การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวินท์ รักธรรมานนท์
อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
และนิสิตปริญญาโทวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

ABSTRACT- The aim of our work is to contribute to an improved quality of engineering graduates by proposing a data mining system that assists students in selecting the appropriate major according to their profile and in their course registering.

KEY WORDS – data mining, knowledge

บทคัดย่อ - บทความนี้เป็นการศึกษาและวิเคราะห์ระบบฐานข้อมูลนิสิต โดยนำความรู้ทางด้านดาต้าไมน์นิงมาประยุกต์ใช้กับข้อมูลนิสิตคณะวิศวกรรมศาสตร์ เพื่อเป็นแนวทางในการแก้ไขปัญหาต่างๆ อาทิเช่น ปัญหาการเลือกสาขาวิชาไม่ตรงกับความสามารถที่แท้จริง ปัญหาผลการเรียนของนิสิตตกต่ำจนต้องออกจากสถาบันการศึกษา อันเป็นผลทำให้ไม่ได้มาซึ่งบุคลากรที่มีความสามารถสูงสุด

คำสำคัญ – ดาต้าไมน์นิง, ความรู้

1. บทนำ

การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: KDD) หรือที่เรียกกันว่า ดาต้าไมน์นิง (Data Mining) เป็นสาขาหนึ่งในวิทยาศาสตร์คอมพิวเตอร์ที่กำลังได้รับความสนใจอย่างสูงในปัจจุบัน [2, 5, 8] เมื่อใช้เทคนิคดาต้าไมน์นิง ข้อมูลขนาดใหญ่จะถูกวิเคราะห์และสืบค้นความรู้หรือสิ่งที่สำคัญออกมา จากนั้นจะรวบรวมความรู้ที่ได้ให้อยู่ในรูปฐานความรู้ (Knowledge Base) เพื่อนำไปใช้ประโยชน์ต่อไป โดยในปัจจุบันได้มีการนำเทคนิคดาต้าไมน์นิงไปประยุกต์ใช้ในงานด้านต่างๆ มากขึ้น ทั้งในด้านการส่งเสริมการขายสินค้าในห้างสรรพสินค้า, ด้านการวิเคราะห์เครดิตลูกค้าในธนาคาร และในด้านอื่นๆ อีกมาก แต่ไม่มีการนำมาประยุกต์กับด้านการศึกษาอย่างจริงจัง ทั้งที่ในปัจจุบันตามสถาบันการศึกษาส่วนใหญ่มีข้อมูลนิสิตที่ได้จัดเก็บไว้เป็นเวลานาน แต่มิได้ถูกนำมาใช้ให้เกิดประโยชน์เท่าที่ควร โดยในบทความนี้ได้นำเสนอเทคนิคต่างๆ ที่สำคัญของดาต้าไมน์นิงมาประยุกต์ใช้ในการสืบค้นสิ่งที่น่าสนใจออกมาจากข้อมูลนิสิต

ข้อมูลนิสิตที่นำมาวิจัยนี้เป็นข้อมูลของนิสิตคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ โดยประกอบไปด้วยข้อมูล 2 ส่วน ส่วนแรกคือ ฐานข้อมูลการลงทะเบียนเรียนของนิสิต ที่แต่และแถวแสดงถึงวิชาที่นิสิตได้ลงทะเบียนเรียนและผลการเรียนในวิชาต่างๆ และส่วนที่ 2 คือ ฐานข้อมูลประวัติส่วนตัวของนิสิต เช่น อายุ เพศ ที่อยู่ ประวัติการศึกษาก่อนเข้ามาในมหาวิทยาลัย เกรดเฉลี่ยสะสม เป็นต้น

ในบทความนี้ เริ่มต้นด้วยบทนำที่กล่าวถึงความสำคัญและแนวทางในการพัฒนาคุณภาพการศึกษาโดยใช้เทคนิคดาตาไมน์นิงในส่วนแรก ในส่วนที่ 2 กล่าวถึงรายละเอียดของแนวคิดและวิธีการที่นำมาประยุกต์ใช้ในงานนี้ ได้แก่ การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ ตามด้วยเทคนิคดาต้าไมน์นิงที่สำคัญ 3 ประการคือ การค้นหากฎความสัมพันธ์ (association rule discovery) , การจำแนกข้อมูล (data classification) และ การพยากรณ์ข้อมูล (data prediction) ในส่วนที่ 3 กล่าวถึงการนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้ในการช่วยในการช่วยนิสิตเลือกสาขาวิชาที่เหมาะสม ส่วนที่ 4 กล่าวถึงการนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้ในการทำนายเกรดแต่ละรายวิชาในภาคการศึกษาต่อไป ส่วนที่ 5 เป็นการสรุปงานวิจัยนี้และกล่าวถึงการปรับปรุงงานต่อไปในอนาคต

2. แนวคิดและวิธีการเบื้องต้น

การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: KDD) หรือที่เรียกกันว่าดาต้าไมน์นิงเป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่ โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้ หรือสิ่งที่สำคัญออกมา เพื่อใช้ในการวิเคราะห์ หรือทำนายสิ่งต่างๆ ที่จะเกิดขึ้น กระบวนการ KDD ประกอบไปด้วย 3 ส่วนหลักๆ คือ

1) Pre-processing คือ ขั้นตอนการเตรียมข้อมูลให้เหมาะสมและให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้

2) Data Mining คือ ขั้นตอนการเลือกเทคนิคที่เหมาะสมสำหรับงานที่ต้องการ โดยสามารถรวมเทคนิคได้มากกว่าหนึ่งเทคนิคมาประมวลผลเพื่อดึงความรู้หรือสิ่งที่น่าสนใจจากข้อมูลที่ผ่านขั้นตอน Pre-precessing แล้ว โดยผลลัพธ์ที่ได้จากขั้นตอนนี้คือฐานความรู้

3) Post-processing คือขั้นตอนการนำฐานความรู้ที่ได้จากขั้นตอนดาต้าไมน์นิงมาทดสอบและพิจารณาว่าถูกต้องตามความต้องการหรือไม่ ซึ่งบางครั้งอาจต้องปรับแก้ค่าและนำเข้าสู่ขั้นตอนดาต้าไมน์นิงใหม่อีกครั้ง จนกว่าจะได้ความรู้หรือสิ่งที่น่าสนใจตามที่ต้องการออกมา

จะเห็นได้ว่าดาต้าไมน์นิงเป็นองค์ประกอบหลักที่สำคัญในกระบวนการ KDD ในลำดับต่อไป เราจะกล่าวถึงเทคนิคดาต้าไมน์นิงที่สำคัญ 3 เทคนิค ที่ได้นำมาประยุกต์ใช้ คือ การค้นหากฎความสัมพันธ์ (Association rule discovery) , การจำแนกประเภทข้อมูล (data classification) และ การพยากรณ์ข้อมูล (data prediction)

2.1 การค้นหากฎความสัมพันธ์ (Association rule discovery)

การค้นหากฎความสัมพันธ์ [7] เป็นการค้นหาความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่างๆ โดยเทคนิคนี้ใช้กันอย่างแพร่หลายในการขายสินค้า หรือการวิเคราะห์ข้อมูลที่เป็นทรานแซกชัน เราขอยกตัวอย่างการนำเทคนิคการค้นหากฎความสัมพันธ์มาประยุกต์ใช้ในข้อมูลการขายสินค้าดังตารางที่ 1

ตารางที่ 1 แสดงตัวอย่างข้อมูลการขายสินค้า

หมายเลขทรานแซกชัน	สินค้าที่ซื้อ
1	น้ำตาล, ขนมปัง
2	นมสด, น้ำตาล, ขนมปัง
3	ขนมปัง
4	นมสด, น้ำตาล

