การเปรียบเทียบประสิทธิภาพของแบบจำลองในการพยากรณ์ความสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพ

The Comparison of Performance Models for Predicting Students Success in Vocation Education

พัฒนพงษ์ ดลรัตน์[1*], จารี ทองคำ[2]

Pattanaphong Donrat[1*], JareeThongkam[2]

Received: 27 March 2017 ; Accepted: 23 November 2017

บทคัดย่อ

ปัจจุบันประเทศไทยนั้นอยู่สภาวะขาดแคลนแรงงานฝีมือ กระทรวงศึกษาธิการจึงได้มีนโยบายในปีพุทธศักราช 2554 การเปลี่ยนแปลงสัดส่วนของผู้เรียนอาชีวศึกษาต่อผู้เรียนสายสามัญจากเดิม 40:60 เป็น 60:40 ภายในปีพุทธศักราช 2561 ซึ่งเป็นปีสิ้นสุดของการปฏิรูปการศึกษาทศวรรษที่ 2 ซึ่งสถานศึกษาอาชีวศึกษาเอกชนก็เป็นหน่วยงานหนึ่งที่ดำเนินการสนองนโยบายของรัฐบาลเพื่อแก้ไขปัญหาสภาวะขาดแคลนแรงงานฝีมือ เพิ่มจำนวนผู้สำเร็จการศึกษา พัฒนาการศึกษาให้เกิดประสิทธิภาพสูงสุด ดังนั้นงานวิจัยนี้จึงมีจุดประสงค์เพื่อการเปรียบเทียบประสิทธิภาพของแบบจำลองในการพยากรณ์ความสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพ งานวิจัยนี้ได้ใช้ 6 เทคนิคที่มีประสิทธิภาพในการสร้างแบบจำลอง คือ C4.5, Random Forest, Random Tree, Reduced Error Pruning (REP Tree), k-Nearest Neighbors (k-NN) และ Support Vector Machine (SVM) และวัดประสิทธิภาพการพยากรณ์ของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) จากการศึกษาพบว่า แบบจำลอง C4.5 มีประสิทธิภาพในการพยากรณ์ความสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพมากที่สุดถึง 95.36%

คำสำคัญ แบบจำลองการพยากรณ์ความสำเร็จการศึกษา ประกาศนียบัตรวิชาชีพ ต้นไม้การตัดสินใจ เหมืองข้อมูล

Abstract

Currently, Thailand has a shortage of skilled labor. Ministry of Education policy requires that between B.E 2554 and 2561, the proportion of vocational students compare with the common line should increase from 40:60 to 60:40, the year of the second decade of education reform education. The private Vocational College was working to meet the government's policy to solve the shortage of skilled labor, increase the number of graduates, and develop the best performance. This research compares performance of models for predicting student success in vocation education. This research using six powerful techniques in modeling is C4.5, Random Forest, Random Tree, Reduced Error Pruning (REP Tree), k-Nearest Neighbors (k-NN) and Support Vector Machine (SVM) A set of test data performance measurement and prediction of models with accuracy, precision and recall found that the C4.5 model was effective in predicting educational success. Most vocational certificate levels reached 95.36%

Keywords: Models predicting graduation, Vocational Education Certificate, Decision tree, Data mining

---

[1] *นิสิต, สาขาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม, มหาสารคาม, 44150.*

[2] *ผู้ช่วยศาสตราจารย์, อาจารย์ที่ปรึกษา หน่วยวิจัยสารสนเทศประยุกต์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม, มหาสารคาม, 44150.*

[1] *Student, Department of Information Technology, Faculty of Informatics, Mahasarakham University, Maha Sarakham, 44150.*

[2] *Assistant Professor, Applied informatics Research Unit, Faculty of Informatics, Mahasarakham University, Maha Sarakham, 44150.*

** Corresponding author: Tel: +66 086 5803169*

Email address: s2s9890@gmail.com

บทนำ

ปัจจุบันประเทศไทยนั้นอยู่สภาวะขาดแคลนแรงงานฝีมือ[1] กระทรวงศึกษาธิการจึงได้มีนโยบายในปีพุทธศักราช 2554 การเปลี่ยนแปลงสัดส่วนของผู้เรียนอาชีวศึกษาต่อผู้เรียนสายสามัญจากเดิม 40:60 เป็น 60:40 ภายในปีพุทธศักราช 2561 ซึ่งเป็นปีสิ้นสุดของการปฏิรูปการศึกษาทศวรรษที่ 2

การเพิ่มจำนวนผู้สำเร็จการศึกษาตามระยะเวลาของหลักสูตรของอาชีวศึกษา ซึ่งหลักสูตรอาชีวศึกษาแบ่งออกได้เป็น 2 หลักสูตรคือ 1)หลักสูตรประกาศนียบัตรวิชาชีพ (ปวช.) มีระยะเวลาการศึกษาตามหลักสูตร 3 ปี 2) หลักสูตรประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) มีระยะเวลาการศึกษาตามหลักสูตร 2 ปี ซึ่งการจัดให้ผู้เรียนสำเร็จการศึกษาตามระยะเวลาของหลักสูตรนั้น เป็นตัวบ่งชี้ตัวหนึ่งที่แสดงว่าสถานศึกษาได้จัดการศึกษาอย่างมีคุณภาพและประสิทธิภาพ ตามตัวบ่งชี้ของการประกันคุณภาพการศึกษาทั้งการประกันคุณภาพภายในและการประกันคุณภาพภายนอก

