การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ

จิราภา เลาหะวรนันท์, รชต ลิ้มสุทธิวันภูมิ และ บัณฑิต ฐานะโสภณ

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

Emails: lao.jirapa@gmail.com, rachataaa95@gmail.com

บทคัดย่อ

ในปัจจุบัน ทางคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ได้เปิดหลักสูตรการเรียนตามแขนงวิชา โดยมีการแบ่งเป็น 4 แขนงวิชา (วิศวกรรมซอฟต์แวร์, เทคโนโลยีเครือข่ายและระบบ, การพัฒนาสื่อประสมและเกม และ อัจฉริยะทางธุรกิจ) อย่างไรก็ตามนักศึกษาจำนวนมากไม่ทราบถึงความถนัดของตนเองที่แท้จริง ทำให้ผู้เขียนเกิดความสนใจที่จะพัฒนา “ระบบแนะนำแขนงวิชา”เพื่อเป็นเครื่องมือช่วยตัดสินใจเลือกแขนงวิชา โดยโครงงานชิ้นนี้ได้ใช้ข้อมูลผลการเรียนและผลการวัดความสามารถด้านต่างๆ ที่เกี่ยวข้องมาสร้างแบบจำลองพยากรณ์โดยเปรียบเทียบเทคนิคการทำเหมืองข้อมูล 5 เทคนิค และพยากรณ์ผ่านเทคนิค “Ensemble” โดยพบว่าการพยากรณ์มีความแม่นยำอยู่ที่ 72.92% โดยแขนงวิศวกรรมซอร์ฟแวร์สามารถทำนายได้แม่นยำถึง 86.67% ซึ่งประโยชน์จากการพัฒนาระบบนี้ทำให้นักศึกษาทราบถึงแขนงวิชาที่เหมาะสมกับตนเองมากที่สุดถึงน้อยที่สุดพร้อมระดับความน่าจะเป็น ซึ่งจะส่งผลให้นักศึกษาตัดสินใจเลือกสาขาที่เหมาะกับตนเองได้ และมีโอกาสประสบความสำเร็จในการเรียน

คำสำคัญ – *การทำเหมืองข้อมูล (Data Mining); การคัดเลือกแขนงวิชา; การพยากรณ์ผลการเรียน (Study Performance Prediction); การจำแนกหมวดหมู่ (Classification)*

Abstracts

The Bachelor of Science Program in Information Technology of Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang (KMITL) is a program that focuses on four different subject areas, namely software engineering, network and system technology, multimedia and game development, and business intelligence. However, many students are not yet to realize their strengths and preferences making it difficult to choose the one that works best for them. To alleviate this problem, the authors had developed “major subject selection decision support system” which can be used as a tool to help students choose their most suitable subject area. The system employs relevant information determining a student’s academic performance such as GPA and several other factors. It uses the information to create a predictive model based on five data mining techniques and an additional technique called “Ensemble”. As a result, we found that the average accuracy is at 72.92 percent. The subject area of software engineering achieves the best accuracy at 86.67 percent. Since the results are ranked in descending order, students will easily know which of four subject areas are the best and worst for them. The authors believe that the aforementioned system will help students make a better decision and improve their chances in successful learning.

Keywords - *Data mining; Academic major selection; Study Performance Prediction ; Classification*

1. บทนำ

แม้ว่าหลายประเทศกำลังผันตัวเข้าสู่โลกยุคดิจิทัล แต่ปัจจุบันยังขาดบุคลากรที่มีคุณภาพทางด้านไอทีอย่างต่อเนื่อง ถือเป็นวิกฤติของอุตสาหกรรมอย่างยิ่งในเวลานี้ ซึ่งเมื่อดูสถิติที่แสดงจำนวนนักศึกษาที่เข้าศึกษาและจำนวนบัณฑิตที่สำเร็จการศึกษาในสาขาเทคโนโลยีสารสนเทศ จากสำนักงานคณะกรรมการอุดมศึกษา (สกอ.) ประเทศไทย พบว่าปีการศึกษา 2556 มีบัณฑิตที่จบออกไปเพียง 6,262 คน จากนักศึกษา 9,230 คน และมีส่วนน้อยที่พร้อมเข้าสู่ตลาดแรงงาน ปัญหาอาจเกิดจากบัณฑิตด้านไอทีส่วนมากไม่สามารถหางานตรงกับสาขาที่เรียนมาได้เนื่องจากจบมาไม่ตรงกับความต้องการ ซึ่งในต่างประเทศได้มีการนำเอาเทคนิคการทำเหมืองข้อมูลมาช่วยแก้ไขและพัฒนาในการศึกษาจำนวนมากอย่างเช่น การนำเทคนิคทำเหมืองข้อมูลมาทำนายนักศึกษาสาขาคอมพิวเตอร์เพื่อดูว่าจะมีโอกาสสอบหรือไม่ โดยดูจากปัจจัยเกรดและเพศ [1] หรือการเปรียบเทียบเทคนิคต่างๆของการทำเหมืองข้อมูลที่ช่วยพยากรณ์ความสำเร็จของนักเรียน [2] เป็นต้น เพื่อปรับหลักสูตรให้สอดคล้องกับความต้องการและมีคุณภาพ

ทางสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังนั้นเป็นสถาบันอีกแห่งหนึ่งที่มุ่งเน้นผลิตบัณฑิตที่มีคุณภาพทางด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศด้วยเช่นกัน โดยในทางหนึ่ง ได้มีการจัดตั้งคณะเทคโนโลยีสารสนเทศขึ้น ซึ่งหลักสูตรการเรียนการสอนแตกต่างจากภาควิชาคอมพิวเตอร์ ทั้งของคณะวิศวกรรมศาสตร์ และ คณะวิทยาศาสตร์ โดยหลักสูตรมีการมุ่งเน้นให้นักศึกษารู้ลึก รู้จริง และสามารถประยุกต์ เพื่อให้นักศึกษาที่สนใจได้เข้ามาศึกษาเรียนรู้ ปฏิบัติจริง โดยศึกษาในส่วนของพื้นฐานความรู้โดยทั่วไปใน 2 ปีแรกและเมื่อเข้าสู่ปีการศึกษาที่ 3 นักศึกษาจะต้องเลือกศึกษาใน 4 แขนงวิชาหลักได้แก่แขนงวิศวกรรมซอฟต์แวร์ (Software Engineering) หรือที่รู้จักในชื่อแขนง Software ที่เน้นเรียนด้านการพัฒนาซอฟต์แวร์ทั้งในภาคส่วนธุรกิจและเกม ทั้งที่เป็นเว็บแอพพลิเคชันและบนมือถือ เพื่อตอบสนองความต้องการในยุคปัจจุบัน โดยจะสอนทั้งในส่วนทฤษฎีและการปฏิบัติจริง เพื่อให้นักศึกษาเกิดองค์ความรู้ที่สามารถจัดการงานได้ครบครบกระบวนการ แขนงเทคโนโลยีเครือข่ายและระบบ (Network and System Technology) หรือแขนง Network เน้นเรียนเกี่ยวกับการออกแบบ พัฒนาและบริหารระบบเครือข่าย ซึ่งนักศึกษาจะได้รู้ตั้งแต่พื้นฐานจนถึงการดูแลระบบเครือข่าย เพื่อให้มองเห็นภาพรวมและสามารถนำไปประยุกต์ใช้กับงานอื่นๆได้ ต่อมาคือแขนงการพัฒนาสื่อประสมและเกม (Multimedia and Game Development) หรือแขนง Multimedia มีการเน้นสอนการพัฒนาซอฟต์แวร์สื่อประสมและประยุกต์ใช้ ซึ่งนักศึกษาจะได้เรียนการใช้สื่อประสมตลอดจนวิธีการใช้เครื่องมือต่างๆในการพัฒนา จนสามารถพัฒนาเป็นเกมที่ใช้งานได้จริง และแขนงอัจฉริยะทางธุรกิจ (Business Intelligence) หรือแขนง Business ที่มีการเน้นในด้านการวิเคราะห์และนำเสนอข้อมูลเพื่อการตัดสินใจ ทำให้นักศึกษาได้รู้จักการวิเคราะห์ข้อมูลในหลายๆประเภทและสามารถนำมาสังเคราะห์จนเกิดเป็นสารสนเทศที่นำไปประยุกต์ใช้งานอื่นๆ หรือนำไปช่วยในการตัดสินใจของผู้บริหารระดับสูงได้ โดยการเลือกเรียนในแขนงวิชานั้นจะมีการคัดเลือกจากผลการเรียนเฉลี่ยโดยรวมแล้วนำมาเรียงอันดับคัดเลือก ซึ่งนักศึกษาที่มีผลการเฉลี่ยโดยรวมสูงยิ่งมีโอกาสได้ศึกษาในแขนงวิชาที่ต้องการมากขึ้น ทำให้นักศึกษาจำนวนหนึ่ง ไม่ได้คัดเลือกเข้าในแขนงวิชาที่เหมาะสมกับความถนัดของตนเอง ส่งผลให้อาจไม่ประสบความสำเร็จทางการเรียนในแขนงวิชานั้นๆ ในอีกมุมหนึ่งนักศึกษาบางส่วนที่ได้ศึกษาในแขนงวิชาที่ตนเองสนใจ แต่เมื่อได้สัมผัสกับการเรียนการสอนจริง พบว่าไม่เหมาะกับความถนัดหรือความสามารถของตนเอง ซึ่งทำให้ไม่มีแรงจูงใจในการศึกษา มีผลให้ระดับผลสัมฤทธิ์ทางการเรียนไม่ดีเท่าที่ควร และอาจไม่ประสบผลสำเร็จด้วยเช่นกัน ดังนั้นเมื่อจบการศึกษา นักศึกษาเหล่านี้อาจทำงานในแขนงที่ศึกษาได้ไม่เต็มความสามารถเท่าที่ควรหรืออาจย้ายไปทำงานสายอื่นที่ไม่เกี่ยวข้องกับเทคโนโลยี

ด้วยเหตุนี้ทางผู้พัฒนาจึงได้เล็งเห็นถึงความสำคัญและมีแนวคิดที่ทำการศึกษาการใช้เทคนิคการทำเหมืองข้อมูลเพื่อจำแนกและคัดเลือกแขนงวิชาของนักศึกษาคณะเทคโนโลยีสารสนเทศขึ้น โดยผลจากโครงการนี้จะเป็นแนวทางเพื่อช่วยในการตัดสินใจเลือกแขนงวิชาที่เหมาะสมกับความถนัดของนักศึกษาคณะเทคโนโลยีสารสนเทศแต่ละบุคคล ซึ่งอาจส่งผลให้จบการศึกษาออกมาเป็นบัณฑิตที่มีคุณภาพและสามารถประยุกต์ความรู้เข้ากับตัวงานได้อย่างดีและเต็มความสามารถ และเป็นที่ต้องการขององค์กรต่างๆ โดยงานวิจัยนี้จะใช้ปัจจัยจากทั้งเกรดรายวิชาต่างๆ เพศ และแบบทดสอบความถนัดด้านต่างๆ มาประกอบรวมกันเพื่อนำไปพยากรณ์ผ่านเทคนิคการทำเหมืองข้อมูล 5 เทคนิคและนำมาโหวตคำตอบร่วมกันเพื่อใช้ในการพยากรณ์แขนงวิชาที่เหมาะสมกับนักศึกษาแต่ละคน

2. การทบทวนวรรณกรรม

2.1 ปัจจัยที่ส่งผลต่อความสำเร็จทางการเรียนคอมพิวเตอร์

จากการทบทวนวรรณกรรมซึ่งเกี่ยวข้องกับการพยากรณ์การประสบความสำเร็จทางการเรียน พบว่าปัจจัยที่ส่งผลต่อการประสบผลสำเร็จทางด้านการเรียนคอมพิวเตอร์นั้นมีอยู่หลายปัจจัยที่ควรให้ความสนใจ

ซึ่งปัจจัยสำคัญที่ผู้เขียนได้ค้นพบจากศึกษาวรรณกรรมที่เกี่ยวข้องนั้น สามารถสรุปดังนี้

1) เกรด (Grade) [2][3] ถูกพบว่าเป็นปัจจัยที่สำคัญที่สุดในหลายงานวิจัย เนื่องจากเป็นค่าที่บ่งบอกถึงผลการดำเนินงานที่ผ่านมาว่าประสิทธิภาพการเรียนรู้เป็นอย่างไร ซึ่งอาจมองไปถึงความตั้งใจ ความรับผิดชอบในการเรียนด้วย เกรดจึงเป็นค่าที่มองเห็นและจับต้องได้ ทั้งยังเป็นค่าที่มีอิทธิพลที่ส่งผลต่อความสำเร็จมาก
2) เพศ (Sex) [3][4] เป็นอีกปัจจัยหนึ่งที่มีความเกี่ยวข้องกับความสำเร็จ เนื่องจากเพศชายกับเพศหญิงจะมีลักษณะและความสามารถในการเรียนรู้ที่แตกต่างกัน บางลักษณะส่งผลต่อผู้หญิงให้เกิดความสำเร็จมากกว่า บางอย่างส่งผลต่อผู้ชายมากกว่า เช่น ผู้ชายจะมีความสามารถในการค้นคว้าด้วยตนเองมากกว่าผู้หญิง
3) ความสามารถในตนเอง (Self-Efficacy) [4][5] เป็นอีกปัจจัยหนึ่ง โดยความสามารถในตนเองมีความเกี่ยวข้องกับระดับความสบายใจอย่างมาก เพราะคนที่สามารถทำได้ดี จะรู้สึกพอใจหรือมีความสบายมากกว่า ซึ่งเมื่อดูจากสถิติแล้ว ความสามารถในตนเองนั้นจะมีในผู้ชายมากกว่าผู้หญิง
4) ระดับความสบายใจ (Comfort Level) [4][5][6] เป็นปัจจัยสำคัญที่ก่อให้เกิดการเรียนรู้ที่มีประสิทธิภาพ โดยนักเรียนที่สามารถศึกษาหรือปฏิบัติที่ดีกว่ามักมีระดับความสบายใจสูง ในทางตรงกันข้าม หากมีระดับความสบายใจที่ต่ำ จะทำให้แสดงออกมาได้แย่กว่าปกติ ซึ่งระดับความสบายใจมีความสำคัญต่อผู้หญิงมากกว่าผู้ชาย
5) ประสบการณ์การเขียนโปรแกรม (Previous Computer Programming Experiences) [4][5][7] ส่งผลให้เกิดความสำเร็จในการเรียนรู้ ซึ่งมักจะมองควบคู่กับความรู้พื้นฐานทางคณิตศาสตร์ แต่โดยส่วนใหญ่แล้ว ผู้ที่มีประสบการณ์ในการเขียนโปรแกรมสูง ย่อมประสบความสำเร็จได้มากกว่า
6) พื้นฐานทางคณิตศาสตร์ (Mathematics Background) [4][5][7] ผู้ที่มีความรู้พื้นฐานคณิตศาสตร์ดี จะมีทักษะในการคิด และเข้าใจเรื่องของโครงสร้างและแนวทางการดำเนินงาน ซึ่งเป็นพื้นฐานสำคัญของการเขียนโปรแกรมที่ดี

