การใช้เทคนิคเหมืองข้อมูลในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อระดับปริญญาตรี
The Use of Data Mining in Selecting Areas of Study for Further Education Opportunity

อนันต์ ปินะเต[1]

Anan Pinate[1]

Received: 18 January 2017 ; Accepted: 6 March 2017

บทคัดย่อ

มหาวิทยาลัยมหาสารคาม มีกระบวนการรับสมัครคัดเลือกบุคคลเข้าศึกษาในระดับปริญญาตรี ซึ่งในแต่ละปีการศึกษาจะมีผู้สมัครเป็นจำนวนมาก ปัญหาที่สำคัญของการรับสมัครคัดเลือกบุคคลเข้าศึกษา คือการมีผู้สมัครจำนวนไม่น้อยที่ไม่มีสิทธิ์เข้าศึกษาในสาขาวิชาที่สมัคร และมีบางสาขาวิชาที่มีผู้สมัครจำนวนน้อยกว่าแผนการรับเข้าศึกษา จากข้อมูลการรับเข้าศึกษาที่ผ่านมาพบว่ามีบางสาขาวิชาที่มีผู้สมัครเป็นจำนวนมากเมื่อเปรียบเทียบกับสัดส่วนแผนการรับเข้าศึกษาที่มีจำนวนน้อยซึ่งผู้สมัครส่วนใหญ่จะเลือกสมัครสาขาวิชาตามความชอบ ความรู้สึกของตนเอง โดยไม่ได้คำนึงถึงผลคะแนนของตนซึ่งเมื่อพิจารณาเข้าศึกษาตามผลคะแนนทำให้มีผู้สมัครจำนวนมากที่ไม่ผ่านคัดเลือกเนื่องจากมีผลคะแนนที่ต่ำเมื่อแข่งขันกับผู้สมัครที่มีคะแนนสูงกว่าในสาขาวิชาเดียวกัน งานวิจัยนี้ได้นำเสนอการใช้เทคนิคเหมืองข้อมูลด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) และการเรียนรู้เบย์อย่างง่าย (Naïve Bayes) เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองจากค่าความถูกต้อง (Accuracy) และนำแบบจำลองที่ได้ไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรีจากการวิจัยมีจำนวน 69 สาขาวิชาที่มีค่าความถูกต้องสูงสุดจากการทดลองด้วยวิธีต้นไม้ตัดสินใจ และมีจำนวน 1 สาขาวิชาที่มีค่าความถูกต้องสูงสุดด้วยวิธีการเรียนรู้เบย์อย่างง่าย

คำสำคัญ: เหมืองข้อมูล เทคนิคต้นไม้ตัดสินใจ เทคนิคการเรียนรู้เบย์อย่างง่าย

Abstract

MahaSarakham University has an application procedure in place that helps students select subjects to study in the bachelor degree program; there are a large number of applicants each semester. The main problem with the application procedure is the number of unqualified students recommended to study in the applied fields: some fields have fewer applicants than shown in the admission plan. According to the admission data, there are a lot of applicants in some fields when compared with fewer in the admission plan's proportion. Most applicants appliy for fields they are fond of regardless of their admission scores. As a result, many applicants were not selected are using their score for placement.This research has shown how to use the Data Mining via the Decision Tree and Naïve Bayes techniques in order to compare the efficiency model accuracy and develop the model as a decision support system that selects study field opportunities in the bachelor's degree program. As to the research, there were 69 fields of study with the highest accuracy from the Decision Tree technique and there was 1 field of study with the highest accuracy from the Naïve Bayes technique.

Keywords: Data Mining, Decision tree Technique, Naïve bayes Technique

[1] *นักวิชาการคอมพิวเตอร์ชำนาญการ กองบริการการศึกษา มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150*

[1] *Computer Technical Officer Professional Level, Division of Academic Affair, Mahasarakham University, Kantharawichai District, MahaSarakham44150 Thailand*

บทนำ

มหาวิทยาลัยมหาสารคาม เป็นสถาบันอุดมศึกษาของรัฐ ที่มีการจัดการเรียนการสอนในระดับปริญญาตรี และบัณฑิตศึกษา กองบริการการศึกษา ซึ่งเป็นหน่วยงานที่มีหน้าที่ดำเนินการรับสมัครร่วมไปถึงการคัดเลือกบุคคลเข้าศึกษาต่อในระดับปริญญาตรี[1] และยังมีหน้าที่ในการแนะแนวการศึกษาต่อร่วมไปถึงการประชาสัมพันธ์ข้อมูลข่าวสารเกี่ยวกับการเข้าศึกษาต่อในระดับปริญญาตรีของมหาวิทยาลัยมหาสารคาม ซึ่งการประชาสัมพันธ์และการให้ข้อมูลประกอบการตัดสินใจแก่ผู้สมัครในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในสาขาวิชาที่ผู้สมัครสนใจเข้าศึกษาถือเป็นเรื่องที่สำคัญอย่างยิ่งต่อการแนะแนวการศึกษาต่อ