จากตารางที่ 1 สามารถบอกได้ว่าน้ำตาลและขนมปังจะถูกซื้อด้วยกันในทรานแซกชันที่ 1 หลังจากที่นำข้อมูลไปผ่านกระบวนการดาต้าไมน์นิงแล้ว จะได้ความสัมพันธ์อยู่ในรูป X -> Y หมายความว่า เมื่อซื้อ X แล้วจะซื้อY ด้วย ยกตัวอย่างเช่น นมสด -> น้ำตาล หมายความว่าเมื่อลูกค้าซื้อนมสดแล้วจะซื้อน้ำตาลด้วย

กฎหรือความสัมพันธ์ต่างๆ ที่ได้มาจากกระบวนการดาต้าไมน์นิงนั้นจะมีความน่าสนใจต่างกัน จึงต้องมีเกณฑ์ในการวัดความน่าสนใจของกฎ การค้นหากฎความสัมพันธ์นี้ มีเกณฑ์ในการวัดความน่าสนใจ 2 แบบ ได้แก่ ค่าสนับสนุน (support) และค่าความมั่นใจ (confidence)

- ค่าสนับสนุน (support) แสดงถึงเปอร์เซ็นต์ของข้อมูลที่เป็นไปตามกฎจากข้อมูลทั้งหมด จากกฎ นมสด -> น้ำตาล เมื่อพิจารณาจากข้อมูลในตารางที่ 1 จะเห็นว่าข้อมูลในทรานแซกชันที่ 2 และ 4 ที่เป็นไปตามกฎ จากข้อมูลทั้งหมด 4 ทรานแซกชัน ดังนั้น จะได้ว่าค่าสนับสนุนจากกฎนี้คือ 2/4 คือ 50 %
- ค่าความมั่นใจ (confidence) แสดงถึงความน่าเชื่อถือของกฎ จากกฎ นมสด -> น้ำตาล จะพบว่า มีทรานแซกชันที่ซื้อนมสด 2 ทรานแซกชัน และมีทรานแซกชันที่ซื้อทั้งนมสดและน้ำตาลพร้อมกัน 2 ทรานแซกชัน ดังนั้น จากกฎนี้ มีค่าความมั่นใจ เท่ากับ 100 %

เราจะพิจารณาเฉพาะความสัมพันธ์ที่มีค่าสนับสนุนและค่าความมั่นใจสูงกว่าค่าสนับสนุนต่ำสุด (minimum support) และ ค่าความมั่นใจต่ำสุด (minimum confidence) ตามลำดับ จากตารางที่ 1 สมมติกำหนดค่าสนับสนุนต่ำสุดเท่ากับ 50 % และความมั่นใจต่ำสุดเท่ากับ 70 % จะได้ว่ามีเพียงกฎ นมสด -> น้ำตาลเท่านั้นที่ผ่านเกณฑ์ตามที่กำหนดไว้

2.2 การจำแนกประเภทข้อมูล (Data Classification)

การจำแนกประเภทข้อมูล [3] เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ โดยจะนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ (training data) เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ ผลลัพธ์ที่ได้จากการเรียนรู้คือ โมเดลจำแนกประเภทข้อมูล (classifier model) และจะนำข้อมูลส่วนที่เหลือจากข้อมูลสอนระบบเป็นข้อมูลที่ใช้ทดสอบ (testing data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้องและปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้น เมื่อมีข้อมูลใหม่เข้ามา เราจะนำข้อมูลมาผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลนี้ได้

ต้นไม้ช่วยการตัดสินใจ (Decision tree) [4] เป็นวิธีหนึ่งที่สำคัญในการจำแนกประเภทข้อมูล โดยต้นไม้ช่วยการตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ที่แต่ละโหนดแสดงคุณลักษณะ (attribute), แต่ละกิ่งแสดงผลในการทดสอบ และลีฟโหนด (leaf node) แสดงกลุ่มที่กำหนดไว้ ซึ่งต้นไม้ช่วยการตัดสินใจนี้ง่ายต่อการปรับเปลี่ยนเป็นกฎการจำแนกประเภทข้อมูล (classification rule)

2.3 การพยากรณ์ข้อมูล (Data Prediction)

การพยากรณ์ข้อมูล [3] เป็นกระบวนการสร้างโมเดลเพื่อทำนายหาค่าที่ต้องการจากข้อมูลที่มีอยู่ โดยมีกระบวนการสร้างโมเดลคล้ายกับการจำแนกประเภทข้อมูลดังที่ได้กล่าวมาข้างต้น ต่างกันตรงที่การพยากรณ์ข้อมูลไม่มีการจัดข้อมูลเข้ากลุ่มตามที่ได้กำหนด แต่การพยากรณ์ข้อมูลนี้เป็นการพยากรณ์หาค่าที่ต้องการออกมาเป็นตัวเลข ตัวอย่างเช่น หายอด

ขายของเดือนถัดไปจากข้อมูลการขายทั้งหมดที่ผ่านมา หรือทำนายเกรดเฉลี่ยของนักเรียนในปีการศึกษาหน้า จากข้อมูลการลงทะเบียนของนิสิตทั้งหมด เป็นต้น

3. การนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้ในการช่วยนิสิตเลือกสาขาวิชาที่เหมาะสม

การเลือกสาขาวิชาที่เหมาะสมของนิสิตนั้น เนื่องด้วยนิสิตยังขาดประสบการณ์ และไม่รู้จักแต่ละสาขาวิชามากพอ นิสิตส่วนใหญ่จึงใช้ความรู้สึกความชอบ หรือสภาพแวดล้อมทั้งเพื่อน หรือผู้ปกครองเป็นหลักใหญ่ โดยอาจไม่ทราบถึงสาขาวิชาที่เหมาะสมกับความสามารถและลักษณะเฉพาะของตัวเอง จึงอาจทำให้เมื่อเข้าไปเรียนจริงๆ แล้วเพิ่งค้นพบตัวเองในภายหลังว่าไม่เหมาะสมกับสาขาวิชานี้ จนอาจทำให้เกิดผลกระทบต่างๆ ตามมา โครงงานนี้ได้นำเทคนิคใหม่ทางวิทยาการคอมพิวเตอร์ คือ เทคนิคดาต้าไมน์นิงมาประยุกต์ใช้กับข้อมูลนิสิต เพื่อช่วยในการชี้แนะแนวทางการเลือกสาขาวิชาที่เหมาะสมกับนิสิตแต่ละคนให้มากที่สุด

ในการที่จะบรรลุให้ได้ตามวัตถุประสงค์ คือ การใช้เทคนิคดาต้าไมน์นิงเพื่อเลือกสาขาวิชาที่เหมาะสมให้กับนิสิตนั้น สามารถทำได้หลายแนวทาง ในบทความนี้ขอเสนอ 2 แนวทางที่มีผลการทดสอบความถูกต้องค่อนข้างสูง คือ การทำเทคนิคการจำแนกประเภทข้อมูล และเทคนิคการพยากรณ์ข้อมูลมาสร้างโมเดล

การสร้างโมเดลสำหรับการเลือกสาขาวิชานั้นสามารถทำได้หลายแนวทาง แนวทางหนึ่งคือการนำข้อมูลของนิสิตที่เรียนดีในทุกวิชามา สร้างโมเดลกลางการจำแนกประเภทข้อมูล [6] โดยแต่ละโหนดภายในต้นไม้ช่วยการตัดสินใจ (decision tree) บ่งบอกถึงลักษณะและผลการเรียนในรายวิชาต่างๆ ของนิสิตและคลาสปลายทางแทนสาขาวิชาต่างๆ เพื่อที่จะทำนายว่าลักษณะของนิสิตแต่ละคนนั้นคล้ายคลึงกับลักษณะของนิสิตที่เรียนดีในสาขาวิชาใดมากที่สุด วิธีดังกล่าวมีข้อดีคือสามารถทำนายสาขาวิชาที่เหมาะสมที่สุดให้กับนิสิตได้ แต่วิธีการดังกล่าวนั้นมีข้อเสียบางประการ เช่น โมเดลจะทำนายแนวโน้มโอนเอียงไปทางสาขาวิชาที่มีจำนวนนิสิตมากเป็นผลทำให้ความถูกต้องของโมเดลที่ได้ค่อนข้างต่ำ คือประมาณ 50 % ดังนั้นเราจึงได้คิดวิธีการสร้างโมเดลแบบอื่นเพื่อปรับปรุงประสิทธิภาพให้มากขึ้นกว่าเดิม