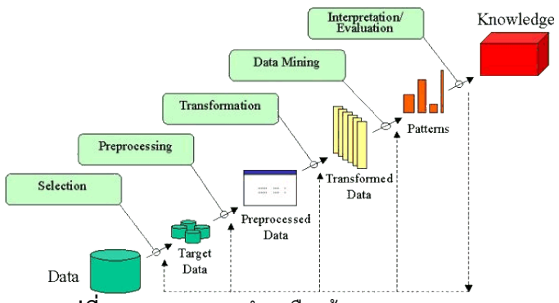
จากการทบทวนวรรณกรรมที่เกี่ยวข้อง ทางผู้พัฒนาได้พิจารณาเลือก 4 ปัจจัยนำมาใช้ในการสร้างแบบจำลอง ซึ่งประกอบด้วยปัจจัยด้านเกรด เพศ พื้นฐานทางคณิตศาสตร์ และประสบการณ์การเขียนโปรแกรม โดยมองจากปัจจัยที่มีการกล่าว ถึงในบทวรรณกรรมเป็นจำนวนมาก อีกทั้งปัจจัยเหล่านี้สามารถวัดหรือประเมินผลจากข้อมูลหรือผลลัพธ์ที่มีอยู่ได้ทันที แต่สำหรับปัจจัยด้านความสามารถในตนเอง และระดับความสบายใจ จะไม่นำมาพิจารณาเนื่องจากเป็นปัจจัยที่ผู้วิจัยต้องใช้การสังเกตพฤติกรรมหรือวิธีการเรียนรู้ของผู้เรียนในขณะทำการเรียนการสอน ทำให้เกิดความลำบากในการรวบรวมข้อมูล และนำมาประมวลผล

2.2 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (Data Mining) [8] คือเทคนิคการวิเคราะห์ข้อมูลจากมุมมองที่แตกต่างกันและสามารถสรุปผลเป็นข้อมูลที่มีประโยชน์ กระบวนการทำเหมืองข้อมูลใช้หลายหลักการเช่น เทคนิคการเรียนรู้ของเครื่อง (Machine Learning), สถิติ, เทคนิคการสร้างภาพ (Visualization Techniques) เพื่อค้นพบและนำเสนอความรู้ในรูปแบบที่เข้าใจได้ง่าย โดยอีกความหมายหนึ่งคือกระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อสกัดสารสนเทศ รวมถึงรูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลขนาดใหญ่นั้น

2.2.1 กระบวนการทำงานของ Knowledge Discovery in Database (KDD) หรือ Data Mining

กระบวนการ Knowledge Discovery in Database (KDD) [9] หมายถึงกระบวนการในการค้นหาความรู้สารสนเทศจากกลุ่มข้อมูลที่มีจำนวนมาก ซึ่งมีขั้นตอนการทำเหมืองข้อมูลเป็นกระบวนการหลักเพื่อค้นหาลักษณะที่น่าสนใจของข้อมูลเหล่านี้ ทำให้ได้มาซึ่งรูปแบบข้อมูลที่เป็นเหตุเป็นผล เป็นประโยชน์และเข้าใจได้ง่าย ซึ่งการนำเอาเทคนิคการทำเหมืองข้อมูลไปใช้นั้นจะมีวิธีการต่างกัน ขึ้นอยู่กับวัตถุประสงค์ที่จะนำไปใช้งานว่าใช้เพื่อทำอะไร สิ่งที่ต้องการรู้คืออะไร ดังนั้นจึงมีการนำเสนอวิธีการที่หลากหลายสำหรับเป้าหมายที่ต่างกัน เพื่อให้ได้ผลลัพธ์ที่เหมาะสมต่อความต้องการหลังจากนำไปใช้งาน โดยขั้นตอนการทำเหมืองข้อมูลสามารถแสดงเป็นรูปได้ดังนี้



รูปที่ 1: กระบวนการทำเหมืองข้อมูล (source: [9])

1) Business Objective Determination เป็นกระบวนการระบุปัญหาทางธุรกิจหรือกำหนดวัตถุประสงค์สำหรับการค้นหาสารสนเทศว่าต้องการหาอะไร และหาไปเพื่ออะไร

2) Data Selection เป็นกระบวนการคัดเลือกแหล่งข้อมูล โดยทำการวิเคราะห์แหล่งข้อมูลจากทั้งภายในและภายนอก และทำการคัดเลือกเพื่อนำมาใช้
3) Data Preprocessing เป็นการเตรียมข้อมูลสำหรับการทำเหมืองข้อมูล เนื่องจากข้อมูลที่มาจากหลายแหล่งอาจจะมีรูปแบบที่ไม่เหมือนกัน จึงต้องมีการปรับคุณภาพให้เหมาะสมและสามารถนำไปใช้งานได้
4) Data Transformation เป็นการแปลงข้อมูลเพื่อให้ใช้ได้กับ model ต่างๆ เช่นการแปลงให้อยู่ในช่วงที่กำหนด แปลงจากตัวอักษรเป็นตัวเลข เป็นต้น
5) Data Mining เป็นขั้นตอนที่สำคัญที่สุดในการสกัดสารสนเทศจากข้อมูลโดยใช้เทคนิคต่างๆเช่น (1) Predictive Modeling การค้นหาโมเดลเพื่ออธิบายและจำแนกคลาสต่างๆได้ สำหรับโมเดลที่นิยมใช้คือ ต้นไม้ตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียม (Neural Network), (2) Data Segmentation การนำข้อมูลมาแบ่งเป็นส่วนๆ สำหรับโมเดลที่นิยมใช้คือ K-mean และ โครงข่ายประสาทเทียม (Neural Network) และ (3) Association Rule Discovery ซึ่งเกี่ยวข้องกับการค้นหากฎความ สัมพันธ์ของข้อมูลที่มี สำหรับโมเดลที่นิยมใช้คือ “Apriori Algorithm”
6) Evaluation of Result เป็นการวิเคราะห์และประเมินผลลัพธ์ที่ผ่านกระบวนการทำเหมืองข้อมูล มาสรุปและตีความหมายที่ได้ออกมาเป็นความรู้ใหม่ (Knowledge) ที่สามารถนำไปเป็นสารสนเทศที่ช่วยในการตัดสินใจได้

สำหรับข้อมูลที่ถูกเก็บรวบรวมมาเพื่อใช้งานนั้นจะนำมาวิเคราะห์โดยใช้เทคนิค ensemble เนื่องจากเทคนิคนี้เป็นเทคนิคที่มีประสิทธิภาพสูง มีการใช้โมเดล classification หลายๆ โมเดล (model) มาช่วยในการหาคำตอบ โดยวิธีการของเทคนิค ensemble สามารถอธิบายได้เป็น 3 รูปแบบ [10] ดังนี้

