การพยากรณ์ภาวะการมีงานทำของบัณฑิตด้วยเทคนิคการจำแนกข้อมูลด้วยกฎร่วมกับวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยบนชุดข้อมูลที่ไม่สมดุล

Graduates Employability Prediction through Rule-Based Classification Techniques with SMOTE in Imbalanced Data Sets

ภรัณยา ปาลวิสุทธิ์[1]* อภินันท์ จุ่นกรณ์[1] มงคล รอดจันทร์[2] และประภาพรรณ เพียรชอบ[3]
Paranya Palwisut[1]*, Apinan Junkorn[1], Mongkol Rodjan[2] and Prapapan Pienchob[3]
[1]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[1]Department of Data Science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
[2]สาขาวิชาเทคโนโลยีคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[2]Department of Computer Technology, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
[3]สาขาวิชาวิทยาศาสตร์และเทคโนโลยีการอาหาร คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[3]Department of Food Science and Technology, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
*Corresponding author: paranya@npru.ac.th
Received: June 15, 2023
Revised: July 26, 2023
Accepted: August 3, 2023

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบเทคนิคการพัฒนาแบบจำลองการพยากรณ์ภาวะการมีงานทำของบัณฑิตด้วยเทคนิคการจำแนกข้อมูลด้วยกฎร่วมกับวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยบนชุดข้อมูลที่ไม่สมดุล ผลจากการวิเคราะห์ข้อมูลพบว่า ข้อมูลมีความไม่สมดุลของกลุ่มข้อมูล โดยมีจำนวนกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งเป็นจำนวนมาก ผู้วิจัยจึงได้ทำการแก้ปัญหาโดยปรับความสมดุลของข้อมูลด้วยวิธีเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-sampling Technique--SMOTE) แล้วพัฒนาตัวแบบด้วยเทคนิคการจำแนกข้อมูลด้วยกฎ ได้แก่ RIPPER PART และ PRISM โดยใช้ 5-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ และได้ใช้ค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าความระลึก (recall) และค่าความเหวี่ยง (F-measure) ในการวัดประสิทธิภาพการพยากรณ์ของตัวแบบ ผลการทดลองประสิทธิภาพในการพยากรณ์ของตัวแบบพบว่า ตัวแบบที่ได้จากอัลกอริทึม PRISM ร่วมกับการปรับสมดุลข้อมูลด้วยวิธี SMOTE มีประสิทธิภาพสูงที่สุด โดยมีค่าความถูกต้อง เท่ากับร้อยละ 85.69 ค่าความแม่นยำเท่ากับร้อยละ 85.60 ค่าความระลึกเท่ากับร้อยละ 85.70 และค่าความเหวี่ยงเท่ากับร้อยละ 85.60 ตามลำดับ

คำสำคัญ: ภาวะการมีงานทำของบัณฑิต การจำแนกข้อมูลด้วยกฎ การสุ่มเพิ่มตัวอย่างกลุ่มน้อย

Abstract

This research aims to develop graduates' employability prediction models through rule-based classification techniques with SMOTE in imbalanced data sets. After analyzing the data, a class imbalance problem was found. In order to improve the quality of the data, SMOTE was used to increase the minority class. Then rules-based classification techniques (RIPPER, PART, and PRISM) were used to build the prediction models. Moreover, 5-fold cross-validation was utilized to split the data into the training and test sets. This research has measured performance models with accuracy, precision, recall, and f-measure. The experimental results demonstrated that the PRISM algorithm combined with SMOTE had the highest efficiency, with an accuracy of 85.69%, precision of 85.60%, recall of 85.70%, and f-measure of 85.60%, respectively.

Keywords: graduates employability, rule-based classification, SMOTE

บทนำ

กระแสความเปลี่ยนแปลงของสภาพสังคมและเศรษฐกิจในปัจจุบัน ส่งผลให้ประชากรทุกวัยต้องปรับตัวเพื่อความอยู่รอด ต้องเผชิญกับภาวะการแข่งขันในหลาย ๆ ด้าน เช่น การเข้าทำงานในหน่วยงานหรือองค์กรต่าง ๆ ซึ่งความต้องการของนายจ้างในตลาดแรงงานล้วนต้องการทรัพยากรมนุษย์ที่มีคุณภาพ คือ มีความพร้อมทั้งความรู้ความสามารถ ทักษะความชำนาญในงาน ความสามารถในการทำงานร่วมกับผู้อื่น มีความรับผิดชอบ สามารถช่วยพัฒนาองค์กรให้มีการดำเนินงานอย่างมีประสิทธิภาพ

สถาบันการศึกษามีหน้าในการผลิตบัณฑิตที่มีคุณภาพเป็นที่ต้องการของตลาดแรงงาน ได้เปิดการเรียนการสอนหลักสูตรต่าง ๆ โดยมีจุดมุ่งหมายเพื่อที่จะผลิตบัณฑิตที่สามารถทำงานตามหน้าที่ในสายการเรียนด้านต่าง ๆ ซึ่งแต่ละปีการศึกษา จะมีบัณฑิตผู้สำเร็จการศึกษาเพื่อเข้าสู่ตลาดแรงงานในสาขาวิชาชีพตามกลุ่มอาชีพต่าง ๆ และตามประเภทงานที่บัณฑิตสนใจ สถาบันการศึกษาจึงมีการจัดเก็บข้อมูลภาวะการมีงานทำของบัณฑิต ว่าสามารถหางานได้หรือไม่ ได้ทำงานที่ตรงกับสาขาที่ศึกษาหรือไม่ ตำแหน่งงานและรายได้เหมาะสมกับวุฒิการศึกษาหรือไม่ อย่างไร เพื่อนำข้อมูลที่ได้มาเป็นแนวทางในการพัฒนาหลักสูตรการเรียนการสอนเพื่อผลิตบัณฑิตที่มีคุณภาพ มีศักยภาพในการแข่งขันเข้าทำงาน และสามารถทำงานเพื่อตอบสนองความต้องการของตลาดแรงงานในปัจจุบันและในอนาคตต่อไป

สถาบันอุดมศึกษามีข้อมูลที่ถูกจัดเก็บอยู่เป็นจำนวนมากข้อมูลต่าง ๆ เพิ่มมากขึ้นตามการเติบโตควบคู่กับความทันสมัยของเทคโนโลยีทำให้สามารถเก็บจนเป็นฐานข้อมูลขนาดใหญ่และมีข้อมูลสารสนเทศที่มีองค์ความรู้สำคัญซ่อนอยู่ การทำเหมืองข้อมูลจะเป็นกระบวนการค้นหาความสัมพันธ์ รูปแบบ หรือกฎเกณฑ์ โดยนำข้อมูลปริมาณมากจากแหล่งจัดเก็บข้อมูลขนาดใหญ่ที่เก็บไว้นำมาเข้าสู่กระบวนการวิธีทางคณิตศาสตร์ และสถิติมาช่วยในการวิเคราะห์ข้อมูลและจัดการข้อมูล การทำเหมืองข้อมูลนั้นมีเทคนิคที่สามารถประยุกต์ใช้ได้มากมาย โดยได้มีงานวิจัยด้านการศึกษาจากข้อมูลที่มีอยู่ในสถาบันอุดมศึกษาเป็นจำนวนมาก และการทำเหมืองข้อมูลเป็นเครื่องมือช่วยให้ทราบสถานการณ์รูปแบบต่าง ๆ ช่วยเพิ่มโอกาสเลือกที่ส่งเสริมให้ทราบปัจจัยสำคัญเพื่อวางแผนกลไกรองรับการพัฒนาและชี้แจงแนวทาง ตลอดจนสร้างกระบวนการเรียนรู้ระหว่างความสัมพันธ์ต่าง ๆ อย่างมีเป้าหมาย

จากความสำคัญของปัญหา ผู้วิจัยจึงมีแนวคิดในการใช้ประโยชน์จากข้อมูลที่มีมาสกัดเข้ากระบวนการทำเหมืองข้อมูลให้ได้มาซึ่งข้อมูลที่มีประสิทธิภาพมากที่สุด โดยนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้ในการ

พัฒนาแบบจำลองกฎการการพยากรณ์ภาวะการมีงานทำของบัณฑิตด้วยเทคนิคเหมืองข้อมูล เพื่อให้สถานศึกษาได้นำไปประยุกต์ในการคาดการณ์ปัจจัยที่มีผลต่อภาวะการมีงานทำและช่วงเวลาในการได้งานทำของบัณฑิตได้อย่างแม่นยำและมีประสิทธิภาพ

วัตถุประสงค์การวิจัย

เพื่อศึกษาและเปรียบเทียบเทคนิคการพัฒนาแบบจำลองการพยากรณ์ภาวะการมีงานทำของบัณฑิตด้วยเทคนิคการจำแนกข้อมูลด้วยกฎร่วมกับวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยบนชุดข้อมูลที่ไม่สมดุล

