ส่วนใบจะแสดงระดับของผลการเรียนของนักศึกษาซึ่งสามารถอธิบายได้ดังนี้

R1: ถ้า นศ. เข้าใช้งานเว็บไซต์ประเภท social network และ education และ newsgroups and message boards และ streaming media and download แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **สูง**

R2: ถ้า นศ. เข้าใช้งานเว็บไซต์ประเภท social network และ education และ newsgroups and message boards แต่ไม่ได้เข้าใช้งาน streaming media and download แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ปานกลาง**

R3: ถ้า นศ. เข้าใช้งานเว็บไซต์ประเภท social network และ Education แต่ไม่ได้เข้าใช้งาน newsgroups and message boards แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ปานกลาง**

R4: ถ้า นศ. เข้าใช้งานเว็บไซต์ประเภท social network และ shopping and auction แต่ไม่ได้เข้าใช้งาน education แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ปานกลาง**

R5: ถ้า นศ. เข้าใช้งานเว็บไซต์ประเภท social network และ internet radio and tv แต่ไม่ได้เข้าใช้งาน education และ shopping and auction แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ปานกลาง**

R6: ถ้า นศ. เข้าใช้งานเว็บไซต์ประเภท social network แต่ไม่ได้เข้าใช้งาน education และ shopping and auction และ internet radio and tv แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ต่ำ**

R7: ถ้า นศ. ไม่ได้เข้าใช้งานเว็บไซต์ประเภท social network และ internet radio and tv แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ต่ำ**

R8: ถ้า นศ. ไม่ได้เข้าใช้งานเว็บไซต์ประเภท social network และ streaming media and download แต่เข้าใช้งาน internet radio and tv แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ต่ำ**

R9: ถ้า นศ. ไม่เข้าใช้งานเว็บไซต์ประเภท social network และ job search แต่เข้าใช้งาน internet radio and tv และ streaming media and download แล้วทำนายว่า นศ. กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ต่ำ**

R10: ถ้า นศ ไม่เข้าใช้งานเว็บไซต์ประเภท social network แต่เข้าใช้งาน internet radio and tv และ streaming media and download และ job search แล้วทำนายว่า นศ กลุ่มนี้จะมีผลการเรียนอยู่ในระดับ **ปานกลาง**

สรุปผลการวิจัย

การสร้างแบบจำลองการพยากรณ์ผลการเรียนของนักศึกษา จากพฤติกรรมข้อมูลพฤติกรรมการใช้งานอินเทอร์เน็ต โดยการทำเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยราชภัฏยะลา สามารถสรุปได้ดังนี้ 1) ชุดข้อมูลที่ผ่านการกำจัดข้อมูลที่สูญหาย (Missing Value) และผ่านกระบวนคัดเลือกคุณลักษณะของแอตทริบิวต์ (Feature Selection) ก็คือ ชุดการทดลอง D จะให้ค่าเฉลี่ยความถูกต้องสูงที่สุด 2) แบบจำลองที่สร้างด้วยต้นไม้ตัดสินใจ J48 จะให้ค่าเฉลี่ยความถูกต้องสูงสุด และนอกจากนี้ยังสามารถแสดงความสัมพันธ์ของพฤติกรรมการใช้งานอินเทอร์เน็ตเพื่อพยากรณ์ผลการเรียนของนักศึกษาได้อีกด้วย 3) จากรูปพฤติกรรมการใช้งานอินเทอร์เน็ตทำให้มหาวิทยาลัยสามารถกำหนดนโยบายในการบริหารจัดการเครือข่ายอินเทอร์เน็ตได้ถูกต้องตามวัตถุประสงค์ในการใช้งานอินเทอร์เน็ตภายในองค์กรได้อย่างมีประสิทธิภาพ

เอกสารอ้างอิง

ชเนตตี สยนานนท์. (2555). *พฤติกรรมและปัญหาการใช้อินเทอร์เน็ตของนักศึกษาระดับปริญญาตรี*. วิทยานิพนธ์ปริญญาการศึกษามหาบัณฑิต, มหาวิทยาลัยศรีนครินทร์วิโรฒ.

พัชราภรณ์ หงส์สิบสอง และ นันทา เติมสมบัติถาวร. (2557). พฤติกรรมการค้นคว้าข้อมูลบนเครือข่ายอินเทอร์เน็ต กรณีศึกษา: นักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เขตพื้นที่น่าน. *PULINET Journal*, Vol. 1, No. 2, May-August 2014: pp.14-20.

อับดุลเลาะ บากา, วิภาดา เวทย์ประสิทธิ์ และศิริรัตน์ วณิชโยบล.(2555). การสร้างแบบจำลองพยากรณ์น้ำท่วมโดยใช้เทคนิคการทำเหมืองข้อมูลของอำเภอหาดใหญ่. *2012 Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 53-58.

Baitharu, T. R., and Pani, S. K. (2016). Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset. *Procedia Computer Science 85*, pp. 862-870.

Corani, G., and Guario., G. (2005). Coupling Fuzzy Modeling and Neural Networks for River Flood Prediction. *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews*, vol. 35, no. 3, pp .382-388.

Crockett., K., Latham., A., and Whitton., N. (2016). On predicting learning styles in conversational intelligent tutoring systems using fuzzay decision trees. *Int.J. Human-Computer Studies*, pp. 98-115.

Farid, D.M., Zhang, L., Rahman, C.M., Hossain, M., and Strachan, R. (2014). Hybrid decision tree and Naive Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*; 41 (4, Part 2) pp. 1937-1946.

Greiff., S., Wustenberg., S., and Avvisati., F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computer & Education*, pp. 92-105.

Greiff., S., Niepel., C., Schere., R., and Martin., R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, pp. 36-46.

Kavuk., M., Keser., H., and Teker., N. (2011). Reviewing unethical behaviors of primary education students' internet usage. *Procedia – Social and Behavioral Sciences 28*, pp. 1043-1052.

Marbouti., F., Diefes-Dux., H. A., and Krishavan., M. (2016). Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education, pp. 1-15.

Perveen., S., Shahbaz., M., Guergachi., A., and Keshavjee., K. (2016). Perfomance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science, Symposium on Data Mining Applications, SDMA*, 30 March 2016, Riyadh, Saudi Arabia, pp. 115-121.

Renno., C., Petito., F., and Gatto., A. (2016). ANN model for predicting the direct normal irradiance and the global radiation for a solar application to a residential building. *Journal of Cleaner Production*, pp.1298-1316.

Roiger., R. J. and Geatz,. M. W. (2003). *Data Mining a Tutorial-Based Primer*, Pearson Education, Inc., 2003.

Shahiri., A. M., Husain., W., and Rashid., N. A. A. (2015). A Review on Predicting Student's Performance using Data Mining Techniques. *Procedia Computer Science 72*, pp. 414-422.

Kemp., S. (2016). *Digital in 2016*. สืบค้นเมื่อวันที่ 5 เดือน พฤษภาคม พ.ศ. 2559, http://wearesocial.com/special-reports/digital-in-2016.