

การใช้เทคนิคเหมืองข้อมูลในการเลือกกลุ่มสาขาวิชาที่เหมาะสมสำหรับการศึกษาระดับปริญญาตรี

A Selected-proper Major group of Bachelor Degrees using Data Mining Techniques

อนันต์ ปิณะเต¹,ฉัตรเกล้า เจริญผล², แกมกาญจน์ สมประเสริฐศรี³

Anan Pinate¹,Chatklaw Jareanpon², Gamkan Somprasertsri³

Received: 12 December 2013 ; Accepted: 20 February 2014

บทคัดย่อ

มหาวิทยาลัยมหาสารคามมีการจัดการเรียนการสอนในระดับปริญญาตรี และบัณฑิตศึกษา การดำเนินการรับบุคคลเข้าศึกษาในระดับปริญญาตรี มีกระบวนการในคัดเลือกผู้สมัครที่มีคุณสมบัติตรงตามกลุ่มสาขาวิชาที่มหาวิทยาลัยกำหนด จากปัญหาที่สำคัญในการคัดเลือกเข้าศึกษา คือ การเลือกกลุ่มสาขาวิชาที่ไม่ตรงกับความรู้ความสามารถของผู้สมัครเมื่อผู้สมัครเข้ามาศึกษาในมหาวิทยาลัยแล้วค้นพบตัวเองในภายหลังว่าไม่เหมาะกับการเรียนในกลุ่มสาขาวิชานั้น ส่งผลกระทบต่อผลการเรียนที่ไม่ผ่านตามเกณฑ์ที่มหาวิทยาลัยกำหนด ทำให้ต้องพ้นสภาพการเป็นนิสิต บทความนี้จึงได้นำเสนอการใช้เทคนิคเหมืองข้อมูลในการเลือกกลุ่มสาขาวิชาที่เหมาะสมในระดับปริญญาตรี เพื่อการประชาสัมพันธ์ข้อมูลการสมัครให้ผู้สมัครได้เลือกกลุ่มสาขาวิชาได้ตรงตามทักษะ ความรู้ความสามารถของผู้สมัคร การวิจัยพบว่าประสิทธิภาพของการจำแนกข้อมูลแบบวิธีต้นไม้ตัดสินใจมีประสิทธิภาพดีที่สุดมีค่าความถูกต้องสูงสุด จากข้อมูลการทดลองตามกลุ่มประกันคุณภาพการศึกษาของมหาวิทยาลัย กลุ่มมนุษยศาสตร์และสังคมศาสตร์ ร้อยละ 82.22 กลุ่มวิทยาศาสตร์เทคโนโลยี ร้อยละ 88.26 และกลุ่มวิทยาศาสตร์สุขภาพ ร้อยละ 87.73

คำสำคัญ : การพยากรณ์การเลือกกลุ่มสาขาวิชา เทคนิคต้นไม้ตัดสินใจ เทคนิคเหมืองข้อมูล

Abstract

Maharakham University has provided several courses for both undergraduate and graduate levels. For the undergraduate level, there are procedures for selecting qualified applicants to study in an appropriate area of study since problems have arisen when applicants choose to study in an area that is not suitable for their capability. The impact from this problem is a higher rate of retirement when the students cannot complete their degree. This journal presents a data mining technique for selecting appropriate groups to study for undergraduate level. This method aims to give students the information to assist them in selecting an appropriate area of study which matches their skills. Results from this research indicates that 'Decision Tree Classifier' is the most effective data classification since it gives the most adequate value. The experiment from Educational Quality Assurance group indicates that result which are a) Humanities and Social Sciences group (82.22%), b) Science and Technology group (88.26%), and c) Health Sciences group (87.73%).

Keywords: Major area selected forecasting, Decision tree Technique, Data mining Technique

¹ นิสิตระดับบัณฑิตศึกษา, สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ, มหาวิทยาลัยมหาสารคาม จังหวัดมหาสารคาม 44150

² อาจารย์,สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ, มหาวิทยาลัยมหาสารคาม จังหวัดมหาสารคาม 44150

³ ผู้ช่วยศาสตราจารย์, สาขาวิชาสารสนเทศศาสตร์ คณะวิทยาการสารสนเทศ, มหาวิทยาลัยมหาสารคาม จังหวัดมหาสารคาม 44150

¹ Graduate Student in Information technology, Faculty of Informatics, Maharakham University, Maharakham 44150.

² Lecturer, Department of Computer science, Faculty of Informatics, Maharakham University, Maharakham 44150.

³ Assist.Prof., Department of Information science, Faculty of Informatics, Maharakham University, Maharakham 44150.