แนวคิดทฤษฎีที่เกี่ยวข้อง

1) ภาวะการมีงานทำของบัณฑิต

ภาวะการมีงานทำของบัณฑิต จากสำนักงานคณะกรรมการการอุดมศึกษา ได้กล่าวถึงบัณฑิตปริญญาตรีที่ได้งานทำหรือประกอบอาชีพอิสระ ในคู่มือการประกันคุณภาพการศึกษาภายใน ระดับอุดมศึกษา ฉบับปีการศึกษา 2558 สรุปใจความสำคัญได้ว่า บัณฑิตปริญญาตรีที่สำเร็จการศึกษาในหลักสูตรภาคปกติ ภาคพิเศษ และภาคนอกเวลาในสาขานั้น ๆ ที่ได้งานทำหรือมีกิจการของตนเองที่มีรายได้ประจำสามารถเลี้ยงชีพตนเองได้การนับการมีงานทำนับกรณีการทำงานสุจริตทุกประเภทที่สามารถสร้างรายได้เข้ามาเป็นประจำเพื่อเลี้ยงชีพตนเองได้ ซึ่งสอดคล้องกับคู่มือการประเมินคุณภาพภายนอกรอบสามระดับอุดมศึกษา สำนักงานรับรองมาตรฐานและประเมินคุณภาพการศึกษา (องค์การมหาชน) และสำนักงานคณะกรรมการการอุดมศึกษาคู่มือการประกันคุณภาพการศึกษาภายใน ระดับอุดมศึกษา พ.ศ. 2557 กล่าวว่า บัณฑิตปริญญาตรีที่สำเร็จการศึกษาในหลักสูตรภาคปกติ และภาคพิเศษ และภาคนอกเวลา ในสาขานั้น ๆ ที่ได้งานหรือมีกิจการเป็นของตนเองที่มีรายได้ประจำภายในระยะเวลา 1 ปี นับจากวันที่สำเร็จการศึกษา เมื่อเทียบกับบัณฑิตที่สำเร็จการศึกษาในปีนั้น การนับการมีงานทำนับกรณีการทำงานสุจริตทุกประเภทที่สามารถสร้างรายได้เข้ามาเป็นประจำเพื่อเลี้ยงชีพตัวเองได้

2) เหมืองข้อมูล (data mining)

การทำเหมืองข้อมูล Kijsirikul (2002) คือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์รวมทั้งในด้านเศรษฐกิจและสังคม การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย ๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล ซึ่งจุดประสงค์ของการทำเหมืองข้อมูลสามารถใช้ค้นข้อมูลสำคัญที่ปะปนกับข้อมูลอื่น ๆ ในฐานข้อมูลที่ไม่ใช่แค่การสุ่มหา เรียกว่า KDD--Knowledge Discovery in Database หรือการค้นหาข้อมูลด้วยความรู้ มีด้วยกัน 5 รูปแบบ คือ (1) การค้นหากฎความสัมพันธ์ (2) การจำแนกประเภทและการพยากรณ์ (3) การจัดกลุ่มข้อมูล (4) การหาค่าผิดปกติที่เกิดขึ้น และ5) การวิเคราะห์แนวโน้ม

3) การจำแนกข้อมูลด้วยกฎ (rule-based classification)

การจำแนกข้อมูลด้วยกฎ Amphawan (2022) จะเป็นแบบจำลองที่จะแสดงผลด้วยเซตของกฎที่มีลักษณะแบบ ‘IF-THEN’ โดยกฎหนึ่ง ๆ จะถูกแสดงอยู่ในรูปฟอร์มดังนี้

IF condition THEN conclusion

ตัวอย่างเช่น

R1: IF age = youth AND student = yes THEN buys_computer = yes

จากรูปแบบฟอร์มของกฎและตัวอย่างกฎ R1 ข้างต้นจะประกอบไปด้วยข้อมูลสองส่วนด้วยกัน คือ ส่วนแรก ส่วนเงื่อนไข ‘IF’ จะเป็นส่วนที่เรียกว่า ‘rule antecedent’ หรือ ‘precondition’ ประกอบไปด้วยเซตของแอทริบิวต่าง ๆ ประกอบกันเป็นเงื่อนไข ซึ่งจากกฎ R1 จะประกอบไปด้วยแอทริบิวที่บ่งบอกถึงคุณลักษณะของคน 2 แอทริบิวด้วยกัน คือ อายุอยู่ในช่วงหนุ่มสาว (age=youth) และเป็นนักเรียน

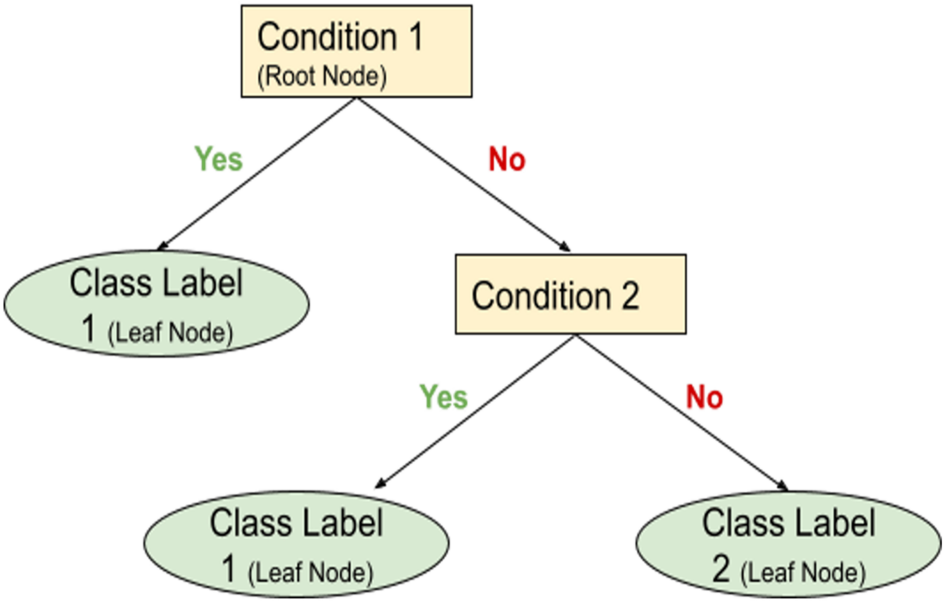
(student=yes) โดยส่วนเงื่อนไขของกฎจะทำการเชื่อมโยงแอทริบิวต่าง ๆ เข้าด้วยกันด้วยเครื่องหมาย AND ที่ซึ่งจะเป็นการบ่งบอกว่าข้อมูลจะต้องเป็นไปตามเงื่อนไขที่ถูกกำหนดไว้ทั้งหมด ในส่วนที่สอง จะเรียกว่า 'rule consequence' ที่จะมีหมวดหมู่ข้อมูลบรรจุอยู่ จากตัวอย่างกฎ R1 ข้างต้นจะทำการทำนายว่า ลูกค้าที่มีคุณลักษณะเช่นไรที่จะทำการซื้อคอมพิวเตอร์ เป็นต้น รูปแบบของการจำแนกข้อมูลด้วยกฎ แสดงได้ดังภาพ 1

การจำแนกข้อมูลด้วยกฎที่นำมาใช้เปรียบเทียบในงานวิจัย จำนวน 3 เทคนิค มีดังนี้

- RIPPER--Repeated Incremental Pruning to Produce Error Reduction Thakur (2020) เป็นการเรียนรู้กฎ ซึ่งมีกระบวนการตัดแต่งกิ่งที่เพิ่มขึ้นซ้ำ ๆ เพื่อลดข้อผิดพลาดที่เกิดจาการสร้างกฎ ประกอบด้วย 4 ขั้นตอน คือ (1) การเจริญเติบโต (growth) สร้างลำดับของกฎ โดยเพิ่มกฎจนเป็นที่น่าพอใจแล้วจึงหยุดเพิ่ม (2) การตัดแต่งกิ่ง (pruning) ทำการตัดกฎที่ลดประสิทธิภาพการเรียนรู้กฎออก (3) การเพิ่มประสิทธิภาพ (optimization) มีการเพิ่มคุณลักษณะเข้าไปในแต่ละกฎเดิมหรือกฎที่ถูกสร้างขึ้นใหม่ในขั้นที่ (1) และ (2) และ (4) การเลือกกฎ (selection) เป็นการเลือกกฎที่ดีที่สุดเก็บไว้ และตัดกฎอื่น ๆ ออกไป

- PART Mohamed Salleh and Omar (2012) เป็นเทคนิคการเรียนรู้กฎที่พัฒนาจาก C4.5 และ RIPPER โดยรวมทั้ง 2 เทคนิคเข้าด้วยกัน มีจุดเด่น คือ สามารถเรียนรู้กฎได้เองเหมือนเทคนิค RIPPER และสามารถจัดการกับข้อมูลที่หายไปและคุณลักษณะทางตัวเลขที่ต่างกันได้ดี