ในลำดับต่อมา เราได้สร้างโมเดลจำแนกประเภทข้อมูลสำหรับแต่ละสาขาวิชาโดยพิจารณาว่านิสิตเหมาะสมกับสาขาวิชานั้นๆ หรือไม่ และโมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชาเพิ่มเติมเพื่อลดข้อผิดพลาดที่เกิดขึ้นจากโมเดลกลางการจำแนกประเภทข้อมูล ในส่วนที่ 3.1 กล่าวถึงโมเดลจำแนกประเภทข้อมูลสำหรับแต่ละสาขาวิชาโดยพิจารณาว่านิสิตเหมาะสมกับสาขาวิชานั้นๆ หรือไม่โดยการใช้เทคนิคการจำแนกประเภทข้อมูล (Classification) และส่วนที่ 3.2 จะนำเสนอโมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชาโดยการใช้เทคนิคการพยากรณ์ข้อมูล (Prediction) การนำเสนอในแต่ละส่วนจะกล่าวถึงการเตรียมข้อมูล การสร้างโมเดล การแปลความหมายจากโมเดล การนำความรู้ที่ได้มาใช้งานจริง และการเปรียบเทียบประสิทธิภาพของแต่ละโมเดล ในรายละเอียดต่อไป

3.1 โมเดลการจำแนกประเภทข้อมูลสำหรับแต่ละสาขาวิชาโดยพิจารณาว่านิสิตเหมาะสมกับสาขาวิชานั้นๆ หรือไม่ (Classification data model for each major)

3.1.1 การเตรียมข้อมูล

ในงานวิจัยนี้ใช้ข้อมูลการลงทะเบียนเรียนของนิสิตคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ตั้งแต่ปี 2535-2542 รวมทั้งสิ้น 476,085 แถว จากนิสิตกว่า 10,000 คน โดยข้อมูลชุดนี้ ประกอบไปด้วยข้อมูล 2 ส่วน ส่วนแรกคือ ข้อมูลประวัติส่วนตัวของนิสิต เช่น ชื่อ ที่อยู่ ภูมิลำเนา อายุ ฯลฯ ดังตารางที่ 2

ตารางที่ 2 แสดงตัวอย่างข้อมูลประวัติส่วนตัวนิสิต

Stu_code	Sex	Address	SchoolGPA	…	GPA
37058063	male	Bangkok	2.5	….	2.3
37058167	male	Songkla	3.4	….	3.2
…….	….	…….	……	….	….

ข้อมูลอีกส่วนหนึ่งคือ ข้อมูลการลงทะเบียนของนิสิต เนื่องมาจาก คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์นี้จะมีการเลือกสาขาวิชาขึ้นในปี 2 ดังนั้น ในการวิจัยนี้จึงนำเฉพาะวิชาในปีที่ 1 ของการศึกษา (9 วิชา) มาเป็นตัวพิจารณา ในแต่ละแถวประกอบไปด้วยวิชาและผลการเรียนของนิสิตดังตารางที่ 3

ตารางที่ 3 แสดงตัวอย่างข้อมูลการลงทะเบียนเรียนของนิสิต

Stu_code	Sub_code	Section	Term	Year	Grade
37058063	204111	2	1	2537	C+
37058063	403111	6	1	2537	D
37058063	208111	1	1	2537	B+
……	……..	……	…….	……	…

จากตารางที่ 3 ข้อมูลอยู่ในระดับรายวิชาจำเป็นต้องแปลงข้อมูลให้อยู่ในระดับของนิสิต เพื่อให้ได้ลักษณะโครงสร้างข้อมูลตรงตามเป้าหมายที่ต้องการคือศึกษาพฤติกรรมและลักษณะของนิสิตแต่ละคน โดยใช้วิธีแบ่งกลุ่มของวิชาต่างๆ ที่ลงทะเบียนตามรหัสนิสิต จากนั้น จึงแปลงให้แต่ละแถวแทนนิสิตแต่ละคน และคอลัมน์แทนรายชื่อวิชาต่างๆ นอก

จากนี้เพื่อต้องการลดการกระจายของข้อมูลเกรดของนิสิต จึงจัดกลุ่มเกรดของนิสิตเป็น 3 กลุ่ม ตามวิธีที่ได้เคยกล่าวมาในส่วน 3.1 แล้ว ผลของตารางแสดงได้ดังในตารางที่ 4

ตารางที่ 4 แสดงตัวอย่างข้อมูลที่จัดในระดับนิสิต

Stu_code	Sex	204111	403111	…	GPA
37058063	male	Medium	Low	…	2.3
37058167	male	High	High	…	3.2
……	….	…..	…..	…	…..

ในการสร้างแต่ละโมเดล แบ่งข้อมูล 70 % จากข้อมูลทั้งหมดเป็นข้อมูลสอนระบบ (training data) เพื่อสร้างโมเดล และข้อมูล 30 % ที่เหลือเป็นข้อมูลที่ใช้ทดสอบ (testing data) เพื่อทดสอบความถูกต้องของโมเดล

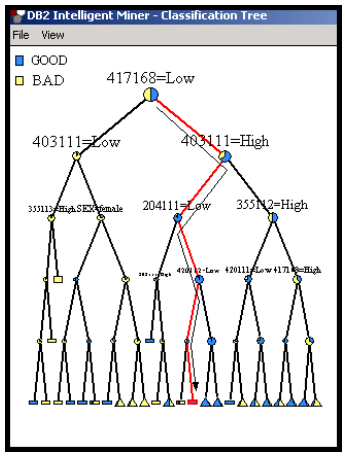
3.1.2 การสร้างโมเดล

ในโมเดลการจำแนกประเภทข้อมูลนี้ แต่ละกิ่งแทนลักษณะต่างๆ ของนิสิตดังที่ได้กล่าวมา และลีฟโหนด (คลาสปลายทาง) มี 2 คลาส คือ GOOD และ BAD โดยมีเกณฑ์ว่า GOOD คือนิสิตที่มีเกรดเฉลี่ยสะสมอยู่ในอันดับ 40% แรกของแต่ละสาขาวิชา และ BAD คือ นิสิตที่มีเกรดเฉลี่ยสะสมอยู่ในอันดับ 40% สุดท้ายในแต่ละสาขาวิชา โดยเปอร์เซ็นต์นี้สามารถปรับเปลี่ยนให้เหมาะสมได้ตามลักษณะของข้อมูล ในการวิจัยนี้พบว่า ที่ 40% เป็นอัตราส่วนที่เหมาะสมที่สุด เพราะจะได้ข้อมูลที่นำมาวิจัยมาก และมีการเว้นช่วงผิดพลาดที่สามารถเกิดได้ในช่วงกลางของนิสิต (ช่วงระหว่างเกรด GOOD และ BAD) ดังนั้น ในการวิจัยนี้ข้อมูลสอนระบบคือข้อมูลนิสิตที่มีลักษณะทั้ง GOOD และ BAD

จากวิธีการที่ได้นำเสนอข้างต้น ต้องสร้างโมเดลดังกล่าวกับทุก สาขาวิชา ในบทความนี้ขอยกตัวอย่างกับการนำข้อมูลนิสิตภาคไฟฟ้ามาสร้างโมเดล ผลที่ได้จากข้อมูลและวิธีข้างต้นแสดงดังรูปที่ 1