จากปัญหาของการรับสมัครเพื่อคัดเลือกบุคคลเข้าศึกษาในระดับปริญญาตรี มหาวิทยาลัยมหาสารคาม ปัญหาที่สำคัญที่สุด[1] คือจำนวนผู้มีสิทธิ์เข้าศึกษาไม่เป็นไปตามแผนการรับเข้าศึกษาในบางสาขาวิชา ทำให้บางสาขาวิชามีผู้มีสิทธิ์เข้าศึกษาน้อยกว่าแผนการรับเข้าศึกษาจากการวิเคราะห์ข้อมูลการรับสมัครของกองบริการการศึกษา พบว่าในบางสาขาวิชามีการกระจุกตัวมากในจำนวนของผู้สมัคร เช่น การสมัครคัดเลือกบุคคลเข้าศึกษาในระดับปริญญาตรี ระบบรับตรง ปีการศึกษา 2559 สาขาวิชาพย.บ.พยาลาลศาสตร์ คณะพยาบาลศาสตร์ มีแผนการรับเข้าศึกษาจำนวน 40 คน และมีผู้สมัครทั้งหมดของสาขาพยาบาลศาสตร์ 3,259 คนเมื่อพิจารณาเข้าศึกษาตามผลคะแนนของผู้สมัครแล้วมีผู้สมัครจำนวนไม่น้อยที่ไม่สามารถเข้าศึกษาในสาขาวิชานั้นได้ เนื่องจากผู้สมัครมีผลคะแนนไม่ถึงเกณฑ์ที่สาขาวิชากำหนดเข้าศึกษา และผู้สมัครมีผลคะแนนที่ต่ำจึงไม่สามารถแข่งขันกับผู้สมัครที่มีผลคะแนนที่สูงกว่าได้ในสาขาวิชาของผู้สมัคร โดยผู้สมัครส่วนใหญ่ยังขาดประสบการณ์ในการสมัครซึ่งผู้สมัครส่วนใหญ่จะเลือกสมัครในสาขาวิชาตามความรู้สึก ความชอบ และสภาพแวดล้อม เช่น เพื่อนหรือผู้ปกครองเป็นหลัก โดยไม่ได้คำนึงถึงความรู้และทักษะของตนเอง เช่น ผลการเรียนเฉลี่ยสะสม (GPAX) ผลการเรียนกลุ่มสาระการเรียนรู้ (GPA) ผลการทดสอบวิชาความถนัดทั่วไป (GAT) และผลการทดสอบวิชาความถนัดทางวิชาการและวิชาชีพ (PAT)ซึ่งเป็นองค์ประกอบหลักในการเข้าศึกษาต่อในระดับปริญญาตรี

จากข้อมูลการรับสมัครคัดเลือกบุคคลเข้าศึกษาในระดับปริญญาตรี ระบบรับตรง มหาวิทยาลัยมหาสารคาม ปีการศึกษา 2557 – 2559 พบว่ามีจำนวนผู้สมัครที่ไม่มีสิทธิ์เข้าสอบสัมภาษณ์จำนวนมากในแต่ละปีการศึกษาคิดเป็นร้อยละ 44.04, 29.85 และ 25.67 ตามลำดับจากข้อมูลดังกล่าวเห็นได้ว่ามีผู้สมัครจำนวนไม่น้อยที่ไม่มีสิทธิ์เข้าสอบสัมภาษณ์เพื่อพิจารณาเข้าศึกษาในระดับปริญญาตรี มหาวิทยาลัยมหาสารคามต่อไป

ปัญหาดังกล่าวผู้วิจัยได้มีแนวคิดในการนำเทคนิคเหมืองข้อมูล (Data Mining) เพื่อวิเคราะห์หารูปแบบที่ดีที่สุดเพื่อนำรูปแบบที่ได้ไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจเลือกสาขาวิชา เพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรี มหาวิทยาลัยมหาสารคาม โดยให้ผู้สมัครได้ทำการทดสอบผลคะแนนของตนเองผ่านระบบเพื่อให้ทราบว่ามีโอกาสมากน้อยเพียงใดในการเข้าศึกษาต่อในสาขาวิชานั้นๆ ก่อนที่จะทำการสมัครจริงในสาขาวิชานั้น

วัตถุประสงค์

เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรี ระหว่างอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) และพัฒนาระบบระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรีจากรูปแบบหรือแบบจำลองที่มีประสิทธิภาพสูงสุด

ทฤษฎีที่เกี่ยวข้อง

งานวิจัยนี้ผู้วิจัยได้นำเอาข้อมูลการรับเข้าศึกษาในระดับปริญญาตรี ระบบรับตรง มหาวิทยาลัยมหาสารคาม ระหว่างปีการศึกษา 2557 – 2559 มาทำการทดลองด้วยการใช้เทคนิคเหมืองข้อมูลจากการสร้างแบบจำลองในการทำนาย (Predictive Modeling) และการเรียนรู้แบบมีตัวแบบ (Supervised Modeling) โดยการจำแนกข้อมูล (Classification)[2,3,4] อัลกอริทึมต้นไม้ตัดสินใจ (Decision TreeC4.5) และวิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) มาเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมซึ่งกระบวนการแต่ละวิธีมีรายละเอียดดังต่อไปนี้

1. เหมืองข้อมูล

เหมืองข้อมูล (Data Mining)[5] คือการนำเทคนิคการเรียนรู้ของเครื่อง (Machine Learning)หรือวิธีการทางสถิติ (Statistical Methods)[6]มาทำการวิเคราะห์ข้อมูลที่จัดเก็บไว้ในฐานข้อมูลหรือจัดเก็บไว้ในรูปแบบอื่น จุดประสงค์ของการทำเหมืองข้อมูล คือการวิเคราะห์แนวโน้มความสัมพันธ์ หรือรูปแบบของข้อมูลซึ่งเป็นความรู้ที่ถูกซ่อนอยู่ภายในข้อมูลขนาดใหญ่ สามารถนำสารสนเทศที่ได้มาใช้ในการวางแผน การตัดสินใจ หรือการแก้ปัญหาด้านต่างๆ

2. ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree)[7,8]เป็นเทคนิคที่

มีรูปแบบที่นิยม โครงสร้างต้นไม้ตัดสินใจเป็นแบบลำดับชั้น โดยมีการตัดสินใจซึ่งประกอบด้วยโหนดที่ใช้ในการตัดสินใจ (Decision node) และโหนดใบ (Leaf node or Terminal node) แต่ละโหนดตัดสินใจนั้นจะมีการสร้างฟังก์ชันที่เอาไว้สำหรับทดสอบทางเลือกจากการป้อนข้อมูลเข้าจะทดสอบตามทางเลือกไปเรื่อยๆ ไปจนถึง Terminal node จะได้คำตอบดัง Figure 1

