การเปรียบเทียบอัลกอริทึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษา

On Comparison of Data Mining Algorithms for Analysis of Factors Affecting the Academic Performance of Students

เยาวภา ภารสำเร็จ, [1] จิรัฏฐา ภูบุญอบ, [2] วิรัตน์ พงษ์ศิริ[3]
Yaowapa Pansumret, [1] Jirata Phuboon-ob, [2] Wirat Pongsiri[3]

บทคัดย่อ

งานวิจัยนี้เป็นการนำเสนอการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลของอัลกอริทึมเหมืองข้อมูล 3 แบบ คือ C4.5, Naïve Bayes และ k-Nearest Neighbor อัลกอริทึมที่ให้ค่าประสิทธิภาพสูงที่สุดจะถูกนำมาใช้ในการค้นหาปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษาโดยการลดการนำเข้าทีละตัวแปร ข้อมูลที่ใช้ในการวิจัย เป็นข้อมูลนักศึกษาจากมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสกลนคร ระหว่างปีการศึกษา 2553-2555 จำนวน 4,591 ชุดข้อมูล 17 แอทริบิวต์ ผลการเปรียบเทียบพบว่า แบบ C4.5 ให้ค่าประสิทธิภาพสูงที่สุด ร้อยละ 73.55 ซึ่งมากกว่าแบบ k-Nearest Neighbor และแบบ Naïve Bayes ซึ่งมีค่าร้อยละ 66.63 และร้อยละ 49 ตามลำดับ ผลการค้นหาปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษาพบตัวแปรที่มีความสำคัญเรียงจากสำคัญมากไปหาน้อย คือ ชั้นปี จำนวนพี่น้องที่กำลังศึกษา อายุ จำนวนพี่น้องทั้งหมด และสาขาวิชา นอกจากการค้นหาปัจจัยแล้วสามารถนำกฎการจำแนกข้อมูลที่ได้มาใช้ในการพัฒนาระบบพยากรณ์ระดับผลการเรียนของนักศึกษาได้อีกด้วย

คำสำคัญ: เหมืองข้อมูล, จำแนกข้อมูล, เปรียบเทียบประสิทธิภาพแบบจำลอง

Abstract

This Research presents the performance comparison of classification of data mining algorithms 3 models there are C4.5, Naïve Bayes and k-Nearest Neighbor. The most efficient algorithms was used to determine the factors that affect the academic performance of students by reducing imports by one variable. Data used in the research. As a data of student in Rajamangala University of Technology Isan Sakonnakhon campus. During the 2553-2555 academic years, 4,591 data sets at 17 attribute. The results showed that the performance of C4.5 is 73.55% higher than the k-Nearest Neighbor and Naïve Bayes with 66.63% and 49%. Factors that affect the academic performance of the students is an important variable. Sort of important descending the grade, number of siblings who are studying, age, number of siblings, and all disciplines in

[1]*นิสิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคม 44150 เบอร์โทรศัพท์ 0847910064 Email : hangjai_bee@hotmail.com*

[2]*assistant Professor สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามอำเภอกันทรวิชัย จังหวัดมหาสารคม 44150 เบอร์โทรศัพท์ 089275997 Email : jiratta.p@msu.ac.th*

[3]*รองศาสตราจารย์ สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามอำเภอกันทรวิชัย จังหวัดมหาสารคม 44150 เบอร์โทรศัพท์ 081871177 Email : wirat@msu.ac.th*

[1]*Student, Information technology, Faculty of Informatics, Mahasarakham University.*

[2]*Assistant Professor, Information technology, Faculty of Informatics, Mahasarakham University.*

[3]*Associate Professor, Information technology, Faculty of Informatics, Mahasarakham University.*

addition to the factors can apply the classification data were used to develop predictive learning level of scholar.

Keyword: Data Mining, Classification, Algorithm

บทนำ

การศึกษาเป็นกระบวนการที่มีความสำคัญในการพัฒนาคน ซึ่งคนเป็นตัวแปรที่สำคัญในการพัฒนาชาติ การจัดการศึกษาที่ดีนั้น วัดได้จากคุณภาพของบัณฑิต ทั้งผลสัมฤทธิ์ทางการเรียนที่เป็นตัวชี้วัดเชิงคุณภาพ และระดับผลการเรียนที่เป็นตัวชี้วัดเชิงปริมาณแต่จากการศึกษาข้อมูลระดับผลการเรียนของนักศึกษา มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสกลนคร ระหว่างปี 2553-2555 พบว่า นักศึกษาที่มีระดับผลการเรียนต่ำเสี่ยงต่อการพ้นสภาพ มีจำนวนค่อนข้างสูง[1] และอีกด้านหนึ่ง ข้อมูลร้อยละของบัณฑิตที่ได้งานทำหลังสำเร็จการศึกษา 1 ปี ประจำปีการศึกษา 2555 พบว่ามีทั้งหมด 10 สาขาวิชาที่ไม่ผ่านเกณฑ์เป้าหมายคุณภาพของมหาวิทยาลัยมีค่าเฉลี่ยเพียงร้อยละ 65.91 ซึ่งสาขาวิชาที่ไม่ผ่านเกณฑ์ [2] ส่วนใหญ่เป็นสาขาวิชาด้านวิทยาศาสตร์และเทคโนโลยี และเป็นสาขาวิชาที่มีจำนวนนักศึกษาเยอะที่สุด