บทนำ

มหาวิทยาลัยมหาสารคามเป็นสถาบันอุดมศึกษาของรัฐ มีหลักสูตรการเรียนการสอน 19 คณะมีสาขาวิชา 85 สาขาวิชาในระดับปริญญาตรี มีการดำเนินการรับสมัครคัดเลือกบุคคลเข้าศึกษา 3 ช่องทางการรับเข้าศึกษา ได้แก่ ระบบรับตรงระบบกลาง (Admissions) และระบบพิเศษ มหาวิทยาลัยมหาสารคามดำเนินการรับสมัครเองในระบบรับตรง และระบบพิเศษ ส่วนระบบกลาง สำนักงานคณะกรรมการการอุดมศึกษา (สกอ.) เป็นฝ่ายดำเนินการ ปัญหาที่สำคัญประการหนึ่งในการรับนิสิตเข้าศึกษาในระดับปริญญาตรี คือ การเลือกกลุ่มสาขาวิชาของผู้สมัครที่ไม่ตรงตามทักษะความรู้ ความสามารถของผู้สมัครโดยผู้สมัครส่วนใหญ่จะเลือกกลุ่มสาขาวิชาตามความรู้สึก ความชอบ หรือสภาพแวดล้อมทั้งเพื่อน และผู้ปกครองเป็นหลักโดยอาจไม่ทราบถึงกลุ่มสาขาวิชาที่เหมาะสมกับความรู้ ความสามารถและลักษณะเฉพาะของตนเอง จึงส่งผลให้เมื่อเข้ามาศึกษาในระดับอุดมศึกษาแล้วค้นพบตัวเองในภายหลังว่าไม่เหมาะสมกับกลุ่มสาขาวิชานั้นจนส่งผลกระทบต่อให้นิสิตมีผลการเรียนตกต่ำ เกิดปัญหาหาการขอย้ายสาขาวิชาเรียน การขอพักการศึกษา การลาออกจากการเป็นนิสิต รวมถึงปัญหาการผันสภาพการเป็นนิสิต และปัญหาอีกประการ คือ การผลิตบุคคลากรในสาขาวิชานั้นมีประสิทธิภาพไม่สูงเท่าที่ควร ปัญหาที่นิสิตผันสภาพการเป็นนิสิตในระดับปริญญาตรีมีอัตราสูงขึ้นทุกปีสาเหตุของการผันสภาพนิสิตเกิดจากนิสิตมีผลการเรียนไม่ผ่านตามเกณฑ์ที่กำหนดมีคะแนนเฉลี่ยรวม (GPA) อยู่ในระดับต่ำทำให้นิสิตไม่สามารถสำเร็จการศึกษาตามหลักสูตรที่มหาวิทยาลัยกำหนด จากรายงานประจำปีการศึกษาของทะเบียนและประมวลผล มหาวิทยาลัยมหาสารคาม พบว่ามีจำนวนนิสิตที่ผันสภาพการเป็นนิสิตปีการศึกษา 2552 - 2555 ร้อยละ 5.22, 10.30, 10.32 และ 10.48 ตามลำดับ จากข้อมูลดังกล่าวเห็นได้ว่ามีจำนวนนิสิตที่ผันสภาพการเป็นนิสิตเพิ่มขึ้นทุกปีการศึกษา จากรายงานข้อมูลศูนย์พัฒนาและประกันคุณภาพการศึกษามหาวิทยาลัยมหาสารคาม พบว่ามาตรฐานด้านคุณภาพบัณฑิตที่สำเร็จการศึกษาจากมหาวิทยาลัยมหาสารคามมีอัตราการได้งานทำภายในหนึ่งปี มีจำนวนร้อยละ 68.16 และผลการประเมินคุณภาพการศึกษาจากภายนอกกรอบสามสำนักงานรับรองมาตรฐานและประเมินคุณภาพการศึกษา (องค์กรมหาชน : สมศ.) ปี 2554 ถึง 2558 ระดับอุดมศึกษา พบว่า กลุ่มสาขาวิชาวิทยาศาสตร์สุขภาพ กลุ่มสาขาวิชาวิทยาศาสตร์และเทคโนโลยี และกลุ่มสาขาวิชามนุษยศาสตร์สังคมศาสตร์มีผลการประเมินในระดับดี³

ด้วยเหตุผลดังกล่าว มหาวิทยาลัยมหาสารคามมีความต้องการที่จะแก้ไขปัญหาที่เกิดขึ้น ผู้วิจัยจึงมีความต้องการที่จะศึกษารูปแบบข้อมูลที่ส่งผลต่อการเลือกกลุ่มสาขาวิชาที่เหมาะสมในระดับปริญญาตรี โดยนำข้อมูลประวัติ นิสิตที่สำเร็จการศึกษา และข้อมูลประวัติการสมัครเข้าศึกษา เพื่อหาความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลด้วยวิธีการเทคนิคเหมืองข้อมูล (Data Mining) และนำผลการทดลองที่ได้เปรียบเทียบกับประสิทธิภาพ เพื่อได้ผลลัพธ์ที่ดีที่สุดมาพัฒนา รูปแบบระบบพยากรณ์การเลือกกลุ่มสาขาวิชาที่เหมาะสมในระดับปริญญาตรี

งานวิจัยฉบับนี้ผู้วิจัยได้นำอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree C4.5) วิธีการเรียนรู้แบบอย่างง่าย (Naïve Bayesian Learning) และวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine :SVM) มาเปรียบเทียบกับเพื่อให้เห็นถึงประสิทธิภาพของอัลกอริทึมทั้งสามแบบดังกล่าว ซึ่งทฤษฎีและกระบวนการในการทำเหมืองข้อมูลของแต่ละอัลกอริทึมมีรายละเอียดดังต่อไปนี้

เหมืองข้อมูล

เหมืองข้อมูล (Data Mining)⁴คือ การนำเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) วิธีการทางสถิติ (Statistical Methods) วิธีการปัญญาประดิษฐ์ (Artificial Intelligence) หรือวิธีอื่นๆ มาทำการวิเคราะห์และสกัดความรู้จากข้อมูลที่จัดเก็บไว้ในฐานข้อมูลหรือจัดเก็บไว้ในรูปอื่น ๆ โดยจุดประสงค์ของการทำเหมืองข้อมูล คือ การวิเคราะห์แนวโน้มความสัมพันธ์ กฎหรือรูปแบบของข้อมูลซึ่งเป็นความรู้ที่ถูกซ่อนอยู่ภายใต้ข้อมูลขนาดใหญ่ และนำเสนอสารสนเทศที่ได้มาใช้เพื่อช่วยในการวางแผน การตัดสินใจในการบริหารหรือแก้ปัญหาในด้านต่างๆ ซึ่งถือได้ว่าเป็นเครื่องมือที่ช่วยเพิ่มคุณค่าให้กับข้อมูลที่มีอยู่ เหมืองข้อมูลจะแก้ไขปัญหาได้บางปัญหาเท่านั้นตามเทคนิควิธีการที่เลือกใช้ได้เท่านั้น ประโยชน์หลักของการทำเหมืองข้อมูล คือ การค้นหาความรู้ที่ซ่อนอยู่ในฐานข้อมูลเพื่อให้ได้ซึ่งความรู้มาช่วยในการตัดสินใจ

ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree)^{4,5}เป็นโมเดลที่มีรูปแบบที่นิยม โครงสร้างต้นไม้ตัดสินใจเป็นแบบลำดับชั้น โดยมีการตัดสินใจ ซึ่งประกอบด้วย โหนดที่ใช้ในการตัดสินใจ (Decision node) และโหนดใบ (Leaf node or Terminal node) แต่ละโหนดตัดสินใจนั้นจะมีการสร้างฟังก์ชันที่เอาไว้สำหรับทดสอบทางเลือกจากการป้อนข้อมูลเข้าจะทดสอบตามทางเลือกไปเรื่อยๆ ไปจนถึง Terminal node จะได้คำตอบในที่สุด

การเรียนรู้แบบเบย์อย่างง่าย

อัลกอริทึมวิธีการเรียนรู้แบบเบย์อย่างง่าย (Naïve Bayesian Learning)⁶ เป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง ใช้งานได้ดีและเหมาะสมกับกรณีของเซตตัวอย่างที่มีจำนวนมากและมีคุณลักษณะ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน

คุณลักษณะเด่นของการจำแนกประเภทด้วยวิธีแบบเบย์อย่างง่าย (Naïve Bayes) คือ การเรียนรู้สามารถใช้กับข้อมูลที่เป็นค่าว่างของข้อมูล (Missing Value) โดยไม่เอาคุณลักษณะนั้นมาคำนวณและมีความเสถียรต่อค่ารบกวน เนื่องจากทฤษฎีเบย์นั้นมีแนวคิดพื้นฐานมาจากความน่าจะเป็น มีการนำการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่นเนื่องจากวิธีการจำแนกข้อมูลที่มีประสิทธิภาพและมีการทำงานไม่ซับซ้อนเหมือนวิธีการอื่น

ซัพพอร์ตเวกเตอร์แมชชีน

เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine :SVM)⁷ เป็นเทคนิคเพื่อใช้งานด้านการแบ่งข้อมูลที่เกินระนาบให้แบ่งกลุ่มข้อมูลที่เหมาะสมที่สุดเพียงระนาบเดียวเท่านั้น ที่สามารถรักษาระยะห่างระหว่างข้อมูลสองกลุ่มที่ใกล้กันมากที่สุด

การหาระนาบเกินที่เหมาะสมที่สุดของซัพพอร์ตเวกเตอร์แมชชีนนั้นจะทำการหาตำแหน่งของซัพพอร์ตเวกเตอร์ (Support Vector) เพื่อให้เป็นตัวแทนของกลุ่มข้อมูลทั้งคู่ เพื่อใช้เป็นข้อมูลในการแบ่งกลุ่มโดยหลักการ คือ จะใช้ระนาบเกินที่เป็นระยะห่างที่สุทธระหว่างข้อมูลสองกลุ่ม ที่มีอยู่ใกล้กันมากที่สุดเพียงระนาบเดียวเท่านั้น

งานวิจัยที่เกี่ยวข้อง

กฤษณะ ไวยมัย และคณะ⁸ ได้นำเสนอเทคนิคเหมืองข้อมูลเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์ การค้นหากฎการจำแนกข้อมูลโดยการพยากรณ์ข้อมูลมาประยุกต์ใช้สำหรับช่วยแนะนำนิสิตเลือกสาขาที่เหมาะสมที่สุด พร้อมทำนายผลการเรียนแต่ละรายวิชาในภาคการศึกษาถัดไป โดยนำเสนอแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ ข้อมูลที่ใช้ทดลอง ข้อมูลประวัติส่วนตัว และข้อมูลการลงทะเบียนนิสิต คณะวิศวกรรมศาสตร์ จากงานวิจัยสามารถทำนายสาขาวิชาที่เหมาะสมที่สุดให้กับนิสิตได้และมีความถูกต้องค่อนข้างสูง

ชิตชนก ส่งศิริ และคณะ⁹ ใช้เทคนิคเหมืองข้อมูลเพื่อค้นหาภาควิชาที่เหมาะสมที่สุดให้กับนิสิตงานวิจัยนี้ได้นำเสนอ

ปัญหาการเลือกสาขาวิชาที่ไม่ตรงกับความสามารถที่แท้จริง ซึ่งส่งผลให้การผลิตบุคลากรในสาขานั้น ๆ มีประสิทธิภาพไม่สูงเท่าที่ควร โดยได้เสนอการใช้เทคนิค Classification ด้วยวิธีต้นไม้ตัดสินใจซึ่งข้อมูลที่ใช้ในนำทดสอบ ประกอบด้วยข้อมูลสองส่วน คือข้อมูลประวัติส่วนตัวของนิสิต และข้อมูลการลงทะเบียนเรียนของนิสิต ผลการทดลองเป็นที่น่าพอใจมีค่าความถูกต้อง ร้อยละ 80.00

Pathom Pumpuang และคณะ¹⁰ ได้ทำการวิจัยเปรียบเทียบวิธีการ Bayesian Network, C4.5, Decision Forest และ NBTree ในการวางแผนหลักสูตรการลงทะเบียนเรียนของนักศึกษา จากข้อมูลนักศึกษาสาขาวิทยาการคอมพิวเตอร์ และสาขาวิชาวิศวกรรมคอมพิวเตอร์ จำนวน 1,533 คน งานวิจัยนี้จึงได้มีการวางแผนการลงทะเบียนเรียนของนักศึกษา โดยใช้ข้อมูลผลการเรียน (GPA) เพื่อเปรียบเทียบประสิทธิภาพความถูกต้อง ผลการทดลองพบว่า Decision Forest (DF) มีค่าความถูกต้อง ร้อยละ 80.33, Bayesian Network (BN) มีค่าความถูกต้อง ร้อยละ 83.78, C4.5 มีค่าความถูกต้อง ร้อยละ 81.86 และ NBTree มีค่าความถูกต้อง ร้อยละ 87.99