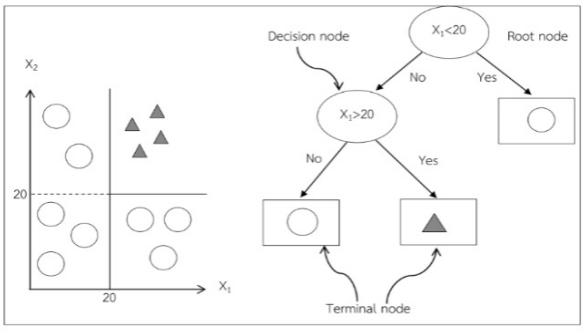

Figure 1 Example of tree map built from learing platform[7]

3. อัลกอริทึม C4.5

เป็นอัลกอริทึมที่ถูกพัฒนามาจาก ID3[5,7,8]โดย Ross Quinlan เป็นอัลกอริทึมที่ใช้หลักการสร้างต้นไม้โดยเลือกแอทริบิวต์ (Attribute) ที่สำคัญที่สุดมาเป็นโหนดราก (Root Node) โดยใช้ Gain Ratio ที่สูงที่สุดเป็นโหนดรากและโหนดถัดไปในการหาค่า Gain Ratio ต้องทำการหาค่า Split Information และค่า Entropy ก่อนวิธีหาค่าสมการดังต่อไปนี้

3.1 สมการ *Entropy* การหาค่าสมการที่ใช้ในการหาค่าสารสนเทศของข้อมูลดังสมการที่1

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

โดย S คือ แอทริบิวต์ที่นำมาวัดค่า

P_i คือ สัดส่วนของจำนวนสมาชิกในกลุ่ม i เท่ากับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

3.2 สมการ *Information Gain* เป็นสมการที่ใช้ในการหาค่าสารสนเทศก่อนนำไปใช้ในการหาค่ามาตรฐานอัตราส่วนดังสมการที่ 2

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

โดย A คือ แอทริบิวต์ A

$|S_v|$ คือ สมาชิกของแอทริบิวต์ A ที่มีค่า v $|S|$ คือ จำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

3.3 สมการ *Split Information* เป็นสมการที่ใช้ในเพิ่มค่าสารสนเทศการแบ่งแยกข้อมูลจะบอกถึงลักษณะการกระจายของข้อมูล เป็นการแก้ปัญหาความโน้มเอียงดังสมการที่ 3

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

โดย S_i คือ สัดส่วนของจำนวนสมาชิกในกลุ่ม i

3.4 สมการ *Gain Ratio* เป็นสมการในการคำนวณหาค่ามาตรฐานอัตราส่วนค่าเกน (*Gain*) เพื่อลดความลำเอียงของข้อมูลดังสมการที่ 4

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

4. อัลกอริทึมการเรียนรู้เบย์อย่างง่าย

อัลกอริทึมการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning)[5,6] เป็นวิธีการจำแนกประเภทข้อมูลที่มีการเรียนรูปแบบมีผู้สอนโดยใช้ความน่าจะเป็นเพื่อคำนวณการแจกแจงความน่าจะเป็นตามสมมติฐานให้กับข้อมูลจากการคำนวณตัวอย่างใหม่ที่ได้จะถูกนำมาปรับเปลี่ยนการแจกแจงซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นของข้อมูล ข้อมูลใหม่ที่เกิดขึ้นจะถูกปรับเปลี่ยนไปตามข้อมูลใหม่โดยผนวกกับข้อมูลเดิมที่มี

หลักการของการเรียนรู้เบย์อย่างง่ายใช้ในการคำนวณหาค่าความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล เป็นวิธีการในการจำแนกที่สามารถคาดการณ์ผลลัพธ์ได้ ซึ่งจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์เป็นวิธีการที่ไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมาก มีคุณสมบัติ (Attribute) ไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม V_i สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = \{a_1, a_2, ..., a_n\}$ หรือ $P(a_1, a_2, ..., a_n \mid v_j)$ ดังสมการที่ 5

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

โดย $P(v_j)$ คือ ความน่าจะเป็นของข้อมูลที่ให้คลาส

$P(a_i \mid v_j)$ คือ ค่าความน่าจะเป็นของข้อมูลคุณลักษณะที่ i มีค่า a_i และให้คลาส v_j

$P(a_1, a_2, a_3, ..., a_n)$ คือ ค่าความน่าจะเป็นของข้อมูลทดสอบที่มีคุณลักษณะ $a_1, a_2, a_3, ..., a_n$ ที่จะให้คลาส v_j กลุ่ม v_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว

$\prod$ คือ ผลคูณของค่า $P(a_i \mid v_j)$ ทั้งหมด $i = 1, 2, 3, ..., n$ และ $j = 1, 2, 3, ..., n$

วิธีการนำเทคนิคการเรียนรู้เบย์อย่างง่ายไปใช้โดยหาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2, a_3, ..., a_n \mid v_j)$ จากสมการที่ 5 มาคูณกับค่าความน่าจะเป็นของของกลุ่มนั้นๆคือ $P(v_j)$ ได้เท่ากับ V_{NB} และนำค่าที่ได้มาเปรียบ กลุ่มที่มีค่าความน่าจะเป็นสูง คือคำตอบที่ได้ ดังนั้นจะได้วิธีการจำแนกประเภทการเรียนรู้เบย์อย่างง่ายดังสมการที่ 6

$$V_{NB} = \arg\max P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j) \quad (6)$$

จากสมการที่ 6 สามารถเขียนเป็นอัลกอริทึมการเรียนรู้แบบเบย์อย่างง่ายดัง Figure 2

Naïve Bayes Learn (examples)

FOR EACH target value v DO

$\overline{P}(v_j) \leftarrow estimate\, P(v_j)$

FOR EACH attribute value a of each attribute DO

$\overline{P}(a_j \mid v_j) \leftarrow estimate\, P(a_j \mid v_j)$

Classify NEW Example(x)

$$V_{NB} = \arg\max P(v_j) * \prod_{i=1}^{n} P(a_i \mid v_j)$$

Figure 2 Naive Bayes learning algorithm[6]