ปัจจุบันการทำเหมืองข้อมูล ได้เข้ามามีบทบาทในการวิเคราะห์ข้อมูลด้านการศึกษา เพื่อการวางแผนดูแลนักเรียน-นักศึกษา โดยนำข้อมูลประวัตินักศึกษา ข้อมูลประวัติการเรียน มาใช้ในการวิเคราะห์หาสาเหตุหรือปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ทางการเรียน ระดับผลการเรียนของนักศึกษา ดังเช่น งานวิจัยของ Baradwaj B และคณะ[3] ได้วิเคราะห์ความสำเร็จของนักศึกษาในระดับอุดมศึกษา โดยใช้เทคนิคต้นไม้ตัดสินใจอัลกอริทึม ID3 จากงานวิจัยนี้ทำให้พบว่าการนำเทคนิคทางด้านเหมืองข้อมูลมาใช้ในการวิเคราะห์ข้อมูลทางด้านการศึกษา สามารถนำผลการวิจัยไปใช้ประโยชน์ในการพยากรณ์การแบ่งนักศึกษา เพื่อวางแผนช่วยเหลือนักศึกษาและอาจารย์ในการให้คำแนะนำดูแลเป็นพิเศษสำหรับนักศึกษาที่มีผลการเรียนที่ต่ำกว่าเกณฑ์ส่วนเทคนิควิธีในการทำเหมืองข้อมูลนั้นมีหลากหลายอัลกอริทึมที่นิยมนำมาใช้ดังเช่น ในงานวิจัยของ CheewaprakobkitP[4] ได้ศึกษาปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ในการเรียนของนักศึกษาระดับปริญญาตรีในหลักสูตรนานาชาติ ใช้เทคนิคเหมืองข้อมูลประเภทต้นไม้ตัดสินใจ อัลกอริทึม C4.5 เปรียบเทียบประสิทธิภาพกับโครงข่ายประสาทเทียม จากการทดลองพบว่า C4.5 ให้ความแม่นยำในการทำนายมากกว่าโครงข่ายประสาทเทียม ที่ร้อยละ 85.18 และร้อยละ 83.87 ตามลำดับ แต่ในงานวิจัยของ S. B. Kotsiantis และคณะ[5] ได้ศึกษาประสิทธิภาพของอัลกอริทึมในการทำเหมืองข้อมูลซึ่งได้แก่ Naïve Bayes, 3-NN, Ripper, C4.5 และ Winnow เพื่อพยากรณ์ประสิทธิภาพของนักศึกษาในระบบการศึกษาทางไกล ผลการศึกษาพบว่า อัลกอริทึม Naïve Bayes ให้ค่าประสิทธิภาพที่ร้อยละ 74.70 ซึ่งสูงกว่าประสิทธิภาพของอัลกอริทึมอื่น ๆ รวมถึงอัลกอริทึม C4.5 ด้วย และนอกจากนั้นยังมีงานวิจัยของกริชสมกันธา และคณะ[6] ได้นำอัลกอริทึม k-Nearest Neighbor มาเปรียบเทียบประสิทธิภาพกับอัลกอริทึม Naïve Bayes เพื่อพัฒนาระบบทำนายผลการเรียนนักศึกษาออนไลน์สำหรับประเมินผลการเรียนของนักศึกษา ผลการทดลองแสดงให้เห็นว่าอัลกอริทึม k-Nearest Neighbor ให้ค่าประสิทธิภาพมากกว่าอัลกอริทึม NaïveBayes ที่ร้อยละ 89.44 และร้อยละ 82.21 ตามลำดับ จากการทบทวนงานวิจัยที่กล่าวมาแล้วนั้น จะเห็นได้ว่าเทคนิคเหมืองข้อมูลเป็นเทคนิคที่สามารถวิเคราะห์ข้อมูลด้านการศึกษา และนำผลการวิเคราะห์ไปใช้ประโยชน์ในการวางแผนช่วยเหลือนักศึกษาได้ และอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูลนั้นมีหลากหลายอัลกอริทึม สรุปดังนี้อัลกอริทึม C4.5 เมื่อนำมาเปรียบเทียบกับโครงข่ายประสาทเทียม จะให้ค่าประสิทธิภาพที่สูงกว่า[4] แต่เมื่อนำมาเปรียบเทียบกับ Naïve Bayes กลับพบว่า Naïve Bayesให้ค่าประสิทธิภาพที่สูงกว่า C4.5[5] และเมื่อนำ Naïve Bayes มาเปรียบเทียบกับ k-Nearest Neighbor กลับพบว่า k-Nearest Neighbor ให้ค่าประสิทธิภาพที่สูงกว่า