3.1.3 การแปลความหมายข้อมูลและการนำความรู้ที่ได้มาใช้

เมื่อต้องการทราบสาขาวิชาที่เหมาะสมของนิสิตคนหนึ่ง จะพิจารณาลักษณะต่างๆ ของนิสิตที่แสดงเงื่อนไขกำกับไว้ในโหนดต่างๆ ในต้นไม้ช่วยการตัดสินใจ (decision tree) โดยพิจารณาลงมาตามทางของต้นไม้ทีละโหนดที่ตรงกับลักษณะของนิสิต จนกระทั่งถึงคลาสปลายทาง หากคลาสปลายทางที่ได้ประกอบด้วยสัดส่วน GOOD มากกว่า BAD มากๆ ย่อมแสดงว่า นิสิตที่มีลักษณะเดียวกับนิสิตคนนี้เข้าภาคไฟฟ้าแล้วจะมีผลการเรียนดี (GOOD) มากกว่าผลการเรียนไม่ดี (BAD) ดังนั้น สาขาวิชาไฟฟ้านี้ เป็นสาขาวิชาหนึ่งที่นิสิตคนนี้ควรพิจารณา โดยจะทดสอบข้อมูลนิสิตคนนี้กับโมเดลของทุกสาขาวิชา และเลือกเฉพาะสาขาวิชาที่คลาสปลายทางมีสัดส่วนของ GOOD มากกว่า BAD ถ้าผลออกมาว่ามีหลายสาขาวิชาที่เหมาะสม สามารถเลือกสาขาวิชาที่ดีที่สุดได้โดยพิจารณาจากสัดส่วนของ GOOD ในโมเดลที่มากกว่าเป็นหลัก หลังการทดสอบ ได้ผลลัพธ์เป็นที่น่าพอใจ คือ ได้ผลการทดสอบถูกต้องเฉลี่ย 84.58 % ในทุกโมเดล



รูปที่ 1 แสดง Decision tree การเลือกสาขาวิชาในโมเดลการจำแนกประเภทข้อมูลของสาขาวิศวกรรมไฟฟ้า

3.2 โมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชา (Prediction data model for each major)

3.2.1 การเตรียมข้อมูล

ปรับเปลี่ยนข้อมูลจากการเตรียมข้อมูลที่นำไปสร้างโมเดลการจำแนกประเภทข้อมูลเล็กน้อย คือ แทนผลการเรียนในวิชาต่างๆ ของโมเดลการพยากรณ์ข้อมูลนั้นด้วยตัวเลขที่แทนเกรดจริงๆ ที่นิสิตได้โดยไม่ต้องจัดกลุ่มเพื่อลดการกระจายของเกรด และประวัติส่วนตัวของนิสิตแทนด้วยตัวเลข 1 หรือ 2 นอกจากนี้คอลัมน์ที่ทำนายโดยโมเดลการจำแนกประเภทข้อมูลนั้นเป็นกลุ่มของข้อมูล ดังตารางที่ 4 แต่สำหรับตารางข้อมูลในโมเดลการพยากรณ์ข้อมูลนี้ คอลัมน์การทำนายคือ เกรดเฉลี่ยที่เป็นตัวเลข ดังตารางที่ 5

ตารางที่ 5 แสดงตัวอย่างข้อมูลสำหรับโมเดล Prediction ในแต่ละสาขาวิชา

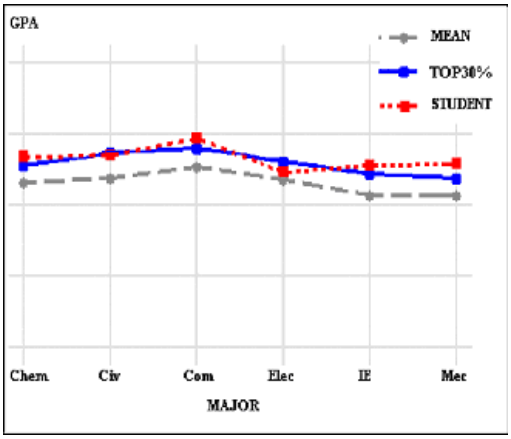
Stu_code	Sex	Address	204111	…	GPA
37058063	1	1	1	…	2.3
37058167	1	1	3.5	…	3.2
……	….	…..	…..	…	….

3.2.2 การสร้างโมเดล แปลความหมายข้อมูลและการนำความรู้ที่ได้มาใช้

สร้างโมเดลโดยนำผลการเรียนและลักษณะต่างๆ ของนิสิตแต่ละสาขาวิชามาเป็นตัวพิจารณา และสร้างโมเดลเพื่อทำนายเกรดเฉลี่ยตามลักษณะในแต่ละสาขาวิชา หลังจากสร้างโมเดลของทุกสาขาวิชาแล้ว เมื่อมีนิสิตคนหนึ่งต้องการทราบว่าตนเหมาะสมกับสาขาวิชาใด จะนำข้อมูลต่างๆ ของนิสิตคนนั้น ทั้งข้อมูลการศึกษา และข้อมูลประวัติส่วนตัวมาเป็นปัจจัยในแต่ละโมเดลเพื่อที่จะทำนายค่าเกรดเฉลี่ยที่ต้องการออกมา เมื่อนำข้อมูลนิสิตมาผ่านทุกโมเดลแล้ว จะได้เกรดเฉลี่ยของนิสิตคนนั้นตามโมเดลของแต่ละสาขาวิชา นั่นก็คือ เป็นการทำนายผลการเรียนของนิสิตในทุกๆ สาขาวิชา

จากนั้นจะหาเกณฑ์ผลการเรียนในทุกสาขาวิชาเพื่อที่จะทดสอบความถูกต้องของการทำนายเกรดเฉลี่ยในแต่ละโมเดล 1) ค่า MEAN คือ ค่าที่บอกผลการเรียนโดยเฉลี่ยของแต่ละสาขาวิชา 2) ค่า TOP30% คือ ค่าที่บอกถึงผลการเรียนของนิสิตที่ถือได้ว่าเรียนได้เป็นอันดับ 30 % แรกของแต่ละสาขาวิชา โดยถ้าโมเดลทำนายเกรดให้ไปอยู่ในเกณฑ์หนึ่ง แต่เกรดจริงๆ แล้วไปอยู่อีกในเกณฑ์หนึ่งนั้น แสดงว่าโมเดลทำนายผิดพลาด ผลการทดสอบความถูกต้องได้ค่อนข้างสูงมาก ข้อดีของโมเดลการพยากรณ์ข้อมูล

การนำเสนอสาขาวิชาให้กับนิสิตโดยใช้โมเดลการพยากรณ์ข้อมูล เป็นดังรูปที่ 2



รูปที่ 2 แสดงตัวอย่างการนำเสนอสาขาวิชาให้กับนิสิต โดยโมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชา

จากกราฟในรูปที่ 2 รายละเอียดต่างๆ ของกราฟเป็นดังนี้

- กราฟเส้นประห่าง แสดงเกรดเฉลี่ยสะสมของนิสิตทุกคนในแต่ละสาขาวิชา
- กราฟเส้นทึบ แสดงเกรดเฉลี่ยสะสมสูงสุด 30 % แรกของนิสิตในแต่ละสาขาวิชา ซึ่งแสดงให้เห็นถึงผลการเรียนของนิสิตที่เรียนดีในแต่ละสาขาวิชา
- กราฟเส้นประถี่ แสดงเกรดเฉลี่ยเมื่อจบการศึกษาที่ทำนายได้ของนิสิตในแต่ละสาขาวิชา