สถานศึกษาอาชีวศึกษาเอกชน[2] ก็เป็นหน่วยงานหนึ่งในการจัดการศึกษาอาชีวศึกษาที่มีบทบาท สำคัญในการให้บริการการศึกษาทางด้านวิชาชีพเช่นเดียวกับสถานศึกษาอาชีวศึกษารัฐบาลซึ่งล้วนแต่มีบทบาทและความสำคัญอย่างยิ่งต่อการพัฒนาประเทศ เพราะเป็นการศึกษาที่จัดเตรียมบุคคลให้มีอาชีพเป็นหลักในอนาคต และช่วยให้อาชีพที่มีอยู่แล้วมีความก้าวหน้าในอาชีพตน

แต่ในความเป็นจริงนั้นสถานศึกษาอาชีวศึกษาเอกชนนั้นมีข้อเสียเปรียบสถานศึกษาอาชีวศึกษารัฐบาลเช่น งบประมาณในการจัดซื้อสื่อการเรียนการสอนที่ชัดเจนคือประเภทวิชาอุตสาหกรรมเป็นต้น เนื่องจากงบประมาณส่วนใหญ่ของสถานศึกษาอาชีวศึกษาเอกชนนั้นมาจากเงินอุดหนุนรายบุคคลที่รัฐสนับสนุนโดยเงินที่ได้จะมากหรือน้อยนั้นก็ขึ้นอยู่จำนวนนักเรียนเป็นหลักดังนั้นหากสถานศึกษาอาชีวศึกษาเอกชนมีจำนวนนักเรียนมากก็จะได้งบประมาณมาพัฒนาการเรียนการสอนให้ได้ประสิทธิภาพมากยิ่งขึ้น ซึ่งสถานการณ์ปัจจุบันของสถานศึกษาอาชีวศึกษาในจังหวัดกาฬสินธุ์ซึ่งมีการรวมเอาสถานศึกษาอาชีวศึกษาทั้งภาครัฐและเอกชนมาไว้ภายใต้หน่วยงานเดียวกันคืออาชีวศึกษาจังหวัดกาฬสินธุ์โดยมีสถานศึกษาทั้งสิ้น 18 สถานศึกษา แบ่งเป็นสถานศึกษารัฐบาลจำนวน 6 แห่ง และสถานศึกษาของเอกชน 12 แห่ง แม้นจำนวนสถานศึกษาของเอกชนจะมากกว่าจำนวนสถานศึกษาของรัฐบาลแต่ก็ยังมีจำนวนนักเรียน-นักศึกษา น้อยกว่าสถานศึกษาของรัฐบาลอยู่มาก

เหมืองข้อมูลเป็นกระบวนการในการค้นหาความรู้จากข้อมูล นักวิจัยหลายท่านได้นำเอากระบวนการของเหมืองข้อมูลมาใช้ในการสร้างแบบจำลองเพื่อการพยากรณ์ เช่น R.K.Kavitha และ Dr. D.DoraiRangasamy[3] ได้เสนอการทำนายการรอดชีวิตจากมะเร็งเต้านมโดยใช้เทคนิค Naive Bayes และ C4.5 ผลการทดลองพบว่า เทคนิค C4.5 ให้ความถูกต้อง 97.9% ซึ่งสูงกว่า Naive Bayes

เพียงฤทัย หนูสวัสดิ์[4] ได้เสนอการสร้างโมเดลทำนายอัตราการใช้พลังงานของแบตเตอรี่มือถือโดยใช้เทคนิคเหมืองข้อมูล 2 เทคนิค คือ Perceptron Neural Network และ SVM แบบ kernel ผลการวิจัยพบว่าเทคนิค SVM แบบ kernel ให้ประสิทธิภาพความแม่นยำมากที่สุด เทคนิคในเหมืองข้อมูลที่เป็นที่นิยมและมีประสิทธิภาพ[5] เช่น C4.5, The *k*-means, SVM, The Apriori algorithm, The EM algorithm, PageRank, AdaBoost, *k*-NN, Naive Bayes, CART, Random Forest, Random Tree และ REP Tree เป็นต้น

วัตถุประสงค์เพื่อศึกษาปัจจัยที่มีผลต่อการสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์และพัฒนาแบบจำลองที่มีประสิทธิภาพในการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ โดยใช้ข้อมูลนักเรียนที่เรียนครบตามระยะเวลาของหลักสูตรประกาศนียบัตรวิชาชีพ (ปวช.) ในปีการศึกษา 2557-2558 เพื่อช่วยในการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ซึ่งสามารถช่วยส่งเสริมให้ครูแนะแนวและผู้บริหารมีข้อมูลสารสนเทศเพื่อวางแผนในการแนะแนวให้นักเรียน-นักศึกษาเข้าศึกษาต่อในสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ในงานวิจัยนี้ได้ใช้ 6 เทคนิคที่มีประสิทธิภาพในการสร้างแบบจำลอง คือ C4.5, Random Forest, Random Tree, REP Tree, k-NN และ SVM คณะผู้วิจัยได้ใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพการพยากรณ์ของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. สถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์

สถานศึกษาอาชีวศึกษาเอกชนสังกัดสำนักงานคณะกรรมการส่งเสริมการศึกษาเอกชน (สช.) และมีหน่วยงานต้นสังกัดคือสำนักงานเขตพื้นที่การศึกษาประถมศึกษากาฬสินธุ์ เขต 1 เขต 2 และเขต 3 ตามพื้นที่ที่สถานศึกษาตั้งอยู่ แต่ใช้หลักสูตรของสำนักงานคณะกรรมการการอาชีวศึกษา (สอศ.) จนมาถึงวันที่ 12 กุมภาพันธ์ 2559 เว็บไซต์ราชกิจจานุเบกษา

ได้เผยแพร่คำสั่งหัวหน้าคณะรักษาความสงบแห่งชาติที่ 8/2559 เรื่องการบริหารจัดการรวมสถานศึกษาอาชีวศึกษาภาครัฐและภาคเอกชน โดยให้รวมสถานศึกษาทั้งภาครัฐและเอกชนไว้ด้วยกันภายใต้สังกัดอาชีวศึกษาจังหวัดกาฬสินธุ์ (อศจ.กาฬสินธุ์)