Naïve Bayes[6] ดังนั้น จึงสรุปไม่ได้ว่าอัลกอริทึมใดที่ให้ค่าประสิทธิภาพสูงที่สุด

จากเหตุผลดังกล่าว จึงทำให้งานวิจัยนี้มุ่งที่จะเปรียบเทียบค่าประสิทธิภาพของอัลกอริทึมทั้ง 3 แบบ เพื่อศึกษาปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษาระดับปริญญาตรีและผลการศึกษาจะสามารถนำไปใช้สำหรับการวางแผนดูแล ให้คำแนะนำนักศึกษาที่คาดว่าจะอยู่ในกลุ่มที่มีระดับผลการเรียนต่ำ ซึ่งหากนักศึกษามีผลการเรียนอยู่ในระดับดี จะสามารถลดอัตราการพ้นสภาพของนักศึกษาลงได้ และอาจทำให้การสมัครเข้าทำงานได้รับการตอบรับเพิ่มมากขึ้น ลดอัตราการว่างงานของบัณฑิตลงได้

วัตถุประสงค์

1. เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลระหว่างอัลกอริทึม C4.5, อัลกอริทึม Naïve Bayes และอัลกอริทึม k-Nearest Neighbor
2. เพื่อศึกษาปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษาระดับปริญญาตรี มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสกลนคร

ทฤษฎีที่เกี่ยวข้อง

1. เหมืองข้อมูล

นักวิจัยได้นิยามคำว่า “เหมืองข้อมูล” ไว้หลายความหมายดังนี้ Daniel[7] กล่าวว่า การทำเหมืองข้อมูลเป็นกระบวนการของการค้นพบรูปแบบความสัมพันธ์ใหม่ที่มีความหมายและแนวโน้มจากข้อมูลจำนวนมากที่เก็บไว้ โดยใช้การจดจำรูปแบบเทคโนโลยีที่เป็นเทคนิคทางสถิติและคณิตศาสตร์ และ กิตติ ภักดีวัฒนกุล[8] ได้ให้ความหมายของเหมืองข้อมูลไว้ว่า เหมืองข้อมูล หมายถึง การวิเคราะห์ข้อมูลเพื่อแยกประเภท จำแนกรูปแบบและความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่หรือคลังข้อมูล และนำสารสนเทศที่ได้ไปใช้ในการตัดสินใจทางธุรกิจดังนั้นจึงสรุปได้ว่า เหมืองข้อมูล หมายถึง การวิเคราะห์ข้อมูลจากฐานข้อมูลขนาดใหญ่เพื่อกลั่นกรอง ค้นหา ด้วยวิธีทางสถิติและคณิตศาสตร์ เพื่อให้ได้สารสนเทศที่มีรูปแบบ (pattern) มีความสัมพันธ์ โดยสารสนเทศนั้นได้ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ และนำสารสนเทศที่ได้มาใช้เพื่อช่วยในวางแผนการตัดสินใจในการบริหารหรือแก้ปัญหาด้านต่าง ๆ ซึ่งถือได้ว่าเป็นเครื่องมือที่ช่วยเพิ่มคุณค่าให้กับข้อมูลที่มีอยู่

2. อัลกอริทึม C4.5

อัลกอริทึม C4.5 เป็นอัลกอริทึมที่ใช้ในการจำแนกประเภท นำเสนอโดย Ross Quinlan[9] ใช้หลักการสร้างต้นไม้โดยคัดเลือกแอทริบิวต์ที่สำคัญที่สุดมาเป็นโหนดราก (Root Node) โดยใช้ค่า Gain Ratio ที่สูงที่สุดเป็นโหนดราก (Root Node) และโหนดถัดไปในการหาค่า Gain Ratio ต้องทำการหาค่า Split Information และค่า Entropy ก่อน ดังนี้

2.1 สมการ Entropy[3]

$$Entropy(s) = \sum_{i=1}^{c} -P_i \log_2 P_i \qquad (1)$$

โดย S คือ แอทริบิวต์ที่นำมาวัดค่า

P_i คือ สัดส่วนของจำนวนสมาชิกในกลุ่มเท่ากับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

2.2 สมการ Information Gain[3]

$$Gain(S,A) = Entropy(s) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

โดย A คือ แอทริบิวต์ A

$|S_v|$ คือ สมาชิกของแอทริบิวต์ A ที่มีค่า v

$|S|$ คือ จำนวนสมาชิกของกลุ่มตัวอย่าง

2.3 สมการ Split Information[3]

$$Split\ Information(S,A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \qquad (3)$$

2.4 สมการ Gain Ratio[3]

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{Split\ Information(S,A)} \qquad (4)$$

3. อัลกอริทึม Naïve Bayes

อัลกอริทึม Naïve Bayes[10,11] เป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพอีกวิธีหนึ่ง โดยที่ใช้งานได้ดีและเหมาะสมกับกรณีของเซตตัวอย่างที่มีจำนวนมากและมีแอทริบิวต์ของตัวอย่างไม่ขึ้นต่อกัน นิยม

นำไปใช้ในด้านการจำแนกประเภทข้อความ การวินิจฉัย เป็นอัลกอริทึมที่ใช้งานง่าย ไม่ซับซ้อน วิธีการวิเคราะห์อัลกอริทึม Naïve Bayes ดังนี้

กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม v_j สำหรับข้อมูลที่มีแอทริบิวต์ทั้งหมด n ตัว $P=\{a_1,a_2.....a_n\}$ หรือใช้สัญลักษณ์ว่า $P\left(a_1,a_2.....a_n\,|V_j\right)$ คือ[11]

$$P(a_1,a_2,.....a_n\,|V_j)=\prod_{i=1}^{n}P(a_1\,|V_j) \qquad (5)$$

โดย $\prod$ คือ ผลคูณของค่าทั้งหมด

i คือ 1, 2, 3,..., n

J คือ 1, 2, 3,..., n

นำค่าที่ได้มาเปรียบเทียบกับกลุ่มที่มีค่าความน่าจะเป็นสูงสุด[11] ดังนี้

$$V_{NB}=\arg_{v_j\grave{o}v}\max P(V_j)x\prod_{i=1}^{n}P(a_i\,|V_j) \qquad (6)$$

จากสมการที่กล่าวมาเขียนเป็นอัลกอริทึมการเรียนรู้แบบเบย์อย่างง่ายได้ดัง Figure 1

•Naive_Bayes_Learn (examples)

FOR EACH target value v Do
$\bar{P}(v_j)\leftarrow$extimate$P(v_j)$
FOR EACH attribute value a of each attribute Do
$\bar{P}(a_i|v_j)\leftarrow$extimate$P(a_i|v_j)$

•Naive_Bayes_Learn (examples)

$$V_{NB}=\arg_{v_j\grave{o}v}\max P(V_j)x\prod_{i=1}^{n}P(a_i\,|V_j)$$

Figure1 Naïve Bayes Learning Algorithms[11]

4. อัลกอริทึม k-Nearest Neighbor

k-Nearest Neighbor[6,7,10,12] เป็นวิธีการในการจัดแบ่งคลาส โดยจะตัดสินใจว่าคลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง เป็นวิธีการหนึ่งสำหรับแก้ปัญหาประมาณค่าฟังก์ชั่นนอนพาราเมตริก สำหรับการจำแนกกลุ่มของข้อมูลที่ไม่เป็นรูปร่างที่ดี หรือข้อมูลที่กระจัดกระจาย โดยทำการตรวจสอบจำนวนบางจำนวนของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด เท่ากับจำนวน k ที่ต้องการ โดยการหาระยะทางที่ใกล้ที่สุด ด้วยสมการ Euclidean Distance[7] ดังนี้

$$d_{Euclidean}(x_i,y_i)=\sqrt{\sum_{k=1}^{n}(x_{i,k}-y_{i,k})^2} \qquad (7)$$

โดย $d_{Euclidean}(x_i,y_i)$ คือ ระยะห่างระหว่างตัวอย่าง x_i และตัวอย่าง y_i

$x_{i,k}$ คือ คุณสมบัติตัวที่ k ของตัวอย่าง x_i

5. การวัดประสิทธิภาพ

การวัดค่าประสิทธิภาพของเทคนิควิธีต่าง ๆ จะต้องทำการเลือกข้อมูลสำหรับเรียนรู้ (Training Set) และข้อมูลสำหรับทดสอบ (Testing Set) ในงานวิจัยนี้เลือกใช้วิธีสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง k กลุ่ม (k-Fold Cross Validation)[13] โดยเริ่มจากการแบ่งชุดข้อมูลออกเป็นส่วน ๆ เท่า ๆ กัน นำข้อมูลบางส่วนมาทำการเรียนรู้ และนำข้อมูลบางส่วนมาทำการทดสอบแบบจำลองที่ได้จากการเรียนรู้ โดยในการทำงานจะทำการเลือกสุ่มข้อมูลออกเป็น k ชุดเท่ากัน ในการทดลองครั้งแรกข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดเรียนรู้ ในการทดลองครั้งที่ 2 ข้อมูลชุดที่ 2 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดเรียนรู้ ทำจนกระทั่งข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบและชุดเรียนรู้ ซึ่งจะมีการทดลองทั้งหมด kครั้ง ในงานวิจัยนี้เลือกใช้ค่า k = 10 โดยอธิบายดัง Figure1

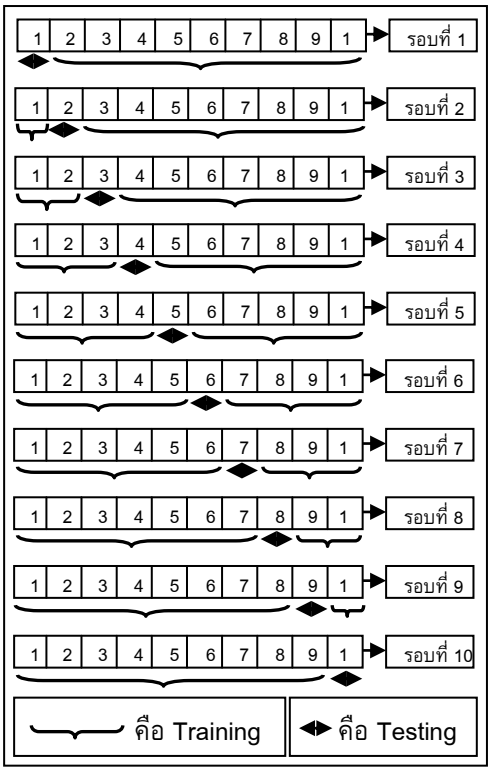

Figure 2 Analysis Data Method

10-Fold Cross Validation

5.1 ค่าความแม่นยำ (Accuracy)