จากรายละเอียดดังกล่าว จะเห็นได้ว่า เส้นกราฟเส้นประห่างและเส้นทึบของนิสิตทุกคนจะมีลักษณะเหมือนกัน เพราะเป็นกราฟที่แสดงผลการเรียนของนิสิตในสาขาวิชาต่างๆ ที่คงที่ ส่วนกราฟเส้นประถี่ของนิสิตแต่ละคนนั้นมีลักษณะแตกต่างกันไปตามค่าเกรดเฉลี่ยที่ทำนายได้ของนิสิตแต่ละคนในแต่ละสาขาวิชา การนำเสนอกราฟเส้นประห่างและเส้นทึบให้กับนิสิตนั้นเพื่อที่นิสิตจะได้ทราบว่าผลการเรียนของนิสิตเป็นอย่างไรเมื่อเปรียบเทียบกับนิสิตส่วนใหญ่ในแต่ละสาขาวิชา จากรูปที่ 2 แสดงให้เห็นว่าถ้านิสิตเรียนในสาขาวิชาวิศวกรรมคอมพิวเตอร์แล้วจะมีแนวโน้มผลการเรียนตอนจบการศึกษาสูงกว่าสาขาวิชาอื่นๆ แต่สาขาวิชาที่นิสิตคนนี้เรียนแล้วมีแนวโน้มผลการเรียนสูงกว่านิสิตส่วนใหญ่ และสูงกว่านิสิตที่เรียนดีมากที่สุด คือ สาขาวิชาวิศวกรรมเครื่องกล เพราะจากกราฟจะเห็นได้ว่าระยะระหว่างกราฟเส้นประถี่กับกราฟเส้นทึบของสาขาวิชาวิศวกรรมเครื่องกลนี้จะมากกว่าของสาขาวิชาอื่นๆ ซึ่งการนำเสนอแบบนี้จะไม่บอกนิสิตว่าสาขาวิชาที่เหมาะสมที่สุด แต่ระบบจะนำเสนอผลที่ได้ในทุกสาขาวิชาให้กับนิสิตเพื่อที่นิสิตจะได้สามารถนำไปพิจารณาประกอบกับความต้องการของนิสิตอย่างเหมาะสม

3.3 การเปรียบเทียบโมเดล

เปรียบเทียบโมเดลที่ได้นำเสนอมาทั้ง 3 โมเดลดังตารางที่ 6

ตารางที่ 6 แสดงการเปรียบเทียบโมเดลทั้งสามโมเดล

โมเดลกลางการจำแนกประเภทข้อมูล	โมเดลการจำแนกประเภทข้อมูลในแต่ละสาขาวิชา	โมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชา
1. ความน่าเชื่อถือน้อย ประมาณ 50% เนื่องจากกลุ่มเป้าหมายมาก	1. ผลการทดสอบที่ได้มีความถูกต้องสูง 84.58 %	1. ผลการทดสอบที่ได้มีความถูกต้องสูง 96.84 %
2. ข้อมูลนิสิตในสาขาวิชาต่างๆ มีจำนวนแตกต่างกันมาก ทำให้โมเดลการทำนายโอนเอียงไปทางสาขาวิชาที่มีนิสิตมาก	2. ข้อมูลแต่ละสาขาวิชาไม่ส่งผลกระทบต่อกัน เนื่องมาจากวิธีนี้ได้สร้างโมเดลแยกกันในแต่ละสาขาวิชา	2. ข้อมูลแต่ละสาขาวิชาไม่ส่งผลกระทบต่อกัน เนื่องมาจากวิธีนี้ได้สร้างโมเดลแยกกันในแต่ละสาขาวิชา
3. ต้องมีการจัดกลุ่มผลการเรียนในแต่ละราย วิชา (High, Medium, Low) เพื่อลดการกระจายตัวของข้อมูล ทั้งนี้ ถ้าไม่มีการจัดกลุ่มข้อมูล จะทำให้โมเดลที่ได้กระจายตัว ข้อมูลใน	3. ต้องมีการจัดกลุ่มผลการเรียนในแต่ละรายวิชา (High, Medium, Low) เพื่อลดการกระจายตัวของข้อมูล ทั้งนี้ ถ้าไม่มีการจัดกลุ่มข้อมูล จะทำให้โมเดลที่ได้กระจายตัว ข้อมูลในแต่ละเส้นทางของ	3. ข้อมูลผลการเรียนในแต่ละรายวิชาเป็นข้อมูลผลการเรียนจริง (A, B+, B, C+, C, D+, D, F) ที่มิได้มีการจัดกลุ่ม ทำให้ข้อมูลที่นำมาสร้างโมเดล Prediction นั้นมีความ

แต่ละเส้นทางของโมเดลมีจำนวนน้อย เป็นผลทำให้ความถูกต้องของโมเดลลดลงอย่างมาก	โมเดลมีจำนวนน้อย เป็นผลทำให้ความถูกต้องของโมเดลลดลงอย่างมาก	ละเอียดและแม่นยำมากกว่าการจัดกลุ่มดัง เช่นโมเดลการจำแนกประเภทข้อมูล
4. โมเดลนำเสนอเพียงสาขาวิชาเดียวที่เหมาะสม ซึ่งส่งผลกระทบโดยตรงกับการตัดสินใจของนิสิต	4. โมเดลนำเสนอเฉพาะสาขาวิชาที่เหมาะสมให้กับนิสิตเท่านั้น สำหรับนิสิตบางส่วนที่มีผลการเรียนดี โมเดลจะเสนอทุกสาขาวิชาให้กับนิสิตเป็นสาขาวิชาที่เหมาะสม และสำหรับนิสิตบางส่วนที่มีผลการเรียนไม่ดี โมเดลจะไม่นำเสนอสาขาวิชาใดๆ ที่เหมาะสมให้กับนิสิตเลย ทำให้การตัดสินใจทั้งหมดไปตกอยู่กับนิสิต โดยที่โมเดลมิได้ช่วยนิสิตในกลุ่มเหล่านี้เลย	4. โมเดลนำเสนอแนวโน้มเกรดเฉลี่ยสะสมเมื่อจบการศึกษาของนิสิตในทุกสาขาวิชา ทำให้นิสิตได้เห็นแนวโน้ม และเห็นความแตกต่างของผลการเรียนของตน เมื่อเข้าไปศึกษาในสาขาวิชาที่แตกต่างกัน นอกจากนี้การนำเสนอได้เพิ่มในส่วนของ MEAN และ TOP30% ทั้งที่ได้เคยกล่าวไป ทำให้ช่วยให้นิสิตได้เห็นความแตกต่างในการเรียนในแต่ละสาขาวิชามากยิ่งขึ้น

4. การนำเทคนิคดาตาไมน์นิงมาประยุกต์ใช้ในการช่วยทำนายแนวโน้มเกรดรายวิชาต่างๆ ในภาคเรียนต่อไป

ปัญหาผลการเรียนตกต่ำของนิสิตนั้น มีสาเหตุหลายประการ เช่น นิสิตไม่มีความตั้งใจในการเรียน , การเตรียมตัวสอบไม่ดีพอ , นิสิตลงวิชาที่ไม่เหมาะกับความสามารถของตน หรือวิชาพื้นฐานที่จำเป็นไม่ดีพอ ในโครงงานนี้ได้เสนอแนวทางการแก้ปัญหาเหล่านี้บางส่วน คือ การทำนายแนวโน้มเกรดแต่ละวิชาในภาคเรียนต่อไป เพื่อเป็นแนวทางให้กับนิสิตในการเลือกลงรายวิชา และสามารถปฏิบัติตนในการเรียนในแต่ละวิชาได้อย่างเหมาะสม ซึ่งโมเดลที่จัดทำขึ้นนี้เป็นสิ่งที่ช่วยแนะแนวในการปฏิบัติให้กับนิสิตเท่านั้น การแก้ปัญหาหลักนั้นขึ้นอยู่กับตัวของนิสิต

งานวิจัยได้นำเทคนิคดาตาไมน์นิงหลายเทคนิคมาประยุกต์ใช้ในการทำนายแนวโน้มเกรด ในบทความนี้ขอนำเสนอการนำเทคนิคการค้นหากฎความสัมพันธ์ (Association rule discovery) มาประยุกต์ใช้ เพราะจากการทดสอบแล้วพบว่าโมเดลที่ได้ให้ผลการทดสอบที่น่าเชื่อถือมากที่สุด ในลำดับต่อไปนี้จะได้กล่าวถึงรายละเอียดต่างๆ ของการใช้เทคนิคการค้นหาความสัมพันธ์มาช่วยทำนายแนวโน้มเกรด อันได้แก่ การเตรียมข้อมูล การสร้างโมเดล การแปลความหมายข้อมูล และการนำความรู้ที่ได้มาใช้

