การเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ เพื่อวิเคราะห์ปัจจัยที่ส่งผลกับการเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐม โดยใช้เทคนิค ENSEMBLE

The Comparison of Model Efficiency to Forecast the Number of New Students for Analysis of Factors Affecting on Admission to Nakhon Pathom Rajabhat University by using ENSEMBLE

พงษ์ดนัย จิตตวิสุทธิกุล[1] และจรัญ แสนราช[2]

[1]*สาขาวิชาคอมพิวเตอร์ศึกษา คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม*
[2]*ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*
vazabizatan@gmail.com

บทคัดย่อ

ในปัจจุบันการรับสมัครสอบคัดเลือกนักศึกษาใหม่ของแต่ละสถาบันในระดับอุดมศึกษาได้มีรูปแบบเป็นนโยบายเชิงรุกมากยิ่งขึ้น เนื่องจากการลดลงของจำนวนนักเรียนที่มีอยู่ในระบบการศึกษา การประชาสัมพันธ์หลักสูตรและการแนะแนวตามโรงเรียนต่าง ๆ ให้ตรงกับกลุ่มเป้าหมายจึงเป็นสิ่งที่มีความสำคัญเป็นอย่างยิ่ง ผู้วิจัยจึงได้นำเทคนิคเหมืองข้อมูล (Data Mining) มาใช้วิเคราะห์ข้อมูลของผู้สมัครที่สอบผ่านการคัดเลือกเป็นนักศึกษาใหม่ระดับปริญญาตรี มหาวิทยาลัยราชภัฏนครปฐม ระหว่างปีการศึกษา 2559-2560 โดยศึกษาวิธีการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ด้วยเทคนิค ENSEMBLE ทั้ง 3 เทคนิค คือ Vote Ensemble, Bootstrap Aggregating (Bagging) และ Random Forest เพื่อนำอัลกอริทึมที่มีประสิทธิภาพสูงสุดมาใช้ในการวิเคราะห์ข้อมูล และนำผลที่ได้จากการวิเคราะห์ข้อมูลไปประกอบการวางแผนการรับนักศึกษาใหม่ในปีการศึกษาถัดไปได้อย่างมีประสิทธิภาพ

ผลการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ พบว่าการสร้างตัวแบบพยากรณ์ด้วยเทคนิค Bootstrap Aggregating โดยใช้อัลกอริทึม Rule Induction มีค่าประสิทธิภาพความถูกต้องสูงสุดเท่ากับ 83.96% สูงกว่าเทคนิค Vote Ensemble และ Random Forest ซึ่งมีค่าเท่ากับ 83.19% และ 81.95% ตามลำดับ และผลการวิเคราะห์ปัจจัยที่ส่งผลกับการเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐมมากที่สุด คือ รูปแบบการสอบคัดเลือก เกรดเฉลี่ยรวม เพศ โรงเรียนที่สำเร็จการศึกษา จังหวัด สายการเรียน และสาขาวิชาที่เลือก ตามลำดับ

คำสำคัญ: ประสิทธิภาพ ตัวแบบพยากรณ์ เทคนิค ENSEMBLE

Abstract

Nowadays, a recruitment examination of new students for each institution of higher education has a more aggressive policy toward the decrease in the number of students available in the education system. Publicity and curriculum counseling according to various schools, matching people is very important. Researchers have used data mining techniques to analyze the data of applicants who have passed the

selection of new undergraduate students Nakhon Pathom Rajabhat University during the academic year 2015-2016. By focused on the comparative accuracy of models to forecast number of new students using 3 techniques of ENSEMBLE: Vote Ensemble, Bootstrap Aggregating (Bagging) and Random Forest to implement the most efficient algorithm for data analysis and the results from the data analysis to the planning of new students in the next academic year effectively.

The result found that the model was created by Bootstrap Aggregating with Rule Induction algorithms had the highest accuracy at 83.96% higher than the Vote Ensemble and Random Forest techniques, 83.19% and 81.95%, respectively and factors affecting admission to NPRU was type of examination.