2. สภาวะการขาดแคลนแรงงานฝีมือในประเทศไทย

พิเชษฐ์ สุขเสกสรรค์[6] ได้ทำการสำรวจประสบการณ์ของหน่วยงานกลุ่มรับเหมาก่อสร้างในการเผชิญปัญหาการขาดแคลนแรงงานก่อสร้างกลุ่มช่างฝีมือหรือช่างเทคนิค คือผู้ที่สำเร็จการศึกษาระดับประกาศนียบัตรวิชาชีพและประกาศนียบัตรวิชาชีพชั้นสูงโดยมีผลการสำรวจดังนี้ 1) หน่วยงานของท่านเคยประสบปัญหาการขาดแคลนแรงงานกลุ่มนี้คิดเป็นร้อยละ 89.7 2) หน่วยงานของท่านกำลังประสบปัญหาการขาดแคลนแรงงานกลุ่มนี้คิดเป็นร้อยละ 60.3 3) หน่วยงานของท่านได้เคยประเมินสถานการณ์หรือได้ประเมินสถานการณ์การขาดแคลนแรงงานกลุ่มนี้ในอนาคตคิดเป็นร้อยละ 75.6

จงจิตต์ ฤทธิรงค์, รีนา ต๊ะดี[7] กล่าวว่า ความต้องการแรงงานฝีมือในอุตสาหกรรมที่สำคัญ 3 ประเภท คือ อุตสาหกรรมยานยนต์ การผลิตอาหารและการท่องเที่ยว นั้นยังมีมากโดยเฉพาะแรงงานที่อยู่ในระดับปฏิบัติงานคือกลุ่มผู้ที่สำเร็จการศึกษาระดับประกาศนียบัตรวิชาชีพและประกาศนียบัตรวิชาชีพชั้นสูง เนื่องจากบุคลากรกลุ่มนี้มีทักษะที่จำเป็นต่อการปฏิบัติงาน จากการฝึกประสบการณ์วิชาชีพซึ่งมีความร่วมมือระหว่างสถานบันอาชีวศึกษาและภาคเอกชน ในการฝึกทักษะให้แก่ผู้เรียนมีความสามารถและทักษะตรงตามความต้องการของตลาดแรงงาน แต่อย่างไรก็ตาม การผลิตแรงงานฝีมือยังไม่เพียงพอและมีแนวโน้มว่าจะขาดแคลนอันเนื่องมาจากค่านิยมที่ให้ความสำคัญกับปริญญาบัตรมากกว่าความสามารถและทักษะในการทำงาน ดังสะท้อนออกมาเป็นอัตราค่าจ้างที่แปรผันตามระดับวุฒิการศึกษา

3. เทคนิคในเหมืองข้อมูล

เทคนิคในเหมืองข้อมูลได้ถูกนำมาใช้ในการสร้างแบบจำลองกันอย่างแพร่หลายและมีประสิทธิภาพ เช่น C4.5, Random Forest, Random Tree, REP Tree, k-NN และ SVM ซึ่งสามารถอธิบายได้ดังต่อไปนี้

1) เทคนิค C4.5[18] เป็นเทคนิคในการสร้างต้นไม้การตัดสินใจพัฒนาโดย J. Ross Quinlan ในปี 1993 โดยนำเอา ID3 มาปรับปรุงให้มีความสามารถมากขึ้นใช้วิธีการ Information Gain เพิ่มเติมการจัดการกับข้อมูล, ตัวเลข, ข้อมูลที่ขาดไปและไม่สมบูรณ์ และการ Prune ด้วยการแทนกิ่ง (Branch) ที่ไม่ช่วยในการตัดสินใจด้วย Leaf Node ที่ตัดสินใจได้ดีกว่า การแบ่งของ tree ในการทำงานขั้นตอนแรกคล้ายกับการทำงานด้วย ID3 คือต้องหา Info และ Gain ออกมาก่อน ซึ่งมีนักวิจัยหลายท่านได้นำเอาเทคนิคนี้มาใช้ในการพยากรณ์ เช่น Abdelghani Bellaachia และ Erhan Guven[8] ได้เสนอการพยากรณ์การรอดชีวิตจากมะเร็งเต้านมโดยใช้เทคนิคการทำเหมืองข้อมูล โดยใช้ 3 เทคนิค คือ Naive Bayes, BP-ANN และ C4.5 ผลการทดลองพบว่าการพยากรณ์การรอดชีวิตจากมะเร็งเต้านมโดยใช้เทคนิค C4.5 ให้ความถูกต้องร้อยละ 86 มากกว่าการพยากรณ์ด้วยเทคนิค BP-ANN ที่ให้ค่าความถูกต้องร้อยละ 85.5 และเทคนิค Naive Bayes ที่ให้ค่าความถูกต้องร้อยละ 84.5

2) เทคนิค Random Forest[9] เป็นเทคนิคการสุ่มเลือกใช้ข้อมูลและคุณลักษณะ Decision Tree ซึ่งถูกสร้างจากการนำข้อมูลไปสุ่มเลือกตัวอย่างแบบเลือกแล้วใส่กลับ (Sampling with Replacement) แล้วนำมาสร้างเป็น Tree ซึ่งจะมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือก ซึ่งข้อมูลส่วนนี้เรียกว่า Out-of-Bag (OOB) จะถูกนำมาใช้ในการทดสอบ Decision Tree วิธีการดังกล่าวนี้เรียกว่า Bagging ผลลัพธ์ที่ได้อย่างอิสระจาก Decision Tree ในแต่ละต้นถูกนำมาคิดเป็นผลการโหวต ผลโหวตที่มากที่สุดจะใช้ระบุสถานะของคลาสดัง Figure 1 เทคนิค Random Forest ไม่จำเป็นต้องมีข้อมูลทดสอบ เพื่อประมาณความผิดพลาดเพราะข้อมูล OOB นั้นถูก นำมาใช้ทดสอบ Decision Tree นั้นแล้ว