4.1 การใช้เทคนิคการค้นหากฎความสัมพันธ์มาช่วยทำนายแนวโน้มเกรด

เทคนิคการสืบค้นกฎความสัมพันธ์ (Association rule discovery) เป็นการหาความสัมพันธ์ของข้อมูลที่มีอยู่ ในงานวิจัยนี้ได้ประยุกต์ใช้เทคนิคการสืบค้นกฎความสัมพันธ์กับข้อมูลผลการเรียนนิสิต โดยหาความสัมพันธ์ของผลการเรียนในแต่ละวิชาที่ส่งผลต่อกัน ซึ่งจะทำให้ได้ว่าวิชาใดบ้างที่มีผลต่อวิชาที่ต้องการจะทำนายเกรดล่วงหน้า

4.1.1 การเตรียมข้อมูล

งานวิจัยส่วนนี้ได้นำข้อมูลการลงทะเบียนเรียนของนิสิตคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ตั้งแต่ปี 2535-2542 รวมทั้งสิ้น 476,085 แถว จากนิสิตกว่า 10,000 คน เช่นเดียวกับในส่วนที่ 3 มาใช้ โดยการเตรียมข้อมูลนั้นคล้ายกับการเตรียมข้อมูลในตารางที่ 4 แต่ในการเตรียมข้อมูลครั้งนี้จะคัดเลือกเฉพาะคอลัมน์ผลการเรียนของนิสิต โดยจะไม่นำคอลัมน์ประวัติส่วนตัวของนิสิตมาพิจารณา ทั้งนี้เพราะความสัมพันธ์ของการเรียนในแต่ละวิชานั้นส่งผลต่อผลการเรียนในอีกวิชาหนึ่งมากกว่าประวัติส่วนตัวของนิสิต นอกจากนี้ในการเตรียมข้อมูลครั้งนี้จะนำผลการเรียนในทุกวิชา ของทุกภาคการศึกษามาเป็นตัวพิจารณาด้วย เพราะเราต้องการหาความสัมพันธ์ของทุกวิชาในทุกปีการศึกษา รูปแบบตารางที่อยู่ในรูปแบบที่เหมาะสมกับเทคนิคการสืบค้นกฎความสัมพันธ์เป็นดังตารางที่ 7 และเนื่องมาจากในสาขาวิชาต่างๆ มีรายวิชาแตกต่างกันออกไป ดังนั้นจึงต้องแบ่งข้อมูลออกตามแต่ละสาขาวิชา แล้วจึงสร้างโมเดลการสืบค้นกฎความสัมพันธ์ สำหรับแต่ละสาขาวิชานั้นๆ

ตาราง 7 แสดงตัวอย่างข้อมูลสำหรับการใช้เทคนิคการค้นหากฎความสัมพันธ์ในการทำนายเกรด

Stu_code	Subject1	Subject2	…
37058063	204111Medium	403111Low	…
37058167	204111High	403111High	…
……	…..	…..	…

4.1.2 การสร้างโมเดล

จากตารางที่ 7 แสดงผลการเรียนในรายวิชาต่างๆ ของนิสิตแต่ละคน จากตารางที่ 7 ได้ตัวอย่างความสัมพันธ์หนึ่งเป็น

[204111Medium] + [403111Low] -> [417167Low]

จากความสัมพันธ์นี้ สามารถอธิบายได้ว่าเมื่อเรียนวิชา 204111 ได้เกรดอยู่ในช่วง C+, C และเรียนวิชา 403111 ได้เกรดอยู่ในช่วง D+, D, F แล้วจะได้ผลการเรียนวิชา 417167 อยู่ในช่วง D+, D, F ซึ่งจากความสัมพันธ์นี้เป็นเพียงตัวอย่างหนึ่งของนิสิตคนหนึ่งเท่านั้น แต่ต้องหาความสัมพันธ์

ทั้งหมดของทุกรายวิชาและของนิสิตทุกคน ซึ่งจะได้ความสัมพันธ์ที่มีค่าสนับสนุน (support) และค่าความมั่นใจ (confidence) แตกต่างกันออกไป ผลที่ได้จากการใช้เทคนิคการค้นหากฎความสัมพันธ์จะอยู่ในรูปความสัมพันธ์ของวิชาต่างๆ มากมาย ดังรูปที่ 3

Support(%	Confidence(%	Type	Lift	Rule Body		Rule Head
3.1769	84.6200	.	4.9...	[417167High]+[204111High]+[417168High]+[355113High]	=..	[403111High]
4.1155	77.0300	+	4.4...	[417167High]+[417168High]+[355113High]	=..	[403111High]
3.7545	74.2900	+	4.3...	[417167High]+[204111High]+[417168High]	=..	[403111High]
3.4657	73.8500	+	4.2...	[204111High]+[417168High]+[355113High]	=..	[403111High]
3.5379	70.0000	+	4.2...	[417167High]+[420112High]	=..	[417168High]
3.4657	67.6100	+	4.1...	[417167High]+[420111High]	=..	[417168High]
3.8989	69.2300	.	4.0...	[417167High]+[204111High]+[355112High]+[355113High]	=..	[403111High]
5.0542	67.9600	+	3.9...	[417167High]+[204111High]+[355113High]	=..	[403111High]
3.5379	66.2200	.	3.8...	[417167High]+[208111High]+[355113High]	=..	[403111High]
3.1769	62.8600	.	3.8...	[417167High]+[204111High]+[355113High]+[403111High]	=..	[417168High]
3.3213	65.7100	+	3.8...	[417167High]+[420112High]	=..	[403111High]
3.1769	64.7100	.	3.7...	[417168High]+[355112High]+[355113High]	=..	[403111High]
3.7545	59.7700	.	3.6...	[417167High]+[204111High]+[403111High]	=..	[417168High]
3.0325	62.6900	.	3.6...	[420111Medium]+[417167High]+[204111High]+[355113High]	=..	[403111High]
3.0325	62.6900	+	3.6...	[420112High]+[355113High]	=..	[403111High]
4.2599	62.1100	+	3.6...	[204111High]+[417168High]	=..	[403111High]
3.4657	96.0000	+	3.5...	[417168High]+[420111High]	=..	[417167High]
4.6209	60.9500	+	3.5...	[417168High]+[355113High]	=..	[403111High]
3.0325	60.0000	.	3.4...	[208111Medium]+[417167High]+[355112High]+[355113High]	=..	[403111High]
3.6823	60.0000	.	3.4...	[420111Medium]+[417167High]+[355112High]+[355113High]	=..	[403111High]
4.1877	59.7900	+	3.4...	[417167High]+[204111High]+[355112High]	=..	[403111High]
3.6101	56.8200	+	3.4...	[420111High]	=..	[417168High]
5.7040	58.5200	.	3.3...	[417167High]+[355112High]+[355113High]	=..	[403111High]

รูปที่ 3 แสดงตัวอย่างความสัมพันธ์ของผลการเรียนในรายวิชาต่างๆ

ความสัมพันธ์ที่ได้ในแต่ละสาขาวิชามีมากกว่า 200,000 กฎ ซึ่งต้องตัดความสัมพันธ์บางส่วนโดย

1) กำหนดค่าสนับสนุนต่ำสุด (minimum support) และค่าความมั่นใจต่ำสุด (minimum confidence) ไว้เพื่อเลือกเฉพาะความสัมพันธ์ที่มีจำนวนนิสิตมาก และน่าเชื่อถือ มานำเสนอให้กับนิสิต

2) กำจัดความสัมพันธ์ที่ลำดับของวิชาไม่เป็นไปตามข้อกำหนดหลักสูตร และคัดเลือกเฉพาะความสัมพันธ์ที่ทางด้านซ้ายมือของกฎเป็นวิชาที่นิสิตเคยเรียน และทราบผลการเรียนแล้ว