Keywords: efficiency, model to forecast, ENSEMBLE technique

1. บทนำ

ในปัจจุบันการรับสมัครนักศึกษาใหม่ของแต่ละสถาบันอุดมศึกษาได้มีรูปแบบเป็นนโยบายเชิงรุกและมีการแข่งขันกันมากขึ้น เนื่องจากการลดลงของจำนวนนักเรียนในระบบการศึกษา ซึ่งเป็นผลมาจากอัตราการเกิดของประชากรในประเทศไทยลดน้อยลง การรับนักศึกษาใหม่ของมหาวิทยาลัยราชภัฏนครปฐมนั้นมีการสอบคัดเลือกอยู่ 2 รูปแบบ คือ การสอบคัดเลือกด้วยวิธีการสอบตรง จำนวน 3 รอบ และการสอบคัดเลือกแบบโควตา เพื่อให้มหาวิทยาลัยได้นักศึกษาใหม่ที่มีความรู้ ความสามารถความถนัดตรงตามหลักสูตรที่ต้องการศึกษา โดยมีองค์ประกอบในการพิจารณาคัดเลือกนักศึกษาจากผลรวมของคะแนนสอบวิชาศึกษาทั่วไป และวิชาชีพเฉพาะของแต่ละหลักสูตร การประชาสัมพันธ์หลักสูตรและการแนะแนวตามโรงเรียนต่าง ๆ ให้ตรงกับกลุ่มเป้าหมายจึงถือได้ว่าเป็นกระบวนการที่มีความสำคัญเป็นอย่างยิ่งที่จะก่อให้เกิดกระบวนการรับนักศึกษาใหม่ได้อย่างคุ้มค่ากับทรัพยากร และมีประสิทธิภาพสูงสุด

การนำเทคนิคเหมืองข้อมูล (Data Mining) มาใช้วิเคราะห์ข้อมูลในอดีต ปัจจุบัน เพื่อพยากรณ์สิ่งที่จะเกิดขึ้นในอนาคตนั้นเป็นสิ่งที่กำลังมีบทบาทเป็นอย่างมากไม่ว่าจะเป็นเรื่องของธุรกิจ สภาพแวดล้อม สุขภาพของมนุษย์ รวมไปถึงด้านอื่น ๆ ซึ่งเทคนิคในการวิเคราะห์หรือการพยากรณ์นั้นมีอยู่มากมาย เทคนิค Ensemble ก็เป็นหนึ่งในเทคนิคดังกล่าวที่มีประสิทธิภาพสูง ดังจะเห็นได้จากงานวิจัยเรื่อง Boosting-based ensemble learning with penalty profiles for automatic Thai unknown word recognition (Jakkrit, Cholwich and Thanaruk, 2012) โดยใช้โมเดล Classification หลาย ๆ โมเดลในการวิเคราห์พยากรณ์ ซึ่งสามารถแบ่งออกได้ 3 ประเภท คือ Vote Ensemble, Bootstrap Aggregating (Bagging) และ Random Forest

ดังนั้นผู้วิจัยจึงควรที่จะนำข้อมูลของผู้สมัครที่สอบผ่านการคัดเลือกเป็นนักศึกษาใหม่ระดับปริญญาตรี มหาวิทยาลัยราชภัฏนครปฐม ระหว่างปีการศึกษา 2559-2560 ที่สำเร็จการศึกษาจากสถานศึกษาในจังหวัดนครปฐม และที่มีรอยต่อติดกับจังหวัดนครปฐมเท่านั้น จำนวน 11,338 ชุด มาวิเคราะห์โดยการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ด้วยเทคนิค ENSEMBLE ทั้ง 3 เทคนิค คือ Vote Ensemble, Bootstrap Aggregating (Bagging) และ Random Forest เพื่อนำอัลกอริทึมที่มีประสิทธิภาพสูงสุดมาใช้ในการวิเคราะห์ข้อมูลเพื่อพยากรณ์จำนวนนักศึกษาใหม่ และวิเคราะห์ปัจจัยที่ส่งผลต่อการมารายงานตัวเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐม เพื่อนำผลที่ได้จากการวิเคราะห์ข้อมูลไปประกอบการวางแผนการรับนักศึกษาใหม่ในปีการศึกษาถัดไปได้อย่างมีประสิทธิภาพ

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ของมหาวิทยาลัยราชภัฏนครปฐม โดยใช้เทคนิค Ensemble

2.2 เพื่อวิเคราะห์ปัจจัยที่ส่งผลกับการเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐม โดยใช้อัลกอริทึมที่มีประสิทธิภาพสูงสุด