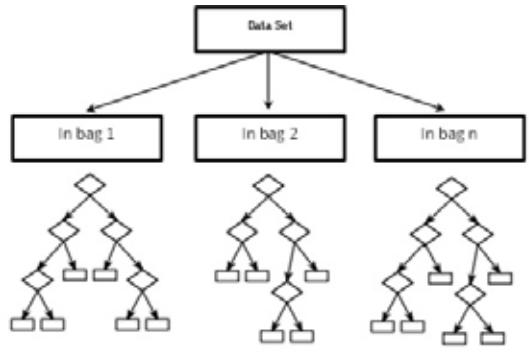

Figure 1 Characteristics of Random Forest

มีนักวิจัยจำนวนมากได้นำเทคนิค Random Forest มาใช้ในการจำแนก เช่น Krishnaveni[10] ได้พัฒนาแบบจำลองสำหรับทำนายการบาดเจ็บที่เกิดจากอุบัติเหตุบนท้องถนน โดยใช้เทคนิควิธี Naive, PART Rule, J48 Decision tree และ Random forest พบว่าการสร้างแบบจำลองด้วยเทคนิควิธี Random forest ให้ความถูกต้องคิดเป็นร้อยละ 74.34 ซึ่งดีกว่าวิธีอื่นๆ

ภรัณยา ปาลวิสุทธิ์[11] ได้เสนอการเพิ่มประสิทธิภาพเทคนิค Decision tree บนชุดข้อมูลที่ไม่สมดุล

โดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต ด้วยเทคนิค Decision Tree J48, ID3, LMT, CART และ Random Forest ผลการทดลองประสิทธิภาพในการพยากรณ์ของตัวแบบพบว่าเทคนิค Random Forest มีค่าความแม่นยำร้อยละ 87.15 สามารถพยากรณ์ได้ดีกว่า J48 ID3 LMT และ CART

3) เทคนิค Random Tree[12] คือ เทคนิคที่ใช้ในการจำแนกหมวดหมู่เช่นเดียวกับ C4.5 โดยมีหลักการสร้าง Tree จากการสุ่ม Tree หลายๆแบบ ในแต่ละโหนดแล้วเลือกมาประมวลผลโดยไม่ใช้การ Prune และเนื่องจากจำนวนของ Tree เพิ่มขึ้นอย่างรวดเร็ว ซึ่งยากแก่การแก้ปัญหา การสร้าง Tree ที่เป็นไปได้ทั้งหมด ส่วนประกอบสามารถสร้างชุดสุ่มของ Tree ออกมาจากการกระจายชุดต่างๆของ Tree Random Tree เป็นการสุ่มวาดที่สุ่มจากชุดของ Tree ที่เป็นไปได้ ในบริบทนี้ “สุ่ม” หมายความว่า Tree ในชุดของ Tree แต่ละ Tree มีโอกาสเท่าเทียมกันของการเป็นตัวอย่าง วิธีที่บอกนี้ก็คือว่าการกระจายของ Tree คือ “ชุด” Tree สุ่มแบบต่อเนื่อง (CRT) เป็นแบบสุ่ม Tree จริง T ๐ และมีระยะในการเดินทางสั้นที่สุด Tree ชุดย่อย ประกอบด้วย Tree ที่มี n จุด แต่ละองค์ประกอบของโครงสร้างที่แท้จริงคือราก ซึ่งมีนักวิจัยจำนวนมากได้นำเทคนิค Random Tree มาใช้ในการจำแนกเช่น ฐิติมา ช่วงชัย[13] ได้เสนอการวิเคราะห์หารูปแบบการเรียนรู้โดยใช้เหมืองข้อมูลของนักศึกษาต่อการจัดทำปริญญานิพนธ์ โดยใช้วิธีวิเคราะห์ผลด้วยรูปแบบ Rule Based Classification ด้วยวิธี Decision Table, Jrip และ PART และรูปแบบ Decision Tree Classification ด้วยวิธี LMT, J48 และ Random Tree จากการวิเคราะห์ผลทั้งหมดพบว่า รูปแบบของ Decision Tree ด้วยวิธีการ Random Tree ให้ค่าความถูกต้องสูงสุด(100%) ส่วนรูปแบบของ Rule Based ด้วยวิธี PART ให้ค่าความถูกต้องสูงสุด (84.12%)

Sushikumar Kalmegh[14] นำเสนอการเปรียบเทียบการวิเคราะห์อัลกอริทึมในโปรแกรม Weka ได้แก่ REP Tree, Simple Cart และ Random Tree ในการจำแนกข่าวอินเดียพบว่า Random Tree มีค่าความถูกต้อง 100% ซึ่งมากกว่า REP Tree และ Simple Cart

4) เทคนิค Reduced Error Pruning (REP Tree)[12] คือเทคนิคที่ใช้ regression tree logic และสร้าง Tree หลายๆ ต้นที่แตกต่างกัน หลังจากนั้นก็เลือกที่ดีที่สุดจาก Tree ที่สร้างทั้งหมดมาเป็นตัวแทนของ Tree ทั้งหมด ในการตัดกิ่งใช้ค่า mean square error ในการพยากรณ์เป็นพื้นฐานการวัด REP Tree เป็น Decision Tree ที่มีการเรียนรู้และสร้างแบบจำลองอย่างรวดเร็วบนพื้นฐานของ Information gain หรือ reducing the variance และตัดกิ่งโดยใช้การลดข้อผิดพลาดในการตัดแต่ใช้ได้เฉพาะตัวแปรที่เป็นตัวเลขเท่านั้น มีนักวิจัยจำนวนมากได้นำเทคนิค REP Tree มาใช้ในการจำแนก เช่น Prerna Kapoor และReena Rani[15] ได้เสนอประสิทธิภาพการตัดสินใจเทคนิค Decision Tree โดยใช้อัลกอริทึม J48 และ Reduced Error Pruning ผลการทดลองพบว่าทำให้การพยากรณ์มีความแม่นยำมากขึ้น