จากการศึกษางานวิจัยที่เกี่ยวข้องซึ่งได้กล่าวถึงการใช้เทคนิคเหมืองข้อมูลในการเลือกสาขาวิชาที่เหมาะสม โดยใช้ข้อมูลนักศึกษาชั้นปีที่ 1 คณะวิศวกรรมศาสตร์ในการทดลองในการเลือกสาขาวิชาที่เหมาะสมในชั้นปีที่ 2^{8,9} ซึ่งงานวิจัยดังกล่าวมีวัตถุประสงค์คล้ายกับงานวิจัยนี้ซึ่งใช้ข้อมูลนักเรียนที่สมัครเข้าศึกษาในระดับปริญญาตรีในการทดลอง และได้เปรียบเทียบประสิทธิภาพแบบจำลองการเลือกกลุ่มสาขาวิชา ระหว่างวิธีต้นไม้ตัดสินใจ อัลกอริทึม C4.5 วิธีการเรียนรู้แบบเบย์อย่างง่าย (Naïve Bayes) และวิธีซัพพอร์ตเวกเตอร์แมชชีน (SVM) จากการเปรียบเทียบประสิทธิภาพแบบจำลองได้พัฒนาระบบพยากรณ์การเลือกกลุ่มสาขาวิชาของผู้สมัครเข้าศึกษาในระดับปริญญาตรี มหาวิทยาลัยมหาสารคามจากแบบจำลองที่ดีที่สุด

การวัดประสิทธิภาพแบบจำลอง

ในการวัดประสิทธิภาพของแบบจำลองในแต่ละเทคนิควิธีต่าง ๆ โดยใช้ค่าความแม่นยำของแบบจำลอง (Accuracy) จะต้องทำการเลือกข้อมูลสำหรับเรียนรู้ (Training Set) และข้อมูลสำหรับทดสอบ (Testing Set) งานวิจัยนี้เลือกใช้วิธีการสุ่มเลือกข้อมูลแบบความเที่ยงตรง 10กลุ่ม (10-fold Cross Validation)¹¹

การสุ่มข้อมูลแบบความเที่ยงตรงเป็นการสุ่มตัวอย่างข้อมูลโดยเริ่มจากการแบ่งชุดข้อมูลออกเป็น ส่วน ๆ กัน นำข้อมูลบางส่วนมาทำการเรียนรู้ และนำข้อมูลส่วนที่เหลือมา

ทำการทดสอบแบบจำลองที่ได้จากเรียนรู้ โดยในการทำงานจะทำการแบ่งกลุ่มข้อมูลออกเป็นสิบชุดเท่าๆ กันในการทดลอง ครั้งแรกจะเลือกข้อมูลชุดที่หนึ่ง เป็นข้อมูลชุดทดสอบและข้อมูลที่เหลือเป็นชุดเรียนรู้ ในการทดลองครั้งถัดไปจะเลือกข้อมูลชุดที่สอง เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นชุดเรียนรู้ ทำจนกระทั่งข้อมูลได้ถูกนำมาเป็นข้อมูลทดสอบและชุดเรียนรู้ ซึ่งจะมีการทดลองทั้งหมด 10 ครั้ง

งานวิจัยเพื่อทำการเปรียบเทียบประสิทธิภาพของแบบจำลองในแต่ละวิธีในการเลือกกลุ่มสาขาวิชาที่เหมาะสมสำหรับศึกษาต่อระดับปริญญาตรี การวัดประสิทธิภาพของการจำแนกข้อมูลตามแนวคิดทางด้านการค้นคืนสารสนเทศซึ่งการวัดค่าประสิทธิภาพของแบบจำลองนั้นจะอาศัย Confusion matrix ในการคำนวณค่าดังรูปที่ 1

		Predicted Class	
		Class C ₁	C ₂
Actual	C ₁	TP	FN
	C ₂	FP	TN

Figure 1 Confusion matrix

ค่าความถูกต้อง (Accuracy) คือ ค่าที่บอกถึงความถูกต้องของค่าที่วัดได้ว่าใกล้เคียงค่าจริงเพียงใดสามารถอธิบายดังสมการที่ 1

ค่าความแม่นยำ (Precision) คือ ค่าสัดส่วนของข้อมูลที่ถูกต้อง และตรงตามความต้องการด้วยข้อมูลที่ถูกต้องค้นคืนทั้งหมดหรือเป็นค่าที่แสดงถึงการค้นคืนสารสนเทศและมีความเป็นไปได้ตรงตามความต้องการมากหรือน้อยเพียงใด ดังสมการที่ 2

ค่าความระลึก (Recall) คือ ค่าสัดส่วนของข้อมูลที่ตรงตามความต้องการที่ถูกค้นคืนกับข้อมูลที่ต้องการทั้งหมด หรือค่าที่แสดงถึงความครอบคลุมในการสืบค้นสารสนเทศที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการมากที่สุด ดังสมการที่ 3

ค่าความเหวี่ยง (F-Measure) คือ เป็นการหาเฉลี่ยค่าความแม่นยำในการตรวจพบและค่าความครบถ้วนในการตรวจพบเข้าด้วยกัน จึงเปรียบเสมือนค่าวัดความแม่นยำโดยรวมดังสมการที่ 4