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 เอกสารที่เกี่ยวข้อง

3.1.1 เหมืองข้อมูล

เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่งซึ่งมาจากคำว่า “เหมืองข้อมูล” นั่นคือ เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่าง ๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลแล้วจะได้สิ่งที่เป็นประโยชน์ เช่น ลูกค้าส่วนใหญ่ที่ซื้อเบียร์มักจะซื้อผ้าอ้อมด้วย จะเห็นว่าข้อมูลนี้เป็นข้อมูลที่ไม่เคยคิดว่ามีความสัมพันธ์กัน และเมื่อได้ความรู้แบบนี้ก็อาจจะนำเป็นออกโปรโมชั่นหรือช่วยในการจัดวางชั้นสินค้า หรือเป็นแนวทางในการสั่งซื้อสินค้าในซุปเปอร์มาร์เก็ตต่อไปได้ นอกจากนี้การทำเหมืองข้อมูลยังมีเทคนิคในการประยุกต์ใช้งานได้อย่างดี เช่น เทคนิคการแบ่งกลุ่มข้อมูล โดยข้อมูลที่มีลักษณะคล้าย ๆ กัน อยู่กลุ่มเดียวกัน และข้อมูลที่อยู่คนละกลุ่มจะมีลักษณะที่แตกต่างกันมาก แต่ละกลุ่มจะเรียกว่าคลัสเตอร์ (Cluster) ซึ่งมีหลายเทคนิค และ Clustering Validity เป็นการวัดประสิทธิภาพของ Clustering เพื่อดูว่าเทคนิคใดสามารถทำให้การแบ่งกลุ่มมีประสิทธิภาพสูงที่สุด และควรจัดข้อมูลออกมาเป็นกี่กลุ่ม เทคนิคการจำแนกประเภทข้อมูลเป็นการนำข้อมูลเดิมที่มีคำตอบที่เราสนใจมาสร้างเป็นโมเดลเพื่อหาคำตอบให้กับข้อมูลใหม่ การประมาณค่าข้อมูล (Regression) การสร้างโมเดลและการวัดประสิทธิภาพของโมเดล โดยการดูค่าความแม่นยำว่า โมเดลใดให้ความแม่นยำในการทายข้อมูลได้ถูกมากที่สุด ดังนั้นหากข้อมูลใดที่มีขนาดใหญ่หรือมีจำนวนมาก การทำเหมืองข้อมูลก็จะเป็นเทคนิคหนึ่งที่จะช่วยในการจัดการข้อมูลให้เป็นประโยชน์ได้ดี (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557) ซึ่งในการวิจัยครั้งนี้ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลของผู้สมัครที่สอบผ่านการคัดเลือกเป็นนักศึกษาใหม่ระดับปริญญาตรี มหาวิทยาลัยราชภัฏนครปฐม ระหว่างปีการศึกษา 2559-2560 ที่สำเร็จการศึกษาจากสถานศึกษาในจังหวัดนครปฐม และที่มีรอยต่อติดกับจังหวัดนครปฐมเท่านั้น จำนวน 11,338 ชุด 8 แอทริบิวต์ ได้แก่ เพศ, รูปแบบการสอบคัดเลือก, เกรดเฉลี่ยรวม, โรงเรียนที่สำเร็จการศึกษา, จังหวัด, สายการเรียน, สาขาวิชาที่เลือก และสถานะการรายงานตัวเข้าศึกษาต่อ

3.1.2 กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM

กระบวนการมาตรฐานในการวิเคราะห์ข้อมูลด้านการทำเหมืองข้อมูลพัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือกันของ 3 บริษัท คือ DaimlerChrysler, SPSS และ NCR กระบวนการทำงานนี้เรียกว่า Cross-Industry Standard Process for Data Mining หรือเรียกย่อว่า CRISP-DM ซึ่งประกอบด้วย 6 ขั้นตอน ดังนี้ (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

1) ความเข้าใจในธุรกิจ (Business Understanding) เป็นขั้นตอนแรกในกระบวนการ CRISP-DM ซึ่งเน้นไปที่การเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางการทำเหมืองข้อมูล พร้อมทั้งวางแผนในการดำเนินงานคร่าว ๆ

2) ความเข้าใจในข้อมูล (Data Understanding) ขั้นตอนนี้เริ่มจากการเก็บรวบรวมข้อมูล หลังจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้ เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์

3) การเตรียมข้อมูล (Data Preparation) เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (Raw Data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจต้องมีการทำข้อมูลให้ถูกต้อง (Data Cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (Scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป เป็นต้น โดยขั้นตอนนี้จะเป็นขั้นตอนที่ใช้เวลามากที่สุดของกระบวนการ CRISP-DM

4) การจัดทำตัวแบบ (Modeling) ขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางการทำเหมืองข้อมูล เช่น การจำแนกประเภทข้อมูล หรือ การแบ่งกลุ่มข้อมูล ซึ่งในขั้นตอนนี้หลายเทคนิคจะถูกนำมาใช้เพื่อให้ได้คำตอบที่ดีที่สุด ดังนั้นในบางครั้งอาจจะต้องมีการย้อนกลับไปที่ขั้นตอนการเตรียมข้อมูล เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย

5) การประเมินผล (Evaluation) ขั้นตอนนี้จะได้ผลการวิเคราะห์ข้อมูลด้วยเทคนิคทางการทำเหมืองข้อมูลแล้ว แต่ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานต่อไปก็ต้องมีการวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรกหรือ มีความน่าเชื่อถือมากน้อยเพียงใด ซึ่งอาจจะย้อนกลับไปยังขั้นตอนก่อนหน้าเพื่อเปลี่ยนแปลงแก้ไขเพื่อเปลี่ยนแปลงแก้ไขให้ได้ผลลัพธ์ตามที่ต้องการได้