Kittipol Wisaeng[16] ได้เสนอการเปรียบเทียบอัลกอริทึมของ Decision Tree ในการจำแนก UCI Repository ได้แก่ Nursery, Iris, Anneal, Shuttle_trn, Voting, Waveform และ Sick โดยมีอัลกอริทึมคือ อัลกอริทึม functional tree, อัลกอริทึม logistic model trees, อัลกอริทึม REP Tree และอัลกอริทึม best-first decision tree ผลการทดลองพบว่า REP Tree ให้ค่าความถูกต้องที่ 92.87%

5) เทคนิค K-Nearest Neighbour (k-NN)[17] เป็นขั้นตอนวิธีการในการหาสมาชิกที่ใกล้ที่สุดเป็นอัลกอลิทึมที่ใช้ในการจัดกลุ่มข้อมูล โดยการจัดข้อมูลที่อยู่ใกล้กันให้เป็นกลุ่มเดียวกันซึ่งเทคนิคนี้จะทำให้ตัดสินใจได้ว่า คลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง โดยการตรวจสอบจำนวน k ซึ่งถ้าหากเงื่อนไขของการตัดสินใจมีความซับซ้อน วิธีนี้จะสามารถสร้างโมเดลที่มีประสิทธิภาพได้ แต่ขั้นตอนวิธีการหาสมาชิกที่ใกล้ที่สุดจะใช้ระยะเวลาในการคำนวณนาน ถ้าตัวแปร (แอตทริบิวต์) มีจำนวนมากจะเกิดปัญหาในการคำนวณค่าและค่อนข้างที่จะใช้ปริมาณงานในการคำนวณสูงมากบนคอมพิวเตอร์ เพราะเวลาที่ใช้สำหรับการคำนวณจะเพิ่มขึ้นแบบแฟคทอเรียลตามจำนวนจุดทั้งหมด ดังนั้นเพื่อจะเพิ่มความรวดเร็วสำหรับเทคนิคขั้นตอนวิธีการหาสมาชิกที่ใกล้ที่สุดให้มากขึ้น ข้อมูลทั้งหมดที่ใช้บ่อยจะต้องถูกเก็บไว้ในหน่วยความจำ โดยวิธีการเข้าถึงหน่วยความจำพื้นฐานอย่างมีเหตุผล (Memory-Based Reasoning) ซึ่งจะเป็นวิธีที่นำมาอ้างถึงเป็นประจำในการจัดเก็บกลุ่มคลาสของขั้นตอนวิธีการหาสมาชิกที่ใกล้ที่สุดในหน่วยความจำ และถ้าหากข้อมูลที่ต้องการหาคำตอบมีตัวแปรอิสระเพียงไม่กี่ตัวแล้ว จะทำให้เราสามารถเข้าใจโมเดลขั้นตอนวิธีการสมาชิกที่ใกล้ที่สุดได้ง่ายขึ้น ตัวแปรเหล่านี้ยังมีประโยชน์สำหรับนำมาสร้างโมเดลต่าง ๆ ที่เกี่ยวข้องกับชนิดของข้อมูลที่ไม่เป็นมาตรฐาน เช่น ข้อความเพียงแต่อาจต้องมีมาตรฐานการวัดค่าสำหรับชนิดของข้อมูลดังกล่าวที่เหมาะสมด้วย นอกจากนี้ประสิทธิภาพของขั้นตอนวิธีการสมาชิกที่ใกล้ที่สุดนี้ จะขึ้นอยู่กับจำนวนระยะห่าง การอธิบายระหว่างข้อมูลทั้งคู่ที่สามารถแบ่งแยกอย่างมีประสิทธิภาพระหว่างข้อมูลปกติ และข้อมูลผิดปกติ การอธิบายจำนวนระยะห่างระหว่างข้อมูลเป็นความท้าทายอย่าง

มากเมื่อข้อมูลมีความซับซ้อน อย่างเช่น ข้อมูลกราฟ และข้อมูลแบบลำดับเป็นต้น

6) เทคนิค Support Vector Machines (SVM)[18] คือ ขั้นตอนวิธีการที่มีความรวดเร็วและเป็นเทคนิคที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล ใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด

แนวความคิดของเทคนิควิธี SVM นั้นเกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกัน โดยจะทำการสร้างเส้นแบ่งที่เป็นเส้นตรงขึ้นมา เพื่อให้ทราบว่าเส้นตรงที่แบ่งกลุ่มสองกลุ่มออกจากกันนั้น เส้นใดเป็นเส้นที่ดีที่สุด สำหรับ SVM นั้นเดิมได้มีการนำมาใช้กับข้อมูลที่เป็นเชิงเส้น แต่ในความเป็นจริงแล้วข้อมูลที่นำมาใช้ในระบบการสอนให้ระบบเรียนรู้ส่วนใหญ่มักเป็นข้อมูลแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยการนำ Kernal Function มาใช้การจำแนกข้อมูลบนระนาบหลายมิติ จะใช้ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า โครงสร้างในการคัดเลือกซึ่งโครงสร้างในการคัดเลือกมาจากข้อมูลที่สอนให้ระบบเรียนรู้จำนวนเซตของโครงสร้างที่ใช้อธิบายในกรณีหนึ่ง เรียกว่า เวกเตอร์ ดังนั้นจุดมุ่งหมายของตัวแบบ SVM คือ แบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วยหนึ่งกลุ่มของตัวแปรของเป้าหมายที่อยู่ข้างหนึ่งของระนาบและกรณีของกลุ่มอื่นที่อยู่ทางระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดเรียกว่า ซัพพอร์ตเวกเตอร์ ซึ่งวิธีการนี้เหมาะสำหรับข้อมูลที่มีมิติของข้อมูลสูง