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

โดย True Positive (TP) คือ จำนวนข้อมูลแบบจำลองจำแนกกลุ่ม C₁ และคำตอบเป็นกลุ่ม C₁

True Negative (TN) คือ จำนวนข้อมูลแบบจำลองจำแนกกลุ่ม C₂ และคำตอบเป็นกลุ่ม C₂

False Positive (FP) คือ จำนวนข้อมูลแบบจำลองจำแนกกลุ่ม C₁ และคำตอบเป็นกลุ่ม C₂

False Negative (FN) คือ จำนวนข้อมูลแบบจำลองจำแนกกลุ่ม C₂ และคำตอบเป็นกลุ่ม C₁

วิธีการดำเนินการวิจัย

งานวิจัยนี้ใช้ข้อมูลประวัติการศึกษาในอดีตที่สำเร็จ การศึกษาตั้งแต่ปีการศึกษา 2549-2554 และข้อมูลการสมัครเข้าศึกษาในระดับปริญญาตรี มหาวิทยาลัยมหาสารคาม โดยข้อมูลทั้งสองส่วนนี้ต้องเป็นข้อมูลที่มีความสัมพันธ์กันเป็นข้อมูลบุคคลเดียวกัน ในการทดลองได้แบ่งกลุ่มข้อมูลตามกลุ่มประกันคุณภาพการศึกษาของมหาวิทยาลัย ได้แก่ กลุ่มมนุษยศาสตร์และสังคมศาสตร์ จำนวน 11,643 ชุดข้อมูล กลุ่มวิทยาศาสตร์เทคโนโลยี จำนวน 4,610 ชุดข้อมูล และกลุ่มวิทยาศาสตร์สุขภาพ จำนวน 2,626 ชุดข้อมูล การทดลองทำการเตรียมข้อมูลให้เหมาะสมและอยู่ในรูปแบบที่สามารถนำไปใช้กับโปรแกรม WEKA (Waikato Environment for knowledge Analysis)¹² ที่ประกอบด้วยข้อมูลส่วนทดสอบ (Data Testing) และข้อมูลเรียนรู้ (Data Training) โดยใช้วิธี 10-fold Cross-Validation ในการแบ่งกลุ่มข้อมูลในการทดลอง จากนั้นนำผลจากการทดลองมาวิเคราะห์และสรุปผลการทดลองหาอัลกอริทึมที่เหมาะสมในการที่จะพัฒนาเป็นระบบต้นแบบสำหรับพัฒนาระบบสนับสนุนการตัดสินใจการสมัครเข้าศึกษาในกลุ่มสาขาวิชาที่เหมาะสมระดับปริญญาตรี แสดงกรอบแนวคิดได้ Figure 2

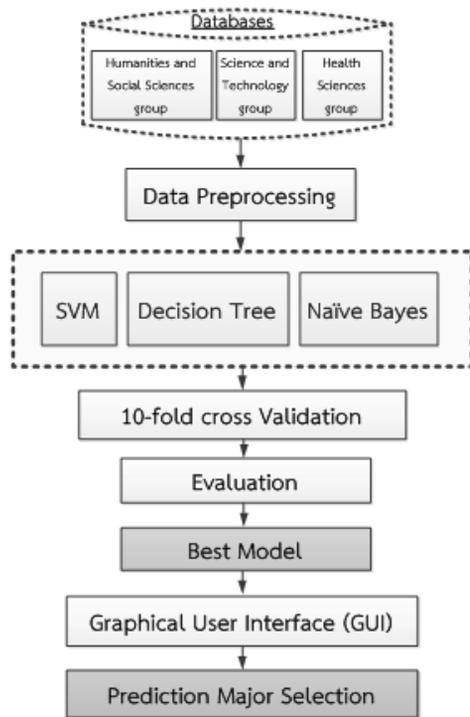


Figure 2 Conceptual framework

ข้อมูลกลุ่มประกันคุณภาพการศึกษามหาวิทยาลัยมหาสารคาม มีการแบ่งออกเป็น 3 กลุ่มประกันคุณภาพซึ่งประกอบไปด้วยคณะสาขาวิชาในแต่ละกลุ่มประกันคุณภาพมหาวิทยาลัย รายละเอียดกลุ่มประกันคุณภาพดังตารางที่ 1

Table 1 Group Quality Assurance Mahasarakham University.

Group	Faculty/College
Humanities and Social Sciences	<ul style="list-style-type: none"> - Humanities and Social sciences - Education - Business School - Fine and Applied Arts - Tourism and Hotel Management - College of Politics and Governance - College of Music - Cultural Science
Science and Technology	<ul style="list-style-type: none"> - Science - Technology - Informatics - Engineering - Architecture Urban Design And Creative Arts - Environment and Resource Studies

Health Sciences	<ul style="list-style-type: none"> - Nursing - Pharmacy - Public Health - Medicine - veterinary Medicine
-----------------	---

ข้อมูลที่ใช้ทำการทดลองนั้นประกอบด้วยข้อมูลตัวแปรโดยได้เลือกข้อมูลที่มีความสำคัญต่อการเลือกกลุ่มสาขาวิชาในระดับปริญญาตรี ข้อมูลตัวแปร (Attribute) มีดังนี้ เพศ, ขนาดสถานศึกษา, ผลการเรียนเฉลี่ยสะสม, ผลการเรียนกลุ่มสาระวิชา ภาษาไทย, คณิตศาสตร์, วิทยาศาสตร์, สังคมศาสตร์, สุขศึกษา, ศิลปะ, การงานอาชีพ, ภาษาต่างประเทศ และตัวแปรคลาส (Class) คือ ผลการเรียนที่สำเร็จการศึกษาจากมหาวิทยาลัย รายละเอียดข้อมูลตัวแปรการทดลองดัง Table 2 การทดลองจะแยกชุดข้อมูลออกเป็นกลุ่มข้อมูลตามกลุ่มประกันคุณภาพมหาวิทยาลัย ได้แก่ กลุ่มมนุษยศาสตร์และสังคมศาสตร์ กลุ่มวิทยาศาสตร์เทคโนโลยี

Table 2 Variable.

Variable	Explanation
Sex	Sex
Size	School size
GPAX	Grade point average
GPA1	Grade point average Thai Language
GPA2	Grade point average Mathematics
GPA3	Grade point average Science
GPA4	Grade point average Social studies
GPA5	Grade point average Health education
GPA6	Grade point average Art
GPA7	Grade point average Home working and technology
GPA8	Grade point average English
GPA	Grade point average in Mahasarakham University

และกลุ่มวิทยาศาสตร์สุขภาพ หลังจากได้ข้อมูลตัวแปรที่สำคัญในการทดลองแล้วลักษณะข้อมูลจะมีรายละเอียด Table 3 ตัวอย่างชุดข้อมูลกลุ่มมนุษยศาสตร์และสังคมศาสตร์

Table 3 Data sample

Sex	Size	GPAX	GPA1	GPA2	GPA3	GPA4	GPA5	GPA6	GPA7	GPA8	GPA
Male	Extra	3.00	3.00	2.14	2.54	3.47	3.83	3.90	3.85	3.45	3.74
Female	Big	2.44	2.60	2.20	1.68	3.30	3.10	3.90	3.60	2.40	2.19
Female	Big	2.63	2.16	2.12	2.75	2.45	3.70	3.70	3.53	2.11	2.91
Female	Small	2.28	1.42	1.84	3.15	3.60	3.10	3.07	2.60	2.10	2.45
Male	Extra	3.37	3.68	2.40	3.77	3.12	3.62	3.81	3.50	3.27	3.31
Female	Me- dium	2.47	2.00	1.43	2.13	2.26	3.66	2.91	3.16	2.84	2.16
Female	Extra	3.46	3.74	2.96	4.00	3.70	4.00	4.00	3.30	3.30	3.68
Female	Extra	2.77	2.16	2.12	2.75	2.45	3.70	3.70	3.53	2.11	2.91
Male	Extra	3.00	3.00	2.14	2.54	3.47	3.83	3.90	3.85	3.45	3.74