6) การนำเอาตัวแบบไปใช้งาน (Deployment) ในกระบวนการทำงานของ CRISP-DM นั้นไม่ได้หยุดเพียงแค่ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลเท่านั้น แม้ว่าผลลัพธ์ที่ได้จะแสดงถึงองค์ความรู้ที่มีประโยชน์ แต่จะต้องนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริงในองค์กรหรือบริษัท

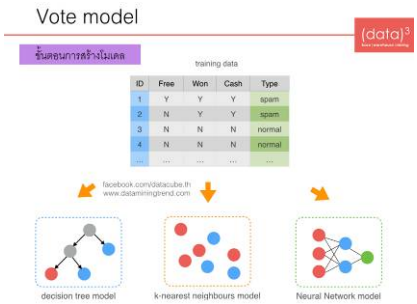
3.1.3 การจำแนกประเภทข้อมูล (Classification)

เป็นกระบวนการจัดแบ่งข้อมูลตามลักษณะของวัตถุประสงค์นั้น ๆ ด้วยการวิเคราะห์เซ็ตของกลุ่มข้อมูล (Data Object) ที่ยังไม่ได้จัดแบ่งประเภท เพื่อสร้างโมเดลจัดการข้อมูลให้อยู่ในรูปชุดข้อมูล (Class) ที่กำหนด โดยจะนำข้อมูลส่วนหนึ่งจากข้อมูลทั้งหมดเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ (Training Data) เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ ผลลัพธ์ที่ได้จากการเรียนรู้ คือ โมเดลจัดประเภทข้อมูล (Classifier Model) แล้วจึงนำข้อมูลส่วนที่เหลือมาใช้สำหรับทดสอบ (Testing Data) และนำผลที่ได้มาใช้เปรียบเทียบกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง โดยกฎการจำแนกที่ได้มีรูปแบบ IF <Conditions> THEN <Class> หรือถ้า <เงื่อนไข> แล้ว <คลาส> นั่นเอง (Jaiwei Han and Micheline Kamber, 2006)

3.1.4 เทคนิค ENSEMBLE

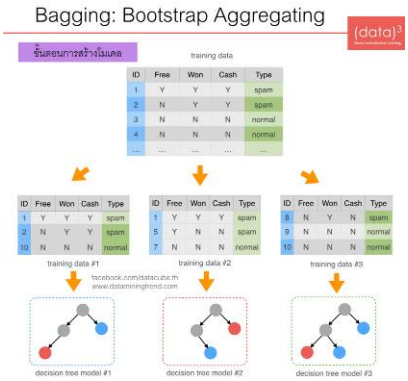
หลักการสร้างโมเดล Ensemble คือ โมเดลที่สร้างควรจะมีความหลากหลายเพื่อให้ทำนายข้อมูลแบบต่าง ๆ กัน การสร้างโมเดลที่หลากหลายนี้ สามารถทำได้โดยการใช้เทคนิค Classification หลาย ๆ ประเภท หรือการสร้าง Training Data ที่มีลักษณะต่าง ๆ โดยเทคนิค Ensemble สามารถจำแนกออกเป็น 3 ประเภท ดังนี้ (https://goo.gl/JDiYLO, 2560)

1) Vote Ensemble เป็นการใช้ Training Data ชุดเดียวกันแต่สร้างโมเดลด้วยเทคนิคต่าง ๆ กัน ดังแสดงในภาพที่ 1



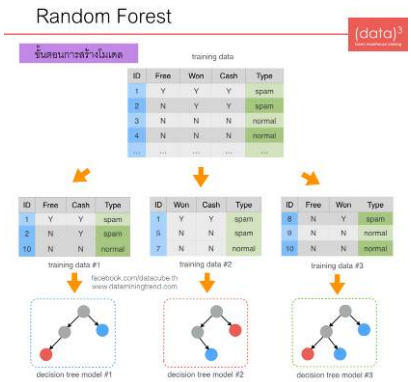
ภาพที่ 1 Vote Ensemble

2) Bootstrap Aggregating (Bagging) เป็นการสุ่ม Training Data ให้เป็นหลาย ๆ ชุด แต่สร้างโมเดลด้วยเทคนิคเดียวกันทั้งหมด ดังแสดงในภาพที่ 2



ภาพที่ 2 Bootstrap Aggregating (Bagging)

3) Random Forest เป็นเทคนิคที่คล้าย ๆ กับ Bagging แต่แทนที่จะสุ่มข้อมูลอย่างเดียวก์ทำการสุ่มเลือกแอตทริบิวต์ต่าง ๆ ออกมาเป็นหลาย ๆ ชุดด้วย และสร้างโมเดลด้วยเทคนิคต้นไม้ตัดสิน (Decision Tree) หลาย ๆ ต้น ดังแสดงในภาพที่ 3



ภาพที่ 3 Random Forest

3.2 งานวิจัยที่เกี่ยวข้อง

ธีรพงษ์ สังข์ศรี (2557). ได้ทำการศึกษาวิจัยเรื่องการวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิคการทำเหมืองข้อมูล ผลการวิจัยพบว่าผลการ