4.1.3 การแปลความหมายข้อมูลและการนำความรู้ที่ได้มาใช้

เมื่อนิสิตต้องการทำนายผลการเรียนในรายวิชาหนึ่ง โมเดลจะทำนายผลการเรียนให้กับนิสิตโดยพิจารณาวิชาที่นิสิตได้เคยลงทะเบียนเรียนที่ทราบผลการเรียนมาแล้วทุกวิชา กับกฎความสัมพันธ์ต่างๆ ที่หาออกมาได้ ซึ่งจะเห็นได้ว่าสามารถหากฎความสัมพันธ์ที่ตรงกับเงื่อนไขความต้องการ (คือกฎความสัมพันธ์ที่ด้านขวาของกฎเป็นวิชาที่เราต้องการทำนาย และด้านซ้ายของกฎเป็นวิชาและผลการเรียนในวิชาต่างๆ ของนิสิตที่เคยเรียนมา) ออกมาได้หลายกฎความสัมพันธ์

ตัวอย่างเช่น จากตารางที่ 7 ถ้าต้องการทำนายเกรดในวิชา 417168 ของนิสิตรหัส 37058063 เมื่อพิจารณากฎตามเงื่อนไขความต้องการข้างต้นแล้วได้ความสัมพันธ์ดังนี้

(1) [204111Medium]+[403111High]+[417167Medium] -> [417168Medium] confidence = 86.5

(2) [403111Low] + [417167Low] -> [417168Low] confidence = 80.3

(3) [403111Medium] + [204111Medium] -> [417168Medium] confidence = 84.2

จากตัวอย่างความสัมพันธ์ จะเห็นได้ว่า กฎที่ได้มาทั้งหมดนั้นสามารถทำนายผลการเรียนในวิชาที่ต้องการทำนาย (417168) ได้ออกมาหลายแบบด้วยกัน ในกรณีนี้ทำนายได้เป็น Medium และ Low ซึ่งควรนำเสนอผลการเรียนเพียงช่วงเดียวให้กับนิสิต ดังนั้นต้องกำหนดหลักเกณฑ์ในการตัดสินว่าจะเลือกความสัมพันธ์หรือนำเสนอผลการเรียนใดให้กับนิสิตจึงจะถูกต้องมากที่สุด การลำดับความสำคัญของเกณฑ์การเลือกความสัมพันธ์เป็นดังนี้ [1]

1) เลือกความสัมพันธ์ที่ทางด้านซ้ายมือของกฎมีวิชาและผลการเรียนตรงกับนิสิตคนนั้นมากที่สุดเป็นอันดับแรก
2) เลือกความสัมพันธ์ที่มีค่าความมั่นใจ (confidence) สูงสุดเมื่อเกณฑ์ในข้อ 1) เท่ากัน
3) เลือกความสัมพันธ์ที่มีค่าสนับสนุน (support) สูงสุดเมื่อเกณฑ์ในข้อ 2) เท่ากัน

จากตัวอย่าง เมื่อพิจารณาตามเกณฑ์ข้อ 1) เป็นอันดับแรก จะเห็นได้ว่าความสัมพันธ์ (1) มีวิชาและผลการเรียนที่ตรงตามเกณฑ์ที่ 1) 1 วิชาคือ [204111Medium] ส่วนความสัมพันธ์ (2) และ (3) มีวิชาและผลการเรียนที่ตรงตามเกณฑ์เท่ากับ 2 และ 1 ตามลำดับ ดังนั้น จากความสัมพันธ์นี้ จะสรุปได้ว่า ความสัมพันธ์ที่ (2) นั้นตรงตามหลักเกณฑ์มากที่สุด และทำนายได้ว่าเกรดในวิชา 417168 ของนิสิตคนนี้ในภาคการศึกษาต่อไปจะอยู่ในช่วง Low (D+, D, F) นอกจากนี้ ระบบได้นำเสนอเปอร์เซ็นต์ความเป็นไปได้ที่นิสิตคนนั้นจะได้เกรดอยู่ในช่วงที่เราทำนายโดยนำมาจากค่าความมั่นใจ (confidence) ให้กับนิสิตด้วย ทั้งนี้เพราะโมเดลทำนายโดยอ้างอิงมาจากข้อมูลเดิมของนิสิตที่เคยเรียนมาและได้ผลการเรียนเช่นเดียวกันกับนิสิตคนนั้น เช่นถ้ามีความสัมพันธ์ที่มีค่าความมั่นใจเท่ากับ 100 % นั้น มิได้หมายความว่านิสิตคนนั้นจะได้เกรดตามที่ทำนาย 100 % แต่หมายความว่า ข้อมูลของนิสิตทั้งหมดที่ได้นำมาสร้างโมเดลที่มีผลการเรียนในรูปแบบเดียวกันกับนิสิตคนนี้ได้ผลการเรียนในวิชาดังกล่าวเป็นแบบเดียวกัน 100 %

โมเดลที่ได้นี้เป็นเพียงสิ่งที่ช่วยแนะแนวทางให้กับนิสิต ในกรณีที่นิสิตสามารถเลือกวิชาเรียนได้ เช่น วิชาเลือก โมเดลจะเป็นสิ่งที่ช่วยนิสิตตัดสินใจในการเลือกลงทะเบียนเรียนในวิชาที่เหมาะสมกับความสามารถของตน ในกรณีที่เป็นวิชาบังคับและนิสิตต้องลงทะเบียนเรียน โมเดลนี้จะเป็นสิ่งที่ช่วยชี้แนะแนวทางในการปฏิบัติตนที่ดีให้กับนิสิตต่อวิชานั้นๆ ถ้าโมเดลทำนายออกมาว่าผลการเรียนจะออกมาไม่ดี (D+, D, F)

นั้นหมายความว่า นิสิตควรจะใส่ใจ และตั้งใจในการเรียนวิชานี้เป็นพิเศษ ถ้าโมเดลทำนายผลการเรียนได้ว่าค่อนข้างดี (A, B+, B) ก็มิได้หมายความว่าจะให้นิสิตละเลย หรือไม่ต้องใส่ใจในวิชานี้ แต่สิ่งที่ได้จากโมเดลหมายความว่า นิสิตคนนั้นมีโอกาสที่จะเรียนในวิชานั้นได้ดี เนื่องมาจากผลการเรียนในบางวิชาที่มีผลต่อวิชานั้นของนิสิตคนนี้ค่อนข้างดี แสดงว่านิสิตมีพื้นฐานได้เปรียบกว่านิสิตคนอื่น ดังนั้นนิสิตควรที่จะตั้งใจเรียนเพื่อให้ได้ผลการเรียนดีตามที่โมเดลได้ทำนายไว้

ตัวอย่างการนำเสนอการทำนายแนวโน้มผลการเรียนให้กับนิสิตแสดงได้ดังรูปที่ 4

จากรูปที่ 4 แสดงแนวโน้มผลการเรียนในรายวิชาต่างๆ ที่นิสิตต้องการทราบ โดยแสดงชื่อรายวิชา ช่วงเกรดที่ทำนายได้ เปอร์เซ็นต์ความเป็นไปได้ในการทำนายในรายวิชานั้นๆ จำนวนนิสิตที่อ้างอิงโดยมีลักษณะการเรียนคล้ายคลึงกับนิสิตในรายวิชานั้นๆ และจำนวนนิสิตทั้งหมดในสาขาวิชานั้นที่นำมาสร้างโมเดล