- PRISM Chiroma, Liu and Cocea (2018) ได้ถูกแนะนำโดย Cendrowska ที่ได้ศึกษาปัญหาเกี่ยวกับอัลกอริทึม ID3 ของ Quinlan ซึ่งได้ระบุข้อจำกัดที่สำคัญบางประการของอัลกอริทึม ID3 ซึ่งทำให้การใช้งานไม่เหมาะสมสำหรับหลายโดเมน ดังนั้น เพื่อลดปัญหาดังกล่าวจึงได้นำเสนออัลกอริทึมที่เรียกว่า PRISM ซึ่งอัลกอริทึมนี้มีจุดมุ่งหมายหลักเพื่อหลีกเลี่ยงการสร้างกฎที่ซับซ้อนที่มีเงื่อนไขซ้ำซ้อนจำนวนมาก



ภาพ 1 Rule-Based Classification

Note. From *Rule based data mining classifier: A comprehensive guide 101*, by Hevo Data Inc., 2024, retrieved from https://hevodata.com/learn/rule-based-data-mining/

4) เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

Synthetic Minority Over-sampling Technique--SMOTE Chawla (2002) เป็นเทคนิคที่ใช้ในการแก้ปัญหาที่ต้องการจำแนกข้อมูลไม่สมดุล ซึ่งข้อมูลมีจำนวนตัวอย่างแตกต่างกันมากในแต่ละคลาส เมื่อทำการจำแนกประเภทจะทำให้มีการเรียนรู้แต่ข้อมูลกลุ่มที่มาก ผลที่ได้ก็จะจำแนกไปในข้อมูลกลุ่มมาก วิธี SMOTE เป็นวิธีการเพิ่มจำนวนข้อมูลประเภทที่มีข้อมูลน้อย ให้เพิ่มปริมาณข้อมูลใกล้เคียงกับประเภทที่มีมากที่สุด โดยสุ่มค่าขึ้นมาหนึ่งค่า และหาค่าระยะห่างระหว่างค่าที่เลือกกับทุก ๆ ค่า เลือกค่าที่ใกล้เคียงที่สุด เช่น กำหนดไว้ 5 ค่า สุ่มค่าจากที่เลือก 1 ใน หลัง เพื่อนำค่าที่ได้มาเพิ่มจำนวนข้อมูล ดังสมการที่ 1

$$x_{new} = x_i + (\hat{x_i} - x_i) \times \delta \quad (1)$$

โดยที่

x_{new} คือ ข้อมูลใหม่

x_i คือ ข้อมูลที่สุ่มในตอนแรก

$\hat{x_i}$ คือ ข้อมูลที่สุ่มมาอีก เช่น สุ่มมาอีก 5 จุด

δ คือ ค่าสุ่มตั้งแต่ 0-1

5) การวัดประสิทธิภาพของตัวแบบ

การวัดประสิทธิภาพของตัวแบบ ในแต่ละเทคนิควิธีที่นำมาเปรียบเทียบ โดยใช้ค่าความถูกต้องของแบบจำลอง (accuracy) จะต้องทำการเลือกข้อมูลสำหรับเรียนรู้ (training set) และข้อมูลสำหรับทดสอบ (testing set) งานวิจัยนี้เลือกใช้วิธีการสุ่มเลือกข้อมูลแบบความเที่ยงตรง k กลุ่ม (k-fold cross validation) การวัดค่าประสิทธิภาพของแบบจำลองนั้นจะอาศัย Confusion matrix ในการคำนวณค่า

5.1 Confusion Matrix Markoulidakis et al. (2021)

การประเมินผลลัพธ์การทำนาย หรือ ผลลัพธ์จากโปรแกรมเปรียบเทียบกับผลลัพธ์จริง ซึ่งจะเป็นตารางแสดงจำนวนของข้อมูลในแต่ละประเภทที่ถูกจำแนกด้วยตัวแบบ

ตาราง 1

ตาราง Confusion Matrix แสดงจำนวนของข้อมูลในแต่ละประเภท

Actual Class	Predicted Class	
	Class = Yes	Class = No
Class = Yes	TP	FN
Class = No	FP	TN

โดยที่

True Positive--TP เป็นจำนวนข้อมูลประเภทเป็นบวกและถูกจำแนกประเภทว่าเป็นบวก

False Positive--FP เป็นจำนวนข้อมูลประเภทเป็นลบ แต่ถูกจำแนกประเภทว่าเป็นบวก

True Negative--TN เป็นจำนวนข้อมูลประเภทเป็นลบและถูกจำแนกประเภทว่าเป็นลบ

False Negative--FN เป็นจำนวนข้อมูลประเภทเป็นบวกแต่ถูกจำแนกประเภทว่าเป็นลบ

5.2 ตัวชี้วัดที่สามารถคำนวณได้จาก Confusion Matrix Markoulidakis et al. (2021)

- ค่าความแม่นยำ (precision) เป็นการวัดความแม่นยำของตัวแบบ โดยพิจารณาแยกทีละคลาส

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

- ค่าความระลึก (Recall) เป็นการวัดความถูกต้องของตัวแบบ โดยพิจารณาแยกทีละคลาส

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

ค่าความเหวี่ยง (F-measure) เป็นการวัดค่า Precision และ Recall พร้อมกันของตัวแบบ โดยพิจารณาแยกทีละคลาส

$$\text{F-measure} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

- ค่าความถูกต้อง (accuracy) เป็นการวัดความถูกต้องของตัวแบบ โดยพิจารณารวมทุกคลาส คือ จำนวน True Positive ของทุกคลาสรวมกัน

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

6) งานวิจัยที่เกี่ยวข้อง

Chinjoho Jabjone and Jongmuanwai (2021) ได้ทำการวิจัยเกี่ยวกับการพัฒนาแบบจำลองเทคโนโลยีสารสนเทศในการคาดการณ์อาชีพอนาคตของบัณฑิตสาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏนครราชสีมา การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองเทคโนโลยีสารสนเทศในการคาดการณ์อาชีพอนาคตของบัณฑิตสาขาเทคโนโลยีสารสนเทศ และศึกษาผลการคาดการณ์อาชีพอนาคตของบัณฑิตสาขาเทคโนโลยีสารสนเทศการดำเนินการกับ 2 กลุ่ม คือ กลุ่มเป้าหมาย ได้แก่ บัณฑิตที่จบการศึกษาระหว่างปี พ.ศ. 2546-2560 สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏนครราชสีมา ได้มาจากการเลือกแบบเจาะจง และกลุ่มตัวอย่าง จำนวน 1,178 คน ได้มาจากการสุ่มเลือกแบบแบ่งกลุ่มชั้นหลายขั้นตอน การเก็บรวบรวมข้อมูลได้จากการศึกษาเอกสาร แนวคิด ทฤษฎี และเทคนิคเหมืองข้อมูลด้วยแบบศึกษาไม่มีโครงสร้าง ในการสังเคราะห์แบบจำลอง ส่วนผลการคาดการณ์อาชีพอนาคตของบัณฑิต ด้วยแบบรายการบันทึกข้อมูลด้านอาชีพของบัณฑิต การวิเคราะห์ข้อมูลเชิงคุณภาพ โดยใช้การวิเคราะห์เชิงเนื้อหาและข้อมูลเชิงปริมาณ ได้แก่ ค่าร้อยละ และใช้หลักการวิเคราะห์แบบสมการเชิงเส้น ในด้านค่าความถูกต้องและค่าความเชื่อมั่น ผลการวิจัย พบว่า องค์ประกอบแบบจำลองเทคโนโลยีสารสนเทศในการคาดการณ์อาชีพของบัณฑิตสาขาเทคโนโลยีสารสนเทศประกอบด้วย 3 องค์ประกอบหลัก (1) เทคนิคการแบ่งกลุ่มข้อมูล (2) เทคนิคการจัดกลุ่มข้อมูล (3) เทคนิคการคาดการณ์ข้อมูล โดยแบบจำลองมีประสิทธิภาพในการคาดการณ์ข้อมูลภาพรวมเฉลี่ยร้อยละ 77.40 ซึ่งอยู่ในระดับสูง และผลการคาดการณ์อาชีพอนาคตของบัณฑิต สาขาเทคโนโลยีสารสนเทศ ประกอบด้วย การเรียงลำดับคะแนนเฉลี่ยสูงสุด (1-3) ได้แก่ อาชีพโปรแกรมเมอร์ อาชีพเจ้าหน้าที่ธุรการ และอาชีพอิสระ ทั้งนี้แบบจำลองสามารถเป็นประโยชน์ของข้อมูลเพื่อประกอบการตัดสินใจเลือกอาชีพในอนาคตสำหรับบัณฑิตสาขาเทคโนโลยีสารสนเทศได้