ดำเนินงานได้โดยแบ่งเป็น 2 ส่วน ดังนี้ 1) การเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ การวัดประสิทธิภาพของตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ด้วยเทคนิคต้นไม้ตัดสินใจและโครงข่ายประสาทเทียมจะ ใช้ค่าความถูกต้อง (Accuracy) ในการทำนายผล โดยมีการทดสอบตัวแบบพยากรณ์ด้วยวิธีการแบ่งชุดข้อมูลเท่ากับ 70:30 ผลการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ ต้นไม้ตัดสินใจมีความถูกต้อง ร้อยละ 93.76 โครงข่ายประสาทเทียมมีความถูกต้อง ร้อยละ 93.60 แสดงให้เห็นว่าตัวแบบพยากรณ์จำนวนยอดของนักศึกษาใหม่ที่ถูกสร้างขึ้นด้วยเทคนิคต้นไม้ตัดสินใจมีค่าความถูกต้องเท่ากับ 93.76% ซึ่งสูงกว่าตัวแบบพยากรณ์ที่ถูกสร้างขึ้นด้วยโครงข่ายประสาทเทียมเพียงเล็กน้อย

ธาดา จันตะคุณ (2559). ได้ทำการศึกษาวิจัยเรื่องตัวแบบการจำแนกการเลือกหลักสูตรการศึกษา คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม โดยใช้เทคนิคเหมืองข้อมูล สรุปผลการดำเนินการวิจัยครั้งนี้โดยพัฒนาและเปรียบเทียบ ตัวแบบการจำแนกทั้ง 4 เทคนิค ได้แก่ Decision Tree, Naïve Bayes, k-NN, Rule Induction ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ Decision Tree ซึ่งได้ค่าที่สูงที่สุดจากการแบ่งข้อมูลทดสอบออกเป็น 10 ชุด ค่าความถูกต้องได้ 83.97% จึงสรุปได้ว่า Decision Tree เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำไปใช้ประชาสัมพันธ์หลักสูตร

อัจฉราภรณ์ จุฑาผาด (2556). ได้ศึกษาวิจัยเรื่องการพัฒนาระบบสารสนเทศเพื่อการพยากรณ์จำนวนนักศึกษาใหม่ โดยใช้กฎการจำแนกต้นไม้ตัดสินใจ ผลการวิจัยพบว่า 1) การพัฒนาระบบสารสนเทศเพื่อใช้ในการพยากรณ์จำนวนนักศึกษาใหม่ที่จะเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและการบัญชี มหาวิทยาลัยราชภัฏร้อยเอ็ด โดยใช้กฎการจำแนกเทคนิคต้นไม้ตัดสิน ซึ่งจากการพัฒนาตัวต้นแบบในการพยากรณ์แบ่งออกเป็น 3 วิธีได้แก่ การตรวจสอบไขว้ การแบ่งข้อมูลแบบสุ่มด้วยการ แบ่งร้อยละ และการแบ่งชุดข้อมูลและการทดสอบออกจากกัน แสดงว่าวิธีการแบ่งชุดข้อมูล และการทดสอบออกจากกันสามารถนำไปใช้ในการพัฒนาตัวต้นแบบในการพยากรณ์นักศึกษาใหม่ โดยใช้กฎการจำแนกเทคนิคต้นไม้ตัดสินใจที่มีความถูกต้องแม่นยำสูง และเหมาะสมกว่าวิธีอื่น การนำชุดข้อมูลของผู้สมัครในปีถัดไปมาทดสอบกับตัวต้นแบบหรือกฎที่ได้จะช่วยให้สามารถทราบจำนวนนักศึกษาใหม่ในปีนั้น ๆ ได้ 2) การวัดค่าประสิทธิภาพต่างๆที่วัดได้ จะมีค่าใกล้เคียงกันหรือมีค่าเท่ากันในบางตัวต้นแบบ โดยตัวต้นแบบการพยากรณ์ที่พัฒนาด้วยวิธีการแบ่งชุดข้อมูล และการทดสอบออกจากกัน วัดค่าความถูกต้องได้เท่ากับร้อยละ 97.34 ค่าความแม่นยำเท่ากับร้อยละ 98.56 ค่าความระลึกเท่ากับร้อยละ 97.00 และค่าความถ่วงดุลเท่ากับร้อยละ 97.13 ซึ่งมีค่าประสิทธิภาพสูงทุกค่า แสดงว่าตัวต้นแบบที่ใช้ในการพยากรณ์ที่ผู้วิจัยพัฒนาขึ้นมีความถูกต้องแม่นยำในการพยากรณ์ในการรับสมัครนักศึกษาใหม่ โดยที่ผู้วิจัยใช้กฎการจำแนกเทคนิคต้นไม้ตัดสินใจมากที่สุด