วิธีการดำเนินการวิจัย

ในการทำวิจัยคณะผู้วิจัยได้มีแบ่งวิธีการดำเนินการวิจัยออกเป็น 4 ขั้นตอนหลัก 1) การเตรียมข้อมูล 2) กระบวนการก่อนการสร้างแบบจำลอง 3) การสร้างแบบจำลอง และ 4) การวัดประสิทธิภาพแบบจำลอง ดัง Figure 2

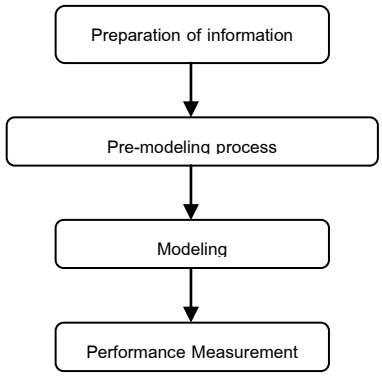

Figure 2 The process of data mining

1. การเตรียมข้อมูล

การเตรียมข้อมูลของการศึกษาค้นคว้าอิสระครั้งนี้ได้ข้อมูลมาจากการสำรวจข้อมูลจำนวนนักเรียนจากสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ประจำปีการศึกษา 2557-2558 จำนวน 12 แห่ง ดัง Table 1

Table 1 The number of students in the academic year 2014-2015

School	2014	2015
1. Kalasin Commercial Technology College	176	197
2. Kamalasai Technology College	38	35
3. Perm Poon Business Administration Technology College	21	19
4. Lampao Vocational College	13	14
5. Sahassakun Technology College	15	16
6. Samchai Technology College	9	11
7. Thai Tech Kalasin Vocational College	12	15
8. Thai Tech Esan Vocational College	23	12
9. Somdet Commercial Technology College	132	268
10. Pattanabandit Technology College	18	10
11. Thai Tech Asia Vocational College	130	82
12. Natchavin Technology College	-	-
Included	587	679
Total	1,266	

ซึ่งข้อมูลที่ทำการสำรวจได้แก่

1. เพศ
2. อายุ
3. จำนวนพี่น้อง
4. สถานบิดา มารดา
5. อาชีพบิดา
6. อาชีพมารดา
7. หมู่บ้าน
8. ตำบล
9. อำเภอ
10. จังหวัด
11. เกรดเฉลี่ยกลุ่มภาษาไทย
12. เกรดเฉลี่ยกลุ่มคณิตศาสตร์
13. เกรดเฉลี่ยกลุ่มวิทยาศาสตร์
14. เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม
15. เกรดเฉลี่ยกลุ่มสุขศึกษา พลศึกษา
16. เกรดเฉลี่ยกลุ่มศิลปะ
17. เกรดเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี
18. เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ
19. ประเภทวิชาที่จบ
20. จบการศึกษา/ไม่จบการศึกษา

2. การทำกระบวนการก่อนการสร้างแบบจำลอง

ในขั้นตอนนี้ผู้วิจัยได้ทำการแปลงข้อมูลและทำการวิเคราะห์ปัจจัย

1. การแปลงข้อมูล (Data transformation) จากข้อมูลที่ได้จากการสำรวจข้อมูลจำนวนนักเรียนจากสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ประจำปีการศึกษา 2557-2558 จำนวน 12 แห่ง มีทั้งหมด 20 ตัวแปร ดัง Table 2

Table 2 Variable details

Variables	Variable types	Description	Variable value
Sex	Nominal	Male	1
		Female	2
Age	Numeric	Age	Actual data
Number of siblings	Numeric	Number of siblings	Actual data
Parent Status	Nominal	Family	1
		Divorced	2
		Deceased	3
Father Career	Nominal	Government	1
		State Enterprises	2
		Trade	3
		Agriculture	4
		Contractors	5
		Government Employee	6
		Retired Government	7
		Priest	8
		Jobless	9
		Others	10
		Deceased	11
Mother Career	Nominal	Government	1
		State Enterprises	2
		Trade	3
		Agriculture	4
		Contractors	5
		Government Employee	6
		Retired Government	7
		Priest	8
		Jobless	9
		Others	10
		Deceased	11

Table 2 Variable details (continue)

Variables	Variable types	Description	Variable value
Village	Nominal	Village	Actual data
Tambol	Nominal	Tambol	Actual data
Amphoe	Nominal	Amphoe	Actual data
Province	Nominal	Province	Actual data
Thai	Nominal	Thai	Actual data
Math	Nominal	Math	Actual data
Science	Nominal	Science	Actual data
Social	Nominal	Social	Actual data
Health	Nominal	Health	Actual data
Art	Nominal	Art	Actual data
Career and Technology	Nominal	Career and Technology	Actual data
Foreign language	Nominal	Foreign language	Actual data
Type of course	Nominal	Commercial	1
		Industry	2
Graduate (Class)	Nominal	Successfully	1
		Unsuccessful	2

2. การวิเคราะห์ปัจจัยเป็นการคัดเลือกแอตทริบิวต์ที่สามารถเป็นตัวแทนของกลุ่มแอตทริบิวต์เพื่อลดจำนวนแอตทริบิวต์ในการพยากรณ์ซึ่งงานวิจัยนี้ได้นำเทคนิค Gain Ratio Attribute Evaluation มาใช้ในการวิเคราะห์ปัจจัยผลการวิเคราะห์ทำการตัดแอตทริบิวต์ที่มีผลต่อการพยากรณ์น้อยออกจำนวน 3 แอตทริบิวต์คือ Age, Number of siblings และ Province คงเหลือ 17 แอตทริบิวต์ ดัง Table 3