Wanon, Arreerard and Sanrach (2018) ได้ทำการศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีสาขาคอมพิวเตอร์โดยใช้เทคนิคเหมืองข้อมูล มีวัตถุประสงค์ (1) เพื่อศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีโดยใช้เทคนิคเหมืองข้อมูลที่เหมาะสม (2) เพื่อเปรียบเทียบผลการวิเคราะห์พยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีโดยใช้เทคนิคเหมืองข้อมูล งานวิจัยนี้ได้นำข้อมูลภาวะการมีงานทำของบัณฑิต และข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา จากสำนักงานคณะกรรมการการอุดมศึกษาย้อนหลัง 5 ปี คือปี 2555–2559 จำนวน 65,335 ระเบียน ในสาขาวิชาทางด้านคอมพิวเตอร์ และมีคุณลักษณะประกอบด้วย ผลการเรียนความสามารถพิเศษ อาชีพของบิดามารดา รายได้ของบิดามารดา เพศ ตำแหน่งงาน ความสอดคล้องสาขา สาขาวิชาทดลองวัดความแม่นยำด้วยเทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจ เทคนิคการจำแนกข้อมูลด้วยวิธีแรนดอม ฟอร์เรส และเทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิง ผลการวิจัยพบว่า ความแม่นยำในการจำแนกประเภทข้อมูลจาก 3 เทคนิค (1) เทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจเท่ากับร้อยละ 81.91 (2) เทคนิคการจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรสเท่ากับร้อยละ 84.29 และ (3) เทคนิคการจำแนกข้อมูลด้วยวิธีแบ็กกิงเท่ากับร้อยละ 81.71 พบว่า เทคนิคแรนดอมฟอร์เรสให้ความถูกต้องในการจำแนกประเภทข้อมูลสูงที่สุด

Vilailuck, Jaroenpuntaruk and Wichadakul (2015) ได้วิจัย “การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียนโรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา” เพื่อพัฒนาคลังข้อมูลและสร้างแบบจำลองทำนายผลการเรียน ของนักเรียนโรงเรียนสาธิตมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา โดยใช้ข้อมูลนักเรียนระดับมัธยมศึกษาตอนต้น ระหว่างปีการศึกษา 2548–2556 นำมาพัฒนาคลังข้อมูลด้วยโครงสร้างแบบ Snowflake Schema จากนั้นใช้ข้อมูลนักเรียนระดับมัธยมศึกษาตอนปลาย ระหว่างปีการศึกษา 2556–2556 จำนวน 525 ระเบียน ประกอบด้วย 16 คุณลักษณะ นำมาพัฒนาแบบจำลองทำนาย โดยใช้ชุดข้อมูล 2 แบบ คือ แบบไม่จัดกลุ่ม (original data) และแบบจัดกลุ่ม (cluster data) จำกนั้นทำการคัดเลือกคุณลักษณะด้วยวิธี Correlation-based Feature Selection และ วิธี Information Gain แล้วใช้เทคนิค Neural Network

แบบ Multi-Layer Perceptron, เทคนิค Support Vector Machine และเทคนิค Decision Tree มาพัฒนาแบบจำลองทำนายพร้อมเปรียบเทียบ และวัดประสิทธิภาพด้วยวิธี 10-fold Cross Validation พบว่า คลังข้อมูลที่พัฒนาขึ้นผู้ใช้งานมีความพึงพอใจอยู่ในระดับดี และจากการเปรียบเทียบแบบจำลองทำนายโดยชุดข้อมูลแบบไม่จัดกลุ่ม (original data) ที่คัดเลือกด้วยวิธี Correlation-based Feature Selection ร่วมกับเทคนิค Neural Network แบบ Multi-Layer Perceptron จำนวน 5 คุณลักษณะ คือ แผนการเรียน ผลการเรียนเฉลี่ยวิทยาศาสตร์ ผลการเรียนเฉลี่ยสังคมศึกษา ผลการเรียนเฉลี่ย ภาษาอังกฤษ และผลการเรียนเฉลี่ยระดับชั้นมัธยมศึกษาปีที่ 3 มีความเหมาะมากที่สุด โดยมีค่าความถูกต้องสูงสุดที่ร้อยละ 94.48 และมีค่ารากที่สองของความคลาดเคลื่อน (RMSE) น้อยที่สุดที่ 0.1880 จากนั้นนำมาพัฒนาระบบพยากรณ์ผลการเรียนด้วยภาษา PHP

Kladkaew (2014) ได้วิจัย “ระบบสนับสนุนการตัดสินใจการเลือกตำแหน่งงานให้สอดคล้องกับความสามารถของบัณฑิต” เพื่อศึกษาปัจจัยที่มีผลต่อการเลือกตำแหน่งงานที่สอดคล้องกับความสามารถบัณฑิตด้วยเทคนิค Logistic Regression Analysis โดยการจำแนกประเภทข้อมูลด้วยเทคนิค Decision Tree และเปรียบเทียบความถูกต้องการทำนายระหว่างเทคนิค Logistic Regression Analysis และการจำแนกประเภทข้อมูลด้วยเทคนิค Decision Tree จากนั้น ได้ทำการพัฒนาแบบจำลองช่วยในการตัดสินใจเลือกตำแหน่งงานให้สอดคล้องกับความสามารถของบัณฑิต จากข้อมูลภาวะมีงานทำบัณฑิตที่สำเร็จปีการศึกษา 2555–2557 จำนวน 1,933 ระเบียน จำนวน 2,825 คน เป็นกลุ่มตัวอย่าง และได้รวบรวมแบบสำรวจข้อมูลภาวะมีงานทำของศูนย์คอมพิวเตอร์ มหาวิทยาลัยราชภัฏหมู่บ้านจอมบึงเป็นเครื่องมือที่ใช้ศึกษา และได้แบ่งตำแหน่งงานหรือกลุ่มอาชีพเป็น 8 กลุ่ม ผลการวิจัยพบว่าปัจจัยจาก 12 คุณลักษณะ มีผลต่อการเลือกตำแหน่งงานหรือกลุ่มอาชีพมี 4 คุณลักษณะ คือ เพศ คุณวุฒิการศึกษาที่สำเร็จการศึกษา สาขาวิชาที่เรียน และความสามารถพิเศษ ซึ่งมีระดับนัยสำคัญทางสถิติอยู่ที่ .05 และจากการวัดค่าประสิทธิภาพตัวแบบด้วยค่าความถูกต้อง ค่าความแม่นยำและค่าความระลึก ซึ่งได้วิเคราะห์ผลเปรียบเทียบค่าความถูกต้องของการพยากรณ์เทคนิคการจำแนกประเภทข้อมูล พบว่า เทคนิค Decision Tree อยู่ที่ร้อยละ 57.37 โดยมีค่ามากกว่าเทคนิค Logistic Regression Analysis เล็กน้อยที่ร้อยละ 56.30 จึงได้พัฒนาแบบจำลองระบบสนับสนุนการตัดสินใจการเลือกตำแหน่งงานให้สอดคล้องกับความสามารถของบัณฑิตโดยใช้เงื่อนไข 4 คุณลักษณะดังกล่าวมาสร้างเป็นแบบจำลองระบบโดยออกแบบส่วนติดต่อประสานงานผู้ใช้ด้วย Microsoft Visual Basic

กรอบแนวคิดการวิจัย

ในการศึกษาวิจัยครั้งนี้ นำเสนอตัวแบบการพยากรณ์ภาวะการมีงานทำของบัณฑิตด้วยเทคนิคการจำแนกข้อมูลด้วยกฎร่วมกับวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยบนชุดข้อมูลที่ไม่สมดุล (SMOTE) ซึ่งเป็นการสังเคราะห์ข้อมูลเพิ่มโดยอ้างอิงจากข้อมูลที่มีอยู่ สามารถสร้างความหลากหลายของข้อมูลได้และส่งผลดีต่อการสร้างแบบจำลองต่างจากวิธีการสุ่มลด (under sampling) ที่ทำการลดจำนวนข้อมูลซึ่งอาจไม่เหมาะกับการสร้างแบบจำลองที่ต้องการความหลากหลายของข้อมูล และวิธีการสุ่มเกิน (over sampling) ที่ทำการสุ่มข้อมูลที่มีอยู่แล้วเพิ่มขึ้นมาซึ่งอาจจะส่งผลดีกับข้อมูลที่มีลักษณะไม่ซับซ้อนมากเกินไป โดยมีกรอบแนวคิดแสดงดังภาพ 2