ผลการทำนายเกรดรายวิชาต่าง ๆ จากจำนวนนิสิต 437 คน

รหัสวิชา	วิชา	เกรดที่ทำนาย	% ความเป็นไปได้	จำนวนนิสิต
417168	คณิตศาสตร์วิศวกรรม II	C+ หรือ C	68.7000	78
417267	คณิตศาสตร์วิศวกรรม III	A หรือ B+ หรือ B	95.5600	39
417268	คณิตศาสตร์วิศวกรรม IV	C+ หรือ C	77.7800	13
208111	การเขียนแบบวิศวกรรม	A หรือ B+ หรือ B	60.8700	69
208221	กลศาสตร์วิศวกรรม I	A หรือ B+ หรือ B	88.8900	39
208222	กลศาสตร์วิศวกรรม II	A หรือ B+ หรือ B	71.8800	21
208281	การฝึกงานโรงงาน	A หรือ B+ หรือ B	100.0000	148
204331	การโปรแกรมระ บบ	A หรือ B+ หรือ B	90.9100	48
204341	วิศวกรรมซอฟต์แวร์	A หรือ B+ หรือ B	82.3500	13
204351	ฐานข้อมูลและการสืบค้นสารสนเทศ	A หรือ B+ หรือ B	95.3800	61
204371	เทคนิคการแปลงและการวิเคราะห์สัญญาณ	A หรือ B+ หรือ B	87.5000	26
204497	สัมมนา	A หรือ B+ หรือ B	96.0000	43
204499	โครงงานวิศวกรรมคอมพิวเตอร์	A หรือ B+ หรือ B	94.4400	13
205211	การวิเคราะห์วงจรไฟฟ้า I	D+ หรือ D หรือ F	78.0000	34
205213	ปฏิบัติการวงจรไฟฟ้า	A หรือ B+ หรือ B	100.0000	13
205251	วงจรและระ บบอิเล็กทรอนิกส์ I	D+ หรือ D หรือ F	82.3500	13

รูปที่ 4 แสดงตัวอย่างการนำเสนอการทำนายแนวโน้มผลการเรียน

สาเหตุที่นำเสนอเกรดให้นิสิตเป็นช่วงนั้น เนื่องมาจาก วัตถุประสงค์ของงานวิจัยนี้ที่ต้องการชี้แนะแนวทางในการปฏิบัติตนในการเรียนในวิชานั้นๆ อย่างเหมาะสม ซึ่งการที่นิสิตทราบว่า ผลการเรียนมีแนวโน้มจะออกมาดี ปานกลาง หรือไม่ดีนั้น เป็นการนำเสนอที่มีผลต่อดีต่อการปฏิบัติตนของนิสิต โดยถ้านำเสนอเกรดให้กับนิสิตแบบเจาะจงว่านิสิตจะได้ A นั้นอาจส่งผลเสียทำให้นิสิตเกิดความลำพองใจ ไม่ตั้งใจเรียน ทำให้ผลการเรียนตกต่ำกว่าที่ควรจะเป็น หรือได้ทำนายว่านิสิตจะได้ F อาจเป็นการบั่นทอนกำลังใจของนิสิตลงไปมาก ทำให้นิสิตเกิดความท้อแท้ในการเรียนรายวิชานั้นๆ ได้

บทสรุป

การพัฒนาคุณภาพการศึกษาสามารถทำได้ในหลายรูปแบบ ไม่ใช่ว่าทำเพียงแบบใดแบบหนึ่งแล้วจะสำเร็จได้ แต่ต้องอาศัยวิธีการต่างๆ มาประกอบกันเพื่อให้ได้ผลที่ดีที่สุด อีกทั้งต้องคำนึงถึงปัจจัยต่างๆ หลายประการ ที่จะส่งผลกระทบต่อกัน ซึ่งในบางครั้งปัจจัยเหล่านี้มีมากเกินกว่าที่จะพิจารณาได้ด้วยตาเปล่า

ในบทความนี้ได้นำเสนอการนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้เพื่อช่วยพิจารณาหารูปแบบความสัมพันธ์ของข้อมูลในมุมมองต่างๆ เพื่อที่จะได้นำสิ่งที่เป็นประโยชน์ที่ได้จากดาต้าไมน์นิงไปเป็นส่วนหนึ่งในการพัฒนาคุณภาพการศึกษาต่อไป ซึ่งผลลัพธ์ที่ได้จากงานวิจัยนี้ค่อนข้างเป็นที่น่าพอใจ โดยมีเปอร์เซ็นต์ความถูกต้องค่อนข้างสูง แต่มีปัญหาบางประการ ได้แก่ จำนวนข้อมูลในบางสาขาวิชามีปริมาณค่อนข้างน้อยทำให้โมเดลที่ได้ไม่แม่นยำเท่าที่ควร หากต้องการกำจัดความผิดพลาดที่เกิดจากปริมาณข้อมูลน้อยเกินไปจำเป็นต้องใช้ข้อมูลอย่างน้อยพันคนในแต่ละสาขาวิชาที่ต้องการทำนาย, วิธีที่นำเสนอไปนั้นอาจยังไม่ใช่วิธีที่ดีที่สุดในการวิเคราะห์แนวโน้ม, หน่วยงานการศึกษาที่ต้องการนำโมเดลนี้ไปใช้งาน จะต้องเตรียมข้อมูลให้ตรงตามโครงสร้างที่โมเดลนี้ได้ออกแบบไว้ แต่สามารถปรับปรุงโมเดลให้เหมาะสมกับโครงสร้างข้อมูลในแต่ละหน่วยงานได้

เอกสารอ้างอิง

[1] B.Liu and W.Hsu and Y.Ma. “Integrating Classification and Association Rule Mining” , .In Proceeding of International Conference KDD’98, 1998.

[2] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. 1996. “Advances in knowledge discovery and data mining”, AAAI/MIT Press.

[3] Han, J., Kamber, M. 2000., “Data Mining Concepts and Techniques”, Morgan Kaufmann.

[4] J.Gehtke, R.Ramakrishnan, and V.Ganti. Rainforest., “A framework for fast decision tree construction of large datasets”, In Proceeding of International Conference Very Large Database, p.416-427, 1998.

[5] Kitsana Waiyamai and Lotfi Lotfi Lakhal. 2000., “Knowledge Discovery from Very Large Databases Using Frequent Concept Lattices”, p.437-445, 2000.

[6] Kitsana Waiyamai, Chidchanok Songsiri and Thanawin Rakthanmanon., "A Data Mining based Approach for Improving Quality of Engineering Graduates". In Proceeding of 4th International Conference UNESCO International center for Engineering Education (UICEE ' 2001), p.84-88, 2001.

[7] R.Agrawal, H. Mannila, R. Srikant and A. I. Verkamo., "Fast Discovery of Association Rules". In U.M. Fayyad, G.Piatetsky-Shapiro, P.Smyth, and R.Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AASAi/MIT Press, p. 307-328, 1996.

[8] กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และ ธนาวินท์ รักธรรมานนท์, มารู้จักกับ Data Mining, "ไมโครคอมพิวเตอร์" (MICRO COMPUTER), Vol. 18, No. 187, p. 179-181, December 2000.

กฤษณะ ไวยมัย จบปริญญาตรีและปริญญาโททางวิทยาศาสตร์คอมพิวเตอร์ จาก University of Picardie ประเทศฝรั่งเศส จบปริญญาเอกทางด้านวิทยาการคอมพิวเตอร์จาก University of Clermont ประเทศฝรั่งเศส งานวิจัยหลักประกอบด้วยสามส่วนสำคัญ ส่วนแรกคือ ระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: Data Mining) ส่วนที่สองคือ ระบบสารสนเทศ (Information System) และ ส่วนที่สามคือ ระบบฐานข้อมูล (Database Management System) ปัจจุบัน ดร.กฤษณะ ไวยมัย ดำรงตำแหน่งอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

ชิดชนก ส่งศิริ จบปริญญาตรีทางวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์ งานวิจัยหลักประกอบด้วยการวิจัยและการพัฒนาระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (knowledge Discovery from very large Databases: Data Mining) , ระบบคลังข้อมูล (Data Warehousing) และระบบจัดการฐานข้อมูล (Database management) ปัจจุบันศึกษาปริญญาโทวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์

ธนาวินท์ รักธรรมานนท์ จบปริญญาตรีทางวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์ งานวิจัยหลักประกอบด้วยการวิจัยและการพัฒนาระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (knowledge Discovery from very large Databases: Data Mining) , ระบบคลังข้อมูล (Data Warehousing) และระบบจัดการฐานข้อมูล (Database management) ปัจจุบันศึกษาปริญญาโทวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์