การแทนค่าข้อมูลในการทดลอง ข้อมูลที่จะทำการวิเคราะห์ทดลองนั้น เพื่อให้คอมพิวเตอร์ประมวลผลข้อมูลและเข้าใจ รวดเร็ว และถูกต้องแม่นยำนั้น ข้อมูลต้องอยู่ในรูปแบบที่คอมพิวเตอร์เข้าใจความหมายของข้อมูลนั้นเสียก่อน ผู้วิจัยได้มีการแทนค่าให้กับข้อมูลในส่วนข้อมูลผลการเรียนเฉลี่ยสะสม (GPAX) ผลการเรียนกลุ่มสาระการเรียนรู้ และผลการเรียนที่สำเร็จการศึกษามหาวิทยาลัยมหาสารคาม (GPA) ซึ่งข้อมูลผลการเรียนที่ใช้ในการทดลองเป็นข้อมูลแบบทศนิยมต่อเนื่อง เพื่อลดการกระจายของข้อมูล (Binning Data) ผู้วิจัยได้แบ่งช่วงของข้อมูลผลการเรียนเฉลี่ยสะสม และผลการ

เรียนกลุ่มสาระการเรียนรู้ออกเป็น 3 ช่วง คือ Low, Medium และ High ซึ่งช่วงผลการเรียน 0.00 - 2.50 เป็นระดับต่ำ แทนค่า L ช่วงผลการเรียน 2.51 - 3.25 เป็นระดับกลาง แทนค่า M และช่วงผลการเรียน 3.26 - 4.00 เป็นระดับสูง แทนค่า H และมีการแบ่งช่วงข้อมูลผลการเรียนที่สำเร็จการศึกษาจากมหาวิทยาลัยมหาสารคาม(GPA)มีการแบ่งช่วงข้อมูลออกเป็น 3 ช่วง คือ ช่วงผลการเรียน 2.00 - 2.74 เป็นระดับพอใช้ แทนค่า Fair ช่วงผลการเรียน 2.75 - 3.24 เป็นระดับดี แทนค่า Good และช่วงผลการเรียน 3.25 - 4.00 เป็นระดับดีเยี่ยม แทนค่า Excellent รายละเอียดตัวอย่างการแทนค่าข้อมูลดัง Table 4

Table 4 Substitution information

Sex	Size	GPAX	GPA1	GPA2	GPA3	GPA4	GPA5	GPA6	GPA7	GPA8	GPA
Male	S1	M	H	L	M	H	H	H	H	H	Excellent
Female	S2	L	L	L	L	H	M	H	H	L	Fair
Female	S2	M	M	L	M	L	H	H	H	L	Good
Female	S4	L	L	L	M	H	M	M	L	L	Excellent
Male	S1	H	M	L	H	M	H	H	H	M	Good
Female	S3	L	L	L	L	H	M	M	M	L	Fair
Female	S1	H	H	M	H	H	H	H	H	H	Excellent
Male	S4	H	M	L	H	M	M	M	H	M	Good

ผลการศึกษาวิจัย

จากการทดลองการสร้างแบบจำลองโดยใช้เทคนิคเหมืองข้อมูลในการเลือกกลุ่มสาขาวิชาที่เหมาะสมในระดับปริญญาตรีโดยได้แบ่งกลุ่มข้อมูลการทดลองออกเป็นสามกลุ่มตามประกันคุณภาพของมหาวิทยาลัย คือ กลุ่มมนุษยศาสตร์และสังคมศาสตร์ กลุ่มวิทยาศาสตร์เทคโนโลยี และกลุ่มวิทยาศาสตร์สุขภาพ และได้ใช้อัลกอริทึมในการทดลอง ได้แก่

ต้นไม้ตัดสินใจ (Decision Tree C4.5) การเรียนรู้แบบเบย์อย่างง่าย (Naïve Bayes) และซัพพอร์ตเวกเตอร์แมชชีน (SVM) ผลการทดลองได้ผลลัพธ์ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความเหวี่ยง (F-Measure) ผลการทดลองของแต่ละอัลกอริทึม รายละเอียดดัง Table 5

Table 5 Results

Group Quality Assurance	Algorithm	Accuracy (%)	Precision	Recall	F-Measure
Humanities and Social Sciences	Decision Tree	82.22	0.821	0.822	0.821
	Naïve Bayes	78.09	0.785	0.781	0.782
	SVM	80.78	0.801	0.808	0.801
Science and Technology	Decision Tree	88.26	0.892	0.883	0.881
	Naïve Bayes	84.31	0.845	0.843	0.843
	SVM	87.41	0.880	0.874	0.873
Health Sciences	Decision Tree	87.73	0.881	0.877	0.877
	Naïve Bayes	78.75	0.783	0.788	0.783
	SVM	85.87	0.857	0.859	0.858

จาก Table 5 ตารางสรุปผลการทดลองจากกลุ่มข้อมูลที่ใช้ในการทดลองทั้งสามกลุ่มข้อมูล โดยใช้วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีการเรียนรู้แบบเบย์อย่างง่าย (Naïve Bayes) และวิธีซัพพอร์ตเวกเตอร์แมชชีน (SVM) สามารถอธิบายได้ว่าในทุกกลุ่มข้อมูล การทดลองด้วยวิธีต้นไม้ตัดสินใจมีค่าความถูกต้อง (Accuracy) สูงสุดโดยกลุ่มมนุษยศาสตร์และสังคมศาสตร์มีค่าความถูกต้องร้อยละ 82.22 ค่าความแม่นยำ (Precision) 0.821 และค่าการค้นคืนข้อมูล (Recall) 0.822 กลุ่มวิทยาศาสตร์เทคโนโลยี มีค่าความถูกต้องร้อยละ 88.26 ค่าความแม่นยำ (Precision) 0.892 และค่าการค้นคืนข้อมูล (Recall) 0.883 และกลุ่มวิทยาศาสตร์สุขภาพ มีค่าความถูกต้องร้อยละ 87.73 มีค่าความแม่นยำ (Precision) 0.881 และค่าการค้นคืนข้อมูล (Recall) 0.877 แสดงให้เห็นว่าค่าข้อมูลการค้นคืนข้อมูลในการพยากรณ์การเลือกกลุ่มสาขามีค่าความแม่นยำสูง