วัสดุอุปกรณ์และวิธีการศึกษา

การวิจัยครั้งนี้ได้นำเสนอการใช้เทคนิคเหมืองข้อมูลมาวิเคราะห์การเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อระดับปริญญาตรี จากการนำข้อมูลการรับสมัครคัดเลือกบุคคลเข้าศึกษาต่อในระดับปริญญาตรี ระบบรับตรง ระหว่างปีการศึกษา 2557 – 2559 จำนวน 74,602 ชุดข้อมูลการทดลองโดยทำการเตรียมข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปใช้กับโปรแกรม WEKA (Waikato Environment for Knowledge Analysis)[9,10] ซึ่งประกอบด้วยการแบ่งข้อมูลออกเป็น 2 กลุ่ม ได้แก่ข้อมูลการทดสอบ (Data Testing) และข้อมูลการเรียนรู้ (Data Training) โดยใช้หลักการเลือกแบบความเที่ยงตรง 10 กลุ่ม (10-fold Cross Validation) จากนั้นทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพแบบของแบบจำลองในแต่ละอัลกอริทึม คือ อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree C4.5) และอัลกอริทึมการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) จากนั้นทำการเลือกแบบจำลองที่มีประสิทธิภาพสูงสุดจากค่าความถูกต้อง (Accuracy) มาพัฒนาเป็นระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรีต่อไป

จากการวิจัยครั้งนี้ผู้วิจัยได้แสดงขั้นตอน และกรอบแนวคิดการวิจัยเพื่อให้ทราบกระบวนการในการวิจัยในครั้งนี้ โดยมีรายละเอียดดัง Figure 3

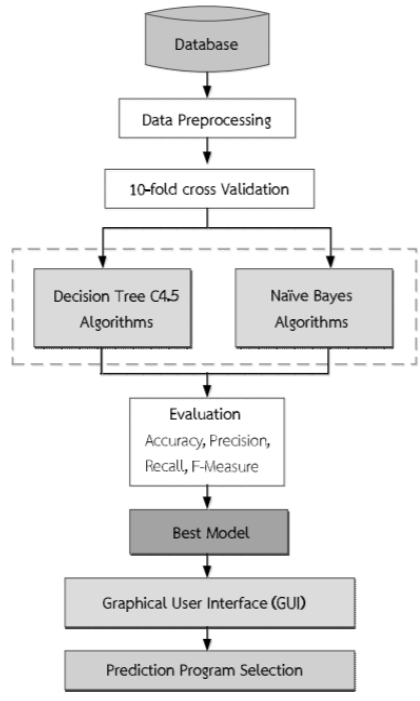

Figure 3 Conceptual framework

งานวิจัยนี้ได้มีวิธีการวัดประสิทธิภาพของแบบจำลองในแต่ละข้อมูลตามรายสาขาวิชา โดยได้ใช้ค่าความถูกต้องของแบบจำลอง (Accuracy) ค่าความแม่นยำของแบบจำลอง (Precision) ค่าความระลึกของแบบจำลอง (Recall) และค่าความเหวี่ยงของแบบจำลอง (F-Measure) การวัดประสิทธิภาพของการจำแนกข้อมูลตามแนวคิดทางด้านค้นคืนสารสนเทศดังสมการที่ 7, 8, 9 และ 10 ตามลำดับ[7,8]

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

$$\Pr ecision = \frac{TP}{TP + FP} \quad (8)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (9)$$

$$F - Measure = \frac{2 * \Pr ecision * \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \quad (10)$$

โดย TP คือ จำนวนข้อมูลที่ถูกดึงออกมาอย่างถูกต้อง

FP คือ จำนวนข้อมูลที่ผิดพลาดที่ถูกดึงออกมา

TN คือ จำนวนข้อมูลที่ถูกต้องแต่ไม่ถูกดึงออกมา

FN คือ จำนวนข้อมูลที่ผิดพลาดแต่ไม่ถูกดึงออกมา

ข้อมูลที่ใช้ทำการทดลองเป็นข้อมูลที่ใช้เป็นองค์ประกอบหลักที่ใช้ในการพิจารณาเข้าศึกษา ได้แก่ข้อมูลผลการเรียนเฉลี่ยสะสม (GPAX) ข้อมูลผลการเรียนกลุ่มสาระการเรียนรู้ (GPA) และข้อมูลคะแนนมาตรฐานการทดสอบความถนัดทั่วไป (GAT) และความถนัดทางวิชาการและวิชาชีพ (PAT) มีรายละเอียดดังนี้

GPAX คือ ผลการเรียนเฉลี่ยสะสม
GPA1 คือ ผลการเรียนกลุ่มสาระวิชาภาษาไทย
GPA2 คือ ผลการเรียนกลุ่มสาระวิชาคณิตศาสตร์
GPA3 คือ ผลการเรียนกลุ่มสาระวิชาวิทยาศาสตร์
GPA4 คือ ผลการเรียนกลุ่มสาระวิชาสังคมศึกษา
GPA5 คือ ผลการเรียนกลุ่มสาระวิชาสุขศึกษา
GPA6 คือ ผลการเรียนกลุ่มสาระวิชาศิลปะ
GPA7 คือ ผลการเรียนกลุ่มสาระวิชาการงานอาชีพ
GPA8 คือ ผลการเรียนกลุ่มสาระวิชาภาษาต่างประเทศ
GAT คือ คะแนนมาตรฐานความถนัดทั่วไป
PAT1 คือ คะแนนมาตรฐานทางคณิตศาสตร์
PAT2 คือ คะแนนมาตรฐานทางวิทยาศาสตร์
PAT3 คือ คะแนนมาตรฐานทางวิศวกรรมศาสตร์
PAT4 คือ คะแนนมาตรฐานทางสถาปัตยกรรมศาสตร์
PAT5 คือ คะแนนมาตรฐานทางวิชาชีพครู
PAT6 คือ คะแนนมาตรฐานทางศิลปะกรรมศาสตร์
Score คือ ผลคะแนนรวม

Table 1 Data sample of B.Eng. (Engineering)