1) Vote ensemble เป็นการสร้างโมเดลด้วยเทคนิคต่างๆกัน โดยใช้ชุดข้อมูลสำหรับสอน (Training Data) ชุดเดียวกัน
2) Bootstrap aggressive (Bagging) เป็นการสร้างโมเดลด้วยเทคนิคเดียวกันทั้งหมด โดยการสุ่มชุดข้อมูลสำหรับสอนออกเป็นหลายชุด
3) Random forest เป็นเทคนิคที่คล้ายๆ กับ bagging แต่นอกจากสุ่มข้อมูล ได้ทำการสุ่มเลือกแอตทริบิวต์ต่างๆ ออกมาเป็นหลายๆชุดด้วย และสร้างโมเดลด้วยเทคนิค decision tree หลายๆ ต้น

โดยหลักการที่เลือกนำมาใช้สำหรับงานวิจัยชิ้นนี้ คือ vote ensemble โดยเป็นการโหวตผลพยากรณ์ของเทคนิคการทำเหมืองข้อมูล 5 เทคนิค อันได้แก่ เทคนิคต้นไม้ตัดสินใจ (Decision Tree), Naïve Bayesian, โครงข่ายประสาท (Neural Network), ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) และการถดถอยโลจิสติกส์ (Logistic Regression)

2.2.2 ต้นไม้ตัดสินใจ (Decision Tree) [11]

Decision tree หรือต้นไม้ตัดสินใจเป็น 1 ในเทคนิคการทำเหมืองข้อมูลในรูปแบบวิธีการจัดหมวดหมู่ที่รู้จักกันดีที่สุดโดยมักใช้ตรวจสอบข้อมูลและสร้างต้นไม้เพื่อการพยากรณ์ สำหรับโครงสร้างของต้นไม้ตัดสินใจ จะมีลักษณะคล้ายโครงสร้างต้นไม้ทั่วไป โดยการแตกแขนงไปตามเงื่อนไขหรือเส้นทางของกิ่งไม้ และข้อมูลที่คาดคะเนไว้ว่าจะเกิดขึ้น ซึ่งจะใช้กฎในรูปแบบ “ถ้า (เงื่อนไข) แล้ว (ผลลัพธ์)” (If-then Rule) มาประกอบการสร้างโครงสร้างต้นไม้ตัดสินใจ สำหรับโครงสร้างต้นไม้ตัดสินใจจะประกอบด้วย

- โหนดภายใน (Internal Node) คือโหนดที่แสดงถึงคุณลักษณะ (Feature) ที่นำมาใช้ในการแบ่งกลุ่มของข้อมูลซึ่งมีโหนดราก (Root Node) อยู่บนสุดของโครงสร้าง ซึ่งเป็นโหนดที่มีอิทธิพลต่อการจำแนกกลุ่มมากที่สุด
- กิ่ง (Branch) เป็นตัวเชื่อมระหว่างโหนดที่ใช้เป็นเงื่อนไขหรือทางเลือกของการกระทำ ซึ่งมาจากผลลัพธ์ แต่ละตัวของทุกตัวทำนาย (Predictor) หรือคุณสมบัติ (Feature)
- โหนดใบ (Leaf Node) เป็นโหนดที่แสดงผลลัพธ์ของเงื่อนไขหรือการกระทำตามเงื่อนไขที่เกิดขึ้น

สำหรับการสร้าง Decision Tree ของแต่ละอัลกอริทึมนั้นจะมีลักษณะที่คล้ายกันคือ เริ่มต้นทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของต้นไม้ (Root Node) หลังจากนั้นจะทำการแตกกิ่งแอตทริบิวต์ออกไปเรื่อยๆ จนสามารถแบ่งข้อมูลออกเป็นคลาสได้ชัดเจน

2.2.3 Naïve Bayesian [12]

เป็นเทคนิคการทำเหมืองข้อมูลอย่างง่าย โดยนำโมเดลมาใช้ในการคัดแยกประเภทข้อมูลผ่านหลักความน่าจะเป็นที่อยู่บนพื้นฐานของทฤษฎี Bayes และสมมติฐานของการเกิดของเหตุการณ์เป็นอิสระต่อกัน เทคนิค Naïve Bayes เป็นเทคนิคที่ไม่มีการหมุนวนที่ซับซ้อน ส่งผลให้สามารถทำงานได้ดีและมีประโยชน์กับชุดข้อมูลขนาดใหญ่อย่างมาก ทำให้เทคนิคนี้ได้รับความไว้ใจและถูกใช้งานเป็นวงกว้าง สำหรับหลักการคำนวณของเทคนิคนี้ จะอ้างอิงจากทฤษฎีโดยสันนิษฐานว่าผลลัพธ์หรือค่าที่เกิดจากตัวที่ใช้ทำนาย (predictor) เป็นอิสระต่อกัน โดยเขียนเป็นสมการได้ดังนี้

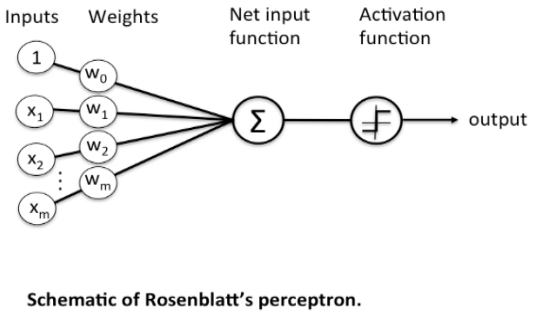
$$P(c \mid x) = P(x \mid c)\, P(c) / P(x) \qquad (1)$$

โดย P(c | x) คือค่าความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส c; P(x | c) คือ ค่าความน่าจะเป็นที่ข้อมูลในชุดข้อมูลสอนที่มีคลาส c และมีแอตทริบิวต์ x โดยที่ $x = x_1 \cap x_2 \ldots \cap x_M$ โดยที่ M คือ จำนวนแอตทริบิวต์; P(c) คือ ค่าความน่าจะเป็นของคลาส C; และ P(x) คือ ค่าความน่าจะเป็นของแอตทริบิวต์ x

2.2.4 โครงข่ายประสาท (Neural Network) [13]

เป็นหนึ่งในเทคนิคการทำเหมืองข้อมูลที่ใช้โมเดลทางคณิตศาสตร์มาประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) โดยได้แนวคิดมาจากการจำลองการทำงานของเซลล์สมองมนุษย์ ที่แต่ละเซลล์ประสาทจะประกอบไปด้วยเดนไดรท์ (Dendrite) หรือปลายในรับกระแสประสาท ซึ่งเป็นตัว input ของเซลล์ และ แอกซอน (Axon) เป็นเสมือน output ของเซลล์ที่จะส่งกระแสประสาทไปยังเซลล์ตัวอื่น ถ้าสมองได้รับกระแสไฟฟ้ามากพอจะทำให้เซลล์ส่งต่อกระแสประสาทไปเรื่อยๆ