เป็นการทดสอบหาค่าที่ทำนายค่าข้อมูลว่ามีความถูกต้องมากน้อยเพียงใด[14] โดยคิดเป็นค่าร้อยละสูตรการคำนวณ ดังนี้

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (8)$$

โดย TP คือ ค่าที่ทำนายถูกต้องเชิงบวก

TN คือ ค่าที่ทำนายถูกต้องเชิงลบ

FP คือ ค่าที่ทำนายผิดพลาดเชิงบวก

FN คือ ค่าที่ทำนายผิดพลาดเชิงลบ

5.2 ค่าสัมบูรณ์ของค่าคลาดเคลื่อนเฉลี่ย[15] (Mean Absolute Error : MAE)

$$MAE = \frac{\sum_{i=1}^{n} | e_i |}{n} \quad (9)$$

โดย e_i คือ ผลต่างระหว่างค่าข้อมูลจริงและค่าพยากรณ์

n คือ ข้อมูลในการพยากรณ์

วิธีการศึกษาวิจัย

งานวิจัยที่เกี่ยวข้องกับการเลือกใช้อัลกอริทึม C4.5, Naïve Bayes และ k-Nearest Neighbor ดังต่อไปนี้

CheewaprakobkitP[4] ได้ศึกษาปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ในการเรียนของนักศึกษาระดับปริญญาตรีในหลักสูตรนานาชาติ ใช้เทคนิคเหมืองข้อมูลประเภทต้นไม้ตัดสินใจ อัลกอริทึม C4.5 เปรียบเทียบประสิทธิภาพกับโครงข่ายประสาทเทียม จากการทดลองพบว่า C4.5 ให้ความแม่นยำในการทำนายมากกว่าโครงข่ายประสาทเทียม ที่ร้อยละ 85.18 และร้อยละ 83.87 ตามลำดับ

S.B. Kotsiantis และคณะ[5] ได้เปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำเหมืองข้อมูลซึ่งได้แก่ Naïve Bayes, 3-NN, Ripper, C4.5 และ Winnow เพื่อพยากรณ์ประสิทธิภาพของนักศึกษาในระบบการศึกษาทางไกล ผลการศึกษาพบว่าอัลกอริทึม Naïve Bayes ให้ค่าประสิทธิภาพที่ร้อยละ 74.70 ซึ่งสูงกว่าประสิทธิภาพของอัลกอริทึมอื่น ๆ รวมถึงอัลกอริทึม C4.5 ด้วย

กริชสมกันธา และคณะ[6] ได้นำอัลกอริทึม k-Nearest Neighbor มาเปรียบเทียบประสิทธิภาพกับอัลกอริทึม Naïve Bayes เพื่อพัฒนาระบบทำนายผลการเรียนนักศึกษาออนไลน์สำหรับประเมินผลการเรียนของนักศึกษา ผลการทดลองแสดงให้เห็นว่าอัลกอริทึม k-Nearest Neighbor ให้ค่าประสิทธิภาพมากกว่าอัลกอริทึม NaïveBayes ที่ร้อยละ 89.44 และร้อยละ 82.21 ตามลำดับ

วิธีการศึกษาวิจัย

1. ทำความเข้าใจปัญหา

ศึกษาข้อมูลของหน่วยงาน เช่น ศึกษาข้อมูลผลการเรียนของนักศึกษา จากงานส่งเสริมวิชาการและงานทะเบียน ข้อมูลร้อยละของบัณฑิตที่ได้งานทำหลังสำเร็จการศึกษา 1 ปี จากฝ่ายแนะแนวการศึกษาและอาชีพ กองพัฒนานักศึกษา จากปัญหาดังกล่าว ทำให้เห็นปัญหาด้านจำนวนของนักศึกษาที่มีระดับผลการเรียนอยู่ในระดับต่ำ เสี่ยงต่อการพ้นสภาพ และส่งผลให้สมัครเข้าทำงาน ไม่ได้รับการตอบรับ ทำให้ค่าเฉลี่ย

บัณฑิตที่ได้งานหลังสำเร็จการศึกษา 1 ปี ไม่ผ่านเกณฑ์เป้าหมายคุณภาพของหน่วยงาน

2.ทำความเข้าใจข้อมูล

ทำการรวบรวมข้อมูลประวัตินักศึกษา จากงานส่งเสริมวิชาการและงานทะเบียน มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสกลนคร ระหว่างปีการศึกษา 2553-2555 ทำการสำรวจและตรวจสอบข้อมูล ถึงคุณภาพของข้อมูล

3. เตรียมข้อมูล

3.1คัดเลือกข้อมูล โดยวิเคราะห์ความสัมพันธ์ระหว่างแต่ละแอทริบิวต์ (ตัวแปรต้น) กับแอทริบิวต์คลาส (ตัวแปรตาม) ผลการคัดเลือกดัง Table 1