GPAX	GPA1	GPA2	GPA3	GPA4	GPA8	GAT	PAT1	PAT3	Score
3.87	3.80	4.00	3.88	3.86	3.68	65.75	78.87	66.06	89.64
3.90	3.68	3.98	3.97	3.73	3.78	60.11	83.40	81.79	81.99
3.75	3.90	2.92	3.61	3.69	4.00	66.54	54.99	67.60	77.52
2.41	3.20	1.62	2.53	2.08	2.15	37.55	44.93	42.45	44.88
3.69	3.90	3.72	3.49	3.92	3.70	56.12	51.29	52.20	67.60
3.21	3.50	3.14	3.01	3.35	3.33	53.04	54.39	42.40	61.84
2.18	2.50	2.03	1.58	2.76	2.06	37.35	49.17	34.68	43.49

ในการทดลองโดยจะแยกข้อมูลที่ใช้เป็นรายสาขาวิชา ตัวอย่างข้อมูลที่ใช้ในการทดลองดังรายละเอียดตัวอย่างสาขาวิชาวิศวกรรมศาสตร์ (Table 1) การแทนค่าข้อมูลในการทดลองเพื่อให้การวิเคราะห์ข้อมูลในการทดลองนั้นถูกต้อง และมีความแม่นยำในการทดลองข้อมูลที่จะใช้ในการทดลองต้องเป็นข้อมูลที่อยู่ในรูปแบบที่คอมพิวเตอร์เข้าใจถึงความหมายของข้อมูลนั้นเสียก่อน การวิจัยนี้ได้มีการแทนค่าให้กับข้อมูลโดยข้อมูลที่ใช้ในการทดลองจะเป็นข้อมูลแบบทศนิยมแบบต่อเนื่อง (Binning data) โดยมีการแบ่งข้อมูลส่วนที่เป็นตัวแปรคุณลักษณะ (Attribute) 2 ส่วน ได้แก่ส่วนที่ 1 คือข้อมูล GPAX, GPA1, GPA2, GPA3, GPA4, GPA5, GPA6, GPA7 และ GPA8 มีการแทนค่าข้อมูลผลการเรียนออกเป็น 4 ระดับ คือช่วงผลการเรียนระหว่าง 0.00 – 2.50 อยู่ในระดับต่ำแทนค่าเป็น (L), ช่วงผลการเรียนระหว่าง 2.51 – 3.00 อยู่ในระดับกลาง แทนค่าเป็น (M), ช่วงผลการเรียนระหว่าง 3.01 – 3.50 อยู่ในระดับสูงแทนค่าเป็น (H) และช่วงผลการเรียนระหว่าง 3.51 – 4.00 อยู่ในระดับสูงมาก แทนค่าเป็น (VH) ส่วนที่ 2 คือข้อมูล GAT, PAT1, PAT2, PAT3, PAT4, PAT5 และ PAT6 มีการแทนค่าข้อมูลผลคะแนนมาตรฐานออกเป็น 4 ระดับคือช่วงผลคะแนนมาตรฐานระหว่าง 0.00 – 39.08 อยู่ในระดับต่ำ แทนค่าเป็น (L), ช่วงผลคะแนนมาตรฐานระหว่าง 39.09 – 51.48 อยู่ในระดับกลาง แทนค่าเป็น (M), ช่วงผลคะแนนมาตรฐานระหว่าง 51.49 – 63.88 อยู่ในระดับสูง แทนค่าเป็น (H) และช่วงผลคะแนนมาตรฐานระหว่าง 63.89 – 100 อยู่ในระดับสูงมาก แทนค่าเป็น (VH)และส่วนข้อมูลที่เป็นคำตอบ (Class) มีการแทนค่าช่วงผลคะแนนรวม (Score) ออกเป็น 4 ระดับ คือช่วงผลคะแนนรวมระหว่าง 0.00 – 46.77 โอกาสระดับน้อย แทนค่าเป็น (Low), ช่วงผลคะแนนรวมระหว่าง 46.78 – 62.93 โอกาสระดับปานกลาง แทนค่าเป็น (Moderate), ช่วงผลคะแนนรวมระหว่าง 62.94 – 79.09 โอกาสระดับมาก แทนค่า

เป็น (High) และช่วงผลคะแนนรวมระหว่าง 79.10 – 100 โอกาสระดับมากสุด แทนค่าเป็น (Most) ตัวอย่างการแทนค่าข้อมูลในการทดลองในสาขาวิชาวิศวกรรมศาสตร์ B.Eng. (Engineering) ดัง Table 2

Table 2 Variable sample of B.Eng. (Engineering)

GPAX	GPA1	GPA2	GPA3	GPA4	GPA8	GAT	PAT1	PAT3	Score
VH	VH	VH	VH	VH	VH	VH	VH	VH	Most
VH	VH	VH	VH	VH	VH	H	VH	VH	Most
VH	VH	M	VH	VH	VH	VH	H	VH	High
L	H	L	M	L	L	L	M	M	Low
VH	VH	VH	H	VH	VH	H	M	H	High
H	H	H	H	H	H	H	H	M	Moderate
L	L	L	L	M	L	L	M	L	Low

ผลการศึกษาวิจัย

จากการทดลองข้อมูลในแต่ละสาขาวิชาทั้งหมด 70 สาขาวิชา โดยใช้วิธีต้นไม้ตัดสินใจ (Decision tree C4.5) และวิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayes) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในแต่ละวิธีการ โดยใช้โปรแกรม WEKA ในการทดลองวิเคราะห์ข้อมูล โดยการหาค่าความถูกต้อง (Accuracy)จากผลการศึกษาตามกลุ่มประกันคุณภาพการศึกษาของมหาวิทยาลัย รายละเอียดผลการทดลองโดยแยกตามกลุ่มดัง Figure 4 – 6