ผลการทดลองสามารถแสดงผลการทดลองให้อยู่ในรูปแบบของต้นไม้ตัดสินใจ (Decision Rule) ในการเลือกกลุ่มสาขาวิชาของผู้สมัครทั้งสามกลุ่มข้อมูลที่ใช้ในการทดลองได้ Figure 3 ตัวอย่างในบางกิ่งต้นไม้ของข้อมูลกลุ่มมนุษยศาสตร์และสังคมศาสตร์

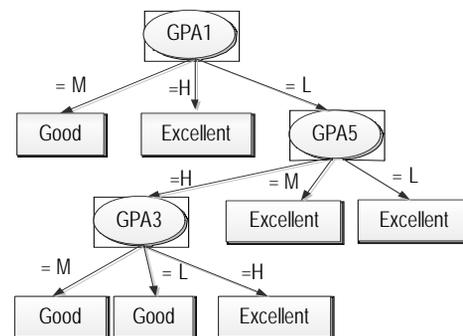


Figure 3 Decision tree

จาก Figure 3 ตัวอย่างในบางกิ่งต้นไม้สามารถสร้างเป็นกฎการตัดสินใจ (Decision Rule) ดังนี้

- IF GPA1='Medium(M)' Then GPA=Good
- IF GPA1='High(H)' Then GPA=Excellent
- IF GPA1='Low(L)' AND GPA5='Medium(M)' OR 'Low(L)' Then GPA=Excellent
- IF GPA1='Low(L)' AND GPA5='High(H)' AND GPA3='Medium(M)' OR 'Low(L)' Then GPA=Good

IF GPA1='Low(L)' And GPA5='High(H)'
AND GPA3='High(H)' Then GPA=Excellent

จากตัวอย่างกฎสามารถอธิบายได้ดังนี้
กฎข้อที่ 1 ผู้สมัครที่จะเลือกเรียนในกลุ่มมนุษยศาสตร์
และสังคมศาสตร์ มีผลการเรียนกลุ่มสาระวิชาภาษาไทย
(GPA1) ระดับกลาง (Medium) ระหว่าง (2.51 - 3.25) เมื่อ
ศึกษาในกลุ่มวิชานี้จะมีโอกาสสำเร็จการศึกษาที่ได้ผลการ
เรียนเฉลี่ยรวม (GPA)ระดับดี (Good) ระหว่าง 2.75 - 3.24

กฎข้อที่ 2 ผู้สมัครที่จะเลือกเรียนในกลุ่มมนุษยศาสตร์
และสังคมศาสตร์ มีผลการเรียนกลุ่มสาระวิชาภาษาไทย
(GPA1) ระดับสูง (High) ระหว่าง (3.26 - 4.00)เมื่อศึกษาใน
กลุ่มวิชานี้จะมีโอกาสสำเร็จการศึกษาที่ได้ผลการเรียนเฉลี่ย
รวม (GPA)ระดับดีเยี่ยม (Excellent) ระหว่าง 3.25 - 4.00

กฎข้อที่ 3 ผู้สมัครที่จะเลือกเรียนในกลุ่มมนุษยศาสตร์
และสังคมศาสตร์ มีผลการเรียนกลุ่มสาระวิชาภาษาไทย
(GPA1) ระดับต่ำ (Low) ระหว่าง (0.00 - 2.50) และมีผลการ
เรียนกลุ่มสาระวิชาสุขศึกษา(GPA5) ระดับกลาง (Medium)
ระหว่าง (2.51 - 3.25) หรือระดับต่ำ (Low)ระหว่าง (0.00 -
2.50) เมื่อศึกษาในกลุ่มวิชานี้จะมีโอกาสสำเร็จการศึกษาที่ได้
ผลการเรียนเฉลี่ยรวม (GPA)ระดับดีเยี่ยม (Excellent) ระหว่าง
3.25 - 4.00

กฎข้อที่ 4 ผู้สมัครที่จะเลือกเรียนในกลุ่มมนุษยศาสตร์
และสังคมศาสตร์ มีผลการเรียนกลุ่มสาระวิชาภาษาไทย
(GPA1) ระดับต่ำ (Low) ระหว่าง (0.00 - 2.50) และมีผลการ
เรียนกลุ่มสาระวิชาสุขศึกษา (GPA5) ระดับสูง (High)ระหว่าง
(3.26 - 4.00)และผลการเรียนกลุ่มสาระวิชาสังคมศึกษา
(GPA3)ระดับกลาง (Medium)ระหว่าง(2.51 - 3.25)หรือ
ระดับต่ำ (Low) ระหว่าง (0.00 - 2.50) เมื่อศึกษาในกลุ่มวิชานี้
จะมีโอกาสสำเร็จการศึกษาที่ได้ผลการเรียนเฉลี่ยรวม (GPA)
ระดับดี (Good) ระหว่าง 2.75 - 3.24

กฎข้อที่ 5 ผู้สมัครที่จะเลือกเรียนในกลุ่มมนุษยศาสตร์
และสังคมศาสตร์ มีผลการเรียนกลุ่มสาระวิชาภาษาไทย
(GPA1) ระดับต่ำ (Low) ระหว่าง (0.00 - 2.50) และมีผล
การเรียนกลุ่มสาระวิชาสุขศึกษา (GPA5) ระดับสูง (High)
ระหว่าง (3.26 - 4.00) และผลการเรียนกลุ่มสาระวิชา
สังคมศึกษา(GPA3) ระดับสูง (High) ระหว่าง (3.26 - 4.00)
เมื่อศึกษาในกลุ่มวิชานี้จะมีโอกาสสำเร็จการศึกษาที่ได้ผล
การเรียนเฉลี่ยรวม (GPA) ระดับดีเยี่ยม (Excellent) ระหว่าง
2.75 - 3.24