จากทฤษฎีและงานวิจัยที่เกี่ยวข้องดังกล่าวข้างต้น ผู้วิจัยเห็นควรที่จะทำการวิจัยโดยเทคนิค Ensemble ด้วยโปรแกรม Rapid Miner Studio โดยได้เลือกใช้อัลกอริทึม Decision Tree, Naïve Bayes และ Rule Induction ในการทำ Vote Ensemble และ Bootstrap Aggregating (Bagging) และทำ Random Forest โดยกำหนดการสร้างโมเดลไว้ที่ 10 ต้น เพื่อเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ของมหาวิทยาลัยราชภัฏนครปฐม และวิเคราะห์ปัจจัยที่ส่งผลต่อการมารายงานตัวเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐม

4. ระเบียบวิธีการวิจัย

งานวิจัยนี้ผู้วิจัยใช้กระบวนการของ CRISP-DM ดังนี้

4.1 ความเข้าใจในธุรกิจ จากสถานการณ์ในปัจจุบันที่จำนวนนักเรียนในระบบการศึกษาลดน้อยลง ทำให้สถาบันอุดมศึกษามีการแข่งขันในการรับนักศึกษาเพื่อเข้าศึกษาต่อเป็นอย่างมาก การประชาสัมพันธ์หลักสูตรให้ตรงกับความสนใจของกลุ่มเป้าหมายเป็นเรื่องที่สำคัญ เพื่อรับมือกับสถานการณ์ดังกล่าวจึงได้พัฒนาตัวแบบการจำแนกพฤติกรรมการรายงานตัวของผู้สมัครที่สอบผ่านการคัดเลือก โดยวิเคราะห์ว่าปัจจัยใดมีผลต่อการมารายงานตัวของนักศึกษาใหม่มหาวิทยาลัยราชภัฏนครปฐม เพื่อนำไปประกอบการวางแผนประชาสัมพันธ์หลักสูตรได้อย่างมีประสิทธิภาพต่อไป

4.2 ความเข้าใจในข้อมูล รวบรวมข้อมูลจากสำนักส่งเสริมวิชาการและงานทะเบียนของมหาวิทยาลัยราชภัฏนครปฐม โดยเลือกเฉพาะข้อมูลของผู้สมัครที่สอบผ่านการคัดเลือกเป็นนักศึกษาใหม่ระดับปริญญาตรี มหาวิทยาลัยราชภัฏนครปฐม ระหว่างปีการศึกษา 2559-2560 ที่สำเร็จการศึกษาจากสถานศึกษาในจังหวัดนครปฐม และที่มีรอยต่อติดกับจังหวัดนครปฐมเท่านั้น ซึ่งมีจำนวน 11,338 ชุด 10 แอทริบิวต์ ได้แก่ ปีการศึกษา, รหัสประจำตัวสอบ, เพศ, รูปแบบการสอบคัดเลือก, เกรดเฉลี่ยรวม, โรงเรียนที่สำเร็จการศึกษา, จังหวัด, สายการเรียน, สาขาวิชาที่เลือก และสถานะการรายงานตัวเข้าศึกษาต่อ ดังแสดงในตารางที่ 1

ตารางที่ 1 ข้อมูลของผู้สมัครที่สอบผ่านการคัดเลือกเป็นนักศึกษา

ปีการศึกษา	รหัสประจำตัวสอบ	เพศ	รูปแบบการสอบคัดเลือก	เกรดเฉลี่ยรวม	โรงเรียน	จังหวัด	สายการเรียน	สาขาที่เลือก	สถานะการรายงานตัว
2559	48939	ชาย	โควต้า	3.36	อู่ทอง	สุพรรณบุรี	อาชีพ	การตลาด	ไม่มา
2559	48941	หญิง	รอบที่ 1	3.51	บางเลนวิทยา	นครปฐม	สามัญ	การบัญชี	มา
...	...	...	...	...	...	...	...	...	...

4.3 การเตรียมข้อมูล นำข้อมูลไปปรับปรุงในส่วนของค่าข้อมูลที่ขาดหายไป (Missing Value) โดยใช้วิธีการตัดเรคคอร์ดที่ข้อมูลไม่สมบูรณ์ทิ้งทั้งหมด และทำการเลือกเฉพาะแอตทริบิวต์เฉพาะที่สำคัญและเกี่ยวข้อง โดยทำการตัดแอตทริบิวต์ออก 2 แอตทริบิวต์ ได้แก่ ปีการศึกษา และรหัสประจำตัวสอบ จากนั้นทำการกำหนดแอตทริบิวต์ “สถานะการรายงานตัว” เป็นประเภท Label โดยมีการปรับเปลี่ยนรูปแบบของข้อมูลเพื่อลดขนาดของข้อมูลที่ใช้ในการประมวลผล จากการปรับปรุงข้อมูลเพื่อนำไปวิเคราะห์ในงานวิจัยครั้งนี้ประกอบด้วย 8 แอตทริบิวต์ และมีจำนวนเรคคอร์ดเท่ากับ 11,288 เรคคอร์ด ดังตารางที่ 2