Table 3 Factors

Variables	
1. Amphoe	10. Thai
2. Tambol	11. Science
3. Father Career	12. Parent Status
4. Mother Career	13. Health
5.Vilage	14. Sex
6. Career and Technology	15. Type of course
7. Math	16. Foreign language
8. Social	17.Graduate(Class)
9. Art	

3. การสร้างแบบจำลอง

การสร้างแบบจำลองพยากรณ์การสำเร็จการศึกษาของผู้ที่สนใจเข้าศึกษาต่อสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ เพื่อช่วยในการตัดสินว่าเมื่อเข้าศึกษาต่อแล้วจะสามารถสำเร็จการศึกษาได้หรือไม่ด้วยตัวแปรทั้ง 21 ตัวแปรที่ได้จากการสำรวจข้อมูลจำนวนนักเรียนในปีการศึกษา 2557-2558 นั้น ด้วยการทำเหมืองข้อมูลและเทคนิคที่นำมาใช้ในการสร้างแบบจำลองมีจำนวน 6 เทคนิคคือ

เทคนิค C4.5
เทคนิค Random Forest
เทคนิค Random Tree
เทคนิค REP Tree
เทคนิค k-NN
เทคนิค SVM

4. การวัดประสิทธิภาพของแบบจำลอง

ในการวัดประสิทธิภาพของแบบจำลองนั้น ได้มีการใช้เทคนิคแบบ 10-fold cross validation โดยจะทำการแบ่งข้อมูลออกเป็น 10 ชุดเท่าๆกัน จากนั้นจะทำการทดสอบทั้งหมด 10 รอบ โดยในแต่ละรอบจะใช้ข้อมูล 1 ชุดเป็นชุดทดสอบและอีก 9 ชุดที่เหลือเป็นชุดฝึกสอน ในรอบต่อไปก็ใช้ชุดข้อมูลถัดไปเป็นชุดทดสอบจนครบทั้ง 10 ชุดข้อมูล ซึ่งข้อมูลทั้งหมด จำนวน 1266 ข้อมูล แบ่งออกเป็น 10 ชุดข้อมูล ซึ่งคิดเป็นอัตราข้อมูลทดสอบต่อข้อมูลฝึก เป็นอัตราส่วน 10:90

การวิเคราะห์ประสิทธิภาพ คือ การวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล โดยค่าของผลลัพธ์ที่ได้จากการจำแนกคือ

ค่า True Positive (TP) ค่า True Negative (TN) ค่า False Positive (FP) ค่า False Negative (FN) และสามารถหาค่าความถูกต้อง (Accuracy) จากสมการ

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

ค่าความแม่นยำ (Precision) จากสมการ

$$Precision = \frac{TP}{TP+FP}$$

และค่าความระลึก (Recall) ได้จากสมการ

$$\mathrm{Re}call = \frac{TP}{TP+FN}$$

ผลการศึกษา

ในการศึกษาประสิทธิภาพของแบบจำลองนี้คณะผู้วิจัยได้นำเอาโปรแกรม WEKA เวอร์ชัน 3.9.1 มาเป็นเครื่องมือมาใช้ในการสร้างแบบลองด้วยเทคนิค C4.5, Random Forest, Random Tree, REP Tree, k-NN, และ SVM วัดประสิทธิภาพของแบบจำลองด้วย Accuracy, Precision และ Recall ผลการทดลองสามารถแสดงได้ดัง Figure 3

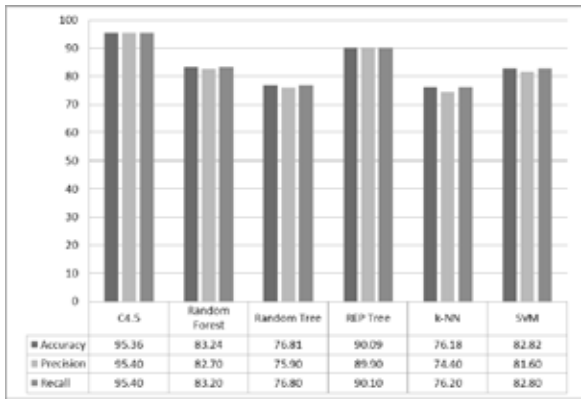

Figure 3 The performance of prediction models

Figure 3 แสดงให้เห็นค่า Accuracy ของแบบจำลองได้แก่เทคนิค C4.5 ให้ค่า Accuracy มากที่สุด 95.36% ต่อมาเทคนิค REP Tree ให้ค่า Accuracy 90.09% เทคนิค Random Forest ให้ค่า Accuracy 83.24% เทคนิค SVM ให้ค่า Accuracy 82.82% เทคนิค Random Tree ให้ค่า Accuracy 76.81% และเทคนิค k-NN ให้ค่า Accuracy 76.18% ตามลำดับ

ค่า Precision ของแบบจำลองได้แก่เทคนิค C4.5 ให้ค่า Precision มากที่สุด 95.4% ต่อมาเทคนิค REP Tree ให้ค่า Precision 89.9% เทคนิค Random Forest ให้ค่า Precision 82.7% เทคนิค SVM ให้ค่า Precision 81.6% เทคนิค Random Tree ให้ค่า Precision 75.9% และเทคนิค k-NN ให้ค่า Precision 74.4% ตามลำดับ

ค่า Recall ของแบบจำลองได้แก่เทคนิค C4.5 ให้ค่า Recall มากที่สุด 95.4% ต่อมา เทคนิค REP Tree ให้ค่า Recall 90.1%