จากกฎที่ได้ทั้งสามกลุ่มตามกลุ่มข้อมูลประกัน
คุณภาพการศึกษาของมหาวิทยาลัยมหาสารคามสามารถพัฒนา

เป็นระบบพยากรณ์การเลือกกลุ่มสาขาวิชาที่เหมาะสม
Figure 4



Figure 4 Major area selected forecasting

อภิปรายผลการวิจัย

บทความนี้ได้นำเสนอแบบจำลองการเลือกกลุ่มสาขาวิชา
ที่เหมาะสมในระดับปริญญาตรี ซึ่งแบ่งกลุ่มข้อมูลตาม
กลุ่มประกันคุณภาพของมหาวิทยาลัย โดยเปรียบเทียบ
ประสิทธิภาพการจำแนกข้อมูลของอัลกอริทึมที่ใช้ใน
การทดลองทั้งสามวิธี พบว่าวิธีการแบบต้นไม้ตัดสินใจมี
ประสิทธิภาพสูงสุด ผลลัพธ์ที่ได้ในแต่ละกลุ่มข้อมูล ดังนี้
กลุ่มมนุษยศาสตร์และสังคมศาสตร์ ค่าความถูกต้องที่ได้
ร้อยละ 82.22กลุ่มวิทยาศาสตร์เทคโนโลยี ร้อยละ 88.26
กลุ่มวิทยาศาสตร์สุขภาพ ร้อยละ 87.73 แสดงให้เห็นว่า
ประสิทธิภาพการจำแนกข้อมูลโดยใช้อัลกอริทึมแบบต้นไม้
ตัดสินใจมีประสิทธิภาพสูงสามารถนำไปประยุกต์ในการสร้าง
แบบจำลองในการเลือกกลุ่มสาขาวิชาที่เหมาะสมในระดับ
ปริญญาตรี ในโอกาสต่อไป

ข้อเสนอแนะ

จากงานวิจัยนี้ได้ทำการทดลองกับกลุ่มข้อมูลประกันคุณภาพ
การศึกษาของมหาวิทยาลัยมหาสารคาม ซึ่งไม่ได้ทดลองกับ
ข้อมูลในรายสาขาวิชา อันเนื่องมาจากปัจจัยเรื่องจำนวนข้อมูล
ในบางรายสาขาวิชาที่มีไม่เพียงพอต่อการทดลอง ผู้วิจัยจึงได้
เลือกข้อมูลตามกลุ่มประกันคุณภาพเพื่อให้ได้จำนวนข้อมูล
เพียงพอต่อการทดลอง ในอนาคตผู้วิจัยมีแนวคิดที่จะเลือก
ข้อมูลในบางสาขาวิชา เช่น สาขาวิชาที่มีจำนวนเพียงพอใน
การทดลอง สาขาวิชาที่มีจำนวนน้อยสภาพการเป็นนิสิต
หรือสาขาวิชาที่มีจำนวนนิสิตขยับสาขาเป็นจำนวนมาก มา
ทำการทดลองเพื่อให้สามารถทำนายผลการเลือกสาขาวิชาได้
มากยิ่งขึ้น

กิตติกรรมประกาศ

โครงการวิจัยนี้ได้รับการสนับสนุนจากเงินอุดหนุนการวิจัยงบประมาณเงินรายได้ ประจำปีงบประมาณ 2557 มหาวิทยาลัยมหาสารคาม

เอกสารอ้างอิง

1. กองบริการการศึกษา มหาวิทยาลัยมหาสารคาม. ระเบียบการรับสมัครคัดเลือกบุคคลเข้าศึกษาในระดับปริญญาตรี ระบบรับตรง ปี 2549-2554.
2. กองทะเบียนและประมวลผล มหาวิทยาลัยมหาสารคาม. รายงานประจำปี 2552-2555.
3. ศูนย์พัฒนาและประกันคุณภาพการศึกษา มหาวิทยาลัยมหาสารคาม. รายงานการประเมินคุณภาพภายนอกรอบสาม สำนักงานรับรองมาตรฐานและประเมินคุณภาพการศึกษา ปี 2554-2558.
4. ฉัตรเกล้า เจริญผล. เอกสารประกอบการสอนรายวิชา Introduction to Data Mining; 2013.
5. Mark H, Ian W, Eibe F, Data Mining Practical Machine Learning Tools and Techniques; 2011.
6. บุญเสริม กิจศิริกุล. รายงานวิจัยฉบับสมบูรณ์โครงการย่อยที่ 7 อัลกอริทึมการทำเหมืองข้อมูล; ปี 2545.
7. Han J, Kamber M, Data Mining Concepts and Techniques; 2003.
8. กฤษณะ ไวยมัย, ชิตชนก ส่งศิริ, ธนาวิทย์ รักธรรมานนท์, การใช้เทคนิคดาต้าไมนิ่งเพื่อพัฒนาคุณภาพการศึกษา คณะวิศวกรรมศาสตร์; NECTEC Technical Journal, Vol. 3, No. 11; 2001 July-October 2001.
9. ชิตชนก ส่งศิริ, ธนาวิทย์ รักธรรมานนท์, กฤษณะ ไวยมัย. การใช้เทคนิค Data Mining เพื่อค้นหาภาควิชาที่เหมาะสมที่สุดให้กับนิสิต. การประชุมวิชาการของมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 39; ปี 2544.
10. Pathom Pumpuang. Comparisons of Classifier Algorithm : Bayesian Network, C4.5, Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students. 2008 IEEE International Conference on Systems, Man and Cybernetics (SMC 2008), 2008.
11. Tomas B, Marcel J, Pavel K, Selecting Representative Data Sets; INTECH 2012.
12. Zdravko M, Ingrid R, An Introduction to the WEKA Data Mining System.