ตารางที่ 2 ข้อมูลที่ผ่านการเตรียมข้อมูล

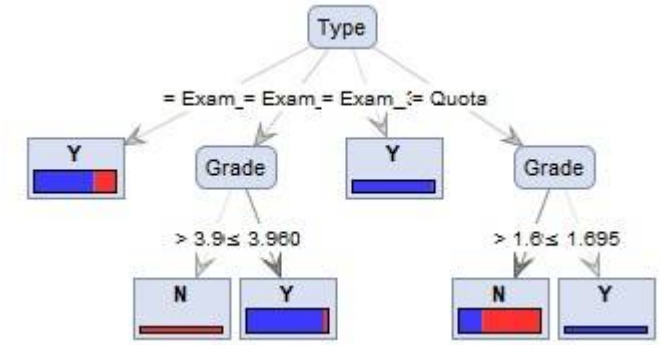
เพศ	รูปแบบการสอบคัดเลือก	เกรดเฉลี่ยรวม	โรงเรียน	จังหวัด	สายการเรียน	สาขาที่เลือก	สถานะการรายงานตัว
M	Quota	3.36	อู่ทอง	สุพรรณบุรี	V	การตลาด	N
F	Exam_1	3.51	บางเลนวิทยา	นครปฐม	S	การบัญชี	Y
...	...	...	...	...	...	...	...

4.4 การจัดทำตัวแบบ ทำการสร้างตัวแบบโดยใช้เทคนิค Ensemble ด้วยโปรแกรม Rapid Miner Studio จำนวน 3 ประเภทดังนี้

1) Vote Ensemble โดยได้เลือกใช้อัลกอริทึม Decision Tree, Naïve Bayes และ Rule Induction ในการหาประสิทธิภาพของเทคนิค Vote Ensemble

2) Bootstrap Aggregating (Bagging) โดยได้เลือกใช้อัลกอริทึม Rule Induction เนื่องจากมีค่าความถูกต้องมากที่สุดจากอัลกอริทึมที่ใช้ในการ Vote ในโมเดล Vote Ensemble และยังทำงานกับแอตทริบิวต์ที่เป็น Nominal อีกด้วย

3) Random Forest เป็นอัลกอริทึมที่ใช้โมเดลเทคนิค Decision Tree โดยสุ่มข้อมูลและสุ่มเลือกแอตทริบิวต์ต่าง ๆ จำนวน 10 ชุด และสร้างโมเดลตามที่ผู้วิจัยได้กำหนดไว้ที่ 10 ต้น ดังแสดงในภาพที่ 4



ภาพที่ 4 ตัวอย่าง Decision Tree สร้างด้วย Random Forest

4.5 การประเมินผล ทำการเปรียบเทียบประสิทธิภาพของตัวแบบที่สร้างขึ้น ซึ่งตรวจสอบจากค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) จะพบว่าตัวแบบที่สร้างขึ้นด้วยเทคนิค Bootstrap Aggregating โดยใช้อัลกอริทึม Rule Induction มีค่าความถูกต้องสูงที่สุด ซึ่งเท่ากับ 83.96% ดังแสดงในตารางที่ 3

ตารางที่ 3 ผลการหาประสิทธิภาพของตัวแบบที่สร้างขึ้น

Algorithms	Accuracy	Precision	Recall
Vote Ensemble	83.19%	73.98%	69.90%
Bootstrap Aggregating (Rule Induction)	83.96%	73.88%	73.94%
Random Forest	81.95%	69.87%	72.59%

4.6 การนำเอาตัวแบบไปใช้งาน นำตัวแบบที่พัฒนาด้วยเทคนิค Bootstrap Aggregating (Rule Induction) ซึ่งได้ค่าความถูกต้องที่สุดไปใช้ในการวิเคราะห์ปัจจัยที่มีผลต่อการรายงานตัวของนักศึกษาใหม่ของมหาวิทยาลัยราชภัฏนครปฐม ซึ่งผลการวิเคราะห์ปัจจัยพบว่าปัจจัยที่มีผลมากที่สุด คือ รูปแบบการสอบคัดเลือก เกรดเฉลี่ยรวม เพศ โรงเรียนที่สำเร็จการศึกษา จังหวัด สายการเรียน และสาขาวิชาที่เลือก ตามลำดับ ดังแสดงในตารางที่ 4 เพื่อนำผลที่ได้จากการวิเคราะห์ข้อมูลไปประกอบการวางแผนการรับนักศึกษาใหม่ในปีการศึกษาถัดไปได้อย่างมีประสิทธิภาพ และจากตัวแบบที่ได้นี้สามารถนำไปพัฒนาเป็นแอพพลิเคชั่นสำหรับพยากรณ์หรือทำนายจำนวนนักศึกษาใหม่ในปีการศึกษานั้น ๆ