Table 1 Attribute after Selection

Attribute	Description
GENDER	Gender of students.
AGE	Age of students.
CURR	Curriculum of study.
PROGRAM	Program of student, Field of study.
CURR_TYPE	Curriculum type.
REGULAR_YEAR	Regular years.
CLASS	Class of students.
TALENT	Talent of students.
SON_NUM	Son number.
SON_STDNUM	Son study number.
FAT_STATUS	Father status.
FAT_REVENUE	Father revenue.
FAT_OCCUP	Father Occupation.
MOT_STATUS	Mother status.
MOT_REVENUE	Mother revenue.
MOT_OCCUP	Mother Occupation.
PAR_STATUS	Parent's status.
PAR_REVENUE	Parent revenue.
PAR_OCCUP	Parent occupation.
ACC_GPA	Class 0 - 2.50 =LOW 2.51 - 3.00 =MIDDLE 3.01 - 4.00 =HIGH

3.2 ทำความสะอาดข้อมูล หลังจากสำรวจข้อมูลแล้วพบว่า ข้อมูลยังไม่สมบูรณ์ เช่น ค่าว่าง (Missing Value) และมีสิ่งรบกวน (Noisy Data) แก้ไขโดยการแทนค่าข้อมูลที่ผิดปกติดังกล่าว ซึ่งหากค่าข้อมูลเป็นตัวเลข แทนค่าโดยใส่ค่าเฉลี่ย และหากข้อมูลเป็นตัวอักษร แทนค่าโดยใส่ค่าฐานนิยมของค่าข้อมูลในแอทริบิวต์นั้น ๆ

3.3 แปลงข้อมูล เนื่องจากข้อมูลมีทั้งที่เป็นตัวเลข และข้อมูลที่เป็นตัวอักษร ไม่อยู่ในรูปแบบที่สามารถวิเคราะห์ได้ จึงต้องทำการแทนค่าข้อมูลให้อยู่ในรูปแบบที่สามารถวิเคราะห์ได้

4. สร้างแบบจำลอง

นำข้อมูลวิเคราะห์ตามอัลกอริทึม C4.5, Naïve Bayes และ k-Nearest Neighborกำหนด k=2 ซึ่งกำหนดรูปแบบการทดสอบผลลัพธ์ตามวิธี k-fold cross validation กำหนด k=10 ในโปรแกรม Weka 3.7.5

5. ทดสอบแบบจำลอง

ทดสอบแบบจำลอง โดยชุดข้อมูลทดสอบ (Testing Data) โดยวัดประสิทธิภาพด้วยค่าความแม่นยำ (Accuracy) ค่าความสัมบูรณ์ความคลาดเคลื่อนเฉลี่ย (MAE) จากนั้นทำการเปรียบเทียบประสิทธิภาพแต่ละอัลกอริทึม เพื่อให้ได้แบบจำลองจากอัลกอริทึมที่ดีที่สุด

6. ค้นหาปัจจัย

ค้นหาปัจจัยโดย นำอัลกอริทึมที่สร้างแบบจำลองที่ดีที่สุดจากการเปรียบเทียบ มาค้นหาปัจจัยที่ส่งผลต่อระดับผลการเรียน ด้วยการตรวจสอบปัจจัยย้อนกลับโดยลดการนำเข้าทีละแอทริบิวต์[16] ตรวจสอบค่าความแม่นยำ (Accuracy) ที่ลดลงมากที่สุด คือ แอทริบิวต์ที่มีความสำคัญ หรือเป็นปัจจัยที่มีความสำคัญที่สุด

ผลการศึกษา

1. ผลการเปรียบเทียบประสิทธิภาพ

Table 2 Performance of Model

Algorithms	Accuracy	MAE
C4.5	73.55	0.23
k-NN	66.63	0.26
Naïve Bayes	49	0.37

จาก Table 2 แสดงให้เห็นผลการเปรียบเทียบค่าประสิทธิภาพของทั้ง 3 อัลกอริทึม ซึ่งพบว่าอัลกอริทึม

แบบ C4.5 ให้ค่าประสิทธิภาพสูงที่สุด ที่ร้อยละ 73.55 ค่า MAE 0.23 ซึ่งมากกว่าแบบ k-Nearest Neighbor (k=2) ที่ได้ค่าประสิทธิภาพร้อยละ 66.63 ค่า MAE 0.26 และแบบ Naïve Bayes ได้ค่าประสิทธิภาพที่ร้อยละ 49 ค่า MAE 0.37 ดัง Figure 3 และ 4