จากกรอบแนวคิดได้มีดำเนินการ 4 ขั้นตอน คือ

1) รวบรวมข้อมูล

2) การเตรียมข้อมูลสำหรับการทำเหมืองข้อมูล

3) พัฒนาตัวแบบพยากรณ์

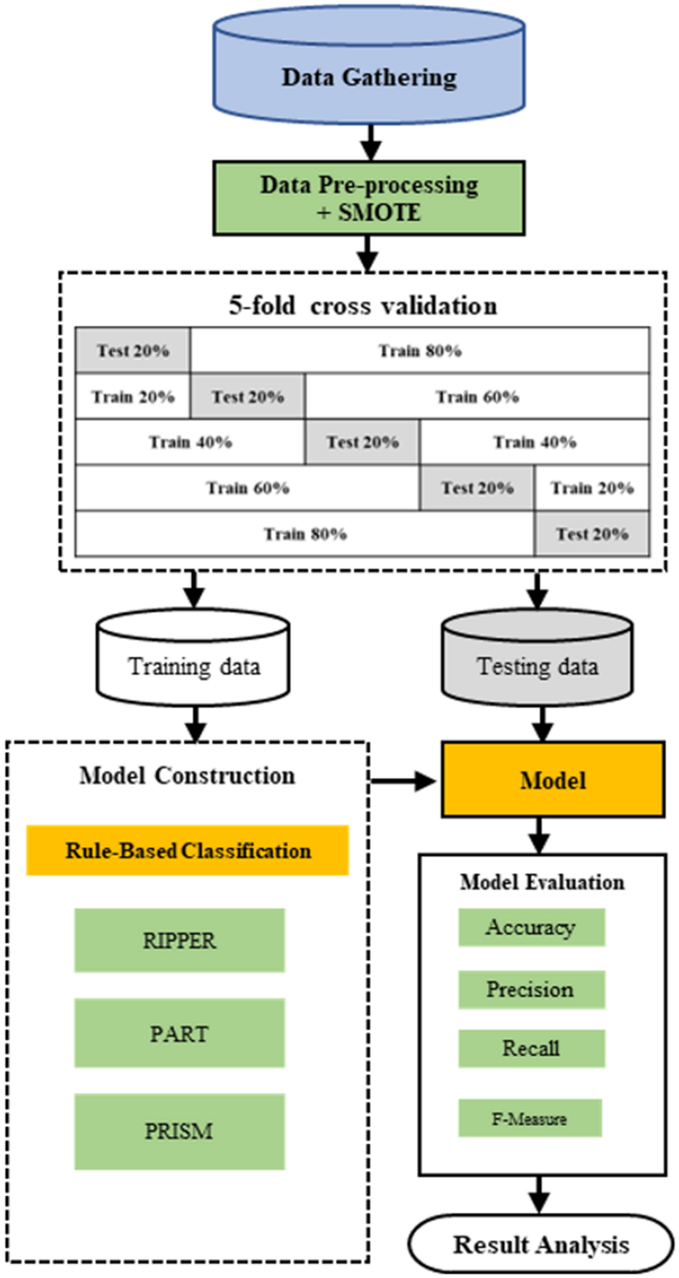
4) การตรวจสอบประสิทธิภาพของตัวแบบ

วิธีดำเนินการวิจัย

ตัวแบบการพยากรณ์ภาวะการมีงานทำของบัณฑิตด้วยเทคนิคการจำแนกข้อมูลด้วยกฎร่วมกับวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยบนชุดข้อมูลที่ไม่สมดุล (SMOTE) มีรายละเอียดการดำเนินการวิจัย ดังนี้

1) การรวบรวมข้อมูล

เป็นขั้นตอนที่เริ่มต้นจากการเก็บรวบรวมข้อมูลภาวะการมีงานทำที่จะนำมาสร้างตัวแบบ ซึ่งข้อมูลที่ในมาใช้ในการสร้างตัวแบบได้ทำการรวบรวมจากข้อมูลภาวะการมีงานทำของบัณฑิต มหาวิทยาลัยราชภัฏนครปฐม จำนวน 11,971 ระเบียน หลังจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูล บางส่วนมาใช้ในการวิเคราะห์ รายละเอียดข้อมูลที่ใช้ในการวิจัย คือ (1) ข้อมูลส่วนบุคคลและข้อมูลการศึกษา และ (2) ข้อมูลภาวะมีงานทำของบัณฑิต ดังตาราง 2 โดยตรวจสอบความสมบูรณ์ของข้อมูล ถึงจำนวนข้อมูล คุณลักษณะ ความเป็นไปได้ที่จะนำมาวิเคราะห์ความครบถ้วนในแต่ละข้อมูล



ภาพ 2 กรอบแนวคิดการวิจัย

ตาราง 2

รายละเอียดคุณลักษณะ

ลำดับ	ชื่อ	คำอธิบาย
1	Gender	เพศ
		F = หญิง
		M = ชาย
2	Major	สาขาวิชา ยึดสาขาวิชาตาม
		ระบบของ ISCED
		1 = โปรแกรมทั่วไปและคุณสมบัติ
		2 = การศึกษา
		3 = ศิลปะและมนุษย์
		4 = สังคมศาสตร์ วารสารศาสตร์
		และสารสนเทศ
		5 = การบริหารธุรกิจและกฎหมาย
		6 = วิทยาศาสตร์ ธรรมชาติ
		คณิตศาสตร์และสถิติ
		7 = เทคโนโลยีสารสนเทศและ
		การสื่อสาร
		8 = วิศวกรรม, อุตสาหกรรมและ
		การก่อสร้าง
		9 = เกษตรศาสตร์ วนศาสตร์
3	Faculty	การประมงและสัตวแพทย์
		10 = สุขภาพและสวัสดิการ
		คณะที่ศึกษา
		01 = กลุ่มสาขาวิทยาศาสตร์และ
		เทคโนโลยี
		02 = กลุ่มสาขาวิทยาศาสตร์สุขภาพ
		03 = กลุ่มสาขาสังคมศาสตร์และ
		มนุษยศาสตร์
4	Gpax	เกรดเฉลี่ยสะสม
		G1 = 0.00-2.00
		G2 = 2.01-2.50
		G3 = 2.51-3.00
		G4 = 3.01-3.50
		G5 = 3.51-4.00

ตาราง 2 (ต่อ)

ลำดับ	ชื่อ	คำอธิบาย
5	Honor	เกียรตินิยม
		First=เกียรตินิยมอันดับ 1
		Second=เกียรตินิยมอันดับ 2
		Non=สำเร็จการศึกษา
6	Hometown	ภูมิภาคของจังหวัดที่เป็น
		ภูมิลำเนา
		Bangkok=กรุงเทพมหานคร
		Central=ภาคกลาง
		North=ภาคเหนือ
		South=ภาคใต้
		Northeast=ภาคะวันออกเฉียงเหนือ
		Eastern=ภาคตะวันออก
		West=ภาคตตะวันตก
7	Jobstatus	สถานภาพการทำงาน
		Full=ทำงานประจำ
		Free=ทำงานอิสระ
8	Jobtype	ประเภทงานที่ทำ
		Gov=ข้าราชการ/เจ้าหน้าที่หน่วยงานรัฐ
		State=รัฐวิสาหกิจ
		Comp=พนักงานบริษัท/องค์กรธุรกิจ/
		เอกชน
		Free=เจ้าของกิจการ/อาชีพอิสระ
		Agri=เกษตรกรรม
		Etc=อื่น ๆ
9	Talent 1	ความสามารถพิเศษ1
		A=ภาษาต่างประเทศ
		B=การใช้คอมพิวเตอร์
		C=กิจกรรมสันทนาการ
		D=ศิลปะ/วัฒนธรรม
		E=นาฏศิลป์/ดนตรีขับร้อง
		F=การสื่อสาร
		G=การคำนวณ
		H=ไม่มี

ตาราง 2 (ต่อ)

ลำดับ	ชื่อ	คำอธิบาย
10	Talent 2	ความสามารถพิเศษ2
		A=ภาษาต่างประเทศ
		B=การใช้คอมพิวเตอร์
		C=กิจกรรมสันทนาการ
		D=ศิลปะ/วัฒนธรรม
		E=นาฏศิลป์/ดนตรีขับร้อง
		F=การสื่อสาร
		G=การคำนวณ
		H=ไม่มี
11	Talent 3	ความสามารถพิเศษ3
		A=ภาษาต่างประเทศ
		B=การใช้คอมพิวเตอร์
		C=กิจกรรมสันทนาการ
		D=ศิลปะ/วัฒนธรรม
		E=นาฏศิลป์/ดนตรีขับร้อง
		F=การสื่อสาร
		G=การคำนวณ
		H=ไม่มี
12	Workplace	ภูมิภาคจังหวัดที่ได้งานทำ
		Bangkok=กรุงเทพมหานคร
		Central=ภาคกลาง
		North=ภาคเหนือ
		South=ภาคใต้
		Northeast=ภาคะวันออกเฉียงเหนือ
		Eastern=ภาคตะวันออก
		West=ภาคตะวันตก
13	Salary	เงินเดือน/รายได้เฉลี่ยต่อเดือน
		Low=ต่ำกว่า 10,000
		Medium=10,001 – 30,000
		High=มากกว่า 30,001 ขึ้นไป
14	Match	งานที่ทำตรงกับสาขาที่เรียน
		Yes=ตรง No=ไม่ตรง