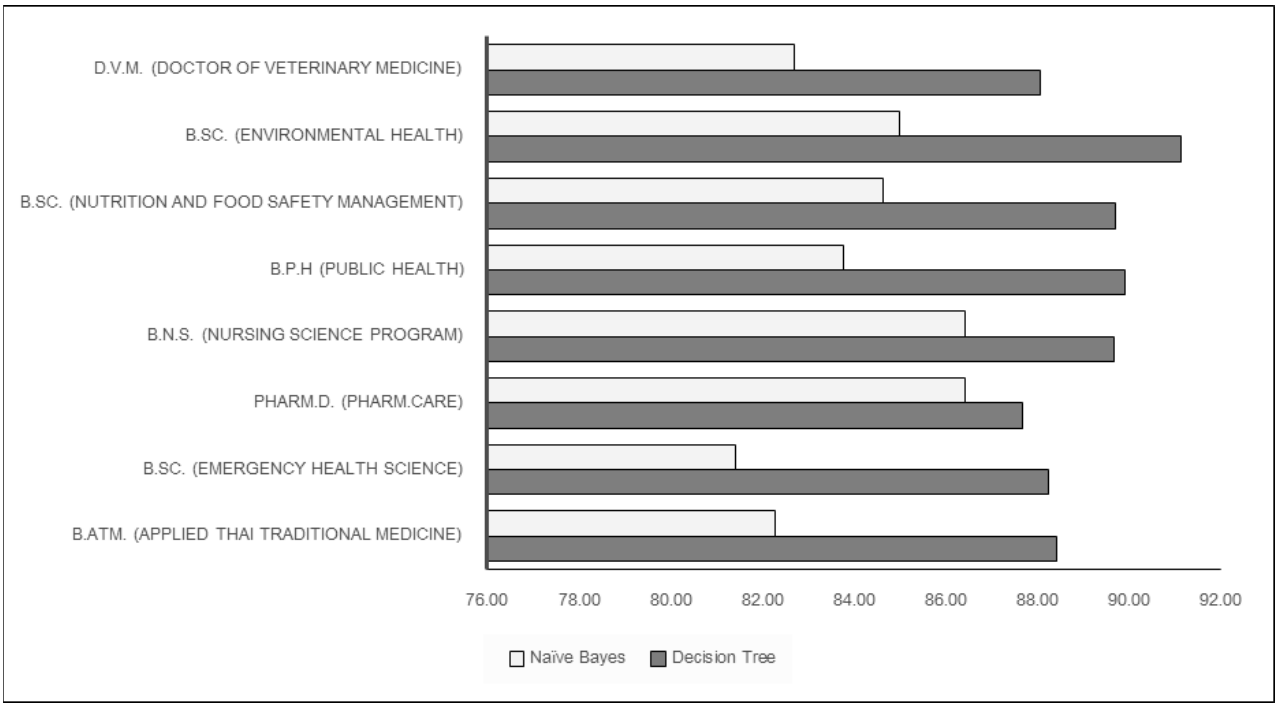

Figure 4 Results of Group Health Science

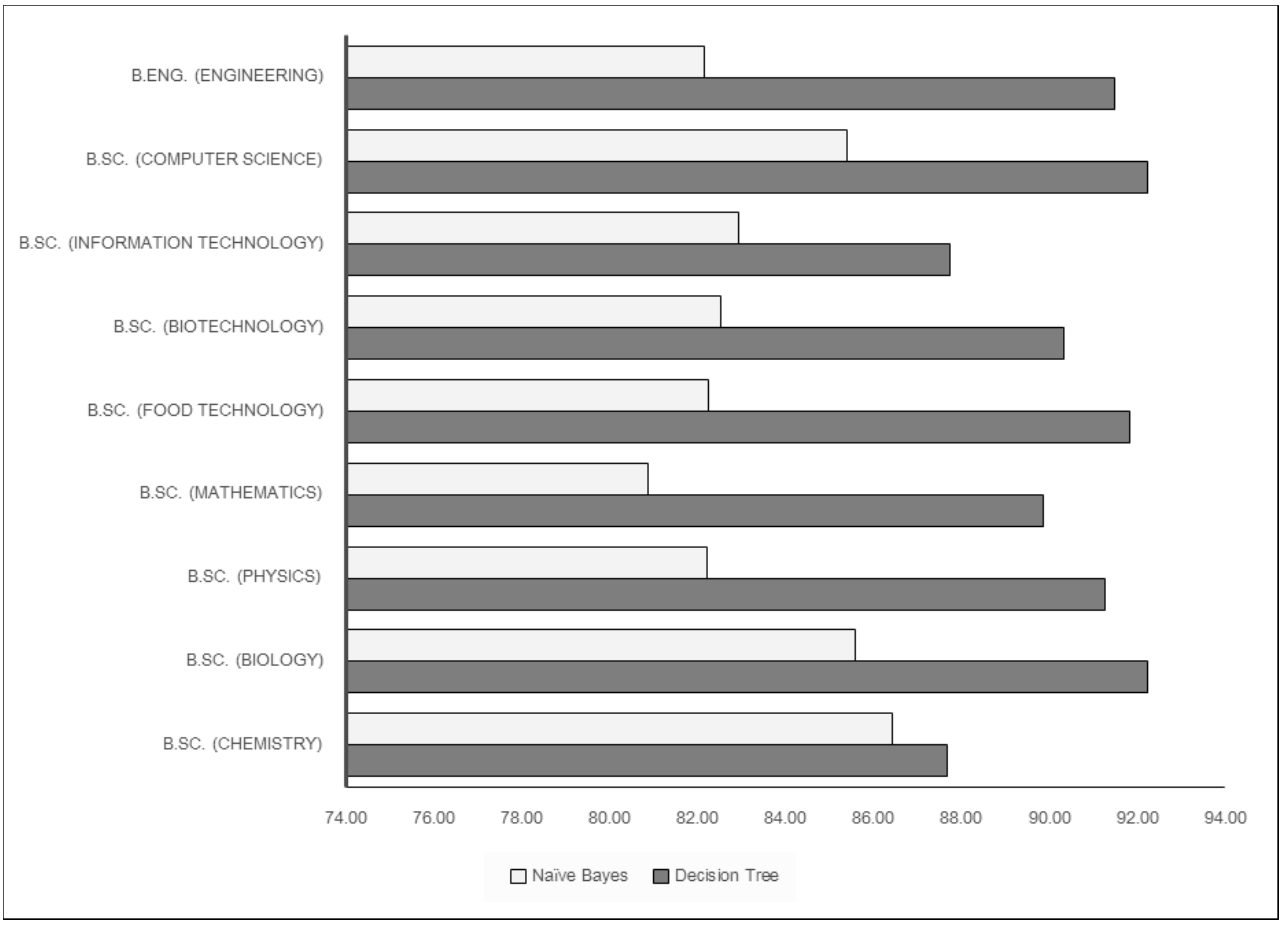

Figure 5 Results of Group Science and Technology

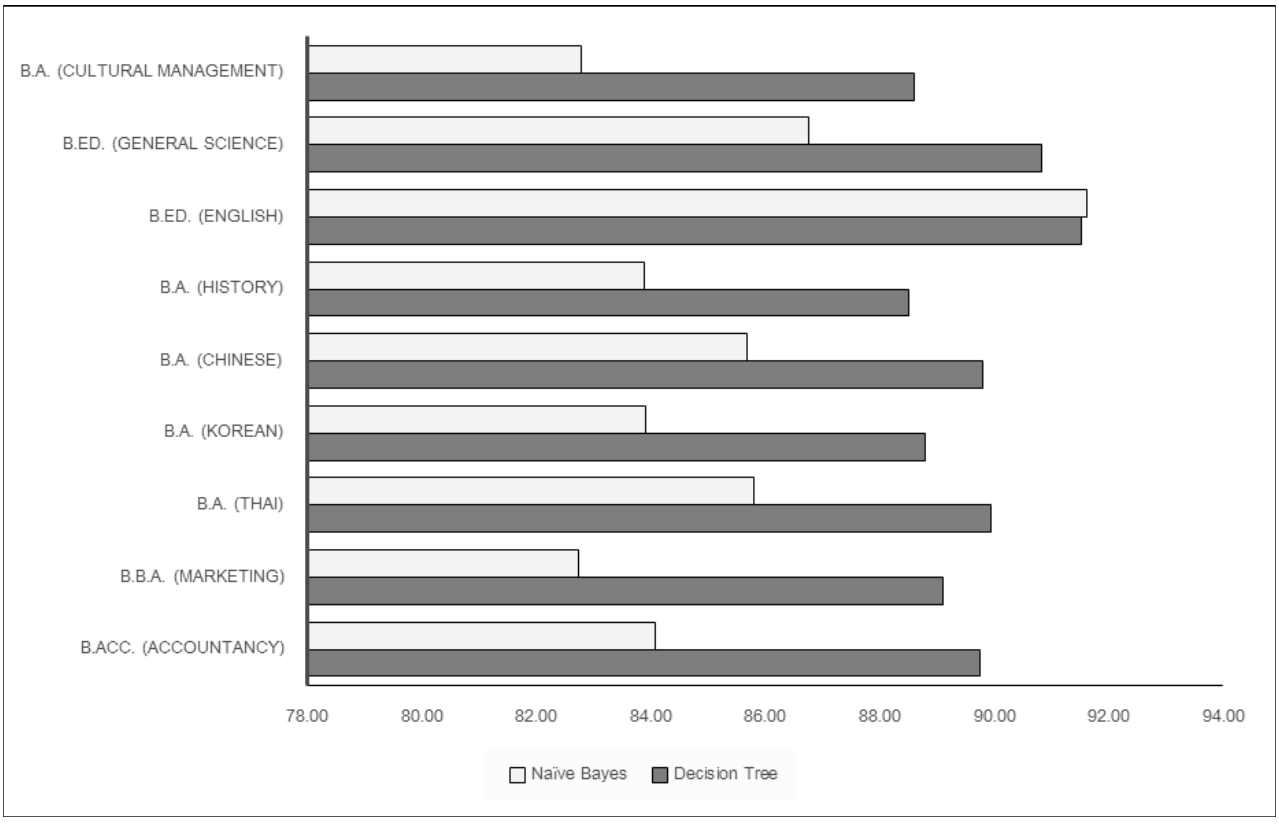

Figure 6 Results of Group Humanities and Social Sciences

จากผลการทดลองด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) และวิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayes) จากข้อมูลการทดลองในแต่ละสาขาวิชาจำนวนทั้งหมด 70 สาขาวิชาสามารถอธิบายการทดลองได้ว่าการทดลองด้วยวิธีต้นไม้ตัดสินใจ (Decision Tree) มีค่าความถูกต้อง (Accuracy) สูงสุดใน 69 สาขาวิชา และมีเพียงสาขาวิชาเดียวที่การทดลองด้วยวิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayes) ที่มีค่าความถูกต้องสูงสุด คือสาขาวิชา กศ.บ. ภาษาอังกฤษ คณะศึกษาศาสตร์ (B.Ed. English)

ผลการทดลองสามารถสร้างเป็นกฎการตัดสินใจ (Decision Rule) ในแต่ละสาขาวิชาเพื่อนำรูปแบบที่ได้ไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรีจากตัวอย่างสาขาวิชา ศศ.บ. ภาษาจีน รายละเอียดกฎการตัดสินใจดังนี้