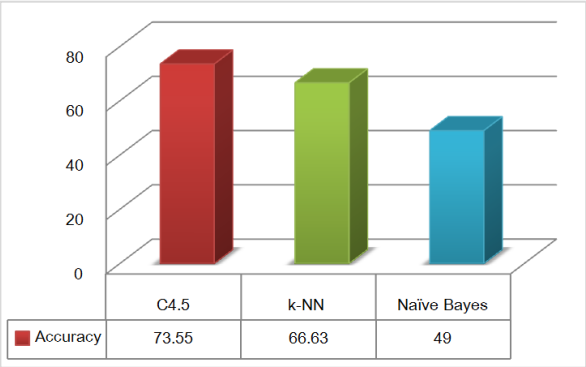

Figure 3 Comparison Accuracy of Model

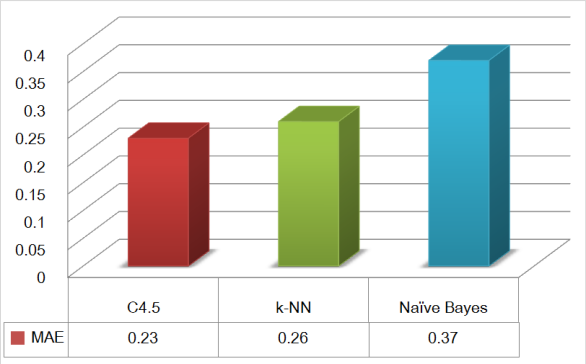

Figure 4 Comparison MAE of Model

2. ผลการค้นหาปัจจัย

จากการนำอัลกอริทึมที่ให้ค่าประสิทธิภาพสูงที่สุดมาค้นหาปัจจัย โดยการลดการนำเข้าทีละแอทริบิวต์แสดงค่าแอทริบิวต์ที่มีความสำคัญ 5 ลำดับ ดังนี้

Table 3 Factors Affecting the Academic Performance of Students

Factors	Accuracy	No.
CLASS	69.30	1
SON_STDNUM	70.35	2
AGE	70.63	3
SON_NUM	70.72	4
PROGRAM	71.03	5

จาก Table 3 แสดงให้เห็นปัจจัยที่ส่งผลกระทบต่อระดับผลการเรียน โดยเรียงจากผลกระทบมากไปหาน้อย ซึ่งพิจารณาจากค่าความแม่นยำที่ลดลงมากที่สุดไปหาน้อย 5 ลำดับ ได้แก่ ชั้นปี จำนวนพี่น้องที่กำลังศึกษา อายุ จำนวนพี่น้องทั้งหมด และสาขาวิชา ดัง Figure 5

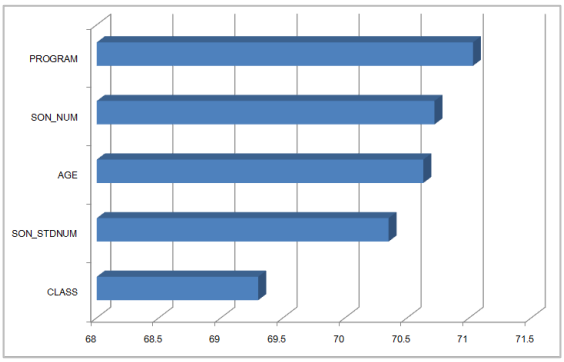

Figure 5 Factors Affecting the Academic Performance of Students

สรุปผลและวิจารณ์

จากการศึกษา เปรียบเทียบอัลกอริทึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษาระดับปริญญาตรีโดยได้นำอัลกอริทึม 3 แบบมาเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูล ได้แก่ C4.5, Naïve bayes, k-Nearest Neighbor ผลการวิจัยสรุปได้ว่า อัลกอริทึม C4.5 ให้ค่าประสิทธิภาพสูงที่สุด ที่ร้อยละ 73.55 มีประสิทธิภาพในการจำแนกข้อมูลดีที่สุดในชุดข้อมูลนี้ และถูกเลือกนำมาใช้ในการวิเคราะห์ปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษานักศึกษาระดับปริญญาตรี ส่วนปัจจัยที่พบ เรียงตามสำคัญมากที่สุดไปหาน้อยที่สุด 5 ลำดับได้แก่ ชั้นปี จำนวนพี่น้องที่กำลังศึกษา อายุ จำนวนพี่น้องทั้งหมด และสาขาวิชา จากผลการวิจัยถ้าต้องการให้ได้ค่าประสิทธิภาพสูงขึ้น ควรมีการจัดเก็บข้อมูลอื่น ๆ ที่ไม่รวมอยู่ในชุดข้อมูลนี้ด้วย เช่น ข้อมูลพฤติกรรมนักศึกษา ข้อมูลความรู้พื้นฐานในแต่ละรายวิชา ซึ่งข้อมูลดังกล่าวไม่มีอยู่ในฐานข้อมูลพื้นฐานที่จัดเก็บในข้อมูลชุดนี้ และมหาวิทยาลัยไม่ได้รวบรวมเก็บเป็นชุดเดียวกันกับชุดข้อมูลนี้ และถ้าจะวิเคราะห์ให้ละเอียดขึ้นนั้น สามารถใช้เครื่องมือในการจัดเก็บ