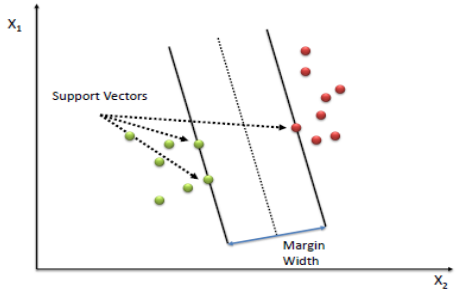
สำหรับโครงสร้างของประสาทเทียมจะประกอบด้วย input และ output เช่นกัน โดยแบ่งเป็นชั้นหรือ layer ซึ่งจะมีชั้นคั่นตรงกลางคือ hidden layer โดยโครงสร้างประสาทเทียมจะมีหน่วยย่อยเรียกว่า perceptron ซึ่งเทียบเท่าได้กับเซลล์สมองของมนุษย์หนึ่งเซลล์



รูปที่ 2: perceptron (source: [14])

2.2.5 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) [15]

เป็นอัลกอริทึมที่นำมาใช้ในการวิเคราะห์และจำแนกข้อมูลอย่างกว้างขวาง โดยอาศัยโมเดลทางคณิตศาสตร์ในเรื่องของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นเลือกเส้นที่แบ่งข้อมูลได้ดีที่สุด โดยไม่จำเป็นต้องเป็นเส้นตรงเสมอไป

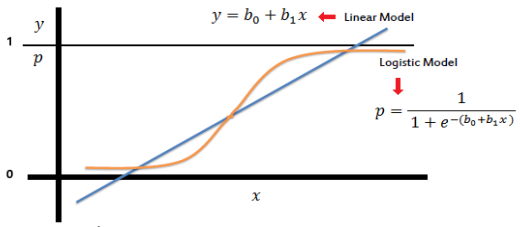


รูปที่ 3: support vector machine (source: [15])

หลักการของอัลกอริทึมคือ การนำข้อมูลที่ใช้สอนมากระจายเป็นเวกเตอร์ในระนาบ หรือ สเปซ N มิติ (Feature Space) จากนั้นคำนวณหาเส้นไฮเปอร์เพลน (Hyperplane) ที่จะแยกกลุ่มของเวคเตอร์อินพุทออกเป็นประเภทต่างๆ โดยเส้นไฮเปอร์เพลนต้องเป็นเส้นที่มีค่า margin หรือระยะห่างระหว่างจุดของข้อมูลที่อยู่ใกล้กับไฮเปอร์เพลนทั้งสองด้าน คือ d+ และ d- มากที่สุด เพราะถ้าค่า margin น้อย แสดงว่าอาจมีจุดหนึ่งของไฮเปอร์เพลนที่อยู่ใกล้ข้อมูลของแต่ละคลาสมากเกินไป ทำให้ข้อมูลใหม่ที่อยู่ห่างออกไปเล็กน้อยเกิดการทำนายผิดพลาดได้ โดยข้อมูลที่อยู่บนขอบของ margin เรียกว่า support vector

2.2.6 การถดถอยโลจิสติกส์ (Logistic Regression) [16]

การถดถอยโลจิสติกส์ (Logistic Regression) เป็นการพยากรณ์ความน่าจะเป็นของผลลัพธ์ที่จะเกิดขึ้นซึ่งสามารถเป็นได้เพียง 2 ค่า เช่น ใช่หรือไม่ใช่ โดยการพยากรณ์จะขึ้นกับตัวแปรที่ส่งผลต่อเหตุการณ์นั้นๆ ซึ่งอาจจะมีเพียงหนึ่งตัวหรือมากกว่า



รูปที่ 4: การถดถอยโลจิสติกส์ (source: [16])

logistic regression สามารถสร้าง curve จากการใช้ลอการิทึม “odds” ในตัวแปรเป้าหมาย ซึ่งจำกัดค่า 0 ถึง 1

2.3 งานวิจัยที่เกี่ยวข้อง

2.3.1 DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE [2]

รายงานการศึกษาฉบับนี้ทำการค้นหาเทคนิคการทำเหมืองข้อมูลที่เหมาะสมต่อการจำแนกโดยเปรียบเทียบจากเทคนิค Bayesian

Classifier, Neural Network และ Decision Tree ซึ่งใช้ตัวแปรที่สำรวจมาว่าส่งผลต่อประสิทธิภาพการเรียนรู้หรือการดำเนินการ รวมทั้งหมด 12 ปัจจัยคือ เพศ ระยะทาง(ระหว่างบ้านกับโรงเรียน) เกรดเฉลี่ย(GPA) ทุนการศึกษา อุปกรณ์การเรียน ความสำคัญของเกรด จำนวนสมาชิกในครอบครัว ประเภทของโรงเรียนมัธยมที่เคยศึกษา คะแนนสอบเข้า เวลาเรียนเฉลี่ยของแต่ละสัปดาห์ อินเทอร์เน็ต เงินที่ได้รับ ซึ่งผลทดสอบปรากฏว่าปัจจัยที่ส่งผลมากสุดคือ เกรดเฉลี่ย (GPA) สำหรับปัจจัยที่ส่งผลน้อยสุดคือจำนวนสมาชิกภายในครอบครัว โดยเทคนิคที่ใช้งานได้ดีที่สุดคือ Naive Bayes ซึ่งให้ผลที่ถูกต้องและสามารถเข้าใจได้ง่าย โดยมีความแม่นยำ (Accuracy) 76.65%

2.3.2 A Review on Predicting Student's Performance using Data Mining Techniques [3]

บทความนี้รวบรวมงานวิจัยซึ่งมีการใช้เทคนิคการทำเหมืองข้อมูลในรูปแบบต่างๆ เพื่อช่วยประเมินการเรียนของนักศึกษา ซึ่งบทความส่วนมากเลือกใช้เทคนิคต้นไม้ตัดสินใจเพื่อค้นหาคุณลักษณะสำคัญที่ใช้ในการทำนายผลการศึกษาของนักศึกษา ปรากฏว่าคุณลักษณะสำคัญที่นำมาใช้ทำนายมากที่สุด คือ เกรดเฉลี่ยรวมสะสม (Cumulative Grades Point Average-CGPA) เนื่องจากว่าเป็นค่าที่มองเห็นและจับต้องได้ อีกทั้งเป็นค่าที่มีอิทธิพล หรือส่งผลต่อตัวอื่นๆ มากสุด ส่วนคุณลักษณะทางประชากร (เช่น เพศ อายุ) ได้นำมาใช้เป็นปัจจัยเช่นเดียวกัน โดยเพศถูกคัดเลือกมาใช้มากที่สุดในบรรดาคุณลักษณะ ซึ่งเพศเป็นสามารถบ่งบอกถึงผลการดำเนินงานได้ เช่น เพศหญิงจะมีการเรียนรู้ที่เป็นระบบมากกว่า