ตาราง 2 (ต่อ)

ลำดับ	ชื่อ	คำอธิบาย
15	Period	ระยะเวลาการได้งานทำ
		P0=ได้งานทันที
		P1=1–3 เดือน
		P2=4–6 เดือน
		P3=7–9 เดือน
		P4=10–12 เดือน
		P5=มากกว่า 1 ปี
		P6=ได้งานระหว่างศึกษา

จากตาราง 2 คุณลักษณะที่ 15 ระยะเวลาการได้งานทำ เป็นคุณลักษณะที่เป็นกลุ่มของข้อมูลที่ใช้ในการจำแนกข้อมูล โดยแบ่งออกได้ 7 กลุ่มได้แก่

P0=ได้งานทันที

P1=1–3 เดือน

P2=4–6 เดือน

P3=7–9 เดือน

P4=10–12 เดือน

P5=มากกว่า 1 ปี

P6=ได้งานระหว่างศึกษา

2) การเตรียมข้อมูล (Data Preparation)

ในขั้นตอนการเตรียมข้อมูลสำหรับทำเหมืองข้อมูลที่จะนำไปใช้ในการวิเคราะห์ ผู้วิจัยได้นำข้อมูลที่ได้จากการเก็บรวบรวมข้อมูลจากขั้นตอนก่อนหน้า ดังต่อไปนี้

2.1 การทำความสะอาดข้อมูลและแปลงข้อมูล เมื่อพิจารณาข้อมูลที่ได้แล้วนำข้อมูลมาทำการศึกษา พบว่า

- ข้อมูลมีค่าที่หายไป (missing value) กรณีระเบียนนั้นมีข้อมูลที่ขาดหายจำนวนมากได้ใช้วิธีตัดระเบียน (Ignore the tuple) เหล่านั้นออกไป และกรณีระเบียนข้อมูลขาดหายจำนวนเล็กน้อยแก้ไขด้วยการเติมเต็มข้อมูลด้วยการประมาณค่าข้อมูลใช้ค่าเฉลี่ยคุณลักษณะเติม

ค่าข้อมูลส่วนที่หายไป เช่น ข้อมูลภูมิภาคของจังหวัดที่เป็นภูมิลำเนา ที่ขาดหาย จากเดิม “ ” เปลี่ยนเป็น “ภาคกลาง” เป็นต้น

- ข้อมูลที่มีความผิดปกติ (outliers) ข้อมูลรบกวน (noisy data) ข้อมูลมีความไม่สอดคล้องกัน (inconsistent data) โดยจัดการให้มีความถูกต้องสมบูรณ์

- การแปลงข้อมูล ในขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน ในงานวิจัยครั้งนี้ ผู้จัดทำได้เลือกใช้โปรแกรม Weka มาใช้ในการสร้างตัวแบบซึ่งต้องจัดเตรียมแฟ้มงาน ให้อยู่ในรูปแบบไฟล์ arff ตัวอย่างดังภาพ 3 มาใช้ในงานวิจัย

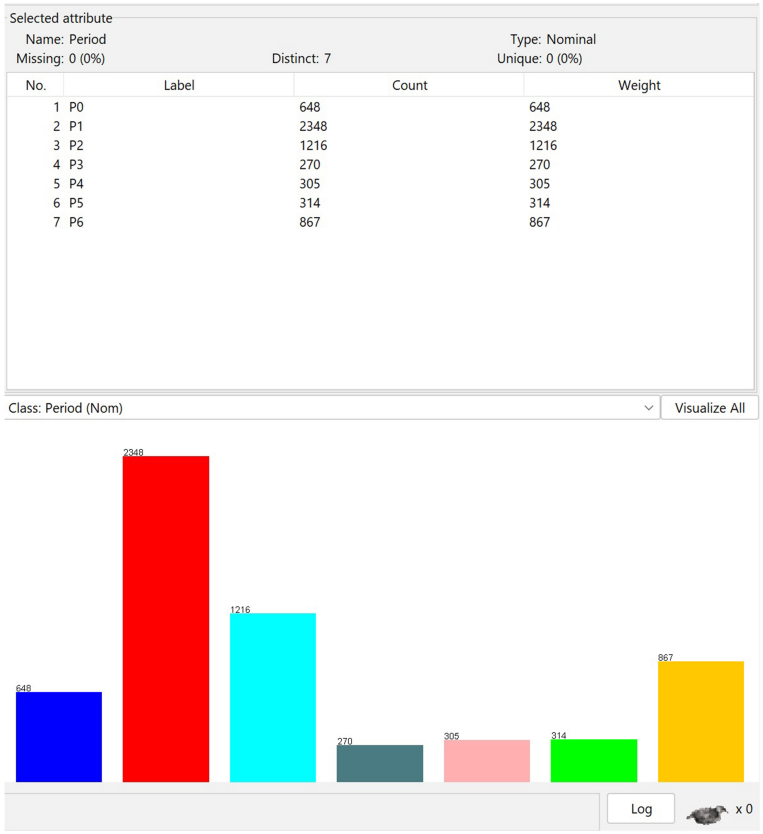
2.2 การสุ่มเพิ่มตัวอย่างกลุ่มน้อย SMOTE ซึ่งในงานวิจัยได้ทำการเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย SMOTE ใน class ที่มีจำนวนข้อมูลน้อย ได้แก่ class P0 P2 P3 P4 P5 และ P6 ตัวอย่างดังภาพ 4

หลังจากทำการสุ่มเพิ่มตัวอย่างกลุ่มน้อย ด้วยเทคนิค SMOTE ผลที่ได้ class ที่ทำการสุ่มเพิ่มตัวอย่างจะมีข้อมูลที่ไม่ต่างกันมาก ดังภาพ 5

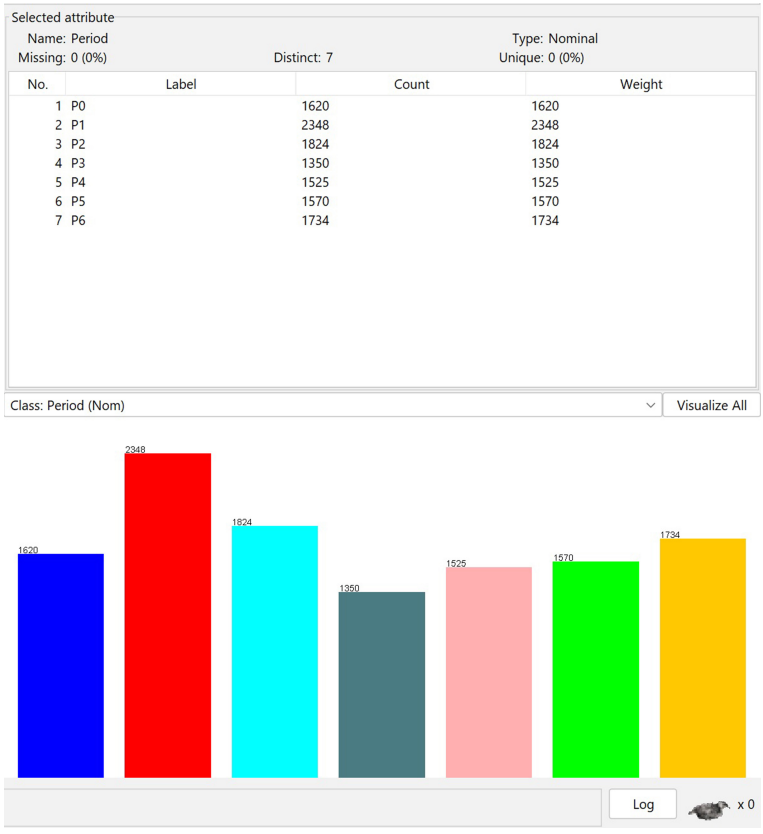
```
@relation npru-finnish

@attribute Gender {M,F}
@attribute Major {1,2,3,4,5,6,7,8,9,10}
@attribute Faculty {1,2,3}
@attribute Gpax {G4,G5,G2,G1,G3}
@attribute Honor {Non,Second,First}
@attribute Hometown {West,Central,Eastern,Bangkok,South,Northeast,North}
@attribute Jobstatus {Full,Free}
@attribute Jobtype {Comp,Free,Gov,Etc,State,Agri}
@attribute Talent1 {A,D,H,B,C,G,E,F}
@attribute Talent2 {H,C,B,F,G,D,E,A}
@attribute Talent3 {F,D,B,H,A,G,C,E}
@attribute Workplace {Central,Northeast,West,Bangkok,South,North,Eastern}
@attribute Salary {Medium,High,Low}
@attribute Match {Yes,No}
@attribute Period {P0,P1,P2,P3,P4,P5,P6}

@data
M,2,1,G4,Non,West,Full,Comp,A,H,F,Central,Medium,Yes,P0
F,3,2,G5,Non,Central,Full,Comp,D,H,D,Central,Medium,Yes,P0
F,4,1,G4,Non,Eastern,Full,Free,A,C,B,Central,High,No,P0
F,2,1,G4,Non,Bangkok,Full,Comp,H,H,H,Central,High,No,P0
M,2,1,G2,Non,West,Free,Free,B,B,H,Northeast,Medium,Yes,P0
F,6,1,G2,Non,Eastern,Full,Gov,C,F,A,West,High,No,P0
F,3,1,G4,Non,Central,Full,Comp,D,B,G,Bangkok,High,No,P0
M,1,1,G4,Non,South,Full,Etc,C,B,B,South,Medium,No,P0
F,4,1,G2,Non,Bangkok,Full,Comp,H,G,H,Central,Medium,No,P0
F,2,3,G1,Non,Bangkok,Full,Comp,H,F,H,Central,Medium,No,P0
F,6,1,G3,Non,West,Free,Comp,C,G,B,Bangkok,High,No,P0
F,3,3,G4,Second,Central,Free,State,H,F,H,Central,Medium,Yes,P0
M,4,3,G4,Non,Central,Free,Comp,C,H,A,Bangkok,High,Yes,P0
M,2,1,G3,Non,Central,Free,Comp,D,D,B,Central,Low,Yes,P0
M,4,2,G3,Non,West,Free,Free,H,H,C,Central,Medium,Yes,P0
F,8,3,G4,Non,Northeast,Full,Agri,B,E,C,South,High,No,P0
M,2,2,G2,Non,West,Free,Free,G,B,F,West,High,Yes,P0
```