เทคนิค Random Forest ให้ค่า Recall 83.2% เทคนิค SVM ให้ค่า Recall 82.8% เทคนิค Random Tree ให้ค่า Recall 76.8% และเทคนิค k-NN ให้ค่า Recall 76.2% ตามลำดับ

วิจารณ์และสรุป

ในการศึกษาและพัฒนาแบบจำลองที่เหมาะสมสำหรับการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ โดยแบบจำลองที่ใช้ในการเปรียบเทียบ 6 เทคนิคได้แก่ C4.5, Random Forest, Random Tree, REP Tree, k-NN และ SVM ผลการทดลองพบว่าเทคนิค C4.5 มีความเหมาะสมมากที่สุดในการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ โดยการวัดประสิทธิภาพการทำงานของแบบจำลองด้วยค่า Accuracy ได้ 95.36% ค่า Precision ได้ 95.4% และค่า Recall ได้ 95.4%

จากผลการทดลอง สรุปได้ว่า C4.5 มีความเหมาะสมในการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ หากผู้ที่สนใจศึกษาหรือสนใจที่จะพัฒนางานวิจัยนี้ ควรเพิ่มจำนวนข้อมูลชุดฝึกสอนให้มากขึ้น และเพิ่มในลักษณะของคำแนะนำให้แก่นักเรียนว่าเหมาะสมที่จะเรียนประเภทวิชาอะไร เป็นต้น เพื่อที่จะได้เป็นแนวทางในเลือกตัดสินใจในการเลือกเรียนต่อไป

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณ สำนักงานอาชีวศึกษาจังหวัดกาฬสินธุ์ สำนักงานอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ และคณาจารย์คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่ให้ความอนุเคราะห์ ข้อมูลและคำปรึกษาในการดำเนินงานวิจัยครั้งนี้

เอกสารอ้างอิง

1. ข่าวการศึกษา. การเพิ่มสัดส่วนสายอาชีพเป็น 60% สายสามัญเหลือ 40%. Available: http://www.unigang.com/Article/5170,(2559, 28 สิงหาคม).
2. วันฉัตร ทิพย์มาศ, "ปัจจัยที่มีผลต่อการตัดสินใจเข้าศึกษาต่อวิทยาลัยอาชีวศึกษาเอกชนของนักเรียนในเขตภาคใต้ตอนบน," วารสารศึกษาศาสตร์, vol. 24, กุมภาพันธ์-สิงหาคม 2556.
3. R.K.Kavitha and Dr.D.DoraiRangasamy, "Predicting Breast Cancer Survivability Using Naive baysein Classifier And C4.5 Algorithm," *Elysium Journal Engineering Research & Management,* 2014.
4. เพียงฤทัย หนูสวัสดิ์, "การสร้างโมเดลทำนายอัตราการใช้พลังงานของแบตเตอรี่มือถือโดยใช้เทคนิคเหมืองข้อมูล," มหาวิทยาลัยศิลปากร, 2556.
5. X. Wu, "Top 10 Algorithms in Data Mining," *Springer-Verlag London Limited,* 2007.
6. พิเชษฐ์ สุขเสกสรรค์, "บทบาทใหม่ของวิศวกรในภาวการณ์การขาดแคลนแรงงานช่างฝีมือหรือช่างเทคนิค," มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, Ed., ed, 2549.
7. จงจิตต์ ฤทธิรงค์ and รีนา ต๊ะดี, "ข้อท้าทายในการผลิตแรงงานฝีมือไทยเพื่อเข้าสู่ตลาดแรงงานประชาคมเศรษฐกิจอาเซียน," วารสารสถาบันวิจัยประชากรและสังคม มหาวิทยาลัยมหิดล pp. 129-147, 2558.
8. Abdelghani, Bellaachia, and E. Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques," The George Washington University, 2006.
9. L. Breiman, "Random Forest," *Machine Learning,* vol. 45, pp. 5-32, 2001.
10. K. S and H. M, "A Perspective Analysis of Traffic Accident using Data Mining Techniques," *International Journal of Computer Applications,* vol. 23, pp. 40-48, 2011.
11. ภรัณยา ปาลวิสุทธิ์ "การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุล โดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูล การเป็นโรคติดอินเทอร์เน็ต" วารสารเทคโนโลยีสารสนเทศ, vol. 12, pp. 54-63, 2559.
12. *Weka*. Available: http://www.cs.waikato.ac.nz/ml/index.html,(2016, 3 December).
13. ฐิติมา ช่วงชัย, "การวิเคราะห์หารูปแบบการเรียนรู้โดยใช้เหมืองข้อมูลของนักศึกษาต่อการจัดทำปริญญานิพนธ์," วารสารบัณฑิตศึกษา มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ในพระบรมราชูปถัมภ์, vol. 10, pp. 53-62, 2559.
14. S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News," *International Journal of Innovative Science, Engineering & Technology,* vol. 2, pp. 438-446, 2015.
15. P. Kapoor and R. Rani, "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning," *International Journal of Engineering Research and General Science,* vol. 3, pp. 1613-1621, 2015.
16. K. Wisaeng, "A Comparison of Decision Tree Algorithms For UCI Repository Classification," *International Journal of Engineering Trends and Technology,* vol. 4, pp. 3393-3397, 2013.
17. ศลิษา หนูเสมียน, "ระบบแนะนาการเลือกสาขาเพื่อศึกษาต่อระดับอาชีวศึกษาโดยเทคนิคการคัดกรองข้อมูลแบบผสมระหว่างการคัดกรองข้อมูลแบบอิงเนื้อหากับการคัดกรองแบบพึ่งพาผู้ใช้ร่วม กรณีศึกษา วิทยาลัยสารพัดช่างระยอง," มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2554.
18. J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second Edition ed.: Morgan Kaufmann, 2006.