ในงานวิจัยนี้ผู้วิจัยได้ทำการพัฒนาโมเดลพยากรณ์แขนงวิชาที่เหมาะสมโดยการนำเทคนิดการทำเหมืองข้อมูล 5 เทคนิคมาพยากรณ์ร่วมกันภายใต้เทคนิค vote ensemble เพื่อทำการหาความน่าจะเป็นที่นักศึกษาคนหนึ่งจะประสบผลสำเร็จในการเรียนแต่ละแขนงวิชา ซึ่งแตกต่างจากงานวิจัยอื่นๆ ที่ทำกันมาก่อนหน้านี้อยู่สองประเด็นใหญ่ๆ ได้แก่ การใช้เทคนิค vote ensemble ซึ่งจากการทบทวนวรรณกรรม งานวิจัยที่ทำการพยากรณ์ผลการเรียนโดยส่วนมากจะทำการเปรียบเทียบเทคนิดต่างๆ เพื่อหาว่าเทคนิคใดดีที่สุด พร้อมทั้งนำเสนอปัจจัย (feature) ที่ส่งผลกระทบต่อการพยากรณ์ และ อีกประเด็นหนึ่งคือการที่ตัวแปรตาม (dependent variable) ในงานวิจัยชิ้นนั้นมีถึง 8 คลาสคือ เรียนดี-ไม่ดีของแต่ละแขนงวิชา ซึ่งแตกต่างจากงานวิจัยอื่นซึ่งมักจะสนใจทำนายว่านักเรียนจะเรียนจบ-ไม่จบ หรือทำนายเกรดเฉลี่ยเมื่อจบการศึกษา

3.วิธีดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์หลักเพื่อพัฒนาระบบสนับสนุนการตัดสินใจเลือกแขนงวิชาของนักศึกษาสำหรับศึกษาต่อในชั้นปีที่ 3 คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง โดยนำเทคนิคการทำเหมืองข้อมูลมาใช้เป็นเครื่องมือในการพยากรณ์ เพื่อสกัดให้ได้ซึ่งข้อมูลที่เป็นประโยชน์และสามารถนำมาสนับสนุนการตัดสินใจ โดยรายละเอียดขั้นตอนการดำเนินงานวิจัย มีดังนี้

3.1 การออกแบบโครงสร้างการทำงานของระบบ

โครงสร้างและสถาปัตยกรรมของระบบจะแบ่งออกเป็น 3 ส่วนประกอบหลัก ซึ่งแต่ละส่วนประกอบมีการทำงานเข้าด้วยกัน โดยมีรายละเอียดดังนี้

1) ส่วนฐานข้อมูล (Database) เป็นการนำข้อมูลที่เก็บรวบรวมจาก 2 แหล่ง ทั้งหน้าเว็บแอพพลิเคชั่นที่ผู้ใช้ป้อนเข้ามา และส่วนประมวลผลหลังจากการประมวลเสร็จ โดยนำมาจัดเก็บเป็นฐานข้อมูลในรูปแบบ Relational Database Management System (RDMS) เพื่อที่จะนำไปใช้สำหรับการวิเคราะห์ผลในส่วนต่างๆ
2) ส่วนการประมวลผล (Processing) ใช้รวบรวมข้อมูลของนักศึกษาแต่ละบุคคล เพื่อนำไปเป็นปัจจัยที่ใช้สร้างแบบจำลองการพยากรณ์หลายๆเทคนิค เพื่อนำผลการพยากรณ์ทุกเทคนิคมาวิเคราะห์และโหวตคำตอบร่วมกันให้ได้ซึ่งผลลัพธ์สุดท้ายที่ต้องการจะนำไปใช้ประโยชน์และช่วยในการตัดสินใจของผู้ใช้งาน โดยใช้ไลบรารี่ scikit-learn เข้ามาช่วยในการสร้างแบบจำลอง
3) ส่วนการแสดงผลข้อมูล (User Interface) เป็นส่วนที่ใช้ติดต่อกันระหว่างผู้ใช้งานกับระบบ โดยจัดทำในรูปแบบของเว็บไซต์ พัฒนาโดยใช้ HTML, CSS และ JavaScript ภายใต้การจัดการของ Bootstrap Framework

3.2 การออกแบบโมเดลสำหรับการพยากรณ์ผล

3.2.1 ประชากรและกลุ่มตัวอย่าง

ประชากรที่ใช้ในการวิจัยครั้งนี้ คือ นักศึกษาคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังที่ได้รับการศึกษาตามหลักสูตรแขนงวิชา ซึ่งเป็นนักศึกษาชั้นปีการศึกษา 2555- 2557 เนื่องจากหลักสูตรการเรียนตามแขนงวิชา ได้ริเริ่มมาตั้งแต่ปี 2555

สำหรับกลุ่มตัวอย่างที่ใช้ในการวิจัยในครั้งนี้คือ นักศึกษาคณะเทคโนโลยีสารสนเทศชั้นปีที่ 4 (ปีการศึกษา 2556) จำนวน 133 คน ซึ่งมีผลการเรียนวิชาแขนงในชั้นปีที่ 3 ครบถ้วนซึ่งใช้เป็นตัวแปรที่บ่งชี้ว่านักเรียนคนใดทำได้ดีและคนใดที่ทำได้ไม่ได้เมื่อเข้าไปเรียนในแขนงวิชาที่เลือกแล้ว

3.2.2 การเก็บรวบรวมข้อมูล

สำหรับขั้นตอนการจัดเก็บข้อมูล ได้ทำการวางแผนพัฒนาเว็บแอพพลิเคชันเพื่อใช้ในการจัดเก็บผลปัจจัยต่างๆลงในฐานข้อมูลและดำเนินการวิเคราะห์ตามแบบแผนที่วางไว้ โดยข้อมูลที่นำมาใช้ในการวิจัยมาจาก 2 แหล่ง ได้แก่ (1) แหล่งปฐมภูมิ (Primary Data) มาจากการจัดทำแบบทดสอบความถนัดทางสาขาคอมพิวเตอร์ผ่านเว็บแอพพลิเคชันเพื่อวัดประเมินความสามารถทางการเขียนโปรแกรมและคณิตศาสตร์ของนักศึกษาแต่ละคน และ (2) แหล่งทุติยภูมิ (Secondary Data) ประกอบด้วยข้อมูลผลการเรียนและข้อมูลเพศของนักศึกษา โดยทางผู้จัดทำได้ติดต่องานทะเบียนและประมวลผลเพื่อขอความอนุเคราะห์ในการเก็บรวมรวมข้อมูลของกลุ่มตัวอย่างเพื่อใช้ในการวิจัยครั้งนี้