ตารางที่ 4 ผลการวิเคราะห์ปัจจัยที่มีผลต่อการรายงานตัวเข้าศึกษาของนักศึกษาใหม่

Factors	Accuracy	No.
รูปแบบการสอบคัดเลือก	81.93%	1
เกรดเฉลี่ยรวม	69.29%	2
เพศ	69.27%	3
โรงเรียน	69.27%	3
จังหวัด	69.27%	3
สายการเรียน	69.27%	3
สาขาที่เลือก	69.27%	3

5. อภิปรายผลการวิจัย

การวิจัยครั้งนี้เป็นการนำเสนอผลการเปรียบเทียบประสิทธิภาพของตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิค ENSEMBLE ทั้ง 3 เทคนิค คือ Vote Ensemble, Bootstrap Aggregating (Bagging) และ Random Forest ซึ่งตัวแบบที่สร้างโดยเทคนิค Bootstrap Aggregating (Rule Induction) มีค่าความถูกต้องสูงที่สุดที่ 83.96% ซึ่งสูงกว่าเทคนิค Vote Ensemble และ Random Forest ที่มีค่าความถูกต้องเท่ากับ 83.19% และ 81.95% ตามลำดับ และผลการวิเคราะห์ปัจจัยที่ส่งผลกับการเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐมมากที่สุด คือ รูปแบบการสอบคัดเลือก เกรดเฉลี่ยรวม เพศ โรงเรียนที่สำเร็จการศึกษา จังหวัด สายการเรียน และสาขาวิชาที่เลือก ตามลำดับ ซึ่งสาเหตุที่ปัจจัยรูปแบบการสอบคัดเลือกมีผลกับการรายงานตัวเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐมมากที่สุด เพราะว่า รูปแบบการสอบคัดเลือกการสอบคัดเลือกด้วยวิธีการสอบตรง ในรอบที่ 3 นั้น มีผู้ที่สอบผ่านการคัดเลือกมารายงานเป็นนักศึกษาตัวเกือบ 100% และการสอบคัดเลือกแบบโควตานั้น ส่วนใหญ่ผู้ที่สอบผ่านการคัดเลือกไม่มารายงานตัวเป็นนักศึกษา เพราะเป็นผู้ที่มีเกรดเฉลี่ยสูงสามารถที่จะเลือกศึกษาต่อในสถานศึกษาอื่นๆ ได้ ผู้วิจัยเห็นควรว่าจะต้องมีนโยบายในการสอบคัดเลือกแบบโควตาที่หลายหลากมากขึ้น ไม่เฉพาะเจาะจงกับคุณสมบัติในเรื่องของเกรดเฉลี่ยเท่านั้น อาจจะต้องพิจารณาในประเด็นอื่น ๆ เพิ่มเติม เพื่อให้ได้นักศึกษาที่ต้องการมาศึกษาต่อยังมหาวิทยาลัยราชภัฏนครปฐมอย่างแท้จริง และเพื่อให้การประชาสัมพันธ์หลักสูตรและแนะแนวตามโรงเรียนต่าง ๆ ได้ตรงกับกลุ่มเป้าหมายมากยิ่งขึ้น

6. เอกสารอ้างอิง

ธาดา จันตะคุณ. (2559). **รายงานการวิจัยเรื่อง ตัวแบบการจำแนกการเลือกหลักสูตรการศึกษา คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม โดยใช้เทคนิคเหมืองข้อมูล.** กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ธีรพงษ์ สังข์ศรี. (2557). **รายงานการวิจัยเรื่อง การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิคการทำเหมืองข้อมูล**. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). **การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิงเบื้องต้น**. (พิมพ์ครั้งที่ 2). กรุงเทพฯ: เอเซียดิจิตอลการพิมพ์ จำกัด.

อัจฉราภรณ์ จุฑาผาด. (2556). **รายงานการวิจัยเรื่อง การพัฒนาระบบสารสนเทศเพื่อการพยากรณ์จำนวนนักศึกษาใหม่โดยใช้กฎการจำแนกต้นไม้ตัดสินใจ.** กรุงเทพฯ: มหาวิทยาลัยนเรศวร.

Data Mining Trend. **การสร้างโมเดล Ensemble.** เข้าถึงได้จาก https://goo.gl/JDiYLO สืบค้นเมื่อ มิ.ย. 2560.

Jaiwei Han and Micheline Kamber. (2006). **Data Mining Concepts and Technique.** San Francisco, Morgan Kqufmann Publishers.

Jakkrit Techo, Cholwich Nattee and Thanaruk Theeramunkong. (2012). Boosting-based ensemble learning with penalty profiles for automatic Thai unknown word recognition. **International Journal of ELSEVIER**, Vol.63, No.6, 1117-1134.