Normal text file | length : 332,320 lines : 5,988 | Ln : 1 Col : 16 Pos : 16 | Unix (LF) | UTF-8 | IN

ภาพ 3 แปลงข้อมูลให้อยู่ในไฟล์ arff



ภาพ 4 แสดงจำนวนข้อมูลในแต่ละคลาส



ภาพ 5 จำนวนข้อมูลคลาสที่ได้ปรับด้วย SMOTE

3) การพัฒนาตัวแบบพยากรณ์ (modeling)

พัฒนาตัวแบบพยากรณ์ โดยได้ทำการเปรียบเทียบเพื่อหาเทคนิคในการคาดการณ์ภาวะการมีงานทำด้วยเทคนิคเหมืองข้อมูลที่เหมาะสมกับข้อมูลที่ให้ประสิทธิภาพที่ดีสุด การจำแนกข้อมูลด้วยกฎที่นำมาใช้เปรียบเทียบในงานวิจัย จำนวน 3 เทคนิค ได้แก่ RIPPER PART และ PRISM ซึ่งทำงานร่วมกับ เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย SMOTE

4) การตรวจสอบประสิทธิภาพตัวแบบ (evaluation)

ในการวัดประสิทธิภาพของตัวแบบในแต่ละเทคนิควิธีที่นำมาเปรียบเทียบ โดยใช้ค่าความถูกต้องของแบบจำลอง จะต้องทำการเลือกข้อมูลสำหรับเรียนรู้ และข้อมูลสำหรับทดสอบ งานวิจัยนี้เลือกใช้วิธีการสุ่มเลือกข้อมูลแบบความเที่ยงตรง 5 กลุ่ม (5-fold cross validation) เป็นการวัดประสิทธิภาพตัวแบบซึ่งจะจำแนกประเภทข้อมูลด้วยการแบ่งข้อมูลออกเป็น 5 ส่วนจำนวนเท่า ๆ กัน แบ่งเป็นข้อมูลสอน 4 ส่วนและอีก 1 ส่วนเพื่อทดสอบจะทำวนซ้ำสลับครบ 5 รอบ วิธีนี้เป็นที่นิยมเนื่องจากมีความน่าเชื่อถือ นอกจากนี้ยังเหมาะสมกับข้อมูลที่มีปริมาณไม่มาก

ผลการวิจัย

การพัฒนาตัวแบบพยากรณ์ภาวะการมีงานทำของบัณฑิตด้วยเทคนิคการจำแนกข้อมูลด้วยกฎร่วมกับวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยบนชุดข้อมูลที่ไม่สมดุล (SMOTE) ในงานวิจัยนี้ได้เลือกการจำแนกข้อมูลด้วยกฎที่นำมาใช้เปรียบเทียบในงานวิจัย จำนวน 3 เทคนิค ได้แก่ RIPPER PART และ PRISM ซึ่งทำงานร่วมกับ เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย SMOTE โดยการวัดประสิทธิภาพของตัวแบบจากค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความเหวี่ยง ซึ่งได้ทำการเปรียบเทียบข้อมูลแบบดั้งเดิมและข้อมูลที่ได้ปรับปรุงด้วยเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย SMOTE สรุปได้ดังตาราง 3

เทคนิค RIPPER ผลการวัดประสิทธิภาพอัลกอริทึม RIPPER โดยใช้ข้อมูลแบบดั้งเดิม ได้ดังนี้ ค่าความถูกต้องเท่ากับร้อยละ 52.19 ค่าความแม่นยำเท่ากับร้อยละ 58.70 ค่าความระลึกเท่ากับร้อยละ 52.20 และค่าความเหวี่ยงเท่ากับร้อยละ 48.70 และข้อมูลที่ปรับสมดุลข้อมูลด้วยวิธี SMOTE ได้ดังนี้ ค่าความถูกต้อง เท่ากับร้อยละ 73.74 ค่าความแม่นยำเท่ากับร้อยละ 72.70 ค่าความระลึกเท่ากับร้อยละ 73.70 และค่าความเหวี่ยงเท่ากับร้อยละ 73.10

เทคนิค PART ผลการวัดประสิทธิภาพอัลกอริทึม PART โดยใช้ข้อมูลแบบดั้งเดิม ได้ดังนี้ ค่าความถูกต้องเท่ากับร้อยละ 62.62 ค่าความแม่นยำเท่ากับร้อยละ 60.70 ค่าความระลึกเท่ากับร้อยละ 62.60 และค่าความเหวี่ยงเท่ากับร้อยละ 61.40 และข้อมูลที่ปรับสมดุลข้อมูลด้วยวิธี SMOTE ได้ดังนี้ ค่าความถูกต้อง เท่ากับร้อยละ 78.62 ค่าความแม่นยำเท่ากับร้อยละ 79.80 ค่าความระลึกเท่ากับร้อยละ 78.60 และค่าความเหวี่ยงเท่ากับร้อยละ 78.90

เทคนิค PRISM ผลการวัดประสิทธิภาพอัลกอริทึม PRISM โดยใช้ข้อมูลแบบดั้งเดิม ได้ดังนี้ ค่าความถูกต้องเท่ากับร้อยละ 73.84 ค่าความแม่นยำเท่ากับร้อยละ 73.60 ค่าความระลึกเท่ากับร้อยละ 73.00 และค่าความเหวี่ยงเท่ากับร้อยละ 73.10 และข้อมูลที่ปรับสมดุลข้อมูลด้วยวิธี SMOTE ได้ดังนี้ ค่าความถูกต้อง เท่ากับร้อยละ 85.69 ค่าความแม่นยำเท่ากับร้อยละ 85.60 ค่าความระลึกเท่ากับร้อยละ 85.70 และค่าความเหวี่ยงเท่ากับร้อยละ 85.60

จากการเปรียบเทียบประสิทธิภาพของอัลกอริทึม พบว่า ตัวแบบที่ได้จากอัลกอริทึม PRISM ร่วมกับการปรับสมดุลข้อมูลด้วยวิธี SMOTE มีประสิทธิภาพสูงที่สุด โดยมีค่าความถูกต้อง เท่ากับร้อยละ 85.69 ค่าความแม่นยำเท่ากับร้อยละ 85.60 ค่าความระลึกเท่ากับร้อยละ 85.70 และค่าความเหวี่ยงเท่ากับร้อยละ 85.60 และตัวอย่างกฎที่สกัดได้จากอัลกอริทึม PRISM ดังภาพ 6

ดังนั้น คณะผู้วิจัยจึงเลือกตัวแบบจากอัลกอริทึม PRISM ร่วมกับการปรับสมดุลข้อมูลด้วยวิธี SMOTE ไปพัฒนาตัวแบบพยากรณ์ภาวะการมีงานทำของบัณฑิตต่อไป

ตาราง 3

ผลการทดลองในแต่ละอัลกอริทึม (ร้อยละ)

ค่าประสิทธิภาพจากการทดสอบตัวแบบ				
Algo-rithm	Ac-cu-racy	Pre-cision	Re-call	F-measure
Rule-Based Classification				
RIPPER	52.19	58.70	52.20	48.70
PART	62.62	60.70	62.60	61.40
PRISM	73.84	73.60	73.00	73.10
Rule-Based Classification + SMOTE				
RIPPER	73.74	72.70	73.70	73.10
PART	78.62	79.80	78.60	78.90
PRISM	85.69	85.60	85.70	85.60