ผู้วิจัยได้ใช้โปรแกรม RapidMiner Studio เวอร์ชั่น 7.5 ในการทดลองสร้างโมเดลด้วยทั้ง 5 เทคนิค และการทำ Vote Emsemble โดยขั้นตอนการดำเนินงานนั้นมีดังนี้ ในขั้นตอนแรกผู้วิจัยได้แบ่งชุดข้อมูลสอนออกเป็น 4 ชุดสำหรับแขนงทั้ง 4 แขนง แต่ละชุดจะมีเฉพาะข้อมูลของนักศึกษาแขนงนั้นๆ สำหรับใช้ในการสร้างโมเดลพยากรณ์ซึ่งประกอบด้วย ข้อมูลเกรดรายวิชาพื้นฐาน 14 ตัว คะแนนแบบทดสอบความถนัดเฉพาะด้าน เพศ และระดับผลการเรียนในแขนงวิชา ซึ่งใช้เป็นตัวแปรทำนาย (Label) เพื่อให้ทราบถึงความเหมาะสมและความเป็นไปได้ของการเรียนแต่ละแขนง จากนั้นแต่ละชุดข้อมูลจะถูกนำไปสร้างโมเดลพยากรณ์โดยใช้เทคนิคการทำเหมืองข้อมูล 5 เทคนิคที่ได้กล่าวไว้ข้างต้น (ดังนั้นจะได้โมเดลพยากรณ์ทั้งหมด 20 โมเดล) ซึ่งผลการทำนายที่ได้จากแต่ละเทคนิคจะมี 2 ค่า คือ ดีและไม่ดี จากนั้นจึงนำผลมาโหวตดูว่าคำตอบส่วนมากคืออะไร ซึ่งในระบบนี้ได้พิจารณาเฉพาะความน่าจะเป็นที่จะเรียนแขนงใดๆ ประสบผลสำเร็จ โดยดูจากจำนวนเทคนิคที่แสดงผลออกมาว่าเรียนแขนงนั้นแล้วเป็นผลสำเร็จ เช่น ถ้า 4 ใน 5 เทคนิคแสดงผลว่าสำเร็จ ผลที่ได้จะเป็น 80% ซึ่งหมายถึงมีความน่าจะเป็น 80% ที่จะเรียนแขนงนี้ประสบผลสำเร็จ ดังนั้น เมื่อนำข้อมูลของนักศึกษาที่ต้องการทำนายผล ผ่านเข้าสู่กระบวนการพยากรณ์ทั้งหมด ผลที่ได้จะทำให้นักศึกษาทราบถึงความน่าจะเป็นที่จะประสบผลสำเร็จของการเรียนแต่ละแขนง และ แขนงวิชาที่เหมาะสมกับความถนัดของตนเองมากที่สุดไปยังน้อยที่สุด ไล่เรียงเป็นอันดับ

ตารางที่ 1. แสดงความแม่นยำในการทำนายผลของ 5 เทคนิค

Track/Technique	Decision Tree	Naïve Bayes	Neural Network	Support Vector Machine	Logistic Regression
Software	81.67%	83.00%	71.67%	80.00%	80.83%
Network	63.33%	70.00%	50.00%	65.00%	70.00%
Multimedia	70.83%	71.67%	57.50%	70.83%	64.17%
Business	60.83%	65.00%	71.67%	68.33%	59.17%

3.2.3 การวิเคราะห์ข้อมูลและออกแบบตัวโมเดล

สำหรับในการพัฒนาระบบได้ทำการเลือกใช้เทคนิค vote ensemble ซึ่งเป็นวิธีการที่ใช้ชุดข้อมูลสอนชุดเดียวกัน มาสร้างโมเดลจากเทคนิคการพยากรณ์จาก 5 เทคนิค เพื่อให้โมเดลมีความหลากหลายมากขึ้น ซึ่งเทคนิคที่เลือกใช้ ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree), Naïve Bayesian, โครงข่ายประสาท (Neural Network), ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine), และ การถดถอยโลจิสติกส์ (Logistic Regression)โดยเหตุผลที่เลือกทั้ง 5 เทคนิคนี้ก็เนื่องจากทั้ง 5 เทคนิคนี้สามารถใช้ได้กับตัวแปรที่เป็นหมวดหมู่ (Categorical Variable) ซึ่งช่วยให้สามารถจำแนกได้ว่านักศึกษาควรอยู่ในกลุ่มหรือแขนงใด และจากการทบทวนวรรณกรรมของผู้เขียน ผู้เขียนมีความรู้สึกว่าทั้ง 5 เทคนิคเป็นที่นิยมใช้ในการทำนายผลการเรียน

โดยผลที่ได้จากการทำการทดลองโดยซอร์ฟแวร์ RapidMiner นั้นจะเป็นผลการตั้งค่าที่ดีที่สุด โดยดูจากผลการทดลองที่ได้ และจะถูกนำมาใช้ในการตั้งค่าการสร้างโมเดลพยากรณ์ในระบบต่อไป

3.2.4 การทดสอบความแม่นยำและความน่าเชื่อถือของโมเดล

ในขั้นตอนนี้จะเป็นการนำผลลัพธ์ที่ได้จากกระบวนการพยากรณ์ มาทำการทดสอบความถูกต้องแม่นยำ และความน่าเชื่อถือของโมเดลนั้นๆ โดยในงานวิจัยครั้งนี้ จะนำผลลัพธ์แท้จริงของข้อมูลที่ใช้ทดสอบ (Testing Data) มาเปรียบเทียบกับข้อมูลผลลัพธ์ที่ได้จากโมเดลการ

ทำนาย โดยในการวัดประสิทธิภาพของข้อมูลนั้นจะอาศัย Confusion Matrix เพื่อค้นหาความแม่นยำ

4. ผลการทดลอง

เมื่อทดสอบความแม่นยำในการทำนายผลของแต่ละเทคนิคในทุกแขนงวิชาดังตารางที่ 1 พบว่าความแม่นยำของการทำนายในทุกๆ เทคนิคมีค่ามากกว่า 50 % ทำให้สามารถนำเทคนิค vote ensemble มาใช้ในการพยากรณ์แขนงวิชาที่เหมาะสมได้ เนื่องจากเทคนิค Vote Ensemble จะสามารถพยากรณ์ได้ดีและมีความน่าเชื่อถือเมื่อทุกๆเทคนิคมีความแม่นยำอย่างต่ำ 50 % หลังจากนั้นจึงนำผลการพยากรณ์ที่ได้จากแต่ละเทคนิคมาร่วมโหวต เพื่อให้ทราบถึงความน่าจะเป็นในการเรียนแขนงนั้นๆให้ประสบผลสำเร็จ โดยการทดสอบผลจะใช้วิธี cross-validation แบบวนรอบทั้งหมด 10 ครั้ง เช่นเดียวกับการตรวจหาความแม่นยำของแต่ละเทคนิค และแสดงค่าความแม่นยำโดยรวม รวมถึงค่า precision และ recall ดังตารางที่ 2

ตารางที่ 2. แสดงความแม่นยำของการทำนายโดย Vote Ensemble

Track/Performance	Accuracy	Precision	Recall
Software	86.67%	82.35%	87.50%
Network	70.00%	73.33%	73.33%
Multimedia	74.17%	76.92%	66.67%
Business	60.83%	59.09%	68.42%

จะเห็นได้ว่าแขนงวิชาวิศวกรรมซอฟต์แวร์ (Software) มีความแม่นยำสูงถึง 86.67% รองลงมาคือแขนงวิชาการพัฒนาสื่อประสมและเกม (Multimedia) แขนงเทคโนโลยีและเครือข่ายระบบ (Network) และแขนงวิชาอัจริยะทางธุรกิจ (Business) ทำนายได้แม่นยำน้อยที่สุดเพียง 60.83%