ข้อมูลเพิ่มขึ้น เช่น แบบสอบถามสาเหตุการได้เกรดเฉลี่ยของนักศึกษา ซึ่งผลการทดลองอาจมีข้อแตกต่างจากงานวิจัยนี้ได้ ดังนั้นผลการทดลองในงานวิจัยนี้สามารถนำมาใช้เป็นข้อมูลในการวางแผนดูแล ให้คำแนะนำสำหรับนักศึกษาที่คาดว่าจะอยู่ในกลุ่มเสี่ยงผลการเรียนต่ำ นอกจากนี้ยังสามารถนำกฎการจำแนกข้อมูลที่ได้จากแบบจำลองที่ดีที่สุดมาพัฒนาระบบพยากรณ์ระดับผลการเรียนของนักศึกษา และนำเทคนิคนี้ไปประยุกต์ใช้ในการวิเคราะห์จำแนกข้อมูลอื่น ๆ ได้

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสกลนคร ที่ให้ความอนุเคราะห์ข้อมูลในการศึกษาครั้งนี้

เอกสารอ้างอิง

1. ฝ่ายสารสนเทศเพื่อการบริหาร กองนโยบายและแผน มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน. สารสนเทศเพื่อการบริหารมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน. [ข้อมูล ณ 30 เมษายน 2556]

2. ฝ่ายแนะแนวการศึกษาและอาชีพ กองพัฒนานักศึกษา มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน. สรุปการติดตามผลผู้สำเร็จการศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลอีสานรุ่นปีการศึกษา 2554 ระดับปริญญาตรี.[ข้อมูล ณ 30 เมษายน 2556]

3. Baradwaj B, Pal S. Mining Education Data to Analyze Students' Performance. In: Proceeding of International Journal of Advanced Computer Science and Applications; United States of America; 2011. p.63-69.

4. Cheewaprakobkit P. Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. In: Proceeding of Second International MultiConference of Engineers and Computer Scientists (IMECS 2013); March 13-15, 2013; Hong Kong. ISSN 2078-0958(Print). ISSN 2078-0966 (online).

5. S. B. Kotsiantis, C. J. Pierakeas, I. D. Zaharakis, P. E. Pintelas. Efficiency of Machine Learning Techniques in Predicting Student's Performance in Distance Learning System. In: Recent Advances in Mechanics and Related Fields University of Patras Greece. 2003.

6. กริช สมกันธา,วิไลพร กุลตั้งวัฒนา,ธีระวัฒน์ หัสโก, จิระพงศ์ รอดชมภู. ระบบทำนายผลการเรียนนักศึกษาออนไลน์โดยใช้เคเนียเร็ซเนเบอร์. ใน : Proceedding of The 3rd International Conference on Knowledge and Smart Technologies. BuraphaUniversity.Chonburi Thailand;2554.

7. Daniel T. Larose. Discovering Knowledge in Data. New York : A JOHN WILEY & SONS, INC., ; 2005.p.2.

8. กิตติ ภักดีวัฒนกุล. คัมภีร์ระบบสนับสนุนการตัดสินใจและระบบผู้เชี่ยวชาญ. กรุงเทพมหานคร : เคทีพี คอมพ์ แอนด์ คอนซัลท์; 2550.

9. Quinlan J.R. Simplifying Decision Trees. Technology Square,Cambridge, MA. 1999.

10. Xindong Wu, Vipin Kumar, J. Ross Quinlan, JoydeepGhosh, Qiang Yang, Hiroshi Motoda et al. Top 10 algorithms in data mining. KnowlInfSyst (2008) 14:1-37.

11. บุญเสริม กิจศิริกุล. อัลกอริทึมการทำเหมืองข้อมูล. รายงานวิจัยฉบับสมบูรณ์ โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ 2545 คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย; 2546.

12. ภาคิน สกุลเจริญ. ระบบแบ่งกลุ่มลูกค้าอุตสาหกรรมโดยใช้เทคนิคเหมืองข้อมูลแบบการหาสมาชิกที่ใกล้กันที่สุด. [ปัญหาพิเศษปริญญาวิทยาศาสตรมหาบัณฑิต]. กรุงเทพมหานคร: บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ; 2554.

13. พัชราภรณ์ ราชประดิษฐ์. ระบบเชี่ยวชาญเพื่อวินิจฉัยโรคข้าว. [ค้นคว้าอิสระ ปริญญาวิทยาศาสตรมหาบัณฑิต] กรุงเทพมหานคร :บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร;2553
14. ธานินทร์ศิลป์จารุ. การวิเคราะห์ข้อมูลทางสถิติ SPSS. กรุงเทพฯธรรมสาร 2553.
15. รวินรัฒน์จารุกำเนิดกนก. การพยากรณ์มูลค่าการส่งออกอัญมณีและเครื่องประดับโดยวิธีอารีม่า. [รายงานส่วนหนึ่งของวิชาแบบฝึกหัดการวิจัยปัญหาเศรษฐกิจปัจจุบัน]เชียงใหม่ : มหาวิทยาลัยเชียงใหม่; 2552.
16. ปรีดา ไวยราบุตร. การประยุกต์อัลกอริทึมแบบแพร่ย้อนกลับในการศึกษาผลกระทบต่อรายได้ครัวเรือน : กรณีศึกษา อำเภอร่องคำ จังหวัดกาฬสินธุ์. [ค้นคว้าอิสระ ปริญญาวิทยาศาสตรมหาบัณฑิต] มหาสารคาม : บัณฑิตวิทยาลัย มหาวิทยาลัยมหาสารคาม ; 2555.