IF GAT85=*VH* Then Score=*High*
IF GAT85=*M* Then Score=*Moderate*
IF GAT85=*L*And GPAX=*L* Then Score=*Low*
IF GAT85=*L* And GPAX=(*VH* Or *H*)
Then Score=*Moderate*
IF GAT85=*L*And GPAX=*M*And GPA1=*L*
Then Score=*Low*
IF GAT85=*L* And GPAX=*M* And GPA1=(*VH* Or *H*)
Then Score=*Moderate*
IF GAT85=*L*And GPAX=*M*And GPA1=*M* And
GPA8=(*VH* Or *H*) Then Score=*Moderate*
IF GAT85=*L*And GPAX=*M*And GPA1=*M* And
GPA8=(*M* Or *L*) Then Score=*Low*
IF GAT85=*H*And GPAX=*VH* Then Score=*High*
IF GAT85=*H* And GPAX=(*M* Or *L*)
Then Score=*Moderate*
IF GAT85=*H* And GPAX=*H* And GPA1=(*VH* Or *H*)
Then Score=*High*
IF GAT85=*H*And GPAX=*H*And GPA1=*L*
Then Score=*Moderate*
IF GAT85=*H*And GPAX=*H*And GPA1=*M* And
GPA8=*VH* Then Score=*High*
IF GAT85=*H*And GPAX=*H*And GPA1=*M* And
GPA8=(*M* Or *L*) Then Score=*Moderate*
IF GAT85=*H*And GPAX=*H*And GPA1=*M* And
GPA8=*H*And GPA4=*VH* Then Score=*High*
IF GAT85=*H*And GPAX=*H*And GPA1=*M* And
GPA8=*H* And GPA4=(*H* Or *M* Or *L*)
Then Score=*Moderate*

จากรูปแบบที่ได้ในแต่ละสาขาวิชาสามารถนำมาพัฒนาเป็นระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรีโดยมีรายละเอียดดัง Figure 7

Figure 7 Home Program

ตัวอย่างการทดสอบข้อมูลผ่านระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรี รายละเอียดดัง Figure 8

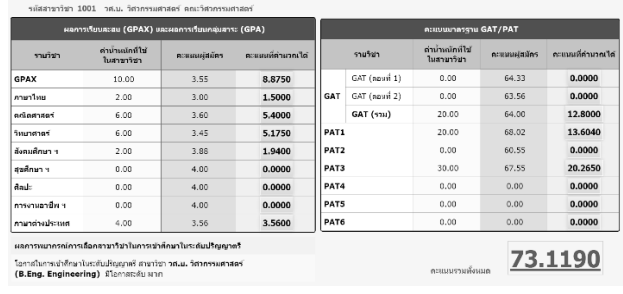

Figure 8 Test Results

อภิปรายผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรี ระหว่างอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึมการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) และพัฒนาเป็นระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรีจากรูปแบบ หรือแบบจำลองที่มีประสิทธิภาพสูงสุด

จากผลการวิจัยพบว่าอัลกอริทึมต้นไม้ตัดสินใจมีค่าความถูกต้อง (Accuracy) สูงสุด และนำแบบจำลองที่ได้ในแต่ละสาขาวิชามาพัฒนาเป็นระบบสนับสนุนการตัดสินใจในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อในระดับปริญญาตรีได้

ข้อเสนอแนะ

1. ข้อมูลที่ใช้ในการทดลองในแต่ละสาขาวิชาต้องมีจำนวนมากพอที่ใช้สำหรับการทดลอง

2. ข้อมูลในการทดลองการวิจัยนี้ เป็นข้อมูลแบบทศนิยมแบบต่อเนื่อง (Binning data)การแทนค่าข้อมูลควรแทนค่าช่วงระหว่างข้อมูลให้มีความเหมาะสม

3. ข้อมูลที่ใช้ในการทดลองคือข้อมูลการคัดเลือกบุคคลเข้าศึกษาในระดับปริญญาตรี ระหว่างปีการศึกษา 2557 – 2559 หากมีจำนวนข้อมูลที่เพิ่มมากขึ้นผลการทดลองอาจจะมีการปรับเปลี่ยน

กิตติกรรมประกาศ

โครงการวิจัยนี้ได้รับการสนับสนุนจากเงินอุดหนุนการวิจัยจากงบประมาณรายได้ ประจำปีงบประมาณ 2560 มหาวิทยาลัยมหาสารคาม

เอกสารอ้างอิง

1. กองบริการการศึกษา มหาวิทยาลัยมหาสารคามระเบียบการการสมัครคัดเลือกบุคคลเข้าศึกษาในระดับปริญญาตรีระบบรับตรงประจำปีการศึกษา 2557 - 2559.
2. บุญมา เพ่งช่วน. การใช้เทคนิคเหมืองข้อมูลเพื่อพัฒนาระบบสนับสนุนการตัดสินใจด้านการผลิตบัณฑิตระดับปริญญาตรี. ปริญญานิพนธ์ ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์. กรุงเทพฯ : บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร; 2548.
3. Han J, Kamber M, Data Mining Concepts and Techniques; The Morgan Kaufmann Publishers,2001.
4. Olson D, Shi Y. Introduction to Business Data Mining; McGraw Hill International Edition, 2007.
5. ฉัตรเกล้า เจริญผล. เอกสารประกอบการสอนรายวิชา Introduction to Data Mining 2013.
6. บุญเสริม กิจศิริกุล. อัลกอริธึมการทำเหมืองข้อมูล. รายงานวิจัยฉบับสมบูรณ์โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย. 2546
7. อนันต์ ปินะเต. การพัฒนาระบบสนับสนุนการตัดสินใจในการเลือกสมัครในสาขาวิชาโดยใช้เทคนิคต้นไม้ตัดสินใจ; วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, ปีที่ 35, ฉบับที่ 4, ประจำเดือน : กรกฎาคม – สิงหาคม 2559.
8. อนันต์ ปินะเต, ฉัตรเกล้า เจริญผล, แกมกาญจน์ สมประเสริฐศรี. การใช้เทคนิคเหมืองข้อมูลในการเลือกกลุ่มสาขาวิชาที่เหมาะสมสำหรับการศึกษาต่อระดับปริญญาตรี; วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, ปีที่ 33, ฉบับที่ 6, ประจำเดือน : พฤศจิกายน – ธันวาคม 2557.
9. Remco, R. Bouckert,et al WEKA Manual for Version 3-6-4. (Online). Available : http://cdnetworks~kr~1.dl.sourceforge.net/project/wekadocumentation/3.6.x/WekaManual-3-6-4.pdf
10. Zdravko M, Ingrid R, An Introduction to the WEKA Data Mining System.