5. อภิปรายผลการทดลอง

จากผลการทดลอง จึงสรุปได้ว่าเทคนิคที่พยากรณ์ผลได้แม่นยำมากที่สุดคือ เทคนิค Naïve Bayes รองลงมาคือเทคนิค ต้นไม้ตัดสินใจ และซัพพอร์ตเวกเตอร์แมชชีน ต่อมาคือ เทคนิคการถดถอยโลจิสติกส์ และเทคนิคที่แม่นยำน้อยสุดคือเทคนิคโครงข่ายประสาท และเมื่อมองในส่วนแขนงวิชาร่วมด้วย จะพบว่าแขนงวิชา software ให้ความแม่นยำมากสุด โดยความแม่นยำที่กล่าวถึงในที่นี้ หมายความว่าระบบสามารถพยากรณ์ได้ตรงกับระดับประสิทธิภาพของแขนงวิชาที่นักศึกษาเรียนอยู่ ซึ่งผู้วิจัยมีความเห็นว่าการที่ความแม่นยำในการทำนาย รวมถึงค่า precision และ recall มีความแตกต่างกันในแต่ละแขนงวิชาอาจจะเกิดจากลักษณะของกลุ่มบุคคลในแต่ละแขนงวิชากล่าวคือ ถ้าบุคคลที่มีลักษณะใกล้เคียงกันจำนวนมากมาเรียนในแขนงเดียวกันแล้วมีผลสัมฤทธิ์ทางการเรียนที่เหมือนกัน ย่อมส่งผลให้เกิดเอกลักษณ์ของผู้ที่เรียนในแขนงวิชานี้ และทำให้ระบบทำนายได้แม่นยำมากกว่า อย่างเช่น บุคคลที่อยู่ในแขนง Software อาจมีลักษณะคล้ายกันคือ ได้คะแนนเต็มด้านการเขียนโปรแกรม หรือ เกรดวิชาโปรแกรมมิ่งอยู่ในระดับดี ทำให้พยากรณ์ได้ง่ายกว่า ซึ่งต่างจากแขนงวิชา Business ที่มีลักษณะข้อมูลที่ค่อนข้างแตกต่างและกระจายในแต่ละบุคคลจึงพยากรณ์ได้แม่นยำน้อยกว่า ทั้งนี้ ถ้าในอนาคตมีปริมาณชุดข้อมูลสอนที่มากขึ้นหรือมีปัจจัยอื่นๆที่บ่งเฉพาะความถนัดเฉพาะแขนงวิชาที่มากขึ้นก็จะช่วยลดความแตกต่างของข้อมูลและทำให้พยากรณ์ได้แม่นยำมากยิ่งขึ้น

6. ข้อสรุปและข้อเสนอแนะ

โครงงานนี้จัดทำขึ้นเพื่อนำเสนอระบบที่ช่วยสนับสนุนการตัดสินใจการคัดเลือกแขนงวิชาของนักศึกษาคณะเทคโนโลยีสารสนเทศ สำหรับการพยากรณ์ผลการทำนายแขนงวิชาเราได้ใช้หลักการโหวตภายใต้เทคนิค vote ensemble ซึ่งได้ทำการพัฒนาโมเดลการคัดเลือกแขนงของนักศึกษาจากเทคนิค 5 เทคนิคอันได้แก่ Decision Tree, Naïve Bayesian, Neural Network, Support Vector Machine และ Logistic Regression จากผลการทดลองพบว่าการทำนายมีความแม่นยำในระดับที่ต่ำกว่าที่คาดหวังไว้ ซึ่งอาจเนื่องด้วยข้อจำกัดด้านข้อมูลกลุ่มตัวอย่างที่ใช้ในการสร้างแบบจำลองมีปริมาณน้อย อย่างไรก็ตามทางผู้พัฒนามีแผนที่จะเก็บข้อมูลเพิ่มเติมในปีการศึกษาหน้า ซึ่งเมื่อมีข้อมูลกลุ่มตัวอย่างเพิ่มขึ้นก็น่าจะส่งผลให้การพยากรณ์มีแม่นยำมากยิ่งขึ้น ข้อจำกัดในอีกประการหนึ่งสำหรับตัวระบบคือ ระบบสามารถใช้ได้เฉพาะกับนักศึกษาคณะเทคโนโลยีสารสนเทศ สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง หลักสูตร 4 แขนงวิชา

สำหรับประโยชน์จากการทำระบบแนะนำแขนงวิชานี้สามารถนำไปประยุกต์ใช้ร่วมกับงานอื่นๆได้หลายแนวทาง อาทิเช่น การนำไปใช้ในการแนะนำภาคหรือแขนงวิชาของการศึกษาสาขาวิชาอื่นๆได้ หรือการปรับเปลี่ยนพัฒนาเพื่อใช้ในการจัดการทรัพยากรบุคคล หรือประยุกต์ใช้กับงานวิจัยที่เกี่ยวข้องกับการศึกษาอื่นๆได้เป็นต้น

เอกสารอ้างอิง

[1] G.S. Abu-Oda and ,A. M. El-Harees, “DATA MINING IN HIGHER EDUCATION: UNIVERSITY STUDENT DROP OUT

CASE STUDY," *International Journal of Data Mining & Knowledge Management Process,* vol. 5, no. 1, Jan 2015.

[2] E. Osmanbegović and M. Suljic, "DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE," *Journal of Economics and Business,* vol. 10, no. 1, 2012

[3] A. M. Shahiri, W. Husain and N. A. Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques," in *The Third Information Systems International Conference*, 2015.

[4] B.C. Wilson and S. Shrock, " Contributing to Success in an Introductory Computer Science Course: A Study of Twelve Factors," in Proceedings of *thirty-second SIGCSE technical symposium on Computer Science Education,* New York, 2001, pp. 184-18

[5] B. C. Wilson, "A Study of Factors Promoting Success in Computer Science Including Gender Differences, " *Computer Science Education,* vol. 12, no. 1-2, pp. 141-164, 2002.

[6] A. Paloheimo, and J.Stenman, "Gender, Communication and Comfort Level in Higher Level Computer Science Education" – Case Study in Proc. Frontiers in Education. 36th Annual Conference, San Diego, CA, Oct. 2006, pp. 13-18

[7] Y. ERDOGAN, E. AYDIN, and T. KABACA, "Identifying Predictors of Programming Achievement" in Conf. 6th WSEAS International Conference on EDUCATION and EDUCATIONAL TECHNOLOGY, Venice, Italy, Nov. 2007, pp. 600-115

[8] J. Han and M. Kamber, Data mining concepts and techniques (2nd ed.)., Diane Cerra, 2006.

[9] U. Fayyad, P. Smyth and G. P. Shapiro, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine,* vol. 17, no. 3, 1996

[10] EAKASIT PACHARAWONGSAKDA, Data Mining Trend. "การสร้างโมเดล Ensemble แบบต่างๆ" [Online]. Available: http://dataminingtrend.com/2014 /data-mining-techniques/ ensemble-model/

[11] ชินพัฒน์ แก้วชินพร. การจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจและการจัดกลุ่ม[ปริญญาวิทยาศาสตร์บัณฑิต]. กรุงเทพฯ :สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง; 2553.

[12] Saed Sayad. "Naive Bayesian" [Online]. Available: http://www.saedsayad.com/naive_ bayesian.htm

[13] Wittaya Pornpatcharapong. "โครงข่ายประสาทเทียม (Artificial Neural Networks – ANN)" [Online]. Available: https://www.gotoknow.org/posts /163433

[14] Sebastian Raschka. "Single-Layer Neural Networks and Gradient Descent" [Online]. Available: http://sebastianraschka.com/Articles/2015_singlelayer_ neurons.html

[15] Saed Sayad. "Support Vector Machine - Classification (SVM)" [Online]. Available: http://www.saedsayad .com/support_ vector _ machine.htm

[16] Saed Sayad. "Logistic Regression" [Online]. Available: http://www.saedsayad.com/ logistic_ regression.htm