If Major = 10 and Talent3 = B and Hometown = North then P0
If Gpax = G5 and Major = 5 and Talent1 = F then P0
If Honor = First and Talent2 = A and Major = 5 then P1
If Honor = First and Major = 3 and Talent1 = C then P2
If Faculty = 1 and Talent2 = F and Gpax = G2 and Talent1 = H and Major = 6 then P3
If Major = 10 and Talent1 = G and Faculty = 3 and Gender = F then P4

ภาพ 6 ตัวอย่างกฎที่ได้จากอัลกอริทึม PRISM

การอภิปรายผล

จากการทดลองโดยใช้เทคนิคเหมืองข้อมูลในการพัฒนาระบบพยากรณ์ภาวะการมีงานทำของบัณฑิตและแก้ปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE พบว่า

การแก้ปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE สามารถเพิ่มประสิทธิภาพให้ตัวแบบเพิ่มขึ้นในทุกเทคนิคที่ได้นำมาเปรียบเทียบ เมื่อพิจารณาเป็นด้าน สามารถสรุปได้ว่า ค่าความถูกต้อง เพิ่มขึ้นเฉลี่ยร้อยละ 16.47 ค่าความแม่นยำเพิ่มขึ้นเฉลี่ยร้อยละ 15.03 ค่าความระลึกเพิ่มขึ้นเฉลี่ยร้อยละ 16.73 และค่าความเหวี่ยงเพิ่มขึ้นเฉลี่ยร้อยละ 18.13

ตัวแบบคาดการณ์จากเทคนิค PRISM เป็นตัวแบบที่มีประสิทธิภาพในการคาดการณ์สูงที่สุด ซึ่งเทคนิค PRISM เป็นเทคนิคที่มีจุดมุ่งหมายหลักเพื่อหลีกเลี่ยงการสร้างกฎที่ซับซ้อนที่มีเงื่อนไขซ้ำซ้อนจำนวนมาก และเมื่อได้ทำการปรับสมดุลของข้อมูลด้วยวิธี SMOTE แล้วทำให้ประสิทธิภาพการพยากรณ์เพิ่มขึ้นอีก โดยมีค่าความถูกต้อง เท่ากับร้อยละ 85.69 ค่าความแม่นยำเท่ากับร้อยละ 85.60 ค่าความระลึกเท่ากับร้อยละ 85.70 และค่าความเหวี่ยงเท่ากับร้อยละ 85.60 จึงเป็นตัวแบบที่เหมาะสมต่อการนำไปประยุกต์ใช้สำหรับพัฒนาระบบพยากรณ์ภาวะการมีงานทำของบัณฑิตต่อไป

เมื่อพิจารณาจากกฎที่ได้ ตัวอย่างเช่น กรณีที่ได้งานทำทัน จากตัวอย่างกฎที่ได้สามารถแปลได้ ดังนี้

- if major =10 and talent3 = b and Hometown = north then P0

ถ้า สาขาวิชา = สุขภาพและสวัสดิการ ความสามารถพิเศษ 3 = การใช้คอมพิวเตอร์ ภูมิลำเนา = ภาคเหนือ แล้ว ระยะเวลาการได้งานทำ คือ ได้งานทันที

- Gpax = g5 and Major =5 and Talent1 =f then P0

ถ้า เกรดเฉลี่ยสะสม = 3.51-4.00 สาขาวิชา = การบริหารธุรกิจและกฎหมาย ความสามารถพิเศษ 1 = การสื่อสาร แล้ว ระยะเวลาการได้งานทำ คือ ได้งานทันที

- if Hometown = West and Major = 1 and Jobtype = Comp and Gpax = G4 then P0

ถ้า ภูมิลำเนา = ภาคตะวันตก สาขาวิชา = โปรแกรมทั่วไป ประเภทงานที่ทำ = พนักงานบริษัท/องค์กรธุรกิจ/เอกชน เกรดเฉลี่ยสะสม = 3.01 - 3.50 แล้ว ระยะเวลาการได้งานทำ คือ ได้งานทันที

ข้อเสนอแนะ

การสร้างตัวแบบเพื่อพยากรณ์คำตอบหรือผลลัพธ์นั้น ปัจจัยที่มีความสำคัญซึ่งต้องคำนึงถึง คือ จำนวนข้อมูลของกลุ่มคำตอบหรือผลลัพธ์ต้องมีสัดส่วนใกล้เคียงกัน หรือที่เรียกว่าปัญหาข้อมูลไม่สมดุล แบ่งออกเป็น 2 อย่าง คือ ข้อมูลกลุ่มน้อย และข้อมูลกลุ่มมาก ซึ่งเป็นปัญหาที่เกิดขึ้นกับได้กับข้อมูลทุกประเภท เมื่อต้องการทำการจำแนกประเภทข้อมูลผลลัพธ์ที่ได้เมื่อทำการจำแนกข้อมูลจะถูก

จำแนกไปในกลุ่มข้อมูลกลุ่มมาก ซึ่งผู้วิจัยต้องหาเทคนิคหรือวิธีการเพื่อแก้ปัญหาเพื่อให้สามารถพยากรณ์คำตอบหรือผลลัพธ์ได้มีประสิทธิภาพ อย่างไรก็ตามตัวแบบคาดการณ์ที่ผู้วิจัยได้นำเสนอนั้นเป็นเพียงการใช้เทคนิคหนึ่งของการทำเหมืองข้อมูลจากหลากหลายเทคนิคที่มีอยู่ ในอนาคตหากมีการพัฒนางานวิจัยให้มีประสิทธิภาพมากขึ้นอาจจะนำเทคนิคอื่น ๆ และเพิ่มกลุ่มตัวอย่างจากมหาวิทยาลัยราชภัฏต่าง ๆ หรือมหาวิทยาลัยของรัฐในกลุ่มอื่นเช่น ราชมงคล หรือ มหาวิทยาลัยในกำกับของรัฐเพิ่มเติมมาทำการทดลองเปรียบเทียบเพื่อให้ได้ตัวแบบที่มีประสิทธิภาพมากยิ่งขึ้น

References

Amphawan, K. (2022). *Data mining*. Retrieved from https://staff.informatics.buu.ac.th/~komate/886464/%5B6%5D-Classification.pdf. (in Thai)

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmayer, W. P., (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligent Research, 16*(2002), 321-357. https://doi.org/10.1613/jair.953

Chinjoho, P., Jabjone S., & Jongmuanwai, B. (2021). Developing information technology model into future careers forecast of graduate information technology major in Nakhonratchasima Rajabhat University. *Journal of Technology Management Rajabhat Maha Sarakham University, 8*(2), 20–32. (in Thai)

Chiroma, F., Liu, H., & Cocea, M. (2018). Suiciderelated text classification with prism algorithm. *2018 International Conference on Machine Learning and Cybernetics (ICMLC), Chengdu, China, 2018,* (pp. 575-580). China: IEEE. doi: 10.1109/ICMLC.2018.8527032.

Hevo Data Inc. (2024). *Rule based data mining classifier: A comprehensive guide 101*. Retrieved from https://hevodata.com/learn/rule-based-data-mining/

Li, X., & Liu, B. (2014). *Rule-based classification*. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8557fde1e865f0dcc209468ccaf94f12a04b7835.

Kijsirikul, B. (2002). *Data mining algorithm*. Bangkok: Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand. (in Thai)

Kladkaew, S. (2014). *Decision support systems for selecting careers in accordance with the ability of the graduates* (Master's thesis). Ramkhamhaeng University. Bangkok (in Thai)

Markoulidakis, J., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multiclass confusion matrix reduction method and Its application on net promoter score classification problem. *Technologies 2021, 9*, 81. doi: 10.3390/technologies9040081

Mohamed, W. N. H. W., Salleh, M. N. M., & Omar, A. H. (2012). A comparative study of reduced error pruning method in decision tree algorithms. *2012 IEEE International Conference on Control System, Computing and Engineering, Penang, Malaysia, 2012* (pp. 392-397). Malaysia: IEEE. doi: 10.1109/ICCSCE.2012.6487177.

Thakur, S. (2020). Detection of malicious URLs in big data using RIPPER algorithm. *International Journal of Creative Research Thoughts (IJCRT), 8*(4), 767-771. https://ijcrt.org/papers/IJCRT2004096.pdf

Vilailuck, S., Jaroenpuntaruk, V., & Wichadakul, D. (2015). Utilizing data mining techniques to forecast student academic achievement of Kasetsart University Laboratory School Kamphaeng Saen Campus Educational Research and Development Center. *Veridian E-Journal, Science and Technology Silpakorn University, 2*(2), 1-17. (in Thai)

Wanon, S., Arreerard T., & Sanrach, C. (2018). A study of techniques in predicting career counseling for undergraduate students of the computer program by using data mining technique. *Journal of Technology Management Rajabhat Maha Sarakham University, 5*(1), 164–171. (in Thai)